Journal of the Korean Data & Information Science Society
2016, **27**(4), 899–910

http://dx.doi.org/10.7465/jkdi.2016.27.4.899
한국데이터정보과학회지

특정 범주에 대한 평가자간 카파 일치도의 퍼뮤테이션 p값

엄용환[1]

[1]성결대학교 산업경영공학부

접수 2016년 3월 21일, 수정 2016년 6월 7일, 게재확정 2016년 7월 13일

요 약

근사검정은 종종 표본이 작은 순서척도의 범주를 갖는 분할표를 분석할 때 그 p값이 과대추정 되거나 과소추정 되기 때문에 적절하지 못한 것으로 여겨진다. 본 논문에서는 순서화된 범주를 갖는 $k \times k$ 분할표에서 특정 범주에 대한 가중 일치도에 대해 정확한 p값과 재표본 기법에 의해 p값을 구하는 퍼뮤테이션 방법을 제시한다. 이를 위해 두 명의 평가자가 특정의 범주에서 얼마나 일치된 평가를 하는 지를 측정하기 위해 Kvålseth가 제안한 특정 범주에 대한 가중 일치도 (weighted specific-category kappa)를 사용한다. 사례 데이터로서 3×3 분할표 형태의 실제 데이터와 가상데이터 그리고 4×4 분할표 형태의 가상데이터를 이용하며, 정확한 퍼뮤테이션 p값과 재표본 퍼뮤테이션 p값 그리고 근사검정의 p값을 계산하여 비교한다.

주요용어: 분할표, 특정 범주에 대한 가중 일치도, 퍼뮤테이션, p값.

1. 서론

두 명 이상의 평가자들이 동일한 대상에 대해 평가할 때, 이 평가가 얼마나 일치하는지를 나타내는 일치도 (measure of agreement)는 심리학, 교육학, 사회학 등의 사회과학 분야에서 연구되어 온 중요한 통계적 관심사로서 오늘날에는 의학, 정보기술 분야에까지 그 활용도가 커지고 있다. 두 명의 평가자 사이의 일치도를 측정하는 일치도 kappa (unweighted kappa)가 Cohen (1960)에 의해 처음 소개되었고 그 후 1968년에는 순서척도의 범주형 데이터에 대해 가중치를 부여하여 일치도를 측정하는 weighted kappa가 제시되었다 (Cohen, 1968). Weighted kappa는 평가자간의 평가가 일치하지 않을 때 이 불일치하는 정도를 반영하는 일치도이며 임상평가, 검사-재검사 신뢰도, 사회연구 등의 분야에서 널리 사용되고 있다. 이 후 weighted kappa와 unweighted kappa는 여러 연구자들에 의해 활발히 연구가 진행되어 왔으며 (Kraemer, 1983; Upton과 Cook, 2002; Shoukri, 2004) 일치도로서 kappa가 갖고 있는 몇가지 결함에도 불구하고 여전히 널리 사용되고 있다 (Feinstein과 Cicchetti, 1990; Oleckno, 2008; Zhao, 2011; Han과 Park, 2012)

Weighted kappa와 unweighted kappa는 모든 범주에 걸쳐서 평가자들이 전체적으로 일치하는 정도를 측정하는 반면에 어느 특정 범주에 대해서 평가자들이 얼마나 일치하는지를 평가하는 것도 중요한 관심사가 되고 있다. 이를 위하여 Spitzer 등 (1967)과 Fleiss (1981)는 kappa에 기초한 일치도 specific-category kappa를 제안하였는데, 이 일치도는 순서척도의 범주형 데이터에서 평가대상을 k개의 범주 중에 어느 한 범주에 속하는 것으로 평가할 때 특정 범주에서 평가자들의 일치도를 나타낸다. Spitzer 등은 이 kappa를 얻기 위해 $k \times k$ 분할표로 요약된 데이터를 2×2 분할표로 병합하였는데 이 때 2×2

[1] (14097) 경기도 안양시 만안구 성결대학로 53, 성결대학교 산업경영공학부, 교수,
Email: uyh@sungkyul.ac.kr

분할표는 관심있는 범주 s와 다른 k - 1개의 범주들을 연합하여 만든 범주로 구성된 표이다. 그러나 이 kappa는 가중치를 부여하지 않은 일치도이기 때문에 이를 보완하기 위해 Kvålseth (1989)는 평가자들의 불일치 정도에 따라서 가중치를 다르게 부여하여 일치도를 측정하는 특정 범주에 대한 가중 일치도 (weighted specific-category kappa)를 제시하였다. Kvålseth는 이 kappa를 측정하기 위해 Fleiss와 달리 범주들을 병합하는 대신 $k \times k$ 분할표를 그대로 사용하였다. Kvålseth이 제시한 특정 범주에 대한 가중 일치도의 자세한 것은 본 논문의 제 2절에서 소개한다.

본 연구는 퍼뮤테이션 검정 (permutation test)을 이용하여 Kvålseth가 제안한 특정 범주에 대한 두 평가자간의 가중 일치도에 대해서 p값과 경험적인 분위수 한계 (empirical quantile limit)를 산출하고 이것을 전통적인 근사검정의 결과와 비교한다. 퍼뮤테이션 검정은 Fisher (1935)가 최초로 제안한 이후로 꾸준히 발전해 왔는데, 표본의 크기가 작거나 근사적인 분포에 기초한 p값이 정확하지 않을 때 널리 사용된다. 특히 표본이 작을 때 퍼뮤테이션 검정은 큰 표본에 대해 전통적으로 사용되는 근사적인 검정보다 더 정확한 p값을 제공하는 것으로 알려져 있다 (Holms, 1979, 1990; Mielke와 Berry, 2001; Good, 2000, 2001). 퍼뮤테이션 검정이 갖는 장점은 첫째 퍼뮤테이션 검정은 데이터에 의존하는데 이것은 분석에 필요한 모든 정보가 관찰된 데이터 안에 포함되어 있다는 뜻이다, 둘째로 퍼뮤테이션 검정은 모집단에 대한 정규성을 비롯한 어떤 이론적인 분포를 가정하지 않으며 셋째로 퍼뮤테이션 검정은 이산형의 퍼뮤테이션 분포에 기초하여 p값을 제시하고 비임의 표본 (nonramdom sample)에 대해서도 사용될 수 있다는 것이다. 퍼뮤테이션 방법의 자세한 내용은 제 3절에서 소개한다.

2. Weighted kappa and weighted specific kappa

두 명의 평가자가 n명의 평가 대상을 순서척도로 이루어진 k개의 범주 중 어느 한 범주로 분류할 때, 이들의 평가 결과는 $k \times k$ 분할표로 요약된다. 여기서 두 평가자가 각각 대상들을 i번째 범주와 j번째 범주로 분류한 대상들의 빈도수를 n_{ij}, 첫 번째 평가자에 의해 i번째 범주로 분류된 대상들의 수를 $n_{i.}$, 두 번째 평가자에 의해 j번째 범주로 분류된 대상들의 수를 $n_{.j}$라 하자 (i와 $j = 1, 2, \cdots, k$). 그러면 Cohen (1968)의 weighted kappa 일치도 KW는 다음과 같이 정의된다.

$$KW = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i.}p_{.j}}, \tag{2.1}$$

여기서 $p_{ij} = n_{ij}/n$, $p_{i.} = n_{i.}/n$, $p_{.j} = n_{.j}/n$, w_{ij}는 두 평가자들이 불일치하는 정도를 나타내는 가중치이고 $\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{ij}$는 관찰된 평가자간의 일치비율, $\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i.}p_{.j}$은 귀무가설 ($H_0$: $p_{ij} = p_{i.}p_{.j}$, i, $j = 1, \cdots, k$)하에서 두 평가자간의 평가가 우연히 일치하게 되는 비율을 의미한다. 가중치 w_{ij}는 0과 1 사이의 값이며 i= j이면 $w_{ij} = 0, i \neq j$이면 w_{ij} >0이다. 가중치로서 널리 사용되는 것은 $w_{ij} = |i - j|$ (Agresti, 2002)와 $w_{ij} = (i - j)^2$ (Cohen, 1968), 그리고 $w_{ij} = (i - j)^2/(k - 1)^2$ (Fleiss와 Cohen, 1973)와 $w_{ij} = |i - j|/(k - 1)$ (Cicchetti와 Allison, 1971)이 있다. 여기서 k는 상수이므로 가중치 $|i-j|$와 $(i-j)^2$는 각각 $|i-j|/(k-1), (i-j)^2/(k-1)^2$와 동일한 가중치가 된다. 만일 $i \neq j$일 때 w_{ij}=1 이고 $i = j$일 때 w_{ij}=0이면 weighted kappa는 unweighted kappa가 된다. 본 논문의 일치도 연구를 위해서는 일차형 가중치 $w_{ij} = |i-j|/(k-1)$와 이차형 가중치 $w_{ij} = (i-j)^2/(k-1)^2$를 사용하였다.

한편 Kvålseth (1989)는 가중치가 w_{ij} ($0 \leq w_{ij} \leq 1, i = j$이면 w_{ij}=0) 일 때 특정 범주 s ($s = 1, \cdots, k$)에 대해 평가자들의 일치하는 정도를 나타내는 특정 범주에 대한 가중 일치도 (weighted

specific-category kappa) KW_s를 다음과 같이 제안하였다.

$$KW_s = 1 - \frac{AW_s}{BW_s}, \tag{2.2}$$

여기서 $AW_s = \sum_{i=1}^{k} w_{is}p_{is} + \sum_{j=1}^{k} w_{sj}p_{sj}$, $BW_s = p_{s.}(\sum_{i=1}^{k} w_{si}p_{.i}) + p_{.s}(\sum_{j=1}^{k} w_{js}p_{j.})$이고, AW_s와 BW_s는 각각 평가자들에 의해 특정 범주 s로 분류된 대상들에 대해 관찰된 불일치 정도와 우연히 불일치하는 정도를 나타낸다. 여기서 말하는 불일치는 $k \times k$ 분할표에서 s번째 행과 s번째 열에 있는 모든 셀들 중에서 셀 (s, s)를 제외한 나머지 셀들에 해당되는 불일치이다. 예를 들면 k= 3일 경우 범주 s=1에 대한 평가자들의 불일치가 발생하는 셀들은 (1,2), (1,3), (2,1), (3,1)이고, s=2인 경우의 불일치 셀들은 (2,1), (2,3), (1,2), (3,2), s=3일 때의 셀은 (3,1), (3,2), (1,3), (2,3)이 된다. 그리고 s=1일 때의 일치도 KW_1= 1 - AW_1/BW_1이고 AW_1= $w_{21}p_{21} + w_{31}p_{31} + w_{12}p_{12} + w_{13}p_{13}$, BW_1= $w_{12}p_{1.}p_{.2} + w_{13}p_{1.}p_{.3} + w_{21}p_{2.}p_{.1} + w_{31}p_{3.}p_{.1}$이다. KW_s값은 1부터 0 그리고 음의 값까지 가질 수 있는데, 범주 s에 대해 관찰된 불일치 정도가 0일 때 (즉, AW_s= 0 일 때) KW_s= 1이 되고 관찰된 불일치 정도가 우연한 불일치 정도와 같을 때 (즉, 모든 i에 대해 $p_{si} = p_{s.}p_{.i}$이고 모든 j에 대해 $p_{js} = p_{j.}p_{.s}$일 때)는 KW_s= 0이며 관찰된 불일치 정도가 우연한 불일치 정도를 초과할 때 KW_s <0이 된다.

또한 KW_s는 n이 클 때 근사적으로 정규분포를 따르는 것으로 밝혀졌다 (Kvålseth, 2003).

$$KW_s \approx N(E(KW_s),\ \sigma^2)$$

여기서 KW_s의 분산 σ^2의 추정치 $\widehat{\sigma}^2$는 다음과 같이 주어진다.

$$\widehat{\sigma}^2 = \frac{1}{nD^2}\left(\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}C_{ij}^2 - [KW_s - (1-D)(1-KW_s)]^2\right) \tag{2.3}$$

그리고

$$C_{ij} = 1 - w_{ij} - \left(2 - \sum_{j=1}^{k} w_{ij}p_{.j} - \sum_{i=1}^{k} w_{ij}p_{i.}\right)(1 - KW_s)$$

$$D = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i.}p_{.j}$$

이다. 따라서 예를 들면, KW_s에 대한 근사적인 95% 신뢰구간은 $KW_s \pm 1.96 \times \widehat{\sigma}$로 주어지고 $E(KW_s)$ = 0에 대한 가설검정은 근사적으로 표준정규분포를 따르는 Z를 이용한다.

$$Z = \frac{KW_s}{\widehat{\sigma}}. \tag{2.4}$$

이 때 우측검정과 좌측검정에 대한 p값은 각각 우측검정 p값 $\doteq P(Z \geq KW_s^0/\widehat{\sigma})$, 좌측검정 p값 $\doteq P(Z \leq KW_s^0/\widehat{\sigma})$이며, KW_s^0은 관찰된 weighted specific-category kappa를 의미한다.

3. 퍼뮤테이션 검정

두 명의 평가자가 독립적으로 n명의 대상을 순서척도로 이루어진 k개의 범주 중 어느 한 범주로 평가할 때 얻어지는 $k \times k$ 분할표에서 퍼뮤테이션 검정은 주변 합계 (marginal frequency totals)가 일정하다는 조건하에서 n명의 대상을 k^2셀들에 배정하는 모든 가능한 배열들을 고려한다. 이때 이 배열들

을 모두 다 이용하여 검정통계량을 계산할 때 정확한 퍼뮤테이션 검정 (exact permutation test)이라 하고 이 배열들의 수가 매우 클 경우에는 전체 배열들 중에서 L개의 배열만을 복원 추출하여 검정통계량을 계산하는 것을 재표본 퍼뮤테이션 검정 (resampling permutation test)라 한다. 보통 검정의 정확도를 높이기 위해 L=1,000,000을 사용한다 (Johnston 등, 2007). Mielke와 Berry (2001)는 분할표에서 주변 합계가 주어질 때 생성될 수 있는 배열의 수를 제시하였다. 예를 들면, $k = 3$, $n_{1.}$=16, $n_{2.}$=14, $n_{.1}$=15, $n_{.2}$=10일 때 모든 가능한 배열의 수 = 5,225이고 이 때는 정확한 검정과 재표본 검정이 모두 가능하지만 $n_{1.}$=338, $n_{2.}$=325, $n_{.1}$=337, $n_{.2}$=329 일 때 가능한 배열의 수 = 1,504,687,715은 매우 큰 수이므로 재표본에 의한 퍼뮤테이션 검정을 이용하는 것이 실제적이라 하겠다 (Mielke와 Berry, 2001). 이차 분할표 (two-way contingency table)로부터 임의로 셀 빈도수의 배열을 생성하는 재표본 알고리즘은 Patefield (1981)에 의해 소개된 바 있다.

각 셀 빈도수의 배열이 주어지면 검정통계량 KW_s와 귀무가설 하에서 정확한 확률 $p(n_{ij}|n_{i.}, n_{.j})$이 다음과 같이 계산된다 (Mielke와 Berry, 2001).

$$p(n_{ij}|n_{i.}, n_{.j}) = \frac{\left(\prod_{i=1}^{k} n_{i.}!\right)\left(\prod_{j=1}^{k} n_{.j}!\right)}{n! \prod_{i=1}^{k}\prod_{j=1}^{k} n_{ij}!} \tag{3.1}$$

이 때 만일 관찰된 검정통계량을 KW_s^0라 하면 정확한 우측감정의 p값은 KW_s^0보다 크거나 같은 KW_s에 대한 $p(n_{ij}|n_{i.}, n_{.j})$들의 합이고 정확한 좌측감정의 p값은 KW_s^0보다 작거나 같은 KW_s에 대한 $p(n_{ij}|n_{i.}, n_{.j})$들의 합으로 계산한다. 즉, 전체 배열의 수 = M 일 때

$$\text{우측검정 } p\text{값} = \sum_{m=1}^{M} \Phi_m(KW_s) p(n_{ij}|n_{i.}, n_{.j}), \quad \text{여기서 } \Phi_m(KW_s) = \begin{cases} 1 & KW_s \geq KW_s^0 \text{ 일때} \\ 0 & \text{아닐때} \end{cases}$$

$$\text{좌측검정 } p\text{값} = \sum_{m=1}^{M} \Phi_m(KW_s) p(n_{ij}|n_{i.}, n_{.j}), \quad \text{여기서 } \Phi_m(KW_s) = \begin{cases} 1 & KW_s \leq KW_s^0 \text{ 일때} \\ 0 & \text{아닐때} \end{cases}$$

이고 각각의 n_{ij}는 m번째 배열에 따라 정해지는 셀 빈도수이다. 우측검정의 p값이 작으면 일치도가 높은 것이고 좌측검정의 p값이 작으면 불일치도가 높은 것을 의미한다. 마찬가지로 재표본에 의한 우측검정의 p값은 KW_s^0보다 크거나 같은 KW_s들의 비율이고, 좌측검정의 p값은 KW_s^0보다 작거나 같은 KW_s들의 비율로 계산된다.

$$\text{우측검정 } p\text{값} = \frac{1}{L}\sum_{m=1}^{L} \Phi_m(KW_s), \quad \text{여기서 } \Phi_m(KW_s) = \begin{cases} 1 & KW_s \geq KW_s^0 \text{ 일때} \\ 0 & \text{아닐때} \end{cases}$$

$$\text{좌측검정 } p\text{값} = \frac{1}{L}\sum_{m=1}^{L} \Phi_m(KW_s), \quad \text{여기서 } \Phi_m(KW_s) = \begin{cases} 1 & KW_s \leq KW_s^0 \text{ 일때} \\ 0 & \text{아닐때} \end{cases}$$

또한 퍼뮤테이션 검정은 모집단 분포에 의존하지 않기 때문에 전통적인 방법으로 $1-\alpha$ 신뢰구간을 구할 수 없으나 $1-\alpha$ 경험적인 분위수 한계 (empirical quantile limits)를 구할 수 있다. 먼저 재표본 기법을 사용하여 얻은 L개의 KW_s를 작은 값에서 큰 값으로 정렬한 후 하한값 ($Q_{\alpha/2}$)과 상한값

$(Q_{1-\alpha/2})$을 구한다. 이 $Q_{\alpha/2}$과 $Q_{1-\alpha/2}$은 각각 $(KW_s)_1, (KW_s)_2, \cdots, (KW_s)_L$에 대응되는 순서통계량을 $T_1 \le T_2 \le \cdots \le T_L$이라 할 때 다음과 같이 주어진다.

$$Q_\alpha/2 = T_{\text{최대}[1,\text{정수}[(\alpha/2)L+0.5]]}$$
$$Q_{1-\alpha/2} = T_{\text{최소}[L,\text{정수}((1-\alpha/2)L+0.5)]}$$

여기서 정수$[(\alpha/2)L+0.5]$은 괄호([ ])안의 계산결과에서 정수만을 취한다는 의미이다.

4. 예제 데이터

퍼뮤테이션 검정에 의한 p값 계산을 예시하기 위하여 세 개의 예제 데이터를 사용하였다. 이 데이터들은 모두 2명의 평가자가 평가 대상들을 3개 또는 4개의 범주로 분류하여 얻은 결과이며 특정 범주에 대한 가중 일치도 (weighted specific-category kappa)를 계산하기 위하여 일차형 가중치 $w_{ij} = |i-j|/(k-1)$와 이차형 가중치 $w_{ij} = (i-j)^2/(k-1)^2$를 사용하였고 추가적으로 가중치 없는 특정 범주에 대한 비가중 일치도 (unweighted specific-category kappa)를 계산하였다. 정확한 퍼뮤테이션 검정과 재표본 퍼뮤테이션 검정에 의한 p값과 근사적인 검정에 의한 p값을 비교하였으며 이 계산을 위해 퍼뮤테이션 관련 R 프로그램의 패키지를 사용하였다 (Kim과 Lee, 2014).

4.1. 예제 데이터 1 (실제 데이터)

Table 4.1은 Berry 등 (2006)의 연구에서 사용된 데이터로서 n=41명의 평가 대상을 두 명의 평가자가 k = 3개의 범주 (high, medium, low)로 분류한 결과이다. 주변 빈도수가 고정되어 있을 때 셀 빈도수의 가능한 배열의 수는 단지 M=5,225 이므로 정확한 퍼뮤테이션 검정으로 충분하지만 L=1,000,000 개의 배열을 이용한 재표본 퍼뮤테이션 검정도 함께 실시하였다. Table 4.1의 데이터와 각각의 가중치를 이용하여 KW_1, KW_2, KW_3퍼뮤테이션 검정의 p값, 근사검정에 의한 p값과 분위수 한계 ($Q_{0.025}$, $Q_{0.975}$) 그리고 식 (2.2)에 의해 계산한 분산 ($\widehat{\sigma}^2$)과 근사적인 95% 신뢰구간을 얻었다 (Table 4.2∼Table 4.8, Figure 4.1). Table 4.2와 Table 4.3에서 가중치가 이차형일 때 KW_1= 0.3014, KW_2=-0.1648, KW_3=0.3295이고 KW_1에 대응되는 정확한 우측검정의 p값 = 0.0483, 재표본 우측검정의 p값 = 0.0483, 근사 우측검정의 p값 = 0.0561 이고 $Q_{0.025}$= -0.3574, $Q_{0.975}$= 0.3612, $\widehat{\sigma}^2$(KW_1) = 0.0345, 95% 신뢰구간 = [-0.0737, 0.6765] 이다. 그리고 KW_2에 대응되는 정확한 우측검정의 p값 = 0.7537, 재표본 우측검정의 p값 = 0.7536, 근사 우측검정의 p값 = 0.8578, $Q_{0.025}$= -0.2813, $Q_{0.975}$= 0.3011, $\widehat{\sigma}^2(KW_2)$ = 0.0231, 95% 신뢰구간 = [-0.4717, 0.1422] 이며 KW_3에 대응되는 정확한 우측검정의 p값 = 0.0319, 재표본 우측검정의 p값 = 0.0317, 근사 우측검정의 p값 = 0.0408, $Q_{0.025}$= -0.3816, $Q_{0.975}$= 0.3702, $\widehat{\sigma}^2(KW_3)$ = 0.0340, 95% 신뢰구간 = [-0.0433, 0.7024] 이다. 따라서 평가자들의 일치도는 세 개의 범주 중 세 번째 범주에서 가장 높게 나타냈고, 두 번째 범주에서의 일치도가 음의 값을 가짐으로 두 번째 범주에서는 평가자들의 의견이 우연히 일치하게 되는 것보다 일치 정도가 작다는 것을 보여준다.

KW_1, KW_2와 KW_3의 각각에 대한 퍼뮤테이션 검정에서는 정확한 p값과 재표본 p값은 매우 유사한 값을 보이나 근사검정의 p값은 이보다 큰 값을 보이고 있다.

가중치가 일차형일 경우 (Table 4.4, Table 4.5)와 가중치를 사용하지 않은 경우 (Table 4.6, Table 4.7)에도 가중치가 이차형일 때와 비슷한 결과를 나타내었으며 근사검정의 p값은 퍼뮤테이션 검정의 p값 보다 크거나 작게 나타났다.

Table 4.1 Example 3×3 real data set

Judge 1	Judge 2		
	High	Medium	Low
High	7	6	3
Medium	5	2	7
Low	3	2	6

Table 4.2 Exact, resampling and asymptotic p-values for example 1 with quadratic weight

Agreement	Observed value	Exact p-value	Resampling p-value	Asymptotic p-value
KW_1	0.3014	0.0483	0.0483	0.0561
KW_2	-0.1648	0.7537	0.7536	0.8578
KW_3	0.3295	0.0319	0.0317	0.0408

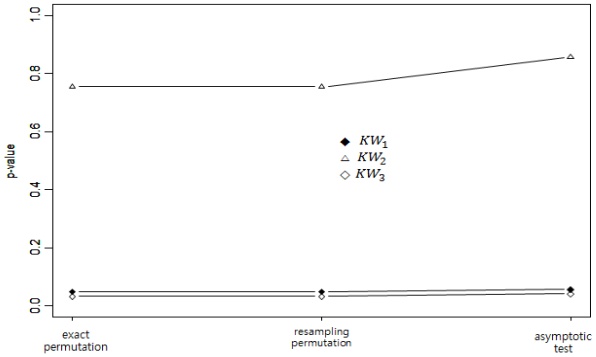

Figure 4.1 Exact, resampling and asymptotic p-values for example 1 with quadratic weight

Table 4.3 Quantile limits and 95% confidence interval for example 1 with quadratic weight

Agreement	Quantile Limits ($Q_{0.025}$, $Q_{0.975}$)	Variance	95% Confidence Interval
KW_1	(-0.3574, 0.3612)	0.0345	(-0.0737, 0.6765)
KW_2	(-0.2813, 0.3011)	0.0231	(-0.4717, 0.1422)
KW_3	(-0.3816, 0.3702)	0.0340	(-0.0433, 0.7024)

Table 4.4 Exact, resampling and asymptotic p-values for example 1 with linear weight

Agreement	Observed value	Exact p-value	Resampling p-valuee	Asymptotic p-value
KW_1	0.2219	0.0780	0.0780	0.0941
KW_2	-0.1648	0.7537	0.7536	0.8578
KW_3	0.2679	0.0651	0.0652	0.0537

Table 4.5 Quantile limits and 95% confidence interval for example 1 with linear weight

Agreement	Quantile Limits ($Q_{0.025}$, $Q_{0.975}$)	Variance	95% Confidence Interval
KW_1	(-0.3193, 0.3234)	0.0275	(-0.1132, 0.5571)
KW_2	(-0.2813, 0.3011)	0.0231	(-0.4717, 0.1422)
KW_3	(-0.3248, 0.3376)	0.0264	(-0,0608 0.5965)

Table 4.6 Exact, resampling and asymptotic p-values for example 1 without weight

Agreement	Observed value	Exact p-value	Resampling p-value	Asymptotic p-value
KW_1	0.1188	0.3322	0.3318	0.2253
KW_2	-0.1648	0.7537	0.7536	0.8578
KW_3	0.1854	0.0563	0.0563	0.1123

Table 4.7 Quantile limits and 95% confidence interval for example 1 without weight

Agreement	Quantile Limits ($Q_{0.025}$, $Q_{0.975}$)	Variance	95% Confidence Interval
KW_1	(-0.2958, 0.3262)	0.0243	(-0.1964, 0.4341)
KW_2	(-0.2813, 0.3011)	0.0231	(-0.4717, 0.1422)
KW_3	(-0.2490, 0.2940)	0.0226	(-0,1184, 0.4892)

4.2. 예제 데이터 2 (가상 데이터)

Table 4.8은 Johnston 등 (2008)의 연구에서 사용된 가상 데이터로서 n=990의 평가 대상을 k = 3개의 범주로 평가한 자료이다. 주변 빈도수가 고정되어 있을 때 셀 빈도수의 가능한 배열의 수 M=1,504,687,715는 매우 큰 값이므로 L=1,000,000개의 배열을 이용한 재표본 퍼뮤테이션 검정만을 실시하였다. Table 4.9와 Table 4.10에서 가중치가 이차형일 때 KW_1= 0.0348, KW_2=0.0502, KW_3=0.0351이므로 평가자들의 일치도는 두 번째 범주에서 가장 높게 나타냈고 첫 번째 범주에서의 일치도가 가장 낮았다. 또한 KW_1에 대응되는 재표본 우측검정의 p값 = 0.1913, 근사 우측검정의 p값 = 0.1879이고 $Q_{0.025}$= -0.0768, $Q_{0.975}$= 0.0771, $\widehat{\sigma}^2(KW_1)$ = 0.001545, 95% 신뢰구간 = [-0.0423, 0.1120]이다. 그리고 KW_2에 대응되는 재표본 우측검정의 p값 = 0.0494, 근사 우측검정의 p값 = 0.0572, $Q_{0.025}$= -0.0640, $Q_{0.975}$= 0.0639, $\widehat{\sigma}^2(KW_2)$ = 0.001010, 95% 신뢰구간 = [-0.0122, 0.1126]이며 KW_3에 대응되는 재표본 우측검정의 p값 = 0.1895, 근사 우측검정의 p값 = 0.1863, $Q_{0.025}$= -0.0773, $Q_{0.975}$= 0.0769, $\widehat{\sigma}^2(KW_3)$ = 0.001553, 95% 신뢰구간 = [-0.0422, 0.1125]이다.

가중치가 일차형일 (Table 4.11, Table 4.12)때와 가중치를 사용하지 않을 때 (Table 4.13, Table 4.14)도 유사한 결과가 나타났으며 각 범주에서의 일치도 KW_1, KW_2, KW_3에 대한 재표본 p값과 근사검정의 p값의 대소 관계는 가중치에 따라 다르게 나타났다.

Table 4.8 Example 3×3 artificial data set

Judge 1	Judge 2		
	Category 1	Category 2	Category 3
Category 1	124	108	106
Category 2	104	119	102
Category 3	109	102	116

Table 4.9 Exact, resampling and asymptotic p-values for example 2 with quadratic weight

Agreement	Observed value	Resampling p-value	Asymptotic p-value
KW_1	0.0348	0.1913	0.1879
KW_2	0.0502	0.0494	0.0572
KW_3	0.0351	0.1895	0.1863

Table 4.10 Quantile limits and 95% confidence interval for example 2 with quadratic weight

Agreement	Quantile Limits ($Q_{0.025}$, $Q_{0.975}$)	Variance	95% Confidence Interval
KW_1	(-0.0768, 0.0771)	0.0015	(-0.0423, 0.1120)
KW_2	(-0.0640, 0.0639)	0.0010	(-0.0122, 0.1126)
KW_3	(-0.0773, 0.0769)	0.0016	(-0.0422, 0.1125)

Table 4.11 Exact, resampling and asymptotic p-values for example 2 with linear weight

Agreement	Observed value	Resampling p-value	Asymptotic p-value
KW_1	0.0372	0.1343	0.1389
KW_2	0.0502	0.0494	0.0572
KW_3	0.0378	0.1308	0.1357

Table 4.12 Quantile limits and 95% confidence interval for example 2 with linear weight

Agreement	Quantile Limits ($Q_{0.025}$, $Q_{0.975}$)	Variance	95% Confidence Interval
KW_1	(-0.0663, 0.0672)	0.0012	(-0.0300, 0.1045)
KW_2	(-0.0640, 0.0640)	0.0010	(-0.0122, 0.1126)
KW_3	(-0.0670, 0.0666)	0.0012	(-0.0296, 0.1052)

Table 4.13 Exact, resampling and asymptotic p-values for example 2 without weight

Agreement	Observed value	Resampling p-value	Asymptotic p-value
KW_1	0.0402	0.1164	0.1031
KW_2	0.0502	0.0494	0.0572
KW_3	0.0411	0.1111	0.0981

Table 4.14 Quantile limits and 95% confidence interval for example 2 without weight

Agreement	Quantile Limits ($Q_{0.025}$, $Q_{0.975}$)	Variance	95% Confidence Interval
KW_1	(-0.0632, 0.0627)	0.0010	(-0.0222, 0.1026)
KW_2	(-0.0640, 0.0640)	0.0010	(-0.0122, 0.1126)
KW_3	(-0.0642, 0.0640)	0.0010	(-0.0213, 0.1035)

4.3. 예제 데이터 3 (가상 데이터)

Table 4.15는 n=90의 평가 대상을 $k = 4$개의 범주로 평가한 4×4 가상 데이터이다. 주변 빈도수가 고정되어 있을 때 셀 빈도수의 가능한 배열의 수 M=3,131,026,869는 매우 큰 값이므로 L=1,000,000개의 배열을 이용한 재표본 퍼뮤테이션 검정만을 실시하였다. Table 4.16과 Table 4.17에서 가중치가 이차형일 때 KW_1= 0.0758, KW_2=0.0622, KW_3=0.2485, KW_4=-0.0447이므로 평가자들의 일치도는 세 번째 범주에서 가장 높게 나타냈고 네 번째 범주에서의 일치도가 가장 낮았다. 또한 KW_1에 대응되는 재표본 우측검정의 p값 = 0.3169, 근사 우측검정의 p값 = 0.3249이고 $Q_{0.025}$= -0.2939, $Q_{0.975}$= 0.2925, $\widehat{\sigma}^2(KW_1)$ = 0.0277, 95% 신뢰구간 = [-0.2547, 0.4063]이고 KW_2에 대응되는 재표본 우측검정의 p값 = 0.3248, 근사 우측검정의 p값 = 0.1102, $Q_{0.025}$= -0.2951, $Q_{0.975}$= 0.2855, $\widehat{\sigma}^2(KW_2)$ = 0.0025, 95% 신뢰구간 = [-0.0380, 0.1623]이다. 그리고 KW_3에 대응되는 재표본 우측검정의 p값 = 0.0518, 근사 우측검정의 p값 = 0.0557, $Q_{0.025}$= -0.2967, $Q_{0.975}$= 0.2927, $\widehat{\sigma}^2(KW_3)$ = 0.0239, 95% 신뢰구간 = [-0.0585, 0.5556]이며, KW_4에 대응되는 재표본 우측검정의 p값 = 0.6266, 근사 우측검정의 p값 = 0.6125, $Q_{0.025}$= -0.2980, $Q_{0.975}$= 0.2908, $\widehat{\sigma}^2(KW_4)$ = 0.0244, 95% 신뢰구간 = [-0.3549, 0.2654]이다..

가중치가 일차형일 (Table 4.18, Table 4.19) 때와 가중치를 사용하지 않을 때 (Table 4.20, Table 4.21)도 유사한 결과가 나타났으며 각 범주에서의 일치도 KW_1, KW_2, KW_3, KW_4에 대한 재표본 p값과 근사검정의 p값의 대소 관계는 가중치에 따라 다르게 나타났다.

Table 4.15 Example 4×4 artificial data set

Judge 1	Judge 2			
	Category 1	Category 2	Category 3	Category 4
Category 1	8	5	4	7
Category 2	4	7	3	5
Category 3	3	7	8	6
Category 4	5	6	7	5

Table 4.16 Exact, resampling and asymptotic p-values for example 3 with quadratic weight

Agreement	Observed value	Resampling p-value	Asymptotic p-value
KW_1	0.0758	0.3169	0.3249
KW_2	0.0622	0.3248	0.1102
KW_3	0.2485	0.0518	0.0557
KW_4	-0.0447	0.6266	0.6125

Table 4.17 Quantile limits and 95% confidence interval for example 3 with quadratic weight

Agreement	Quantile Limits ($Q_{0.025}$, $Q_{0.975}$)	Variance	95% Confidence Interval
KW_1	(-0.2939, 0.2925)	0.0277	(-0.2547, 0.4063)
KW_2	(-0.2951, 0.2855)	0.0025	(-0.0380, 0.1623)
KW_3	(-0.2967, 0.2927)	0.0239	(-0.0585, 0.5556)
KW_4	(-0.2980, 0.2908)	0.0244	(-0.3549, 0.2654)

Table 4.18 Exact, resampling and asymptotic p-values for example 3 with linear weight

Agreement	Observed value	Resampling p-value	Asymptotic p-value
KW_1	0.1197	0.1496	0.1865
KW_2	0.0825	0.2124	0.1470
KW_3	0.1862	0.0704	0.0663
KW_4	-0.0445	0.6160	0.6399

Table 4.19 Quantile limits and 95% confidence interval for example 3 with linear weight

Agreement	Quantile Limits ($Q_{0.025}$, $Q_{0.975}$)	Variance	95% Confidence Interval
KW_1	(-0.2384, 0.2391)	0.0179	(-0.1460, 0.3853)
KW_2	(-0.2307, 0.2392)	0.0061	(-0.0728, 0.2379)
KW_3	(-0.2317, 0.2302)	0.0150	(-0.0575, 0.4299)
KW_4	(-0.2357, 0.2350)	0.0153	(-0.2903, 0.2014)

Table 4.20 Exact, resampling and asymptotic p-values for example 3 without weight

Agreement	Observed value	Resampling p-value	Asymptotic p-value
KW_1	0.1600	0.0402	0.0751
KW_2	0.1023	0.2365	0.1419
KW_3	0.1245	0.1825	0.1256
KW_4	-0.0513	0.5748	0.6860

Table 4.21 Quantile limits and 95% confidence interval for example 3 without weight

Agreement	Quantile Limits ($Q_{0.025}$, $Q_{0.975}$)	Variance	95% Confidence Interval
KW_1	(-0.2000, 0.2200)	0.0122	(-0.0590, 0.3790)
KW_2	(-0.1960, 0.2226)	0.0091	(-0.0868, 0.2928)
KW_3	(-0.2257, 0.1829)	0.0116	(-0.0897, 0.3387)
KW_4	(-0.2265, 0.1823)	0.0111	(-0.2607, 0.1582)

5. 결론

본 논문에서는 특정 범주에서의 평가자간 가중 일치도 (weighted specific-category kappa)에 대해 정확한 퍼뮤테이션 검정과 재표본 퍼뮤테이션 검정에 의해 p값을 산출하는 방법을 소개한 것이다. 퍼뮤테이션 검정은 모집단 분포에 의존하지 아니하고 통계치의 실제 확률분포를 구축하기 때문에 전통적인 근사검정 (asymptotic tests)보다 선호하는 검정으로 인식되고 있다. 특히 표본의 크기가 작을 때 근사검정에 의한 p값이 부정확할 수 있기 때문에 종종 퍼뮤테이션 검정을 사용한다. 퍼뮤테이션 검정은 정확한 퍼뮤테이션 검정과 재표본 퍼뮤테이션 검정으로 나누는데 정확한 퍼뮤테이션 검정은 비교적 표본 크기가 작은 데이터에 대해 사용되지만 재표본 퍼뮤테이션 검정은 표본이 크기나 범주의 수에 제한 받지 않고 사용된다.

본 논문은 퍼뮤테이션에 의한 p값 산출을 위해 두 개의 3×3 데이터와 한 개의 4×4 데이터를 예제로 사용하였다. 데이터 1 (실제 데이터)의 분석결과에서는 가중치에 관계없이 그리고 모든 KW_1, KW_2, KW_3에서 (정확한 퍼뮤테이션 p값) $\doteq$ (재표본 퍼뮤테이션 p값)로 나타났다. 그리고 가중치가 이차형일 때는 KW_1에 대한 (퍼뮤테이션 p값) $<$ α=0.05이나 (근사검정의 p값) $>$ α=0.05이고, KW_2에 대해서는 α=0.05 $<$ (퍼뮤테이션 p값) $<$ (근사검정의 p값)이고, KW_3경우에는 (퍼뮤테이션 p값) $<$ (근사검정의 p값) $<$ α=0.05이다. 가중치가 일차형일 때는 KW_1, KW_2의 경우에 α=0.05 $<$(퍼뮤테이션 p값) $<$ (근사검정의 p값)이지만. KW_3의 경우는 α=0.05 $<$ (근사검정의 p값) $<$ (퍼뮤테이션 p값)이다. 가중치를 사용하지 않을 때는 KW_2와 KW_3의 경우 α=0.05 $<$ (퍼뮤테이션 p값) $<$ (근사검정의 p값)이지만 KW_1에 대해서는 α=0.05 $<$ (근사검정의 p값) $<$ (퍼뮤테이션 p값)이다.

데이터 2 (가상 데이터)의 분석결과에서는 KW_2의 경우에는 가중치에 관계없이 (재표본 퍼뮤테이션 p값) $<$ α=0.05이나 (근사검정의 p값) $>$ α=0.05이고, KW_1과 KW_3의 경우에는 가중치가 이차형일 때와 가중치를 사용하지 않을 때는 (재표본 퍼뮤테이션 p값) $>$ (근사검정의 p값) $>$ α=0.05, 가중치가 일차형일 때는 (근사검정의 p값) $>$ (재표본 퍼뮤테이션 p값) $>$ α=0.05로 나타났다.

데이터 3 (가상 데이터)의 경우에는 가중치를 사용하지 않을 때의 KW_1에 대한 재표본 p값=0.0402를 제외하고 모든 p값이 0.05보다 크게 나타났다. 그리고 가중치에 관계없이 KW_1의 경우에는 (재표본 퍼뮤테이션 p값) $<$ (근사검정의 p값)이고, KW_2의 경우에는 (재표본 퍼뮤테이션 p값) $>$ (근사검정의 p값)로 나타났으며 KW_3와 KW_4의 경우에는 가중치에 따라 재표본 p값과 근사 p값 사이의 대소 관계가 달리 나타났다.

4절의 예제 데이터에서는 퍼뮤테이션 검정에 의한 p값이 대응되는 근사검정의 p값보다 작거나 큰 값을 보이고 있어 전체적으로 퍼뮤테이션 검정은 근사검정보다 보수적 (conservative)이거나 리버럴 (liberal)한 결과를 보여주고 있다. 그러나 주목할 점은 $k = 3$인 데이터 1과 데이터 3의 결과를 비교할 때 주어진 가중치에서 데이터 1에서의 재표본 p값과 근사검정 p값 사이의 차이가 데이터 3에서의 차이보다 크다는 것인데, 이는 데이터 1의 표본 크기가 데이터 3보다 더 작기 때문이라고 말할 수 있다. 예를들면 가중치가 이차형일 때 데이터 1에서 각각 KW_1, KW_2, KW_3의 재표본 p값과 근사검정 p값의 차이는 0.0078, 0.1042, 0.0091이고 데이터 3에서의 차이는 0.0034, 0.078, 0.0032이다. 따라서 표본의 크기가 작을 때 퍼뮤테이션 검정은 근사검정보다 더 정확한 p값을 제공한다는 여러 연구결과에 따라 퍼뮤테이션 검정이 셀 빈도수가 작은 $k \times k$ 분할표에서 일치도를 분석하는 좋은 방법이라 할 수 있다. 또한 두 예제 데이터에서 정확한 퍼뮤테이션 p값과 1,000,000개의 재표본에 의한 퍼뮤테이션 p값은 거의 동일한 값을 보이므로 재표본 p값에 의해 정확한 p값의 대체가 가능하다고 할 수 있으며, 게다가 재표본 p값을 계산하는데 걸리는 시간이 단지 40.45초 (데이터 1의 경우), 40.93초 (데이터 2의 경우)와 72.12초 (데이터 3의 경우)에 불과하다는 점은 재표본 퍼뮤테이션의 장점이라 할 수 있다.

References

Agresti, A. (2002). *Categorical data anaysis*, 2nd Ed., Wiley, New York.

Berry, K. J., Johnston, J. E. and Mielke, P. W. (2006). Exact and resampling probability values for measures associated with ordered R by C contingency tables. *Psychological Reports*, **99**, 231-238.

Cicchetti, D. V. and Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep redordings. *The American Journal of EEG Technology*, **11**, 101-109.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213-220.

Feinstein, A. R. and Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, **43**, 543-549.

Fisher R. A. (1935). *A design of experiments*, Oliver & Boyd, Edinburgh.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* , 2nd Ed., Wiley, New York.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, **2**, 113-117.

Good, P. I. (2000). *Permutation tests : A practical guide to resampling to resampling methods for testing hypotheses*, 2nd Ed., springer-Verlag, New York.

Good, P. I. (2001). *Resampling methods : A practical guide to data analysis*, 2nd Ed., Birkhauser, Massachusetts.

Han, K. D. and Park Y. G. (2012). A simulation study of rater agreement measures. *Journal of the Korean Data & Information Science Society*, **23**, 25-37.

Holms, C. B. (1979). Sample size in psychological research. *Perceptual and Motor Skills*, **49**, 283-288.

Holms, C. B. (1990). *The honest truth about lying with statistics*, Thomas Springfield, Illinois.

Johnston, J. E., Berry, K. J. and Mielke, P. W. (2007). Permutation tests: Precision in estimating probability values. *Perceptual and Motor Skills*, **105**, 915-920.

Johnston, J. E., Berry, K. J. and Mielke, P. W. (2008). Resampling permutation probability values for weighted kappa. *Psychological Reports*, **103**, 467-475.

Kim, J. and Lee, J. D. (2014). Independence tests using coin package in R. *Journal of the Korean Data & Information Science Society*, **25**, 1039-1055.

Kraemer, H. C. (1983). Kappa coefficient. In *Encyclopedia of Statistical Sciences 4*, Wiley, New York, 352-354.

Kvålseth, T. O. (1989). Note on Cohen's kappa. *Psychological Reports*, **65**, 223-226.

Kvålseth, T. O. (2003). Weighted specific-category kappa measure of interobserver agreement. *Psychological Reports*, **93**, 1283-1290.

Mielke, P. W. and Berry, K. J. (2001). *Permutation methods : A distance function approach. 2001*, Springer-Verlag, New York.

Oleckno, W. A. (2008). *Epidemiology : Concepts and methods*, Waveland Press, Inc., Illinois.

Patefield, W. M. (1981). Algorithm AS 159: An efficient method of generating random R $\times$ C tables with given row and column totals. *Journal of the Royal Statistical Society C*, **30**, 91-97.

Shoukri, M. M. (2004). *Measures of intererobserver agreement*, CRC Press, Florida.

Spitzer, R. L., Cohen, J., Fleiss, J. L. and Endicott, J. (1967). Quantization ofagreement in psychiatric diagnosis. *Archives of General Psychiatry*, **17**, 83-87.

Upton, G. and Cook, I. (2002). *Oxford dictionary of statistics*, Oxford University Press, United Kingdom.

Zhao, X. (2011). *When to use Cohens K, if ever?* International Communication Association 2011 Conference.

Journal of the Korean Data & Information Science Society
2016, **27**(4), 899–910

http://dx.doi.org/10.7465/jkdi.2016.27.4.899
한국데이터정보과학회지

Permutation p-values for specific-category kappa measure of agreement

Yonghwan Um[1]

[1]Division of Industrial and Management Engineering, Sungkyul University

Received 21 March 2016, revised 7 June 2016, accepted 13 July 2016

Abstract

Asymptotic tests are often not suitable for the analysis of sparse ordered contingency tables as asymptotic p-values may either overestimate or underestimate the true p-values. In this pater, we describe permutation procedures in which we compute exact or resampling p-values for a weighted specific-category agreement in ordered $k \times k$ contingency tables. We use the weighted specific-category kappa proposed by Kvålseth to measure the extent to which two independent raters agree on the specific categories. We carried out comparison studies between exact p-values, resampling p-values and asymptotic p-values using 3×3 contingency data (real and artificial data sets) and 4×4 artificial contingency data.

Keywords: Contingency tables, permutation, p-values, weighted specific category agreement,

[1] Professor, Division of Industrial and Management Engineering, Sungkyul University, Anyang, Kyunggi 14097, Email: uyh@sungkyul.ac.kr