The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.4.513
30(4), 513–528

Power comparison of distribution-free two sample goodness-of-fit tests

Seon Bin Kim[a] · Jae Won Lee[a,1]

[a]Department of Statistics, Korea University

(Received February 8, 2017; Revised March 20, 2017; Accepted June 14, 2017)

Abstract

Statistics are often used to test two samples if they have been drawn from the same underlying distribution. In this paper, we introduce several well-known distribution-free tests to compare distributions and conduct an extensive Monte-Carlo simulation to specify their behaviors. We consider various circumstances of when two distributions vary in (1) location, (2) scale, (3) symmetry, (4) kurtosis, (5) tail weight. A practical guideline for two-sample goodness-of-fit test is presented based on the simulation result.

Keywords: Kolmogorov-Smirnov test, Cramér-von-Mises test, Anderson-Darling test, Epps-Singleton test, Mann-Whitney U test, goodness-of-fit test

1. 서론

두 표본 집단이 동일한 분포를 따르는지를 통계적으로 검정하기 위한 분포무관 검정법에는 여러가지가 있다. 대표적인 이표본 분포 동일성 문제의 분포무관 검정은 Kolmogorov-Smirnov (KS) 검정, Cramér-von-Mises (CVM) 검정, Anderson-Darling (AD) 검정, Mann-Whitney (MW) 검정이 있고 상대적으로 잘 알려지지 않은 Epps-Singleton (ES) 검정이 존재한다.

이표본 KS 검정은 Smirnoff (1939)가 제안한 검정법으로 두 표본의 경험적분포간 거리의 최대값에 근거한다. 이표본 CVM 검정은 Cramér (1928)와 von Mises (1928)의 일표본 CVM 검정법을 Anderson (1962)이 이표본 검정으로 변형한 것이다. 이표본 CVM 통계량은 두 경험적분포간 거리의 적분제곱을 바탕으로 한 것으로 일반적인 경우에 CVM 검정은 KS 검정보다 높은 검정력을 가진다고 알려져 있다. 이표본 AD 검정은 Anderson과 Darling (1952)이 제안한 일표본 AD 검정의 확장이다. AD 통계량은 CVM 통계량과 같이 두 분포간 거리의 적분제곱에 바탕을 두지만 분포의 꼬리 부분의 차이에 대해 더 큰 가중치를 두는 특성이 있다. ES 검정은 Epps와 Singleton (1986)에 의해 제안되었으며 경험적 특성 함수의 차이에 기반한다. 연속형 자료를 가정하는 KS, AD, CVM 검정과는 달리 ES 검정은 이산형 자료에 대해서도 검정을 시행할 수 있다는 장점을 가지고 있으며, Epps와 Singleton (1986)에 의하면 특정 조건에서 ES 검정은 AD 검정 또는 CVM 검정과 비슷하거나 더 나은 성능을 보인다고

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2016943438).
[1]Corresponding author: Department of Statistics, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul 02841, Korea. E-mail: jael@korea.ac.kr

알려져 있다. MW 검정은 Wilcoxon (1945)이 제안한 동일 표본 크기에 대한 순위합검정을 Mann과 Whitney (1947)가 다른 표본 크기에 대해서도 적용할 수 있는 다소 다른 방식으로 변형하여 제안한 것이다. MW 검정은 ES 검정과 같이 이산형 자료에 대해서도 적용 가능하다는 장점을 가지고 있고 (Goerg와 Kaiser, 2009), 두 표본의 독립성을 검정하는 문제에서 널리 사용되지만 위치(location)에 대해서만 검정한다는 특성을 가지고 있다.

이처럼 이표본 분포 동일성 문제를 검정하는 많은 분포 무관 검정법이 존재한다. 하지만 일표본 문제와는 달리 이표본 문제에 대해서는 여러 검정법을 망라한 비교 연구가 존재하지 않고 각 연구마다 각기 다른 실험 조건 하에서 각기 다른 결과를 제시하고 있어 실제 검정법의 적용에 명확한 기준점이 존재하지 않고 있다. 본 논문에서는 두 표본 분포의 동질성 검정법을 소개하고 여러 가지 상황 하에서 검정력을 평가하는 모의실험을 진행하여 여러 상황을 고려한 실용적인 지침을 제공하려고 한다. 2절에서는 두 표본 분포의 동질성을 검정하는 분포무관 검정법들을 소개한다. 3절에서는 지금까지의 비교 연구에 대해서 다루고, 4절에서는 모의실험 결과를 제시하도록 한다. 5절에서는 두 개의 실제 자료에 적용해 보기로 한다.

2. 두 표본 간 동일성을 검정하는 분포무관 검정

연속누적분포함수 $F(\cdot)$과 $G(\cdot)$에서 독립적으로 추출된 확률 표본 $X_1, \ldots, X_m$과 $Y_1, \ldots, Y_n$을 고려하자. X와 Y가 따르는 경험적누적분포함수(empirical distribution function)를 각각 $F_m(x) = P[X_i \leq x]$, $i = 1, \ldots, m$과 $G_n(x) = P[Y_j \leq x]$, $j = 1, \ldots, n$이라고 하면 두 표본 분포의 동일성 검정의 귀무가설은

$$H_0 : F = G \tag{2.1}$$

이다. 다음에 제시되는 모든 검정에 대해, H_0는 충분히 큰 통계량에 의해 기각된다.

2.1. 이표본 Kolmogorov-Smirnov 검정법

KS 검정은 Smirnoff (1939)에 의해 제안된 검정법으로 두 표본 간 분포의 동일성을 검정하는 데에 가장 널리 사용되는 검정이다. KS 검정은 다음과 같은 통계량을 사용한다.

$$D = \sup_x |F_m(x) - G_n(x)|, \tag{2.2}$$

여기서 $F_m(x)$와 $G_n(x)$은 각각 X와 Y에 대응하는 경험적분포함수이다. KS 통계량의 임계값은 Massey (1951)의 연구에 잘 정리되어 있다. KS 통계량은 분포의 형상에 민감한 특성을 가지고 있고 (Darling, 1957), 적은 표본 수에 대해서도 적용 가능하다고 알려져 있다 (Lilliefors, 1969).

2.2. 이표본 Cramér-von-Mises 검정법

CVM 검정 역시 널리 이용되는 이표본 분포무관 검정이다. 이표본 CVM 검정법은 Cramér (1928)와 von Mises (1928)가 도입한 일표본 CVM 검정법을 Anderson (1962)이 변형한 것이다. CVM 검정은 다음과 같이 정의된 T_2 통계량에 기반한다.

$$T_2 = \frac{mn}{(m+n)^2} \left\{ \sum_{i=1}^{m} (F_m(X_i) - G_n(X_i))^2 + \sum_{j=1}^{n} (F_m(Y_j) - G_n(Y_j))^2 \right\}, \tag{2.3}$$

여기서 F_m과 G_n은 각각 표본 X와 표본 Y에 대응하는 경험적분포함수이다. $X_1, \ldots, X_m$과 $Y_1, \ldots, Y_n$의 결합된 순서 표본 $\xi_1, \ldots, \xi_m$와 $\eta_1, \ldots, \eta_n$에 대응하는 경험적 분포함수 F_m^*과 G_n^*을 대응하는 경험적분포함수라고 했을때 T_2는 m, n에만 의존하는 상수가 되며 F_m^*과 G_n^*의 L_2-거리의 제곱에 해당한다. 귀무가설 하의 T_2의 분포는 Anderson (1962)과 Burr (1964)의 연구에서 자세히 다루고 있다.

2.3. 이표본 Anderson-Darling 검정법

이표본 AD 검정은 Darling (1957)과 Pettitt (1976)에 의해 제안된 검정으로 다음의 통계량을 갖는다.

$$\mathrm{AD} = \frac{1}{mn} \sum_{i=1}^{n+m} \left(N_i Z_{(n+m-ni)}\right)^2 \frac{1}{iZ_{(n+m-i)}}, \tag{2.4}$$

여기서 $Z_{(n+m)}$는 X과 Y의 결합된 순서 표본이다. N_i는 $Z_{(m+n)}$의 i번째 관측치보다 작거나 같은 X에 속한 관측치의 개수이다. AD 검정의 분포에 대한 연구와 임계값은 Pettitt (1976)의 연구에 제시되어 있다. AD 검정은 Scholz와 Stephens (1987)에 의해 k-표본 검정으로 일반화되었다.

2.4. Epps-Singleton 검정

ES 검정은 Epps와 Singleton (1986)에 의해 제안되었다. KS 검정, CVM 검정과 AD 검정은 경험적 분포를 비교하는 데에 비해, ES 검정은 두 표본에 대응하는 경험적특성함수(empirical characteristic function)의 차이를 검정한다.경험적특성함수는 경험적분포함수의 푸리에(Fourier) 변환이다. 자료가 이산형일 때, 분포함수는 특정 점에서만 정의되는 데 비해 특성함수는 모든 점에서 정의되기 때문에 특성함수에 기반한 ES 검정은 이산형 자료에 대해서도 적용할 수 있다. 식 (2.1)의 귀무가설은 특성함수를 사용하여 다음과 같이 나타낼 수 있다.

$$H_0 : \phi_1(t) = \phi_2(t), \quad -\infty < t < \infty, \tag{2.5}$$

여기서 $\phi_1(t)$와 $\phi_2(t)$는 각각 X와 Y의 t점에서의 특성함수이다. ES 검정을 적용하기 위해서 경험적특성함수의 모수 $t_1, t_2, \ldots, t_J$가 선택되어야 한다. Epps와 Singleton (1986)은 9개의 분포(정규분포, 균일분포, 코시분포, 라플라스분포, 대칭안정분포, 감마분포, 포와송분포, 이항분포, 음이항분포)에 대해서 표본 크기 30 하에서 모의실험을 실시하여 일반적인 상황에서 적용할 수 있는 모수를 밝혔다. 그 결과, 표본 크기 30인 조건 하에서 $t_1 = 0.4$, $t_2 = 0.8$ ($J = 2$)가 최적으로 드러났다. 검정은 t_j는 척도(scale)의 추정치 $\hat{\sigma}$로 표준화한 $\tilde{t_j} = t_j/\hat{\sigma}$, $j = 1, 2$를 이용하는데 Epps와 Singleton (1986)은 준사분범위(semi-interquartile range)를 $\hat{\sigma}$에 대한 좋은 추정치라고 제시하고 있다. 각각의 X_{km}에 대해 4×1 벡터 $g(X_{km})$을 다음과 같이 생성한다.

$$g(X_{km}) = (\cos t_1 X_{km}, \sin t_1 X_{km}, \cos t_2 X_{km}, \sin t_2 X_{km})', \tag{2.6}$$

여기서 X_{km}, $m = 1, 2, \ldots, n_k$는 k번째 모집단에서 추출된 표본의 m번째 관측치이다. 표본에서 계산된 특성함수의 실수부와 가수부를 포함하도록 g_k를 다음과 같이 정의한다.

$$g_k = n_k^{-1} \sum_{m=1}^{n_k} g(X_{km}).$$

각 벡터 간의 차를 $G_2 = g_1 - g_2$라고 한다면 H_0가 참일 경우, $\sqrt{n_1 + n_2} G_2$는 점근적으로 다변량 정규분포 $N(\underline{0}, \Omega)$를 따른다. Epps와 Singleton은 공분산행렬 Ω에 대한 추정치를 아래 같이 제시하고 있다.

$$\hat{\Omega} = \frac{1}{\nu_1} \hat{S}_1 + \frac{1}{\nu_2} \hat{S}_2,$$

여기서 $\hat{S_k} = (n_k - 1)/n_k \text{cov}\{g(X_{km})\}$, $\nu_k = n_k/(n_1 + n_2)$이다. ES 통계량은

$$W_2 = (n_1 + n_2) \cdot G_2^{'} \cdot \hat{\Omega}^{+} \cdot G_2$$

이다. 여기서 $\hat{\Omega}^{+}$은 $\hat{\Omega}$의 일반화 역행렬이다. $\hat{\Omega}^{+}$의 계수(rank)를 r이라고 하면, W_2는 점근적으로 r의 자유도를 갖는 카이제곱 분포를 따른다. ES의 통계량 W_2은 두 표본에서 계산된 경험적특성함수의 거리를 분산-공분산 행렬로 표준화한 것으로 이해할 수 있다.

2.5. Mann-Whitney U검정

MW U검정은 두 표본이 같은 모분포를 따르는지 검정하기 위해 널리 이용되는 검정이다. MW 검정은 ES 검정과 마찬가지로 이산형 자료에 대해서도 적용할 수 있다 (Goerg와 Kaiser, 2009). 언급된 다른 검정법과는 달리 MW 검정의 대립가설은 두 표본의 모집단분포는 다른 위치 모수를 가진다는 것이다. $\triangle$를 위치 이동(location shift)이라고 하자. 그러면 다음과 같은 위치 이동 모형에 대해

$$G(t) = F(t - \triangle), \quad \text{for every } t.$$

대립 가설은 $H_1 : \triangle \neq 0$이다. MW 검정은 $\triangle > 0$ 또는 $\triangle < 0$의 단측 검정도 가능하다. MW 검정은 다음과 같은 통계량 U를 이용한다.

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(X_i, Y_j),$$

여기서

$$\phi(X_i, Y_j) = \begin{cases} 1, & \text{if } X_i < Y_j, \\ 0, & \text{otherwise.} \end{cases}$$

귀무가설 H_0는 큰 U 값에 의해 기각된다.

3. 선행 비교 연구

3.1. 선행 연구

Engmann과 Cousineau (2011)는 표본 크기가 $(16, 16), (32, 32), (64, 64)$인 조건 하에서 Weibull 분포를 따르는 누적분포함수 F와 G 사이에 위치(location) 모수, 척도(scale) 모수, 대칭도(symmetry) 모수를 달리하여 표본을 생성한 후 AD 검정과 KS 검정 간의 검정력 비교실험을 진행하였다. 그 결과 KS 검정은 과도하게 보수적인 제 1종 오류를 보였고, 위치, 척도, 대칭도의 모수 변화에 대해 AD 검정이 더 나은 검정력을 보였다. 또한, 분포의 극단 부분에서의 차이를 KS 검정보다 AD 검정이 더 잘 식별하며, 충분한 검정력을 갖기 위해 필요한 표본 크기 역시 KS 검정보다 AD 검정이 더 적었다.

Epps와 Singleton (1986)은 AD 검정과 CVM 검정, KS 검정을 자신들이 제안한 ES 검정과 비교하는 실험을 실시하여 다음과 같은 결과를 얻었다. (1) 자료가 이산형일 경우에는 ES를 적용한다. (2) KS 검정의 검정력이 가장 낮다. (3) 자료가 연속형이면서 표본 크기가 25 미만일 경우에는 AD 검정 또는 CVM 검정의 검정력이 뛰어나다. (4) 자료가 연속형이고, 모분포의 위치 모수가 차이날 경우 AD 또는 CVM의 검정력이 뛰어나다. (5) 자료가 연속형이면서 표본 크기가 25 이상일 경우에는 ES의 검정력이 좋다. (6) AD와 CVM의 검정력은 비슷하며, 간혹 CVM이 상당히 높은 모습을 보인다.

Table 3.1. Power of Kolmogorov-Smirnov test in different sample sizes

	m	Significance level (critical value)		
		0.1	0.05	0.01
$n = m$	20	0.081 (0.35)	0.030 (0.40)	0.004 (0.50)
	40	0.098 (0.25)	0.029 (0.30)	0.007 (0.35)
	60	0.077 (0.22)	0.047 (0.23)	0.009 (0.28)
	80	0.081 (0.19)	0.034 (0.21)	0.008 (0.25)
	100	0.078 (0.17)	0.036 (0.19)	0.010 (0.22)
$n = m - 1$	20	0.098 (0.38)	0.049 (0.4)	0.010 (0.49)
	40	0.100 (0.26)	0.049 (0.29)	0.010 (0.35)
	60	0.100 (0.21)	0.049 (0.24)	0.010 (0.29)
	80	0.100 (0.19)	0.050 (0.21)	0.010 (0.25)
	100	0.100 (0.17)	0.050 (0.19)	0.010 (0.23)

Goerg와 Kaiser (2009)는 실제 자료에 KS 검정과 MW 검정, ES 검정을 비교 적용하여 (1) KS 검정보다 ES 검정의 검정력이 뛰어나고, (2) 분포적인 특성의 차이를 감지하는 데에 MW 검정은 성능이 좋지 않은 반면, ES 검정은 뛰어난 성능을 보임을 밝혔다.

Özçomak 등 (2013)는 KS 검정과 MW 검정의 검정력을 비교하기 위해서 같은 왜도와 다른 첨도를 갖는 분포 사이의 검정과 다른 왜도와 같은 첨도를 갖는 분포 사이의 검정을 실시했다. 실험은 다양한 표본 크기에 대해 진행되었지만 일관된 결과를 얻을 수 없었다.

3.2. 선행 연구의 한계

이상의 검정력 비교 연구에서 공통적으로 발견된 중요한 한계점은 두 표본의 크기가 일치하는 조건 하에서 진행되었다는 것이다. 표본 크기가 같을 경우 KS 검정의 검정력은 심각하게 훼손된다. 이로 인해 KS 검정의 검정력이 낮게 보고되는 편향된 실험 결과를 얻었을 것이다. KS 통계량 D의 값은 불연속적으로 분포하기 때문에 D에 기반한 KS 검정의 유의 확률도 불연속적으로 분포한다. 예를 들어, $m = n = 30$인 경우를 생각해보자. D가 가질 수 있는 가능한 값은 $0, 1/30, 2/30, \ldots, 30/30 = 1$의 31가지이다. $D = 11/30$일 때 유의 확률은 0.03458이고 $D = 10/30$일 때 유의 확률은 0.07089이다. 만일 $\alpha = 0.05$라면 임계치는 $D = 11/30$이고 실제 α 값은 0.03458에서 관리되어 제 1종 오류가 보수적으로 나타날 것이다. 이처럼 D가 취할 수 있는 값이 제한적일수록 KS 검정이 보수성을 띄게 된다. KS 통계량 D가 취할 수 있는 값은 m, n의 공약수가 많을수록 제한적으로 주어지게 되는데, $m = n$의 경우는 D의 값이 가장 극단적으로 제한되는 경우이다. 선행 연구의 실험은 KS 검정이 가장 보수성을 띄는 조건 하에서 진행되었기에 편향된 결과를 주었을 것이다.

Table 3.1은 $n = m$의 경우와 공약수가 존재하지 않는 $n = m - 1$ 상황에서 KS 검정의 임계값과 제 1종 오류를 유의수준 0.1, 0.05, 0.01에 대해 산출한 모의실험 결과이다. 모의실험은 $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$에 대해 X, Y의 가능한 조합을 1,000,000번 임의추출하여 D의 분포를 산출하고 $P(D \geq d) = \alpha$(α는 주어진 유의수준보다 작은 값 중에 가장 큰 $P(D \geq d)$ 값)를 만족하는 d와 α를 산출하였다. 식 2.1의 귀무가설 하에서 1,000,000번의 임의 추출 시 기각 비율에 대한 99%, 95%, 90% 신뢰구간은 각각 [0.04998, 0.05002], [0.04995, 0.05005], [0.0494,0.0506]이다. $n = m$의 경우, 실제 유의수준은 표본 크기가 커짐에 따라 주어진 유의수준에 수렴하는 모습을 보이나 그 속도가 매우 느리고 불안정하여 두 집단의 표본 크기가 100이 되어도 실제 유의 수준은 신뢰 구간에서 한참 벗어나 있음을 확인할 수 있다. 이를 통해 KS 검정이 $n = m$인 경우에 제 1종 오류 관리가 적합하게 되어있지 않고 보수적

인 성질을 띌 수 밖에 없음을 알 수 있다. 따라서 본 연구에서는 보다 객관적인 실험 결과를 얻기 위해 $n = m$인 경우와 $n = m - 1$인 경우를 나누어 실험하려고 한다.

또한 지금까지의 검정력 비교 연구들은 두 가지 혹은 세 가지의 검정법을 각기 다른 제한된 환경에서 비교하였기 때문에 어느 검정법을 사용해야 할 지에 대한 통합된 가이드라인을 주고 있지 못한다. 본 연구에서는 기존의 연구에 사용된 실험 조건을 통합하고 한계점을 보완하여 객관적인 실험을 진행하여 특정 조건에 맞는 최상의 검정법을 제시하려고 한다.

4. 모의실험

4.1. 모의실험 구조

본 절에서는 기존의 검정력 비교 연구를 통합하여 일관된 조건 하에서 2절에서 언급한 검정법들의 성능을 비교하기 위한 모의실험을 실시하려고 한다. 모의실험에 사용된 검정법은 다음과 같다.

- Kolmogorov-Smirnov 검정
- Anderson-Darling 검정
- Cramér-von-Mises 검정
- Epps-Singleton 검정
- Mann-Whitney U 검정

본 연구에서는 기존 검정력 비교 연구에서 고려된 조건을 통합하여 다음과 같은 상황 하에서 실험 자료를 모의 생성하였다.

- 위치(location)가 다른 경우
- 척도(scale)가 다른 경우
- 왜도(skewness)가 다른 경우
- 첨도(kurtosis)가 다른 경우
- 꼬리가중치(tail weight)가 다른 경우

각 조건에 해당하는 분포는 Epps와 Singleton (1986)의 실험 조건을 차용하되 ES 검정의 성능을 좀 더 객관적인 조건에서 실험하기 위해 꼬리가중치가 다른 경우를 추가하였다. 각 조건별 분포는 Table 4.1과 Figure 4.1에 정리되어있다. 실험에 사용된 분포는 정규분포, 감마분포, 코시분포이며 Table 4.1에서 각각 $\mathcal{N}$, $\mathcal{G}$, $\mathcal{C}$로 표기되어 있다. 비록 코시분포의 첨도는 수학적으로 정의되지 않으나, 두 분포의 뾰족함의 정도의 차이는 Figure 4.1의 case 4를 통해 시각적으로 확인할 수 있다. 표본 크기는 Engmann과 Cousineau (2011)와 Epps와 Singleton (1986)의 조건을 차용하되 표본 크기가 동일할 때 나타나는 KS 검정의 보수성을 제거하기 위해 $(16, 15), (32, 31), (64, 63)$로 두었다. 표본 크기가 동일할 때의 효과를 비교하기 위해 $(64, 64)$의 표본 크기에 대해서도 실험하였다. 유의 수준 α는 0.05로 두어 각 조건별 1,000회를 반복하여 표본을 생성하고 기각된 횟수로 검정력을 측정하였다.

4.2. 모의실험 결과

4.2.1. 두 분포가 동일한 경우 Table 4.2는 귀무가설 하에서 생성된 표본에 대해 각 검정이 기각한 횟수의 비를 나타낸 것이다. 유의수준 0.05에서 기각 비율의 95% 신뢰구간은 [0.036, 0.064]으로, AD

Table 4.1. Underlying Distribution for each group

Case	Distributions	X	Y
	Equal	$\mathcal{N}(0,2)$	$\mathcal{N}(0,2)$
1	Different locations	$\mathcal{N}(0,2)$	$\mathcal{N}(1,2)$
2	Different scales	$\mathcal{N}(0,2)$	$\mathcal{N}(0,4)$
3	Different skewnesses	$\mathcal{N}(2,2)$	$\mathcal{G}(2,1)$
4	Different kurtosises	$\mathcal{N}(0,2)$	$\mathcal{C}(0,1)$
5	Different tail weights	$\mathcal{N}(0,2)$	$\mathcal{C}(0,2)$

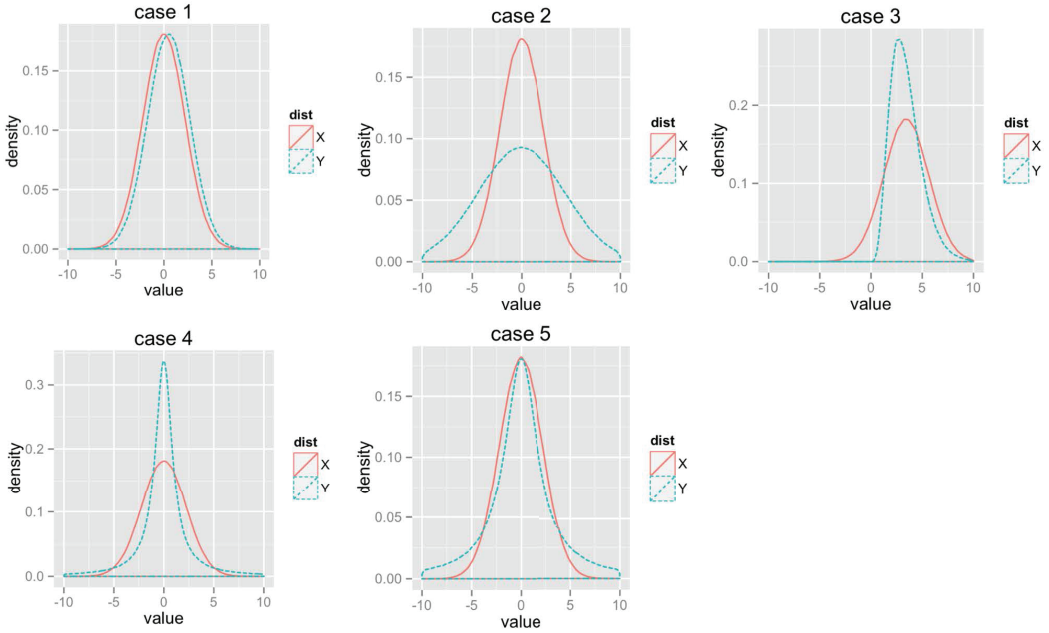

Figure 4.1. Underlying distribution for each group; case 1: different locations, case 2: different scales, case3: different skewnesses, case 4: different kurtosises, case 5: different tail weights.

Table 4.2. Power of test under the null assumption

Test	(m,n)			
	$(16,15)$	$(32,31)$	$(64,63)$	$(64,64)$
KS	0.047	0.060	0.054	0.031
AD	0.054	0.059	0.053	0.036
CVM	0.055	0.056	0.054	0.037
ES	0.055	0.064	0.063	0.054
MW	0.054	0.056	0.059	0.036

KS = Kolmogorov-Smirnov; AD = Anderson-Darling; CVM = Cramér-von-Mises; ES = Epps-Singleton; MW = Mann-Whitney.

검정과 CVM 검정, MW 검정은 이 구간에서 크게 벗어나지 않는다. KS 검정은 $(m,n) = (64,64)$일 경우에 기각 비율이 0.031으로써 95% 신뢰구간에서 벗어나 제 1종 오류가 보수적으로 나타나는 모습을 관찰할 수 있다. 하지만 $(m,n) = (64,63)$일 경우에는 KS 통계량의 기각 비율이 0.054로서, 제 1종 오

Table 4.3. Power of test when two distributions differ in location

Test	(m, n)			
	$(16, 15)$	$(32, 31)$	$(64, 63)$	$(64, 64)$
KS	0.200	0.402	0.680	0.603
AD	0.249	0.472	0.771	0.766
CVM	0.247	0.485	0.764	0.763
ES	0.071	0.165	0.230	0.274
MW	0.253	0.486	0.789	0.776

KS = Kolmogorov-Smirnov; AD = Anderson-Darling; CVM = Cramér-von-Mises; ES = Epps-Singleton; MW = Mann-Whitney.

Table 4.4. Power of test when two distributions differ in scale

Test	(m, n)			
	$(16, 15)$	$(32, 31)$	$(64, 63)$	$(64, 64)$
KS	0.100	0.197	0.577	0.518
AD	0.147	0.427	0.890	0.897
CVM	0.193	0.487	0.900	0.909
ES	0.150	0.226	0.128	0.135
MW	0.056	0.059	0.040	0.055

KS = Kolmogorov-Smirnov; AD = Anderson-Darling; CVM = Cramér-von-Mises; ES = Epps-Singleton; MW = Mann-Whitney.

류의 보수성이 나타나지 않는다. 이는 전 장에서 언급한 KS 통계량의 불연속적인 특성으로 인한 현상이다. 본 실험에서는 동일하지 않은 표본 크기를 조건으로 설정하여 KS 검정의 검정력이 왜곡되는 현상을 방지할 수 있었다.

4.2.2. 두 분포의 위치가 다를 경우 Table 4.3은 위치 모수만 다른 두 분포에서 생성된 표본 사이의 검정을 실시하여 기각한 횟수의 비율을 나타낸 것이다. 모든 표본 크기에 대해 MW 검정이 가장 강력한 검정으로 관찰되었다. AD 검정과 CVM 검정 역시 MW 검정과 비슷한 수준의 검정력을 보였고, KS 검정의 검정력은 약간 뒤쳐지는 모습을 보인다. ES 검정의 기각률은 크게 떨어지는 모습을 보이며 위치 모수의 이동에 대해서는 잘 식별하지 못하고 있음을 알 수 있다.

4.2.3. 두 분포의 척도가 다를 경우 Table 4.4는 척도 모수만 다른 두 분포에서 생성된 표본 사이의 검정을 실시하여 기각한 횟수의 비율을 나타낸 것이다. 모든 표본 수에 대해 CVM 검정이 가장 강력한 검정으로 드러났으며, AD 검정 또한 CVM 검정에 비해 크게 떨어지지 않는 검정력을 보였다. 모수 $t_1 = 0.4$, $t_2 = 0.8$ ($J = 2$)를 사용한 ES 검정의 기각력은 $m = 32$까지는 KS 검정과 비슷한 수준이지만 $m = 64$의 경우에 표본 크기가 늘어났음에도 기각력이 급격히 감소하는 것을 보인다. Epps와 Singleton (1986)이 제안한 모수를 활용할 때 표본 크기 30 근처에서만 이상적인 성능을 내는 것으로 보인다. MW 검정은 두 집단 사이의 위치 모수의 차를 검정하기 때문에 검정력이 유의수준에 머물고 있다. 즉, MW 검정은 척도 모수의 이동을 식별하지 못한다.

4.2.4. 두 분포의 왜도가 다를 경우 Table 4.5는 왜도가 다른 두 분포에서 생성된 표본 사이의 검정을 실시하여 기각한 횟수의 비율을 나타낸 것이다. 모든 검정 중 ES 검정의 기각 비율이 단연 높은 모습을 보였다. 왜도의 차이는 ES 검정이 가장 민감하게 식별하는 것으로 나타났다. AD 검정과 CVM

Table 4.5. Power of test when two distributions differ in skewness

Test	(m, n)			
	$(16, 15)$	$(32, 31)$	$(64, 63)$	$(64, 64)$
KS	0.075	0.161	0.426	0.300
AD	0.096	0.198	0.581	0.599
CVM	0.100	0.210	0.555	0.545
ES	0.165	0.570	0.888	0.902
MW	0.052	0.044	0.058	0.053

KS = Kolmogorov-Smirnov; AD = Anderson-Darling; CVM = Cramér-von-Mises; ES = Epps-Singleton; MW = Mann-Whitney.

Table 4.6. Power of test when two distributions differ in kurtosis

Test	(m, n)			
	$(16, 15)$	$(32, 31)$	$(64, 63)$	$(64, 64)$
KS	0.146	0.329	0.702	0.577
AD	0.062	0.291	0.689	0.692
CVM	0.122	0.283	0.598	0.599
ES	0.245	0.394	0.416	0.398
MW	0.054	0.060	0.058	0.062

KS = Kolmogorov-Smirnov; AD = Anderson-Darling; CVM = Cramér-von-Mises; ES = Epps-Singleton; MW = Mann-Whitney.

검정의 검정력은 비슷하였으나 낮은 표본 크기에서는 CVM의 검정력이 약간 우세하였고, 표본 크기 32를 기점으로 높은 표본 크기에서는 AD가 약간 우세한 검정력을 보였다. MW 검정은 왜도 차이를 식별할 수 없는 것으로 보인다.

4.2.5. 두 분포의 첨도가 다를 경우 Table 4.6은 첨도에 차이가 있는 두 분포에서 생성한 표본 사이의 검정 결과를 나타낸 것이다. 표본 크기가 $m = 32$ 이하일 경우 ES 검정이 가장 우세한 검정력을 보였지만 표본 크기 $m = 64$에서는 검정력 증가가 급격히 둔화되는 것으로 관찰되었다. 표본 크기의 증가에 따라 검정력이 꾸준한 증가 경향을 보이는 KS, AD, CVM 검정과는 대조적이다. $m = 30$에서 멀리 벗어난 표본 크기에 대해서 Epps와 Singleton (1986)이 제시한 모수의 사용은 불안정한 결과를 줄 것으로 예상된다. AD 검정과 CVM 검정의 검정력은 Table 4.5와 마찬가지로 $m = 32$를 기점으로 낮은 표본 크기에서는 CVM 검정이 우세하였고 높은 표본 크기에서는 AD 검정이 우세하였다. 표본 크기가 64일 때 KS 검정의 검정력은 표본 크기가 동일한지 여부에 좌우되어, 표본 크기가 동일할 경우에는 AD 검정 및 CVM 검정의 검정력과 비슷하지만, $n = m - 1$의 조건 하에서는 가장 강력한 검정력을 보였다. MW 검정은 첨도 차이를 식별할 수 없는 것으로 보인다.

4.2.6. 두 분포의 꼬리 부분이 다를 경우 본 연구에서는 Epps와 Singleton (1986)이 실험한 조건 이외에서도 ES의 성능을 살펴보기 위해 추가적으로 G의 분포를 Cauchy$(0, 1)$에서 Cauchy$(0, 2)$로 변경하여 실험을 진행하였다. 이 경우, G의 분포는 F의 분포와 비슷하지만 꼬리 부분만 두터운 모습을 보이게 된다 (Figure 4.1). 실험 결과는 Table 4.7에 제시되어 있다. CVM 검정의 검정력이 가장 우세하였고 낮은 표본 크기에서는 AD의 검정이 CVM 검정보다 약간 높은 검정력을 보였다. 두 분포의 왜도가 다른 경우(Table 4.5)와는 대조적으로 ES 검정의 검정력이 크게 떨어졌음을 알 수 있다. 또한 ES 검정의 검정력은 $n = 32$일 경우에 가장 높았으며 그 외의 표본 크기에 대해서는 더 낮은 검정력을 보였다.

Table 4.7. Power of test when two distributions differ in tail weight

Test	(m, n)			
	$(16, 15)$	$(32, 31)$	$(64, 63)$	$(64, 64)$
KS	0.059	0.078	0.194	0.162
AD	0.084	0.199	0.523	0.535
CVM	0.062	0.237	0.627	0.621
ES	0.059	0.119	0.082	0.065
MW	0.053	0.059	0.050	0.057

KS = Kolmogorov-Smirnov; AD = Anderson-Darling; CVM = Cramér-von-Mises; ES = Epps-Singleton; MW = Mann-Whitney.

Table 5.1. Quartile-based descriptive statistics

	Statistics
Location	$F^{-1}(0.5)$
Scale	$F^{-1}(0.75) - F^{-1}(0.25)$
Skewness	$\frac{F^{-1}(0.75) + F^{-1}(0.25) - 2 * F^{-1}(0.5)}{(F^{-1}(0.75) + F^{-1}(0.25)}$
Kurtosis	$\frac{F^{-1}(0.975) + F^{-1}(0.025)}{F^{-1}(0.75) - F^{-1}(0.25)} - 2.91$

KS 검정의 검정력은 표본 크기가 증가함에 따라 매우 천천히 증가하는 모습을 보인다. MW 검정은 분포의 꼬리 부분의 차이를 식별하지 못하였다.

4.3. 모의실험 결과 정리

모의실험 결과, 관측된 사실을 정리하면 다음과 같다.

- 위치의 식별에 대해서는 MW 검정의 검정력이 가장 높게 나타났고 AD 검정과 CVM 검정의 검정력 역시 높게 나타났다.
- 척도의 식별에 대해서 CVM 검정의 검정력이 가장 높게 나타났고 AD 검정 역시 높게 나타났다.
- 왜도의 식별에 대해서 ES 검정의 검정력이 가장 높게 나타났다.
- 첨도의 식별에 대해서 표본 크기가 같을 때에는 AD 검정, 표본 크기가 다를 때에는 KS 검정의 검정력이 가장 높게 나타났다.
- 분포의 꼬리부분에서만 나타나는 차이에 대해서 CVM 검정의 검정력이 가장 높게 나타났다.
- Epps와 Singleton (1986)이 제시한 모수에 기반해 ES 검정을 실시할 경우, 표본 크기가 30에서 벗어날 수록 검정력이 떨어지는 현상을 보인다.
- 표본 크기가 다른 조건에서 실험한 결과, 이표본 분포 동일성 검정의 비교 연구 중 처음으로 KS 검정이 다른 검정보다 우세한 검정력을 보이는 경우가 보고되었다.

5. 실제 자료 예시

본 절에서는 앞에서 소개된 검정법을 실제 자료에 적용하여 그 특성을 살펴보고자 한다. 누적분포함수 F를 따르는 자료 $X_1, X_2, \ldots, X_N$의 위치, 척도, 왜도, 첨도를 나타내기 위해서 분위수에 기반한 추정량을 Table 5.1과 같이 사용하고자 한다. 위치의 추정량은 표본 중위수, 척도의 추정량은 사분위범

Table 5.2. Salivation data

i	X_i	Y_i
1	−0.15	2.55
2	8.60	12.07
3	5.00	0.46
4	3.71	0.35
5	4.29	2.69
6	7.74	−0.94
7	2.48	1.73
8	3.25	0.73
9	−1.15	−0.35
10	8.38	−0.37

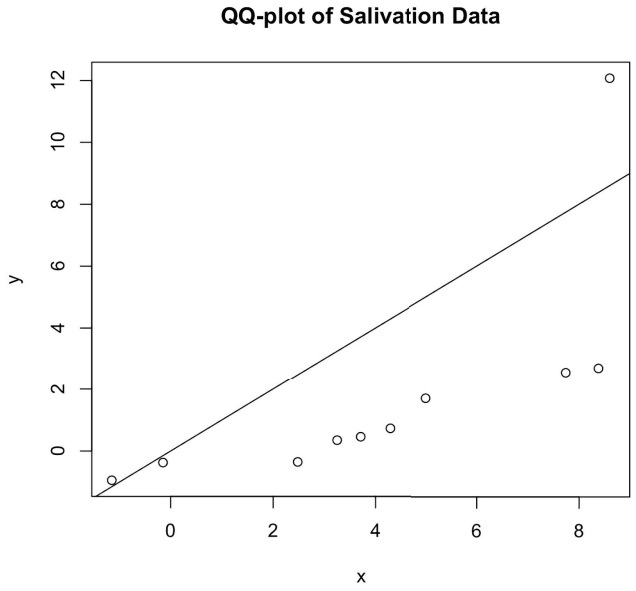

Figure 5.1. Q-Q plot of salivation data.

위(interquartile range), 왜도의 추정량은 Bowley (1920)가 제안한 분위수에 기반한 왜도 계수이고 첨도의 추정량은 Crow와 Siddiqui (1967)가 제안한 척도 계수이다.

5.1. 타액분비자료

Table 5.2의 자료는 Delse와 Feather (1968)가 실시한 타액 분비량의 자기제어에 관한 실험 결과이다. 이 실험에서 20명의 피실험자들은 자신의 타액 분비량을 조절하도록 요구 받았다. 실험군에 해당하는 10명의 피실험자들은 자신의 타액 분비량에 대한 정보를 실시간으로 제공받았고, 대조군에 해당하는 10명의 피실험자들은 아무런 정보를 제공받지 못하였다. Table 5.2에서 X는 실험군, Y는 대조군을 나타낸다. Hollander 등 (2014)는 이표본 KS 검정의 적용의 예시를 위해 이 자료를 이용하였다.

Table 5.2의 자료를 이용하여 모의실험과 동일한 결과를 보이는지 실험하여 보도록 하겠다. Figure 5.1은 X, Y의 Q-Q 도표로써, Y의 자료가 오른쪽으로 치우쳐져 분포함을 알 수 있다. Table 5.3은 위

Table 5.3. Quartile-based descriptive statistics on salivation data

Variable	Location	Scale	Skewness	Kurtosis
X	4.000	4.382	−0.040	1.740
Y	0.595	2.520	0.739	3.630

Table 5.4. Reported p-values of two-sample goodness-of-fit tests on salivation data

	KS	AD	CVM	ES	MW
Original data	0.05244	0.0607	0.0669	0.0044	0.0892
A random Y_i removed	0.04460	0.0913	0.0749	0.0068	0.0892

KS = Kolmogorov-Smirnov; AD = Anderson-Darling; CVM = Cramér-von-Mises; ES = Epps-Singleton; MW = Mann-Whitney.

Table 5.5. Bleeding time data (seconds)

i	X_i	Y_i
1	270	525
2	150	575
3	270	190
4	420	395
5	202	370
6	255	210
7	165	490
8	220	250
9	305	360
10	210	285
11	240	630
12	300	385
13	300	195
14	70	295

치, 척도, 왜도, 첨도의 추정량을 이용한 수치 요약이다. 이를 통해 해당 자료의 X, Y 두 집단이 심한 왜도의 차이를 보인다는 것을 관찰할 수 있고, 따라서 왜도의 차이를 둔 모의실험 4.2.4절과 비슷한 결과를 보일 것으로 예상된다. Table 5.4는 자료 1에 대한 분포의 동일성 검정 결과이다. ES 검정의 유의확률만 유독 낮은 것을 확인할 수 있다. 유의수준 0.05에서 ES 검정만 귀무가설을 기각할 수 있었는데, 이는 앞에서 살펴본 바와 같이 ES 검정이 왜도에 대해 민감한 특성을 지니기 때문으로 보인다.

추가적으로 표본 크기가 동일할 때 KS 검정이 갖는 보수성에 대해서도 실험해보도록 하겠다. 제시된 자료는 X, Y 두 집단이 10개의 동일한 표본 크기를 가지고 있기 때문에 KS 검정이 보수적인 결과치를 주었을 것이다. 이를 확인하기 위해 Table 5.2의 Y에서 임의의 관측치(여기서 Y의 가장 마지막 관측치 −0.37)을 제거한 후의 검정한 결과가 Table 5.4에 제시되어 있다. 다른 검정의 유의확률은 증가하였지만 KS의 유의확률은 낮아졌다는 것을 확인할 수 있다. 이는 KS 검정이 같은 표본 수에 대한 검정에 대해 불리하다는 것을 보여준다.

5.2. 출혈시간자료

Table 5.5는 Bick 등 (1976)의 연구의 출혈 시간 자료이다. 이 실험은 아스피린 복용이 출혈 시간에 미치는 영향을 측정하기 위해서 평범한 실험 자원자들을 대상으로 진행되었다.

Table 5.6. Quartile-based descriptive statistics on bleeding time data

Variable	Location	Scale	Skewness	Kurtosis
X	257.5	88.5	−0.0571	5.4081
Y	370.0	240.0	0.1510	3.3479

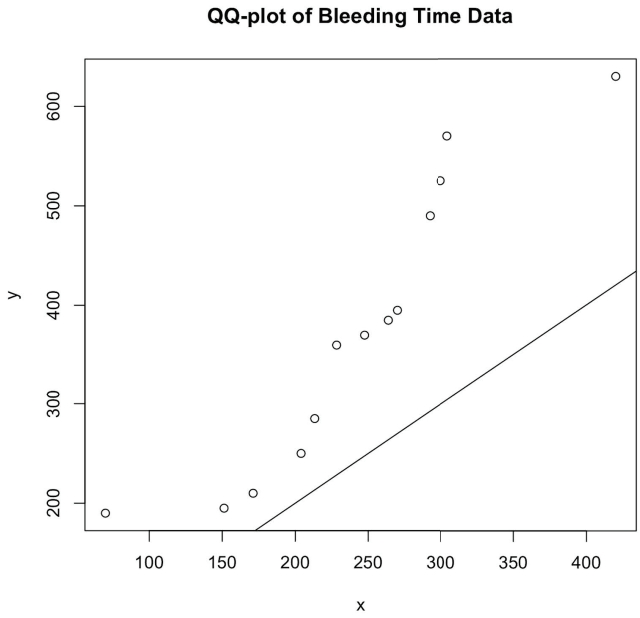

Figure 5.2. Q-Q plot of bleeding time data.

Table 5.7. Reported p-values of two-sample goodness-of-fit tests on bleeding time data

KS	AD	CVM	ES	MW
0.0370358	0.024767	0.007992008	0.1654718	0.03266888

KS = Kolmogorov-Smirnov; AD = Anderson-Darling; CVM = Cramér-von-Mises; ES = Epps-Singleton; MW = Mann-Whitney.

Table 5.5의 자료를 이용하여 모의실험과 동일한 결과를 보이는지 실험하여 보도록 하겠다. 자료의 수치 요약은 Table 5.6, Q-Q 도표는 Figure 5.2에 제시되어 있다. 자료의 X, Y 두 집단은 명백히 다른 분포를 따르는 것으로 보이며 위치와 척도에 큰 차이가 존재하는 것으로 보인다. 따라서 위치와 척도에 차이를 둔 모의실험 4.2.2절과 모의실험 4.2.3절과 비슷한 결과를 보일 것으로 예상된다. 출혈 시간 자료에 대한 검정법별 유의확률은 Table 5.7에 제시되어있다. 유의수준 0.05에서 ES 검정만 귀무가설을 기각하지 못하였는데, 이는 ES 검정이 위치와 척도 차이를 잘 식별하지 못하기 때문으로 보인다. KS, AD, CVM, MW 검정 모두 유의수준 0.05에서 귀무가설을 기각하였고 CVM이 특별히 낮은 유의확률을 보였다. 이는 위치와 척도가 다른 분포에 대해서 CVM이 가장 강력한 검정력을 보인 모의실험 결과와 일치한다.

6. 결론

본 연구에서는 이표본 분포 동일성 검정에 사용되는 분포무관 검정법을 살펴보고 각 검정법의 특성을 모

의실험과 실제 자료 적용을 통해 살펴보았다. 고려된 검정법은 KS 검정 그리고 CVM 검정, AD 검정과 ES 검정이다. 연구 결과 상황 별로 다음과 같은 검정법을 적용할 것을 추천한다.

- 자료가 이산형일 경우, 표본 간 위치의 차이를 검정하고자 한다면 MW 검정, 그렇지 않으면 ES 검정을 적용한다.
- 자료가 연속형일 경우, 자료의 위치나 척도의 차이를 검정하고자 한다면 CVM 검정 혹은 AD 검정을 적용한다.
- 자료가 연속형일 경우, 자료의 치우침이나 퍼짐 정도의 차이를 검정하고자 한다면 AD 검정을 사용한다.
- 자료가 연속형일 경우, 자료의 위치의 차이만을 검정하고자 한다면 MW 검정을 적용한다.
- 자료가 연속형일 경우, 자료의 치우침 정도의 차이만을 검정하고자 한다면 ES 검정을 사용한다.
- 자료가 연속형일 경우, 자료의 꼬리 부분의 차이에 대해서만 검정하고자 한다면 CVM 검정을 사용한다.

전반적으로 CVM 검정이 가장 강력하고 안정적인 검정력을 보였다. 왜도의 차이에 대해서는 ES 검정이 좋은 검정력을 보였으나 Epps와 Singleton (1986)이 제안한 $t_1 = 0.4$, $t_2 = 0.8$ ($J = 2$)를 모수로 이용했을 때에는 표본 크기 30 근처에서만 이상적으로 작동하는 것으로 나타났다. ES 검정을 현실 문제에 적용하기 위해서 다양한 표본 크기에 따른 적정 모수에 대한 연구가 필요할 것으로 보인다. 본 연구의 결과를 바탕으로 주어진 상황에 따라 적절한 검정법을 적용하는 데에 많은 도움이 될 것이라고 기대하며 향후 이표본 분포 동일성 검정에 있어 좋은 참고 자료가 될 것이라고 기대한다.

References

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes, *The Annals of Mathematical Statistics*, **23**, 193–212.

Anderson, T. W. (1962). On the distribution of the two-sample Cramer-von Mises criterion, *The Annals of Mathematical Statistics*, **33**, 1148–1159.

Bick, R. L., Adams, T., and Schmalhorst, W. R. (1976). Bleeding times, platelet adhesion, and aspirin, *American Journal of Clinical Pathology*, **65**, 69–72.

Bowley, A. L. (1920). *Elements of Statistics* (4th ed), P.S. King & Son, London.

Burr, E. J. (1964). Small-sample distributions of the two-sample Cramer-von Mises' W^2 and Watson's U^2, *The Annals of Mathematical Statistics*, **35**, 1091–1098.

Cramér, H. (1928). On the composition of elementary errors: first paper: mathematical deductions, *Scandinavian Actuarial Journal*, **1928**, 13–74.

Crow, E. L. and Siddiqui, M. M. (1967). Robust estimation of location, *Journal of the American Statistical Association*, **62**, 353–389.

Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramér-von Mises tests, *The Annals of Mathematical Statistics*, **28**, 823–838.

Delse, F. C. and Feather, B. W. (1968). The effect of augmented sensory feedback on the control of salivation. *Psychophysiology*, **5**, 15–21.

Engmann, S. and Cousineau, D. (2011). Comparing distributions: the two-sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnoff test, *Journal of Applied Quantitative Methods*, **6**, 1–17.

Epps, T. W. and Singleton, K. J. (1986). An omnibus test for the two-sample problem using the empirical characteristic function, *Journal of Statistical Computation and Simulation*, **26**, 177–203.

Goerg, S. J. and Kaiser, J. (2009). Nonparametric testing of distributions - The Epps-Singleton two-sample test using the empirical characteristic function. *The Stata Journal*, **9**, 454–465.

Hollander, M., Wolfe, D. A., and Chicken, E. (2014). *Nonparametric Statistical Methods* (3rd ed), John Wiley & Sons, Hoboken.

Lilliefors, H. W. (1969). On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown, *Journal of the American Statistical Association*, **64**, 387–389.

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.

Massey, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit, *Journal of the American Statistical Association*, **46**, 68–78.

Özçomak, M. S., Kartal, M., Senger, Ö., and Çelik, A. K. (2013). Comparison of the powers of the Kolmogorov-Smirnov two-sample test and the Mann-Whitney test for different kurtosis and skewness coefficients using the Monte Carlo simulation method, *Journal of Statistical and Econometric Methods*, **2**, 81–98.

Pettitt, A. N. (1976). A two-sample Anderson-Darling rank statistic, *Biometrika*, **63**, 161–168.

Scholz, F. W. and Stephens, M. A. (1987). K-sample Anderson-Darling tests, *Journal of the American Statistical Association*, **82**, 918–924.

Smirnoff, N. (1939). Sur les écarts de la courbe de la distribution empirique, *Receuil Mathémathique (Matematiceskii Sbornik)*, **6**, 3–26.

von Mises, R. (1928). *Wahrscheinlichkeit Statistik und Wahrheit*, Springer-Verlag, Berlin.

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

이표본 분포 동일성에 대한 분포무관 검정법 간 검정력 비교 연구

김선빈[a] · 이재원[a,1]

[a]고려대학교 통계학과

(2017년 2월 8일 접수, 2017년 3월 20일 수정, 2017년 6월 14일 채택)

요 약

두 표본 집단이 동일한 분포를 따르는지 비교하기 위해 분포무관 검정이 많이 사용된다. 하지만 여러 검정법을 체계적으로 비교한 연구가 존재하지 않아서 각 검정법의 특성을 고려하여 연구 상황에 맞는 검정법을 선택하기가 어려웠다. 본 연구에서는 이표본 분포 동일성 검정에 해당하는 여러 분포무관 검정법들을 소개하고 체계적인 모의실험을 통해 그 성능을 비교하고자 한다. 두 표본이 각각 (1) 위치, (2) 척도, (3) 왜도, (4) 첨도, (5) 꼬리가중치가 다른 분포에서 추출된 상황에 대해 실험하였다. 실험 결과를 바탕으로 이표본 분포 동일성 검정법 사용에 대한 실용적인 지침을 제시하려고 한다.

주요용어: 콜모고로프-스미르노프 검정, 크레이머-본미세스 검정, 앤더슨-달링 검정, 앱스-싱글톤 검정, 맨-위트니 U 검정, 적합도 검정

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2016 943438).

[1]교신저자: (02841) 서울특별시 성북구 안암로 145, 고려대학교 통계학과. E-mail: jael@korea.ac.kr