Journal of the Korean Data & Information Science Society
2014, **25**(5), 1039–1055

http://dx.doi.org/10.7465/jkdi.2014.25.5.1039
한국데이터정보과학회지

coin 패키지를 이용한 독립성 검정

김진흠[1] · 이정동[2]

[12]수원대학교 통계정보학과

접수 2014년 6월 30일, 수정 2014년 7월 28일, 게재확정 2014년 8월 5일

요 약

검정통계량의 영가설 분포는 모집단 분포에 의존하는데 모집단의 분포를 모를 때 영가설 분포를 검정통계량의 조건부 분포로 대체하여 검정하는 방법을 순열 검정이라고 한다. Strasser와 Weber (1999)는 순열 검정을 통합하는 이론을 마련하였고, Hothorn 등 (2006, 2008)은 그 이론을 R에 내장된 coin 패키지에 구현하였다. coin 패키지에서 조건부 독립성 검정은 총괄적인 형태의 함수인 independence_test를 통해서 할 수 있지만 대표적인 독립성 검정은 사용자가 편리하도록 간편한 함수를 별도로 제공하고 있다. 본 논문에서는 Strasser와 Weber (1999)의 순열 검정 방법에 대해 소개하고, coin 패키지에 내장된 15개의 간편 함수에 대해 independence_test 함수로 변환하는 절차를 설명하고자 한다. 또한, 정의한 independence_test 함수를 써서 실제 자료의 점근 분포와 순열 검정, 정확 검정에 기초한 p-값을 서로 비교하고자 한다.

주요용어: 독립성, 범주형 자료, 순열 검정, 연속형 자료, coin 패키지.

1. 서론

검정통계량의 영가설 (null hypothesis) 분포는 모집단 분포에 의존하는데, 모집단의 분포를 모르면 결국 검정통계량의 분포도 알 수 없게 된다. 이때 모집단의 분포를 가정하여 영가설 분포를 직접 유도하거나 혹은 자료가 주어졌을 때 검정통계량의 조건부 분포를 영가설 분포로 대체하여 검정할 수 있다. 후자의 방법을 조건부 검정 혹은 순열 검정 (permutation test)이라고 한다 (Fisher, 1935). Strasser와 Weber (1999)는 순열 검정을 통합할 수 있는 이론적 근거를 마련하였다. R에 내장된 coin 패키지는 Strasser와 Weber (1999)의 조건부 검정 이론을 구현한 것이다 (Hothorn 등, 2006, 2008). coin이란 이름은 conditional inference의 줄임말이다. 조건부 독립성 검정은 총괄적인 형태의 함수인 independence_test를 통해서 할 수 있지만, 잘 알려진 몇 가지 독립성 검정에 대해서는 사용자가 편리하도록 간편한 함수가 별도로 패키지에 포함되어 있다. 본 논문에서는 이런 간편 함수에 대응하는 independece_test 함수를 정의하고자 한다. 이를 위해, 변수의 적절한 변환과 관측 값의 가중 값 및 블록 값에 대한 정의가 필요하다.

본 논문에서 다루고자 하는 독립성 검정들은 다음과 같다. 두 개 이상의 모집단의 독립성을 검정하기 위한 Wilcoxon-Mann-Whitney 순위합 검정 (Wilcoxon, 1945; Mann과 Whitney, 1947), van der Waerden 정규 분위수 검정 (van der Waerden, 1952, 1953a, 1953b), 중앙값 검정 (Mood, 1950; Westenberg, 1948), Kruskal-Wallis 검정 (Kruskal, 1952; Kruskal과 Wallis, 1953), Ansari-Bradley

[1] 교신저자: (445-743) 경기도 화성시 봉담읍 와우안길 17, 수원대학교 통계정보학과, 교수.
E-mail: jkimdt65@gmail.com
[2] (445-743) 경기도 화성시 봉담읍 와우안길 17, 수원대학교 통계정보학과, 석사과정생.

검정 (Ansari와 Bradley, 1960), Fligner-Killeen 검정 (Fligner와 Killeen, 1976), 로그 순위 검정 (Kalbfleisch와 Prentice, 2002), Wilcoxon 부호 순위 검정 (Wilcoxon, 1945), Friedman 검정 (Friedman, 1937)과, 두 연속형 변수의 독립성을 검정하기 위한 Spearman 검정 (Spearman, 1904), 분할표 자료에서 독립성을 검정하기 위한 Pearson 카이제곱 검정 (Pearson, 1922), Maxwell-Stuart 검정 (Stuart, 1955; Maxwell, 1970), Cochran-Mantel-Haenszel 검정 (Cochran, 1954; Mantel과 Haenszel, 1959), 선형 대 선형 연관성 검정 (Goodman, 1979) 등이다.

2절에서는 Strasser와 Weber (1999)의 순열 검정 방법에 대해 소개하고, 3절에서는 `coin` 패키지에 내장된 간편 함수 15개에 대해 `independence_test` 함수로 변환하는 절차를 설명하고자 한다. 4절에서는 3절에서 정의한 `independence_test` 함수를 써서 3절에서 소개한 자료의 점근 분포와 순열 검정, 정확 검정 (exact test)에 기초한 p-값을 서로 비교하고 글을 맺고자 한다.

2. 순열 검정

Strasser와 Weber (1999)는 순열 검정에 대한 이론을 정리하였고, Hothorn 등 (2006, 2008)은 그 이론을 프로그램으로 구현하였는데, 본 절에서는 순열 검정 이론을 간략히 소개하고자 한다.

다음과 같이 n개의 관측 값을 가지고 있다고 가정하자.

$$\{(Y_i, X_i, w_i, b_i),\ i = 1, \ldots, n\}.$$

X_i와 Y_i는 각각 표본 공간 $\mathcal{X}$와 $\mathcal{Y}$로부터 얻어진 i번째 관측 값이고, 자료형은 `numeric`이거나 `factor`이다. w_i는 i번째 관측 값의 가중 값이고, 자료형은 `numeric`이며, b_i는 i번째 관측 값의 블록 값이고, 자료형은 k (≥ 1)개의 범주를 가진 `factor`이다. 본 논문에서는 순열 검정을 통해 두 변수 X, Y의 독립성을 검정하고자 한다. j번째 블록에 대해 영가설은 다음과 같고,

$$H_0 : D(Y|X, j) = D(Y|j),\ j = 1, \ldots, k,$$

가설 H_0를 검정하기 위한 통계량은 다음과 같다.

$$\boldsymbol{T} = \sum_{j=1}^{k} \boldsymbol{T}_j \in R^{pq}. \tag{2.1}$$

단, $\boldsymbol{T}_j$는 j번째 블록의 통계량으로 다음과 같고,

$$\boldsymbol{T}_j = \mathrm{vec}\left(\sum_{i=1}^{n} I(b_i = j) w_i g(X_i) h(Y_i)' \right) \in R^{pq},\ j = 1, \ldots, k, \tag{2.2}$$

$I(\cdot)$는 지시함수이고, vec는 행렬의 열들을 열벡터로 쌓는 연산자이다. $g : \mathcal{X} \to R^p$는 관측 값 X를 변환하는 함수이고, $h : \mathcal{Y} \to R^q$는 관측 값 Y를 변환하는 함수이다. 특히 $h(Y_i) = h(Y_i, (Y_1, \ldots, Y_n))$는 영향력 함수라고도 하는데, 관측 값 Y_i에만 의존하고 Y_i들의 순서 (배열)에는 의존하지 않는다. w_i는 (X_i, Y_i, b_i)를 가진 개체수를 의미하는데, 기본 (default) 값은 1이다. Strasser와 Weber (1999)는 모든 순열 $\mathcal{S}$가 주어졌을 때, 영가설 하에서, 통계량 $\boldsymbol{T}$의 조건부 평균 벡터와 공분산 행렬을 유도하였다. 이에 앞서 j번째 블록에 속한 가중 값들의 합을 $w_{\cdot j} = \sum_{i=1}^{n} I(b_i = j) w_i$라고 하고, j번째 블록에 속한 관측 값들의 모든 순열을 $\mathcal{S}_j$라고 하자. j번째 블록에 대해, h의 조건부 평균 벡터는 다음과 같고,

$$E(h|\mathcal{S}_j) = w_{\cdot j}^{-1} \sum_{i=1}^{n} I(b_i = j) w_i h(Y_i),\ j = 1, \ldots, k, \tag{2.3}$$

이에 대응하는 $q \times q$ 공분산 행렬은 다음과 같다.

$$Cov(h|\mathcal{S}_j) = w_{\cdot j}^{-1} \sum_{i=1}^{n} I(b_i = j) w_i (h(Y_i) - E(h|\mathcal{S}_j))(h(Y_i) - E(h|\mathcal{S}_j))', \ j = 1, \ldots, k. \tag{2.4}$$

또한, $\boldsymbol{T}_j$의 조건부 평균 벡터와 $pq \times pq$ 공분산 행렬은 각각 다음과 같다.

$$E(\boldsymbol{T}_j|\mathcal{S}_j) = \text{vec}\left(\left(\sum_{i=1}^{n} I(b_i = j) w_i g(X_i)\right) E(h|\mathcal{S}_j)'\right), \ j = 1, \ldots, k, \tag{2.5}$$

$$\begin{aligned} Cov(\boldsymbol{T}_j|\mathcal{S}_j) \quad = \quad & \frac{w_{\cdot j}}{w_{\cdot j} - 1} Cov(h|\mathcal{S}_j) \otimes \left(\sum_{i=1}^{n} I(b_i = j) w_i (g(X_i) \otimes g(X_i)')\right) \\ & - \frac{1}{w_{\cdot j} - 1} Cov(h|\mathcal{S}_j) \otimes \left(\sum_{i=1}^{n} I(b_i = j) w_i g(X_i)\right) \otimes \left(\sum_{i=1}^{n} I(b_i = j) w_i g(X_i)\right)', \\ & j = 1, \ldots, k. \end{aligned} \tag{2.6}$$

단, $\otimes$는 Kroneker 곱이다. 따라서 통계량 $\boldsymbol{T}$의 조건부 평균 벡터와 $pq \times pq$ 공분산 행렬은 k개 블록의 결과를 합쳐 다음과 같이 얻는다.

$$\boldsymbol{\mu} = E(\boldsymbol{T}|\mathcal{S}) = \sum_{j=1}^{k} E(\boldsymbol{T}_j|\mathcal{S}_j), \ \boldsymbol{\Sigma} = Cov(\boldsymbol{T}|\mathcal{S}) = \sum_{j=1}^{k} Cov(\boldsymbol{T}_j|\mathcal{S}_j). \tag{2.7}$$

통계량 $\boldsymbol{T} \in R^{pq}$를 R로 보내는 단변량 통계량 중에서, $pq = 1$이면, 다음과 같이 정의되는 c_{scalar} 통계량이 사용되고,

$$c_{\text{scalar}}(\boldsymbol{T}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \text{diag}(\boldsymbol{\Sigma})^{-1/2}(\boldsymbol{T} - \boldsymbol{\mu}),$$

$pq > 1$이면, 다음과 같이 정의되는 $c_{\max}$ 통계량이나 c_{quad} 통계량이 사용된다.

$$c_{\max}(\boldsymbol{T}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \max\left|\text{diag}(\boldsymbol{\Sigma})^{-1/2}(\boldsymbol{T} - \boldsymbol{\mu})\right|, \ c_{\text{quad}}(\boldsymbol{T}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{T} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{+}(\boldsymbol{T} - \boldsymbol{\mu}).$$

단, $\boldsymbol{\Sigma}^{+}$는 $\boldsymbol{\Sigma}$의 Moore-Penrose 역행렬이다. 영가설 하에서, c 통계량의 조건부 분포는 다음과 같은데,

$$Pr(c(\boldsymbol{T}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq z|\mathcal{S}),$$

이 값은 z 값을 초과하지 않는 순열의 개수를 전체 순열의 개수로 나누어 구할 수 있다. 단, z는 c 통계량의 관측 값이다.

3. 순열 검정의 응용

3.1. X는 `factor`, Y는 `numeric`인 자료의 독립성 검정

3.1.1. 모든 관측 값이 독립인 자료

Y_i는 i번째 개체의 관측 값이며, X_i는 i번째 개체가 속한 그룹 $(G_l, \ l = 1, \ldots, p)$이다. 모든 개체가 한 블록에 포함되며, 개체의 가중 값은 1로 모두 같다. 즉, $b_i = 0$, $w_i = 1$이다.

두 모집단 $(p = 2)$에서 모평균이 서로 동일한지를 검정하기 위한 Wilcoxon-Mann-Whitney 순위합 검정 (Wilcoxon, 1945; Mann과 Whitney, 1947), van der Waerden 정규 분위수 검정 (van der

Waerden, 1952, 1953a, 1953b), 중앙값 검정 (Mood, 1950; Westenberg, 1948)에 대응하는 `coin` 패키지의 함수는 각각 `wilcox_test`, `normal_test`, `median_test`이며, 모분산이 서로 동일한지를 검정하기 위한 Ansari-Bradley 검정 (Ansari와 Bradley, 1960)에 대응하는 함수는 `ansari_test`이다. 각 함수가 `independence_test`와 서로 동일한 검정이 되도록 함수 g와 h를 정의하면 다음과 같다. g는 다음과 같이 모두 동일하며,

$$g(X_i) = I(i \in G_2),\ i = 1, \ldots, n, \tag{3.1}$$

$l = 1, 2$에 대해, $n_l = \sum_{i=1}^{n} I(i \in G_l)$라고 놓으면, `wilcox_test`, `normal_test`, `median_test`, `ansari_test`에 대응하는 h는 각각 다음과 같다. $i = 1, \ldots, n$에 대해,

$$h(Y_i) = R_i = a_W(i);\ h(Y_i) = \Phi^{-1}\left(\frac{R_i}{n+1}\right) = a_N(i);\ h(Y_i) = I(Y_i \le m) = a_M(i);$$

$$h(Y_i) = \left(\frac{n+1}{2} - \left|R_i - \frac{n+1}{2}\right|\right) = a_{AB}(i).$$

단, $R_i = \mathrm{rank}_i(Y_i)$, $m = \mathrm{med}_i(Y_i)$, Φ는 표준정규분포의 누적분포함수이다. 임의 스코어 벡터 $\boldsymbol{a}_s = (a_s(1), \ldots, a_s(n))'$에 대해,

$$\bar{a}_s = \frac{1}{n}\sum_{i=1}^{n} a_s(i),\ b_s^2 = \frac{1}{n}\sum_{i=1}^{n}(a_s(i) - \bar{a}_s)^2$$

라고 놓자. 단, $a_s(i)$는 실수이다. 모든 i에 대해, $b_i = 0$이고 $w_i = 1$이면, $\bar{a}_s$와 b_s^2은 각각 함수 h의 조건부 평균과 분산에 해당한다. 따라서 스코어 벡터 $\boldsymbol{a}_s$에 대해, 통계량 (2.1)은 각각 다음과 같고,

$$T_s = \sum_{i=1}^{n} I(X_i = 2)a_s(i),\ s = W, N, M, AB,$$

T_s의 조건부 평균과 분산은 식 (2.7)로부터 쉽게 구할 수 있다.

자료 `water_tranfer` (Hollander와 Wolfe, 1999)를 써서 두 임신기간 ('At term': 37 주, '12-26 weeks': 12-26 주)에 따라 태반막의 삼중수소 평균 투과율이 서로 다른지를 각각 Wilcoxon-Mann-Whitney 순위합 검정, van der Waerden 정규 분위수 검정, 중앙값 검정하였고, 자료 `sid` (Hollander와 Wolfe, 1999)를 써서 두 정제 방법 ('Ramsay', 'Jung-Parekh')에 따라 혈청철 결정의 양의 산포가 서로 다른지를 Ansari-Bradley 검정하였다. `wilcox_test`, `normal_test`, `median_test`, `ansari_test` 함수를 각각 대응하는 `independece_test` 함수로 표현하면 `R-code-3.1.1-1`과 같으며, 이때 c 통계량의 형태는 모두 `scalar`이다.

*** R-code-3.1.1-1 ***

```
>water_transfer = data.frame(pd = c(0.80, 0.83, 1.89, 1.04, 1.45, 1.38, 1.91, 1.64, 0.73, 1.46, 1.15,
0.88, 0.90, 0.74, 1.21), age = factor(c(rep("At term", 10), rep("12-26 Weeks", 5)))) #data set
>sid = data.frame(serum = c(111, 107, 100, 99, 102, 106, 109, 108, 104, 99, 101, 96, 97, 102, 107, 113,
116, 113, 110, 98, 107, 108, 106, 98, 105, 103, 110, 105, 104, 100, 96, 108, 103, 104, 114, 114, 113,
108, 106, 99), method = factor(gl(2, 20), labels = c("Ramsay", "Jung-Parekh"))) #data set
```

>wilcox_test(pd ∼ age, data = water_transfer)	>independence_test(pd ∼ age, data = water_transfer, ytrafo = function(data) trafo(data, numeric_trafo = rank)) #Wilcoxon-Mann-Whitney rank-sum test
>normal_test(pd ∼ age, data = water_transfer)	>independence_test(pd ∼ age, data = water_transfer, ytrafo = function(data) trafo(data, numeric_trafo = normal_trafo)) #van der Waerden normal quantiles test
>median_test(pd ∼ age, data = water_transfer)	>independence_test(pd ∼ age, data = water_transfer, ytrafo = function(data) trafo(data, numeric_trafo = median_trafo)) #Median test
>ansari_test(serum ∼ method, data = sid)	>independence_test(serum ∼ method, data = sid, ytrafo = function(data) trafo(data, numeric_trafo = ansari_trafo)) #Ansari-Bradley test

세 개 이상의 모집단 ($p > 2$)에서 모평균이 서로 동일한지를 검정하기 위한 Kruskal-Wallis 검정 (Kruskal, 1952; Kruskal과 Wallis, 1953)에 대응하는 함수는 `kruskal_test`이고, 모분산이 서로 동일한지를 검정하기 위한 Fligner-Killeen 검정 (Fligner와 Killeen, 1976)에 대응하는 함수는 `fligner_test`이다. 각 함수가 `independence_test`와 서로 동일한 검정이 되도록 함수 g와 h를 정의하면 다음과 같다.

$$g(X_i) = (I(i \in G_1), \ldots, I(i \in G_p))', \ i = 1, \ldots, n. \tag{3.2}$$

$l = 1, \ldots, p$에 대해, $n_l = \sum_{i=1}^{n} I(i \in G_l)$라고 놓으면, `kruskal_test`, `fligner_test`에 대응하는 h는 각각 다음과 같다. $i = 1, \ldots, n$에 대해,

$$h(Y_i) = R_i = a_{KW}(i); \ h(Y_i) = \Phi^{-1}\left(\frac{1}{2} + \frac{R_{FK,i}}{2(n+1)}\right) = a_{FK}(i).$$

단, $R_{FK,i} = \sum_{j=1}^{n} I(|Y_j - m_j| \le |Y_i - m_i|)$, $m_i = \mathrm{med}_{\{l:X_i=X_l\}}(Y_l)$이다. 따라서 스코어 벡터 $\boldsymbol{a}_s$에 대해, 통계량 (2.1)은 다음과 같고,

$$\boldsymbol{T}_s = \left(\sum_{i=1}^{n} I(i \in G_1)a_s(i), \ldots, \sum_{i=1}^{n} I(i \in G_p)a_s(i)\right)', \ s = KW, FK, \tag{3.3}$$

$\boldsymbol{T}_s$의 조건부 평균 벡터와 공분산 행렬은 식 (2.7)로부터 쉽게 얻을 수 있다.

자료 `YOY` (Hollander와 Wolfe, 1999)를 써서 미국 Kokosing 호수의 4개 지점('I','II','III','IV')에 따라 어린 전어들의 평균 길이가 서로 다른지를 Kruskal-Wallis 검정하였다. `kruskal_test`, `fligner_test` 함수를 각각 대응하는 `independece_test` 함수로 표현하면 `R-code-3.1.1-2`와 같으며, 이때 c 통계량의 형태는 모두 `quad`이다.

*** R-code-3.1.1-2 ***

```
>YOY = data.frame(length = c(46, 28, 46, 37, 32, 41, 42, 45, 38, 44, 42, 60, 32, 42, 45, 58, 27,
51, 42, 52, 38, 33, 26, 25, 28, 28, 26, 27, 27, 27, 31, 30, 27, 29, 30, 25, 25, 24, 27, 30), site =
factor(c(rep("I", 10), rep("II", 10), rep("III", 10), rep("IV", 10)))) #data set
```

```
>kruskal_test(length ∼ site, data = YOY)
>fligner_test(serum ∼ method, data = sid)
>independence_test(length ∼ site, data = YOY,
ytrafo = function(data) trafo(data, numeric_trafo
= rank), teststat = "quad") #Kruskal-Wallis test
>sub.A = subset(sid, method == "Ramsay",
select = serum); sub.B = subset(sid, method ==
"Jung-Parekh", select = serum)
>dev = sid$serum - ((sid$method == "Ramsay")
* median(sub.A$serum) + (sid$method ==
"Jung-Parekh") * median(sub.B$serum))
>score = qnorm(0.5 + rank(abs(dev)) / (2 *
(length(sid$serum) + 1)))
>independence_test(serum ∼ method, data = sid,
ytrafo = function(data) trafo(data, numeric_trafo
= function(x) cbind(score)), teststat = "quad")
#Fligner-Killeen test
```

3.1.2. 관측 값이 중도절단된 자료

(Y_i, δ_i)에서 Y_i는 i번째 개체의 관측 값이고, δ_i는 우중도절단여부를 나타내는 값으로, $\delta_i = 1$이면, Y_i는 생존시간을 나타내고, $\delta_i = 0$이면, Y_i는 중도절단시간을 나타낸다. 3.1.1절처럼 X_i는 i번째 개체가 속한 그룹 $(G_l,\ l = 1, \ldots, p)$을 나타내고, $b_i = 0$, $w_i = 1$이다.

두 개 이상의 모집단 $(p \geq 2)$의 생존함수가 서로 동일한지를 검정하기 위한 로그 순위 검정 (log rank test)에 대응하는 함수 `surv_test`와 `independence_test`가 서로 동일한 검정이 되도록 함수 g와 h를 정의하면 다음과 같다. g는 (3.2)와 같다. $Y_{(1)} < \cdots < Y_{(c)}$를 관측된 서로 다른 생존시간이라고 놓고, $l = 0, \ldots, c$에 대해, $Y_{l1}, \ldots, Y_{lm_l}$는 구간 $[Y_{(l)}, Y_{(l+1)})$에서 중도절단된 시간이라고 놓자. 단, $Y_{(0)} = 0,\ Y_{(c+1)} = \infty$이다. $l = 1, \ldots, c$에 대해, $Y_{(l)}$에서 이벤트가 발생한 개체수를 d_l이라고 하고, $Y_{(l)}$ 바로 직전까지 위험집합 (risk set)에 남아 있는 개체수를 n_l이라고 하자. $l = 1, \ldots, c$에 대해, $Y_{(l)}$에서 이벤트가 발생한 개체와 구간 $[Y_{(l)}, Y_{(l+1)})$에서 중도절단된 개체들은 각각 다음과 같은 스코어,

$$s_l = 1 - \sum_{h=1}^{l} \frac{d_h}{n_h},\ S_l = -\sum_{h=1}^{l} \frac{d_h}{n_h}$$

를 갖지만, 구간 $(0, Y_{(1)})$에서 중도절단된 개체들은 스코어를 갖지 않는다면 (즉, $S_0 = 0$), h는 다음과 같다.

$$h(Y_i) = \sum_{l=1}^{c} \left\{ s_l I(Y_i = Y_{(l)},\ \delta_i = 1) + S_l \sum_{j=1}^{m_l} I(Y_i = Y_{lj}, \delta_i = 0) \right\} = a_{LR}(i),\ i = 1, \ldots, n.$$

따라서 통계량 (2.1)은 식 (3.3)에 스코어 벡터 $\boldsymbol{a}_{LR} = (a_{LR}(1), \ldots, a_{LR}(n))'$을 대입하여 얻을 수 있고, 그 통계량의 조건부 평균 벡터와 공분산 행렬은 식 (2.7)로부터 쉽게 얻을 수 있다.

자료 `ocarcinoma` (Fleming 등, 1980)는 35명의 난소암 환자들의 생존시간과 중도절단 여부, 병기를 가지고 있는데, 두 병기 ('2기', '2A기')에 따라 난소암 환자들의 생존시간이 서로 다른지를 로그 순위 검정하였다. `surv_test` 함수를 대응하는 `independece_test` 함수로 표현하면 `R-code-3.1.2-1`과 같으며, 이때 c 통계량의 형태는 `scalar`이다.

*** R-code-3.1.2-1 ***

```
>surv_test(Surv(time, event) ∼ stadium, data =
ocarcinoma)
>independence_test(Surv(time, event) ∼ stadium,
data = ocarcinoma, ytrafo = function(data)
trafo(data, numeric_trafo = logrank_trafo)) #Log
rank test
```

3.1.3. 같은 블록 내에 있는 관측 값이 종속된 자료

(Y_i, b_i)에서 Y_i는 i번째 개체의 관측 값이며, b_i는 i번째 개체가 속한 블록 값이다. 따라서 서로 다른 블록에 있는 개체들은 서로 독립이지만 같은 블록 내에 있는 개체들은 서로 종속된 자료이다. X_i는 i번째 개체가 속한 그룹 $(G_l,\ l = 1, \ldots, p)$을 나타내고, $w_i = 1$이다.

블록 내에 있는 두 모집단 $(p = 2)$에서 모평균이 서로 동일한지를 검정하기 위한 Wilcoxon 부호 순위 검정 (Wilcoxon, 1945)에 대응하는 `wilcoxsign_test`가 `independence_test`와 서로 동일한 검정이 되도록 g와 h를 정의하면 다음과 같다. g는 (3.1)과 같다. j번째 블록 내에 있는 두 관측 값 간의 절대 편차 D_j는 다음과 같고,

$$D_j = \sum_{i_1=1}^{n-1} \sum_{i_2=i_1+1}^{n} I(b_{i_1} = j = b_{i_2})|Y_{i_1} - Y_{i_2}|,\ j = 1, \ldots, k,$$

h는 다음과 같다.

$$h_W(Y_i) = R_{W,i} \sum_{l \neq i}^{n} I(b_l = b_i) I(Y_i < Y_l) = a_W(i),\ i = 1, \ldots, n.$$

단, $R_{W,i} = \sum_{j_1=1}^{k} I(b_i = j_1)\{\sum_{j_2=1}^{k} I(D_{j_2} \leq D_{j_1})\}$이다. 임의 스코어 벡터 $\boldsymbol{a}_s = (a_s(1), \ldots, a_s(n))'$에 대해,

$$\bar{a}_{s,j} = \frac{1}{n_j} \sum_{i=1}^{n} I(b_i = j) a_s(i),\ b_{s,j}^2 = \frac{1}{n_j} \sum_{i=1}^{n} I(b_i = j)(a_s(i) - \bar{a}_{s,j})^2,\ j = 1, \ldots, k \tag{3.4}$$

라고 하자. 단, $n_j = \sum_{i=1}^{n} I(b_i = j)$이다. 모든 i에 대해, $w_i = 1$이면, $\bar{a}_{s,j}$와 $b_{s,j}^2$은 각각 j번째 블록 내에서 함수 h의 조건부 평균과 분산에 해당한다. 따라서 스코어 벡터 $\boldsymbol{a}_W = (a_W(1), \ldots, a_W(n))'$에 대해, j번째 블록의 통계량 (2.2)는 다음과 같으며,

$$T_{W,j} = \sum_{i=1}^{n} I(b_i = j) I(i \in G_2) a_W(i),\ j = 1, \ldots, k,$$

$T_{W,j}$의 조건부 평균과 분산은 식 (2.5)과 (2.6)으로부터 각각 다음과 같다.

$$\mu_{W,j} = \bar{a}_{W,j},\ \sigma_{W,j}^2 = b_{W,j}^2,\ j = 1, \ldots, k.$$

그러므로 통계량 (2.1)은 다음과 같고,

$$T_W = \sum_{j=1}^{k} T_{W,j},$$

식 (2.7)로부터 T_W의 조건부 평균과 분산은 쉽게 얻을 수 있다.

쌍으로 얻어진 가상의 자료 `xydat`를 써서 두 변수('x', 'y')의 모평균이 서로 다른지를 Wilcoxon 부호 순위 검정하였다. `wilcoxsign_test` 함수를 대응하는 `independece_test` 함수로 표현하면 `R-code-3.1.3-1`과 같으며, 이때 c 통계량의 형태는 `scalar`이다.

*** R-code-3.1.3-1 ***

```
>x = c(1.83, 0.50, 1.62, 2.48, 1.68, 1.88, 1.55, 3.06, 1.30)
>y = c(0.878, 0.647, 0.598, 2.05, 1.06, 1.29, 1.06, 3.14, 1.29)
>xydat = data.frame(y = c(y, x), x = gl(2, length(x)), block = factor(rep(1:length(x), 2))) #data set
```

```
>wilcoxsign_test(y ~ x | block, data = xydat)
```

```
>a = as.numeric(rep(x > y, rep(2, length(x))))
>b = rep(c(0, 1), length(x))
>arank = as.numeric(a == b) * rep(rank(abs(x −
y)), rep(2, length(x)))
>d = data.frame(d.x = rep(0:1, length(x)),
d.y=c(x, y), block = factor(rep(1 :     length(x),
rep(2, length(x)))))
>independence_test(d.y ~ d.x | block, data = d,
ytrafo = function(data) numeric_trafo = arank)
#Wilcoxon signed rank test
```

같은 블록 내에 있는 세 개 이상의 모집단 ($p > 2$)에서 모평균이 서로 동일한지를 검정하기 위한 Friedman 검정 (Friedman, 1937)에 대응하는 `friedman_test`가 `independence_test`와 서로 동일한 검정이 되도록 g와 h를 정의하면 다음과 같다. g는 (3.2)와 같고, h는 다음과 같다.

$$h(Y_i) = R_{F,i} = a_F(i),\ i = 1, \ldots, n.$$

단, $R_{F,i} = \sum_{l=1}^{n} I(b_l = b_i) I(Y_l \leq Y_i)$이다. 따라서 스코어 벡터 $\boldsymbol{a}_F = (a_F(1), \ldots, a_F(n))'$에 대해, j번째 블록의 통계량 (2.2)는 다음과 같으며,

$$\boldsymbol{T}_{F,j} = \left(\sum_{i=1}^{n} I(b_i = j) I(i \in G_1) a_F(i), \ldots, \sum_{i=1}^{n} I(b_i = j) I(i \in G_p) a_F(i) \right)',\ j = 1, \ldots, k,$$

$\boldsymbol{T}_{F,j}$의 조건부 평균 벡터와 공분산 행렬은 식 (2.5)과 (2.6)으로부터 각각 다음과 같다.

$$\boldsymbol{\mu}_{F,j} = \bar{a}_{F,j} \mathbf{1}_c,\ \boldsymbol{V}_{F,j} = (\boldsymbol{V}_{F,j})_{uv} = \begin{cases} b_{F,j}^2, & u = v, \\ -\frac{1}{n_j - 1} b_{F,j}^2, & u \neq v, \end{cases} \quad j = 1, \ldots, k.$$

단, $\mathbf{1}_c = (1, \ldots, 1)'$는 c-차원 벡터이다. 그러므로 통계량 (2.1)은 다음과 같고,

$$\boldsymbol{T}_F = \sum_{j=1}^{k} \boldsymbol{T}_{F,j},$$

$\boldsymbol{T}_F$의 조건부 평균 벡터와 공분산 행렬은 식 (2.7)로부터 쉽게 얻을 수 있다.

자료 `RoundingTimes` (Hollander와 Wolfe, 1999)를 써서 1루 베이스를 도는 세 가지 방법 ('Round Out', 'Narrow Angle', 'Wide Angle')에 따라 2루 베이스에 도착하는 평균 시간이 서로 다른지를 Friedman 검정하였다. `friedman_test` 함수를 대응하는 `independece_test` 함수로 표현하면 `R-code-3.1.3-2`와 같으며, 이때 c 통계량의 형태는 `quad`이다.

*** R-code-3.1.3-2 ***

```
>RoundingTimes = data.frame(times = c(5.40, 5.50, 5.55, 5.85, 5.70, 5.75, 5.20, 5.60, 5.50, 5.55, 5.50,
5.40, 5.90, 5.85, 5.70, 5.45, 5.55, 5.60, 5.40, 5.40, 5.35, 5.45, 5.50, 5.35, 5.25, 5.15, 5.00, 5.85,
5.80, 5.70, 5.25, 5.20, 5.10, 5.65, 5.55, 5.45, 5.60, 5.35, 5.45, 5.05, 5.00, 4.95, 5.50, 5.50, 5.40,
5.45, 5.55, 5.50, 5.55, 5.55, 5.35, 5.45, 5.50, 5.55, 5.50, 5.45, 5.25, 5.65, 5.60, 5.40, 5.70, 5.65,
5.55, 6.30, 6.30, 6.25), methods = factor(rep(c("Round Out", "Narrow Angle", "Wide Angle"), 22)), block
= factor(rep(1:22, rep(3, 22)))) #data set
>friedman_test(times ~ methods, data =
RoundingTimes)
```

```
>independence_test(times ~ methods | block, data
= RoundingTimes, xtrafo=function(data) trafo(data,
factor_trafo = function(x) f_trafo(x)), ytrafo
= function(data) trafo(data, numeric_trafo =
rank, block = RoundingTimes$block), teststat =
"quad")#Friedman test
```

3.2. X와 Y가 모두 numeric인 자료의 독립성 검정

Y_i와 X_i는 i번째 개체의 관측 값이며, $b_i = 0$, $w_i = 1$이다. 두 변수의 선형관계가 0인지를 검정하기 위한 Spearman 검정 (Spearman, 1904)에 대응하는 spearman_test가 independence_test와 서로 동일한 검정이 되도록 g와 h를 정의하면 다음과 같다. g는 다음과 같고,

$$g(X_i) = \text{rank}_i(X_i) = r_i,\ i = 1, \ldots, n,$$

h는 $a_W(i)$와 같다. 따라서 스코어 벡터 $\boldsymbol{a}_W = (a_W(1), \ldots, a_W(n))'$에 대해, 통계량 (2.1)은 다음과 같고,

$$T_{Sp} = \sum_{i=1}^{n} r_i a_W(i),$$

T_{Sp}의 조건부 평균과 분산은 식 (2.7)로부터 쉽게 얻을 수 있다.

자료 USJudgeRatings (New Haven Register, 14 January, 1977)는 12개 항목에 대해 판사들에 대한 변호사들의 평가 점수를 가지고 있는데, 그 중에서 두 변수 ('CONT': 판사와 만난 횟수,'INTG': 판결 진실성)의 평가 점수가 서로 독립인지를 Spearman 검정하였다. spearman_test 함수를 대응하는 independece_test 함수로 표현하면 R-code-3.2-1과 같으며, 이때 c 통계량의 형태는 scalar이다.

*** R-code-3.2-1 ***

```
>spearman_test(CONT ~ INTG, data =
USJudgeRatings)
```

```
>independence_test(CONT ~ INTG, data =
USJudgeRatings, xtrafo = function(data)
trafo(data, numeric_trafo = rank), ytrafo =
function(data) trafo(data, numeric_trafo = rank))
#Spearman test
```

X의 가능한 모든 값을 기준으로 두 그룹으로 나눈 후 (즉, X의 자료형이 numeric에서 factor로 바뀜) 통계량 값의 최대 값으로 두 그룹의 모평균이 서로 동일한지를 검정하기 위한 최대 선택 통계량 검정 (maximally selected statistic test; Müller와 Hothorn, 2004)에 대응하는 maxstat_test가 independence_test와 서로 동일한 검정이 되도록 g와 h를 정의하면 다음과 같다. $X_{(1)} < \cdots < X_{(p+1)}$를 서로 다른 X의 값이라고 하면, $i = 1, \ldots, n$에 대해,

$$g(X_i) = (I(X_i \le X_{(1)}), \ldots, I(X_i \le X_{(p)}))';\ h(Y_i) = Y_i = a_M(i).$$

$g(X_i) = (1, \ldots, 1)', (0, 1, \ldots, 1)', \ldots, (0, \ldots, 0, 1), (0, \ldots, 0)'$인 개체수가 각각 n'_1, n'_2, $\ldots$, n'_p, n'_{p+1}라고 가정하고, $k = 1, \ldots, p$에 대해, $n_k = \sum_{l=1}^{k} n'_l$로 정의하면, 스코어 벡터 $\boldsymbol{a}_M = (a_M(1), \ldots, a_M(n))'$에 대해, 통계량 (2.1)은 다음과 같고,

$$\boldsymbol{T}_M = \left(\sum_{i=1}^{n} I(X_i \le x_{(1)}) a_M(i), \ldots, \sum_{i=1}^{n} I(X_i \le x_{(p)}) a_M(i) \right)',$$

$\boldsymbol{T}_M$의 조건부 평균 벡터와 공분산 행렬은 식 (2.7)로부터 각각 다음과 같다.

$$\boldsymbol{\mu}_M = \bar{a}_M(n_1, \ldots, n_p)',\ \boldsymbol{\Sigma}_M = (\boldsymbol{\Sigma}_M)_{uv} = \begin{cases} \frac{n_u(n-n_u)}{n-1} b_M^2, & u = v, \\ \frac{n_v(n-n_u)}{n-1} b_M^2, & u > v, \\ \frac{n_u(n-n_v)}{n-1} b_M^2, & u < v. \end{cases}$$

자료 treepipit (Agresti, 2002)는 나무발종다리의 86개 서식지에 대한 10개 변수의 측정 값을 가지고 있는데, 숲이 우거진 정도 ('coverstorey')에 따라 서식지를 두 그룹으로 나누었을 때, 서식하는 나무

밭종다리의 수 ('counts')가 두 그룹 간에 서로 다른지를 최대 선택 통계량 검정하였다. `maxstat_test` 함수를 대응하는 `independece_test` 함수로 표현하면 `R-code-3.2-2`와 같으며, 이때 c 통계량의 형태는 `max`이다.

```
                                    *** R-code-3.2-2 ***
>maxstat_test(counts ~ coverstorey, data =        >independence_test(counts ~ coverstorey, data
treepipit)                                        = treepipit, xtrafo = function(data) trafo(data,
                                                  numeric_trafo = maxstat_trafo), teststat = "max")
                                                  #Maximally selected statistic test
```

3.3. X와 Y가 모두 factor인 자료의 독립성 검정

(X_i, Y_i)는 X와 Y의 범주들의 i번째 쌍을 나타내는 값이다. 따라서 X와 Y의 범주수가 각각 I개, J개라고 하면 총 쌍의 개수는 $I \times J = n$개이다. w_i는 i번째 쌍의 관측 개체수를 나타낸다. X와 Y의 서로 다른 범주를 각각 $r_1, \ldots, r_I$; $c_1, \ldots, c_J$라고 하자.

3.3.1. 모든 쌍이 서로 독립인 자료

2차원 분할표에서 두 범주형 변수의 독립성을 검정하기 위한 Pearson 카이제곱 검정 (Pearson, 1922)에 대응하는 `chisq_test`가 `independence_test`와 서로 동일한 검정이 되도록 g와 h를 정의하면 다음과 같다. $i = 1, \ldots, n$에 대해,

$$g(X_i) = (I(X_i = r_1), \ldots, I(X_i = r_I))'; \tag{3.5}$$

$$h(Y_i) = (I(Y_i = c_1), \ldots, I(Y_i = c_J))' = \boldsymbol{a}_C(i). \tag{3.6}$$

$j = 1, \ldots, J$에 대해, $N_{\cdot j} = \sum_{i=1}^{n} I(Y_i = c_j) w_i$라고 놓고, $i = 1, \ldots, I$에 대해, $N_{i\cdot} = \sum_{l=1}^{n} I(X_l = r_i) w_l$라고 놓으면, 스코어 벡터 $\boldsymbol{a}_C = (a_C(1), \ldots, a_C(n))'$에 대해, h의 조건부 평균 벡터와 공분산 행렬은 식 (2.3)과 (2.4)로부터 각각 다음과 같다.

$$\bar{\boldsymbol{a}}_C = \left(\frac{N_{\cdot 1}}{N}, \ldots, \frac{N_{\cdot J}}{N} \right)', \ \boldsymbol{B}_C = (\boldsymbol{B}_C)_{uv} = \begin{cases} \frac{N_{\cdot u}}{N}\left(1 - \frac{N_{\cdot u}}{N}\right), & u = v, \\ -\frac{N_{\cdot u}}{N}\frac{N_{\cdot v}}{N}, & u \neq v. \end{cases} \tag{3.7}$$

단, $N = \sum_{j=1}^{J} N_{\cdot j} = \sum_{i=1}^{I} N_{i\cdot}$이다. 따라서 통계량 (2.1)은 다음과 같고,

$$\boldsymbol{T}_C = (w_1, \ldots, w_n)',$$

$\boldsymbol{T}_C$의 조건부 평균 벡터와 공분산 행렬은 식 (2.7)로부터 각각 다음과 같다.

$$\boldsymbol{\mu}_C = \left(\frac{N_{1\cdot} N_{\cdot 1}}{N}, \ldots, \frac{N_{I\cdot} N_{\cdot J}}{N} \right)', \ \boldsymbol{\Sigma}_C = \frac{N}{N-1} \boldsymbol{B}_C \otimes \left\{ \mathrm{diag}(N_{1\cdot}, \ldots, N_{I\cdot}) - \frac{1}{N}(m_R \otimes m_R') \right\}.$$

단, $m_R = (N_{1\cdot}, \ldots, N_{I\cdot})'$이다.

자료 `jobsatisfaction` (Agresti, 2002)은 성별에 따라 분류된 수입 ('income')과 직업 만족도 ('job.satisfaction')에 대한 3차원 분할표인데, 'Female'들의 수입과 직업만족도가 서로 독립인지를 Pearson 카이제곱 검정하였다. `chisq_test` 함수를 대응하는 `independece_test` 함수로 표현하면 `R-code-3.3.1-1`과 같으며, 이때 c 통계량의 형태는 `quad`이다.

*** R-code-3.3.1-1 ***

```
>chisq_test(as.table(jobsatisfaction[,,
"Female"]))
```

```
>independence_test(Job.Satisfaction ~ Income,
data = as.table(jobsatisfaction[,,"Female"]),weights=
~ as.vector(as.table(jobsatisfaction[,,"Female"])),
teststat = "quad") #Pearson's chi-square test
```

3차원 분할표에서 두 범주형 변수의 조건부 독립성을 검정하기 위한 Cochran-Mantel-Haenszel 검정 (Cochran, 1954; Mantel과 Haenszel, 1959)에 대응하는 `cmh_test`가 `independence_test`와 서로 동일한 검정이 되도록 하는 g와 h는 각각 (3.5), (3.6)과 같다. 3차원 분할표에서 제어 변수는 블록 변수로 간주되기 때문에 i번째 쌍의 제어 변수의 값은 b_i의 값이 된다. 제어 변수의 서로 다른 범주를 $s_1, \ldots, s_k$라고 하자. $l = 1, \ldots, k$에 대해, h의 조건부 평균 벡터와 공분산 행렬은 식 (2.3)과 (2.4)로부터 각각 다음과 같고,

$$\bar{\boldsymbol{a}}_{CMH,l} = \left(\frac{N_{\cdot 1l}}{N_{\cdot\cdot l}}, \ldots, \frac{N_{\cdot Jl}}{N_{\cdot\cdot l}} \right)', \ \boldsymbol{B}_{CMH,l} = (\boldsymbol{B}_{CMH,l})_{uv} = \begin{cases} \frac{N_{\cdot ul}}{N_{\cdot\cdot l}} \left(1 - \frac{N_{\cdot ul}}{N_{\cdot\cdot l}} \right), & u = v, \\ -\frac{N_{\cdot ul}}{N_{\cdot\cdot l}} \frac{N_{\cdot vl}}{N_{\cdot\cdot l}}, & u \neq v. \end{cases}$$

단, $l = 1, \ldots, k$에 대해, $N_{\cdot jl} = \sum_{i=1}^{n} I(b_i = s_l) I(Y_i = c_j) w_i, \ j = 1, \ldots, J$; $N_{\cdot\cdot l} = \sum_{j=1}^{J} N_{\cdot jl}$이다. l번째 블록의 통계량 (2.2)는 다음과 같고,

$$\boldsymbol{T}_{CMH,l} = (w_{IJ(l-1)+1}, \ldots, w_{IJl})', \ l = 1, \ldots, k,$$

$\boldsymbol{T}_{CMH,l}$의 조건부 평균 벡터와 공분산 행렬은 식 (2.5)와 (2.6)으로부터 각각 다음과 같다.

$$\boldsymbol{\mu}_{CMH,l} = \left(\frac{N_{1\cdot l} N_{\cdot 1l}}{N_{\cdot\cdot l}}, \ldots, \frac{N_{I\cdot l} N_{\cdot Jl}}{N_{\cdot\cdot l}} \right)', \ l = 1, \ldots, k,$$

$$\boldsymbol{V}_{CMH,l} = \frac{N_{\cdot\cdot l}}{N_{\cdot\cdot l} - 1} \boldsymbol{B}_{CMH,l} \otimes \left\{ \mathrm{diag}(N_{1\cdot l}, \ldots, N_{I\cdot l}) - \frac{1}{N_{\cdot\cdot l}} (m_{R,l} \otimes m'_{R,l}) \right\}, \ l = 1, \ldots, k.$$

단, $N_{i\cdot l} = \sum_{i=1}^{n} I(b_i = s_l) I(X_i = r_i) w_i$, $m_{R,l} = (N_{1\cdot l}, \ldots, N_{I\cdot l})'$이다. 그러므로 통계량 (2.1)은 다음과 같고,

$$\boldsymbol{T}_{CMH} = \sum_{l=1}^{k} \boldsymbol{T}_{CMH,l},$$

$\boldsymbol{T}_{CMH}$의 조건부 평균 벡터와 공분산 행렬은 식 (2.7)로부터 쉽게 얻을 수 있다.

자료 `jobsatisfaction` (Agresti, 2002)를 써서 수입과 직업만족도가 서로 조건부 독립인지를 Cochran-Mantel-Haenszel 검정하였다. `cmh_test` 함수를 대응하는 `independece_test` 함수로 표현하면 `R-code-3.3.1-2`와 같으며, 이때 c 통계량의 형태는 `quad`이다.

*** R-code-3.3.1-2 ***

```
>cmh_test(jobsatisfaction)
```

```
>independence_test(Job.Satisfaction ~ Income
| Gender, data = jobsatisfaction, weights
= ~ as.vector(jobsatisfaction), teststat =
"quad")#Cochran-Mantel-Haenzsel test
```

3차원 분할표에서 두 순서형 변수의 조건부 독립성을 검정하기 위한 선형 대 선형 연관성 검정 (Goodman, 1979)에 대응하는 `lbl_test`가 `independence_test`와 서로 동일한 검정이 되도록 g와 h를 정의하면 다음과 같다. X와 Y의 범주별 점수를 각각 $u_1, \ldots, u_I$; $v_1, \ldots, v_J$라고 하면, $i = 1, \ldots, n$에 대해,

$$g(X_i) = I(X_i = r_1) u_1 + \cdots + I(X_i = r_I) u_I = s_x(i);$$

$$h(Y_i) = I(Y_i = c_1)v_1 + \cdots + I(Y_i = c_J)v_J = a_{lbl}(i).$$

스코어 벡터 $\boldsymbol{a}_{lbl} = (a_{lbl}(1), \ldots, a_{lbl}(n))'$에 대해, h의 조건부 평균과 분산은 식 (2.3)과 (2.4)로부터 각각 다음과 같다. $l = 1, \ldots, k$에 대해,

$$\bar{a}_{lbl,l} = \frac{1}{N_{\cdot\cdot l}} \sum_{i=1}^{n} I(b_i = l) w_i a_{lbl}(i),\ b^2_{lbl,l} = \frac{1}{N_{\cdot\cdot l}} \sum_{i=1}^{n} I(b_i = l) w_i (a_{lbl}(i) - \bar{a}_{lbl,l})^2.$$

따라서 l번째 블록의 통계량 (2.2)는 다음과 같고,

$$T_{lbl,l} = \sum_{i=1}^{n} I(b_i = l) w_i s_x(i) a_{lbl}(i),\ l = 1, \ldots, k,$$

$T_{lbl,l}$의 조건부 평균과 분산은 식 (2.5)와 (2.6)으로부터 각각 다음과 같다.

$$\mu_{lbl,l} = \bar{a}_{lbl,l} w^{\otimes 1}_{x,l},\ \sigma^2_{lbl,l} = b^2_{lbl,l} \left(\frac{N_{\cdot\cdot l}}{N_{\cdot\cdot l} - 1} w^{\otimes 2}_{x,l} - \frac{1}{N_{\cdot\cdot l} - 1} (w^{\otimes 1}_{x,l})^2 \right),\ l = 1, \ldots, k.$$

단, $d = 1, 2$에 대해 $w^{\otimes d}_{x,l} = \sum_{i=1}^{n} I(b_i = l) w_i s^d_x(i)$이다. 그러므로 통계량 (2.1)은 다음과 같고,

$$T_{lbl} = \sum_{l=1}^{k} T_{lbl,l},$$

T_{lbl}의 조건부 평균과 분산은 식 (2.7)로부터 쉽게 얻을 수 있다.

자료 `jobsatisfaction` (Agresti, 2002)를 순서형 자료로 간주하고 수입과 직업만족도가 서로 조건부 독립인지를 선형 대 선형 연관성 검정하였다. `lbl_test` 함수를 대응하는 `independece_test` 함수로 표현하면 `R-code-3.3.1-3`과 같으며, 이때 c 통계량의 형태는 `quad`이다.

```
>lbl_test(jobsatisfaction)
```

```
*** R-code-3.3.1-3 ***
>independence_test(Job.Satisfaction ~ Income
| Gender, data = jobsatisfaction, weights = ~
as.vector(jobsatisfaction), teststat = "quad",
scores = list(Job.Satisfaction = 1:4, Income =
1:4))#Linear-by-linear association test
```

3.4. 같은 블록 내에 있는 쌍이 종속된 자료

대응비교처럼 한 쌍 (예, 대조군 대 시험군)으로부터 얻어진 관측 값을 행과 열의 범주 값으로 하는 $I \times I$ 2차원 분할표에서, $(X^*_{i'}, Y^*_{i'})$의 값은 서로 다른 두 가지 처리로부터 얻어진 결과를 나타내고, $w^*_{i'}$는 $(X^*_{i'}, Y^*_{i'})$와 동일한 결과를 가진 블록 (혹은 쌍)의 개수, $\sum_{i'=1}^{n'} w^*_{i'}$은 총 블록의 개수를 각각 나타낸다. 단, $i' = 1, \ldots, n'(= I^2)$이다. 두 범주형 변수의 주변합 독립성 (혹은 주변 동질성)을 검정을 위해 $I \times I$ 2차원 분할표를 $k \times 2$ 2차원 분할표로 바꾼다. 단, $k = \sum_{i'=1}^{n'} w^*_{i'}$이다. 변형된 분할표에서 행은 블록을 나타내고 열은 서로 다른 두 처리를 나타낸다. $j = 1, \ldots, k$번째 블록에서, (X_{2j-1}, Y_{2j-1})=(첫 번째 처리 수준=0, 처리 수준 0에서 관측 값), (X_{2j}, Y_{2j})=(두 번째 처리 수준=1, 처리 수준 1에서 관측 값)으로 각각 정의하면, 변형된 관측 자료는 다음과 같다.

$$(X_i, Y_i, b_i, w_i),\ i = 1, \ldots, 2k = n.$$

단, $w_i = 1$, $j = 1, \ldots, k$에 대해, $b_{2j-1} = b_{2j} = j$이다. 두 범주형 변수의 주변합 독립성을 검정하기 위한 Maxwell-Stuart 검정 (Stuart, 1955; Maxwell, 1970)에 대응하는 `mh_test`가 `independence_test`와 서로 동일한 검정이 되도록 g와 h를 정의하면 다음과 같다. $i = 1, \ldots, n$에 대해,

$$g(X_i) = I(X_i = 1),\ h(Y_i) = (I(Y_i = c_1), \ldots, I(Y_i = c_I))' = \boldsymbol{a}_{mh}(i).$$

스코어 벡터 $\boldsymbol{a}_{mh} = (\boldsymbol{a}_{mh}(1)', \ldots, \boldsymbol{a}_{mh}(n)')'$에 대해, h의 조건부 평균 벡터와 공분산 행렬은 식 (2.3)과 (2.4)로부터 각각 다음과 같다. $j = 1, \ldots, k$에 대해,

$$\bar{\boldsymbol{a}}_{mh,j} = \frac{1}{2}\sum_{i=1}^{n} I(b_i = j)\boldsymbol{a}_{mh}(i),$$

$$\boldsymbol{B}_{mh,j} = \frac{1}{2}\sum_{i=1}^{n} I(b_i = j)(\boldsymbol{a}_{mh}(i) - \bar{\boldsymbol{a}}_{mh,j})(\boldsymbol{a}_{mh}(i) - \bar{\boldsymbol{a}}_{mh,j})'.$$

따라서 j번째 블록의 통계량 (2.2)는 다음과 같고,

$$\boldsymbol{T}_{mh,j} = \sum_{i=1}^{n} I(b_i = j)I(X_i = 1)\boldsymbol{a}_{mh}(i),\ j = 1, \ldots, k,$$

$\boldsymbol{T}_{mh,j}$의 조건부 평균 벡터과 공분산 행렬은 식 (2.5)와 (2.6)으로부터 각각 다음과 같다.

$$\boldsymbol{\mu}_{mh,j} = \boldsymbol{a}_{mh,j},\ \boldsymbol{V}_{mh,j} = \boldsymbol{B}_{mh,j},\ j = 1, \ldots, k.$$

그러므로 통계량 (2.1)은 다음과 같고,

$$\boldsymbol{T}_{mh} = \sum_{j=1}^{k} T_{mh,j},$$

$\boldsymbol{T}_{mh}$의 조건부 평균 벡터와 공분산 행렬은 식 (2.7)로부터 쉽게 얻을 수 있다.

자료 `opinions` (Agresti, 2002)은 혼전 성관계와 혼외 성관계가 얼마나 그른지에 대한 정도를 4점 척도로 응답한 2차원 분할표인데, 두 변수에 대한 응답이 서로 주변 독립인지를 Maxwell-Stuart 검정하였다. `mh_test` 함수를 대응하는 `independece_test` 함수로 표현하면 `R-code-3.4-1`과 같으며, 이때 c 통계량의 형태는 `quad`이다.

*** R-code-3.4-1 ***

```
>opinions = c("always wrong", "almost always wrong","wrong only sometimes", "not wrong at all")
>PreExSex = as.table(matrix(c(144, 33, 84, 126, 2, 4, 14, 29,0, 2, 6, 25, 0, 0, 1, 5), nrow = 4,
dimnames = list(PremaritalSex = opinions, ExtramaritalSex = opinions))) #data set
>mh_test(PreExSex)
```

```
>cw = rep(names(margin.table(PreExSex, 2)),
as.vector(margin.table(PreExSex, 2)))
>rw = rep(rep(rownames(PreExSex), times =
dim(PreExSex)[2]), as.vector(PreExSex))
>y = factor(c(rw, cw), levels =
rownames(PreExSex))
>x = c(rep(1, sum(PreExSex)), rep(0,
sum(PreExSex)))
>block = factor(rep(1:sum(PreExSex), 2))
>mh.PreExSex = data.frame(x = x, y = y, block =
block)
>independence_test(y ~ x | block, data =
mh.PreExSex, teststat = "quad")#Maxwell-Stuart
marginal homogeneity test
```

4. 실제 자료 분석 결과

Table 4.1 P-values of fifteen independence tests based on the asymptotic distribution, the number of permutations(1,000, 10,000, or 100,000), and exact distribution

Test	Data set	No. obs	Asymptotic-based	Permutation-based			Exact-based
				1,000	10,000	100,000	
			In case that X is `factor` and Y is `numeric`,				
Wilcoxon-Mann-Whitney	water_transfer	15	0.2207	0.2370	0.2555	0.2540	0.2544
Normal quantiles	water_transfer	15	0.2564	0.2750	0.2707	0.2662	0.2691
Median	water_transfer	15	0.1573	0.2820	0.2788	0.2828	0.2821
Ansari-Bradley	sid	40	0.1815	0.1890	0.1901	0.1879	0.1881
Kruskal-Wallis	YOY	40	4.34e-05	0.0000	0.0000	0.0000	-
Fligner-Killeen	sid	40	0.2237	0.2270	0.2267	0.2258	-
Log rank	ocarcinoma	35	0.0194	0.0120	0.0164	0.0188	0.0182
Wilcoxon signed rank	xydat	18	0.0382	0.0350	0.0424	0.0403	-
Friedman	RoundingTimes	66	0.0038	0.0060	0.0029	0.0032	-
			In case that both X and Y are `numeric`,				
Spearman	USJudgeRatings	43	0.2527	0.2710	0.2637	0.2585	-
Maximally selected statistic	treepipit	86	0.0001	0.0010	0.0007	0.0005	-
			In case both X and Y are `factor`,				
Pearson's chi-square	jobsatisfaction (female only)	64	0.6669	0.6840	0.6893	0.6880	-
Cochran-Mantel-Haenszel	jobsatisfaction	104	0.3345	0.3400	0.3337	0.3300	-
Linear-by-linear association	jobsatisfaction	104	0.0101	0.0140	0.0120	0.0109	-
Maxwell-Stuart	PreExSex	475	0.0000	0.0000	0.0000	0.0000	-

본 절에서는 3절에서 정의한 `independence_test` 함수를 이용하여 3절에서 소개한 자료들을 분석고자 한다. Table 4.1에 자료 집합의 이름과 자료의 크기, 점근 분포에 기초한 p-값, 순열 검정에 기초한 p-값, 정확 분포에 기초한 p-값을 각각 나타냈다. 특히 순열 검정에서는 1,000번, 10,000번, 100,000번 반복 실험한 결과를 각각 수록하였다. `independence_test` 함수는 점근 분포에 기초한 p-값을 디폴트로 제공하기 때문에 순열 검정과 정확 검정에 의한 p-값을 구하기 위해서는 `independence_test` 함수에 옵션 `distribution = approximate(B = b)` 혹은 `distribution = "exact"`를 추가하여야 한다. 단, b =1,000, 10,000 혹은 100,000이다. 정확 검정에 의한 p-값이 제공되지 않는 검정은 하이픈(-)으로 표시하였다. water_transfer 자료의 경우에는 자료의 크기가 충분히 크지 않아서 점근 분포의 p-값이 순열 검정이나 정확 검정의 결과와 약간 다르게 나왔는데, 중앙값 검정의 결과가 Wilcoxon-Mann-Whitney 순위합 검정이나 van der Waerden 정규 분위수 검정보다 상이하였다. 순열 검정의 결과와 정확 검정의 결과를 비교해보면 예상했던 대로 순열의 횟수가 늘어날수록 정확 검정의 결과와 유사해지는 것을 알 수 있었다. 순열 검정에서 자료의 크기가 작을수록 순열의 횟수에 따라 p-값이 더 크게 변동함을 알 수 있었다.

5. 맺음말

본 논문에서는 `coin` 패키지에 내장된 독립성 검정을 위한 간편 함수를 `independence_test` 함수로 표현하였다. 이를 위해, 두 변수 X, Y를 적절히 변환하였으며, 관측 값의 가중 값 및 블록 값에 대해 정의하였다. 또한, 정의한 `independence_test` 함수를 써서 실제 자료의 점근 분포와 순열 검정, 정확 검정에 기초한 유의확률 값을 구하고 그 결과를 서로 비교하였다. 본 논문에서 다룬 15개의 간편 함수는

wilcox_test, normal_test, median_test, ansari_test, kruskal_test, fligner_test, surv_test, wilcoxsign_test, friedman_test, spearman_test, maxstat_test, chisq_test, cmh_test, lbl_test, mh_test이다. 순열 검정 방법은 검정통계량의 분포를 모를 때 유용한데 독립성 검정에 대한 실제 자료 분석에서 살펴본 것처럼 순열의 횟수가 증가함에 따라 정확 검정의 결과에 가까워질뿐만 아니라 자료의 크기가 크면 점근 분포에 의한 결과와도 유사함을 알 수 있었다. 따라서 본 논문에서 살펴본 15개의 독립성 검정 이외 다른 독립성 검정 문제에 대해서도, 본 논문에서 한 것처럼, 두 변수에 대해 적절한 변환을 정의하고, 가중 값과 블록 값을 정의한 후에, 순열 검정 방법에 의해 영가설 분포를 모르는 검정통계량의 유의확률을 구할 수 있을 것으로 기대된다.

References

Agresti, A. (2002), *Categorical data analysis*, Second Edition, Wiley, New York.

Ansari, A. R. and Bradley, R. A. (1960). Rank-sum tests for dispersion. *The Annals of Mathematical Statistics*, **31**, 1174-1189.

Cochran, W. G. (1954). Some methods for strengthening the common chi-squared tests. *Biometrics*, **10**, 417-451.

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

Fleming, T. R., O'Fallon, J. R., O'Brien, P. C., and Harrington, D. P. (1980). Modified Kolmogorov-Smirnov test procedures with applications to arbitrarily censored data. *Biometrics*, **36**, 607-625.

Fligner, M. A. and Killeen, T. J. (1976). Distribution-free two-sample tests for scale. *Journal of the American Statistical Association*, **71**, 210-213.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, **32**, 675-701.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, **74**, 537-552.

Hollander, M. and Wolfe, D. A. (1999). *Nonparametric statistical methods,* Second Edition, Wiley, New York.

Hothorn, T., Kurt Hornik, K., van de Wiel, M. A. and Zeileis, A. (2006). A Lego system for conditional inference. *The American Statistician*, **60**, 257-263.

Hothorn, T., Kurt Hornik, K., van de Wiel, M. A. and Zeileis, A. (2008). Implementing a class of permutation tests: The coin package. *Journal of Statistical Software*, **28**, 1-23.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The satistical analysis of failure time data*, Second Edition, Wiley, New York.

Kruskal, W. H. (1952). A nonparametric test for the several sample problem. *The Annals of Mathematical Statistics*, **23**, 525-540.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, **47**, 583-621.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, **18**, 50-60.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from the retrospective analysis of disease. *Journal of the National Cancer Institute*, **22**, 719-748.

Maxwell, A. E. (1970). Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry*, **116**, 651-655.

Mood, A. M. (1950). *Introduction to the theory of statistics*, McGraw-Hill, New York.

Müller, J. and Hothorn, T. (2004). Maximally selected two-sample statistics as a new tool for the identification and assessment of habitat factors with an application to breeding bird communities in oak forests. *European Journal of Forest Research*, **123**, 219-228.

Pearson, K. (1922). On the chi-square test of goodness of fit. *Biometrika*, **14**, 186-191.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, **15**, 72-101.

Strasser, H. and Weber, C. (1999). On the asymptotic theory of permutation statistics. *Mathematical Methods of Statistics*, **8**, 220-250.

Stuart, A. A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, **42**, 412-416.

Van der Waerden, B. L. (1952). Order tests for two-sample problem and their power 1. *Indagationes Mathematicae*, **14**, 453-458.

Van der Waerden, B. L. (1953a). Order tests for two-sample problem and their power 2. *Indagationes Mathematicae*, **15**, 303-310.

Van der Waerden, B. L. (1953b). Order tests for two-sample problem and their power 3. *Indagationes Mathematicae*, **15**, 311-316.

Westenberg, J. (1948). Significance test for median and interquartile range in samples from continuous populations of any form. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*, **51**, 252-261.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, **1**, 80-83.

Journal of the Korean Data & Information Science Society
2014, **25**(5), 1039–1055

http://dx.doi.org/10.7465/jkdi.2014.25.5.1039
한국데이터정보과학회지

Independence tests using `coin` package in R

Jinheum Kim[1] · Jung-Dong Lee[2]

[12]Department of Applied Statistics, University of Suwon

Received 30 June 2014, revised 28 July 2014, accepted 5 August 2014

Abstract

The distribution of a test statistic under a null hypothesis depends on the unknown distribution of the data and thus is unknown as well. Conditional tests replace the unknown null distribution by the conditional null distribution, that is, the distribution of the test statistic given the observed data. This approach is known as permutation tests and was developed by Fisher (Fisher, 1935). Theoretical framework for permutation tests was given by Strasser and Weber(1999). The `coin` package developed by Hothon *et al.* (2006, 2008) implements a unified approach for conditional inference via the generic `independence_test`. Because convenient functions for the most prominent problems are available, users will not have to use the extremely flexible procedure. In this article we briefly review the underlying theory from Strasser and Weber (1999) and explain how to transform the data to perform the generic function `independence_test`. Finally it was illustrated with a few real data sets.

Keywords: Categorical data, `coin` package, independence, numeric data, permutation test.

[1] Corresponding author: Professor, Department of Applied Statistics, University of Swuon, Gyeonggi 445-743, Korea. E-mail: jkimdt65@gmail.com
[2] Master student, Department of Applied Statistics, University of Swuon, Gyeonggi 445-743, Korea.