The Korean Journal of Applied Statistics (2019) DOI: http://dx.doi.org/10.5351/KJAS.2019.32.6.879
32(6), 879–887

Testing for stochastic order in interval-valued data

Hyejeong Choi[a], · Johan Lim[a] · Minjung Kwak[b] · Seongoh Park[a,1]

[a]Department of Statistics, Seoul National University;
[b]Department of Statistics, Yeungnam University

(Received September 17, 2019; Revised October 2, 2019; Accepted October 21, 2019)

Abstract

We construct a procedure to test the stochastic order of two samples of interval-valued data. We propose a test statistic that belongs to a U-statistic and derive its asymptotic distribution under the null hypothesis. We compare the performance of the newly proposed method with the existing one-sided bivariate Kolmogorov-Smirnov test using real data and simulated data.

Keywords: stochastic order, two-sample test, interval-valued data, blood pressure data

1. 서론

본 연구에서는 이표본 구간 자료의 확률적 순서(stochastic order)를 검정하기 위한 절차에 대하여 논의한다. 구간 자료란 관심 변수가 단일 값으로 측정되지 않고 상한과 하한을 갖는 구간의 형태로 관찰되는 경우를 의미한다. 예를 들어 회사의 주가를 월별로 하한가, 상한가로 요약하여 보고하는 경우 구간 자료를 관찰하게 된다. 또한 본 논문의 동기가 된 자료는 혈압 자료로, 심장 이완기 혈압(diastolic blood pressure; DBP)과 심장 수축기 혈압(systolic blood pressure; SBP)을 각각 하한과 상한으로 갖는 형태이다. 통계학에서는 전통적으로 두 모집단의 동질성을 검정하는 것과 확률적 순서를 검정하는 것을 중요하게 다루어 왔으나 구간 자료에 대한 논의는 극히 제한적인 상황이다. 심지어 구간 자료의 확률적 순서에 대한 명확한 정의조차도 찾기 힘들다. 이에 본 논문에서는 구간 자료의 확률적 순서를 정의하고 이를 검정하는 문제를 다루고자 한다.

본 논문은 다음과 같이 구성되어 있다. 먼저 2절에서는 구간 자료를 순서 제약이 있는 이차원 자료로 이해하고 이를 통하여 이표본 구간 자료의 확률적 순서를 정의한다. 3절에서는 구간 자료의 확률적 순서를 검정하기 위한 점수 기반 U-통계량을 제안하고, 제안한 통계량에 대한 귀무 가설 하에서의 점근적 분포를 U-통계량의 일반 이론을 바탕으로 유도한다. 4절에서는 모의 실험을 통해 제안한 통계량의 성능을 이차원 Kolmogorov-Smirnov (K-S) 통계량을 적절히 변형한 통계량과 비교한다. 검정력을 비교함에 있어 3절에서 계산한 점근 분포의 정확성도 같이 살펴보았다. 5절에서는 제안한 방법을 미국 청소년 여학생을 대상으로 한 코호트 연구 자료에 적용하여 백인 여학생과 아프리카계 여학생들 사이의 혈압에 대한 확률적 순서 검정을 실시하였다. 마지막으로 6절에서는 요약과 함께 본 논문을 마치기로 한다.

[1]Corresponding author: Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea. E-mail: inmybrain@snu.ac.kr

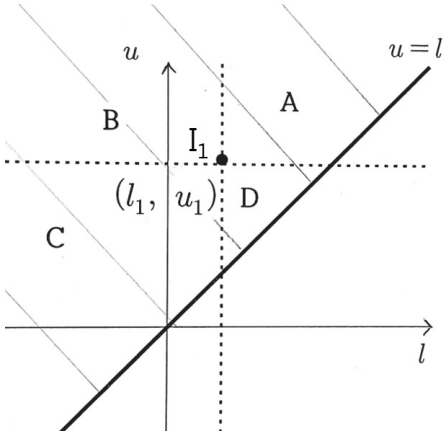

Figure 2.1. A graphical illustration of the order of interval-valued data. $I_1 = (\ell_1, u_1]$을 원점으로 생각하였을 때 제 1, 2, 3, 4분면에 해당하는 공간과 반평면(half-space) $u > \ell$과의 교집합을 구역 A, B, C, D로 정의한다.

2. 단순 확률적 순서(simple stochastic order)

구간 자료의 확률적 순서를 정의하기 전에 일변량 자료의 경우를 살펴보도록 한다. 일변량 확률 변수 X와 Y가

$$\Pr(X > z) \le \Pr(Y > z), \quad \text{모든 } z \in \mathbb{R}$$

를 만족할 때, Y는 확률적으로 X보다 크다고 정의하고 $X \le_{st} Y$처럼 표기한다. 만약 추가로 어떤 z에 대해 $\Pr(X > z) < \Pr(Y > z)$를 만족하면, Y는 확률적으로 X보다 엄격하게(strictly) 크다고 말한다 (Shaked와 Shanthikumar, 2006).

구간 자료의 확률적 순서도 이와 같은 맥락에서 정의할 수 있다. 구간 값 $\mathbf{x} = (\ell_1, u_1]$와 $\mathbf{y} = (\ell_2, u_2]$가 주어졌을 때, 만약 $\ell_1 < \ell_2$이고 $u_1 < u_2$이면 $\mathbf{x} < \mathbf{y}$라고 쓰고 $\mathbf{y}$가 $\mathbf{x}$보다 크다고 읽는다. 이제 구간 확률 변수 $\mathbf{X}$와 $\mathbf{Y}$가 다음의 조건을 만족한다고 하자.

$$\Pr(\mathbf{X} > \mathbf{z}) \le \Pr(\mathbf{Y} > \mathbf{z}), \quad \text{모든 구간 값 } \mathbf{z}. \tag{2.1}$$

이 때, $\mathbf{Y}$가 $\mathbf{X}$보다 확률적으로 크다고 정의하며 $\mathbf{X} \le_{st} \mathbf{Y}$처럼 표기한다. $\mathbf{X}$와 $\mathbf{Y}$의 생존 함수를 각각 $\overline{F}(\mathbf{x}) = \Pr(\mathbf{X} > \mathbf{x})$와 $\overline{G}(\mathbf{y}) = \Pr(\mathbf{Y} > \mathbf{y})$로 표기하면 식 (2.1)은 다음과 동일하다.

$$\overline{F}(\ell, u) \le \overline{G}(\ell, u), \quad \text{for all } (\ell, u) : \ell < u.$$

Figure 2.1은 구간 변수의 순서를 그림으로 묘사한 것이다. 여기서는 구간 $(\ell_1, u_1]$를 좌표 평면 상의 점 (ℓ_1, u_1)으로 표시한다. 그러면 구간 자료는 $\ell < u$와 같은 구조적 제약을 갖기 때문에 관측 가능한 구간 값은 직선 $u = \ell$ 위쪽 공간에 위치하게 된다. $I_1 = (\ell_1, u_1]$와의 순서를 고려하였을 때 임의의 구간 값은 다음의 세 가지 경우 중 하나에 속하게 된다.

1. 구역 A: 해당 구간 값은 $I_1 = (\ell_1, u_1]$보다 크다.
2. 구역 C: 해당 구간 값은 $I_1 = (\ell_1, u_1]$보다 작다.
3. 구역 B 혹은 D: 해당 구간 값은 $I_1 = (\ell_1, u_1]$와 순서를 따질 수 없다.

마지막 경우를 살펴보면, 구역 B에 속한 구간 $(\ell_B, u_B]$은 $(\ell_1, u_1] \subset (\ell_B, u_B]$를 만족하고 구역 D에 있는 $(\ell_D, u_D]$는 $(\ell_D, u_D] \subset (\ell_1, u_1]$를 만족한다.

3. 검정 통계량

독립적으로 관측된 이표본 확률 구간 자료를 생각해보자. 첫 번째 표본 $\mathbf{X}_i = (\ell_{1i}, u_{1i}]$, $i = 1, \ldots, m$은 생존 함수 $\bar{F}$를 갖고, 다른 표본 $\mathbf{Y}_j = (\ell_{2j}, u_{2j}]$, $j = 1, \ldots, n$은 생존 함수 $\bar{G}$를 가진다. 이 때 귀무 가설은 두 표본이 같은 분포에서 나왔는지, 즉, "$\mathcal{H}_0$: 모든 $\mathbf{z}$에 대해 $F(\mathbf{z}) = G(\mathbf{z})$"이고, 대립 가설은 $\mathbf{Y}$가 $\mathbf{X}$보다 확률적으로 엄격하게 크다는 것이다, 즉, "모든 $\mathbf{z}$에 대해 $\overline{F}(\mathbf{z}) \leq \overline{G}(\mathbf{z})$이고 어떤 $\mathbf{z}$에 대해 $\overline{F}(\mathbf{z}) < \overline{G}(\mathbf{z})$"를 의미한다.

본 연구에서 제안하는 확률적 순서 검정 통계량은 다음과 같다.

$$T = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} S_{ij}, \tag{3.1}$$

여기서 S_{ij}는 다음과 같이 정의한다.

$$S_{ij} = \begin{cases} 1, & \text{if } \ell_{1i} < \ell_{2j} \text{ and } u_{1i} < u_{2j}, \\ -1, & \text{if } \ell_{1i} > \ell_{2j} \text{ and } u_{1i} > u_{2j}, \\ 0, & \text{otherwise.} \end{cases}$$

귀무 가설 $\bar{F} = \bar{G}$ 하에서는 $\Pr(S_{ij} = 1) = \Pr(S_{ij} = -1)$이므로 $\mathbb{E}(T) = 0$임을 알 수 있다.

T는 U-통계량에 해당하기 때문에 이에 대한 점근 이론을 이용하여 귀무 가설 하에서 T의 점근 근사 분포를 계산할 수 있다. Lehmann (1999)의 6장에 등장하는 U-통계량의 일반적인 점근 이론을 소개하면 아래와 같다. 우선 $\phi(x_1, \ldots, x_a; y_1, \ldots, y_b)$는 $a + b$개 $(1 \leq a \leq m, 1 \leq b \leq n)$의 입력 변수를 받는 대칭형 커널 함수라고 하자. 여기서 대칭형 커널 함수는 입력 변수 $(x_1, \ldots, x_a)$들의 (또는 $y_1, \ldots, y_b$들의) 순서를 바꾸어도 함수 값이 변하지 않는 함수를 지칭한다. 관심 모수가

$$\theta = \theta(\bar{F}, \bar{G}) = \mathbb{E}[\phi(\mathbf{X}_1, \ldots, \mathbf{X}_a; \mathbf{Y}_1, \ldots, \mathbf{Y}_b)]$$

로 표현되는 경우 이에 대한 U-통계량은 다음과 같이 정의된다.

$$U_{m,n} = \binom{m}{a}^{-1} \binom{n}{b}^{-1} \sum_{C_{m,a}} \sum_{C_{n,b}} \phi(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_a}; \mathbf{Y}_{j_1}, \ldots, \mathbf{Y}_{j_b}), \tag{3.2}$$

여기서 $C_{k,t}$는 크기가 t인 $\{1, 2, \ldots, k\}$의 모든 부분 집합의 모임이고 합 기호에 들어가는 가변수는 $(i_1, \ldots, i_a)$와 $(j_1, \ldots, j_b)$이다. $U_{m,n}$ 는 θ에 대한 불편(unbiased) 추정량이 되며 이의 분산은

$$\mathrm{Var}(U_{m,n}) = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{\binom{a}{i}\binom{m-a}{a-i}}{\binom{m}{a}} \frac{\binom{b}{j}\binom{n-b}{b-j}}{\binom{n}{b}} \sigma_{ij}^2$$

으로 계산된다. 여기서 σ_{ij}^2는

$$\begin{aligned} \sigma_{ij}^2 = \mathrm{Cov}\, [&\phi(\mathbf{X}_1, \ldots, \mathbf{X}_i, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_a; \mathbf{Y}_1, \ldots, \mathbf{Y}_j, \mathbf{Y}_{j+1}, \ldots, \mathbf{Y}_b), \\ &\phi(\mathbf{X}_1, \ldots, \mathbf{X}_i, \mathbf{X}'_{i+1}, \ldots, \mathbf{X}'_a; \mathbf{Y}_1, \ldots, \mathbf{Y}_j, \mathbf{Y}'_{j+1}, \ldots, \mathbf{Y}'_b)\big] \end{aligned}$$

이고 $\mathbf{X}'_i$와 $\mathbf{Y}'_j$는 $\mathbf{X}_i$와 $\mathbf{Y}_j$의 복제 변수로, 원본과 분포가 같으나 독립인 확률 변수를 가리킨다. Lehmann (1999)의 6장에서 발췌한 다음의 정리는 식 (3.2)의 U-통계량의 점근 분포를 설명해준다.

정리 3.1 (Lehmann (1999), Theorem 6.1.3 (ii)) $m/N \to \rho \in (0,1)$이고 $N = m+n \to \infty$일 때, $\sqrt{N}\big(U_{m,n} - \theta\big)$는 평균이 0인 정규 분포로 분포 수렴하고, 이의 분산은 $\sigma^2 = (a^2/\rho)\sigma_{10}^2 + (b^2/(1-\rho))\sigma_{01}^2$이다. 여기서 σ_{10}^2와 σ_{01}^2은 다음과 같이 계산된다.

$$\sigma_{10}^2 = \mathrm{Cov}[\phi(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_a;\ \mathbf{Y}_1, \ldots, \mathbf{Y}_b),\ \phi(\mathbf{X}_1, \mathbf{X}_2', \ldots, \mathbf{X}_a';\ \mathbf{Y}_1', \ldots, \mathbf{Y}_b')] \in (0, \infty),$$
$$\sigma_{01}^2 = \mathrm{Cov}[\phi(\mathbf{X}_1, \ldots, \mathbf{X}_a;\ \mathbf{Y}_1, \mathbf{Y}_2 \ldots, \mathbf{Y}_b),\ \phi(\mathbf{X}_1', \ldots, \mathbf{X}_a';\ \mathbf{Y}_1, \mathbf{Y}_2', \ldots, \mathbf{Y}_b')] \in (0, \infty).$$

위의 U-통계량의 점근 분포에 대한 일반 정리를 이용하여 식 (3.1)의 통계량 T의 점근 분포를 유도할 수 있다.

정리 3.2 귀무 가설 $\mathcal{H}_0 : F = G$ 하에서, 만약 $N = m+n \to \infty$이고 $m/N \to \rho \in (0,1)$라면, 다음의 수렴 결과를 유도할 수 있다.

$$\sqrt{N}T \ \xrightarrow{d}\ N\left(0,\ \frac{\theta_1 + \theta_2 - 2\theta_3}{\rho(1-\rho)}\right),$$

여기서 $\theta_1 = \Pr(\mathbf{X} < \min(\mathbf{Y}, \mathbf{Y}'))$, $\theta_2 = \Pr(\max(\mathbf{Y}, \mathbf{Y}') < \mathbf{X})$, 그리고 $\theta_3 = \Pr(\mathbf{Y}' < \mathbf{X} < \mathbf{Y})$이다.

위 점근 분포의 분산에 이용되는 모수 $\theta_1, \theta_2, \theta_3$은 각 표본 내에서의 순열 방법을 이용하여 근사하여 얻어낼 수 있다. 우선 다음의 관찰이 필요하다. 예를 들어,

$$\begin{aligned}\theta_1 &= \Pr\left(\mathbf{X} < \min\left(\mathbf{Y}, \mathbf{Y}'\right) | F = G\right) = \Pr\left(\mathbf{X} < \min\left(\mathbf{X}', \mathbf{X}''\right) | F = G\right) \\ &= \Pr\left(\mathbf{X} < \min\left(\mathbf{X}', \mathbf{X}''\right)\right) \left(= \Pr\left(\mathbf{Y} < \min\left(\mathbf{Y}', \mathbf{Y}''\right)\right)\right), \end{aligned} \tag{3.3}$$

여기서 $\mathbf{X}, \mathbf{X}', \mathbf{X}''$는 첫 번째 모집단의 독립인 확률 구간들이다. 따라서, θ_1의 근사는 다음과 같다.

$$\hat{\theta}_1 = \frac{\sum_{i,j,k:\mathrm{distinct}} \mathrm{I}\big(\mathbf{X}_i < \min(\mathbf{X}_j, \mathbf{X}_k)\big)}{2m(m-1)(m-2)} + \frac{\sum_{i,j,k:\mathrm{distinct}} \mathrm{I}\big(\mathbf{Y}_i < \min(\mathbf{Y}_j, \mathbf{Y}_k)\big)}{2n(n-1)(n-2)}.$$

계산 (3.3)으로부터 위 근사는 대립 가설 하의 모집단에서도 θ_1의 추정량이 되는 것을 알 수 있다.

증명: $\mathbf{x} = (\ell_1, u_1], \mathbf{y} = (\ell_2, u_2]$에 대해 $\phi(\mathbf{x}; \mathbf{y}) = \mathrm{I}(\mathbf{x} < \mathbf{y}) - \mathrm{I}(\mathbf{x} > \mathbf{y}) = \mathrm{I}(\ell_1 < \ell_2, u_1 < u_2) - \mathrm{I}(\ell_1 > \ell_2, u_1 > u_2)$라고 정의하면, T는 $a = b = 1$인 이표본 U-통계량

$$U_{m,n} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\phi(\mathbf{X}_i; \mathbf{Y}_j)$$

으로 표현 가능하다. 따라서 정리 3.1를 적용하여 다음을 얻는다.

$$\sqrt{N}(U_{m,n} - \theta) \ \xrightarrow{d}\ N\left(0,\ \frac{\sigma_{10}^2}{\rho} + \frac{\sigma_{01}^2}{1-\rho}\right),$$

여기서 $\theta = \mathbb{E}(\phi(\mathbf{X}; \mathbf{Y}))$, $\rho = \lim(m/N) \in (0,1)$, $\sigma_{10}^2 = \mathrm{Cov}\left[\phi(\mathbf{X}; \mathbf{Y}), \phi(\mathbf{X}; \mathbf{Y}')\right]$이고 $\sigma_{01}^2 = \mathrm{Cov}\left[\phi(\mathbf{X}; \mathbf{Y}), \phi(\mathbf{X}'; \mathbf{Y})\right]$이다.

확률 구간 변수를 $\mathbf{X} = (L_1, U_1)$, $\mathbf{Y} = (L_2, U_2)$, 그리고 $\mathbf{Y}' = (L_2', U_2')$로 표기하자. 귀무 가설 $\bar{F} = \bar{G}$ 하에서 $\theta = \mathbb{E}(\phi(\mathbf{X}; \mathbf{Y})) = \Pr(L_1 < L_2, U_1 < U_2) - \Pr(L_1 > L_2, U_1 > U_2) = 0$가 성립한다. 분산의 구성 요소인 σ_{10}^2 $(= \sigma_{01}^2)$는 다음과 같이 계산된다.

$$\begin{aligned}\sigma_{10}^2 &= \mathrm{Cov}\left[\phi\left(\mathbf{X}; \mathbf{Y}\right), \phi(\mathbf{X}; \mathbf{Y}')\right] \\ &= \mathbb{E}\left[\phi(\mathbf{X}; \mathbf{Y})\phi(\mathbf{X}; \mathbf{Y}')\right] \quad (\because \theta = 0) \\ &= \mathbb{E}[\mathrm{I}(\mathbf{X}<\mathbf{Y})\mathrm{I}(\mathbf{X}<\mathbf{Y}')] - \mathbb{E}[\mathrm{I}(\mathbf{X}<\mathbf{Y})\mathrm{I}(\mathbf{X}>\mathbf{Y}')] - \mathbb{E}[\mathrm{I}(\mathbf{X}>\mathbf{Y})\mathrm{I}(\mathbf{X}<\mathbf{Y}')] + \mathbb{E}[\mathrm{I}(\mathbf{X}>\mathbf{Y})\mathrm{I}(\mathbf{X}>\mathbf{Y}')] \\ &= \Pr\big(\mathbf{X} < \min(\mathbf{Y}, \mathbf{Y}')\big) - \Pr\big(\mathbf{Y}' < \mathbf{X} < \mathbf{Y}\big) - \Pr\big(\mathbf{Y} < \mathbf{X} < \mathbf{Y}'\big) + \Pr\big(\max(\mathbf{Y}, \mathbf{Y}') < \mathbf{X}\big).\end{aligned}$$

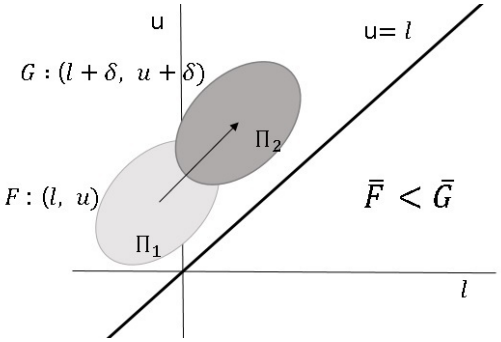

Figure 4.1. A graphical illustration of two populations in the simulation study.

이 때, $\theta_1 = \Pr\big(\mathbf{X} < \min(\mathbf{Y}, \mathbf{Y}')\big)$, $\theta_2 = \Pr\big(\max(\mathbf{Y}, \mathbf{Y}') < \mathbf{X}\big)$, 그리고 $\theta_3 = \Pr\big(\mathbf{Y}' < \mathbf{X} < \mathbf{Y}\big)$로 정의하자. 따라서 $\bar{F} = \bar{G}$일 때,

$$\sigma_{10}^2 = \sigma_{01}^2 = \theta_1 + \theta_2 - 2\theta_3$$

를 얻는다. 그러므로 $\sqrt{N}T(= \sqrt{N}U_{m,n})$의 점근 분포는 다음의 분산을 갖는다.

$$\frac{\sigma_{10}^2}{\rho} + \frac{\sigma_{01}^2}{1-\rho} = \frac{\theta_1 + \theta_2 - 2\theta_3}{\rho(1-\rho)}.$$

□

4. 모의 실험

이번 절에서는 새로 제안한 검정 통계량("U-검정"으로 표기)의 성능을 단측 이변량 Kolmogorov-Smirnov 검정("K-S 검정"으로 표기)의 성능과 비교하기로 한다. 우선 U-검정은 귀무 분포를 근사하는 방법에 따라 두 가지로 나뉜다. U-perm은 U-검정의 귀무 분포를 순열 방법으로 근사한 방법이고 U-asym은 정리 3.2의 근사 분포를 이용한 방법이다. K-S 검정은 대립 가설이 $\overline{F} < \overline{G}$이고 통계량은 다음과 같다 (Feller, 1948).

$$D_{m,n}^{+} = \left(\frac{mn}{m+n}\right)^{\frac{1}{2}} \sup_{s,t\in\mathbb{R},\ s<t} \left(\widehat{F}_m(s,t) - \widehat{G}_n(s,t)\right),$$

여기서 $\widehat{F}_m(s,t) = (1/m)\sum_{i=1}^{m} \mathrm{I}(L_{1i} \le s, U_{1i} \le t)$이고 $\widehat{G}_n(s,t) = (1/n)\sum_{j=1}^{n} \mathrm{I}(L_{2j} \le s, U_{2j} \le t)$이다. $D_{m,n}^{+}$의 귀무 분포는 순열 방법으로 근사하였다 (Gail과 Green, 1976).

본 모의 실험에서는 구간 자료 $(L, U]$를 생성하기 위하여 중심 $C = (L+U)/2$과 범위 $R = (U-L)/2$로 구성된 이변수 $(C, \log R)$의 확률 분포로 다음과 같은 이변량 정규 분포와 자유도 5의 이변량 $t-$분포를 고려한다.

$$N\left(\begin{pmatrix}\mu_{\mathrm{C}}\\ \mu_{\mathrm{R}}\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right) \quad \text{혹은} \quad t_5\left(\begin{pmatrix}\mu_{\mathrm{C}}\\ \mu_{\mathrm{R}}\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$$

그리고 이표본으로 다음의 두 모집단 Π_1와 Π_2를 생각한다.

$$\Pi_1 : \mu_1 = (0, 0),$$
$$\Pi_2 : \mu_2 = (\delta, 0).$$

Table 4.1. Empirical powers of three methods (U-perm, U-asym, and B-KS) for testing stochastic order

Case	(m,n)	δ	$\rho=0$			$\rho=0.4$			$\rho=0.8$		
			U-perm	U-asym	B-KS	U-perm	U-asym	B-KS	U-perm	U-asym	B-KS
(N)	$(30,30)$	0.0	0.045	0.044	0.042	0.043	0.042	0.041	0.045	0.045	0.040
		0.3	0.301	0.293	0.158	0.307	0.306	0.178	0.425	0.425	0.289
		0.5	0.573	0.568	0.321	0.599	0.598	0.366	0.789	0.788	0.630
		1.0	0.980	0.979	0.829	0.988	0.988	0.900	0.999	0.999	0.995
	$(30,120)$	0.0	0.049	0.047	0.052	0.051	0.050	0.058	0.050	0.050	0.046
		0.3	0.396	0.393	0.267	0.422	0.420	0.312	0.578	0.578	0.489
		0.5	0.745	0.744	0.551	0.781	0.780	0.619	0.929	0.928	0.876
		1.0	0.999	0.999	0.979	1.000	1.000	0.991	1.000	1.000	1.000
	$(50,50)$	0.0	0.055	0.055	0.042	0.054	0.054	0.040	0.049	0.049	0.040
		0.3	0.411	0.412	0.252	0.436	0.439	0.287	0.589	0.590	0.476
		0.5	0.756	0.757	0.525	0.790	0.792	0.605	0.936	0.937	0.873
		1.0	0.999	0.999	0.973	1.000	1.000	0.992	1.000	1.000	1.000
	$(50,200)$	0.0	0.052	0.051	0.040	0.052	0.050	0.047	0.057	0.056	0.048
		0.3	0.557	0.556	0.378	0.602	0.590	0.462	0.775	0.776	0.709
		0.5	0.904	0.903	0.733	0.925	0.922	0.831	0.987	0.987	0.975
		1.0	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000
(T)	$(30,30)$	0.0	0.055	0.052	0.042	0.050	0.050	0.045	0.050	0.042	0.047
		0.3	0.239	0.238	0.171	0.253	0.265	0.188	0.334	0.354	0.271
		0.5	0.467	0.488	0.302	0.491	0.518	0.346	0.663	0.664	0.542
		1.0	0.934	0.936	0.752	0.949	0.952	0.810	0.991	0.991	0.919
	$(30,120)$	0.0	0.052	0.048	0.053	0.051	0.048	0.048	0.053	0.044	0.046
		0.3	0.349	0.324	0.225	0.370	0.351	0.246	0.479	0.476	0.344
		0.5	0.650	0.634	0.419	0.685	0.676	0.446	0.843	0.844	0.632
		1.0	0.987	0.988	0.817	0.993	0.992	0.849	1.000	1.000	0.867
	$(50,50)$	0.0	0.053	0.052	0.044	0.049	0.050	0.044	0.046	0.044	0.040
		0.3	0.350	0.333	0.215	0.367	0.362	0.246	0.490	0.486	0.361
		0.5	0.661	0.650	0.426	0.686	0.691	0.490	0.852	0.849	0.687
		1.0	0.993	0.993	0.852	0.996	0.998	0.881	1.000	1.000	0.893
	$(50,200)$	0.0	0.050	0.054	0.052	0.049	0.048	0.051	0.051	0.046	0.056
		0.3	0.482	0.494	0.278	0.499	0.494	0.306	0.690	0.644	0.453
		0.5	0.845	0.860	0.517	0.863	0.860	0.563	0.965	0.958	0.708
		1.0	1.000	1.000	0.825	1.000	1.000	0.834	1.000	1.000	0.843

첫 번째 열은 $(C, \log R)$의 분포를 가리키며 N은 정규 분포, T는 $t-$분포이다. U-perm은 U-검정의 귀무 분포를 순열 방법으로 근사한 방법이고 U-asym은 정리 3.2의 근사 분포를 이용한 방법이다. B-KS는 K-S 검정에 해당한다.

위의 설정에서 δ는 다음의 네 가지 값 $(0, 0.3, 0.5, 1.0)$을 고려한다. $\delta > 0$일 때 대립 가설에 해당하는 것을 알 수 있다. Figure 4.1은 모의 실험 설정을 그림으로 묘사하고 있다. 중심과 범위의 상관성이 미치는 영향을 알아보기 위해 다음의 값 $\rho = (0, 0.4, 0.8)$을 고려하였다. 검정의 유의 수준 α는 0.05로 고정하였다. 검정 절차의 제 1종 오류와 검정력은 2,000번의 반복 실험 중 기각한 비율로 계산하였다. K-S 검정의 귀무 분포는 $\delta = 0$의 가정 하에 생성된 20,000개의 반복 자료를 통하여 경험적으로 근사하였다. 표본의 크기로 $(m,n) = (30,30), (30,120), (50,50), (50,200)$의 네 가지 경우를 고려하였다. 모의 실험 결과는 Table 4.1에 정리되어 있다.

Table 5.1. Descriptive statistics of the BP data by race

	Caucasian	African-American	p-value
Center	78.67 (9.09)	80.13 (8.03)	< 0.001
DBP	56.72 (12.19)	58.03 (11.72)	0.005
SBP	100.62 (9.28)	102.23 (8.65)	< 0.001
Half-range	21.95 (5.89)	22.10 (6.44)	0.279

각 변수의 평균과 표준 편차(괄호 안)를 계산하였다. 마지막 열의 "p-value"는 "아프리카계 미국인의 혈압이 백인보다 높다"라는 대립 가설에 대한 이표본 t-검정의 유의 확률이다.

Table 5.2. Test results of the BP data

	U-perm	U-asym	B-KS
p-value	< 0.001	< 0.001	< 0.001

U-perm은 U-검정의 귀무 분포를 순열 방법으로 근사한 방법이고 U-asym은 정리 3.2의 근사 분포를 이용한 방법이다. B-KS는 K-S 검정에 해당한다.

Table 4.1은 제안한 U-검정과 관련하여 몇 가지 흥미로운 결과를 보여준다. 첫 번째로 가장 중요하게 우리가 고려한 모든 경우에서 U-검정이 K-S 검정보다 더 높은 검정력을 보여주고 있다. 둘 째로 모든 경우에 U-검정에서 귀무 분포를 순열 방법을 이용하여 경험적으로 근사한 결과와 점근 분포를 이용한 결과가 크게 다르지 않은 것을 통해 점근 결과의 정확성을 확인할 수 있다. 셋 째, 중심과 범위의 상관성이 강해지면 검정력이 증가하는 경향도 관찰할 수 있다. 이 현상은 두 모집단 평균 사이의 Mahalanobis 거리로 설명할 수 있다. 두 평균 사이의 거리는 $(\delta, 0)\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1}(\delta, 0) = \delta^2/(1-\rho^2)$이며 ρ에 대한 증가 함수인 것을 알 수 있다. 예를 들어 ρ가 0, 0.4, 그리고 0.8로 커지면 거리는 δ^2, $1.2\delta^2$, 그리고 $2.8\delta^2$처럼 커지게 된다.

5. 실제 자료 분석

이번 절에서는 본 연구에서 제안한 방법을 실제 자료에 적용하여 본다. 자료는 10년 길이의 코호트(cohort) 연구인 National Heart, Lung, and Blood Institute Growth and Health Study (NGHS)의 일부이다. 사용된 자료는 아프리카계 미국인과 백인 여학생들 2,379명을 대상으로 심혈관 위험 인자의 시간적 추세를 연구하기 위해 매년 SBP과 DBP을 측정한 것이다. 혈압 자료는 수축기와 이완기마다 한 번 씩 관측되고 항상 수축기 혈압이 이완기 혈압보다 크므로 구간 자료로 볼 수 있다. 본 연구에서는 첫 번째 방문에 측정한 자료만 사용하고 결측치를 포함한 표본은 제거하였다. 최종적으로 분석에 사용한 표본은 총 $N = 2{,}256$명이며, 그 중 백인 집단은 $m = 1{,}112$명, 아프리카계 미국인은 $n = 1{,}144$명으로 구성된다. 본 분석에서는 "아프리카계 여학생의 혈압이 백인 여학생의 혈압 보다 확률적으로 크다"라는 가설을 검정하고자 한다.

Table 5.1을 보면 수축기 혈압, 이완기 혈압, 두 혈압의 중심에서 아프리카계 미국인이 백인보다 유의미하게 높은 값을 가진다. 반면 범위 값은 두 집단 간의 차이가 뚜렷하지 않은 것으로 확인된다. 이 결과들은 모의 실험의 설정과 비슷하다.

이제 주변 분포가 아닌 구간 자료를 이용하여 아프리카계 미국인의 혈압이 백인보다 확률적으로 큰지 검정을 해보자. Table 5.2는 앞서 비교하였던 세 가지 검정 절차를 적용한 결과이다. 모든 방법에서 유의확률이 0.001보다 작았고 이는 아프리카계 미국인의 혈압이 백인보다 확률적으로 크다는 것을 의미한다.

6. 결론

본 연구에서는 구간 자료의 확률적 순서를 정의하고 U-통계량에 기초한 U-검정 절차를 제안하였고 이에 대한 점근적 귀무 분포를 유도하였다. 모의 실험을 통해 표본이 크지 않더라도 점근 분포가 꽤 정확하게 귀무 분포를 근사하는 것을 알 수 있었다. 또한 고려한 모든 경우에 있어 U-검정 절차가 K-S 검정법보다 우월한 성능을 보여주었다. 따라서 제안한 검정법이 구간 자료의 확률적 순서를 검정하는데 유용하다는 점을 알 수 있었다.

References

Feller, W. (1948). On the Kolmogorov-Smirnov limit theorems for empirical distributions, *The Annals of Mathematical Statistics*, **19**, 177–189.

Gail, M. and Green, S. (1976). Critical values for the one-sided two-sample Kolmogorov-Smirnov Statistic, *Journal of the American Statistical Association*, **71**, 757–760.

Lehmann, E. L. (1999). *Elements of Large Sample Theory*, Springer, New York.

Shaked, M. and Shanthikumar, J. G. (2006). *Stochastic Orders*, Springer, New York.

구간 자료의 확률적 순서 검정

최혜정[a] · 임요한[a] · 곽민정[b] · 박성오[a,1]

[a]서울대학교 통계학과, [b]영남대학교 통계학과

(2019년 9월 17일 접수, 2019년 10월 2일 수정, 2019년 10월 21일 채택)

요 약

본 연구에서는 이표본 구간 자료의 확률적 순서 검정 절차를 제안한다. 제안하는 검정 통계량은 U-통계량에 해당하며 본 연구에서는 이에 대한 점근적 분포를 귀무 가설 하에서 유도하였다. 실제 자료와 모의 실험을 통해 새로 제안한 방법의 성능을 단측 이변량 Kolmogorov-Smirnov 검정법과 비교한다.

주요용어: 확률적 순서, 이표본 검정, 구간 자료, 혈압 자료

[1]교신저자: (08826) 서울시 관악구 관악로 1, 서울대학교 통계학과. E-mail: inmybrain@snu.ac.kr