Korean Journal of Sport Science
2020, Vol. 31. No. 2, 216-229
https://doi.org/10.24985/kjss.2020.31.2.216

체육과학연구
ⓒ Korea Institute of Sport Science

A study on the criticism and alternatives about statistical hypothesis testing of sport management in Korea

Sanghyun Park*
Yonsei University

[Purpose] This study was aimed at laying out criticism regarding statistical hypothesis testing and presenting realistic alternatives focused on published studies of sports management in Korea. **[Methods]** Among 202 studies compiled by the Korean Society for Sport Management, vol. 19, no. 1 through vol. 23, no. 6, 115 studies which used the null hypothesis significant testing were finally selected. After data coding for selected studies, p-curve, the distribution of p-values which reported in individual studies, was schematized, and than adequacy about the sampling method and result descriptions were analyzed. **[Results]** The ratio of p-values close to zero was relatively very high in p-curve although there was no clear evidence of p-hacking. Also, approximately 82% of the studies used convenience sampling method, and incorrect descriptions in part of result and discussion due to lack of understanding of statistical hypothesis testing was found in some studies. In conclusion, the shortcomings of statistical hypothesis testing which commonly used in academic field were depended on not a defect in the method itself but researchers who misused statistical hypothesis testing. **[Conclusions]** Researchers in sport management field need to understand the advantages and disadvantages of statistical hypothesis testing and consider to use both confidence interval and effect sizes to compensate the disadvantages of p-value.

Key words: Sport Management, Statistical hypothesis testing

서 론

p-value가 '.05' 이상으로 도출된 연구결과, 즉 자신의 연구가설(alternative hypothesis)의 옳음을 주장할 만한 충분한 근거가 없다는 결과를 받아든 연구자는 고민에 빠진다. 하지만 그 고민에 대한 해결책은 그리 많지 않다. 연구결과를 책상 서랍에 넣어 출판을 포기하거나(file drawer problem), 자신이 알고 있는 여러 가지 방법을 동원해서 p-value를 '.05' 이하로 낮추려(p-hacking) 노력한다. 그러나 사실 가장 상식적이며 당연하게 받아들여질 해결책이 하나 더 존재한다. 그것은 통계적 유의성이 나타나지 않은 연구결과를 있는 그대로 보고하는 것이다. 이렇게 쉽고 명확한 해결책이 있음에도 불구하고, 대체 p-value가 무엇이기에 수많은 연구자들은 p-value 앞에서 전전긍긍하며 고민에 고민을 거듭해야 하는가?

수많은 사회과학 연구자들이 'p-value의 시대'를 살아가고 있다고 해도 과언이 아닐 만큼, 영가설(null hypothesis)을 중심으로 한 통계적 유의성을 검정(null hypothesis significance testing: NHST) 하는 방법은 양적 연구방법의 보편적인 절차로 인식되어왔다. 이와 관련된 실증연구를 살펴보면 다음과 같다. Sterling, Rosenbaum, and Weinkam(1995)의 연구에서는 심리

논문 투고일 : 2020. 04. 23.
논문 수정일 : 2020. 05. 25.
게재 확정일 : 2020. 06. 01.
* 교신저자 : 박상현(tkdlight@naver.com).

학 분야의 대표적인 8개 저널(Journal of General Psychology, Journal of Personality and Social Psychology, 등)에 1986년부터 1987년까지 출간된 연구물의 95.56%가 NHST를 사용하고 있음을 밝히고 있다. 또한, 국내 스포츠 경영학 분야의 연구 동향을 분석한 Kim, Park, & Won(2015)의 연구결과에서도 나타나듯이 한국스포츠산업경영학회지에 출간된 연구물의 절반 이상이 회귀분석(regression) 및 구조방정식모델(structural equation model)과 같은 NHST에 근거한 연구방법을 활용하고 있다는 점(2001년부터 2010년까지 출간된 연구물의 약 77.8%)을 고려해 본다면, 국내 스포츠 경영학 분야의 연구자들 역시 p-value를 통해 자신들이 수행한 연구결과를 보고하는 방식을 보편적으로 사용하고 있음을 짐작할 수 있다.

이와 같은 p-value의 상용성(常用性)에도 불구하고, 일부 학자들은 NHST를 '가장 멍청한 절차(the most bone-headedly misguided procedure)', '과학적 지식의 성장을 지연하는 방법(retards the growth of scientific knowledge)' 등의 다소 과격한 언어를 사용하며 비판하였다(Rozeboom, 1997; Schmidt & Hunter, 1997). 이러한 견해를 가진 학자들이 밝히고 있는 문제점은 NHST가 표본 크기에 민감한 점, 영가설이 비현실적이라는 점 및 연구자의 잘못된 사용 등이다. 이 중에서 연구자의 잘못된 사용과 관련된 사항들은 다음의 두 가지로 요약될 수 있다.

첫째, 연구자들이 임의적인 기준에 불과한 유의수준 $\alpha=.05$를 지나치게 맹신하고 있어, 마치 p-value가 범주형 변수인 것으로 왜곡하여 인식하는 경향이 있다(Halsey, Curan-Everett, Vowler, & Drummond, 2015). 즉, 유의수준 $\alpha=.05$ 보다 큰 p-value가 나타난 결과는 통계적으로 유의하지 않고, 작은 p-value가 나타난 결과는 통계적으로 유의하다는 이분법적 사고로 자의적인 결론을 도출하는 경향이 팽배해 있다는 것이다.

둘째, 앞선 p-value에 대한 이분법적 사고와 맞물려 연구윤리를 위협하는 행동의 원인으로 작용할 가능성이 있다(Simmons, Nelson, & Simonsohn, 2011). 예를 들어, 연구자가 설정한 유의수준(대부분은 $\alpha=.05$)에 살짝 못 미치는 p-value(예를 들어, $p=.052$)가 도출되어 영가설을 기각하지 못할 때, 연구자는 데이터를 인위적으로 늘리거나(double-dipping), 데이터를 조작(manipulation)하는 'p-hacking'의 유혹에 빠질 가능성이 있다. 또한, 연구부정행위로 단정할 수는 없지만, 자료수집 전 연구모형을 설정하고 이를 검정하는 것이 아닌 여러 개의 독립변수를 확보한 후 통계적으로 유의한 결과를 도출하는 독립변수만을 찾아내려는 'star-fishing' 현상도 발생할 수 있다.

하지만, 이러한 문제점에 대한 모든 책임을 연구자들의 통계적 연구방법에 대한 이해 부족이나 연구윤리의식의 부재로 전가할 수는 없다. 개별연구자들은 자신이 수행한 연구의 출판 가능성(publishability)을 고려하지 않을 수 없기 때문이다. 실제로 통계적으로 유의한 결과가 도출된 연구물이 그렇지 않은 연구물에 비해 출판될 가능성이 크다는 출판 편의(publication bais)의 존재는 많은 연구들을 통해 밝혀졌다(Csada, James, & Espie, 1996; Mahoney, 1977; Park & Kwak, 2018; Stern, & Simes, 1997).

요약하자면, NHST로부터 도출되는 p-value는 수립된 영가설에 근거해 데이터의 적합성을 나타내는 확률[1] 일 뿐이며, 표본 크기에 영향을 받는 값이므로 의사결정을 위한 절대적인 기준이 존재하지 않는다. 하지만, 출판 편의가 존재하는 학계에서 다수의 연구자들은 $p<.05$라는 고착된 기준에 의해 자신의 연구결과를 평가하는 경향을 보이며, 연구자가 p-value에 대한 이해가 충분하지 않거나 p-hacking의 유혹을 뿌리치지 못한 경우, 부적절한 연구결과의 제시로 이어질 수 있다는 점은 매우 심각하게 바라보아야 하는 문제라고 할 수 있다.

이에 다양한 학문 분야의 연구자들이 p-hacking의 존재 여부(Head, Holman, Lanfear, Kahn, & Jennions, 2015; Lakens, 2015), NHST에 대한 문제점(Cohen, 1994; Fidler, Burgman, Cumming, Buttrose & Thomason, 2006; Halsey, Curran-Everett, Vowler, & Drummond, 2015; Lai, Kalinowski, Fidler, & Cumming, 2010; Verdam, Oort, & Sprangers, 2014) 및 연구에서 빈번하게 발생하는 통계적 문제점(Clark & Mulligan, 2011; Hupé, 2015; Kim & Lee, 2018)에 대한 실증연구를 수행하였다. 특히, Kim and

1) $p=P(D|H_0)$

Lee(2018)의 연구에서는 스포츠 경영학 분야의 저널(European Sport Management Quarterly, Journal of Sport Management, Sport Management Review, International Journal of Sports Marketing and Sponsorship)에 출간된 연구물에서 나타나는 통계분석과 관련된 연구자들의 실수를 연구실행 전, 연구시행과정, 연구실행 후로 분류하여 보고하였다.

앞서 언급된 여러 부정적 시선에도 불구하고, NHST는 여전히 양적 연구의 주된 의사결정 방법으로 사용되고 있다. 그 이유는 영가설 기각 여부가 달린 양자택일의 결정상황에서 연구자의 불확실성을 줄일 수 있으며, 개별연구물의 연구결과를 통계적으로 종합하는 메타분석이 활발하게 사용되는 현 상황에서 개별연구물들의 연구결과가 결국 메타분석의 자료로 사용될 수 있기 때문이다(Verdam, Oort, & Sprangers, 2014). 또한, 무엇보다도 현실적인 이유는 연구물의 개재여부를 결정하는 심사위원(reviewer)들 역시 NHST에 너무나 익숙해져 있다는 점일 것이다. 만약 한 저자가 자신의 연구결과를 보고함에 있어 p-value 혹은 통계적 유의성 표기(*)를 생략하고, p-value에 대한 대안으로 여겨지는 효과크기(effect size)를 보고하거나, 영가설 기각여부에 대한 불분명한 입장을 취했다면, 그 연구물에 대해 어떠한 심사평이 오게 될까? 그 답을 상상할 필요도 없는 것이 학계를 스포츠경기에 비유한 Bakker, van Dijk, & Wicherts(2012)의 연구에서 유추가능하다. 선수(개별연구자)가 심판(심사위원)에 의해 점수(출판여부)를 인정받고 승리(임용 혹은 수상)에 이르는 과정에서 선수가 심판의 성향을 파악하고 점수를 쉽고 빠르게 획득하는 방법을 경기에 적용하고자 하는 것은 너무도 당연한 일이다.

이러한 학문적 환경 속에서 NHST에 대한 비판적 시각과 올바른 이해 없이 NHST를 적용한 연구를 수행하는 것이 일반적인 연구자들의 불가피한 선택일지라도 NHST가 어떠한 상황에서건 사용하기만 하면 해결되는 '전가의 보도'가 될 수는 없다. 이에 본 연구의 목적은 스포츠 경영학 분야의 연구물을 중심으로 NHST의 잘못된 사용현황을 선행연구에서 소개된 p-curve와 점검기준을 통해 파악하는 것이며, 나아가 지금까지 학계에서 활발히 논의되었던 대안들의 활용과 그 이론적 토대를 소개하는 것이다. 이를 통해, 스포츠 경영학 분야의 학자들에게 NHST에 대한 올바른 이해 및 보고방법에 대한 지침을 제공할 수 있을 것이다.

연구방법

연구물의 선정

본 연구는 국내 스포츠 경영학 분야의 대표 학술지라고 판단되는 한국스포츠산업경영학회지의 19권 1호(2014년)부터 23권 6호(2018년)까지 출간된 총 202편의 연구물 중 NHST을 주요 분석방법으로 적용하였다고 판단되는 115(56.9%)편의 연구물을 대상으로 하였다. 즉, 202편의 연구물 중 NHST를 연구 내에서 일부 사용되고 있으나, NHST가 연구의 주된 연구방법이 아니라고 판단되는 연구(ex. 사회연결망 분석을 활용하여 연결망의 시각화와 해석을 주요 연구주제로 하였으나, 추가적으로 연결망 변수를 추출하여 회귀분석의 변수로 사용한 연구 및 NHST가 활용되나 통계적 유의성의 판단보다는 평균효과크기를 도출하는데 연구의 주된 목적이 있는 메타분석)는 NHST에 대한 연구자의 잘못한 사용현황을 파악하는 본 연구의 첫 번째 목적에 부합되지 않기 때문에 연구대상에서 제외하였다. 선정된 연구물의 기초정보는 〈Table 1〉과 같다.

Table 1. Basic information of studies

	Vol.19	Vol.20	Vol.21	Vol.22	Vol.23
No. 1	9(5)	5(3)	5(4)	4(2)	9(7)
No. 2	8(5)	5(1)	6(4)	7(5)	7(5)
No. 3	10(6)	9(6)	6(3)	4(2)	6(1)
No. 4	9(6)	9(8)	9(7)	6(5)	4(1)
No. 5	10(7)	6(0)	4(3)	5(1)	4(1)
No. 6	10(4)	5(3)	8(3)	5(2)	8(5)
Sum	56(33)	39(21)	38(24)	31(17)	38(20)
Ratio	58.9%	53.8%	63.1%	54.8%	52.6%

*The number in parentheses indicate the number of studies used NHST.

점검 기준

먼저, 연구대상이 되는 115편의 연구물에서 보고되고 있는 다수의 개별 p-value를 x축으로, 전체연구물 대비 해당 p-value를 보고한 연구물의 비율을 y축으로 하여 p-value의 분포를 나타내는 p-curve를 도식화하였다(Bruns & Ioannidis, 2016). 미리 설정된 효과 크기에 따른 p-value의 분포를 p-hacking이 발생한 경우와 발생하지 않은 경우로 구분하여 비교 모의실험을 수행한 Simonsohn, Nelson, & Simmons(2014)의 연구에 의하면, p-hacking이 발생한 경우의 p-curve가 좀 더 우측편향(right-skewed)됨을 보고하였다. 사실 p-curve의 유효성 및 해석에는 학자들의 다양한 견해(Bishop & Thompson, 2016; Simonsohn, Simmons, & Nelson, 2015; Ulrich & Miller, 2015)가 존재하지만, 기본적으로 x축(p-value) 기준 '.04~.05'의 구간에서 갑작스러운 y값(연구물의 비율)의 증가가 나타나는 p-curve가 그려진다면 p-hacking을 의심해 볼 수 있다. 하지만, 본 연구에서는 p-curve를 통해 p-hacking의 존재 여부만을 판단하지 않고 Simonsohn, Nelson, & Simmons(2014)의 모의실험에서 나타난 다양한 조건(효과 크기, 표본 크기)에 따른 p-curve와의 비교를 통해 본 연구물에서 보고된 p-value 분포의 전반적인 형태를 확인하고자 하였다. 특히, p-hacking이 존재하지 않는 조건에서 효과 크기(d)가 0.9 및 0.6일 때의 p-curve를 본 연구를 통해 그려지는 p-curve를 함께 도식화하여 그 상대적 특성을 파악하고자 하였다.

다음으로, Wicherts et al.(2016)의 연구에서 보고된 p-hacking과 관련된 연구자의 34가지 점검목록을 기준으로 본 연구에서 수집된 115개의 연구물을 평가하였다. 점검목록은 설계(design), 수집(collection), 분석(analyses), 보고(reporting)로 구분되어 있으나, Wicherts et al. (2016)의 연구에서 밝힌 34개의 점검목록은 실험설계(experimental design)에 기초하고 있으며, 점검목록 사이에 본질적 유사성이 존재하므로, 변수들 사이의 상관관계(correlation)에 기초한 연구를 주로 수행하는 스포츠 경영학 분야의 연구물에 모두 적용하기에는 한계가 있다(e.g. 무선배치 여부, 실험대상자들에 대한 적절한 통제 여부, 및 검정력 분석 수행 여부, 등). 또한, 점검목록의 특성상 이미 출간된 논문을 통해 평가할 수 없는 항목도 다수 존재한다(e.g. 효과의 방향성이 모호한 가설설정, 다양한 방법으로 종속변수를 측정, 추정방법의 선정, 구체적인 표집계획 수립, 분석 후 가설설정, 분석 후 변수추가, 등). 이에 본 연구에서는 115개 연구물에서 확인 가능하며, NHST의 사용과 긴밀한 연관성이 있다고 판단되는 2개 항목(표본추출방법, 결과 서술)에 대해서만 평가하였다.

코딩

분석 대상 연구물의 목록을 확보한 후, 앞서 언급된 분석 기준에 해당하는 정보(p-value, 표본추출방법, 결과 서술) 및 기타 중요한 정보(표본 크기, 가설의 수 등)를 통계분석 경험이 풍부한 연구자와 연구보조원 1인이 상호 독립적으로 코딩하였다. 코딩의 과정에서 p-value를 보고하지 않고, 검정 통계량과 통계적 유의성 여부만을 보고한 연구물의 경우에는 연구자가 p-value를 직접 계산하여 코딩하였다.

코딩의 일치도

연구자와 연구보조원이 독립적으로 수행한 코딩이 완료된 후, 평가자 간 신뢰도를 판단할 수 있는 Kappa 지수를 산출하였다(McHugh, 2012). Kappa 지수는 0.9이상으로 도출되었을 때, 거의 완벽한 수준(almost perfect)이라고 할 수 있다. 본 연구에서는 Kappa 지수가 최초 .740로 도출되어, 평가자 간 코딩의 일치성이 다소 떨어지는 것으로 나타났지만 단순한 코딩의 오류(소수점 표기의 실수, 불성실한 응답을 제외하기 전 표본 크기 기입, 표본추출방법의 오기, 등)가 대부분인 것으로 나타났다. 이에 일치하지 않는 부분에 대해서는 평가자 2인이 재확인하였고 이후 합의를 통해 수정하여 일치된 내용으로 코딩하여 완벽한 수준의 평가자 간 일치성을 확보하였다.

연구결과 및 논의

연구결과 및 논의를 기술하기에 앞서, NHST의 사용

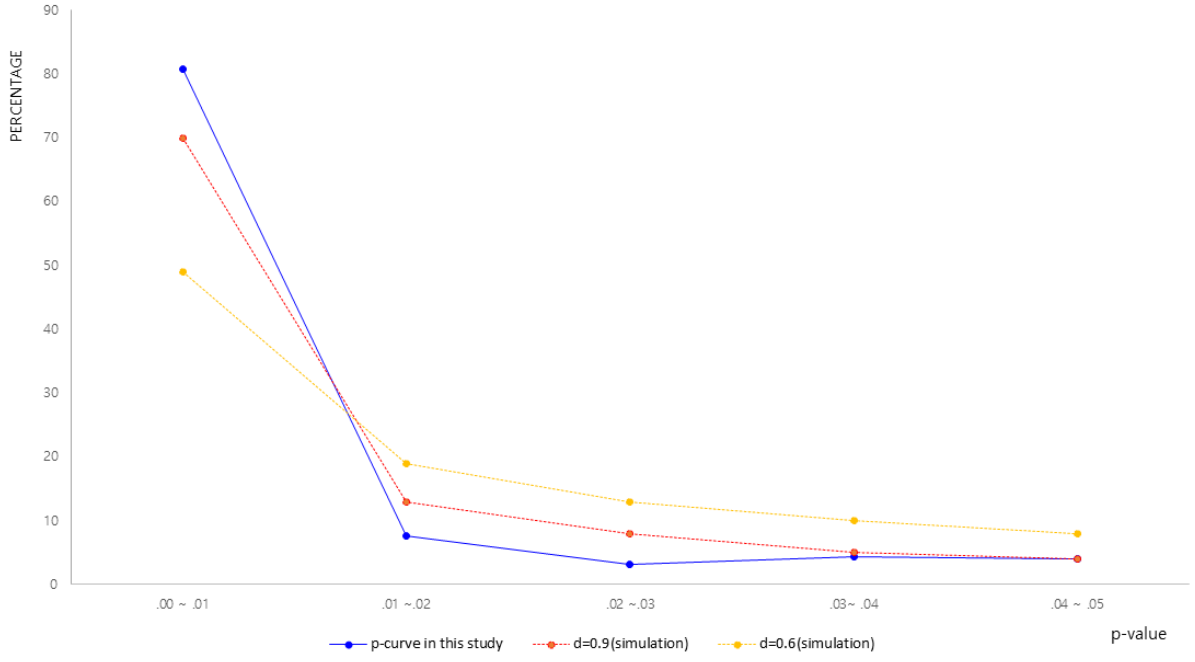

Fig. 1. p-curves

과 관련된 문제들을 자세하게 보고하기 위해, 임의로 선정된 연구의 일부를 발췌하였음을 밝힌다.

p-curve

먼저, p-curve를 도식화하기 위해 115편에서 보고하고 있는 개별 p-value를 코딩한 후, 각 구간에 따라 분류한 후, 그 비율을 계산하였다. 앞선 연구방법에서 언급한 대로 p-value를 코딩하는 과정에서 연구가설과 직접적인 관련이 없다고 판단되는 p-value(회귀분석의 F-test에 대한 p-value, t-test의 등분산 검정, 등)는 코딩되지 않았다. 이와 같은 절차를 통해 〈Fig. 1〉과 같은 p-curve를 도식화하였으며, 특성파악 및 상대비교를 위해 Simonsohn, Nelson, & Simmons(2014)의 모의실험 연구에서 보고된 p-curve(d=0.9, d=0.6)를 함께 도식화하였다.

〈Fig. 1〉과 같이 본 연구의 포함된 115개의 연구물에서 보고되고 있는 622개의 p-value를 통해 도식화한 p-curve에서는 .04~.05 구간의 비율(4.01%)이 인접한 구간에 비해 상대적으로 높다고 할 수 없으므로, 이러한 결과를 통해 p-hacking을 의심할만한 충분한 근거가 있다고 단정할 수 없다. 하지만, p-hacking의 존재 여부와 관계없이, 전체적인 p-curve의 형태를 다른 p-curve와 비교한다면, 0에 근접한(.00~.01) 구간의 p-value 비율이 전체의 약 80.8% 정도를 차지하고 있다는 점에 주목할 필요가 있다. Simonsohn, Nelson, & Simmons(2014)가 수행한 모의실험에서는 모집단에서의 효과 크기(d)가 0일 때, p-curve는 균등분포(uniform distribution)를 갖게 되며, 효과 크기가 증가할수록 '0'에 근접한 p-value의 상대적 비율이 높아진다. 모집단에서의 효과 크기(d)가 매우 큰 경우(빨간색 점선으로 그려진 p-curve)를 예로 들면 효과 크기가 0.9(r=.45)[2] 일 때, .00~.01 구간의 p-value의 비율이 전체의 약 70%를 차지함을 보고하고 있다. 즉, 변수 간 상관을 기초로 한 연구를 주로 수행하는 스포츠 경영학 분야에서 연구자가 설정한 가설을 통해 검정하고자 하는 모집단의 실제 효과 크기가 0.9에 이를 만큼 크다는 비현실적인 가정을 할지라도 본 연구에서

2) $d = 2r/\sqrt{1-r^2}$

나타난 p-curve는 지나칠 정도로 .00~.01 구간에 p-value의 비율이 높은 형태를 가지고 있다고 할 수 있다.

이러한 현상의 원인은 매우 다양하게 존재할 수 있으며, 반복모의실험의 결과를 절대적인 기준인양 무조건 수용하여 본 연구의 결과와 비교 및 판단할 수는 없다(Bruns & Ioannidis, 2016). 하지만, 다수의 반복모의실험에서 그 타당성과 신뢰성이 입증되었으므로 본 연구에서 나타난 비전형적인 p-curve의 원인을 조심스럽게 예측해 보면 다음과 같다.

먼저, 실험설계를 적용한 연구와 비교해 상대적으로 큰 표본 크기 및 모형에 투입되는 변수의 선정에 관한 편의성을 들 수 있다. 표본 크기는 추리통계(inferential statistics)의 핵심인 표준오차(standard error)와 반비례의 함수관계를 가지고 있다. 즉, 표본 크기가 커질수록 표준오차는 감소하게 되어 연구자가 통계적 유의성을 발견하기가 쉬워진다. 본 연구에서 사용된 115편의 연구물 중 패널 자료를 사용한 연구를 제외한 연구물의 평균 표본 크기는 약 351개로 나타났다. 적절한 표본 크기에 관한 많은 선행연구들(Dupont & Plummer, 1990; Jo, Bae, Jhin, Lee, & Park, 2009; Kim & Seok, 2015; MacCallum, Browne, & Sugawara, 1996)을 감안할 때, 351개의 표본 크기는 모수를 추정하기에 부족함이 없는 수치라고 볼 수 있다. 하지만, 설문 조사를 통해 300명이 넘는 설문응답자를 확보하는 것에 대한 현실적 어려움을 고려한다면, 본 연구에서 나타난 .00~.01 구간의 p-value의 비율이 지나치게 높은 비율에 대해서 논의될 충분한 여지가 있다. 연구자는 자신이 수립한 연구가설의 통계적 유의성이 발견되지 않았을 때, 추가적인 표본을 수집을 고려하게 된다. 하지만, 올바르지 못한 절차를 통해 늘어난 표본 크기는 변수 간 실제적 유의성(practical significance)이 없음에도 통계적 유의성(statistical significance)을 쉽게 발견할 수 있게 한다. 요약하자면, 본 연구에서 도식화된 비전형적인 형태의 p-curve만을 근거로 하여, 스포츠 경영학 분야에서 p-hacking의 존재 여부 및 비정상적인 형태로 표본 크기를 증가시키고자 하는 행위가 일어나고 있다고 판단할 수는 없다. 하지만 이러한 비전형적인 p-curve의 형태는 설문 조사의 현실적 어려움과 맞물려, 스포츠 경영학 분야의 연구자들이 표본 크기와 관련된 연구윤리 준수에 경종을 울릴 수 있는 간접적인 지표로써 활용될 수 있을 것이다.

앞서 언급하였듯이, p-hacking의 존재 여부를 떠나, 무엇보다도 중요한 것은 표본조사를 통해 수집된 표본의 크기가 아니라 표본의 대표성이다. 올바른 과정과 절차를 거치지 않고 효율성만을 고려해 쉽게 수집된 자료가 많다면 통계적 유의성이 발견될 수는 있겠지만, 실제적 유의성을 보장할 수 없다. 즉, 표본 크기에 영향을 받는 p-value에 전적으로 의존하지 않고, 올바른 절차를 통해 수행된 표본조사를 토대로 한 영가설 기각의 실패는 연구의 실패가 아닌 새로운 해석과 발견의 시도로서 해석하려는 연구자들의 인식변화가 필요하다고 판단할 수 있다.

다음으로, 모형에 투입되는 변수선정의 편의성 역시 본 연구에서 나타난 p-curve를 유발할 가능성이 있다. 실험설계연구와 달리 변수 간 상관을 기초로 한 관찰연구(observational study)는 이론 및 선행연구 결과의 종합적 고찰에 근거(theory-driven)한 가설설정이 이루어져야 하지만 자료수집의 어려움으로 대표되는 현실적 제약으로 인해, 유의한 결과를 찾아내는 연구가 더욱 많이 수행되고 있는 것이 사실이다. 예를 들어, 관련성이 있다고 여겨지는 다양한 변수들을 하나의 설문지에 포함해 이를 측정한 후, 분석 과정에서 변수를 바꾸어가면서 새로운 인과관계 혹은 모형을 검정하다가 통계적 유의성이 강하게 나타나거나 통계적으로 유의한 관계가 많이 나타나는 모형을 최종적으로 선택한 후에 역으로 가설을 설정하는 행위(hypothesize after the results are known: HARKing)가 우리 주변에 전혀 존재하지 않는다고 누구도 단언할 수 없을 것이다(Kerr, 1988; Murphy & Aguinis, 2019).

실제로, Kerr & Harris(1998)의 연구에서는 HARKing 행위에 대한 설득력 있는 증거를 제공하기 위해 156명의 행동과학 연구자들에게 자신의 분야에서 HARKing 행위를 목격한 적이 있는지 질문하였다. 그 결과, 약 50%의 연구자들이 다양한 형태의 HARKing 행위를 직접 관찰한 경험이 있거나 의심스럽다고 응답하였다. 이러한 결과를 두고, 저자들은 HARKing 행위의 원인이 통계적으로 유의한 결과가 자신들에게 '해피엔딩'을 가져다줄 것이라는 믿음이라고 설명하고 있다. 즉, 낮은 p-value가 출판에서의 유리함을 가져올 수 있다는 이점이 있으므로 연구자들이 다양한 형태로 HARKing을 한다는 것이다. 이러

한 연구결과에 근거할 때, 연구자는 자료로부터 도출된 다양한 결과의 집합을 손에 들고 어떤 결과를 최종 연구물로 해야 할 것인가 하는 선택의 상황에서 강력한 유연성을 가지게 된다. 이러한 상황에서 연구자에 의해 최종적으로 선택되는 결과가 무엇인가에 관해서는 본 연구가 보여주는 비전형적인 p-curve가 대신 답하고 있다고 할 수 있다.

최근에는 연구자들이 연구에 사용하도록 공개된 양적 자료들이 많아지고, 설문 조사 대행업체로 인해 표본의 수집도 편리해졌다. 이로 인해 한 연구에서 선택 가능한 변수의 수도 증가하게 되었지만, 그럴수록 연구자들은 연구모형설정 전반에 걸쳐 변수 간 관계에 관한 선행연구 및 관련 이론들을 철저하게 고찰해야 할 필요성이 있다. 연구자들이 사용하는 분석소프트웨어에서는 하나의 변수가 모형에 추가되고 제외되는 것이 단 한 번의 클릭으로 이루어지는 손쉬운 일이지만, 실제로는 변수의 추가나 제외가 모형 내 다른 변수의 추정치(estimate)를 변화시킬 수 있는 팽창효과(spurious effect) 및 억제효과(suppression effect)의 주요 원인이 될 수 있기 때문이다(Lewis-Beck, Bryman, & Liao, 2003; MacKinnon, Krull, & Lockwood, 2000).

표본추출방법

다음으로, 115개의 연구물에서 보고하고 있는 표본추출방법에 대한 빈도분석을 실시하였다. 이에 편의표본추출법을 활용한 연구가 94개(81.7%)로 나타났다. 표본수집이 따라 필요하지 않은 패널데이터나 2차 자료를 활용한 연구가 7개(6.0%)로 나타났으며, 눈덩이표집법, 판단표집법, 집락표집법을 사용한 연구는 매우 드물게 나타났다.

연구자의 주된 관심인 모집단의 모수를 추정하기 위해 사용되는 표본의 통계량이 정확성을 갖기 위해서는 표본이 모집단을 충분히 대표할 수 있어야 한다(Levy & Lemeshow, 2013). 표본의 대표성은 무작위 표본추출(random sampling)을 통해서 가능하다. 이론적으로 무작위 표본추출은 모집단의 모든 구성원이 표본으로 선택될 확률이 모두 동일함을 의미한다. 하지만 일반적인 사회과학연구의 모집단의 크기는 가늠할 수 없을 정도로 크기 때문에 연구자가 모든 구성원들에게 표본으로 선택될 확률을 동일하게 부여하는 것은 불가능에 가깝다. 더욱이 모집단의 구성원 모두를 파악하는 것이 설령 가능하다 할지라도 시간과 예산이라는 현실적인 벽이 너무나 높다. 그렇기에 본 연구에서 나타난 결과와 같이 많은 연구자들이 비확률 표본 추출법 중에서 편의표본추출법(convenience sampling method)을 활용하여 표본을 수집하고 있다. 이러한 결과는 2006년부터 2008년까지 '한국체육학회지 인문 및 사회과학 분야'에 개재된 약 360편의 연구물의 표본추출방법을 분석한 Jeong & Kim(2009)의 연구에서도 동일하게 나타나고 있다. 편의표본추출법은 연구자의 편의(便宜)로 표본을 수집하기 때문에 표본이 편의(偏倚)될 가능성이 크다는 본질적 위험성을 가지고 있다. 즉, 편의표본추출법을 통해 수집된 표본은 모집단을 대표할 수 없을 가능성이 크기 때문에 추정치의 타당성과 신뢰성을 보장할 수 없다. 즉, 편의표본추출법을 적용한 연구의 결과는 'garbage in, garbage out', 즉, 무가치한 자료가 들어가면 무가치한 결과치가 도출된다는 비판에서 자유로울 수 없다.

그렇다고 해서 편의표본추출법을 사용한 연구의 결과들을 모두 부정적으로 바라볼 수 없는 것이 앞서 언급한 대로 무작위 표본추출이 현실적으로 불가능하기 때문이다. 표집오차(sampling error)가 감소 되는 층화 표집(stratified sampling)이나 군집 표집(clustering sampling)으로 획득된 표본이 무작위표본에 가깝기는 하지만, 모집단에 대한 특성에 대한 사전지식이 필요하므로 이러한 표집 방법의 수행 역시 개별연구자에게 쉽지 않은 일이다. 이에 '무작위'의 현실적인 의미를 연구자들이 새롭게 새길 필요가 있다고 하겠다. 당연히 '무작위'의 현실적인 의미는 편향성이 적은 표본을 확보하려는 연구자의 노력과 맞물려 있다고 하겠다. 즉, 손쉽게 구할 수 있는 표본(예: 자신이 재학 중인 대학에서 교양체육에 참여하는 학생들)보다는 편향성을 줄이기 위해 시간과 노력이 소비되는 표본(예: 서울역 광장의 행인들)을 구하기 위해 노력할 필요가 있다. 나아가 연구자들은 자신의 표본이 가진 한계점을 인정하여 도출된 연구결과를 단정적인 언어를 통해 표현하는 것을 지양해야 할 것이다. 우리가 수행하는 추리통계는 언제나 불확실성(uncertainty)을 포함하기 때문이다(Kim & Lee, 2018; Krzywinski

& Altman, 2013; Kuhberger, Fritz, Lermer, & Scherndl, 2015; Nevill, Holder, & Cooper, 2007).

결과 서술

마지막으로, 도출된 연구결과를 바탕으로 결과를 서술하는 부분에서도 문제점이 발견되었다. 다음은 연구결과를 기술하는 데 흔히 사용되는 문장의 예시이며, 본 연구에서 자료로 사용된 115편의 연구물에서 일부 발췌하였다.

'…… 유의한 영향을 미치는 것으로 나타나 가설 1과 2는 모두 채택되었다.'

'…… 이러한 결과로 가설 1, 2, 3, 4는 채택되었다.'

'…… 통계적 유의성이 검증되지 않아 가설 1은 기각되었다.'

이와 같은 채택(accept)과 기각(reject)이라는 이분법적 의사결정으로 인해 채택된 가설은 변수 간 인과적 영향력 있는 것으로, 기각된 가설은 변수 간 인과적 영향력이 없는 것으로 해석하는 경우가 다수 발견되었다. 본질적으로, 채택과 기각의 대상이 되는 것은 연구자의 연구가설(대립가설)이 아닌 영가설이다. 즉, 영가설이 연구자가 설정한 유의수준에 근거해 기각될 수는 있어도 채택된다는 말은 존재할 수 없으며 대립가설은 아예 통계적 검정이 불가능하다. 그런데도, 많은 연구자들은 자신들이 믿고 있는, 아니 관습적으로 믿어온 'p<.05'에 근거해 영가설이 기각된 사실이 대립가설을 채택하는 강력한 근거라 믿고 자신이 잠정적으로 설정한 인과관계가 자료에 의해 증명(prove)되었다고 주장한다. 이는 아래와 같이 많은 연구에서 '가설검정'을 '가설검증'으로, '유의'를 '유의미'로 잘못 표기한 것과 절대 무관하지 않을 것이다.

'…… H4를 검증한 결과, 관계몰입은 관계 만족에 영향을 미치지 않는 것으로 나타났다.'

'…… 연구가설의 검증을 위해 구조방정식 모형 분석을 실시하였다. …… 연구가설의 검증결과, …… 모기업 이미지에 유의미한 영향을 미치는 것으로 나타나 …… '

'…… 가설 6을 검증한 결과, 구단 평판은 구매 의도에 유의미한 영향을 미치는 것으로 나타나 …… '

이러한 잘못된 믿음은 p-value가 .051이 나온 상황을 가정하면 쉽게 무너지게 된다. 영가설이 옳다고 가정한 표집분포(sampling distribution)에서 검정 통계량이 얻어질 확률이 .051이면 영가설을 기각할 충분한 근거가 없는 것이고, .049면 영가설을 기각할 충분한 근거가 있다고 생각하는 것이 얼마나 위험한 생각인가? 매우 극단적인 예를 들었지만, 유의 수준 .05를 절대적인 기준으로 믿는 것과 고작 0.002%의 확률 차이를 통해 인과관계의 성립 여부를 해석하는 오류는 본질적으로 다르지 않다.

따라서 연구자들은 가설검정에 의한 인과관계를 주장함에 있어서 주의를 기울일 필요가 있다. 완벽히 설계된 무작위 실험설계를 제외한 어떠한 연구방법도 명확한 인과관계를 밝힐 수 없다. 이는 스포츠 경영학 분야에서 빈번하게 사용되고 있는 구조방정식모형에서 받아들일 만한(acceptable) 적합도가 도출된 것이 새로운 인과관계 모형을 증명한 것이 아닌 설정된 모형이 자료를 '그럴듯(plausible)하게 설명'한다고 해석해야 하는 것도 같은 맥락에서 이해할 수 있다(Kline, 2015, Loehlin, 1987). 이에 Kim and Lee(2018)의 연구에서는 자료가 가설과 일치(the evidence is consistent with the hypothesis) 혹은 자료가 가설을 지지(the evidence supports the hypothesis)와 같은 표현을 사용해야 함을 언급하고 있다. 요약하자면, 연구자에 의해 잠정적으로 설정된 인과관계를 주장하기 위해서는 통계적 가설검정을 통해 도출된 p-value가 아닌 명확한 이론적 근거 및 인과관계와 관련된 기본적인 충족조건(시간적 선행, 상관, 혼란 변수의 통제, 등)들을 종합적으로 고려해야 한다(Mulaik, 2009).

대안

'The Earth is Round (p<.05).' 라는 의미심장한 제목으로 NHST에 대해 비판을 시작한 Cohen(1994)의 연구를 필두로 많은 연구자들이 그의 견해를 지지하거나 보완한 연구(Amrhein, Korner-Nievergelt, & Roth, 2017; Guthery, Lusk, & Peterson, 2001; Lew, 2012; Reinhart, 2015)를 수행하였다. 이러한 학문적 흐름에

발맞추어, Basic and Applied Social Psychology라는 저널에서는 아예 NHST를 활용한 연구물을 거절한다는 공식적인 입장을 발표하였다.

앞선 분석에서 나타난 비전형적인 p-curve의 형태, 지나친 편의표본추출법의 사용빈도, 및 적절하지 못한 결과 서술법은 간접적으로 스포츠 경영학 분야의 연구자들이 NHST에 대한 정확한 이해가 부족함을 보여주고 있다고 하겠다. 설령, 이러한 간접적 근거가 실증적 설득력이 부족하고 현실을 제대로 반영하지 못한 우연한 결과라 할지라도, NHST에 대한 비판적 시각을 인정하고 여러 대안을 통해 실제 연구에 반영하려는 시도가 나타나고 있다는 점은 거부할 수 없는 분명한 사실이다.

NHST에 대한 대안으로 제시한 사항들을 총체적으로 고찰해보았을 때, 실질적으로 스포츠 경영학 분야에 적용될 만한 대안들은 신뢰구간(confidence interval)의 보고, 효과 크기(effect size)의 제시 및 메타분석(meta-analysis)의 활용으로 요약될 수 있다. 하지만, 본 연구에서 사용된 115편의 연구물 중 추정치의 신뢰구간이 보고되고 있는 연구물은 7편(6.9%), 효과 크기 제시한 연구물은 1편(0.8%)에 불과하였다. 특히, 신뢰구간을 보고한 7편의 연구들을 세부적으로 살펴보면, 연구 내 모든 가설에 대한 신뢰구간을 보고한 것이 아니라, 무작위 복원추출(bootstrapping) 방법을 활용한 매개 효과 분석에서만 신뢰구간을 보고한다는 특징을 가지고 있었다. 이렇듯, 가설검정의 단점을 보완하는 대안들이 더는 새로운 방법이 아님에도 불구하고, 스포츠 경영학 분야의 연구자들이 이를 적극적으로 활용되지 않고 있다는 사실은 그 대안들의 활용법과 이론적 토대가 연구자들에게 널리 알려지지 않았기 때문일 것이다. 이에 앞서 언급한 세 가지 대안들에 활용과 이론적 근거에 대해 다음과 같이 논의하였다.

첫째, 신뢰구간의 보고는 연구자가 궁극적으로 알고자 하는 모수가 존재할 구간의 범위를 추정하는 방법으로 일반적으로 95% 신뢰구간이 사용된다. 예를 들어, 95% 신뢰구간 하한계(lower limit) 및 상한계(upper limit)를 제시하는 것은 p-value만 제시하여 연구자에게 이분법적 판단을 강제하는 NHST의 단점을 보완할 수 있을 것이다. 다시 말해, 기존의 p-value를 중심으로 한 보고방식보다 신뢰구간을 함께 제시하는 방식이 모수에 대해 더욱 풍부한 정보를 제공하는 방법이다. p-value로부터 알 수 있는 정보는 영가설의 기각 여부뿐이지만, 신뢰구간은 추정치에 대한 신뢰성 및 추정치에 대한 오차도 포함하므로, 실제 여러 학자들이 신뢰구간의 보고를 권장하고 있다(Nakagawa & Cuthill, 2007). 이에 스포츠 경영학 분야의 연구자들도 추정치, 검정 통계량, p-value로 정형화된 가설검정 결과 보고방식에서 벗어나, 더욱 풍부한 정보를 제공할 수 있는 신뢰구간을 자신의 연구에 보고할 필요성이 있다.

둘째, 효과 크기는 연구결과가 실제 모집단에 존재할 가능성의 정도 혹은 비교하려는 집단 간의 차이 및 관계에 대한 표준화된 지표로 정의된다(Cohen, 1988). 연구결과 서술시 효과 크기를 보고해야 하는 이유는 본 연구의 서론에서 언급한 대로 p-value가 표본 크기에 민감한 수치이기 때문이다. 가령, 두 집단 간 평균 차이가 매우 작은 수준(혹은 두 변수의 간의 상관관계가 매우 작은 수준)이어도 표본 크기가 낮은 p-value를 기대할 수 있고, 두 집단 간 평균 차이가 매우 크더라도(혹은 두 변수 간의 상관관계가 매우 큰 수준) 표본 크기가 작다면 통계적 유의성이 발견되지 않을 수 있다. 이와 달리, 효과 크기는 표본 크기와 상호 독립적이며, 통계적 유의성이 아닌 실제적 유의성을 판단하는데 유용한 지표이다(Cooper, 1981). 실제로, 미국심리학회(American Psychological Association: APA)에서는 개별연구물의 연구결과를 보고할 때 효과 크기를 제시할 것을 권고하고 있다(Wilkinson & The Task Force on Statistical Inference, 1999). 즉, 스포츠 경영학 분야의 연구물 대부분이 변수 간 상관에 근거한 회귀분석 및 구조방정식이라는 점에 상기해 본다면 다양한 효과 크기 중에서 아래의 〈Table 2〉에 제시된 효과 크기와 그 해석기준에 대해서 연구자들이 숙지하고 연구결과 보고할 때 활용할 필요성이 있다(Nahm, 2015).

Table 2. Effect size and criteria for interpretation

	criteria for interpretation		
	Small	Medium	Large
Correlation (ρ)	0.1	0.3	0.5
Regression ($f^2=R^2/1-R^2$)	0.02	0.15	0.35

셋째, 메타분석은 개별연구물들의 연구결과를 종합하는 연구방법으로 최근 스포츠 경영학 분야에서도 그 출간빈도가 높아지고 있다. 메타분석은 개별연구물에서 보고되는 연구결과를 일치된 지표로 변환하여야 하므로 앞서 언급된 효과 크기와 긴밀히 연관되어 있다. 즉, 독립적으로 수행된 개별연구의 연구결과들을 체계적 기준에 의해 종합하여 더욱 정확한 효과 크기를 산출할 수 있으며, 표본 수의 충분한 확보로 통계적 검정력을 높일 수 있다는 장점을 가진다(Pillemer & Light, 1980). 또한, 개별연구물들의 특성(설계방법, 표본특성, 출간연도, 척도, 등)에 따른 효과 크기도 비교 가능해 다양한 수준에서 의미있는 정보를 제공할 수 있다. 무엇보다도 같은 주제를 다룬 개별연구들이 상반된 연구결과를 도출하였을 때, 이를 통계적으로 통합하여 더욱 일반화 가능성이 높은 연구결과를 도출할 수 있다는 관점에서 스포츠 경영학 분야의 연구자들도 메타분석을 적극적으로 활용할 필요성이 있다고 하겠다.

하지만, NHST에 단점과 그 대안들의 장점을 인정하면서도 앞서 언급된 세 가지 대안들에도 피할 수 없는 한계점이 존재한다는 견해를 피력한 연구(Cortina & Landis, 2011; Hubbard, Parsa, & Luthy, 1997; Nickerson, 2000)도 꾸준히 출간되고 있다. 예를 들면, 신뢰구간의 보고방법은 여전히 1종 오류를 포함하고 있고, 분포에 대한 가정이 먼저 충족되어야 하는 한계점이 있으며(Attia, 2005), 지금처럼 개별연구들이 NHST를 적극적으로 사용하여 p-value 중심의 연구결과를 도출하여도 추후 메타분석을 통해 이들을 종합할 수 있으므로 큰 문제가 없다는 주장(Robinson & Wainer, 2001) 있다. 무엇보다도, Cohen(1994)의 연구 제목을 인용한 Zhu(2012)의 'Sadly, The Earth is Still Round ($p<0.05$)'에서는 이미 보편적으로 사용되는 p-value가 제공하는 직관적 이해에 관한 효용성 및 범용성이 p-value로부터 파생되는 여러 가지 문제점보다 크다고 주장하였다.

이렇듯, 여전히 NHST의 사용 여부를 두고 벌이는 학자들의 격렬한 논쟁 속에서 스포츠 경영학 분야의 학자들은 어떤 관점을 취해야 할까? 앞선 분석의 결과와 그동안 제시된 세 가지 대안을 종합적으로 고찰해보았을 때, 그동안 널리 활용되어 온 NHST는 장단점을 동시에 가지고 있으나, 최근 그 단점이 더욱 주목받게 된 것은 NHST에 대한 연구자들의 부족한 이해와 잘못된 사용에서 비롯되었다고 할 수 있다. 이에, NHST에 대한 연구자들의 명확한 이해가 선행되어야 하며, NHST에 대한 비판적 시각이 강화되는 최근에 추세를 고려해 볼 때, 스포츠 경영학 분야의 연구자들도 많은 연구에서 대안으로 제시되고 있는 신뢰구간의 보고, 효과 크기의 보고, 메타분석을 적극적으로 연구에 활용할 필요가 있다고 하겠다.

결론 및 제언

본 연구에서는 스포츠 경영학 분야에서 출간된 연구물을 중심으로 통계적 가설검정 방법의 사용과 해석에 대해 비판해보고 선행연구들에서 제시된 대안 중에서 현실적으로 적절하다고 여겨지는 대안을 제시하였다. 요약하면, NHST의 단점은 방법 자체의 결함보다는 NHST를 통한 분석 결과가 의미하는 바를 올바르게 이해하지 못하고, 분석 결과에 관한 해석에서도 NHST가 가지는 본질적 한계점을 고려하지 않는 연구자들의 관행에 근거한다고 볼 수 있다. 여러 연구에서 지적한 대로 NHST의 오용과 남용은 많은 폐해를 가져왔으며 스포츠 경영학 분야의 연구물도 이와 같은 비판에서 완전히 자유로울 수 없다. 특히, 영가설의 기각 여부가 곧 연구 목적달성의 기준으로 오해하는 관습적인 태도는 반드시 지양해야 한다. 하지만 오래된 관습을 단 한 번에 바꾼다는 것은 불가능하며, 신속하고 간편한 의사결정을 위한 객관적인 기준 역시 학문적 기여도가 있다는 점을 고려한다면, NHST의 사용 자체를 비판하는 것은 현실적이지 못하다. 다만 NHST를 통해 도출되는 p-value를 연구자들의 의사결정을 위해 기계적으로 사용하는 것이 아닌 연구결과의 올바른 해석을 위해 다음과 같이 주의하여 사용되어야 할 것이다.

첫째, 영가설에 대한 유의성 검정의 한계점을 명확하게 이해하고 연구결과를 해석할 필요가 있다. 영가설에 대한 통계적 검정은 연구자에 의해 설정된 영가설이 옳다고 가정하였을 때, 주어진 데이터에서 도출된 통계량이 나타날 확률을 의미한다. 즉, 영가설의 기각이 변수 간 관계가 있음을 증명하는 것이 아니라 단지 변수 간 관계가

‘0’일 확률이 적다는 것을 의미한다. 이는 변수 간 관계가 ‘0’일 확률도 일부 존재한다는 것임에도 불구하고 대다수 연구자들은 변수 간 관계에 관한 새로운 발견을 하였다는 듯이 확신에 찬 논의를 서술하곤 한다. 더욱이 모수가 ‘0’이라고 설정하는 영가설의 비현실성(집단 간 차이가 전혀 없거나, 변수 간 상관관계가 전혀 없다는 가정)을 고려해보면 영가설의 기각이 대립가설의 증명으로 이어지는 해석은 지나친 논리적 비약임이 틀림없다. 이에 스포츠 경영학 분야의 연구자들은 p-value에 대한 이분법적 해석을 지양하고, 통계적 유의성과 실제적 유의성을 분리하여 자신의 연구결과를 해석해야 하며, 그 해석의 오류 가능성도 일부 고려하는 논의를 전개할 필요성이 있다.

둘째, 연구자들은 영가설의 기각실패가 연구의 실패가 아닌 연구결과의 풍부한 해석의 기회로 생각을 전환할 필요가 있다. 앞서 언급하였듯이 영가설의 기각실패는 단순히 영가설이 옳다는 비현실적 가정에 근거하여 연구자의 데이터로부터 나온 검정 통계량이 도출될 확률이 그리 낮지 않을 뿐이다. 이에 추정치의 신뢰구간, 효과 크기의 보고를 통해 p-value를 통한 보고방식의 단점을 보완할 수 있을 것이다. 또한, 영가설 기각의 실패는 다양한 원인에 의해 발생 될 수 있으므로, 후속 연구자의 연구에도 학문적 도움이 될 수 있을 것이다. 최근 스포츠 경영학 분야에서 구조방정식을 활용한 연구가 증가하면서 한 연구에서 설정된 가설의 수가 증가하는 경향이 있다. 이에 통계적 유의성이 발견된 가설만을 해석하고 논의하는 연구물이 발견되기도 한다. 이는 결국 p-value에 의해 연구결과에 대한 논의가 취사선택된 것이므로 지양되어야 한다. 통계는 학자들의 생각을 돕는 하나의 도구일 뿐, 그 대체재가 될 수는 없다.

본 연구는 이상의 결론을 통해 스포츠 경영학 분야의 학자들에게 의미 있는 시사점을 도출하였으나 다음과 같은 제한점이 있다.

첫째, 연구의 실행 가능성을 고려하여 본 연구에서는 최근 5년간 한국스포츠산업경영학회지에 게재된 연구물만을 연구대상으로 선정하였다. 후속연구에서는 좀 더 많은 연구물, 다양한 분야의 연구물을 확보하여 연구를 수행한다면 좀 더 폭넓은 연구결과를 도출할 수 있을 것이다.

둘째, NHST에 대한 대안으로 최근 주목받고 있는 베이지안(bayesian) 방법에 대해서는 본 연구에서 다루지 못하였다. 베이지안 방법은 사후확률분포를 사전확률분포와 자료의 확률분포를 통해 도출하는 방법으로 기존의 빈도주의 통계학(frequentist statistics)과는 전혀 다른 관점을 가지고 있다. 이에 대해서 현실적인 사용방법에 대해 논의하기에는 본 연구자가 가진 사전지식이 충분하지 않았다.

참고문헌

Amrhein, V., Korner-Nievergelt, F., & Roth, T. (2017). The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. *PeerJ, 5,* e3544. doi:10.7717/peerj.3544.

Attia, A. (2005). Why should researchers report the confidence interval in modern research. *Middle East Fertility Society Journal, 10*(1), 78-81.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543-554.

Bishop, D. V., & Thompson, P. A. (2016). Problems in using p-curve analysis and text-mining to detect rate of p-hacking and evidential value. *PeerJ, 4*, e1715.

Bruns, S. B., & Ioannidis, J. P. (2016). P-curve and p-hacking in observational research. PloS one, 11(2), e0149144.

Clark, G. T., & Mulligan, R. (2011). Fifteen common mistakes encountered in clinical research. J*ournal of Prosthodontic Research, 55*(1), 1-6.

Cohen, J. (1988). The effect size index: d. *Statistical Power Analysis for the Behavioral Sciences, 2*, 284-288.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49*, 997-1003.

Cooper, H. M. (1981). On the significance of effects and the effects of significance. *Journal of Personality and Social Psychology, 41*(5), 1013-1018.

Cortina, J. M., & Landis, R. S. (2011). The earth is not round (p=.00). *Organizational Research Methods, 14*(2), 332-349.

Csada, R. D., James, P. C., & Espie, R. H. (1996). The “file drawer problem” of non-significant results: does it apply to biological research?. *Oikos, 76*(3), 591-593.

Dupont, W. D., & Plummer Jr, W. D. (1990). Power and sample size calculations: a review and computer program. *Controlled Clinical Trials, 11*(2), 116-128.

Fidler, F., Burgman, M. A., Cumming, G., Buttrose, R., & Thomason, N. (2006). Impact of criticism of null hypothesis significance testing on statistical reporting practices in conservation biology. *Conservation Biology, 20*(5), 1539-1544.

Guthery, F. S., Lusk, J. J., & Peterson, M. J. (2001). The fall of the null hypothesis: liabilities and opportunities. *The Journal of Wildlife Management, 65*(3), 379-384.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature Methods, 12*(3), 179.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature Methods, 12*(3), 179.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS biology, 13*(3), e1002106.

Hubbard, R., Parsa, R. A., & Luthy, M. R. (1997). The spread of statistical significance testing in psychology: The case of the Journal of Applied Psychology, 1917-1994. *Theory & Psychology, 7*(4), 545-554.

Hupé, J. M. (2015). Statistical inferences under the Null hypothesis: common mistakes and pitfalls in neuroimaging studies. *Frontiers in Neuroscience, 9*, 18.

Jeong, J. O., & Kim E. J. (2009). Trends analysis of sampling method through definition of population in the Korean Journal of Physical Education: 2006-2008. *The Korean Journal of Measurement and Evaluation in Physical Education and Sports Science, 11*(3), 35-53.

Jo, S. J., Bea, J. S., Jhun, M. S., Lee, J. W., & Park, S. G. (2009). Sample size computation for one-way analysis of variance. *Journal of The Korean Data Analysis Society. 11*(3), 1429-1441.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217.

Kerr, N. L., & Harris, S. E. (1998). *HARKing: hypothesizing after the results are known: Views from three disciplines.* Unpublished manuscript, Michigan State University, East Lansing.

Kim, H. R., Park, S. H., & Won, D. Y. (2015). An analysis of research trends of sport management: Co-author network and keyword network. *Korean Society for Sport Management, 20*(3), 63-84.

Kim, S. Y., & Seok, H. E. (2015). Determining sample size requirements in latent growth models. *The Korean Journal of Psychology: General, 34*(2), 599-617.

Kim, Y., & Lee, J. L. (2018). Common Mistakes in Statistical and Methodological Practices of Sport Management Research. *Measurement in Physical Education and Exercise Science*, doi: 10.1080/1091367X.2018.1537278.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Krzywinski, M., & Altman, N. (2013). Points of significance: Importance of being uncertain. *Nature Methods, 10*, 809–810.

Kühberger, A., Fritz, A., Lermer, E., & Scherndl, T. (2015). The significance fallacy in inferential statistics. *BMC Research Notes, 8*(1), 84.

Lai, J., Kalinowski, P., Fidler, F., & Cumming, G. (2010). *Dichotomous thinking: A problem beyond NHST.* Data and context in statistics education: Towards an evidence based society.

Lakens, D. (2015). Comment: What p-hacking really looks like: A comment on Masicampo and LaLande (2012). *Quarterly Journal of Experimental Psychology, 68*(4), 829-832.

Levy, P. S., & Lemeshow, S. (2013). *Sampling of populations: methods and applications*. John Wiley & Sons.

Lew, M. J. (2012). Bad statistical practice in pharmacology (and other basic biomedical disciplines): you probably don't know P. *British Journal of Pharmacology, 166*(5), 1559-1567.

Lewis-Beck, M., Bryman, A. E., & Liao, T. F. (2003). *The Sage encyclopedia of social science research methods.* Sage Publications.

Loehlin, J. C. (1987). *Latent variable models: An introduction to factor, path, and structural analysis.* Lawrence Erlbaum Associates, Inc.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological methods, 1*(2), 130.

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science, 1*(4), 173-181.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system.

Cognitive therapy and research, 1(2), 161-175.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica, 22*(3), 276-282.

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. Chapman and Hall/CRC.

Murphy, K. R., & Aguinis, H. (2019). HARKing: how badly can cherry-picking and question trolling produce bias in published results?. *Journal of business and psychology, 34*(1), 1-17.

Nahm, F. S. (2015). Understanding effect sizes. *Hanyang Medical Reviews, 35*(1), 40-43.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews, 82*(4), 591-605.

Nevill, A. M., Holder, R. L., & Cooper, S. M. (2007). Statistics, truth, and error reduction in sport and exercise sciences. *European Journal of Sport Science, 7*(1), 9-14.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 5*(2), 241.

Park, S. H., & Kwak, M. S. (2018). Methodological qualitative evaluation of meta-analysis studies in sport management. *Korean Journal of Physical Education, 57*(1), 247-258

Pillemer, D., & Light, R. (1980). Synthesizing outcomes: How to use research evidence from many studies. *Harvard Educational Review, 50*(2), 176-195.

Reinhart, A. (2015). *Statistics done wrong: The woefully complete guide*. No starch press

Robinson, D.H., & Wainer, H. (2001). *On the past and future of null hypothesis significance testing. Princeton*: Statistics & Re-search Division.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. *What if there were no significance tests*, 335-391.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. *What if there were no significance tests*, 37-64.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology, 144*(6), 1146-1152.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician, 49*(1), 108-112.

Stern, J. M., & Simes, R. J. (1997). Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ, 315*(7109), 640-645.

Ulrich, R., & Miller, J. (2015). p-hacking by post hoc selection with multiple opportunities: Detectability by skewness test?: Comment on Simonsohn, Nelson, and Simmons (2014). *Journal of Experimental Psychology: General, 144*(6), 1137-1145

Verdam, M. G., Oort, F. J., & Sprangers, M. A. (2014). Significance, truth and proof of p values: reminders about common misconceptions regarding null hypothesis significance testing. Quality of Life Research, 23(1), 5-7.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7*, 1832. doi: 10.3389/fpsyg.2016.01832.

Wilkinson, L., & The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Zhu, W. (2012). Sadly, the earth is still round ($p < 0.05$). *Journal of Sport and Health Science, 1*(1), 9-11.

스포츠 경영학 분야의 통계적 가설검정에 대한 비판과 대안

박상현
연세대학교, 강사

〔목적〕 본 연구는 스포츠 경영학 분야에서 출간된 연구물을 중심으로 통계적 가설검정 방법에 대한 비판점을 제시하고 이에 대한 현실적 대안을 제시하는데 근본적인 목적이 있다. **〔방법〕** 연구목적을 달성하기 위해 한국스포츠산업경영학회지 19권 1호부터 23권 6호까지 개재된 202편의 연구물 중 통계적 가설검정방법을 활용한 115편의 연구물이 최종적으로 선택되었다. 선택된 연구물에 대한 자료코딩 후, 개별연구물에서 보고되는 p-value의 분포인 p-curve를 도식화하고, 표본추출방법 및 결과 서술의 적절성에 대해 분석하였다. **〔결과〕** p-curve를 통해 p-hacking의 직접적인 근거는 도출되지 않았으나 0에 근접한 p-value의 비율이 상대적으로 매우 높게 나타났다. 또한, 약 82%의 연구물이 편의표본추출법을 활용하고 있었으며, 결과 서술에서도 통계적 가설검정에 대한 이해 부족으로 인한 잘못된 서술이 일부 연구에서 발견되었다. **〔결론〕** 결론적으로 학계에서 흔히 사용되는 통계적 가설검정의 단점은 방법 자체의 결함이 아닌 방법을 잘못 활용하는 연구자에게 있기에 통계적 가설검정에 대한 장단점을 연구자들이 숙지할 필요가 있으며, 신뢰구간 및 효과 크기를 대안으로 활용하여 기존의 p-value를 중심으로 한 결과 및 논의 서술방식의 단점을 보완할 필요가 있다.

주요어: 스포츠 경영학, 통계적 가설검정