조 영 일 교수 지도

석사학위 청구논문

소표본(small sample)에 활용 가능한 통계 검정법의 비교

: 비모수 검정법과 재표집 방법을 중심으로

2020

성신여자대학교 대학원

심리학과

조 민 지

소표본(small sample)에
활용 가능한 통계 검정법의 비교
: 비모수 검정법과 재표집 방법을 중심으로

조 영 일 교수 지도

이 논문을 석사학위논문으로 제출함

2019년 11월

성신여자대학교 대학원

심리학과

조 민 지

인 준 서

조민지의 석사학위 논문으로 인준함

2019년 11월

심사위원장 ..... (인)

심 사 위 원 ..... (인)

심 사 위 원 ..... (인)

성신여자대학교 대학원

논문개요

본 논문은 표본 크기를 포함한 여러 연구 조건에서 모수 검정법과 비모수적 검정법의 수행을 비교하고, 비모수적 검정법을 중심으로 소표본에 적합한 분석 방법을 탐색하는 시뮬레이션 연구이다. 모수 검정법에는 일원분산분석, 비모수적 검정법에는 비모수 검정법인 크루스칼-왈리스 검정, 재표집 방법인 잭나이프, 임의순열, 붓스트랩 검정이 사용되었다. 이를 위하여 세 개의 조작변수(표본 크기, 효과 크기, 평균 변동성) 수준에 따라 3×5×2의 요인설계(factorial design)를 바탕으로 총 30개의 실험 조건을 설정하였다. 동일한 조건을 10,000번씩 반복하고, 각 조건 아래 통계 검정법의 검정력을 산출하였다. 1종 오류 비율의 경우, 집단 간 차이가 없다는 가정 하에 (효과크기=0, 평균 변동성=∅) 3가지 표본 크기 조건을 10,000번씩 반복하여 표본 크기 수준에 따른 1종 오류 비율을 구하였다. 자료의 생성과 분석에는 통계 프로그램 R이 사용되었다.

분석 결과, 검정법의 종류와 표본 크기에 따라 1종 오류 비율에 차이가 있음이 관찰되었다. 평균적으로 1종 오류 비율은 붓스트랩 검정에서 가장 컸고, 일원분산분석, 임의순열 검정, KW 검정, 잭나이프 검정 순으로 작았다. 표본 크기가 커질수록 1종 오류 비율의 평균이 감소하는 경향이 나타났고, 1종 오류 비율이 점차 유의수준 .05에 가까워지는 것으로 나타났다.

또한, 통계 검정법과 효과 크기에 따라서 검정력에 차이가 있었다. 효과 크기가 커질수록 검정력이 증가하는 경향이 나타났으며, 모든 조건에서 붓스트랩 검정의 검정력이 가장 큰 것으로 관찰되었다. 표본 크기가 3인 경우를 제외하고 평균적으로 가장 작은 검정력을 갖는 검정법은 잭나이프 검정이었다. 평균 변동성에 따라 비교했을 때는 최대 변동성 조건에서 항상 높

은 검정력 수치가 나타났으나, 전반적인 검정력 변화 패턴은 유사했다. 표본 크기에 따라서는 충분한 크기에 가까워질수록 통계 검정법 간 검정력의 차이가 감소하여 비슷한 수행 능력을 가지게 되는 경향을 보였다.

따라서 연구자에게 주어진 데이터의 형태와 조건에 따라 통계 검정법이 가지는 장단점이 다르기 때문에, 상이한 연구문제에 따라 적합한 분석방법을 사용할 필요성에 대해 논의하였다. 본 논문은 두 집단 차이검정에 사용되는 분석법을 비교한 기존의 연구를 3개 이상의 집단에 적용할 수 있도록 일반화 가능성을 높였다는 점과, 비모수적 검정법의 범주에 재표집 방법을 포함하여 보다 확장된 비교 연구를 수행했다는 점에서 의의가 있다. 본 연구의 결과는 추후 연구자의 자료 특성에 맞는 검정법을 선택하는 근거로 활용할 수 있을 것으로 기대된다. 마지막으로 연구의 한계를 보완하기 위한 후속연구 방안을 제언하였다.

주요어 : 비모수 검정법, 재표집 방법, 크루스칼-왈리스 검정, 잭나이프 검정, 임의순열 검정, 붓스트랩 검정, 시뮬레이션, 1종 오류 비율, 검정력

목 차

표 목 차

<표 9> $n_1 = n_2 = n_3 = 5$인 중간 변동성 조건에서 통계 검정법과 효과 크기에 따른 검정력 ····· 71

<표 10> $n_1 = n_2 = n_3 = 10$인 최대 변동성 조건에서 통계 검정법과 효과 크기에 따른 검정력 ····· 75

<표 11> $n_1 = n_2 = n_3 = 10$인 중간 변동성 조건에서 통계 검정법과 효과 크기에 따른 검정력 ····· 76

그 림 목 차

Ⅰ. 서론

1. 연구의 필요성 및 목적

심리학은 인간의 마음과 행동을 탐구하는 학문으로, 다수에게서 나타나는 보편적 개념부터 소수로부터 관찰되는 특이한 개념까지 다양한 심리적 개념을 이해하기 위한 연구가 이루어지고 있다. 원칙적으로 특정 집단에서 나타나는 심리적 구성개념을 명확하게 설명하려면 모집단(population) 전체에 대한 정보가 필요하다. 만약 모집단 대신 이를 잘 대표하는 표본(sample)이 있다면 시간적, 경제적인 제한에서 벗어나 표본의 속성만으로 모집단의 속성을 추론할 수 있다(성태제, 2014; 이재원, 이욱기, 2019). 하지만 모집단을 완벽하게 대표하는 표본을 추출하는 것은 사실상 불가능한 일이다. 표집(sampling)에 따라 표본과 모집단 사이에는 항상 차이가 발생하게 되는데, 표집분포(sampling distribution)의 특성을 활용하면 표본의 불확실성(uncertainty)으로 인한 오차를 고려한 모수 추정이 가능하다.

심리학을 포함한 사회과학 분야에서 대부분의 자료가 표본으로부터 수집되는 만큼, 정확한 연구결과를 얻기 위해서는 통계적 검정 절차에서 표본의 역할에 대한 이해가 필요하다. 특히 사례 수와도 같은 표본 크기(sample size)는 단순히 클수록 모집단을 잘 대표하는 것을 넘어, 결과적으로 추정의 정확성에 영향을 미치기 때문에 중요하게 고려되어야 할 요소이다(김수영, 석혜은, 2015). 표집분포의 특성을 정리한 중심극한정리(central limit theorem)를 통해 표본 크기의 중요성을 더 자세히 살펴볼 수 있다.

표집분포는 평균[1]과 표준오차(standard error)[2]로 표현되는데, 중심극

1) k번의 무수히 많은 반복추출이 이루어졌을 때, 표집분포의 평균($\mu_{\overline{X}}$)은 k개 표본분포 평균들($\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_k$)의 평균과 같다.

한정리를 활용하면 무한한 표집 과정 없이도 표집분포의 특성을 나타내는 모수들을 계산할 수 있다. 중심극한정리에 따르면 표집분포의 평균은 모집단의 평균과 같고(식 (1.1)), 표준오차는 모집단의 표준편차를 표본 크기의 제곱근으로 나눈 것과 같다(식 (1.2)).

$$\mu_{\overline{X}} = \mu \qquad \text{식 (1.1)}$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \qquad \text{식 (1.2)}$$

식 (1.2)에 따르면 표본 크기가 커질수록 표준오차는 감소하게 된다. 즉, 표본의 크기가 커질수록 표본이 전집에 대하여 가지는 대표성이 높아지고, 이는 곧 표본의 특성을 활용하여 모수를 추정함에 오차가 작다는 것을 의미한다. 반대로 표본의 크기가 작아지면 표본은 모집단을 잘 대표할 수 없고, 큰 표집오차 때문에 모수 추정의 정확성이 떨어지게 된다.

표본의 크기가 중요한 또 다른 이유는 분포의 정규성에 있다. 중심극한정리에 의해 설명되는 표집분포의 또 다른 특성은 평균이 μ, 표준편차가 $\frac{\sigma}{\sqrt{n}}$인 표집분포는 표본 크기가 충분히 클 때 모집단의 분포와 상관없이 정규분포(normal distribution)[3]를 따른다는 것이다. 모집단 분포가 대부분 정규분포를 가정하는 것과 다르게, 표본분포는 표집에 따라 편포되어 있기

2) 표준오차는 표집분포의 표준편차이자 표집오차(sampling error)의 표준편차이다. 표집으로 인해 발생하는 표본평균과 모평균의 차이를 표집오차라 하며, 표준오차는 무수히 많은 표집오차를 요약한 값이라 할 수 있다. 표준오차는 표본 통계량이 모수와 얼마나 가까운지 추정의 정확성을 판단하는 기준이 된다.

3) 현실 세계에서 발생하는 현상들은 대개 정규분포를 이루고 있다. 또한, 정규분포곡선의 넓이는 0에서 1사이의 범위를 갖기 때문에, 올바른 의사 결정이 발생하는 사건을 확률의 개념을 통해 논할 수 있다. 따라서, 사회과학 연구에서 다뤄지는 변인들은 대부분 정규분포를 따르는 것으로 가정하는 경우가 많다(성태제, 2014).

도 하다. 표본의 크기가 작을수록 표본분포가 편포될 가능성은 높다. 반면, 표집분포에서는 표본의 크기만 충분하다면 정규분포를 따르는 것으로 볼 수 있어 일반적인 사회 현상을 설명하기에 용이하다. 이때, 충분히 큰 표본이란 주로 표본의 크기가 30 이상인 경우($n \geq 30$)를 기준으로 한다(박승호, 2000; Cohen, 1988; Kreft & De Leeuw, 1998; Maas & Hox, 2005; Sawilowsky & Blair, 1992). 정리하자면 대표성이 높은 표본을 표집하여 모수를 정확히 추정하고, 모수 추론 과정에 정규분포의 이점을 활용하기 위해서는 충분히 큰 표본 데이터를 수집해야 할 필요가 있다.

하지만 적절한 크기의 표본을 얻고자 하는 연구자들의 바람과는 다르게, 연구대상이 지닌 특수성으로 인하여 적은 사례 수만으로 연구를 진행할 수밖에 없는 경우가 종종 발생한다. 가령 조현병을 앓고 있는 임상군 환자나, 특정 발달 단계에 놓인 영유아 등 특수하고 민감한 성질을 지닌 집단을 예로 들 수 있다. 이 경우는 모집단 자체가 소수이기도 하거니와, 접근성이 낮아 수집 가능한 데이터의 수가 극히 일부로 제한될 수밖에 없다(Armsden, McCauley, Greenberg, Burke, & Mitchell, 1990; Hilt & Pollak, 2012; Marsee & Frick, 2007). 또한, 연구대상의 특성상 관찰이나 실험을 필요로 하는 경우가 많기 때문에, 연구 과정에서 중도 탈락이나 통제의 실패 등 연구 방식으로 인한 추가적인 데이터의 누락이 발생할 수도 있다. 이러한 맥락에서 연구자들은 충분하지 못한 표본의 수, 즉 소표본(small sample)[4]으로 인해 추정의 정확성 감소 및 정규분포를 가정할 수 없다는 한계에 직면하게 된다. 이처럼 자료 수집만으로 표본 크기의 문제를 해결하기 어려울 때, 자료 분석 단계에서 소표본에 적합한 검정 방법을 사용함으로써 문제의

4) Hedges와 Olkin(1985), Nachar(2008)은 표본 크기가 20보다 작은 경우($n < 20$)를 소표본이라고 주장했다. 반면, Meek과 Ozgur, Dunning(2007), Streiner(2006), VanVoorhis와 Morgan(2007) 등에 따르면 소표본은 표본 크기가 30보다 작을 때($n < 30$)에 해당한다. 본 논문은 후자의 기준에 따라 표본 크기가 30 미만인 경우를 소표본으로 정의하기로 한다.

해결방안을 모색할 수 있다.

통계 검정법은 모집단의 확률분포(probability distribution)와 모수[5]에 대한 가정의 만족 여부에 따라 두 갈래로 나누어진다. 그중 모수 검정법(parametric methods)은 특정한 연속확률분포를 따르는 모집단으로부터 표본이 표집되었을 것이라는 가정하에 모집단을 추론하는 방법이다(Geisser & Johnson, 2006). 모수 검정법을 사용하기 위해 만족해야 할 분포에 대한 가정(parametric assumptions)으로는 모집단 분포가 정규분포를 따라야 한다는 정규성 가정(assumption of normality), 그리고 자료가 둘 이상의 독립된 집단으로 구성된 경우 각 집단이 표집된 모집단의 분산은 동일해야 한다는 등분산성 가정(homogeneity of variance)이 있다(Gaito, 1959). 선행연구에 따르면 집단의 크기가 같은 경우 등분산성 가정이 위반되더라도 모수 검정의 강건성(robustness)[6]을 유지할 수 있어 비교적 유연한 가정이라고 볼 수 있다(Carroll & Nordholm, 1975; Zimmerman, 1987). 그러나 정규성 가정은 모수 검정법이 가지고 있는 통계적 이점을 극대화하는 조건이자, 모수 검정법 사용을 결정적으로 판단하는 기준이 된다는 Sharp(1979)의 연구결과가 있었다. 소표본의 경우 정규성 가정의 위반을 피하기 어렵기 때문에, 모수 검정법을 적용하는 것이 부정확한 결과를 도출할 가능성이 높다. 따라서, 소표본과 같이 분포에 대한 가정을 만족하지 못하는 자료(이하 비모수 자료)에 대해서 전통적으로 모수 검정법의 대안(alternative) 검정에 해당하는 비모수 검정법의 적용을 고려할 필요가 있다.

비모수 검정법(non-parametric methods)[7]은 척도의 수준이 명명, 서열

---

5) 확률분포란 확률변수 X가 취할 수 있는 각 사건에 대한 확률을 표나 함수식을 이용하여 나타내는 것이다. 이때, 구체적인 확률분포는 모수에 의해 결정지어진다(이재원, 이욱기, 2019).

6) 강건성(robustness)은 통계적 검정이 하나 혹은 그 이상의 기본 가정을 위반하더라도 타당한 추론을 할 수 있는 능력으로, 견고함, 확고함 등으로 표현되기도 한다(박승호, 2000).

7) 비모수적(non-parametric)이라는 용어는 일반적으로 모집단의 분포는 알되, 모수에 대한

척도이거나, 모집단 분포의 가정을 만족할 수 없는 경우에 사용 가능한 통계 검정법이다(이종성, 강계남, 김양분, 강상진, 2007; Siegel, 1957). 비모수 검정법의 활용 가능성은 '분포에 대한 가정' 에 달려 있다. 예를 들어, 범주형 변수처럼 연속확률분포를 이룰 수 없어 정규성 분포의 가정을 만족할 수 없는 자료의 분석에는 비모수 검정법을 사용할 수 있다. 한편, 연속형 자료라고 하더라도 정규분포를 가정하기 어려울 때 역시 비모수 검정법을 활용할 수 있다. 가령, 작은 표본 크기로 인해 정규성 가정을 충족하지 못한다면 비모수 검정법을 활용하여 분포에 상관없이 가설을 검정하는 것이 가능하다. 비모수 검정법은 모수 검정법에 비해 분포의 가정을 위반하는 것에 너그럽지만, 관측치 간 독립성(independence), 확률에 기반한 표집을 전제로 한다는 특징이 있다(Kerlinger & Lee, 2000). 또 다른 특징으로는 자료를 주어진 형태 그대로 사용하지 않고, 기호(sign), 순위(rank)와 같이 질적인 형태로 적절히 변환하여 사용한다는 점을 예로 들 수 있다.

이처럼 상반된 특성을 보이는 모수 검정법과 비모수 검정법 사이에는 대안적인(alternative) 관계가 형성되어 있다(DeCoster, 2006). 구체적인 방법에 차이가 있기는 하지만, 두 검정법 모두 집단 간 차이 여부 등을 포함한 가설에 대해 검정하고자 한다는 점에서 종종 비교 선상에 놓인다. 대표적으로 모수 검정법에 해당하는 분산분석(Analysis of Variance: ANOVA)은 3개 이상의 독립된 표본이 있을 때 집단 간 평균 차이 유무를 확인하는 데 사용된다. 한편, 크루스칼-왈리스 검정(Kruskal-Wallis test)은 J개의 독립표본이 동일한 중앙값을 가진 모집단으로부터 추출되었는지를 검정하는 데 사용되는 분산분석의 대안적 비모수 검정법이다. 그 외에도 집단 간 차이를 비교하는 연구에서 쉽게 찾아볼 수 있는 대응표본 t검정

가정이 없는 상황을 가리킨다. 이와 유사한 분포 무관(distribution free)은 모수와 모집단 분포에 대해 모두 가정할 수 없는 경우를 일컫는 용어로, 오늘날 통계학 연구에서는 두 경우를 아울러 '비모수적' 이라고 지칭하고 있다(Kendall, Stuart, Ord, & Arnold, 1999).

(paired samples t test), 독립표본 t검정(independent samples t test)이 모수 검정법에 해당하며, 윌콕슨 부호 순위 검정(Wilcoxon signed rank test), 만-휘트니 검정(Mann-Whitney test)은 각각 모수 검정법의 대안적 비모수 검정법에 속한다.

사회과학 분야의 연구는 두 검정법 중 주로 모수적 방법에 대한 높은 선호를 보여 왔다. 그 영향으로 분산분석, t검정 등의 모수 검정법이 주요한 자료 분석 방법으로 채택되고, 비모수 검정법에 비해 친숙한 통계 검정법으로 여겨지게 되었다(Cox, 2006). 이러한 경향은 모수 검정법의 검정력(power)과 유효성(efficiency)[8]에 대한 연구자들의 일반적인 믿음으로부터 비롯된 것으로 볼 수 있다(Blair & Higgins, 1985; MacDonald, 1999). 가령 Kerlinger(1964)는 비모수 검정법과 비교했을 때 모수 검정법이 거의 모든 경우에 더 높은 검정력을 갖는다고 주장하며, 정규성이나 등분산성 가성을 심하게 위반한 것이 아닌 이상 비모수 검정법을 대안 검정으로 사용하는 것은 적절하지 못하다고 밝혔다. 이와 같은 주장에는 모수 검정법을 상대적으로 우월한 검정법이라 여기는 연구자들의 인식이 잘 반영되어 있다.

그러나 일부 학자들은 위와 같은 주장이 경험적 근거와 무관하게 제시되었다는 점을 지적하며, 모수 검정법의 우월성은 모수 검정법 사용을 위한 기본 가정이 완벽하게 충족되었을 경우에만 참이라 이야기할 수 있다고 반박하였다(Ary & Jacobs, 1976; Blair & Higgins, 1985; Hays, 1963; Hinkle, Wiersman, & Jurs, 1982; Kirk, 1974; Sharp, 1979). 실제로 소표본과 같은 비모수 자료를 모수 검정법을 사용하여 분석할 경우 추정의

8) 유효성은 추정치(estimator)나 실험 설계, 가설 검정 절차의 질(quality)을 비교할 때 판단의 기준이 되는 수치로, 동일한 성과를 달성하기 위해 더 적은 정보를 필요로 하는 것을 의미한다. 유효성은 일반적으로 분산(혹은 표준오차)을 통해 산출된다. 예를 들어, 두 개의 불편추정치(unbiased estimate)가 있다면 분산(표준오차)이 작은 추정치가 더 유효하다고 할 수 있다(이재원, 이욱기, 2019; Everitt, 2006).

정확도가 감소하고, 검정력이 낮아지는 등 오류를 범할 확률이 증가할 것이라는 문제가 제기되었다(정형찬, 2006). 비모수 자료에 대한 일종의 대안으로 모수 검정법의 가정에 맞게 자료를 변환하려는 시도가 있었으나, 자료 형태를 바꾸는 것은 통계적으로 강건성을 감소하게 할 수 있으며, 대부분 약한 검정력을 가지는 것으로 나타났다(Andrews, Gnanadesikan & Warner, 1971; Zimmerman & Zumbo, 1990b). 이처럼 모수 검정법이 제대로 기능하는 것을 기대하기 어려움에도 불구하고, 모수 검정법에 대한 불확실한 믿음에 기대어 모수 검정법이 기초하고 있는 가정을 만족시키지 못한 자료 분석에 모수 검정법을 사용하는 경우가 많이 발생하고 있다.

예를 들어, 2015년부터 2019년 「Korean Journal of Clinical Psychology」에서 출판된 총 224편의 논문 중 10편의 논문이 표본 크기가 30보다 작은 소표본을 사용하였다. 소표본을 사용한 10편의 연구 중 비모수 검정법을 이용한 연구는 2편에 불과했다. 구체적으로, 김영애(2017)는 현실치료가 외상 후 성장에 미치는 효과를 검정하고자 11명의 외상 경험자를 대상으로 상담적 개입을 시행했다. 상담 전, 후의 외상 후 스트레스, 외상 후 성장, 자기 효능감, 내부 통제성 점수의 변화를 살펴보기 위해 대응표본 t검정의 대안적 비모수 검정인 윌콕슨 부호 순위 검정을 사용하였다. 조은진과 손정락(2016)은 인지행동 집단치료 프로그램이 대학생의 취업 스트레스, 역기능적 태도 및 우울에 미치는 효과를 검정하고자 8명의 인지행동 집단치료 집단, 8명의 통제집단을 선발하여 연구를 진행하였다. 치료집단과 통제집단의 사전 동질성 검정을 위해 취업 스트레스, 역기능적 태도, 우울의 사전 점수로 만-휘트니 검정을 실시하였다. 또한, 사후 및 추적 단계에서 치료 집단과 통제집단 간 차이를 알아보기 위해 만-휘트니 검정을 사용하였다. 마지막으로 인지행동 집단치료 프로그램이 집단 내에서 취업 스트레스, 역기능적 태도, 우울에 미치는 영향을 알아보기 위해 사전-사후,

사전-추적 검사에 대해 윌콕슨 부호 순위 검정을 실시하였다. 10편의 연구 외에도 본 연구에서 선정한 기준에 따르면 소표본이라고 정의할 수는 없지만 표본 크기가 30에 가까운 22편의 연구가 있었다. 22편의 연구 모두 비모수 검정법을 사용하지 않았으나, 충분하지 못한 표본 크기를 연구의 한계점으로 언급했다.

통계적, 실용적 측면에서 소표본과 같은 비모수 자료에 비모수 검정법을 적용해야 하는 이유가 다음과 같이 논의되었다. 첫째, 비모수 검정법의 시행은 1종 오류가 증가할 가능성을 낮출 수 있다. Ryan(1959)은 1종 오류(type I error) 발생의 관점에서 비모수 검정법 사용의 타당성을 제시하였다. 구체적으로 1종 오류는 주어진 데이터에 대해 분석이 수행될 때마다 점차 증가하게 된다. 모수 검정법을 사용하기 위해 모집단 분포를 확인하는 과정에서 역시 1종 오류 발생 가능성이 높아지는 문제가 발생한다. 이를 고려해 볼 때, 분포에 대한 가정에 의존하지 않는 비모수 검정법을 시행한다면 불필요한 1종 오류를 높이지 않고도 자료를 분석하는 것이 가능해진다. 또한, 정규성 가정을 만족하지 않아도 되기 때문에 비교적 표본 크기로부터 자유롭게 사용할 수 있으며, 검정 절차를 간소화할 수 있다는 장점이 있다. 두 번째로 모수 검정법에 비해 더 높은 검정력을 기대할 수 있다. 자료를 그대로 활용하는 모수 검정법과 다르게 비모수 검정법은 주어진 자료를 순위화하는 등 질적인 형태로 변형하기 때문에 극단치(outlier)가 가지는 영향력을 배제할 수 있고, 결과적으로 더 큰 검정력을 가지게 된다(Zimmerman & Zumbo, 1990a). 마지막으로 비모수 검정법은 비교적 신속하고 쉽게 통계량을 구할 수 있으며, 결과에 대한 해석 및 이해가 용이하다는 점에서 실용적인 통계 검정법이라고 할 수 있다.

한편, 비모수 자료의 분석에 적용할 수 있는 또 다른 비모수적 검정 방법으로는 재표집 방법(resampling method)[9]이 있다. 재표집 방법은 연구자

가 가지고 있는 표본을 새롭게 조합하는 과정을 반복함으로써 여러 개의 가상 표본(simulated samples)을 만들어내고, 그것으로 경험적 표집분포를 생성하여 통계적인 추론을 하는 방법이다(Carsey & Harden, 2013; Good, 2005b). 재표집 방법은 비모수 검정법과 유사하게 모집단 분포 가정을 전부 충족하지 못할 때 사용되며, 대표적으로 표본 크기가 충분히 크지 않은 경우에 활용 가능하다.

재표집 방법이 가지는 고유한 특징으로는 광범위한 활용 가능성을 꼽을 수 있다. 비모수 검정법이 두 집단의 차이 비교, 세 집단의 차이 비교 등 정해진 틀을 갖춘 모형 내에서 사용될 수 있는 것과 달리, 재표집 방법은 어떤 모형에도 적용 가능하다. 또한, 컴퓨터의 계산능력을 바탕으로 하고 있다는 점에서 훨씬 복잡한 통계적 문제도 효율적으로 처리할 수 있다. 그뿐만 아니라, 수행능력의 측면에서 근사적으로 정확한 추정치를 제공하여, 1종 오류를 줄이고 검정력을 증가시키는 결과를 도출할 수 있다(Mendeş & Akkartal, 2010).

재표집 방법의 작동 원리는 시뮬레이션과 유사하다. 연구자가 가진 표본에 포함되어 있는 모집단에 관한 정보가 새로 추출된 가상 표본들이 이루고 있는 분포에도 포함되어 있다는 가정에 따라, 관찰된 표본으로부터 재표집하는 것은 이론적인 모집단 분포에서 완전히 새로운 확률표본(random sample)을 생성하는 것과 동일하게 작동한다고 볼 수 있다(Carsey & Harden, 2013). 가상 표본의 추출을 무수히 많이 반복할수록 모집단 분포에 근사한 표집분포를 가지게 되고, 이를 바탕으로 모집단의 분포에 대한 의사 결정을 내리는 것이 가능해진다. 따라서, 재표집 방법은 비모수 검정법 못지않게 모수 검정법의 기능을 충분히 대신할 수 있는 타당한 검정 방법이

9) 재표집 방법은 모집단 분포에 대한 가정을 완벽하게 만족할 필요가 없다는 점에서 비모수적인 검정법이라고 이야기할 수 있다. 하지만, 일반적으로 '비모수 검정법'이 모수 검정법과 짝을 이루는 대안적 비모수 검정을 지칭하는 데 사용되고 있기 때문에, 혼동을 방지하고자 '재표집 방법'을 별도로 분류하는 경우가 많다.

다.

재표집 방법에서 가설 검정을 위한 경험적 표집분포를 생성하는 절차는 실제로 관찰된 표본(observed sample)이 모집단에 대한 정보를 충분히 가지고 있는 표집틀(sampling frame)이라고 가정하는 것으로부터 시작한다. 이후 표본을 재추출할 때 복원추출 방식을 적용하는지, 그리고 재표집 된 표본의 크기[10]가 관찰된 표본 크기와 같은지에 따라 <그림 1>과 같이 재표집 방법을 분류할 수 있다(Rodgers, 1999).

표본 크기 (sample size) / 추출 방법 (resampling method)	부표본 (sub-sample)	완전표본 (full sample)
비복원 추출 (sample without replacement)	**잭나이프 검정 (Jackknife test)**	**임의순열 검정 (Permutation test)**
복원 추출 (sample with replacement)		**붓스트랩 검정 (Bootstrap test)**

<그림 1> 재표집 방법의 분류 (Rodgers, 1999)

일반적으로 가장 많이 사용되는 재표집 방법에는 다음과 같은 세 가지 방법이 있다. 첫째, 잭나이프 검정(jackknife test)은 관찰된 표본으로부터 한 번에 하나의 관측치 혹은 여러 개의 관측치를 순차적으로 제거하여 동일한 크기의 가상 표본을 생성한다. 둘째, 임의순열 검정(permutation test)은

10) 관찰된 표본의 크기가 n일 때, 임의순열 검정과 붓스트랩 검정에서 재표집 된 표본의 크기는 관찰된 표본의 크기와 동일하게 n이다(full sample). 반면, 잭나이프 검정은 경우에 따라 차이가 있으나 재표집 된 표본의 크기가 기본적으로 $n-1$ 이하이다(sub-sample). 이에 대한 자세한 설명은 각 재표집 방법의 이론적 배경에서 논하도록 한다.

기존 관측치가 속해 있던 집단을 무시하고 무작위로 재구성하여 새로운 관계를 가지는 표본을 생성하는 방법이다. 마지막으로 붓스트랩 검정(bootstrap test)은 연구자가 가지고 있는 표본으로부터 반복추출을 허용하여 새로운 가상 표본을 반복적으로 생성하는 방법으로, 재표집 방법 중에서도 가장 다양한 분야에서 활용되는 것으로 알려져 있다. 1930년대를 기점으로 여러 가지 재표집 방법이 고안되어왔으나, 무수히 많은 가상의 표본들을 추출하고 그에 대한 계산을 실행할 수 있는 능력이 현실적으로 부족하다는 점에서 재표집 방법의 사용은 기대만큼 활발하지 못했다. 하지만 과학기술의 진보와 함께 컴퓨터의 성능도 향상되며 매우 복잡하고 방대한 계산과정을 빠른 시간 내에 마칠 수 있는 여건이 조성되었다. 그로 인해 모수검정법을 대신하는 비모수적 검정 방법으로 재표집 방법을 제시하는 연구 또한 지속적으로 증가하는 추세를 보이고 있다.

소표본에 적용 가능한 검정법에 대한 초기의 선행연구는 주로 표본이 2개일 때 여러 조건에서 만-휘트니 검정이나 윌콕슨 부호 순위 검정과 같은 비모수 검정법과 독립표본 t 검정, 대응표본 t 검정 등 모수 검정법의 수행을 비교하는 방향으로 이루어졌다(Blair & Higgins, 1985; Zimmerman, 1987). 이를 토대로 연구의 흐름은 더 복잡한 모형을 대상으로 더욱 다양한 검정법을 사용하는 방향으로 발전해 왔다. 예를 들어, Hecke(2012)은 3개 이상의 집단 간 차이 검정 시 사용되는 모수 검정법인 일원분산분석(one-way ANOVA)과 대안적 비모수 검정법인 크루스칼-왈리스 검정을 비교하였다.

통계 검정법의 비교는 오랜 시간에 거쳐 이루어져 왔고, 연구는 점차 확장되어 오늘날까지도 지속적으로 수행되고 있다. 그럼에도 불구하고, 소표본에 적합한 통계 검정법이 무엇인지에 대한 연구자들의 공통된 의견은 아직까지 논의되지 않고 있다. 검정법 비교 연구가 이러한 한계에 부딪히게 된

이유는 모수 검정법과 비모수 검정법, 재표집 방법을 총체적으로 비교한 선행 연구의 부재 때문으로 볼 수 있다. 단순히 기존의 연구 결과를 아울러 검정법의 수행을 비교하기에는 연구마다 각기 다른 조건을 가정하고 있다는 문제로 인해 정확한 비교에 어려움이 있다.

모수 검정법과 재표집 방법을 포함한 비모수적 검정법을 비교한 Feir-Walsh와 Toothaker(1974), Gleason(2013), Mendeş와 Akkartal(2010)의 시뮬레이션 연구는 연구자가 설정한 조작변수 수준에 따라 월등한 수행을 보이는 검정법에 차이가 있음을 밝혔다. 이로 비추어볼 때, 동일한 조건하에 소표본에 적용 가능한 여러 통계 검정법을 비교 및 분석하고, 주어진 조건의 변화에 따라서 검정법들의 수행 변화를 살펴보는 것은 독자적인 연구로서 의의를 가질 것으로 예상된다.

따라서 본 연구는 모수 검정법의 기본 가정과 관련된 여러 조건에 따라 통계 검정법이 각기 다른 수준의 수행능력을 보인다는 선행연구 결과를 바탕으로, 어떤 조건에서 어떤 분석 방법이 더 효율적으로 기능하는지를 평가해보고자 한다. 나아가, 각 조건에서 어떤 분석 방법을 사용하는 것이 적합한지에 대한 논의를 이어가고자 한다. 특히 표본의 크기에 따른 검정 능력의 차이를 중점적으로 살펴보고자 모집단 분포의 형태를 통제하고 표본 크기, 효과 크기(effect size), 평균의 변동성(mean variability)만을 조작변수로 선정하기로 한다.

기존의 연구들이 주로 2개의 집단 간 차이를 검정하는 통계 검정법을 비교했던 것과 다르게(MacDonald, 1999; Meek, Ozgur, & Dunning, 2000; Meek et al., 2007; Nanna, 2002; Nanna & Sawilowsky, 1998; Siegel & Castellan, 1956), 본 연구는 이를 확장하고 추후 연구결과의 일반화 가능성을 고려하여 3개 이상의 독립된 집단 간 차이를 비교하는 모형을 가정하고자 한다. 따라서 본 연구에서는 3개 이상의 독립표본 간 차이를 검정하

는데 사용되는 모수 검정법인 일원분산분석(one-way ANOVA)과 그의 대안적 비모수 검정법인 크루스칼-왈리스 검정(Kruskal-Wallis test), 연구모형과 무관하게 사용 가능한 재표집 방법인 잭나이프 검정(jackknife test), 임의순열 검정(permutation test), 붓스트랩 검정(bootstrap test)의 1종 오류 비율과 검정력을 비교할 것이다.

연구를 위한 모의 자료의 생성과 분석은 통계 프로그램 R을 활용하여 이루어질 것이다. 구체적으로 프로그램 R에서 연구자가 설정한 조건과 일치하는 이론적 모집단으로부터 임의의 가상 표본을 생성해주고, 가상의 표본들로부터 검정 통계량을 보고해주는 몬테카를로 시뮬레이션(Monte Carlo Simulation) 기능을 실행할 수 있다(이현숙, 김수진, 전수현, 2010; Muthén & Muthén, 2002). 연구 결과를 일반화하기 위해서는 현실 연구 장면에서 나타날 수 있는 문제점이 잘 반영된 조건을 생성하고, 조건에 따라 우수한 수행을 보이는 검정법에 변화가 있는지를 살펴보아야 한다. 따라서 연구자의 의도에 맞는 다양한 조건을 생성할 수 있는 몬테카를로 시뮬레이션 기능을 활용하여, 각 조건에서 통계 검정법의 수행에 어떠한 차이가 나타나는지를 확인할 것이다.

Ⅱ. 이론적 배경

1. 크루스칼-왈리스 검정(Kruskal-Wallis Test)

크루스칼-왈리스 검정은 Kruskal과 Wallis(1952)에 의해 고안된 비모수 검정법으로, 2개의 독립된 표본 간 차이를 확인하는 만-휘트니 검정(Mann-Whitney test)을 *J*개의 독립표본으로 확장한 것이다(Breslow, 1970, DeCoster, 2006). 앞서 설명한 것처럼 수집된 표본 데이터가 일원분산분석(one-way ANOVA)의 가정을 만족하지 못하는 경우에 *F*검정을 대신하여 크루스칼-왈리스 검정을 사용할 수 있다(Pagno, 1994). 요약하자면 크루스칼-왈리스 검정(이하 KW 검정)은 *J*개의 독립표본이 동일한 모집단으로부터 추출되었는지 검정하는 방법이다.

KW 검정은 관측치를 관찰된 형태 그대로 사용하는 것이 아니라, 등위(rank)의 형태로 변환한다는 점에서 모수 검정법과 차이가 있다. 일반적으로 등위로의 변환은 자료가 가지고 있는 정보를 충분히 활용할 수 없도록 만들기 때문에 잘 사용되지 않는 자료 변환 방법이다. 그러나 비모수 검정에서 등위 개념을 사용하는 것에는 몇 가지 장점이 기대된다(Kruskal & Wallis, 1952). 첫째, 계산과정이 비교적 단순하다. 등위를 사용한 검정에서 원래의 점수를 '등위화' 하기만 하면 간단한 함수를 통해 검정 통계량을 쉽게 산출할 수 있다. 둘째, 관측치가 표집된 모집단의 분포에 대해 매우 일반적인(general) 가정만을 전제로 한다. KW 검정은 관측치가 서로 독립적이며, 한 표본에 속한 모든 관측치는 단일 모집단으로부터 비롯된다는 것, 그리고 *J*개의 모집단이 거의 동일한 형태를 취하고 있음을 가정한다. 셋째, 서열척도 수준의 데이터에서도 사용 가능하다. 마지막으로 모수 검정법의 가정이 지나치게 비현실적임에도 이를 사용함으로써 발생할 수 있는 문제를

방지하고, 실제 관심의 대상이 되는 집단 간 차이를 잡아낼 수 있다.

KW 검정이 가지는 또 다른 차이는 평균이 아닌 중앙값을 대푯값으로 한다는 점이다. KW 검정에서 등위가 사용된다는 사실을 고려할 때 평균보다는 중앙값이 자료를 더 잘 요약하는 값에 해당한다. 비모수 검정법이 평균보다 중앙값 차이에 더 민감하게 반응한다는 Howell(1992)의 주장에 따라, 다음과 같은 가설을 통해 집단 간 차이의 유의성을 확인할 수 있다.

H_0 : J개 모집단의 중앙값은 동일하다.

H_1 : J개 모집단의 중앙값 중에서 최소 두 개는 상이하다.

1) KW 검정의 원리 및 시행절차

KW 검정을 시행하기 위해서는 우선 J개 집단에 속한 모든 사례에 대하여 낮은 점수부터 높은 점수에 이르기까지 순차적으로 등위를 결정해야 한다. 등위를 결정한 후에는 집단별로 등위의 합(R_i)을 구할 수 있다. 이때 영가설이 참이라면, 다시 말해 J개 표본이 동일한 모집단으로부터 추출되었다면 각 집단 내 등위합은 크게 다르지 않아야 한다. 이러한 원리를 기초로 하여 각 집단의 등위합 R_i로부터 KW 검정의 검정 통계량인 H를 산출할 수 있다. 검정 통계량 H가 크면 클수록 영가설을 기각할 가능성은 높아진다.

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{J} \frac{R_i^2}{n_i} - 3(N+1) \qquad \text{식 (2.1)}$$

J : 집단의 수

n_i : i번째 집단의 사례 수

N : 전체 사례 수

R_i : i번째 집단의 등위합

만약 동점(tie)을 이루고 있는 관측치들이 있다면, 동점이라는 사실을 고려하지 않았을 때 매겨지는 등위를 평균화하여 동점인 관측치에 공통적으로 부여한다. 동등위는 검정 통계량의 크기에 영향을 미치기 때문에 동점이 있는 표본에 대한 H 검정 통계량은 다음과 같이 교정되어야 한다.

$$C = 1 - \frac{\sum T}{N^3 - N} \qquad \text{식 (2.2)}$$

$T = (t-1)t(t+1) = t^3 - t$

t : 동등위의 개수

$$H' = \frac{H}{C} = \frac{\frac{12}{N(N+1)} \sum_{i=1}^{J} \frac{R_i^2}{n_i} - 3(N+1)}{1 - \sum T/(N^3 - N)} \qquad \text{식 (2.3)}$$

식 (2.1)을 통해 구한 H값을 식 (2.2)로 나누어주면 교정된 H 검정 통계량을 구할 수 있다(식 (2.3) 참조). 이때, T는 동등위의 개수 t를 사용해 구해진다. 예를 들어, 전체 관측치 중 2.0이라는 점수를 갖는 관측치가 총 4개 있다면, T는 60이라는 값을 가지게 된다($T = 4^3 - 4$). 동등위의 개수(t)는 항상 전체 사례수(N)보다 작거나 같다. 결과적으로 C는 0과 1 사이의 값을 갖기 때문에, C를 통해 교정된 H 검정 통계량은 언제나 기존의 H값보다 큰 값으로 나타난다.

표본들이 동일한 연속확률분포를 가진 모집단에서 표집되고, 각 집단의 표본 크기가 너무 작지 않으면($n_i > 5$) H 통계량의 확률표집분포(random sampling distribution)는 자유도가 $J-1$인 χ^2 분포에 가까워진다(식 (2.4)). 따라서, 검정 통계량이 근사χ^2 분포를 따른다는 가정하에 가설을 검정하게 된다.

$$H \sim \chi^2(J-1) \qquad \text{식 (2.4)}$$

이종성 외(2007)의 가상 데이터를 통해 KW 검정의 시행절차를 아래와 같이 구체적으로 설명할 수 있다. 자료는 각 12명의 초등학생이 속한 세 집단의 변별학습과제 성취도 점수로 구성되어 있다. 집단마다 형태, 색깔, 크기라는 서로 다른 단서(cue)를 제공했을 때, 집단 간 성취도 점수에 차이가 있는지 확인하는 것이 KW 검정의 시행 목적이다.

먼저, 세 집단의 성취도 점수를 등위화하고 각 집단의 등위합을 구한 것은 <표 1>과 같다. 위의 가상 데이터에는 동등위가 포함되어 있기 때문에 동등위를 고려한 검정 통계량 산출이 필요하다. 예를 들어, 11점인 관측치 3개는 동점임을 고려하지 않았을 때 6, 7, 8등에 해당하고, 이들의 평균인 7이 11점인 관측치에 부여되는 등위에 해당한다. 정리하자면 두 개 등위가 동등한 경우가 4번(11.5등, 17.5등, 29.5등, 34.5등), 세 개의 등위가 동등위인 경우 1번(7등), 네 개의 등위가 동등위인 경우 1번(26.5등)이다. 따라서, $\sum T$를 활용해 구해진 교정값 C로 H 검정 통계량을 교정해 준 값은 식 (2.8)과 같다(식 (2.5)~(2.7) 참조).

$$\sum T = 4[(2)^3 - 2] + [(3)^3 - 3] + [(4)^3 - 4] = 108 \qquad \text{식 (2.5)}$$

$$C = 1 - \frac{108}{36^3 - 36} = .998 \qquad \text{식 (2.6)}$$

$$H = \frac{12}{36 \cdot 37}\left[\frac{(139)^2}{12} + \frac{(200)^2}{12} + \frac{(327)^2}{12}\right] - 3 \cdot 37 \qquad \text{식 (2.7)}$$

$$= 13.81$$

$$H' = \frac{13.81}{.998} = 13.85 \sim \chi^2(2) \qquad \text{식 (2.8)}$$

<표 1> 단서에 따른 변별학습과제 성취도 점수와 등위를 통한 H 검정 통계량 (이종성 외, 2007)

집단1(형태)		집단2(색깔)		집단3(크기)	
성취도	등위	성취도	등위	성취도	등위
6	1	31	34.5	13	10
11	7	7	2	32	36
12	9	9	4	31	34.5
20	19	11	7	30	33
24	23	16	14	28	31
21	20	19	17.5	29	32
18	16	17	15	25	24
15	13	11	7	26	26.5
14	11.5	22	21	26	26.5
10	5	23	22	27	29.5
8	3	27	29.5	26	26.5
14	11.5	26	26.5	19	17.5
$n_1 = 12$		$n_2 = 12$		$n_3 = 12$	
$R_1 = 139.0$		$R_2 = 200.0$		$R_3 = 327.0$	
$N = 36$					

자유도가 2인 χ^2 분포상에서 교정된 H 검정 통계량인 13.85보다 극단적인 값이 나타나게 될 확률(significance probability) p는 .001로, 유의수준(significance level; α)이 .05일 때 영가설을 기각한다. 즉, 세 집단이 표집된 모집단의 중앙값에는 차이가 있으며, 이는 주어진 단서에 따라 변별학습의 성취도에 차이가 있다고 결론 내릴 수 있다. 세 집단 중 구체적으로 차이가 발생한 집단에 대해 알고 싶을 경우, 만-휘트니 검정을 통해 두 집단씩 비교함으로써 사후 검정(post-hoc)을 시행할 수 있다(DeCoster, 2006).

2) KW 검정의 특성

비모수 검정법은 일반적으로 모수 검정법에 비해 낮은 검정력을 갖는 것으로 알려져, 모수 검정법의 가정을 모두 만족하는 자료에 대해서는 모수 검정법을 사용하는 것이 권장되고 있다(이종성 외, 2007; Blair & Higgins, 1980). 그러나 일부 연구자들은 표본이 특정 조건을 만족하는 경우에 일반적인 예상과 반대되는 연구결과를 얻을 수 있다고 주장했다. 가령 Neave와 Worthington(1988)은 표본이 정규분포를 이루는 모집단으로부터 표집되었다는 사실을 확신할 수 있다면 KW 검정의 검정력이 일원분산분석에서 F 검정의 검정력과 거의 유사하다고 밝혔다.

모수 가정을 만족하지 못하는 자료에 통계 검정법을 적용했을 때에도 대립되는 연구결과가 관찰되었다. 일반적으로 정규분포를 이루기 어려운 소표본의 경우 KW 검정과 같은 비모수 검정법을 사용하는 것이 관행처럼 여겨져 왔다. 하지만, 소표본이더라도 표본 크기가 지나치게 작아지면 검정 결과가 더 이상 정확하지 않다는 연구결과(Kruskal & Wallis, 1952)를 토대로 극단적으로 작은 크기의 표본에는 KW 검정의 적용을 신중히 고려해야 할

필요가 있다. 또한, Zimmerman과 Zumbo(1990a)는 KW 검정을 사용함으로써 극단치가 미치는 영향을 통제할 수 있으므로 오히려 더 믿을만한 결과를 얻을 수 있다고 주장했다. KW 검정이 극단치의 영향력을 통제할 수 있는 이유는 등위의 개념을 사용하기 때문인 것으로 볼 수 있다. 하지만 극단치 통제로 인한 검정력 증가와는 별개로 등위를 사용하는 것 자체가 검정력을 낮아지게 한다는 주장 또한 있었다.

이러한 주장을 뒷받침하듯, 자료가 가진 정보의 수준에 제한을 가하는 등위 검정법보다 원자료의 성격을 그대로 보존할 수 있는 다른 검정법들이 더 우월한 것으로 나타났다. 예를 들어, KW 검정과 임의순열 검정은 모두 비모수적인 검정 방법임에도 불구하고 원자료의 형태를 그대로 유지하며 사용하는 임의순열 검정의 검정력이 더 높다는 연구결과가 있었다(Adams & Anthony, 1996; Ludbrook & Dudley, 1998). 그러나 이와 반대로 자료를 등위화하는 것은 자료 자체에 전혀 영향을 미치지 않고, 단순히 일종의 잡음(noise)을 제거할 뿐이라는 의견도 있었다(Sawilowsky, 1993). 더불어 연구자들은 등위화가 검정력의 감소를 유발하지 않고, 오히려 검정력을 증가시킨다고 주장하기도 했다(Langbehn, Berger, Higgins, Blair & Mallows, 2000).

선행연구를 종합적으로 살펴보면 모수 검정법 시행을 위한 가정을 만족했을 때와 그렇지 않을 때, 모수 검정법과 비모수 검정법의 수행에 대해 대립되는 연구결과가 나타나고 있다. 더불어 어떤 가정이 위배되었는지, 그리고 얼마나 위배되었는지에 따라서도 통계 검정법의 수행이 달라지는 양상이 나타난다. 따라서, 위반된 가정의 종류와 그 정도를 달리했을 때, 통계 검정법의 수행을 비교하여 각 조건에서 더 우수한 방법이 무엇인지를 명확히 해야 할 필요가 있다. 또, 표본 크기에 따라 모수와 비모수 검정법의 수행에 어떤 차이가 나타나는지, 우수한 수행 결과를 산출하기 위한 적절한 표본 크기는

무엇인지에 대해서도 확인해보아야 할 것이다.

2. 잭나이프 검정(Jackknife Test)

대개 모집단의 분포에 대한 정보는 완전하게 알려져 있지 않다. 그러므로 추리통계의 가설 검정을 위해서는 표집분포의 성질을 이용하여 모집단의 속성을 추정하는 절차를 거쳐야 한다. 재표집 방법은 관찰된 표본을 모집단으로 간주하고 이로부터 수많은 가상의 표본(simulated sample)을 반복 추출한다는 점에서 표집분포가 형성되는 원리와 유사하다(Good, 2005b). 안정적이고 덜 편향(bias)된 추정치, 즉 표집오차가 작은 통계량을 얻을 수 있는 유용한 통계 기법으로 알려진 잭나이프 검정 역시 재표집을 활용한 대표적인 검정법 중 하나다. 잭나이프 검정의 재표집은 관찰된 표본(observed sample)에서 얻은 데이터가 곧 모집단과 같다는 가정에 따라 이루어진다. 유사모집단으로부터 무선적으로 표집된 더 작은 크기의 확률표본들은 경험적 분포를 그려내고, 이는 통계량의 편향과 표준오차를 추론하는 데 사용된다(Rodgers, 1999). 통계량의 안정성과 편향은 하나 이상의 관측치(혹은 집단)가 원래의 관찰된 표본으로부터 제거되었을 때 미치는 영향력을 바탕으로 평가된다(Carsey & Harden, 2013; Mosteller & Tukey, 1977; Rodgers, 1999).

잭나이프 검정은 주어진 표본을 활용하여 편향을 추정하는 방법에 대한 고민으로부터 비롯되었다. Quenouille은 시계열 데이터(time series data)에서 계열상관(serial correlation)[11] 추정치의 편향을 줄이기 위한 통계기법을 제안하였고, 이를 보완한 것이 잭나이프 검정의 기저가 되었다(Quenouille, 1949; Quenouille, 1956). 초기의 이론을 바탕으로 Tukey(1958)는 표본을 재표집하는 절차가 동일하게 반복 시행되고, 각각

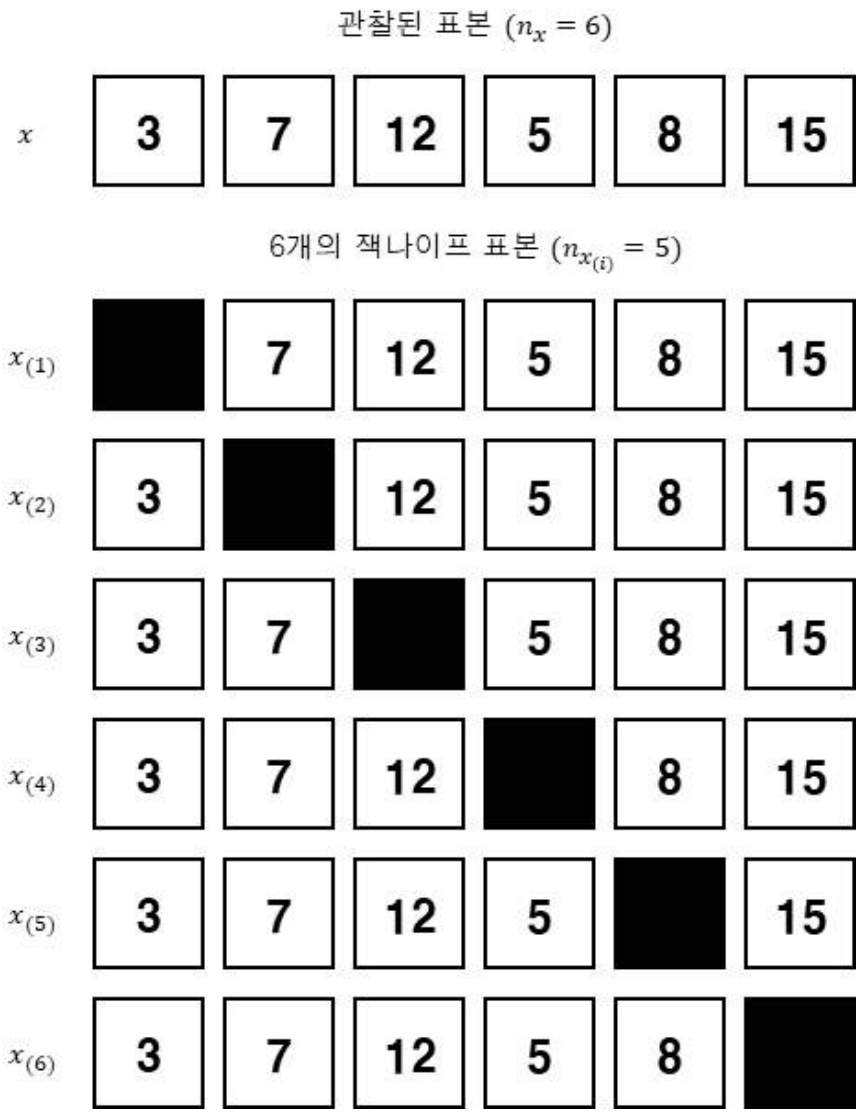
11) 계열상관은 시간 또는 공간적으로 연속된 일련의 관측치들 사이에 존재하는 상관관계로 시계열 자료에 내재된 시점 간의 상관을 의미한다. 자기상관(autocorrelation)이라고도 불린다(이강원, 송호웅, 2016).

독립적인 분포를 이룬다는 가정(independent and identically distributed; i.i.d)이 있다면 표본 통계량의 분산을 추정할 수 있다고 주장했다. 또한, 구해진 표본 통계량의 분산은 자유도가 $n-1$인 t분포를 따른다고 주장하며 Quenouille의 이론을 보다 확장하였다. 잭나이프라는 이름을 통해 연상할 수 있듯이 잭나이프는 한정된 목적으로 사용되는 것이 아니라, 다양한 문제에 대해 빠르고 간단한 해결책을 제시할 수 있는 실용적인 통계 기법이다(Cameron & Trivedi, 2005). 따라서, 어떠한 문제에 특화된 통계 기법을 사용하기 어려운 경우에는 잭나이프 검정을 대신 사용할 수 있다(Mosteller & Tukey, 1977).

1) 잭나이프 검정의 원리 및 시행절차

잭나이프 추정치를 얻는 것은 연구자가 가진 표본으로부터 여러 개의 부표본(sub-sample)을 추출하고, 체계적으로 통계량을 재계산하는 원리를 기반으로 이루어진다. 부표본(이하 잭나이프 표본)을 생성하는 방법은 기존의 데이터에서 한 번에 제거하는 관측치의 수에 따라 달라진다. 가장 기본적인 형태이자 잭나이프 검정을 적용한 선행연구에서 많이 사용되어 온 방법은 한 개의 관측치를 삭제하는 방법(delete-1 observation jackknife)이다(Efron, 1982). 한 개의 관측치를 삭제하는 방법을 시각적으로 표현하면 <그림 2>와 같이 나타낼 수 있다. 즉, 관찰된 표본의 크기가 n이라면 하나의 관측치를 제거하고 남아있는 $n-1$개의 관측치가 하나의 잭나이프 표본을 이룬다. 이후 데이터를 원 상태로 복원하고, 또 다른 관측치를 차례로 제거해나가며 모든 관측치들이 한 번씩 제거될 때까지 재표집 과정을 반복적으로 실시하는 것이 잭나이프 표본을 생성하는 알고리즘이라고 볼 수 있다(Carsey & Harden, 2013; Rodgers, 1999). 위 과정에 따르면 기존 표본

크기가 n일 때 재표집된 잭나이프 표본의 수는 n개로 동일하며, 각 잭나이프 표본의 크기는 $n-1$이다. $x=(x_1, \cdots, x_n)$는 관찰된 확률표본을 의미하며, $x_{(i)}=(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$는 i번째 잭나이프 표본(i번째 관측치가 제거된 부표본)을 의미한다(Rizzo, 2008).



<그림 2> 잭나이프 표본 생성의 예

잭나이프 추정치는 잭나이프 표본에서 구할 수 있는 표본 통계량을 종합하여 산출한다. 우선 i번째 잭나이프 표본으로부터 얻어지는 통계량은 식 (3.1)과 같이 나타낼 수 있다. 만약 <그림 2>와 같이 관찰된 표본 x가 있고, 알고자 하는 통계량이 평균이라고 가정한다면 첫 번째 잭나이프 표본($x_{(1)}$)에서 구한 평균은 $\hat{\theta}_{(1)}$와 같이 표시한다.

$$\hat{\theta}_{(i)} := s(X_{[i]}) \qquad \text{식 (3.1)}$$

이 값들은 모수를 추정할 때 발생하는 편향의 잭나이프 추정치(jackknife estimate of bias)를 구하는 데 사용된다. 잭나이프 편향 추정치는 잭나이프 표본으로부터 구해진 표본 통계량들의 평균(식 (3.3))과 원래의 관찰된 표본에서 나타나는 추정치($\hat{\theta}$) 사이의 차이를 교정한 값이라고 볼 수 있다 (식 (3.2) 참조).

$$\widehat{Bias}_{jack} = (n-1)(\hat{\theta}_{(.)} - \hat{\theta}) \qquad \text{식 (3.2)}$$

$$\hat{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)} \qquad \text{식 (3.3)}$$

n : 표본의 크기

위에서 구해진 잭나이프 편향 추정치를 모수 추정치에서 빼면 식 (3.4)와 같은 편향이 교정된 불편추정치($\hat{\theta}_{jack}$)를 얻을 수 있다.

$$\begin{aligned}\hat{\theta}_{jack} &= \hat{\theta} - \widehat{Bias}_{jack} \\ &= \hat{\theta} - (n-1)(\hat{\theta}_{(.)} - \hat{\theta}) \\ &= n\hat{\theta} - (n-1)\hat{\theta}_{(.)}\end{aligned} \qquad \text{식 (3.4)}$$

한편, Tukey(1958)는 잭나이프 표준오차 추정치(jackknife estimate of standard error)를 식 (3.5)와 같이 정의했다.

$$\widehat{se}_{jack} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^2} \qquad \text{식 (3.5)}$$

식 (3.5)가 표준오차(standard error)를 추정하게 되는 과정을 구체적으로 살펴보자면 다음과 같다. 예를 들어, 추정하고자 하는 모수가 모평균이라고 가정했을 때, 기존의 표본에서 얻은 통계량과 잭나이프 표본에서 얻을 수 있는 통계량 사이에는 아래와 같은 관계가 성립하게 된다(식 (3.6) 참조).

$$\begin{aligned}(\hat{\theta}_{(i)} - \hat{\theta}_{(.)}) &= \frac{n\bar{x} - x_i}{n-1} - \frac{1}{n}\sum_{i=1}^{n}\bar{x}_{(i)} \\ &= \frac{1}{n-1}(n\bar{x} - x_i - \frac{1}{n}\sum_{i=1}^{n} n\bar{x} - x_i) \\ &= \frac{1}{n-1}(\bar{x} - x_i)\end{aligned} \qquad \text{식 (3.6)}$$

표본평균의 표준편차(standard deviation)를 구하는 공식이

$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$이기 때문에, 표준오차는 이를 $\sqrt{n}$으로 나눈 $\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2}$으로 나타낼 수 있다. 표본평균의 표준오차를 식 (3.6)에서 통계량 간 관계를 고려하여 다시 정리하면 식 (3.7)과 같이 표현할 수 있다.

$$\sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)}-\hat{\theta}_{(.)})^2}=\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad \text{식 (3.7)}$$

즉, 잭나이프 표준오차 추정치는 표본평균의 표준오차의 불편추정치(unbiased estimate)라고 볼 수 있다(McIntosh, 2016).

2) 잭나이프 검정의 특성

잭나이프 검정은 다양한 상황에 적용할 수 있지만, 초기의 잭나이프가 시계열 자료에서의 활용을 목적으로 등장했던 것과는 반대로 상관을 이루는 데이터(correlated data)나 시계열 데이터(time series data)에 잭나이프를 사용하는 것이 부적합하다는 연구결과가 있었다(McIntosh, 2016). 잭나이프 검정을 사용하기 위해서는 확률변수 간 독립성, 그리고 관측치 분포의 동일성 가정이 만족되어야 한다(independent and identically distributed; i.i.d). 만약 i.i.d 가정을 만족하지 못하는 자료의 경우에는 의미 없는 결과가 도출되기 때문이다.

잭나이프 검정은 표집으로 인한 오차를 최소화하며 모수를 추정하는 통계기법으로써 다음과 같은 장점이 있다. 첫 번째로 모수 검정법의 가정을 위

반하는 자료에 대해 높은 강건성을 가진다는 비모수적 검정법 고유의 특성이 있다(Carsey & Harden, 2013). 이는 잭나이프 검정 결과가 편향을 교정한 불편추정치로 산출되기 때문으로 볼 수 있다. 두 번째는 극단치나 영향력이 높은 관측치가 무엇인지 찾아내는 데 유용하게 사용될 수 있다는 점이다(Carsey & Harden, 2013; Su & Tsai, 2011). 검정 원리에 따르면 반복시행마다 동일한 수의 관측치가 제거되어 잭나이프 표본은 모두 같은 크기를 갖는다. 따라서 특정 관측치가 표본에 포함된 경우와 그렇지 않은 경우의 통계량을 비교하여, 개별 관측치가 가지는 영향력에 대해 살펴볼 수 있다. 연구자는 각 잭나이프 표본에서 구해지는 통계량이 서로 유사한지 확인하고, 급격한 차이가 발생한다면 해당 관측치를 분석에서 제외하는 것을 고려할 수 있다. 혹은 자료에 적합한 새로운 모형을 적용함으로써 극단치로 인한 문제를 해결하는 것도 가능하다(Tressler & Smotherman, 2009). 셋째, 잭나이프 검정은 시행 횟수와 무관하게 동일한 자료에 대해서는 매번 동일한 결과를 제시한다. 그 이유는 잭나이프 표본 생성 원리에 따라 항상 같은 부표본이 생성되어, 그로부터 얻어지는 추정치 역시 매번 동일할 수밖에 없기 때문이다. 이러한 특성으로 인해 잭나이프 검정은 연구문제를 여러 차례에 걸쳐 반복적으로 입증해야 하는 경우에 유용하게 사용된다. 마지막으로 잭나이프 검정은 다양한 구조를 가진 데이터에 적용 가능하다. 특히 층화 표집, 다단계 표집, 표본 가중치를 반영한 표집 등 복잡하게 표집 설계된 데이터에 보다 쉽게 적용할 수 있다(Carsey & Harden, 2013).

이처럼 잭나이프 검정은 다방면으로 유용하게 사용될 수 있는 가능성이 있지만 동시에 몇 가지 한계점을 가지고 있기도 하다. 우선 표본 크기가 작은 경우 문제가 발생할 수 있다. 잭나이프 검정을 사용하는 데 있어 표본의 크기는 곧 재표집 가능한 횟수를 의미하기 때문이다. 재표집을 반복할 수 있는 횟수, 즉 생성 가능한 잭나이프 표본의 수가 증가할수록 추정치의 정

확성은 높아질 것이다. 하지만, 1개의 관측치를 제거하는 방법에서는 표본 크기가 n인 데이터에 대해 재표집 횟수 역시 n으로 고정되기 때문에, 사례수 n의 크기가 상대적으로 작은 소표본에 사용될 경우 추정치의 정확성에 대한 문제가 예상된다(Carsey & Harden, 2013). 잭나이프 검정의 또 다른 한계는 가용 통계량이 제한적이라는 점이다. 구체적으로 1개 관측치 제거 방법을 사용할 때, 추정하고자 하는 모수는 매끄럽게(smooth) 변화하는 값이어야 한다. 여기서 매끄럽게 변화한다는 것은 데이터에 아주 작은 변화가 있으면 통계량도 그에 대응할 만큼 작게 변화함을 의미한다(Rizzo, 2008). 예를 들어, 동일한 표본에서의 평균과 중앙값을 비교해보면 하나의 관측치가 달라졌을 때 평균에 비해 중앙값이 더욱 급격하게 변화함을 알 수 있다. 그러므로 중앙값처럼 재표집 된 표본마다 통계치가 급격하게 변할 수 있는 값을 추정하고자 하면 잭나이프 검정은 제대로 기능하지 못하게 된다. 이처럼 통계량의 변화가 유연하지 않을 경우에는 한 번에 2개 이상의 관측치를 제거하는 방법(delete-d observation jackknife)을 대안으로 활용할 수 있다(Efron & Tibshirani, 1993). 이 방법은 통계량의 변화에 대한 규제(smoothness requirements)를 완화하여 일관된 분산추정치를 얻도록 함으로써 문제를 해결한다.

3. 임의순열 검정(Permutation Test)

Fisher(1935)는 순열의 원리(permutation principle)를 재표집에 활용하고, 이를 기반으로 가설을 검정하는 임의순열 검정(permutation tests)을 제안하였다. 가장 오래된 형태의 재표집 방법이 등장한 이래로, Good(2000, 2005a), Manly(1997), Pitman(1937a-b, 1938) 등의 연구자들을 통해 임의순열 검정에 관한 연구가 진행되어 왔다.

임의순열 검정은 재표집 방법이자 비모수적인 검정 방법에 속하지만, 전형적인 틀에서는 벗어나 있다. 관측치가 하나의 개인을 나타내고, 관찰된 표본이 k개의 집단($k \geq 2$)으로 구성되어 있다고 가정해보자. 일반적인 재표집 과정에서는 같은 집단에 소속되어 있는 관측치들만을 대상으로 재표집을 하므로, 기존의 소속을 벗어나 다른 집단의 일원으로 표집되는 일이 발생하지 않는다. 반면, 임의순열 검정의 재표집된 표본은 k개 집단을 모두 합친 합동표본(pooled sample)으로부터 재표집된다. 따라서 새로 만들어진 임의순열 표본에서 관측치들은 기존과 다른 새로운 집단에 소속되기도 한다. 또 다른 특이점은 비모수적 검정의 측면에서 확인할 수 있다. 비모수적 검정은 모집단 분포의 영향을 받지 않기 때문에 분포 무관 검정법이라고 불린다. 그러나 임의순열 검정의 경우 비모수적 검정의 특성을 보임과 동시에 분포에 대한 몇 가지 가정을 필요로 하기 때문에 완전한 분포 무관이라고 보기 어렵다(LaFleur & Greevy, 2009).

임의순열 검정의 활용 방안 중 하나는 여러 집단 간 차이를 확인하는 것에 있다. 검정 절차에 모수 검정법을 위한 가정의 만족이 반영되는지와 무관하게 대부분의 집단 비교 연구는 일정한 가정 아래 위치 모수(location parameter)의 차이, 혹은 분포에 대한 차이를 확인하는 절차로 이루어진다. 만약 관찰된 표본이 정규분포를 이루고 분산이 동일한 유한 모집단으로

부터 표집된 것이 확실하다면, 모수 검정법인 t 검정이나 분산분석을 통해 평균 차이에 대한 검정을 시행할 수 있을 것이다. 그러나 모수 가정을 만족하기 어려운 경우에는 상대적으로 유연한 가정을 가지며, 가정의 위반이 있더라도 충분히 타당한 추론을 할 수 있는 임의순열 검정을 사용해야 한다. 두 개 이상의 집단을 비교하는 경우 임의순열 검정의 영가설과 연구가설은 다음과 같다.

$$H_0 : F_1 = \cdots = F_k$$

$$H_1 : F_i \neq F_j \text{ (어떤 } i, j\text{에 대하여)}$$

영가설이 의미하는 바는 k개의 집단이 모두 F라는 분포를 가지는 모집단으로부터 표집된 확률표본이라는 것이다. 즉, 영가설이 참이라는 가정하에 k개의 집단은 같은 모집단으로부터 비롯되었기 때문에 집단 간 차이가 나타나지 않아야 한다. 이와 같은 원리에 비추어 볼 때, 임의순열 검정은 재표집된 표본에서의 통계량이 이루는 분포와 실제 표본에서 관측된 통계량을 비교하여 모집단 분포의 동일성에 대해 검정하는 통계적 방법이라고 할 수 있다. 단순히 둘 이상의 모집단을 비교하는 것 이외에도 회귀모형이나 잠재성장모형에서의 추론에까지 임의순열 검정이 확대 적용되고 있다(Draper & Stoneman, 1966; Kennedy, 1995; Kennedy & Cade, 1996; Raz, 1989; Zerbe, 1979).

임의순열 검정의 시행은 몇 가지 가정을 전제로 한다. 가장 핵심이 되는 가정은 관측치 간 교환 가능성(exchangeability)이다. 교환 가능성이란 문자 그대로 각 관측치가 소속된 집단을 나타내는 표식(label)을 자유롭게 교환할 수 있음을 의미한다. 임의순열 검정이 표식의 교환 가능성을 가정할 수 있는 이유는 모든 개별 관측치가 소속된 집단, 다시 말해 독립변수의 처

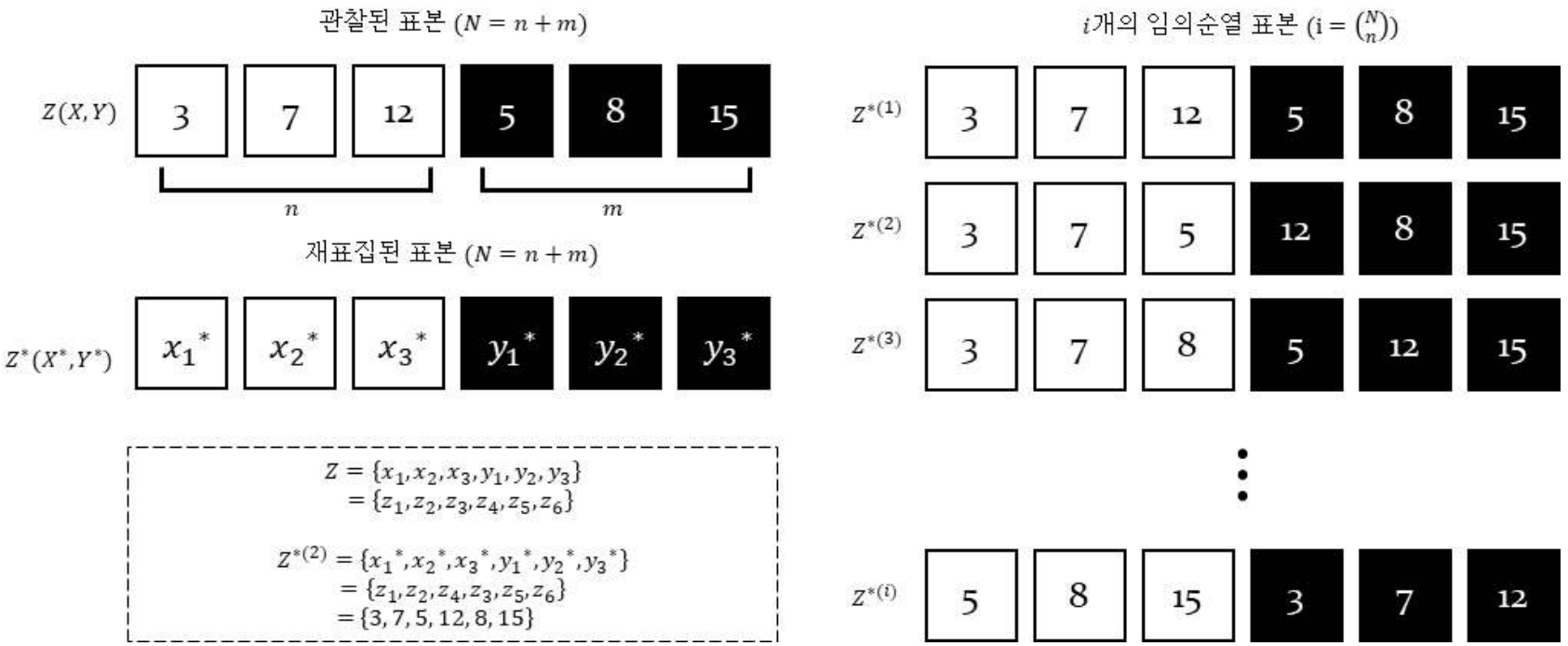
치수준과 관계없이 동일한 분포에서 비롯된다는 영가설 때문이다 (Kennedy, 1995). 이에 따르면, 재표집으로 인해 어떤 관측치가 기존과 다른 집단에 처치된다고 해도 결과에는 변화가 없으므로 임의순열 검정에서 관측치의 처치수준을 무작위로 재배정하는 것이 가능해진다. 잭나이프에서 요구하는 독립동일분포(i.i.d) 가정과 비교하면 이 가정은 비교적 덜 엄격한 가정으로, 관측치 간 독립성이 없는 경우에도 임의순열 검정을 적용할 수 있도록 한다. 또 다른 가정으로는 재표집을 통해 구해진 검정 통계량의 분포가 대칭을 이룬다는 것, 그리고 임의순열 검정이 집단 간 평균 차이와 같은 변화(shifts)를 검정하는 것을 목적으로 한다는 가정이 있다 (LaFleur & Greevy, 2009). 특히 명확한 처치 변수를 가지고 집단이 구분되는 실험연구에서 임의순열 검정이 유용하게 활용된다. 만일, 관찰된 표본이 실험집단과 통제집단으로 나뉜다면 처치수준에 따라 유의한 변화가 나타나는지를 평가할 수 있다.

1) 임의순열 검정의 원리 및 시행절차

임의순열 검정의 재표집 절차는 '관측치의 표식을 물리적으로 변경하는 행위'로 정의된다(LaFleur & Greevy, 2009). 이것이 의미하는 바에 관해 설명하기에 앞서, 크기가 n인 집단 X와 크기가 m인 집단 Y로 구성된 표본이 관찰되었다고 가정하자. 이와 같은 원자료를 바탕으로 처치 조건을 무시하고 두 집단을 통합한 크기 $N(=n+m)$의 합동표본 Z를 가정할 수 있다. 이때, 임의순열 검정에서의 영가설에 따르면 X와 Y는 동일한 분포를 가지고($F_X = F_Y$), 합동표본 Z 역시 같은 분포를 가지게 된다($F_X = F_Y = F_Z$). 합동표본 Z에서 재표집된 표본 Z^*의 통계량들이 이루는 분포도 결국은 동일 선상에 놓인다($F_Z = F_{Z^*}$). 이러한 조건에서 집단 차이를 검정하고자 한다

면, 두 집단의 차이 값이 이루는 분포는 평균이 0인 분포 F_{Z^*}를 따르게 될 것이다.

임의순열 표본은 앞서 설명한 것처럼 표식을 물리적으로 변경함으로써 생성된다. 표식의 물리적 변경을 통한 임의순열 표본의 생성은 <그림 3>을 통해 설명할 수 있다. 관찰된 표본은 집단 X에 속한 3개의 관측치 $\{3, 7, 12\}$와 Y에 속한 3개의 관측치 $\{5, 8, 15\}$로 구성되어 있고, 6개 관측치의 소속을 무시하고 하나의 표본으로 통합시키면 합동표본 Z가 만들어진다. 이후 필요한 절차는 합동표본을 다시 여러 개의 집단으로 구성된 재표집 표본 $Z^*(X^*, Y^*)$로 만들기 위해 자료를 분할(partitioning)하는 것이다. 자료를 분할할 때 유의해야 할 점은 기존 집단의 크기를 유지하여 완전표본(full sample)의 형태를 이룰 수 있도록 해야 한다는 점이다. 따라서, 6개의 관측치 중 새롭게 X^*라는 표식을 부여받을 3개의 값을 임의로 선택하면, 나머지 3개의 값은 저절로 Y^*에 속하게 된다. 이처럼 순열의 원리를 이용해 표식을 바꿀 수 있는 모든 경우를 고려하여 재표집한 것이 곧 임의순열 표본에 해당한다.



<그림 3> 임의순열 표본 생성의 예

전체 사례 수가 N이고 k개의 집단 크기가 각각 $n_1, n_2, \cdots, n_k$라고 하면, 이론적으로 생성 가능한 임의순열 표본의 수는 식 (4.1)과 같이 나타낼 수 있다. 각 집단의 크기가 n, m인 두 집단이 있는 경우에는 재표집 가능한 표본의 수를 다음과 같이 표현할 수도 있다(식 (4.2) 참조).

$$\frac{N!}{n_1! n_2! \cdots n_k!} \qquad \text{식 (4.1)}$$

$$\binom{N}{n} = \frac{N!}{n!\,m!} \qquad \text{식 (4.2)}$$

이론에 따르면 앞선 예시에서 구할 수 있는 서로 다른 재표집된 표본은 총 20개일 것이다($\binom{6}{3} = \frac{6!}{3!\,3!}$). 이처럼 생성 가능한 임의순열 표본의 수가 많지 않을 때 연구자들은 모든 가능한 표본을 사용하여 통계적 검정을 하게 되는데, 이를 정확순열 검정(exact permutation test)이라 부를 수 있다. 한편, 3개의 집단이 각 5개의 개별 관측치를 갖는 경우를 가정해보면 무려 756,756개에 달하는 임의순열 표본을 얻을 수 있다($\frac{15!}{5!\,5!5!}$). 과학 기술의 발전으로 컴퓨터의 성능이 뛰어나게 향상되었음에도, 표본의 크기가 커질수록 모든 순열을 만들어 내는 것에는 어려움이 따른다. 이러한 경우에는 가능한 모든 순열이 아니라 그중 충분히 큰 일부만을 무선적으로 선택하여 사용하는 무선화 검정(randomization test)을 대안으로 적용할 수 있다. 이는 모든 가능한 순열을 만들어 내는 데 한 번, 그중 일부를 선택하는 데 또 한 번, 총 두 차례의 무선화 과정을 거치기 때문에 재 무선화 검정(re-randomization test)으로 불리기도 한다. 한편으로 정확순열 검정에

가까운 결과를 낼 수 있어 근사순열 검정(approximate permutation test)이라고도 이야기할 수 있다(Carsey & Harden, 2013; Good, 2005a; Manly, 2006; Rosenbaum, 2002).

무선화 검정을 사용하는 경우에는 연구자가 적절한 수의 표본을 선택해야 한다. 선행연구가 제안한 적절한 표본의 수는 검정이 이루어지는 유의수준과 모형의 복잡도 등에 따라 최소 200개에서 10,000개까지 다양하다(Fitzmaurice, Lipsitz, & Ibrahim, 2007; LaFleur & Greevy, 2009; Manly, 1997). 추출되는 순열의 개수가 많아질수록 연구결과는 정확순열 검정에 가까워지고, 검정력(power) 또한 증가하게 된다. 비록 무선화 검정의 검정력은 정확 순열 검정에 미치지 못하지만 Dwass(1957), Edgington(1980), Manly(1997)의 연구는 무선적으로 선택한 임의순열 표본만으로도 충분히 타당한 결과를 도출할 수 있음을 보였다.

다음은 가상의 데이터를 활용하여 임의순열 검정을 사용한 가설 검정 절차를 보이고자 한다(LaFleur & Greevy, 2009). 자료는 남녀공학 여학생 3명과 여학교 여학생 3명의 자아 인식 점수로 구성되어 있으며, 연구의 목적은 두 집단의 평균 비교를 통해 자아 인식에 차이가 있는지를 확인하고자 함이다. 가설 검정의 첫 번째 단계는 관찰된 표본으로부터 검정 통계량을 구하는 것이다. 남녀공학 학생(X)들의 점수는 10, 11, 12점, 여학교 학생(Y)들의 점수는 13, 14, 28점으로, 집단의 평균은 각각 11.0점과 18.3점이다. 따라서, 검정 통계량인 평균 차이(D)는 7.3에 해당한다. 두 번째로 집단의 수준(표식)을 무작위로 재배정하여 만들 수 있는 모든 재표집 표본으로부터 검정 통계량을 산출한다. 예시의 경우 6개의 관측치를 두 개의 집단으로 분할 할 수 있는 경우의 수는 20이므로, 총 20개의 서로 다른 임의순열 표본을 얻을 수 있다. 생성 가능한 임의순열 표본과 그로부터 얻어진 검정 통계량들은 <표 2>와 같이 나타낼 수 있다.

<표 2> 여학생의 자아 인식 점수로 생성한 임의순열 표본과 검정 통계량 (LaFleur & Greevy, 2009)

순열 번호 (i)	남녀공학 (X)			여학교 (Y)			X 평균 (Mean X)	Y 평균 (Mean Y)	평균 차이 (D)
1	10	11	12	13	14	28	11.0	18.3	7.3
2	10	11	13	12	14	28	11.3	18.0	6.7
3	10	11	14	12	13	28	11.7	17.7	6.0
4	10	12	13	11	14	28	11.7	17.7	6.0
5	10	12	14	11	13	28	12.0	17.3	5.3
6	11	12	13	10	14	28	12.0	17.3	5.3
7	11	12	14	10	13	28	12.3	17.0	4.7
8	10	13	14	11	12	28	12.3	17.0	4.7
9	11	13	14	10	12	28	12.7	16.7	4.0
10	12	13	14	10	11	28	13.0	16.3	3.3
11	10	11	28	12	13	14	16.3	13.0	−3.3
12	10	12	28	11	13	14	16.7	12.7	−4.0
13	10	13	28	11	12	14	17.0	12.3	−4.7
14	11	12	28	10	13	14	17.0	12.3	−4.7
15	11	13	28	10	12	14	17.3	12.0	−5.3
16	10	14	28	11	12	13	17.3	12.0	−5.3
17	12	13	28	10	11	14	17.7	11.7	−6.0
18	11	14	28	10	12	13	17.7	11.7	−6.0
19	12	14	28	10	11	13	18.0	11.3	−6.7
20	13	14	28	10	11	12	18.3	11.0	−7.3

마지막으로 실제 표본에서 얻은 검정 통계량과 무작위로 생성된 가상의 통계량 분포를 비교하여 차이를 확인한다. 산출된 평균 차이 값을 나열하면 영가설 하에 나타날 수 있는 모든 차이의 정확한 분포가 만들어진다. 일방 검정(one−sided test)은 검정 통계량의 분포에서 기존 표본의 검정 통계량

이 위치한 지점을 파악하여, 그 보다 극단적인 값이 나타날 유의확률(significance probability) p를 확인한다. 유의확률을 추정하는 방법은 임의순열 표본 통계량($\hat{\theta}^{(b)}$) 중 관찰된 검정 통계량($\hat{\theta}$)보다 크거나 같은 값의 개수를 전체 임의순열 표본 통계량의 수(i)로 나눠주는 것이다(식 (4.3)). 만약 양방검정(two-sided test)을 하고자 한다면 검정 통계량의 절댓값을 고려하여 유의확률을 구해야 한다.

$$\hat{p} = \frac{\sum_{b=1}^{i} I(\hat{\theta}^{(b)} \geq \hat{\theta})}{i} \qquad \text{식 (4.3)}$$

<표 2>에 따르면 관찰된 검정 통계량 7.3보다 극단적인 값은 존재하지 않는다. 따라서 유의확률은 1/20＝.05로, 유의수준(significance level; α)이 .05일 때 영가설을 기각하게 된다. 결과적으로 두 여학생 집단의 자기 인식 차이를 비교했을 때, 여학교에 다니는 학생의 자기 인식 수준이 남녀공학에 다니는 학생에 비해 유의하게 높다고 할 수 있다. 한편, 모수 검정법인 독립표본 t 검정을 같은 자료에 적용해보면 유의확률은 약 .10으로 동일한 유의수준에서 영가설 기각에 실패하게 된다. 위와 같은 예시를 통해 표본의 크기가 작고, 모집단 분포에 대한 가정을 만족하는지를 확인하기 어려운 경우에는 임의순열 검정과 같은 비모수적 검정이 보다 정확한 결과를 도출함을 확인할 수 있다.

2) 임의순열 검정의 특성

임의순열 검정은 경험적 데이터를 다룰 때 모수 검정법에서 나타나는 전

형적 문제에 대해 강건하게(robust) 대처할 수 있는 통계 검정법이다. 상대적으로 약한 가정을 전제로 하면서도, 모수 검정법이 지닌 검정력과 정확성에 가까운 결과를 도출할 수 있다는 점은 임의순열 검정을 사용해야 하는 강력한 근거로 작용한다(Potvin & Roff, 1993). Good(2000)에 따르면 소표본 자료를 사용할 때 임의순열 검정의 검정력은 불편향 모수 검정법(unbiased parametric tests)만큼 높은 것으로 나타났다. 또한, 정확순열 검정을 사용하는 경우에 유의수준과 정확히 일치하는 1종 오류 비율(type Ⅰ error rate; false rejection rate)을 얻을 수 있다(Dwass, 1957; LaFleur & Greevy, 2009; Manly, 1997). 이는 영가설이 참일 때 나타날 수 있는 모든 평균의 차이가 임의순열 분포에 정확하게 반영되기 때문이다. 만약 1종 오류 비율이 유의수준에서 벗어난다면 교환 가능성 가정이 충족되지 않은 것으로 볼 수 있다. 정확한 1종 오류 비율은 다른 재표집 방법과 차별화되는 임의순열 검정의 고유한 특징이다. 그 외에도 임의순열 검정은 자료상에 존재하는 극단치나 결측치(missing data)로 인한 문제를 완화할 수 있어, 아동·청소년 이상 심리 연구 등에서의 긍정적 기능이 예상된다(LaFleur & Greevy, 2009). 한편, 임의순열 검정의 시행을 위해서 복잡한 코드와 컴퓨터에 대한 지식이 필요하다는 점, 그리고 고전적인 t 검정을 시행할 때와 마찬가지로 등분산성 가정을 필요로 하기 때문에 유사한 문제가 발생할 수 있다는 점이 임의순열 검정의 한계로 작용한다.

표본의 크기가 매우 작은 경우처럼 모수 검정의 기본 가정을 만족하지 못할 때 재표집 방법은 일종의 대안으로써 제시된다. 대부분의 재표집 방법이 비슷한 목적을 가지고 시행되지만, 임의순열 검정과 다른 검정법의 비교를 통해 몇 가지 방법상의 차이를 확인할 수 있다. 먼저 임의순열 검정에서 재표집으로 새로운 표본을 생성하는 과정은 비복원추출(sampling without replacement) 방식으로 이루어진다. 비복원추출을 사용할 경우 하나의 임

의순열 표본을 생성할 때 같은 관측치를 여러 번 사용하지 않아 모든 관측치가 무선적으로 새로운 집단에 할당될 수 있다. 앞서 설명된 잭나이프 역시 비복원추출 방식을 사용한 재표집 방법이라는 점에서 임의순열 검정과 공통점을 갖는다. 그뿐만 아니라 정확순열 검정은 항상 동일한 $\frac{N!}{n_1!n_2!\cdots n_k!}$ 개의 표본을 생성한다는 점에서 잭나이프처럼 매번 같은 결과를 반복하여 얻을 수 있다. 그러나 잭나이프가 급격한 변화가 없는 검정 통계량에 한해 정상적으로 기능하는 반면, 임의순열 검정은 어떤 검정 통계량에든 적용할 수 있어 두 검정법의 실용성 측면에서 차이가 나타난다. 임의순열 검정을 적용하기에 가장 적합한 데이터의 형태는 명확한 처치 변수가 포함된 실험 데이터이다. 실험 데이터는 처치 수준에 따른 집단 간 비교가 이루어지고, 영가설이 곧 처치 효과가 없음을 의미하기 때문에 임의순열 검정의 원리에 대입하는 것이 용이하다.

수행능력의 측면에서는 재표집 방법인 임의순열 검정과 순위를 반영한 비모수 검정인 크루스칼-왈리스(KW) 검정이 비교되기도 한다. 두 검정법을 비교한 선행연구에 따르면, 임의순열 검정이 KW 검정에 비해 높은 검정력을 갖는 것으로 나타났다(Hunter & May, 1993). 추가적으로 Gleason(2013)의 연구는 정규분포, 혹은 균등분포(uniform distribution)를 가정했을 때, 표본 크기와 무관하게 임의순열 검정이 모수 검정법인 분산분석(ANOVA)과 거의 일치하는 높은 수준의 검정력을 가진다고 밝혔다. 두 연구결과를 종합적으로 살펴보면 비모수 조건에서 임의순열 검정을 사용하는 것은 분산분석의 검정력을 회복시킬 수 있는 바람직한 대안이며, 정규성 가정을 만족할 때에도 모수 검정에 못지않은 결과를 도출하는 검정법임을 알 수 있다.

4. 붓스트랩 검정(Bootstrap Test)

1970년대 후반 Efron(1979a)에 의해 등장한 붓스트랩 검정은 다양한 쓰임새를 가지고 광범위하게 활용되고 있는 재표집 방법의 하나다(Carsey & Harden, 2013). 다른 재표집 방법과 마찬가지로 붓스트랩 또한 연구자가 가진 자료가 비모수적이거나, 혹은 추정치의 계산과정이 매우 복잡할 경우에 비교적 쉽게 모수를 추정할 수 있도록 하는 유용한 통계 기법이다. 붓스트랩 검정과 다른 재표집 방법들의 비교에서 가장 두드러지는 차이는 구간 추정(interval estimation)의 결과를 가설 검정에 활용한다는 점에 있다. 일반적으로 표본이 비모수적일 때, 즉 모수의 분포에 대한 정보 없이 모집단을 추정하기 위해서는 히스토그램과 같은 그림으로 모집단 분포의 형태를 예측하거나 표본의 통계량으로 모수의 참값을 추정하는 것이 최선의 방법일 것이다. 그러나 어떤 표본이 선정되었는지에 따라 모수의 참값이 실제와 달리 추정되는 표집오차는 항상 존재할 수밖에 없다. 이러한 오류를 방지하기 위해 하나의 특정한 값을 추정하는 대신, $100(1-\alpha)\%$의 확신을 가지고 모수의 참값이 포함되어 있을 것으로 예상되는 신뢰구간(confidence interval; CI)을 구할 수 있다(Efron, 1982). 이때, 붓스트랩 검정은 주어진 정보 없이 표본의 불확실성(uncertainty)을 신뢰구간의 추정에 반영한다(백재욱, 윤용운, 1995). 정리하자면 붓스트랩 검정이란 강건한 신뢰구간을 얻어 모수를 추정하고, 나아가 가설을 검정하는 일련의 과정이라고 할 수 있다.

1) 붓스트랩 검정의 원리 및 시행절차

붓스트랩 검정의 시행 과정에는 몇 가지 유의해야 할 사항이 있다. 우선 표본을 구성하고 있는 확률변수들이 독립동일분포(i.i.d)를 따른다는 것은

붓스트랩에서도 빠질 수 없는 가정이다. 재표집 과정에서 각 관측치가 표본으로부터 추출될 확률은 동일해야 하며, 매회 추출이 독립적으로 이루어져야 한다. 또한, 붓스트랩 검정은 다른 재표집 방법들과 다르게 복원추출(sampling with replacement)을 허용한다. 표본이 가진 불확실성을 측정하기 위한 변동성(variation)을 만들어내려면 재표집은 반드시 복원추출로 이루어질 필요가 있다. 마지막으로 붓스트랩 재표집으로 생성된 표본의 정보를 충분히 활용하기 위해서는 재표집된 표본의 크기가 관찰된 표본의 크기와 동일하여 완전표본을 이루어야 한다.

붓스트랩 검정은 복원추출을 허용하는 재표집 방법이라는 큰 틀을 유지하면서도, 재표집 대상이 되는 표본이 어떤 분포를 가정하는지에 따라 여러 가지 유형으로 세분화된다. 가장 기본적이고 널리 사용되는 것은 비모수 붓스트랩(non-parametric bootstrap)으로 특정한 확률분포를 가정하지 않고, 표본의 경험적 분포에 의존하는 방법이다(Hesterberg, 2015). 이 방법은 사례(case)를 하나씩 재표집한다는 점에서 사례 재표집(case resampling)이라고도 불린다. 사례 재표집은 구체적으로 2가지 시행 방법이 있는데, 가능한 모든 경우의 붓스트랩 표본을 생성하는 방법(exhaustive method; exact method)과 몬테카를로 표집 방법(Monte Carlo sampling implementation)에 따라 충분히 많은 재표집된 표본을 생성하는 방법이다. 전자의 경우 사례 수가 n일 때, $\binom{2n-1}{n}$개의 서로 다른 붓스트랩 표본을 얻게 되기 때문에 사례 수가 큰 경우 사용하기 어렵다. 반면, 몬테카를로 표집 방법은 연구자가 재표집되는 표본의 수를 결정할 수 있어 보다 현실적이고 실용적이다. 그 외에도 자료에 모수 검정법의 가정을 만족하는 모형을 합치(fit)시켜 재표집하는 모수 붓스트랩(parametric bootstrap), 재표집한 관측치에 일정한 수를 더하여 연속적인 분포를 이루도록 하는 평활 붓스트랩(smooth bootstrap) 등이 있다(Efron, 1979b; Efron 1981; Efron,

1982; Hesterberg, 2015). 몬테카를로 표집 방법을 활용하여 모수를 추정하는 절차에 대한 자세한 설명은 아래와 같다.

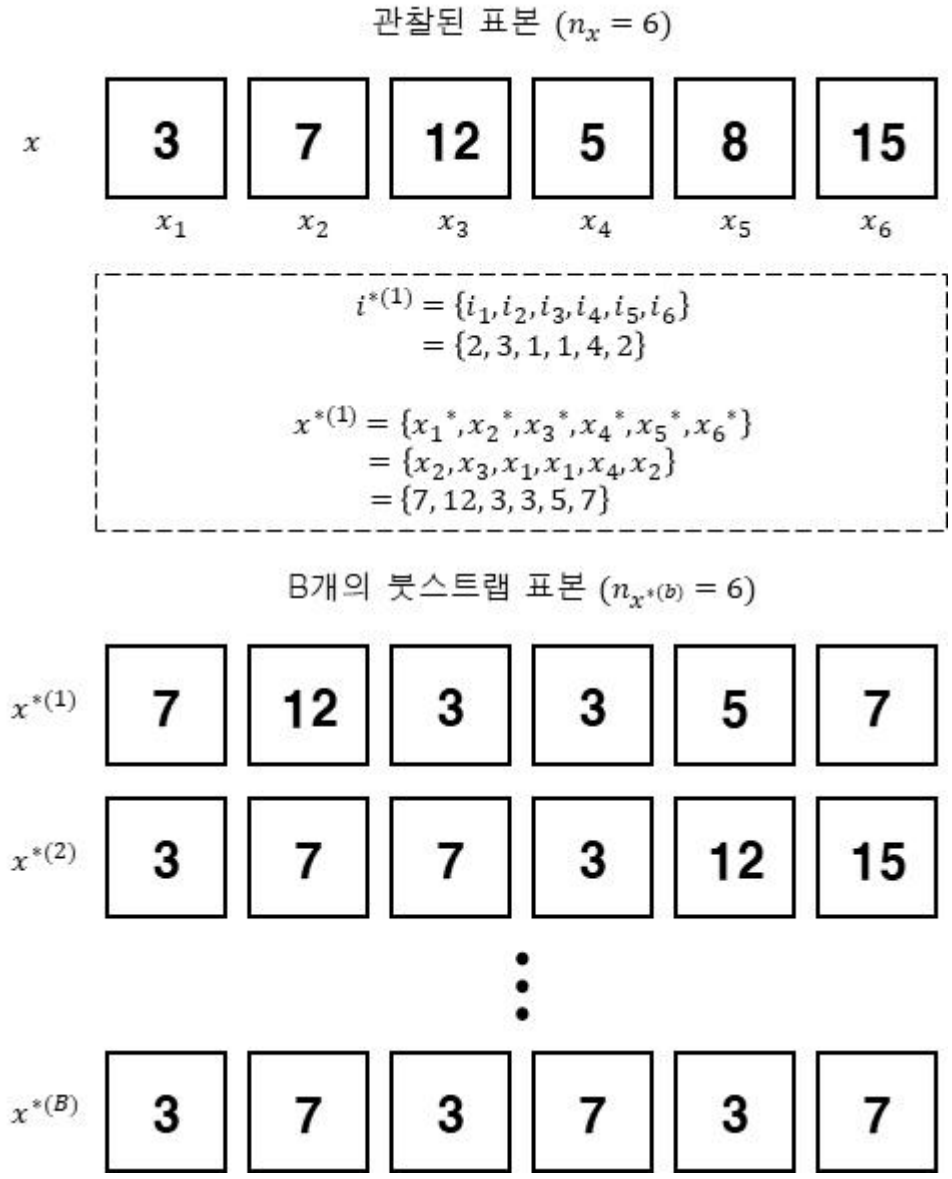
먼저 붓스트랩 표본을 생성하기 위해 n개의 확률변수로 구성된 표본 $x=(x_1, \cdots, x_n)$를 가정한다. 각 관측치는 차례대로 1부터 n까지의 정수를 고유한 번호(x_i)로 가지게 된다. 이때, 주어진 표본에서 하나의 확률변수가 표집될 확률은 n개의 정수 중 하나를 임의로 선택할 확률과 같다.[12] 또한, 두 경우 모두 $\{1,\cdots,n\}$ 사이에서 동일한 균등분포(uniform distribution)를 이룬다. n개의 정수 중 복원추출을 허용하여 n개를 재표집하면 $\{i_1,\cdots,i_n\}$와 같이 정수로 구성된 표본을 만들 수 있다. 이와 같은 정수들의 모음은 표본 x의 실제 관측치 중 어떤 값이 재표집 될지를 결정하는 역할을 한다(식 (5.1)).

$$x_1^* = x_{i_1}, x_2^* = x_{i_2}, \cdots, x_n^* = x_{i_n} \qquad \text{식 (5.1)}$$

이러한 절차를 B회 반복하면 크기가 n인 B개의 붓스트랩 표본을 생성할 수 있다. 붓스트랩 표본 수 B는 연구자의 필요에 따라 달리 설정할 수 있다. 표준오차를 추정하는 데는 보통 50개만으로도 충분하고, 200개를 넘는 경우가 드물다(Efron & Tibshirani, 1993). 반면 신뢰구간의 추정에는 1,000개 정도가 적당하며, 보다 높은 정확성을 위해서 100,000회 이상의 재표집이 권장된다(Hesterberg, 2015). 붓스트랩 표본의 생성은 <그림 4>와 같은 절차로 이루어진다.

---

12) $P(X^*=x_i)=\frac{1}{n}, \quad i=1,\cdots,n$



<그림 4> 붓스트랩 표본 생성의 예

다음으로는 붓스트랩 표본으로부터 모수를 추정하기 위한 붓스트랩 추정치를 산출해야 한다. 관심의 대상이 되는 모수를 θ라고 하면, 표본을 통해 구해지는 모수의 추정치는 $\hat{\theta}$이다. 각 붓스트랩 표본에서 구해진 총 B개의

붓스트랩 추정치 $\hat{\theta}^{(1)},\cdots,\hat{\theta}^{(B)}$는 경험적 붓스트랩 분포를 이룬다. 이 분포를 바탕으로 표준오차와 편향에 대한 정보를 얻을 수 있다. 추정치 $\hat{\theta}$의 붓스트랩 표준오차를 구하는 방법은 식 (5.2)와 같다. 이때, $\overline{\hat{\theta}^{*}}$는 붓스트랩 추정치의 평균 $\overline{\hat{\theta}^{*}}=\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}$ 을 의미한다.

$$\widehat{se}(\hat{\theta}^{*})=\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{(b)}-\overline{\hat{\theta}^{*}})^{2}} \qquad \text{식 (5.2)}$$

마지막으로 붓스트랩 추정치와 표준오차가 구해지면 모수의 신뢰구간을 추정할 수 있다. 신뢰구간은 점추정치 $\hat{\theta}$을 중심으로 일정한 거리만큼 떨어진 구간으로 가장 작은 값은 하한(lower bound), 가장 큰 값을 상한(upper bound)이라고 한다. 붓스트랩 신뢰구간을 추정하는 방법은 중심이 되는 점추정치의 분포와 추정하고자 하는 모수의 종류, 표본 크기 등에 의해 다양하게 나뉜다. 주된 추정법으로는 기본 붓스트랩 신뢰구간(basic bootstrap CI), 표준정규 붓스트랩 신뢰구간(standard normal bootstrap CI), 백분위 붓스트랩 신뢰구간(percentile bootstrap CI), 붓스트랩 t 신뢰구간(bootstrap t CI; studentized bootstrap CI), 가속화 편향교정 신뢰구간(bias−corrected and accelerated CI; BCa CI)이 있다. 그중 백분위 붓스트랩 신뢰구간(percentile bootstrap confidence interval)은 분포의 형태와 무관하게 $\hat{\theta}^{(1)},\cdots,\hat{\theta}^{(B)}$가 이루는 붓스트랩 분포에서 백분위가 $\alpha/2$, $1-\alpha/2$에 해당하는 붓스트랩 추정치를 각각 신뢰구간의 상한, 하한으로 설정하는 방법이다(식 (5.3)).

$$\left(\hat{\theta}^{*}_{\alpha/2}\,,\,\hat{\theta}^{*}_{1-\alpha/2}\right) \qquad \text{식 (5.3)}$$

기본 붓스트랩 신뢰구간, 표준정규 붓스트랩 신뢰구간, 붓스트랩 t 신뢰구간은 점추정치로부터 상한, 하한까지의 거리가 같다는 공통점이 있다. 만약 추정치의 붓스트랩 분포가 정규분포가 아니라면, 대칭을 이루지 않는 분포에 대해 이러한 방법들로 신뢰구간을 구하는 것은 문제가 될 수 있다. 반면 백분위 붓스트랩 신뢰구간은 기준으로부터의 거리와 무관하므로 더 유연하고, 모수의 후보들이 넓은 범위에 걸쳐있을 때도 적절히 신뢰구간을 추정할 수 있다(Carsey & Harden, 2013). 또한, 관찰된 표본분포의 백분위와 추정치의 표집분포의 백분위가 동일하기 때문에 정규분포를 따르지 않을 때는 백분위가 실제 분포를 더 잘 반영한다. 그러므로 백분위를 이용한 추정 방법은 표준정규분포를 사용하는 방법보다 이론적으로 실제 모수에 근접하게 신뢰구간을 추정한다는 장점이 있다(Efron & Tibshirani, 1993). 가설의 검정은 구해진 붓스트랩 신뢰구간이 $0(\theta_0)$을 포함하지 않으면 영가설을 기각하고, 연구가설을 지지하는 것으로 해석할 수 있다.

2) 붓스트랩 검정의 특성

붓스트랩 검정은 분포에 대한 이론적 가정에 의존하지 않고, 현재 소유한 자료와 컴퓨터의 계산능력을 활용하여 불확실성 문제에 대한 답을 얻을 수 있는 통계 기법이다. 그러므로 기본적인 가정이 성립하지 않더라도 통계적 추론 과정에서 발생하는 문제들을 붓스트랩을 이용해 쉽게 풀어낼 수 있다. 표본의 크기가 충분히 크지 않은 경우에도 근사적으로 비교적 정확한 추정치를 구할 수 있다는 점은 붓스트랩 검정의 또 다른 강점에 해당한다(Efron, 1979a). 이처럼 붓스트랩, 잭나이프, 임의순열 등의 재표집 방법들

은 반복적인 표집과 통계량의 산출로 모수를 추정하고자 하는 공통적인 목적을 지닌다. 세 방법이 지닌 유사성으로 인해 붓스트랩 검정과 다른 재표집 방법 간 비교가 지속적으로 이루어져 왔다.

붓스트랩 검정에 대한 아이디어는 잭나이프 검정으로부터 비롯되어 서로 밀접한 관계를 형성하고 있다(Efron, 1982; Rao & Wu, 1988; Sitter, 1992). 붓스트랩 검정은 여러 가지 데이터 조합을 기반으로 주어진 통계량 간의 변동을 측정한다는 점에서 잭나이프 검정의 일반화 된 형태로 볼 수 있다. Efron(1979a)은 실제로 잭나이프 검정이 붓스트랩 검정에 근접하기 위한 선형확장 방법(linear expansion method)임을 보였다. 1개의 관측치를 제거하는 잭나이프 검정법을 사용할 때, 추정된 통계량이 충분히 매끄러운(smooth) 값이라면 잭나이프 추정치는 붓스트랩 추정치에 근사한다(Efron, 1982; Shao & Tu, 2012).

두 방법은 결과적으로는 근사한 추정치를 가지지만, 잭나이프 검정과 붓스트랩 검정의 구조적 차이로 인해 상황에 따라 서로 다른 방법의 사용이 권장된다. 앞서 설명한 잭나이프의 특징과 대조적으로 붓스트랩은 크기가 n인 표본에 대해 $\binom{2n-1}{n}$개 이하의 붓스트랩 표본을 얻을 수 있다. 현실적으로 모든 붓스트랩 표본을 구하는 것에는 한계가 있으므로 가능한 표본 중 일부만을 얻는 과정에서 매번 다른 붓스트랩 추정치가 산출된다. 또한, n개의 잭나이프 표본 통계량만으로는 추정치의 표집분포 전체를 확실히 파악할 수 없지만, 붓스트랩 검정에서는 재표집 횟수가 많을수록 추정치에 대한 정보가 많아져 비교적 정확한 분포를 얻을 수 있다. 정리하자면 잭나이프 검정은 쉽고 간단한 재표집 방법을 사용해야 하는 경우, 혹은 잭나이프 추정치를 반복적으로 검정해야 할 필요가 있는 경우에 사용 가능하다(Shao & Tu, 2012). 반면 붓스트랩 검정은 잭나이프에 비해 직관적이고, 관심 통계치의 종류가 수행에 영향을 미치지 않아서 더욱 쉽고 편리하게 적용할 수

있다. 무엇보다 표본 크기가 충분히 크지 않은 경우에도 근사적으로 정확한 추정치를 제공한다는 선행연구 결과에 따라, 소표본 자료에서 충분히 잘 기능할 것으로 예상할 수 있다(Efron & Tibshirani, 1993).

붓스트랩 검정은 잭나이프뿐만 아니라 임의순열 검정과도 자주 비교된다. 두 검정법 사이에서 확인할 수 있는 단순하면서도 확실한 차이는 재표집된 표본을 추출하는 방법이다. 임의순열 검정이 비복원추출로 표본을 재표집하는 것과 달리, 붓스트랩 검정은 복원추출 방식을 사용한다. 관찰된 표본 내에서 복원추출로 재표집 및 표집분포를 생성하는 붓스트랩 검정의 원리는 모집단에 대한 정보를 가정하고 무작위 복원추출로 표집분포를 만드는 시뮬레이션과 매우 닮아있다. 이러한 사실을 고려했을 때, 붓스트랩 검정은 재표집 방법과 시뮬레이션의 밀접한 관계를 가장 직관적으로 확인할 수 있는 재표집 방법이라는 점에서 사용 가치가 있다.

또 다른 차이점은 두 검정법의 정확성에서 나타난다. 붓스트랩 검정은 표본의 크기나 데이터의 유형에 상관없이 유연하게 적용 가능하고, 강건성(robustness)이 있기 때문에 여러 연구 장면에서 권장되는 재표집 방법이다(Carsey & Harden, 2013). 하지만 붓스트랩 검정이 복원추출을 허용한다는 특성상 동일한 붓스트랩 표본이 반복적으로 사용되어 부정확한 검정결과를 도출하게 될 가능성이 있다. 반면에 임의순열 검정에서는 같은 재표집 표본이 반복 사용될 수 없어 임의순열 표본으로부터 비교적 정확한 정보를 얻을 수 있다. 특히 정확순열 검정의 경우 생성 가능한 모든 재표집 표본을 구하기 때문에 높은 검정력을 가진다. 따라서 정확순열 검정을 적용하기 용이한 소표본 데이터를 사용할 때, 다른 검정법과의 비교에서 임의순열 검정이 가장 우수한 검정력을 가질 것으로 예상할 수 있다(Good, 2005a-b).

Ⅲ. 연구문제 및 가설

연구문제1 : 집단 간 차이가 없는 조건에서 통계 검정법과 표본 크기는 1종 오류 비율에 영향을 미치는가?

[가설 1-1] 세 집단 간 차이가 없는 조건에서 통계 검정법에 따라 1종 오류 비율이 달라질 것이다.

[가설 1-2] 세 집단 간 차이가 없는 조건에서 표본 크기에 따라 1종 오류 비율이 달라질 것이다.

연구문제2 : $n_i = 3$인 소표본 조건에서 통계 검정법과 효과 크기는 검정력에 영향을 미치는가?

[가설 2-1] 세 집단의 표본 크기가 각각 3인 최대 변동성 조건에서 통계 검정법에 따라 검정력이 달라질 것이다.

[가설 2-2] 세 집단의 표본 크기가 각각 3인 최대 변동성 조건에서 효과 크기에 따라 검정력이 달라질 것이다.

[가설 2-3] 세 집단의 표본 크기가 각각 3인 중간 변동성 조건에서 통계 검정법에 따라 검정력이 달라질 것이다.

[가설 2-4] 세 집단의 표본 크기가 각각 3인 중간 변동성 조건에서 효과 크기에 따라 검정력이 달라질 것이다.

연구문제3 : $n_i = 5$인 소표본 조건에서 통계 검정법과 효과 크기는 검정력에 영향을 미치는가?

[가설 3-1] 세 집단의 표본 크기가 각각 5인 최대 변동성 조건에서 통계 검정법에 따라 검정력이 달라질 것이다.

[가설 3-2] 세 집단의 표본 크기가 각각 5인 최대 변동성 조건에서 효과 크기에 따라 검정력이 달라질 것이다.

[가설 3-3] 세 집단의 표본 크기가 각각 5인 중간 변동성 조건에서 통계 검정법에 따라 검정력이 달라질 것이다.

[가설 3-4] 세 집단의 표본 크기가 각각 5인 중간 변동성 조건에서 효과 크기에 따라 검정력이 달라질 것이다.

연구문제4 : $n_i = 10$인 충분한 표본 조건에서 통계 검정법과 효과 크기는 검정력에 영향을 미치는가?

[가설 4-1] 세 집단의 표본 크기가 각각 10인 최대 변동성 조건에서 통계 검정법에 따라 검정력이 달라질 것이다.

[가설 4-2] 세 집단의 표본 크기가 각각 10인 최대 변동성 조건에서 효과 크기에 따라 검정력이 달라질 것이다.

[가설 4-3] 세 집단의 표본 크기가 각각 10인 중간 변동성 조건에서 통계 검정법에 따라 검정력이 달라질 것이다.

[가설 4-4] 세 집단의 표본 크기가 각각 10인 중간 변동성 조건에서 효과 크기에 따라 검정력이 달라질 것이다.

Ⅳ. 연구방법

1. 자료 생성

본 연구에서는 프로그램 R의 몬테카를로 시뮬레이션(Monte Carlo Simulation) 기능을 사용하여 자료를 생성하고, 일원분산분석과 네 가지 비모수적 통계 검정법인 크루스칼-왈리스 검정, 잭나이프 검정, 임의순열 검정, 붓스트랩 검정을 실시하였다. 몬테카를로 시뮬레이션은 불확실한 상황에서 의사 결정을 내리기 위해 확률을 기반으로 하는 모의실험의 일종이다(Stephenson & Holbert, 2003). 연구 목적에 부합하도록 통제된 환경에서 특정 확률 분포를 따르는 표본을 무작위로 생성하고, 이러한 가상의 표본인 모의 표본(simulated sample)에 나타나는 패턴을 탐구하는 것이 몬테카를로 시뮬레이션 연구 절차이다(Mooney, 1997). 연구자는 모의 표본이 표집되는 모집단의 이론적 분포를 모든 측면에서 통제할 수 있으므로 연구 결과를 경쟁 모형이나 통계 추정치의 정밀한 비교에 활용 가능하다(Carsey & Harden, 2013).

잭나이프, 임의순열, 붓스트랩 등의 재표집을 활용한 검정 방법은 복수의 재표집된 표본을 추출하여 모수 추정치에 대한 가정이나 이론을 평가하는 데 사용된다는 점에서 몬테카를로 시뮬레이션과 유사하다. 한편, 연구자에 의해 정의된 이론적 모집단이 아니라 실존하는 데이터로부터의 추출이 이루어지기 때문에 몬테카를로 시뮬레이션과 달리 실제 모집단 분포를 알 수 없다는 차이가 있다. 그러나 관찰된 표본의 관측치 각각을 실제보다 더 큰 모집단에서 무작위로 표집된 확률표본이라고 생각해보면, 관찰된 표본으로부터 재표집된 표본은 실제 모집단에 근사한 분포를 그리게 된다. 따라서 재표집 방법은 몬테카를로 시뮬레이션의 연장선 상에서 논의될 수 있다

(Chernick & LaBudde, 2011; Efron & Tibshirani, 1993; Good, 2005b; Mooney & Duval, 1993).

1) 조작변수(manipulated variables)

본 연구에서는 세 집단의 평균 비교 모형을 모집단으로 설정하고, 세 가지 조작변수에 의해 표본 데이터 생성을 결정하였다. 모집단 분포는 $N(\mu_i, 1)$로 평균이 μ_i이고, 표준편차가 1인 정규분포를 따르는 것으로 가정하였다. 세부적인 조건은 각 집단의 크기를 나타내는 표본 크기(sample size)가 3, 5, 10으로 세 수준, 집단 간 평균의 차이인 효과 크기(effect size)가 0.2, 0.5, 0.8, 1.2, 2.0으로 다섯 수준, 평균의 변동성(mean variability)이 최대(maximum), 중간(intermediate) 두 수준으로 $3\times5\times2$의 요인설계(factorial design)를 바탕으로 자료를 생성하였다. 추가로 영가설 하에서 통계 검정법의 수행을 비교할 수 있도록 효과 크기가 0인 경우(null condition)를 고려하였다. 집단 간 차이가 없어 변동성을 고려할 수 없으므로 표본 크기 수준에 해당하는 3가지 조건을 더한 33개의 조건을 형성하였다. 33개의 조건을 각 10,000번씩 반복(replication)하여 총 330,000개의 자료를 생성하였다.

첫 번째 조작변수인 표본 크기는 소표본(small sample) 조건을 형성하기 위해 세 집단 사례 수의 합이 충분히 큰 표본 크기인 30을 넘지 않도록 하였다. 집단 간 차이를 검정하는 연구에서 충분한 표본 크기를 판가름하는 기준은 한 표본의 크기가 30 이상인 경우(Cohen, 1988; Kreft & De Leeuw, 1998; Maas & Hox, 2005; Sawilowsky & Blair, 1992), 혹은 표본의 크기가 20 이상인 경우(Cooper & Berwick, 2001; Hedges & Olkin, 1985; Nachar, 2008) 등으로 견해의 대립이 있다. 세 집단이 동일

한 모집단으로부터 표집되었을 것이라는 영가설에 초점을 맞추고, 무수히 많은 집단을 동시에 비교하는 것에 현실적 어려움이 있다는 사실을 고려했을 때, 본 연구에서는 전체 사례 수가 30 이상이면 충분히 큰 표본인 것으로 가정하였다. 각 집단의 크기가 같은 경우 분산의 동일성 문제가 추정치에 미치는 영향을 무시할 수 있다는 연구결과에 따라(Carroll & Nordholm, 1975), 세 집단은 모두 동일한 크기를 가지도록 설정하였다. 따라서 표본 크기는 3:3:3, 5:5:5의 소표본 조건과 10:10:10의 충분한 크기를 가진 표본으로 구성하여, 표본 크기에 따른 검정법의 수행을 비교하였다.

두 번째 조작변수인 효과 크기는 0.2, 0.5, 0.8, 1.2, 2.0의 다섯 가지 수준으로 조작하였다. 효과 크기란 집단 사이의 차이나 관계를 나타내는 표준화된 지표로, 평균을 비교하는 경우 집단 간 평균 차이가 클수록 효과 크기는 커진다. 실제 관찰된 데이터를 다루는 연구에서는 수집된 자료가 이미 특정한 효과크기를 지니고 있어 연구자가 이를 임의로 변경할 수 없다. 하지만 시뮬레이션 연구에서는 연구자가 집단 간 차이의 정도를 사전에 지정하여 자료를 생성하고, 자료를 분석한 결과 의도한 수준만큼의 차이가 발견되는지를 확인함으로써 수행을 평가하기 때문에 효과크기는 빼놓을 수 없는 조작변수 중 하나이다(Gleason, 2013; Wright, 2006). 여러 가지 효과 크기의 종류 중 Cohen의 d(Cohen’s d)는 가장 큰 평균과 가장 작은 평균의 차이, 즉 평균의 범위(range)를 표준화한 값이다(식 (6.1) 참조).

$$d = \frac{m_{\max} - m_{\min}}{\sigma} \qquad \text{식 (6.1)}$$

Cohen(1988)은 효과 크기를 직관적으로 이해하기 쉽도록 0.2를 작은(small) 효과, 0.5를 중간(medium) 효과, 0.8을 큰(large) 효과로 구분하

였다. Sawilowsky(2009)는 이에 매우 작은(very small) 효과를 0.01, 매우 큰(very large) 효과를 1.2, 막대한(huge) 효과를 2.0으로 하는 상세한 규준을 제안하였다. 효과 크기가 작을수록 검정력 또한 낮아져 통계 검정법에 따른 차이를 비교하는 데 어려움이 있으므로, 매우 작은 효과 크기인 0.01을 제외한 다섯 가지 수준만을 본 연구에 사용하였다.

세 번째 조작변수인 평균의 변동성은 효과 크기와 함께 여러 집단 사이의 관계를 임의로 조성하기 위해 사용되는 변수로, $k(k \geq 2)$개의 집단이 있을 때 특정 범위 내에서 평균이 흩어져있는 형태를 나타낸다. 변동성은 최소(minimum), 중간(intermediate), 최대(maximum)의 세 가지 수준으로 나뉜다. 효과 크기가 d가 되도록 하는 2개의 평균값을 양극단으로 하고, 그 사이에 나머지 $k-2$개가 놓이면 최소 변동성 조건이 형성된다. 중간 변동성은 k개의 평균이 동일한 간격을 두고 흩어져있는 경우, 최대 변동성은 하나의 평균값에 치우쳐 있는 경우를 의미한다. 시뮬레이션 연구에서 평균 변동성은 집단의 처치 여부를 할당하는 역할을 하기도 한다. 예를 들어, $d=0.8$이고 $k=3$이면 최소 변동성은 $(-0.40, 0.00, 0.40)$, 중간 변동성은 $(0.00, 0.40, 0.80)$, 최대 변동성은 $(0.00, 0.00, 0.80)$과 같은 평균 배치를 통해 나타날 수 있다. 이 때, (0.00)을 통제 집단의 평균이라고 가정하면 (-0.40), (0.40), (0.80)과 같은 값을 평균으로 하는 집단은 처치가 가해져 평균이 달라진 집단이라고 볼 수 있다. 즉, 평균 변동성은 가상의 자료를 활용하는 연구를 시행하는 경우에 보다 현실적인 데이터를 생성하도록 하는 도구에 해당한다(Keselman, 1975; Okada, 2013). 집단이 3개일 때 중간 변동성과 최소 변동성 조건에서의 검정력은 동일하기 때문에, 본 연구에서는 최소 변동성을 제외한 최대, 중간 변동성만을 사용하였다. 효과 크기와 평균 변동성 조건에 따라 세 집단의 평균을 <표 3>과 같이 설정할 수 있다.

<표 3> 효과 크기와 평균 변동성에 따른 세 집단의 평균

효과 크기 (d)	평균 변동성(mean variability)					
	중간 (intermediate)			최대 (maximum)		
	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3
작은 (0.2)	0.00	0.10	0.20	0.00	0.00	0.20
중간 (0.5)	0.00	0.25	0.50	0.00	0.00	0.50
큰 (0.8)	0.00	0.40	0.80	0.00	0.00	0.80
매우 큰 (1.2)	0.00	0.60	1.20	0.00	0.00	1.20
막대한 (2.0)	0.00	1.00	2.00	0.00	0.00	2.00

주. 최대 변동성은 세 집단의 평균이 특정 범위 내에서 가장 많이 흩어져 있는 조건, 중간 변동성은 중간 정도로 흩어진 조건을 의미함.

2) 1종 오류 비율(type Ⅰ error rates)

연구자가 가설 검정 시 범할 수 있는 판단의 착오로는 영가설이 참일 때 영가설을 기각하는 오류, 그리고 영가설이 참이 아닐 때 영가설 기각에 실패하는 오류 두 가지가 있다(성태제, 2014). 이 중 전자를 1종 오류(type Ⅰ error)라고 부르고, 이러한 오류를 허용하는 확률을 α, 즉 유의수준(significant level)이라 한다. 시뮬레이션 연구에서는 집단 간 차이가 없는 조건(null condition)에서 검정의 결과가 유의한 것으로 나타날 때 1종 오류를 범했다고 할 수 있다. 정리하자면 1종 오류 비율은 전체 반복(replication) 횟수 중 1종 오류가 발생하는 횟수가 자치하는 비율이다(Gleason, 2013; Mendeş & Akkartal, 2010).

본 연구에서는 통계 검정법의 수행을 평가하는 기준의 하나로 1종 오류 비율을 사용하였다. 1종 오류 비율은 각 시뮬레이션 시행에서 검정 결과가 유의할 경우 1, 그렇지 않으면 0으로 코딩하고 전체 시행 결과의 합을 시뮬레이션 반복횟수인 10,000으로 나눈 값과 같다. 산출된 1종 오류 비율을 일종의 절대적 기준에 해당하는 통계적 유의수준 $\alpha = .05$와 비교하여, 1종 오류 비율의 수용 가능성을 평가하였다. 또한, 각 통계 검정법의 1종 오류 비율을 상대적으로 비교하여 유의한 차이가 있는지를 검정하고, 각 조작 조건에서 수행능력이 뛰어난 검정법을 확인하였다.

3) 검정력(power)

검정력은 영가설이 참이 아닐 때 영가설을 기각할 확률로, 1종 오류 비율과 함께 통계 검정법의 수행을 평가하는 또 다른 기준이다. 시뮬레이션 연구에서의 검정력은 집단 간 차이가 있는 조건에서 전체 반복횟수 중 유의한 검정 결과가 나오는 횟수가 차지하는 비율과 같다(Gleason, 2013; Mendeş & Akkartal, 2010). 검정력의 수용 가능성 평가는 이론적 검정력인 기대된 검정력(expected power)과의 비교를 통해 이루어진다. 기대된 검정력은 이론적으로 산출되어야 마땅한 검정력이라고 볼 수 있으며, 표본 크기, 효과 크기, 평균 변동성과 같은 조건에 따라 달라진다. 유의수준 $\alpha = .05$에서 기대된 검정력은 <표 4>와 같이 정리할 수 있다.

<표 4> 평균 변동성, 표본 크기, 효과 크기에 따라 기대되는 검정력 ($\alpha = .05$)

평균 변동성	표본 크기	효과 크기 (d)				
		0.2	0.5	0.8	1.2	2.0
최대	3	0.05381	0.07442	0.11497	0.20401	0.48578
	5	0.05796	0.10255	0.19413	0.39098	0.82602
	10	0.06842	0.17768	0.39863	0.74921	0.99490
중간	3	0.05285	0.06818	0.09802	0.16327	0.38089
	5	0.05595	0.08883	0.15556	0.30305	0.70151
	10	0.06372	0.14347	0.30853	0.61630	0.97326

본 연구에서는 검정 결과의 유의성에 따라 0 혹은 1로 코딩한 각 시뮬레이션 시행 결과를 모두 합한 값을 전체 반복횟수인 10,000으로 나눈 것을 검정력으로 하였다. 연구를 통해 구해진 검정력을 기대된 검정력과 비교하여 절대적인 수행 능력을 평가하였다. 더불어, 통계 검정법 간 검정력에 차이가 있는지, 그리고 각 조건에서 가장 우수한 수행을 보이는 검정법이 무엇인지를 확인하였다.

2. 자료 분석

연구문제1을 확인하기 위해 집단 간 차이가 없다고 가정한 조건(null condition)으로부터 생성된 자료를 분석하여 1종 오류 비율을 산출하였다. 산출된 결과를 토대로 각 검정법과 표본 크기에 따른 1종 오류 비율의 평균을 비교하고, 이를 통해 통계 검정법과 표본 크기가 1종 오류 비율에 미치는 영향을 확인하였다. 연구문제2, 연구문제3, 연구문제4는 집단 간 차이를 설정한 자료에 대해 위와 동일한 절차를 통해 검정력을 산출하였다. 이후, 표본 크기와 평균 변동성이 주어졌을 때 검정법의 종류와 효과 크기가 검정력에 미치는 영향을 확인하였다.

Ⅴ. 연구결과

1. 집단 간 차이가 없는 조건에서의 1종 오류 비율

모수 검정법과 비모수적 검정법을 실시한 결과, 통계 검정법과 조작변수에 의해 형성된 조건에 따라 1종 오류 비율이 달라지는지 검정해보고자 하였다. R의 몬테카를로 시뮬레이션 기능을 사용하여 자료를 생성하고, 생성된 자료를 대상으로 일원분산분석, 크루스칼-왈리스 검정, 잭나이프 검정, 임의순열 검정, 붓스트랩 검정을 실시하였다.

효과 크기가 0이고 평균 변동성이 없는 null condition 조건(이하 집단 차이가 없는 조건)에서 조작변수의 수준에 따른 1종 오류 비율을 살펴보기 위해, 각 집단의 표본 크기를 3, 5, 10으로 조작하였다. 설정된 3개의 조건은 각 10,000번씩 반복분석 되었다. 그 결과 산출된 각 조건에서의 집단 간 차이 검정 결과를 바탕으로 1종 오류 비율을 분석하였다.

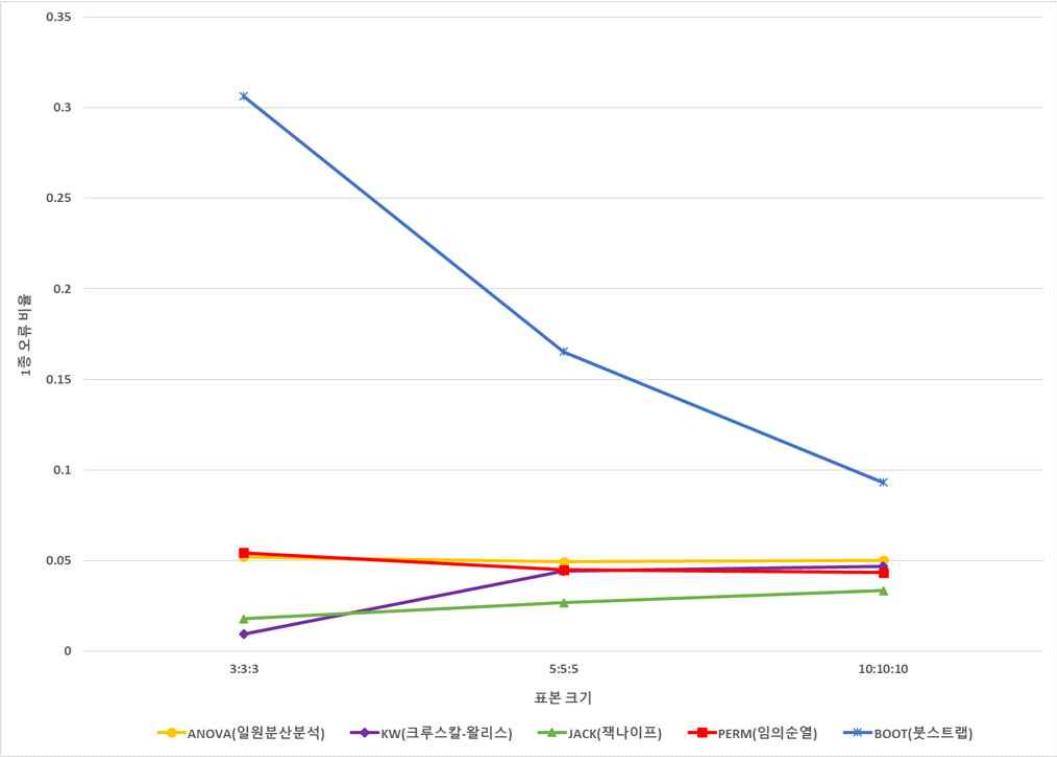
1) 통계 검정법과 표본 크기에 따른 1종 오류 비율

검정법의 종류와 표본 크기에 따른 1종 오류 비율을 <표 5>과 <그림 5>에 제시하였다. 그 결과, 통계 검정법에 따라서는 평균적으로 잭나이프 검정의 1종 오류 비율이 가장 낮았으며, 붓스트랩 검정의 1종 오류 비율이 가장 컸다. 구체적으로 일원분산분석을 제외하고 KW 검정과 잭나이프 검정에서 표본 크기가 증가함에 따라 1종 오류 비율이 증가하는 경향이, 임의순열 검정과 붓스트랩 검정에서는 감소하는 경향이 나타났다. 표본 크기의 경우, 표본이 커질수록 1종 오류 비율은 본 연구에서 설정한 유의수준인 $\alpha = .05$에 가까워지는 것으로 관찰되었다.

<표 5> 통계 검정법과 표본 크기에 따른 1종 오류 비율 ($\alpha = .05$)

검정법 / 표본 크기	모수 검정	비모수 검정			
	ANOVA	KW	JACK	PERM	BOOT
$n_1 = n_2 = n_3 = 3$	0.0521	0.0095	0.0179	0.0544	0.3062
$n_1 = n_2 = n_3 = 5$	0.0494	0.0441	0.0267	0.0448	0.1653
$n_1 = n_2 = n_3 = 10$	0.0502	0.0469	0.0334	0.0435	0.0931

주. ANOVA=일원분산분석, KW=크루스칼-왈리스 검정, JACK=잭나이프 검정, PERM=임의순열 검정, BOOT=붓스트랩 검정



<그림 5> 통계 검정법과 표본 크기가 1종 오류 비율에 미치는 영향

2. 집단 간 차이가 있는 조건에서의 검정력

모수 검정법과 비모수적 검정법을 실시한 결과, 통계 검정법과 조작변수에 의해 형성된 조건에 따라 검정력이 달라지는지를 검정해보고자 하였다. R의 몬테카를로 시뮬레이션 기능을 사용하여 자료를 생성하고, 생성된 자료를 대상으로 일원분산분석, 크루스칼-왈리스 검정, 잭나이프 검정, 임의순열 검정, 붓스트랩 검정을 실시하였다.

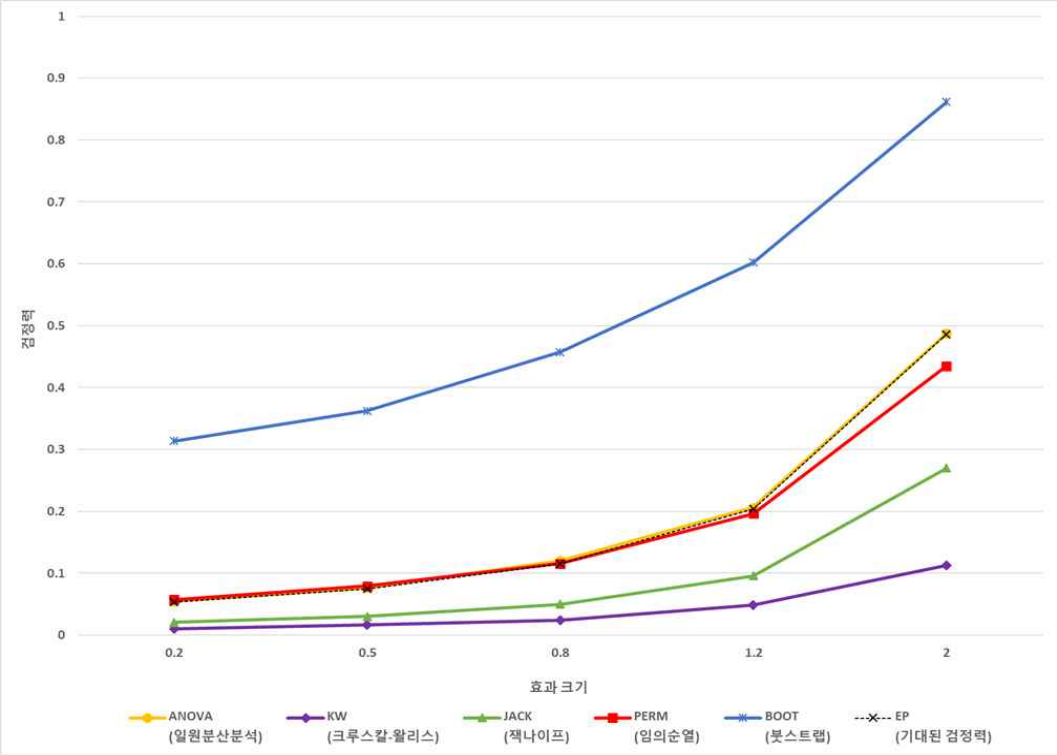
조작변수의 수준에 따른 검정력의 변화를 살펴보기 위해, 표본 크기는 3, 5, 10으로, 효과 크기는 0.2, 0.5, 0.8, 1.2, 2.0으로, 평균 변동성은 최대, 중간으로 조작하였다. 즉, 3(표본 크기)×5(효과 크기)×2(평균 변동성) 요인설계로 모두 30개의 조건으로 구성되어 있으며, 설정된 각 조건은 10,000번씩 반복되었다. 그 결과 산출된 각 조건에서의 집단 간 차이 검정 결과를 바탕으로 검정력을 분석하였다.

세 집단의 차이 검정 시 각 조건에서 이론적으로 산출 가능한 검정력을 구하기 위해 G*Power 3.1.9.4 프로그램을 이용하였다. 집단 수가 3개이고 유의수준이 $\alpha = .05$인 경우, 각 효과 크기와 평균 변동성 조건을 적용한 G*Power 분석결과로 제시되는 '성취된 검정력(achieved power)'을 기대된 검정력(expected power, EP)의 근거로 하였다.

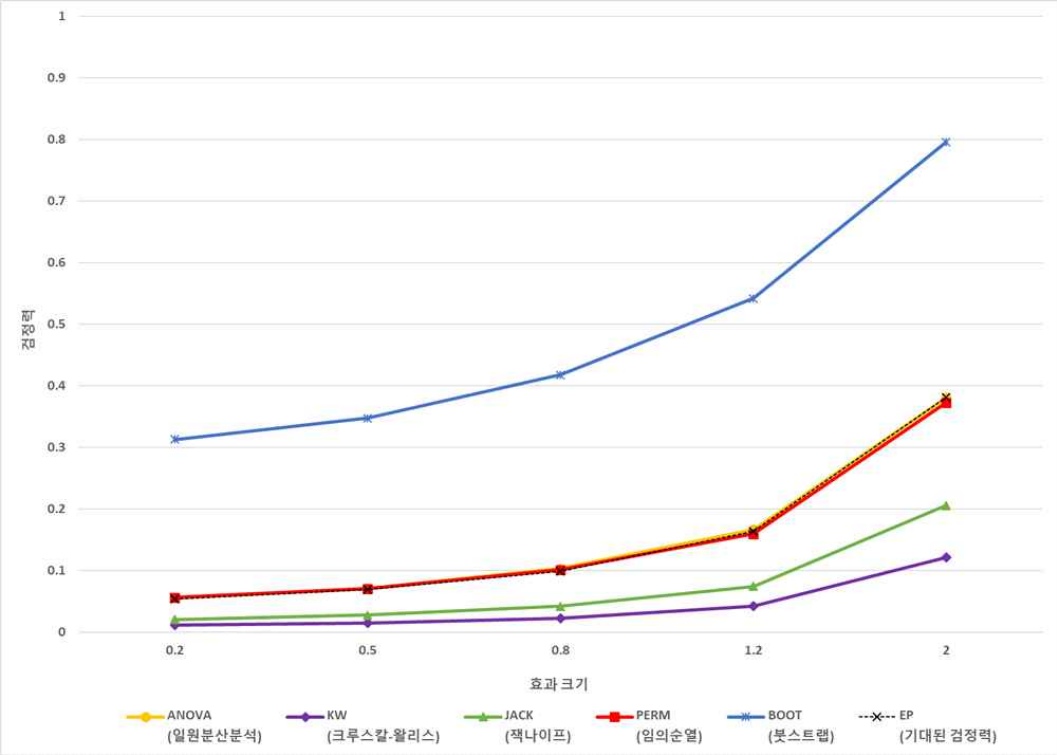
1) $n_i = 3$일 때, 통계 검정법과 효과 크기에 따른 검정력

세 집단의 표본 크기가 각각 3인 조건에서, 범위 내 평균의 흩어진 정도를 최대로 했을 때와 중간 수준으로 했을 때 통계 검정법과 효과 크기에 따른 검정력을 <그림 6>과 <그림 7>에 제시하였다. 최대 변동성 조건에서는 붓스트랩 검정이 평균적으로 가장 큰 검정력, KW 검정이 가장 작은 검정력

을 가졌으며, 이는 모든 효과 크기 조건에서 동일했다. 효과 크기가 커질수록 검정력이 증가하는 경향은 검정법의 종류와 무관하게 나타났고, 중간 변동성 조건에서도 검정법과 효과 크기에 따른 영향력이 일관적으로 관찰되었다.



<그림 6> $n_1 = n_2 = n_3 = 3$인 최대 변동성 조건에서
통계 검정법과 효과 크기에 따른 검정력



<그림 7> $n_1 = n_2 = n_3 = 3$인 중간 변동성 조건에서
통계 검정법과 효과 크기에 따른 검정력

구체적으로 평균 변동성이 최대일 때, 붓스트랩 검정의 검정력은 효과 크기가 증가함에 따라 검정력이 0.5477만큼 증가하여 가장 많은 변화가 있는 것으로 나타났다. 일원분산분석의 검정력은 0.4326, 임의순열 검정의 검정력은 0.3776만큼 변화하여 붓스트랩 검정에 비해서는 검정력 증가의 폭이 좁았다. 잭나이프 검정의 검정력은 효과 크기 증가에 따라 0.2497만큼 변화하였고, KW 검정의 경우 0.1024로 검정력 변화량이 가장 작은 것을 확인할 수 있었다. 통계 검정법 간 검정력의 차이는 효과 크기가 2.0일 때 가장 크게 나타났다. 기대된 검정력과 연구를 통해 얻은 경험적 검정력의 비교에서 붓스트랩 검정은 항상 기대 이상의 수행을 보였다. 일원분산분석과 임의순열 검정의 검정력은 기대된 검정력과 대체로 일치했고, 나머지 검정력은 어떤 조건에서도 기대된 수준에 미치지 못했다(<표 6> 참조).

<표 6> $n_1 = n_2 = n_3 = 3$인 최대 변동성 조건에서 통계 검정법과 효과 크기에 따른 검정력 ($\alpha = .05$)

검정법 / 효과 크기	모수 검정	비모수 검정				기대된 검정력
	ANOVA	KW	JACK	PERM	BOOT	(EP)
0.2	0.0551	0.0105	0.0205	0.0575	0.3136	0.05381
0.5	0.0769	0.0165	0.0304	0.0792	0.3621	0.07442
0.8	0.1203	0.0246	0.0501	0.1157	0.4573	0.11497
1.2	0.2063	0.0487	0.0961	0.1963	0.6019	0.20401
2.0	0.4877	0.1129	0.2702	0.4351	0.8613	0.48578

주. ANOVA=일원분산분석, KW=크루스칼-왈리스 검정, JACK=잭나이프 검정, PERM=임의순열 검정, BOOT=붓스트랩 검정

평균 변동성이 중간 수준일 때, 효과 크기가 가장 작은 경우(0.2)를 제외하고 항상 최대 변동성 조건에서보다 낮은 검정력이 관찰되었다. 그러나 검정법의 종류나 효과 크기에 따른 검정 결과의 패턴은 유사하게 나타났다. 보다 구체적으로 효과 크기가 커질수록 검정력이 증가한 정도는 붓스트랩 검정이 0.4828로 가장 컸다. 일원분산분석은 0.3256, 임의순열 검정은 0.3168로 유사한 수준이었으며, 잭나이프 검정에서는 0.1851만큼의 변화가 관찰되었다. KW 검정의 검정력은 0.1104만큼 증가하여 가장 작은 변화가 있는 것으로 나타났다. 통계 검정법 간 검정력의 차이가 가장 극명하게 나타나는 것은 효과 크기가 2.0인 조건으로, 붓스트랩 검정과 KW 검정 사이에 0.6745의 검정력 차이가 있었다. 기대된 검정력과 실제 검정력을 비교한 결과, 붓스트랩 검정은 항상 기대보다 높은 수행을 보였다. 일원분산분석과 임의순열 검정의 검정력은 기대된 검정력과 대체로 일치했고, 나머지 검정력은 어떤 조건에서도 기대된 수준에 미치지 못했다(<표 7> 참조).

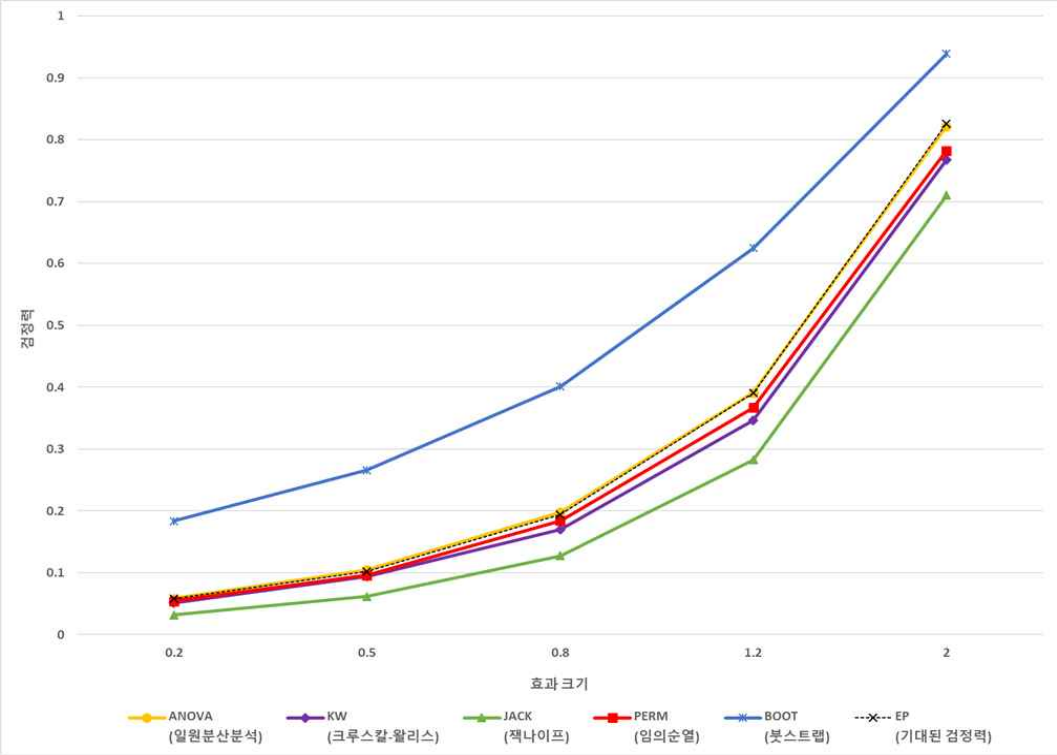
<표 7> $n_1 = n_2 = n_3 = 3$인 중간 변동성 조건에서 통계 검정법과 효과 크기에 따른 검정력 ($\alpha = .05$)

검정법 / 효과 크기	모수 검정	비모수 검정				기대된 검정력
	ANOVA	KW	JACK	PERM	BOOT	(EP)
0.2	0.0552	0.0106	0.0200	0.0558	0.3127	0.05285
0.5	0.0704	0.0146	0.0272	0.0701	0.3470	0.06818
0.8	0.1030	0.0225	0.0416	0.1010	0.4173	0.09802
1.2	0.1664	0.0414	0.0740	0.1589	0.5415	0.16327
2.0	0.3808	0.1210	0.2051	0.3726	0.7955	0.38089

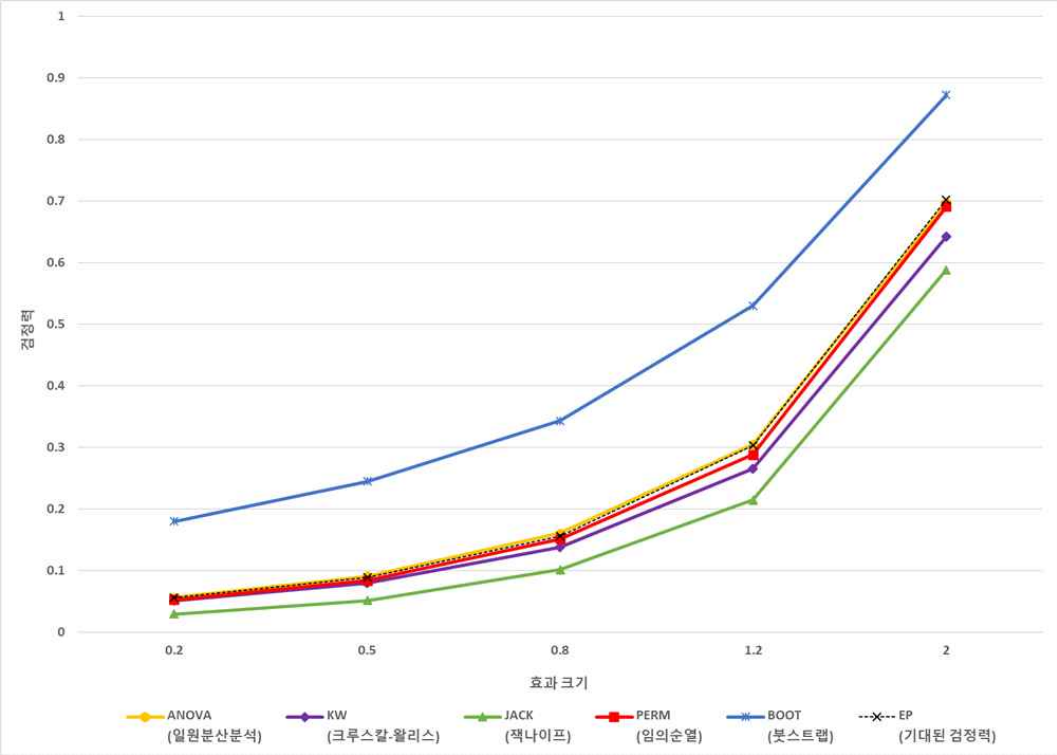
주. ANOVA=일원분산분석, KW=크루스칼-왈리스 검정, JACK=잭나이프 검정, PERM=임의순열 검정, BOOT=붓스트랩 검정

2) $n_i = 5$일 때, 통계 검정법과 효과 크기에 따른 검정력

세 집단의 표본 크기가 각각 5인 조건에서, 평균 변동성을 최대로 했을 때와 중간으로 했을 때 통계 검정법과 효과 크기에 따른 검정력을 <그림 8>과 <그림 9>에 제시하였다. 최대 변동성이 주어졌을 때, 붓스트랩 검정이 가장 큰 검정력, 잭나이프 검정이 가장 작은 검정력을 가지는 것으로 나타났다. 이와 같은 결과는 각 검정법의 검정력 평균에서는 물론, 모든 효과 크기 조건에서 동일하게 관찰되었다. 또한, 효과 크기가 커질수록 검정력이 증가하는 추세는 검정법의 종류를 불문하고 유사하게 나타났다. 통계 검정법과 효과 크기에 따른 검정력의 변화는 중간 변동성 조건에서도 일관적으로 관찰되었다.



<그림 8> $n_1 = n_2 = n_3 = 5$인 최대 변동성 조건에서
통계 검정법과 효과 크기에 따른 검정력



<그림 9> $n_1 = n_2 = n_3 = 5$인 중간 변동성 조건에서
통계 검정법과 효과 크기에 따른 검정력

구체적으로 평균 변동성이 최대일 때, 효과 크기 증가에 따른 검정력의 변화 정도는 일원분산분석이 0.7635로 가장 컸다. 이어서 붓스트랩 검정은 0.7554, 임의순열 검정은 0.7283, KW 검정은 0.7160만큼의 변화가 관찰되었으며, 잭나이프 검정의 검정력은 0.6785만큼 증가하여 가장 작은 변화가 있는 것으로 나타났다. 통계 검정법 간 검정력의 차이가 가장 극명하게 나타나는 것은 효과 크기가 1.2인 조건으로, 붓스트랩 검정과 잭나이프 검정 사이에 0.3425의 검정력 차이가 있었다. 기대된 검정력과 실제 검정력을 비교한 결과, 붓스트랩 검정은 항상 기대보다 높은 수행을 보였다. 일원분산분석과 임의순열 검정의 검정력은 기대된 검정력과 거의 일치했으나, 이외에는 어떤 조건에서도 기대된 수준보다 검정력이 낮았다(<표 8> 참조).

<표 8> $n_1 = n_2 = n_3 = 5$인 최대 변동성 조건에서 통계 검정법과 효과 크기에 따른 검정력 ($\alpha = .05$)

검정법 / 효과 크기	모수 검정	비모수 검정				기대된 검정력
	ANOVA	KW	JACK	PERM	BOOT	(EP)
0.2	0.0585	0.0512	0.0317	0.0542	0.1835	0.05796
0.5	0.1050	0.0942	0.0618	0.0957	0.2658	0.10255
0.8	0.1978	0.1696	0.1276	0.1841	0.4013	0.19413
1.2	0.3908	0.3463	0.2823	0.3668	0.6248	0.39098
2.0	0.8220	0.7672	0.7102	0.7825	0.9389	0.82602

주. ANOVA=일원분산분석, KW=크루스칼-왈리스 검정, JACK=잭나이프 검정, PERM=임의순열 검정, BOOT=붓스트랩 검정

평균 변동성이 중간 수준인 경우, 같은 효과 크기 조건에서 최대 변동성을 가정했을 때보다 항상 낮은 검정력이 관찰되었다. 검정법의 종류나 효과 크기에 따른 검정 결과의 패턴은 두 변동성 조건에서 거의 동일했다. 보다 구체적으로 효과 크기 증가에 따른 검정력의 변화량은 붓스트랩 검정이 0.6926으로 가장 컸다. 일원분산분석은 0.6416, 임의순열 검정은 0.6386, KW 검정은 0.5922만큼의 변화가 관찰되었으며, 잭나이프 검정의 검정력은 0.5588만큼 증가하여 가장 작은 변화가 있는 것으로 나타났다. 통계 검정법 간 검정력의 차이가 가장 뚜렷하게 드러나는 것은 효과 크기가 1.2인 조건으로, 붓스트랩 검정과 잭나이프 검정 사이에 0.3158의 검정력 차이가 있었다. 기대된 검정력과 비교한 결과, 붓스트랩 검정은 항상 기대보다 높은 수행을 보였다. 일원분산분석과 임의순열 검정의 검정력은 기대된 검정력과 거의 일치했고, 그 외 검정법에서는 항상 기대된 수준보다 낮았다(<표 9> 참조).

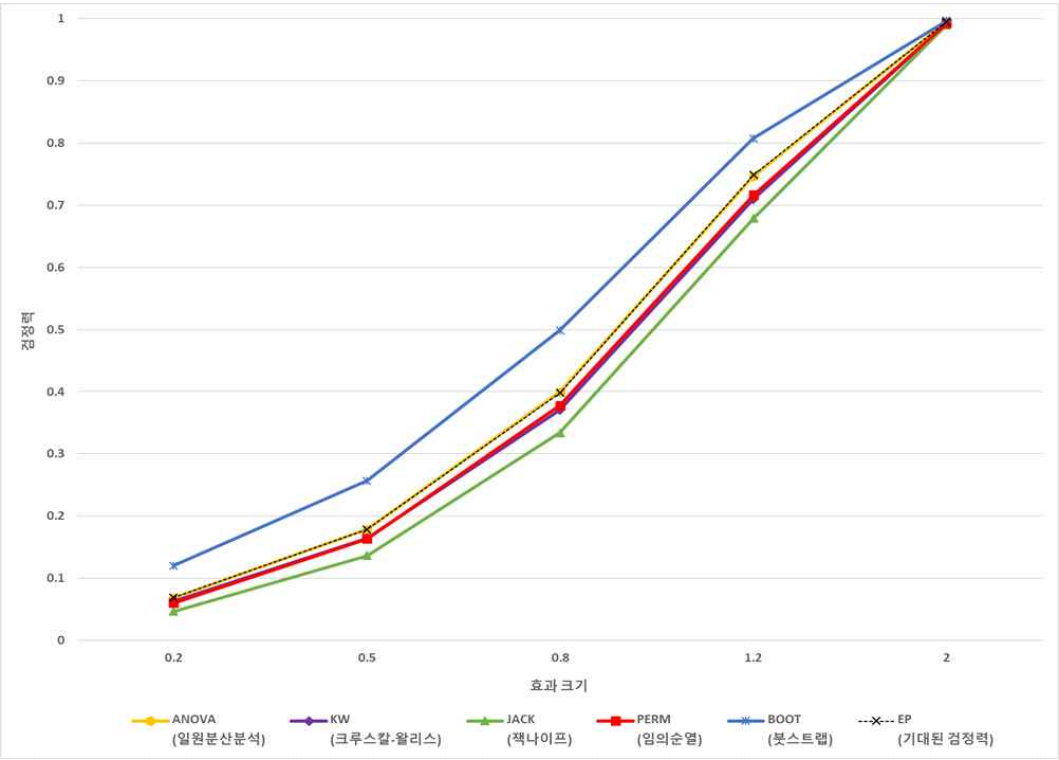
<표 9> $n_1 = n_2 = n_3 = 5$인 중간 변동성 조건에서 통계 검정법과 효과 크기에 따른 검정력 ($\alpha = .05$)

검정법 / 효과 크기	모수 검정	비모수 검정				기대된 검정력
	ANOVA	KW	JACK	PERM	BOOT	(EP)
0.2	0.0558	0.0503	0.0291	0.0523	0.1794	0.05595
0.5	0.0904	0.0790	0.0509	0.0835	0.2447	0.08883
0.8	0.1603	0.1376	0.1011	0.1509	0.3431	0.15556
1.2	0.3045	0.2651	0.2141	0.2879	0.5299	0.30305
2.0	0.6974	0.6425	0.5879	0.6909	0.8720	0.70151

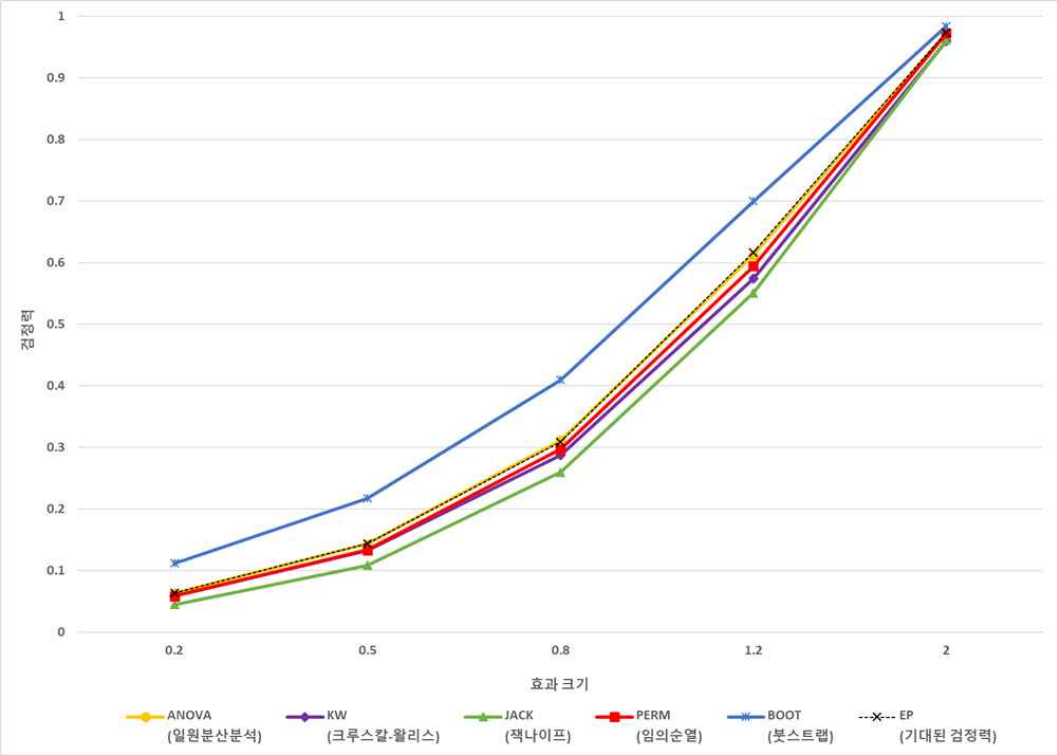
주. ANOVA=일원분산분석, KW=크루스칼-왈리스 검정, JACK=잭나이프 검정, PERM=임의순열 검정, BOOT=붓스트랩 검정

3) $n_i = 10$일 때, 통계 검정법과 효과 크기에 따른 검정력

세 집단의 표본 크기가 각각 10인 조건에서, 집단 간 평균의 흩어진 정도를 최대 수준으로 했을 때와 중간 수준으로 했을 때 통계 검정법과 효과 크기에 따른 검정력을 <그림 10>과 <그림 11>에 제시하였다. 평균 변동성이 최대일 때, 평균적으로 검정력이 가장 큰 검정법은 붓스트랩 검정, 가장 작은 것은 잭나이프 검정이었다. 모든 효과 크기 조건에서도 붓스트랩 검정의 검정력은 항상 최대, 잭나이프 검정의 검정력은 최저에 해당했다. 또, 효과 크기의 증가에 따른 검정력의 향상이 모든 검정법에서 일관적으로 관찰되었다. 이와 같은 경향은 평균 변동성이 중간 수준일 때에도 유사하게 나타났다. 검정법의 종류나 효과 크기에 따른 검정 결과의 패턴은 두 변동성 조건에서 거의 동일했다.



<그림 10> $n_1 = n_2 = n_3 = 10$인 최대 변동성 조건에서
통계 검정법과 효과 크기에 따른 검정력



<그림 11> $n_1 = n_2 = n_3 = 10$인 중간 변동성 조건에서
통계 검정법과 효과 크기에 따른 검정력

구체적으로 최대 변동성 조건에서 효과 크기가 증가함에 따라 검정력이 변화하는 정도는 잭나이프 검정이 0.9434로 가장 큰 변화가 있었다. 임의순열 검정이 0.9325, KW 검정이 0.9294, 일원분산분석이 0.9271만큼 변화했으며, 붓스트랩 검정은 0.8772만큼 증가하여 다섯 가지 통계 검정법 중 가장 작은 검정력 변화가 있는 것으로 나타났다. 통계 검정법 간 검정력 차이가 가장 잘 나타나는 효과 크기 조건은 0.8로, 붓스트랩 검정과 잭나이프 검정의 검정력 사이에 0.1648의 차이가 있었다. 기대된 검정력과 경험적 검정력을 비교해보면 붓스트랩 검정은 항상 기대보다 높은 수행을 보였다. 일원분산분석과 임의순열 검정의 검정력은 기대된 검정력과 거의 유사했지만, 이를 제외하고는 어떤 조건에서도 기대된 수준보다 검정력이 낮았다(<표 10> 참조).

<표 10> $n_1 = n_2 = n_3 = 10$인 최대 변동성 조건에서 통계 검정법과 효과 크기에 따른 검정력 ($\alpha = .05$)

검정법 / 효과 크기	모수 검정	비모수 검정				기대된 검정력
	ANOVA	KW	JACK	PERM	BOOT	(EP)
0.2	0.0680	0.0628	0.0464	0.0601	0.1198	0.06842
0.5	0.1787	0.1637	0.1362	0.1631	0.2568	0.17768
0.8	0.3995	0.3709	0.3340	0.3775	0.4988	0.39863
1.2	0.7466	0.7097	0.6784	0.7161	0.8077	0.74921
2.0	0.9951	0.9922	0.9898	0.9926	0.9970	0.99490

주. ANOVA=일원분산분석, KW=크루스칼-왈리스 검정, JACK=잭나이프 검정, PERM=임의순열 검정, BOOT=붓스트랩 검정

한편, 같은 효과 크기를 가질 때 중간 변동성 조건에서 구해진 검정력은 최대 변동성 조건의 검정력에 비해 항상 작았다. 그러나 통계 검정법이나 효과 크기가 검정력에 미치는 영향은 두 변동성 조건에서 전반적으로 유사하게 나타났다. 자세히 살펴보면 잭나이프 검정에서 효과 크기 증가에 따라 0.9168로 가장 큰 검정력 변화가 있었다. 임의순열 검정에서 0.9140, 일원분산분석에서 0.9101, KW 검정에서 0.9021 정도의 변화가 관찰되었고, 붓스트랩 검정은 0.8722만큼 증가하여 가장 작은 검정력 변화가 있는 것으로 나타났다. 검정법 간 검정력의 차이가 가장 두드러지는 효과 크기 조건은 0.8로, 잭나이프 검정과 붓스트랩 검정 사이에서 0.1502의 차이가 나타났다. 기대된 검정력과 실제 검정력을 비교한 결과, 붓스트랩 검정은 항상 기대보다 높은 수행을 보였다. 일원분산분석과 임의순열 검정의 검정력은 기대된 검정력과 거의 일치했고, 나머지 검정법은 어떤 조건에서도 기대된 수준보다 검정력이 낮았다(<표 11> 참조).

<표 11> $n_1 = n_2 = n_3 = 10$인 중간 변동성 조건에서 통계 검정법과 효과 크기에 따른 검정력 ($\alpha = .05$)

검정법 / 효과 크기	모수 검정	비모수 검정				기대된 검정력
	ANOVA	KW	JACK	PERM	BOOT	(EP)
0.2	0.0636	0.0584	0.0446	0.0584	0.1116	0.06372
0.5	0.1433	0.1325	0.1086	0.1331	0.2174	0.14347
0.8	0.3117	0.2879	0.2595	0.2969	0.4097	0.30853
1.2	0.6126	0.5741	0.5508	0.5944	0.6996	0.61630
2.0	0.9737	0.9605	0.9614	0.9724	0.9838	0.97326

주. ANOVA=일원분산분석, KW=크루스칼-왈리스 검정, JACK=잭나이프 검정, PERM=임의순열 검정, BOOT=붓스트랩 검정

Ⅵ. 논의

본 연구에서는 3개 이상의 독립표본 간 차이 검정에 활용 가능한 일원분산분석, 크루스칼-왈리스 검정, 잭나이프 검정, 임의순열 검정, 붓스트랩 검정의 수행을 비교하였다. 1종 오류 비율과 검정력을 기준으로 각 검정법을 비교함으로써 주어진 조건에 가장 적합한 통계 검정법이 무엇인지를 탐색하고, 추후 자료의 특성에 맞는 검정법 선택의 지표로 활용하고자 하였다. 특히 소표본 자료를 분석할 때 대안적 비모수 검정법이나 재표집 방법과 같은 비모수적 검정을 사용하는 것이 모수 검정을 사용하는 것보다 더 나은지, 어떤 비모수적 검정이 표본 크기로 인한 문제 개선에 가장 효과적인지 등 표본 크기에 대한 이슈를 중점적으로 살펴보고자 하였다. 연구 결과에 따른 가설의 기각 및 지지 여부는 다음과 같다.

연구문제1에서는 집단 간 차이가 없는 조건, 즉 효과 크기가 0이고 평균 변동성이 없을 때 통계 검정법과 표본 크기가 1종 오류 비율에 영향을 미치는지를 관찰하였다. 연구 결과, 검정법에 따라 1종 오류 비율이 서로 다르게 나타났다. 1종 오류 비율의 평균은 붓스트랩 검정에서 가장 컸고, 일원분산분석, 임의순열 검정, KW 검정, 잭나이프 검정 순으로 작았다. 표본 크기에 따라서도 1종 오류 비율에 차이가 있었다. 표본 크기가 3인 조건에서 1종 오류 비율이 가장 높았고, 표본 크기가 커질수록 점점 감소하여 표본 크기가 10인 조건에서는 평균적인 1종 오류 비율이 유의수준 α에 가까워지는 것을 확인할 수 있었다. 이는 가설 1-1인 '세 집단 간 차이가 없는 조건에서 통계 검정법에 따라 1종 오류 비율이 달라질 것이다'와 가설 1-2인 '세 집단 간 차이가 없는 조건에서 표본 크기에 따라 1종 오류 비율이 달라질 것이다'를 지지하는 결과라고 볼 수 있다.

경험적 1종 오류 비율의 수용 가능성을 평가하기 위해 Bradley(1978)가

제시한 1종 오류의 강건성 기준(criteria of robustness)에 따르면, 명목 유의수준(nominal alpha level)이 α일 때 $\alpha_{nominal} \pm 0.1\alpha_{nominal}$를 보수적인 기준(conservative criterion), $\alpha_{nominal} \pm 0.5\alpha_{nominal}$를 관대한 기준(liberal criterion)으로 정의할 수 있다. 이러한 기준에 따르면 일원분산분석과 임의순열 검정의 1종 오류 비율은 표본 크기와 무관하게 항상 보수적 기준의 범위 내에 있었으며, KW 검정과 잭나이프 검정의 1종 오류 비율은 표본 크기가 3인 조건을 제외하고 관대한 기준에 포함되었다. 반면, 붓스트랩 검정은 어떤 경우에도 수용 가능한 범위 내에 포함되지 못했다.

연구문제2에서는 $n_i = 3$인 소표본 조건에서 통계 검정법과 효과 크기가 검정력에 영향을 미치는지 관찰하였다. 그 결과, 평균 변동성이 최대일 때는 모든 검정법의 평균적인 검정력 수준에서 차이가 발견되었다. 구체적으로 크기를 비교했을 때, 검정력이 가장 큰 붓스트랩 검정을 기준으로 일원분산분석, 임의순열 검정, 잭나이프 검정, KW 검정 순으로 검정력의 평균이 작아짐을 확인할 수 있었다. 검정법의 종류뿐만 아니라 효과 크기 조건 또한 검정력에 영향을 미치는 것으로 나타났다. 효과 크기가 증가할수록 검정력의 평균도 증가하였으며, 변화량에 차이가 있기는 하지만 모든 검정 방법에서 효과 크기가 증가할수록 검정력이 증가하는 경향이 관찰되었다. 따라서, 가설 2-1인 '세 집단의 표본 크기가 각각 3인 최대 변동성 조건에서 통계 검정법에 따라 검정력이 달라질 것이다'와 가설 2-2인 '세 집단의 표본 크기가 각각 3인 최대 변동성 조건에서 효과 크기에 따라 검정력이 달라질 것이다'는 지지되었다.

한편, 평균 변동성이 다른 두 조건에서 통계 검정법과 효과 크기에 따른 검정력의 변화 패턴은 거의 유사했다. 각 통계 검정법의 평균 검정력을 크기 순으로 정렬한 결과는 최대 변동성 조건과 중간 변동성 조건에서 서로 일치했고, 효과 크기가 커질수록 검정력이 증가하는 경향도 동일하게 나타

났다. 하지만, 중간 변동성 조건에서의 검정력 크기는 대체로 최대 변동성 조건에 비해 작다는 점에서 차이가 있었다. 일부 효과 크기 조건에서는 변동성 조건에 따른 검정력 크기가 역전되어 나타나기도 했는데, 효과 크기가 0.2인 경우 일원분산분석, KW 검정, 임의순열 검정, 2.0인 경우 KW 검정의 검정력이 그러한 결과를 보였다. 이와 같은 결과는 가설 2-3인 '세 집단의 표본 크기가 각각 3인 중간 변동성 조건에서 통계 검정법에 따라 검정력이 달라질 것이다'와 가설 2-4인 '세 집단의 표본 크기가 각각 3인 중간 변동성 조건에서 효과 크기에 따라 검정력이 달라질 것이다'를 지지하는 결과로 해석할 수 있다.

연구문제3에서는 n_i=5인 소표본 조건에서 통계 검정법과 효과 크기가 검정력에 영향을 미치는지 확인하였다. 연구 결과, 최대 변동성 조건에서는 모든 검정 방법에서 검정력의 평균이 달리 나타났다. 효과 크기 조건에 따라서도 평균적인 검정력 수준에 차이가 나타나 통계 검정법과 효과 크기가 검정력에 영향을 미침을 알 수 있었다. 이처럼 표본 크기가 5인 조건에서의 전반적인 검정 결과는 표본 크기 3인 소표본 조건과 유사했으나, 다소 다른 결과가 나타나는 부분도 있었다. 예컨대, 검정력이 가장 높은 검정법은 붓스트랩으로 동일했지만, 가장 낮은 검정법은 KW 검정에서 잭나이프 검정으로 달라졌다. 표본 크기가 3일 때에 비해 검정력이 더 급격하게 증가하는 것 역시 두 표본 크기 조건에서 나타나는 차이점에 해당됐다. 결과를 종합적으로 고려했을 때, '세 집단의 표본 크기가 각각 5인 최대 변동성 조건에서 통계 검정법에 따라 검정력이 달라질 것이다', 그리고 '세 집단의 표본 크기가 각각 5인 최대 변동성 조건에서 효과 크기에 따라 검정력이 달라질 것이다'라는 가설 3-1과 3-2는 지지되었다.

중간 변동성 조건에서도 통계 검정법과 효과 크기에 따른 검정력의 차이가 나타났다. 전체적인 검정력의 크기는 최대 변동성을 가정했을 때보다 작

았지만, 모든 검정 방법 간 검정력 평균에 차이가 있었고, 효과 크기에 따라 평균적인 검정력 수준에 차이가 관찰되는 등 유사한 결과를 보였다. 평균 변동성이 중간 수준일 때, 표본 크기가 3인 조건과 5인 조건에서의 검정력을 비교한 결과는 앞서 기술한 최대 변동성 조건에서와 비슷했다. 표본 크기가 증가함에 따라 검정력이 가장 낮은 검정법이 KW 검정에서 잭나이프 검정으로 바뀌었고, 전반적인 검정력 변화량의 증가가 관찰되었다. 이와 같은 결과를 통해, 가설 3-3인 '세 집단의 표본 크기가 각각 5인 중간 변동성 조건에서 통계 검정법에 따라 검정력이 달라질 것이다'와 가설 3-4인 '세 집단의 표본 크기가 각각 5인 중간 변동성 조건에서 효과 크기에 따라 검정력이 달라질 것이다'는 지지된 것으로 볼 수 있다.

연구문제4에서는 충분한 표본으로 정의한 $n_i = 10$ 조건에서 통계 검정법과 효과 크기가 검정력에 영향을 미치는지를 확인하였다. 그 결과, 최대 변동성 조건에서 모든 검정 방법이 평균적으로 서로 다른 크기의 검정력을 갖는 것으로 나타났다. 효과 크기의 증가에 따라서는 검정력의 평균이 점차 증가하여, 검정법의 종류와 효과 크기 수준이 검정력에 영향을 미치고 있음이 드러났다. 충분한 표본 조건의 검정력을 소표본 조건과 비교한 결과, 평균적으로 가장 큰 검정력을 갖는 검정법은 일관되게 붓스트랩 검정인 것으로 나타났다. 가장 작은 검정력은 표본 크기가 5인 조건과 일치하게 잭나이프 검정에서 관찰되었다. 표본 크기가 커질수록 검정력이 변화하는 정도는 더욱 커졌다. 이러한 결과를 토대로, '세 집단의 표본 크기가 각각 10인 최대 변동성 조건에서 통계 검정법에 따라 검정력이 달라질 것이다'라는 가설 4-1과 '세 집단의 표본 크기가 각각 10인 최대 변동성 조건에서 효과 크기에 따라 검정력이 달라질 것이다'라는 가설 4-2는 지지되었다.

평균 변동성이 중간일 때에는 최대 변동성 조건에 비해 항상 작은 검정력이 관찰되었지만, 검정력 차이가 나타나는 패턴은 유사하여 마찬가지로 통

계 검정법과 효과 크기에 따른 검정력의 차이가 있었다. 검정력을 표본 크기에 따라 비교한 결과 역시 최대 변동성 조건에서와 비슷했다. 표본 크기가 3인 중간 변동성 조건에서 검정력 평균이 가장 낮은 검정법은 KW 검정이었으나, 표본 크기가 5, 10인 조건에서는 잭나이프 검정에서 가장 낮은 검정력이 관찰되었다. 또한, 표본 크기가 커질수록 조건에 따른 검정력의 변화 정도가 커지는 것으로 나타났다. 이와 같은 결과는 가설 4-3인 '세 집단의 표본 크기가 각각 10인 중간 변동성 조건에서 통계 검정법에 따라 검정력이 달라질 것이다' 와 가설 4-4인 '세 집단의 표본 크기가 각각 10인 중간 변동성 조건에서 효과 크기에 따라 검정력이 달라질 것이다' 를 지지하는 결과라고 해석할 수 있다.

주요 연구 결과를 중심으로 논의하고, 그에 따른 의의를 살펴보면 다음과 같다. 첫 번째로, 본 연구는 세 집단의 차이를 검정할 수 있는 일원분산분석, 크루스칼-왈리스, 잭나이프, 임의순열, 붓스트랩 검정을 비교한 연구로 조작변수에 따른 연구 결과를 다음과 같이 요약할 수 있다. 먼저 표본 크기가 커질수록 1종 오류 비율은 점차 유의수준인 .05에 가까워졌고, 검정력의 변화량이 증가하였다. 효과 크기에 따라서는 효과 크기가 증가할수록 검정력이 증가하는 것으로 나타났다. 마지막으로 평균 변동성은 변동성 수준이 높을수록, 즉 최대 변동성 수준일 때 더 높은 검정력을 가졌다.

이와 같은 결과는 집단이 2개일 때 모수 검정법과 비모수적 검정법의 수행을 비교한 Blair와 Higgins(1985), Zimmerman(1987) 등의 연구를 3개 이상의 집단이 있는 경우로 확장하여 일반화 가능성을 제고하였다는 데 의의가 있다. 그뿐만 아니라 선행연구가 주로 단순하게 모수 검정법과 대안적 비모수 검정법을 비교하거나 재표집 방법의 일부만을 다루었던 것과 달리(Feir-Walsh & Toothaker, 1974; Hecke, 2012; Mendeş & Akkartal, 2010), 대표적인 재표집 방법인 잭나이프, 임의순열, 붓스트랩

검정을 모두 비교 대상에 포함하여 통제된 연구 조건 아래 각 검정법의 효율성을 총체적으로 비교해보았다는 점에서 또 다른 의의를 갖는다.

두번째, 본 연구에서는 표본 크기가 증가함에 따라 통계 검정법들의 수행이 전반적으로 유사해지는 경향이 관찰되었다. 즉, 충분히 큰 표본을 가졌다면 검정법의 선택이 연구 결과에 영향을 미치게 될 가능성이 낮다고 볼 수 있다. 하지만 또 다른 측면에서는 표본 크기가 작을수록 어떤 검정법을 사용하는지가 연구 결과에 결정적인 영향력을 가지게 되는 것으로 해석할 수 있고, 이는 곧 소표본에서 검정법의 선택이 가지는 중요성이 더욱 크다는 점을 시사한다.

따라서 표본 크기가 작다는 이유로 의례적으로 비모수 검정법을 사용하기보다는, 주어진 조건에 따라 검정법의 수행 능력을 면밀히 살펴야 할 필요가 있다. 본 연구는 조건이 달리 부여된 자료들에 대한 통계 검정법의 수행을 비교하고 있다는 점에서, 추후 연구자가 자료의 특성에 맞는 검정법을 선택하는 데 도움을 제공할 수 있을 것으로 기대된다. 연구결과를 바탕으로 각 통계 검정법 사용의 가이드라인을 다음과 같이 제안할 수 있다.

우선 일원분산분석은 표본 크기가 매우 작은 조건에서도 보수적인 1종 오류 강건성 기준을 만족하고 기대된 검정력에 가까운 검정력을 얻어, 결과상으로는 가장 안정적이고 우수한 검정법으로 보였다. 하지만, 이러한 결과는 본 연구에서 가정하고 있는 모집단의 정규성 가정으로부터 기인하였을 가능성이 크다. 이는 정규분포를 벗어난 다양한 분포 조건에서 본 연구의 반복 검정 필요성을 제고하는 부분으로, 본 연구의 한계이자 후속연구 제언과도 연계될 수 있다.

KW 검정은 표본 크기가 극단적으로 작을 때 검정력과 1종 오류가 지나치게 작은 문제를 보였다. 이는 모집단 분포의 정규성을 확신할 수 있을 때 KW 검정의 검정력이 일원분산분석의 검정 결과와 거의 유사하다는 Neave

와 Worthington(1988)의 주장에 반하는 결과이다. 그러나 표본 크기가 증가할수록 붓스트랩을 제외한 나머지 검정법들과 유사한 수행을 보이게 되는 것으로 보아, 일반적으로 표본 크기가 작을 때 KW 검정의 사용이 권장되고 있지만 일정 수준 미만일 때(가령 $n_i < 5$)는 KW 검정이 적합하지 않은 것으로 해석할 수 있다.

잭나이프 검정은 표본 크기와 무관하게 1종 오류 비율이 매우 낮고, $n_i = 3$ 조건을 제외한 표본 크기 조건에서 항상 가장 작은 검정력이 관찰되었다. 이것은 1개의 관측치를 제거하는 잭나이프 검정의 특성상 재표집 횟수가 곧 표본 크기와 동일하므로 소표본에 사용될 경우 추정치의 정확성에 문제가 발생할 가능성이 높다고 주장한 선행 연구를 지지하는 결과이다(Carsey & Harden, 2013). 따라서, 지나치게 작은 표본에서는 잭나이프 검정의 사용이 권장되지 않는다고 볼 수 있다.

임의순열 검정은 모든 조건에서 일원분산분석 결과와 유사한 수준으로 양호한 수행을 보였다. 이는 정규분포를 만족하는 경우 임의순열 검정과 일원분산분석 결과가 거의 일치한다는 Gleason(2013)의 연구결과를 지지하는 것으로 볼 수 있다. 두 검정 결과의 유사성으로 미루어 보아, 임의순열 검정은 다른 비모수적 방법들이 지나치게 작은 표본에서 충분히 기능하지 못하는 부분까지도 보완할 수 있다는 점에서 모수 검정법의 대안적 방법으로써 활용 가능성을 기대해 볼 수 있다. 그러나 현실적으로 소표본에서 정규성 가정이 충족되기 어렵다는 점을 고려했을 때, 임의순열 검정과 일원분산분석의 의미있는 비교를 위해 정규분포 외의 조건에서 두 검정법을 비교하는 추가적인 연구가 필요하다.

마지막으로 붓스트랩 검정 결과 모든 조건에서 가장 높은 검정력을 가지는 것으로 나타났으나, 동시에 1종 오류 비율이 지나치게 높다는 문제가 발견되어 효율적인 검정법으로 판단하는 것에 다소 어려움이 있었다. 붓스트

랩 검정에서 매우 큰 검정력과 1종 오류 비율이 산출된 원인은 붓스트랩의 재표집 방식이 복원추출로 이루어져 동일한 재표집 표본이 반복적으로 사용되었기 때문일 것으로 예상된다. 또 다른 원인으로는 본 연구에 사용된 백분위 붓스트랩 신뢰구간 추정법이 표본 크기에 민감하다는 점을 꼽을 수 있다(Mooney & Duval, 1993). 두 원인 모두 표본 크기가 작은 경우에 문제가 발현되기 때문에, 소표본 데이터보다는 한 집단의 크기가 10보다 큰 충분한 크기의 표본에서 더 효율적인 수행능력을 보일 것으로 예측할 수 있다. 그 외에 붓스트랩 검정의 1종 오류 문제를 해결하는 방안으로는 널리 사용되는 유의수준인 $\alpha = .05$가 아니라, 선행 연구결과를 바탕으로 한 경험적 유의수준을 설정하는 것을 고려해보아야 할 필요가 있다(Franks & Huck, 1986; Pérez & Pericchi, 2014; Pericchi & Pereira, 2016).

한편, 본 연구의 제한점을 살펴보고, 이에 대한 후속 연구를 제언하면 다음과 같다. 첫째, 집단 간 차이 검정을 위해 만들어 낸 가상의 세 집단이 모집단 분포의 정규성을 충족하는 것으로 가정하였기 때문에 모수 검정과 비모수적 검정을 비교하는 데 어려움이 있다. 본 연구에서 각 집단은 주어진 효과 크기와 평균 변동성 조건에 맞게 형성된 평균 μ_i와 표준편차를 1로 하는 정규분포$N(\mu_i, 1)$을 따르는 것으로 설정되었다. 정규분포 가정은 모수 검정법의 활용 가능 여부를 판가름하는 기준으로, 정규성이 완벽히 충족되는 경우 모수 검정법의 통계적 이점이 극대화된다는 강점이 있다. 따라서, 모든 자료가 정규성을 가정한다면 모수 검정법에서 항상 양호한 결과만이 관찰되는 문제가 발생하게 된다. 예를 들어, 실제 연구 장면에서 표본 크기가 $n_i = 3$인 경우는 정규성 가정의 위반을 피하기 어렵기 때문에 모수 검정법을 적용했을 때 부정확한 결과가 도출될 가능성이 높다. 하지만 본 연구에서는 소표본에 대해서도 정규분포를 따르는 것으로 가정하여, 일원분산분석 결과가 양호한 수행 능력을 보였다. 본 연구의 궁극적인 목적이 소표본처럼 현

실성 있는 비모수 자료에 적합한 통계 검정법을 탐색하는 것임을 고려했을 때, 정규분포 이외에도 현실을 잘 반영할 수 있는 다양한 분포 조건을 추가하여 후속연구를 진행해야 할 필요성이 있다.

둘째, 조작 변수의 개수와 각 변수의 수준이 충분하지 않아 현실적인 자료의 특성을 제대로 반영해내기 어렵다. 본 연구에서는 세 집단의 평균 비교 모형을 가정하고, 3가지 표본 크기 조건에서 1종 오류 비율을 비교하고, 표본 크기(3가지)×효과 크기(5가지)×평균 변동성(2가지)의 30가지 조건에서 검정력을 비교하였다. 표본 크기의 경우 중심극한정리에 따라 $n_i = 3$, $n_i = 5$ 조건을 소표본, $n_i = 10$ 조건을 충분한 표본이라 가정하였다. 하지만 연구자의 관점에 따라 $n_i = 10$, 즉 사례 수 30을 충분하지 못한 표본 크기로 보는 경우가 있기 때문에 표본 크기 조건을 확대하여 더 큰 표본 크기를 가질 때의 수행을 살펴볼 필요가 있다. 표본 크기 조건에 관한 또 다른 문제로 모든 집단의 표본 크기를 동일하게 가정했다는 점이 있다. 이는 실제 연구를 통해 수집되는 데이터가 집단 간 동일한 표본 크기를 유지하기 어렵다는 점을 고려했을 때, 다소 현실적이지 못하다. 또한, 모든 집단이 동일한 표본 크기를 가질 경우에 모수 검정법의 1종 오류를 회복하는 효과가 있다는 선행연구 결과로 미루어 보아(Sawilowsky & Blair, 1992), 각 집단의 표본 크기를 다르게 설정하는 조건을 추가하여 본 연구 결과와의 차이를 검토해 보아야 한다. 표본 크기 이외에도 앞서 언급했던 모집단 분포에 대한 가정 등을 새로운 변수로 포함시키는 후속연구가 이루어져야 할 것이다.

참고문헌

김수영, 석혜은 (2015). 잠재성장모형의 사용을 위한 표본크기 결정. 한국심리학회지: 일반, 34(2), 599-617.

박승호 (2000). 교육심리학 용어사전. 서울: 학지사.

백재욱, 윤용운 (1995). Bootstrap 방법의 이해. 統計論文集, -(2), 73-93, 중앙대학교 통계연구소.

성태제 (2014). 현대 기초통계학 : 이해와 적용(제6판). 서울: 학지사.

이강원, 송호웅 (2016). 지형 공간정보체계 용어사전. 서울: 구미서관.

이재원, 이욱기 (2019). (공학인증을 위한) 확률과 통계(제4판), 서울: 북스힐.

이종성, 강계남, 김양분, 강상진 (2007). 사회과학연구를 위한 통계방법(제4판). 서울: 박영사.

이현숙, 김수진, 전수현 (2010). 구조방정식모형 원리와 적용. 서울: 학지사.

정형찬 (2006). 사건연구방법론에서 소표본 문제와 모형의 검정력. 한국증권학회지, 35(3), 107-140.

Adams, D. C., & Anthony, C. D. (1996). Using randomization techniques to analyse behavioural data. *Animal Behaviour, 51(4),* 733-738.

Andrews, D. F., Gnanadesikan, R., & Warner, J. L. (1971). Transformations of multivariate data. *Biometrics, 27,* 825-840.

Armsden, G. C., McCauley, E., Greenberg, M. T., Burke, P. M., & Mitchell, J. R. (1990). Parent and peer attachment in early

adolescent depression. *Journal of abnormal child psychology, 18(6),* 683-697.

Ary, D., & Jacobs, L. C. (1976). *Introduction to statistics.* New York: Holt, Rinehart & Winston.

Blair, R. C., & Higgins, J. J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of student's t statistic under various nonnormal distributions. *Journal of Educational Statistics, 5(4),* 309-335.

Blair, R. C., & Higgins, J. J. (1985). Comparison of the power of the paired samples t test to that of Wilcoxon's signed-ranks test under various population shapes. *Psychological Bulletin, 97(1),* 119-128.

Bradley, J. V. (1978). Robustness?. *British Journal of Mathematical and Statistical Psychology, 31(2),* 144-152.

Breslow, N. (1970). A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. *Biometrika, 57(3),* 579-594.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: methods and applications.* Cambridge university press.

Carroll, R. M., & Nordholm, L. A. (1975). Sampling Characteristics of Kelley's ε and Hays' ω. *Educational and Psychological Measurement, 35(3),* 541-554.

Carsey, T. M., & Harden, J. J. (2013). *Monte Carlo simulation and resampling methods for social science.* Sage Publications.

Chernick, M. R., & Labudde, R. A. (2011). *An introduction to*

bootstrap methods with applications to R. Hoboken, NJ: Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.).* Hillsdale, NJ: Erlbaum.

Cooper, C., & Berwick, S. (2001). Factors affecting psychological well-being of three groups of suicide-prone prisoners. *Current Psychology, 20(2),* 169-182.

Cox, D. R. (2006). *Principles of statistical inference.* Cambridge university press.

DeCoster, J. (2006). Testing group differences using t-tests, ANOVA, and nonparametric measures. Accessed November, 30, 2010 from http://www.stat-help.com/ANOVA%202006-01-11.pdf.

Draper, N. R., & Stoneman, D. M. (1966). Testing for the inclusion of variables in linear regression by a randomization technique. *Technometrics, 8(4),* 695-698.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics, 28(1),* 181-187.

Edgington, E. S. (1980). *Randomization tests.* New York: Marcel Dekker.

Efron, B. (1979a). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7(1),* 1-26.

Efron, B. (1979b). Computers and the theory of statistics: thinking the unthinkable. *SIAM review, 21(4),* 460-480.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika, 68(3),* 589-599.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans (Vol. 38)*. Siam.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York, NY: Chapman & Hall/CRC.

Everitt, B. S. (2006). *The Cambridge dictionary of statistics.* Cambridge University Press.

Feir-Walsh, B. J., & Toothaker, L. E. (1974). An empirical comparison of the ANOVA F-test, normal scores test and Kruskal-Wallis test under violation of assumptions. *Educational and Psychological Measurement, 34(4),* 789-799.

Fisher, R. A. (1935). *The design of experiments.* New York, NY: Hafner.

Fitzmaurice, G. M., Lipsitz, S. R., & Ibrahim, J. G. (2007). A note on permutation tests for variance components in multilevel generalized linear mixed models. *Biometrics, 63(3),* 942-946.

Franks, B. D., & Huck, S. W. (1986). Why does everyone use the. 05 significance level?. *Research Quarterly for Exercise and Sport, 57(3),* 245-249.

Gaito, J. (1959). Non-parametric methods in psychological research. *Psychological Reports, 5(1),* 115-125.

Geisser, S., & Johnson, W. O. (2006). *Modes of parametric*

statistical inference (Vol. 529). John Wiley & Sons.

Gleason, J. H. (2013). Comparative power of the ANOVA, approximate randomization ANOVA, and Kruskal-Wallis test (Doctoral dissertation, Wayne State University).

Good, P. I. (2000). *Permutation tests: A practical guide to resampling methods for testing hypotheses (2nd ed.).* New York: Springer Science & Business Media.

Good, P. I. (2005a). *Permutation, parametric and bootstrap tests of hypotheses (3rd ed.).* New York, NY: Springer.

Good, P. I. (2005b). *Resampling methods: A practical guide to data analysis (3rd ed.)*. Boston, MA: Birkhäuser.

Hays, W. L. (1963). *Statistics.* New York: Holt, Rinehart & Winston.

Hecke, T. V. (2012). Power study of anova versus Kruskal-Wallis test. *Journal of Statistics and Management Systems, 15(2-3),* 241-247.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlanda, FL : Academic.

Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician, 69(4),* 371-386.

Hilt, L. M., & Pollak, S. D. (2012). Getting out of rumination: Comparison of three brief interventions in a sample of youth. *Journal of abnormal child psychology, 40(7),* 1157-1165.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1982). *Basic behavioral statistics.* Boston: Houghton Mifflin.

Howell, D. C. (1992). *Statistical methods for psychology.* Belmont, CA: Duxbury Press.

Hunter, M. A., & May, R. B. (1993). Some myths concerning parametric and nonparametric tests. *Canadian Psychology/Psychologie canadienne, 34(4),* 384.

Kendall, M., Stuart, A., Ord, J., & Arnold, S. (1999). *Kendall's advanced theory of statistics: Vol. 2A-Classical inference and the linear model(6th ed.).* London: Arnold.

Kennedy, P. E. (1995). Randomization tests in econometrics. *Journal of Business & Economic Statistics, 13(1),* 85-94.

Kennedy, P. E., & Cade, B. S. (1996). Randomization tests for multiple regression. *Communications in Statistics B: Simulation and Computation, 25(4),* 923-936.

Kerlinger, F. N. (1964). *Foundations of behavioral research,* New York: Holt, Rinehart & Winston.

Keselman, H. J. (1975). A Monte Carlo investigation of three estimates of treatment magnitude: Epsilon squared, eta squared, and omega squared. *Canadian Psychological Review/Psychologie Canadienne, 16(1),* 44.

Kirk, R. E. (1974). *Experimental design: Procedures for the behavioral sciences.* Belmont, CA: Brooks/Cole.

Kreft, I. G., & De Leeuw, J. (1998). *Introducing multilevel modeling.* Sage.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American*

statistical Association, 47(260), 583-621.

LaFleur, B. J., & Greevy, R. A. (2009). Introduction to Permutation and Resampling-Based Hypothesis Tests∗. *Journal of Clinical Child & Adolescent Psychology, 38(2),* 286-294.

Langbehn, D. R., Berger, V. W., Higgins, J. J., Blair, R. C., & Mallows, C. L. (2000). Letters to the editor. *The American Statistician, 54,* 85-88.

Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician, 52(2),* 127-132.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1(3)*, 86-92.

MacDonald, P. (1999). Power, Type I, and Type III error rates of parametric and nonparametric statistical tests. *The Journal of Experimental Education, 67(4),* 367-379.

Manly, B. F. J. (1997). *Randomization and Monte Carlo methods in biology (2nd ed.).* London: Chapman & Hall/CRC.

Manly, B. F. J. (2006). *Randomization, bootstrap and Monte Carlo methods in biology (3rd ed.).* London: Chapman and Hall/CRC.

Marsee, M. A., & Frick, P. J. (2007). Exploring the cognitive and emotional correlates to proactive and reactive aggression in a sample of detained girls. *Journal of abnormal child psychology, 35(6),* 969-981.

McIntosh, A. (2016). The jackknife estimation method. arXiv

preprint arXiv:1606.00497.

Meek, G. E., Ozgur, C., & Dunning, K. (2000). Does scale of measurement really make a difference in test selection: An empirical comparison of t test vs. Mann Whitney. *Proceedings of the 2000 National Annual Meeting of the Decision Sciences Institute, 951-953.*

Meek, G. E., Ozgur, C., & Dunning, K. (2007). Comparison of the t vs. Wilcoxon signed-rank test for Likert scale data and small samples. *Journal of modern applied statistical methods, 6(1),* 10.

Mendeş, M., & Akkartal, E. (2010). Comparison of ANOVA F and WELCH tests with their respective permutation versions in terms of type I error rates and test power. *Kafkas Univ Vet Fak Derg, 16(5),* 711-716.

Mooney, C. Z. (1997). *Monte carlo simulation (Vol. 116).* Sage Publications.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference.* Thousand Oaks, CA: Sage.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: a second course in statistics.* Addison-Wesley Series in Behavioral Science: Quantitative Methods.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural equation modeling, 9(4),* 599-620.

Nachar, N. (2008). The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology, 4(1),* 13-20.

Nanna, M. J. (2002). Hotelling's T 2 VS. The Rank Transform With Real Likert Data. *Journal of Modern Applied Statistical Methods, 1(1),* 12.

Nanna, M. J., & Sawilowsky, S. S. (1998). Analysis of Likert scale data in disability and medical rehabilitation research. *Psychological Methods, 3(1),* 55.

Neave, H. R., & Worthington, P. L. (1988). *Distribution-free tests.* London: Unwin Hyman.

Okada, K. (2013). Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA. *Behaviormetrika, 40(2),* 129-147.

Pagano, R. R. (1994). *Understanding statistics in the behavioral sciences (4th ed.).* St. Paul, MN: West.

Pérez, M. E., & Pericchi, L. R. (2014). Changing statistical significance with the amount of information: The adaptive α significance level. *Statistics & probability letters, 85,* 20-24.

Pericchi, L., & Pereira, C. (2016). Adaptative significance levels using optimal decision rules: Balancing by weighting the error probabilities. *Brazilian Journal of Probability and Statistics, 30(1),* 70-90.

Pitman, E. J. (1937a). Significance tests which may be applied to

samples from any populations. *Supplement to the Journal of the Royal Statistical Society, 4(1),* 119-130.

Pitman, E. J. G. (1937b). Significance tests which may be applied to samples from any populations. II. The correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society, 4(2)*, 225-232.

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika, 29(3/4),* 322-335.

Potvin, C., & Roff, D. A. (1993). Distribution-free and robust statistical methods: viable alternatives to parametric statistics. *Ecology, 74(6),* 1617-1628.

Quenouille, M. H. (1949, July). Approximate tests of correlation in time-series 3. *In Mathematical Proceedings of the Cambridge Philosophical Society (Vol. 45, No. 3, pp. 483-484).* Cambridge University Press.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika, 43(3/4),* 353-360.

Rao, J. N., & Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the american statistical association, 83(401),* 231-241.

Raz, J. (1989). Analysis of repeated measurements using non-parametric smoothers and randomization tests. *Biometrics, 45(3),* 851-871.

Rizzo, M. L. (2008). *Statistical computing with R.* Boca Raton, FL:

Chapman and Hall/CRC Press.

Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research, 34(4)*, 441-456.

Rosenbaum, P. R. (2002). *Observational studies (2nd ed.).* New York, NY: Springer.

Ryan, T. A. (1959). Multiple comparison in psychological research. *Psychological Bulletin, 56(1)*, 26-47.

Sawilowsky, S. S. (1993). Comments on using alternatives to normal theory statistics in social and behavioural science. *Canadian Psychology/Psychologie canadienne, 34(4),* 432.

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods, 8(2),* 26.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological bulletin, 111(2),* 352.

Shao, J., & Tu, D. (2012). *The jackknife and bootstrap.* Springer Science & Business Media.

Sharp, V. F. (1979). *Statistics for the social sciences.* Boston: Little Brown & Co.

Siegel, S. (1957). Nonparametric statistics. *The American Statistician, 11(3),* 13-19.

Siegel, S., & Castellan, N. J. (1956). *Nonparametric statistics for the behavioral sciences (Vol. 7)*. New York: McGraw-hill.

Sitter, R. R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association, 87(419),* 755-765.

Stephenson, M. T., & Holbert, R. L. (2003). A Monte Carlo simulation of observable versus latent variable structural equation modeling techniques. *Communication Research, 30(3),* 332-354.

Streiner, D. L. (2006). Sample size in clinical research: When is enough enough?. *Journal of Personality Assessment, 87(3),* 259-260.

Su, X., & Tsai, C. L. (2011). Outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(3),* 261-268.

Tressler, J., & Smotherman, M. S. (2009). Context-dependent effects of noise on echolocation pulse characteristics in free-tailed bats. *Journal of Comparative Physiology A, 195(10),* 923-934.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics, 29(1),* 614.

VanVoorhis, C. W., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in quantitative methods for psychology, 3(2),* 43-50.

Weber, M. (2006). Robustness and power of the t, permutation t and Wilcoxon tests. (Doctoral dissertation, Wayne State University).

Wright, D. B. (2006). Comparing groups in a before-after design: When t test and ANCOVA produce different results. *British Journal of Educational Psychology, 76(3),* 663-675.

Zerbe, G. O. (1979). Randomization analysis of the completely randomized design extended to growth curves. *Journal of the American Statsitical Association, 74(365),* 215-221.

Zimmerman, D. W. (1987). Comparative power of Student t test and Mann-Whitney U test for unequal sample sizes and variances. *The Journal of Experimental Education, 55(3),* 171-174.

Zimmerman, D. W., & Zumbo, B. D. (1990a). Effect of outliers on the relative power of parametric and nonparametric statistical tests. *Perceptual and motor skills, 71(1),* 339-349.

Zimmerman, D. W., & Zumbo, B. D. (1990b). The relative power of the Wilcoxon-Mann-Whitney test and Student t test under simple bounded transformations. *The Journal of general psychology, 117(4)*, 425-436.

ABSTRACT

Comparison of Non-parametric Method and Resampling Methods for Small Sample Sizes using Type I Error Rates and Power

Minji Jo
Department of Psychology
The Graduate School of
Sungshin University

This is a simulation study comparing Kruskal-Wallis test, which is non-parametric method using rank, and Jackknife test, Permutation test, and Bootstrap test, which are resampling methods, as alternative statistical methods of one-way ANOVA(analysis of variance). It is conducted through comparing the performance of the statistical methods using type I error rates and power. For this purpose, 30 experimental conditions were set based on the factorial design of 3×5×2 depending on the level of three manipulated variables, sample size, effect size and mean variability. Each experimental conditions were repeated 10,000 times. Then, the error rates and power of each methods were calculated and evaluated. Both of generation and analysis of data were conducted

by using Monte Carlo simulation in the statistical program R.

As a result of the analysis, it showed that there were differences in the type I error rates according to the type of statistical methods and the sizes of sample. On average, the type I error rates were the largest in Bootstrap test, and the smallest in Jackknife test. Furthermore, the larger the sample size was, the smaller the mean of the type I error rates was. As increasing of the sample size, type I error rates of all statistical methods gradually approached to the significance level $\alpha = .05$.

Besides, it showed that there were differences in the power according to the type of statistical methods and the level of the effect size. The larger the effect size was, the bigger the power of the statistical method was. The means of the power were the largest in Bootstrap test, and the smallest in Jackknife test. The power was always higher under the maximum variability condition than the intermediate condition, but the overall pattern of power changes was similar to each other. As the sample size was getting closer to the sufficient level, in addition, the differences of the power among the statistical methods became smaller. Since the statistical methods has their own advantages and disadvantages depending on the form and condition of the data given to the researchers, therefore, it is suggested to use the appropriate method for the different research problems.

The results of this study are meaningful in that they increase the possibility of generalization by expanding the existing comparative

studies of parametric and non-parametric methods. Furthermore, it can be a basis for future researchers to choose the statistical method that is suitable for the characteristics of their data. Finally, a proposal for further researches was discussed.

Key Words: Non-parametric method, Resampling methods, Kruskal-Wallis test, Jackknife test, Permutation test, Bootstrap test, Simulation, Type I error rates, Power