한방안이비인후피부과학회지 제32권 제4호(2019년 11월)
J Korean Med Ophthalmol Otolaryngol Dermatol 2019;32(4):41-61
pISSN 1738-6640 eISSN 2234-4020
http://www.ood.or.kr
http://dx.doi.org/10.6114/jkood.2019.32.4.041

Original Article / 원저

임상연구방법론에서 귀무가설과 대립가설, 귀류법에 대한 고찰과 한방이비인후과에서 베이지안 통계학의 활용

남승표[1] · 배재민[1] · 권 강[2]
부산대학교 한의학전문대학원 ([1]대학원생)
부산대학교 한의학전문대학원 안이비인후피부과 ([2]교수)

A Study on Null Hypothesis and Alternative Hypothesis, Reduction to Absurdity and Application of Bayesian Statistics in Korean Medicine Otolaryngology

Seung-Pyo Nam[1] · Jae-Min Bae[1] · Kang Kwon[2]
[1]Pusan National University School of Korean Medicine
[2]Dep. of Korean Medicine Ophthalmology & Otolaryngology & Dermatology, Pusan National University School of Korean Medicine

Abstract

Background : The current medical statistics used in clinical research are the results of Fisher's significance test and the Neyman-Pearson hypothesis test, which were combined by psychologists. Also, in the philosophical background, it is related to Popper's falsificationism based hypothesis-deductive method and reduction to absurdity.

Objectives : This study was designed to find complementary and alternative methods of null hypothesis and alternative hypothesis used for the clinical research methodology of Korean medicine otolaryngology.

Methods : The body of this paper was divided into seven part. These are historical background, hypothesis test, hypothesis test method used in the design of clinical study, falsificationism and reduction to absurdity, problem and alternative method of the Neyman-Pearson hypothesis test, diagnosis example of sinusitis differentiation syndromes by Bayesian statistics. Through this process, we found out problems of frequentist statistics and suggested alternative methods.

Result & Conclusion : As a solution to the problems of the null hypothesis and the alternative hypothesis, there are effects size, confidence interval, Bayesian statistics and Lakatos methodology of scientific research programmes.

Key words : Null and alternative hypothesis; Reduction to Absurdity; Bayesian Statistics; Lakatos Methodology of Scientific Research Programmes; Korean Medicine Otolaryngology

© 2019 the Society of Korean Medicine Ophthalmology & Otolaryngology & Dermatology
This is an Open Access journal distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/license/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Ⅰ. 서 론

통계학의 Statistics란 단어의 어원은 『Folks 1981』[1]에 의하면 라틴어인 status(국가를 의미하는 state)와 statista(정치인을 의미하는 stateman)에서 온 것이다. 특히 statista는 국가의 업무를 관장하는 사람을 일컫는 말이었다[1,2]. 통계학은 18세기의 영국의 정치산술학과 독일의 국세학에 뿌리를 두고 있으며, 초기 통계학은 국가나 민족을 경영하기 위하여 필요한 인구, 세금, 농지 등에 관한 자료를 효율적으로 구하고 분석하는 학문으로 시작하였다. 그러나 19세기 수학에서 확률론이 발전하면서 초기 통계학은 확률론과 합쳐져 수리통계학으로 발전하였다[2]. 20세기에 접어들면서 피어슨(K. Pearson)에 의하여 생물통계학의 전문학술지 『Biometrika』가 창간되고, 피셔(R. Fisher)에 의한 소표본 이론과 이에 근거한 추론과 실험계획법 등이 연구되면서 현대통계학이 정립되기 시작하였다. 오늘날 피어슨과 피셔를 현대통계학의 아버지라고 부르는 이유도 여기에 있다[2].

가설(hypothesis)은 연구자들이 변수 사이의 기대되는 관계를 분명하게 표현하는 방법이다[3]. 좋은 가설(연구가설)은 연구목적에 적합해야 하며, 개념(변수)들 간의 관계에 대해 진술하고 있어야 하고, 이들 개념(변수)들 간의 관계가 검정 가능해야 한다. 특히 검정 가능성이 가장 중요한 요건이다. 연구가설은 '영향을 미친다, 관계가 있다, 차이가 있다' 등과 같이 대립형태(alternative form)로 표현하는 경우가 일반적이지만 '영향을 미치지 않는다, 관계가 없다, 차이가 없다' 등과 같이 귀무형태(null form)로 표현하는 경우도 있다[4].

귀무가설 또는 영가설의 유의성 검정(null hypothesis significance test)은 피셔의 유의성 검정과 네이만(J. Neyman)과 피어슨(E. Pearson)의 가설검정이 혼합된 것이다. 피셔의 p값은 귀무가설을 기각하기 위한 기준으로 고안되었으며, 네이만과 피어슨(E. Pearson)은 연구가설과 귀무가설 간의 판단 기준으로서 1종 오류의 확률인 α를 사용하였다. 이 두 방법은 철학적 배경이 다르나, 점차 하나의 체계로 진화하였다. 그러나, 귀무가설 검정은 결코 완전하거나 최선의 방법이 아니며, 귀무가설 검정에 심각한 결함이 있다는 사실은 수십 년 전부터 많은 통계학자들에 의해 지적되어 왔다[5]. 그러나 귀무가설 역시 대립가설과 마찬가지로 모집단에 대한 것을 다루고 있으므로 그 타당성여부를 직접 증명하기란 쉬운 일이 아니다. 따라서 우리는 일단 귀무가설이 타당하다는 전제 하에 귀무가설의 타당하지 않음을 나타내 주는 구체적-경험적 증거를 제시함으로써 귀무가설이 타당하다고 가정한 전제(前提)를 뒤엎는 방법을 사용하게 된다. 이러한 논리는 마치 수학에서 어떤 명제가 거짓임을 증명하기 위해, 그 명제가 참이라는 전제로부터 명제가 거짓이 되는 구체적인 경우를 유도해냄으로써 원하는 증명에 도달하려는 '반증(反證)의 원리(principle of falsification)'와 유사하다[6].

한의학계에서 많은 임상연구가 귀무가설과 대립가설을 이용한 가설검정을 기반으로 이루어지고 있으나, 이 방법이 한의학적 패러다임에 적합한지, 또한 그 문제점 및 대안은 무엇인지에 대한 연구는 찾아보기 힘든 실정이다. 임상연구방법의 기반이 되는 귀무가설검정의 문제점과 그 대안을 찾는 논문은 현재까지 치의학계에서 이[5]의 연구가 있었으며, 한의학계에서는 아직 관련 연구를 찾아볼 수 없었다. 따라서 본 연구에서는 이[5]의 연구의 바탕 위에서, 현재 임상연구방법론에서 사용되는 가설검정방법의 장점과 단점을 알아보고 그 보완점에 대하여 살펴볼 예정이다.

Ⅱ. 본 론

1. 역사적 배경

Corresponding author : Kang Kwon, Dep. of Korean Medicine Ophthalmology & Otolaryngology & Dermatology, Pusan National University Korean Medicine Hospital, 20, Geumo-ro, Mulgeum-eup, Yangsan-si, Gyeongsangnam-do, Korea. (Tel : 055-360-5941, E-mail : hanny98@pusan.ac.kr)
• Received 2019/10/23 • Revised 2019/10/29 • Accepted 2019/11/5

1) 수리통계학

① 칼 피어슨(Karl Pearson)의 업적

생물측정학-멘델주의 논쟁은 다윈이 주장한 생물 진화의 연속성과 자연선택을 두고 1890년대 중반부터 약 10여 년 간 치열하게 진행되었던 논쟁이다[7]. 그 논쟁의 한쪽 편에는 칼 피어슨(K. Pearson)과 웰던(W. Weldon)을 중심으로 한 생물측정학자들이 있었고 그 반대편에는 베이트슨을 주축으로 한 멘델주의 유전학자들이 있었다[7]. 멘델주의자들이 무엇의 작용 때문에 유전이라는 결과가 일어나는지 그 생리학적인 인과관계를 보려했다면 생물측정학파는 유전에 관계되는 여러 변수들 사이의 상관관계를 찾으려 하였다[7]. 생물학 연구에 통계학이 필요한 이유는 진화 문제가 오로지 수명, 번식력, 건강 그리고 질병에 대해 다루는 동태통계의 문제이기 때문이다. 진화를 연구하는 사람으로서 이런 통계 없이 연구를 진행하기는 불가능하다[7,8].

적자생존이론의 근간이 되는 통계모형을 인식하고, 19세기 기계론적 세계관과는 다른 새로운 세계관을 이끌어 낸 사람은 칼 피어슨(K. Pearson)이었다[9]. 칼 피어슨은 우리가 관찰하는 것은 임의성이 개입된 그림자에 지나지 않으며, 실제로 존재하는 것은 확률분포라고 생각했다. 과학연구대상은 결코 보거나 만질 수 있는 것이 아니라 관찰되는 임의성을 나타내는 수학적 함수라는 것이다[9]. 특히, 피어슨은 비정규 분포에 관한 연구에 몰두하였다. 비대칭 도수 곡선을 두 개의 정규 곡선의 합으로 나누는 최초의 수리적 방법은 그의 업적 중에서도 가장 중요한 것이었다. 그는 미분 방정식을 이용하여 5개의 곡선군 또는 유형을 제시하였고, 자료의 적률을 이용하여 어느 유형의 곡선을 택해야 하는가를 설명하였다. 이들 유형별 곡선은 통계학을 보다 구체적인 상황과 결부시키게 한 유용한 이론적 결과였고, 실제로 20년 이상 통계학자들의 표준 장비가 되었다. 또한 대량 자료를 정리하는 방법에 주목하는 이른바 '기술 통계(descriptive statistics)' 영역의 대부분은 그가 이론화한 것이다[10,11]. 피어슨은 또한 물리학자들이 주장하는 인과관계란 그들의 필요에 따라 가정하는 순전히 이론적인 극한, 즉 완전한 상관관계일 뿐이라고 주장했다. 그러므로 이전까지 생물, 경제, 사회 현상에 대한 연구가 엄밀한 인과적인 법칙에 따르는 물리학과 비교할 때 과학의 반열에 오르지 못했던 것도 상관관계의 강도가 물리학에서의 강도보다 낮기 때문이므로 과학의 기준을 인과관계가 아닌 상관관계로 바꾼다면 그들도 모두 엄연한 과학이 될 수 있다는 것이다[7].

② 로널드 피셔(Ronald Fisher)의 업적

추측통계학을 바탕으로 하는 현대통계학(modern statistics)은 영국의 피셔(Ronald A. Fisher; 1890~1962)에 의하여 정립되기 시작하였다. 피셔는 그가 1919년부터 일하던 로담스테드 실험연구소(Rothamsted experimental station)에서 농업시험에 통계적 방법을 적용시키는 연구를 계속하여 F-분포를 비롯하여 표본상관계수, 표본회귀계수 등의 많은 통계량의 분포를 유도하여 소표본에 기초를 둔 추론법(inferential method)을 확립하였으며, 분산분석(analysis of variance)법을 창시하여 현대통계학의 기초를 정립하였다[2]. 특히 그는 실험의 계획과 분석에 대한 근대적인 방법들을 발전시켰으며, 불멸의 저서로 『The Design of Experiments』[12]를 출간하여, 오늘날 실험계획법(design of experiments)의 창시자가 되었다. 피어슨(K. Pearson)과 피셔(R. Fisher)를 현대통계학의 아버지라고 지칭할 수 있고, 이들로 인하여 영국이 현대 통계학의 종주국이 된 것이다[2]. 현재 널리 사용되는 대부분의 유의성검정법은 피셔가 개발하였다. 피셔는 유의성을 판단하기 위해 계산하는 확률을 p값이라고 불렀는데, 그는 p값의 의미와 유용성에 대해 확신하고 있었다. 그는 저서 『연구자를 위한 통계적 방법』의 상당 부분을

p값 계산에 할애했다[9]. P값이 어느 정도일 때 유의적이라고 할 것인지에 대해 피셔가 근접하게나마 언급한 것은 1929년 『심령연구학회논문집』에 발표한 논문에서였다[9]. 피셔의 응용 논문을 자세히 읽어보면, 피셔는 유의성검정으로 세 가지 결론을 내린다는 것을 알 수 있다. 만약 p값이 아주 작으면(보통 0.01보다 작으면) 효과가 드러났다는 결론을 내린다. 만약 p값이 크면(보통 0.2보다 크면) 효과가 있더라도 이 정도 규모의 실험으로는 발견할 수 없을 정도로 효과가 작다고 결론을 내린다. 만약 p값이 그 중간이면 다음 실험을 어떻게 하면 효과를 좀 더 잘 알아낼 수 있을까에 대해 생각한다[9].

주어진 자료에서 가장 정확한 추정치를 구하는 방법, 소표본 관찰에 가장 적합한 방법 등은 피어슨 이론에서 간과되거나 완전하게 해결되지 못했고 피셔에 이르러서야 확고한 분석이 이루어진 문제였다. 특히 피셔는 대수적인 방법으로 다루기 어려운 표본에 대해서는 유클리드 공간상의 점에 대응시킴으로써 정확한 분포를 추정해 낼 수 있었다[11]. 피셔가 분산분석과 공분산분석을 위해 개발한 알고리즘은 정말 놀라운 수학적 성과다. 이 알고리즘은 두 분석법에서 추정해야 할 모수의 최대가능도추정량을 구하는 것으로 다차원 공간에서의 치환과 변환을 이용한 것이다[9]. 농학자들은 피셔의 연구를 높이 평가했고 얼마 지나지 않아서 피셔가 제안한 실험설계는 영어권 국가의 농업분야에서 널리 사용되었다. 피셔의 초기 연구를 시작으로 여러 가지 실험설계법이 개발되어 과학문헌에 소개되었으며, 농학 외에도 의학, 화학, 기업체의 품질관리 등에도 적용되었다[9].

③ 예르지 네이만(Jerzy Neyman)과 이곤 피어슨(Egon Pearson)의 업적

현대 통계학의 발전에 기여가 큰 통계학자로는 폴란드인 네이만(Jerzy Neyman;1894~1981)이 있다. 그가 영국에 있는 유니버시티 칼리지 런던(University College London)으로 유학하는 동안 이곤 피어슨(E. Pearson)과 10년(1928~1938) 간 함께 공동연구를 하였으며, 통계적 가설검정에 관한 10개의 논문(예로, 네이만-피어슨 정리)을 발표하여 가설검정 이론을 확고히 정립하였다. 1938년에 네이만은 미국 버클리 대학으로 건너가 통계연구소를 세우고, 세계 각지로부터 통계학자들을 초청하여 1945년부터 'Berkeley Symposia on Mathematical Statistics and Probability'를 매 5년마다 6회에 걸쳐 주관하여 현대 통계학의 발전에 기여하고, 통계학 연구가 영국에서 미국으로 옮겨가는 기폭제 역할을 하였다. 그의 연구업적은 이론 통계학의 근간이 될 뿐만 아니라, 농학, 천문학, 생물학, 기상학 등을 포함한 여러 분야에 있어서 통계학의 광범위한 적용을 제시한 것이다[2].

이곤 피어슨과 네이만이 공동 연구를 시작할 무렵 이곤 피어슨은 네이만에게 유의적인 p값이 나오지 않았다고 해서 자료에 정규분포가 적합하다는 것을 확신할 수 있느냐고 물었다. 그들의 연구는 바로 이 질문에서 시작했다. 하지만 이 질문은 좀 더 광범위한 문제에 대한 시작에 불과했다. 여기서 광범위한 문제란 바로 다음과 같은 질문이다. 유의성검정결과가 유의적이지 않다는 것은 어떤 의미인가? 한 가설을 부정할 수 없다고 그 가설이 옳다는 결론을 내릴 수 있는가?[9]

네이만은 당시에 대표적인 표본추출 방법이었던 유의추출법(purposive sampling)과 랜덤추출법을 사용하지 않고, 신뢰 구간을 이용하여 더욱 정확한 추출 방법을 개발하였다. 이 연구 결과는 현대 표본조사 이론에서 네이만 배분(Neyman allocation)으로 알려져 있다[11]. 네이만과 이곤 피어슨이 정립한 가설검정법은 0.05 같은 고정된 기준값(유의수준이라고 한다)을 설정하고, p값이 이 기준값 이하가 되면 귀무가설을 기각한다. 이렇게 가설검정을 하면 과학자들은 검정하는 횟수

의 5퍼센트에서 옳은 귀무가설을 기각하게 된다[9]. 네이만과 이곤 피어슨이 확립한 가설검정법을 단순화한 방법은 모든 기초통계학 책에 실려 있다. 단순화된 가설검정방법은 거의 성문화되었기 때문에 반드시 그렇게 해야 하고 그것이 유일한 방법인 것처럼 강의되고 있다[9].

④ 피셔와 네이만-피어슨의 관점 차이

네이만과 이곤 피어슨이 정립한 가설검정법이 통계학의 정상에 오르는 과정이 순탄하지만은 않았다. 피셔는 그들이 연구를 시작할 때부터 비판했으며, 사망할 때까지 비판을 멈추지 않았다[9]. 시간이 갈수록 네이만과 이곤 피어슨이 정립한 가설검정법은 교과서에 뿌리를 내렸지만 그 결점을 지적하는 논문도 끊임없이 발표되고 있다[9].

피셔는 자신이 개발한 '유의성검정(significance test)'을, 네이만은 '가설검정(hypothesis test)' 방법을 옹호했는데, 두 방법은 대립가설 유무, 검정결과의 해석 등에서 차이가 있었다. 피셔는 자신의 검정법이야말로 통계적 추론을 통한 귀납적인 과학 연구 방법이라고 주장했던 반면, 네이만과 이곤 피어슨은 대립가설을 설정하는 자신들의 방법이 피셔의 방법을 대폭 개선한 것일 뿐 아니라 분명한 선택을 가능하게 해준다고 주장했다[13]. 오늘날 p값이라고 불리는 것은 피셔의 것인데 그는 p값을 가설들의 채택, 기각을 가르는 기준으로 삼기보다는 가설을 반증(disprove)할 수 있는 통계적 근거(statistical evidence)의 척도로 생각했다. 즉 그는 검정법을 두 가설 중 하나를 선택하는 최종적인 결정의 방법이 아니라 데이터로부터 추론해가는 일련의 과정 가운데 한 단계 정도로, 다시 말해 타당하지 못한 귀무가설을 하나씩 버려나가는 귀납적 과정이 과학적 연구방법이라고 생각하였다. 또한 피셔는 관습적으로 유의수준 0.05를 기준으로 귀무가설을 기각하거나 채택하는 것에 대해서도 통계학이라는 것이 그렇게 단순한 논리로 이루어진 것이 아니라면서 반대하였다[14].

네이만은 피셔의 업적을 실험계획론, 분포 이론, 통계 기초 이론의 세 분야로 나누어 평가했는데, 그 가운데 앞의 두 가지에 대해서는 좋은 평가를 하였지만 통계 기초 이론에 대해서는 입장을 달리하였다. 피셔가 통계학을 귀납적 사고 또는 귀납적 행동으로 설명하려고 한 반면에 네이만은 통계적 사고의 이면에는 여전히 연역적 추론이 강하게 들어 있음에 주목하였다[11].

2) 심리학연구에서 통계학의 적용

근대 이후 흔히 뉴턴주의 물리학이라 일컬어지는 결정론적인, 또는 법칙정립적인 과학관이 널리 받아들여지면서 통계학은 오랫동안 천문학의 그늘을 벗어나지 못했다[14]. 심리학은 그런 분위기 속에서 가장 앞서 통계학적 방법을 받아들인 분야라고 할 수 있다[14].

지거렌저(Gerd Gigerenzer)는 20세기 심리학자들이 통계적 검정법을 그처럼 적극적으로 활용한 이유는 그들이 물리학에서와 마찬가지로 심리학에서도 객관성(objectivity)과 결정론(determinism)을 추구한 때문이라고 주장했다[14-17]. 그런데 지거렌저의 연구 가운데 우리가 주목할 것은 그처럼 통계학적 검정법을 널리 활용한 것이 심리학에 득이 된 것만은 아니라고 평가하는 점이다[14]. 앞서 살펴보았듯이 피셔의 유의성검정과 네이만-피어슨의 가설검정은 수십 년 동안 서로 대립해온 서로 다른 검정법이었는데도 심리학자들은 유일한 방법만 존재하는 것처럼 간주하고 검정법을 이용했다는 것이다. 무엇보다 지거렌저가 강조해서 문제 삼는 것은 교과서에 실리는 검정 방법이 두 가지 대립되는 방법 가운데 하나를 택한 것이 아니라 각 방법을 만든 당사자들이 다르다고 주장했던 차이점을 무시한 채 양쪽을 버무려 만든 잡종검정법이라는 점이다[14].

그가 잡종검정법이라고 일컫는 것은 오늘날 여러 통계학 교과서에도 흔히 볼 수 있는 것으로서 그 설명 순서는 대략 다음과 같이 요약해볼 수 있겠다. 어떤 교과서에는 이 방법을 피셔가 만든 유의성검정법이라고 부르고 또 어떤 책에서는 네이만과 피어슨이 만든 가설검

정법이라고 부르고 있다[14]. 그 순서를 소개하면 다음과 같다. 첫째, 귀무가설 H_0와 대립가설 H_1을 설명한다. 둘째, 제1종 오류와 제2종 오류(때로는 검정력까지 포함하여)에 대해 설명한 다음 유의수준(α)을 설명한다. 셋째, 적절한 검정통계량의 값을 데이터로부터 얻고 그 값에 대한 p값을 구한다. 넷째, p값과 유의수준 α값을 비교하여 p값〈α이면 H_0를 기각한다. 또 그 반대이면 H_0를 채택하거나 H_0를 기각하지 않는다.

3) 의학연구에서 통계학의 적용

19세기 초,중반까지 의학자들은 통계란 다수의 환자들에 대한 것이므로 환자 개인의 세세한 특성은 모두 제거되었다고 생각했다. 그러므로 그런 통계 자료에 바탕을 둔 추론으로부터는 특정 환자의 치료에 도움이 되는 결과가 나올 수 없다고 여겼다. 케틀레의 사회물리학에서 보편적인 사실에 밀려나야했던 개인적인 특성이 역설적이게도 의학에서는 가장 중요한 요인으로 생각되었기 때문에 의학자들은 통계학적인 추론을 의학에 적용하는 것에 반대했던 것이다[18].

19세기 말 이후부터 20세기 전반기까지 활동한 영국의 골턴, 피어슨, 피셔와 같은 거인들 덕분에 20세기 통계학은 이론과 방법을 갖춘 과학으로서 대학에도 학과로 자리 잡기 시작했다[13]. 그렇게 시작된 20세기는 비슷한 시기에 급속히 성장한 의학과 통계학이 서로 상승작용을 일으키는 시기였다. 특히 실험 대상을 임의로 배정하는 실험(randomized experiment)의 설계, 회귀와 상관, 검정방법 등은 20세기 의학에 큰 영향을 미쳤다. 20세기 전반기 영국에서 활동한 대표적인 의학통계학자들(Pearl, Greenwood, A. B. Hill 등)은 모두 당시 세계 유일의 통계학 교수였던 칼 피어슨에게서 통계학을 배운 사람들이었다[13].

1926년 피셔가 농작물의 재배에 관한 연구를 수행하면서 처음으로 실험적 처치방법을 무작위배정(random allocation)하였고, 이러한 개념에 기반하여 1931년 앰버슨(Amberson)이 폐결핵에 대한 sanocrysin의 치료효과를 판정하려는 연구에서 동전던지기 결과에 따라 치료법을 무작위로 배정함으로써 의학연구에 무작위배정을 최초로 도입하였다. 1948년 영국 의학연구원(British Medical Research Council)에서는 최초로 난수표(random number)를 사용하여 치료법을 무작위로 배정하였다[19]. 또한 치료효과의 판정을 객관적으로 하기 위한 눈가림법(blinding)과 관련해서는, 앰버슨의 연구에서 환자를 약물 치료군과 증류수를 주입하는 대조군으로 배정하면서 자신이 어느 치료를 받는지 모르게 하였고, 1936년 디일(Diehl) 등이 감기 백신에 대한 임상시험에서 대조군에게 생리식염수를 위약으로 투여하였다[19].

2. 가설의 검정

추론과 가설은 증명(실험)을 통해 권위를 얻는다. 그리고 무엇보다 중요한 명제, 실험과 일치하지 않는 법칙은 틀린 것이다[20]. 가설(hypothesis)은 연구자들에게 변수 사이의 기대되는 관계를 분명하게 표현하는 방법이다[3]. 기대되는 관계는 (인과관계의 추정이 아닌) 연관성(association)이거나 (독립변수가 종속변수의 변화에 원인이 되는) 인과관계(causal relationship)이다. 검정할 가설은 비교하여야 할 집단, 비교하여야 할 변수와 기대되는 관계를 정의하여야 한다[3]. 가설검정의 목적은 가설이 참인가 거짓인가를 밝히는 것으로 일반적으로 이해된다. 즉 미리 설정된 검정을 통과하는가 통과하지 못하는가에 따라 가설의 진위가 밝혀지게 된다. 그러나 현재 대부분의 과학론자들이 인정하고 있는 점은 경험적 자료에 의한 가설의 확실한 진위판단은 엄격히 따질 경우 불가능하다는 점이다. 이러한 문제점은 검토하려는 가설이 통계적 가설의 형태를 갖고 있을 때 더욱 분명해진다[21].

가설검정에서 발생할 수 있는 두 가지 오류를 이해하기 위해서는 먼저 '채택'과 '기각'이라는 두 가지 용어를 설명할 필요가 있다. 가설검정의 제 절차를 통해 임의의 가설에 대한 타당성이 입증되어 해당 가설을 받아들이는 것을 채택이라고 하며, 반대로 타당성이 입증되지 않아 해당 가설을 받아들이지 않는 것을 기각이라고

부른다[6]. 좋은 가설(연구가설)은 연구목적에 적합해야 하며, 개념(변수)들 간의 관계에 대해 진술하고 있어야 하고, 이들 개념(변수)들 간의 관계가 검정 가능해야 한다. 특히 검정 가능성이 가장 중요한 요건이다. 연구가설은 '영향을 미친다, 관계가 있다, 차이가 있다' 등과 같이 대립형태(alternative form)로 표현하는 경우가 일반적이지만 '영향을 미치지 않는다, 관계가 없다, 차이가 없다' 등과 같이 귀무형태(null form)로 표현하는 경우도 있다. 연구가설을 대립형태로 세워야 하는가 귀무형태로 세워야 하는가는 연구자의 입증하고자 하는 주장에 달려 있다[4]. 귀무가설은 대립가설, 즉 우리가 타당성을 입증하고자 하는 통계적 가설이 무엇인가에 따라 결정되며, 대립가설이 참이면서 동시에 귀무가설이 참일 수는 없고 그 역도 성립한다. 대립가설이 참이면서 동시에 귀무가설이 참일 수 없다는 말은 대립가설이라는 사건이 발생할 때 동시에 귀무가설이라는 사건이 발생할 수 없다는 의미이므로 대립가설과 귀무가설은 상호배반인 사건이라고 말할 수 있다[6].

가설검정은 다음과 같은 다섯 단계에 걸쳐서 수행되게 된다[22]. 첫째, 해당 연구에 관한 귀무가설(null hypothesis)과 대립가설(alternative hypothesis)을 설정한다. 둘째, 각 표본으로부터 적절한 자료를 모은다. 셋째, 귀무가설에 해당하는 검정통계량(test statistic) 값을 계산한다. 넷째, 해당 검정통계량 값을 이미 알려진 확률분포의 값과 비교한다. 다섯째, p-value(p값)를 해석하고 결론을 내린다.

3. 임상연구설계에서 사용되는 가설검정법

임상연구에서 신뢰할 만한 표본수, 즉 대상자의 수는 연구문제의 성격과 통계방법에 따라 달라진다. 표본수가 너무 작으면 모집단의 특성을 대표하는데 어려움이 있어 집단의 차이가 인정되는데 제한이 있고 실제로 존재하는 중요한 효과를 파악해낼 수 있는 검정력이 상대적으로 낮아지게 된다. 표본수가 커지면 추정치의 표준오차는 작아지게 되고 정밀도와 연구의 검정력을 높일 수 있다. 그러나 표본수가 지나치게 많으면 시간과 자원을 불필요하게 소모하게 된다. 그러므로 연구의 시작 단계에서 먼저 제1종 오류와 제2종 오류를 범할 가능성이 균형을 이룬 최적의 표본수를 계산하고 연구를 진행하여야 한다[23]. 귀무가설 또는 영가설의 유의성 검정(null hypothesis significance test)은 피셔의 유의성 검정과 네이만과 피어슨의 가설검정이 혼합된 것이다[5]. 우리는 보통 (표본을 사용해) 모집단에서는 효과가 없다(즉, 평균의 차이가 0이다)는 가정인 귀무가설(H_0)을 검정하게 된다[22]. 다음으로는 귀무가설이 사실이 아닌 경우에 해당하는 대립가설(H_1)을 정의하게 된다. 대립가설은 우리가 연구하고자 하는 이론과 직접적인 관련이 있는 가설이다[22].

영가설(null hypothesis)은 관심이 있는 변수들 간에 차이가 없거나 관련성을 가지고 있지 않다는 것이다. H_0로 쓰여지는 영가설은 통계검정의 기초이다. 통계적으로 가설을 검정할 때 H_0는 관심이 있는 두 변수 사이의 사건의 상태를 정확하게 기술한다고 가정한다. 유의한 차이나 관계가 발견된다면 영가설은 기각된다. 차이나 관련성이 존재하지 않으면 H_0는 채택된다[3]. 표본에서 얻은 경험적 사실을 이용하여 타당성을 입증하고자 하는 통계적 가설을 대립가설이라고 하며, 보통 H_1으로 표시한다[6]. H_1 또는 Ha로 표현되는 대립가설은 실행가설(acting hypothesis) 또는 연구가설(research hypothesis)로 알려져 있다[3].

유의성검정결과는 p값으로 표현된다. 귀무가설이 사실이라는 가정하에서 자료와 관련해서 계산한 확률이다[9]. 제1종 오류(Type Ⅰ error, 위양성)는 모집단내에서는 실제로 참인 귀무가설을 연구에서는 기각하는 경우 발생한다. 제2종 오류(Type Ⅱ error, 위음성)는 모집단 내에서는 실제로 거짓인 귀무가설을 기각하지 않아서 발생한다[25]. 즉, 다음과 같이 설명할 수 있다. 제1종 오류(Type Ⅰ error)는 귀무가설이 사실인데도 불구하고 이를 기각하는 잘못으로, 실제로는 효과가 없는 데 효과가 있다고 결론을 내리게 되는 오류이다. 제1종 오류를 범할 때 최대 가능성(확률)은 α(알파)로 표기한다. 이 값은 검정의 유의수준에 해당한다[22]. 제2종 오류

(Type Ⅱ error)는 귀무가설이 거짓인데도 불구하고 이를 기각하지 않는 잘못으로, 실제로 효과가 존재하는데도 불구하고 효과가 없다고 결론을 내리는 오류이다. 제2종 오류를 범할 가능성은 β(베타)로 표기한다. 이에 대한 여수(餘數) 1-β를 해당 검정의 검정력(power)이라 한다. 따라서 검정력은 거짓인 귀무가설을 기각하게 되는 확률이다. 즉, 검정력이란 주어진 크기의 실제 처리효과를 통계적으로 유의한 것으로 탐지해낼 수 있는 가능성이며, 일반적으로는 백분율로 표기한다[22].

산출된 확률이 작다고 해서 우연히 일어난 차를 유의한 차로 판정해버리는(귀무가설이 올바른데도 기각해버리는) 것을 α 오류라고 하며, 통상은 유의수준을 5%로 설정한다. 한편, 산출된 확률이 크다고 해서 본래의 차를 놓쳐버리는 것(귀무가설이 잘못되었는데도 채택하고 마는)을 β오류라고 한다[26]. 여기서 α가 적을수록 검정의 신빙성은 높아지지만 적절한 β의 크기와 균형이 잡혀 있어야 한다[21]. 극단적으로 α가 0이면 귀무가설은 절대로 기각되지 않고 β의 값이 1이면 대립가설은 절대로 채택되지 않으며 따라서 검정력은 0이 된다[21].

통계적 가설검정의 원칙 및 제 단계[9]는 다음과 같다. 첫째, 귀무가설 H_0와 대립가설 H_1을 세운다. 둘째, 유의수준 α를 결정한다. 셋째, 검정통계량을 결정한다. 넷째, 표본으로부터 얻은 구체적 결과을 이용하여 p값을 계산하거나, p=α가 되는 검정통계량의 값을 구한다. 다섯째, p≤α를 만족하면 유의수준 α에서 귀무가설을 기각하고, 이 때 '통계적으로 의미있는' 결과를 얻었다고 말한다. 또는 셋째와 넷째 과정에서 구한 검정통계량의 값이 해당 검정통계량의 확률분포에서 p≤α를 만족하는 영역에 있으면 유의수준 α에서 귀무가설을 기각할 수 있다. 만일 p〉α라면 귀무가설을 기각할 수 없다.

4. 반증주의와 귀류법

포퍼(K. Poppre)의 과학철학을 한마디로 표현하는 것은 쉬운 일이 아니지만, 그가 주장하는 과학탐구의 방법은 귀납법에서 '가설-연역적' 방법으로, 검정의 논리에서 '반증'의 논리로, 발견의 맥락에서 '정당화'의 맥락으로 대체되어야 한다는 것이다. 그는 '추측'과 '반박'을 통하여 지식이 정교화되고 성장한다고 보고 있으며, 지식의 성장 과정에 이러한 과학적 탐구 방법들이 적용되어야 한다고 말한다[27]. 포퍼에 따르면, 과학적 지식은 추측(conjecture)을 통해서 성장한다. 즉, 포퍼가 말하는 지식의 성장은 지식의 축적이라는 '덧셈'의 방법이 아니라, 잠정적으로 제안된 수많은 추측들 중에서 반증된 것을 제거하는 '뺄셈'의 방법에 의해서 이루어진다. 뺄셈의 방법에 의해 지식이 성장하려면, 가능한 한 많은 설익은 가설들이 제한 없이 요구된다. 즉 자유로운 가설의 생성, 나아가서 실수까지도 권장된다. 포퍼는 자신의 저서 『Conjecture & Refutations: The Growth of Scientific Knowledge』를 관통하는 하나의 주제가 곧 "우리는 실수로부터 배울 수 있다."는 것이라고 말한다[28,29].

반증 가능성(falsifiability)은 쉽게 말하면 '경험적으로 반박할 수 있는 가능성'이라고 할 수 있다[30]. 그런데 포퍼는 모든 진술이 반증의 시도에 놓이는 것은 아니라고 말한다. 즉 아무리 반증을 해보려 해도 반증할 수 있는 사례가 존재하지 않기 때문에 반증 자체가 아예 불가능한 진술들이 존재한다는 것이다. 그는 반증이 가능한 진술과 불가능한 진술을 구분하여 반증이 가능한 진술만 '과학적 진술(scientific statement)'이라고 규정한다[30].

요약하면 반증주의라는 새로운 과학 방법론을 제시한 포퍼는 과학 활동을 다음과 같은 절차로 규정했다[30]. 첫째, 주어진 문제들을 잘 설명하는 듯 보이는 가설을 제시하라(가설 창안의 단계). 둘째, 가설을 반박하는 경험적 사례가 발견되면 그 가설을 곧바로 폐기한다. 그렇지 않은 경우에는 그 가설을 그대로 유지한다. 이때 가설이 입증되었다고 주장해서는 안된다. 그저 몇 차례 혹독한 경험적 시험에 잘 견뎌왔다고 말할 수 있을 뿐이다.

귀무형태로 세운 연구가설은 이것이 검정 결과 기각되지 않을지라도 '참이라고 채택 할 수도 없다'는 부담이 남는다. 편의상 귀무가설이나 귀무형태의 연구가설이 기각되지 않은 경우 '채택되었다'는 표현을 사용하기

도 하지만 그 의미는 '기각할 수 없다'는 것으로 해석해야 하며, 이에 대한 보다 정확한 기술은 '기각할 수 없다 혹은 기각되지 않았다'고 표현하는 것이다[4]. 귀무가설은 절대로 증명되거나 확립될 수 없지만, 아마도 반증하는 것은 가능할 것이다. 모든 실험은 오직 그 사실들로 하여금 귀무가설을 반증시키도록 하는 기회를 부여하기 위해서 존재한다고 말할 수 있다[31,32]. 실제로 통계학에서는 귀무가설을 참이라고 가정한 상태에서 귀무가설에 반(反)하는 경험적 증거의 강약(强弱)을 이용하여 가설검정을 실시하게 되며, 이러한 경험적 증거의 강약은 보통 확률의 형태로 표현된다[6].

귀류법은 고전 논리학에서 두 가지 기본적인 원리인 모순율과 배중률에 근거를 두고 있다. 모순율은 명제 p와 p의 부정은 동시에 참일 수 없다는 것이고, 배중률은 p와 p의 부정 중 하나는 반드시 참임을 뜻한다(즉 제3 또는 중간의 가능성은 없다)[33]. 수학에서의 귀류법은 실제 증명하고자 하는 명제를 직접 증명하기 어려울 때 사용하는 방법으로 가정보다 결론의 부정이 훨씬 쉽게 이해되는 경우에 사용된다[34]. 귀류법은 고전적인 수학 교과서인 유클리드(Euclid) 원론에 빈번하게 등장할 뿐 아니라 에우독소스(Eudoxus)나 아르키메데스(Archimedes) 등 고대 그리스 수학자들에 의해 널리 사용되었다[35]. 귀류법은 증명하기를 원하는 명제의 결론을 부정하였을 때 모순되는 가정이 나온다는 것을 보여 증명하기를 원하는 명제가 참인 것을 증명하는 방법이다. 이는 자신의 이론을 직접 증명하는 것이 힘들 경우에 귀류법과 같은 간접증명법에 의지하는 것이 효과적인 증명방법이 되기 때문이다[36,37]. 귀류법에 의한 증명은 주어진 명제의 가정으로부터 결론을 직접 유도하는 것이 아니라, 실제로는 참인 원래의 명제를 부정하여 거짓이라고 가정한 새로운 명제를 상정함으로써 증명을 전개하기 때문에 이를 수행하는 사람에게는 상당한 인지적 압박을 준다[34].

5. 네이만-피어슨 방식의 가설검정법의 문제점

연구나 조사의 결과로부터 우연히 의도하지 않았던 통계적 유의성을 발견하였을 때, 이러한 통계적 유의성이 해당 대립가설의 타당성을 입증해 주는 것처럼 결론을 내리는 것은 옳지 않다. 모든 통계적 가설의 타당성 여부는, 미리 가설을 전제한 상태에서 표본추출을 실시하여 얻은 자료에 대해 통계적 가설검정의 절차를 거친 후에야 판단할 수 있는 것이다. 따라서 기대하지 않았던 통계적 유의성이 얻어졌다면 이는 가설의 타당성을 확증하는 것이라기 보다, 새로운 가설의 형성을 지지하고 촉진하는 것으로 이해되어야 한다[38]. 유의하지 않은 결과가 나왔다는 것은 현재의 자료에 근거하여 어떤 결정을 내릴 수 없다는 것을 의미한다. 귀무가설이 기각되지 않았다고 해서 귀무가설이 옳다는 것이 확인된 것은 아니다. 유의한 결과가 나왔을 때 귀무가설을 기각할 수는 있지만, 통계적 유의성은 단지 최소한의 기준이며, 필요조건이지 충분조건은 아니다[5]. 아무리 작은 유의수준을 선택하고 이보다 더 작은 p값을 얻었다고 해도 이러한 사실이 곧 기존의 행위나 신념의 변경으로 연결되는 것은 아니다. 이것은 제 과학의 발전과정이 간혹 비약적이지 않은 것은 아니나 그보다는 점진적인 이해의 확장에 의존하는 경우가 더 많기 때문이다[38].

통계적 유의성은 실제적 유의성이 아니다. 귀무가설 검정은 효과의 크기나 중요성을 직접 평가하지 않는다. 통계적 유의성과 실제적 유의성이 같다고 생각하는 이 유의성 오류(significance fallacy)는 연구결과의 중요성을 결정하는 효과 크기(effect size)에 관해서는 p값이 직접적인 정보를 제공하지 않는다는 사실을 간과하는 데에서 기인한다[5,39]. 통계적 가설검정의 결과 일정한 유의수준 이하의 p값이 얻어져야만 그 연구 또는 자료수집이 잘 된 것이라고 믿는 것은 매우 위험한 견해이다. 통계적 유의성을 얻지 못한 연구 결과라 할지라도 그 자체로서 충분한 '의미'가 있을 수 있음을 명심해야 한다[38].

유의한 결과가 중요한 결과가 아닌 이유 중 하나는 p값이 표본 수에 의존하기 때문이다. 귀무가설 기각의 가능성은 표본이 작을 때에는 작고 표본이 클 때에는 크며, 표본이 매우 클 때에는 어떤 결과도 유의하게 나

타난다. 큰 효과 크기를 가진 작은 표본 연구는 작은 효과 크기를 가진 큰 표본 연구와 동일한 p값을 산출할 수 있다[40]. 오늘날의 대규모 자료(big data)를 사용 하는 연구에서는 거의 모든 귀무가설에서 p값이 아주 작게 나타날 수 있다[41]. 그렇기 때문에 p값은 고정된, 객관적 의미를 가지고 있지 않다[5].

P값의 'p'는 확률(probability)을 의미한다. 그리고 p값은 귀무가설이 참이라고 가정했을 때, 관찰된 데이터를 관찰할 확률(the probability of observing the observed data)로 정의될 수 있다. P값은 실제의 수(real number)가 아니다. 그것은 실제의 확률을 반영하는 것이 아니라, 귀무조건(null condition)이 참일 것이라고 우연히 추정(알지 못하므로 추정하는 것이다)할 가능성을 의미한다[32]. 사실 5%는 임의적인 기준이다. 5%라는 값을 사용한다는 것은 우리가 사실인 귀무가설을 기각하게 되는 잘못을 100번 중 5번 정도 저지르게 될 것이라는 의미이다. 만일 귀무가설을 잘못 기각하는 경우가 대립가설을 잘못 채택하게 되는 경우에 비해서 임상적으로 매우 심각한 결과를 초래할 수 있는 상황이라고 한다면 귀무가설을 기각하기 위한 증거를 더 강하게 설정하면 된다(예를 들어 기준점으로 1%나 0.1%를 사용하여 p값이 0.01보다 작거나 0.001보다 작은 경우에 귀무가설을 기각하기로 하면 된다)[22]. P값이 0.05에 가까운 다른 값이 아닌, 0.05여야 하는 과학적인 근거는 없다. 여기서 우리는 이토록 고도로 수학적이고 과학적인 학문인 통계학의 핵심적인 부분이 전혀 과학이나 수학에 근거하지 않고 있다는 것을 알게 된다. 인류의 다른 모든 노력들과 마찬가지로, 통계학도 부분적으로는 개념적 추정에 근거하고 있는 것이다[32]. P값이 0.04이던가 0.12라던가 하는 것은 사실상 많은 것들을 말해주지 않는다. 실제로 임상에 통계학을 적용할 때 벌어지는 문제 가운데 하나가 통계학적인 계산으로부터 나오는 양적 검정력이 임상적인 측면으로 볼때는 큰 의미가 없을 때가 많다는 것이다[32].

네이만 자신은 자신의 가설검정법이 과학연구에 고정관념처럼 자리 잡는데 전혀 관여하지 않았다. 네이만은 1935년 프랑스수학회보에 불어로 발표한 논문에서 최적 가설검정법을 찾지 못할지도 모른다는 회의적인 시각을 토로했으며, 그 후 발표한 논문에서는 가설검정을 거의 사용하지 않았다. 그는 주로 이론적 원리로부터 확률분포를 유도하고 자료로부터 모수를 추정하는 데 전념했다[9]. 네이만의 기본적인 아이디어가 이처럼 왜곡되었음에도 가설검정은 과학연구에서 가장 널리 사용되는 통계적 방법이 되었다. 네이만의 뛰어난 수학적 재능에서 나온 가설검정법은 이제 과학자들의 뇌리에 고정관념처럼 박혀있다. 대부분의 과학 학술지는 자료분석에 가설검정결과를 포함할 것을 요구한다. 이런 현상은 과학계를 넘어 다른 영역으로까지 확산되었다. 미국, 캐나다, 유럽의 약품규제기관도 약품제조허가를 신청할 때 가설검정결과를 요구한다. 법정에서도 가설검정결과를 적절한 증거로 받아들이고 있으며, 원고는 가설검정으로 고용차별을 증명할 수 있다. 가설검정은 통계학의 모든 분야에 스며들었다[32].

6. 네이만-피어슨 방식의 가설검정법의 대안

1) 베이지안 통계학(Bayesian statistics)

통계학 안에는 두 가지 철학이 존재해왔다. 오늘날 통계학의 주류는 단지 데이터를 평가하고 수학적으로 그 데이터를 해석하는 것으로서 이를 빈도주의자의 통계학이라고 한다. 이것과는 다른 접근법이 관찰자가 데이터를 어떻게 해석할 것인지에 대한 것으로 이를 베이지안(Bayesian) 통계학이라고 한다[32]. 베이지안 통계는 베이즈(Bayes)가 발견한 정리에 기초한 통계로서, 피셔와 네이만-피어슨의 빈도주의자 통계(frequentist statistics)와 구별된다[5]. 베이지안 통계는 축적되는 근거(evidence)로부터 배우는 방법이다. 빈도주의자 통계에서는 이전 연구들로부터 얻은 정보를 연구 설계 단계에서 주로 사용하나, 베이지안 통계에서는 이전 연구들의 결과와 새로운 연구들의 결과들을 연속적인 자료 흐름으로 간주하며, 새로운 자료가 얻어질 때마다 가설이 갱신(update)된다. 연구가 반복되어 결과가 축적될수록 베이지안 통계는 정확한 진실에 가까워진다[5,42]. 베이지

안 통계는 우리가 세상을 해석하고 의사결정을 내리는 방식과 비슷하다. 우리의 뇌는 부분적인 정보를 토대로 가장 가능성이 높은 것을 예측하고 그 결과를 분석해서 새로운 정보를 얻어 기존 정보를 갱신(update)하도록 진화되어 왔다[5].

베이지안(Bayesian)의 사고방식은 숫자를 진실로 간주해서 지나치게 흠모하는 것과 개별적 환자에 대한 결정을 내림에 있어 직관적으로 접근하는 방법 사이를 이어주는 가교와 같다[32]. 베이즈(Bayes)의 정리는 사전에 주어진 확률 X와 관찰된 사건인 Y는 사후확률 Z를 만들어 낸다는 것이다[32]. 베이지안(Bayesian) 접근법은 기존에 주어진 사전 확률로부터 시작된다. 그 다음에 관찰이나 실험을 통해 데이터를 생산해낸다. 이렇게 생성된 데이터는 사전확률을 수정하는데 사용되며 결국 사후확률을 만들어낸다[32].

빈도주의자는 연구나 실제상황에서 주관적인 것들을 없애고 싶어 하겠지만 베이지안(Bayesian)은 그런 것들을 인정하고 나서 그것들로 인한 위험은 최소화시키고 객관적 과학 증거의 활용도는 최대화시키고자 한다[32]. 베이지안 통계를 사용한 연구에서는 사전 정보의 선택 및 복수 근원의 사전 정보들을 통합하기 위한 수학적 모델의 선택을 포함한 사전 기획의 영향이 결정적이며, 어떻게 기획을 하는가에 따라 연구의 결과가 달라진다. 그러나, 선택에 따른 영향 역시 수학적으로 검정될 수 있다. 빈도주의자 통계에 비해 베이지안 통계는 매우 복잡한 수학이 사용된다. 연구모형 자체가 유연하고 분석을 위한 계산 기술이 복잡하기 때문에 과거에는 간단한 역학 연구의 분석도 어려웠다[43]. 지금은 컴퓨터의 발달로 계산의 문제가 해결되었고 모의실험(simulation) 모델링 기술로 아주 복잡한 생의학 분야에의 적용도 가능하게 되었다. 이 진보는 베이지안 통계의 인기가 크게 증가하는 결과를 낳았다[5,44,45].

베이즈주의는 일련의 "합리성 원리"에 기초한다. 확률에 대한 주관적 해석, 신념 체계가 확률 공리를 준수해야한다는 규정, 그리고 증거에 의한 믿음의 갱신이 그 핵심 원리를 이룬다[46,47]. 베이즈주의는 형식인식론의 많은 부분을 대표할 수 있을 만큼 확률론, 통계학, 의사결정론, 그리고 형식적 학습이론 등 여러 분야에 독자적인 입장을 발전시키고 있다. 베이즈 신부가 제시한 베이즈의 정리와 믿음의 정도라는 확률의 해석에 기초하여 경험으로부터의 학습과정을 형식화하는 베이즈주의는 베이즈의 사상도 아니고 확률론의 법칙 중 하나인 베이즈주의 정리 자체도 아니다. 베이즈주의는 여러 가지 일상적 판단을 확률적으로 분석하고 평가하는 과학방법론이다[48]. 베이즈주의에서 사전확률이 기존의 귀납추리에 덧붙여진 전제라고 해석한다면 사전확률값의 부여가 적절히 이루어졌는지 아닌지를 판단하여 베이즈주의의 귀납추리가 신뢰할만한지 아닌지를 판단할 수 있는 것이다[49]. 베이즈주의에서 사전확률이 객관적이 아니라 주관적으로 결정되며 0부터 1사이의 어떤 값도 부여받을 수 있다고 말하더라도 그 값이 임의적으로 제멋대로 결정된다는 것을 의미하는 것인지는 다시 생각해볼 문제이다[50].

호손(J. Hawthone)은 베이즈주의가 본질적으로 제거적 귀납의 확률론적 형태라고 주장한다. 호손에 따르면 베이즈주의는 제거적 귀납을 정교하게 발전시킨 것으로, 그 본질은 참된 가설의 잘못된 경쟁자들을 반박하는 데에 있다고 한다. 이와 같은 주장을 옹호하기 위해 호손은 베이즈주의가 제시하고 있는 수렴성 정리(Bayesian convergence theorems)의 의미를 분석한다. 베이즈주의의 수렴성 테제가 이미 경쟁가설을 비교하여 제거할 수 있는 매커니즘을 갖추고 있다는 것이다[49]. 베이즈주의의 수렴성 정리는 모두 가설의 신빙성에 관한 초기의견이 행위자들마다 서로 다르다고 하더라도 그 초기의견이 극단적인 경우가 아니라면, 즉 그 초기의견의 확률값이 0 또는 1을 부여받지 않는다면, 그 초기의견의 차이는 결국 증거에 의한 확률값 조정을 거치는 과정에서 사라지게 될 것이라는 점을 제시한다. 즉 가설의 신빙성에 관한 초기의견이 사전확률로 제시될 때, 그 사전확률이 사후확률에 이르는 과정은 적절한 증거가 조건화법칙에 의해서 참인 이론에 대한 확률값을 높이는 과정이고, 이 과정은 결국 행위자들 사이의

다양한 초기의견을 하나로 모으게 한다는 것이다[49].

2) 라카토스의 연구프로그램(Lakatos methodology of scientific research programmes)

포퍼(K. Popper)의 반증주의의 문제점을 누구보다도 분명히 밝히면서, 반증가능성의 의의를 살린 사람은 포퍼와 함께 런던정경대학교에서 근무했던 라카토스(I. Lakatos)이다. 쿤(T. Kuhn)의 혁명적 변화론에 대한 라카토스(Imre Lakatos;1922-1974)의 반응은 포퍼보다 복합적이다[51]. 수학의 논증을 전공한 라카토스는 수학분야에서마저도 포퍼의 주장과는 달리 어떤 이론이 단지 하나의 반증되는 증명에 의해서 간단히 폐기되지 않는다는 점을 인지하고 있었다[51,52]. 라카토스는 포퍼의 반증주의를 '소박한 반증주의'라고 칭하고, 과학 이론은 하나의 틀이란 점을 강조하며, 포퍼의 반증주의에 대비하여 자신의 입장을 '세련된 방법론적 반증주의'라고 부른다[51]. 다시 말하면, 라카토스는 포퍼와 같이 반증가능성을 이론 발전의 중요한 요소로 보면서도, 쿤처럼 이론은 하나의 연계된 진술들로 이루어진 구조를 이루고 있어, 쉽사리 반증되지 않으며, 어떤 사실 혹은 진술에 의하여 반증되었다고 해서, 해당 이론이 간단히 폐기되는 일은 일어나지 않는다는 점을 분명히 하였다. 라카토스는 과학자의 연구활동을 이해하려면 그들이 명시적으로든 혹은 암묵적으로든 따르고 있는 과학적 연구프로그램(Scientific Research Program)을 이해해야 한다고 주장하며, 그 연구프로그램에 의한 과학관을 제시한다[51].

과학적 평가와 진보의 단위는 개별 이론이 아니라, '일련의 이론들'이어야 한다는 주장은 라카토스의 기본적인 통찰이다[53]. 그는 과학 이론이 마치 고립된 섬처럼 단독으로 존재하는 것이 아니라고 말한다. 오히려 그것은 여러 종류의 가설과 믿음들이 촘촘히 연결되어 있는 그물과도 같다. 더욱이 그는 그런 그물 체계가 시간에 따라 조금씩 수정되어 간다고 주장했다. 정확하게 말하면 그가 말하는 '연구 프로그램'은 일련의 이론들의 집합으로서 크게는 '견고한 핵(hard core)'이라고 하는 핵심 이론과 '보호대(protective belt)'라는 여러 유형의 보조 가설들(auxiliary hypotheses)로 구성되어 있다[30]. 라카토스 연구프로그램(methodology of scientific research program)은 견고한 핵(hard core), 보호대(protective belt), 변칙사례(anomaly), 부정적 발견법(negative heuristics), 긍정적 발견법(positive heuristics), 세련된 예측(novel prediction)으로 구성되어 있다[54].

라카토스에 따르면 과학 이론의 모든 부분은 동등한 중요성을 가지고 있는 요소로 구성되어 있지 않다. 과학 이론을 구성하는 특정한 법칙이나 원리는 다른 구성요소들보다 이론에서 근본적인 역할을 한다. 라카토스는 이처럼 과학적 연구프로그램을 특징짓는 근본적인 원리를 '견고한 핵(hard core)'이라고 정의한다[55,56]. 이론을 구성하는 자연법칙, 초기 조건, 보조 가설이 어떤 사실 명제와 충돌하는 경우, 이를 변칙사례라고 한다[55,56]. 보호대는 이론이 변칙사례에 직면했을 때, 견고한 핵의 내용이 반증되지 않게 하기 위해 수정·보완 가능한 보조가설을 일컫는다[54]. 라카토스는 '발견법(heruistic)'을 통해 연구프로그램의 특징을 기술한다[57]. 부정적 발견법과 긍정적 발견법은 연구프로그램의 진행 방향을 지시하는 연구지침이다. 연구프로그램에서 부정적 발견법은 변칙 사례에 직면하여도 견고한 핵이 '반박 불가능하다'는 사실을 명확하게 지시하는 연구 지침이다[55,56]. 라카토스에 의하면 긍정적 발견법은 연구프로그램의 "반증 가능한 보호대를 어떻게 수정하고, 정교하게 할 것인가를 명확하게 설명하고 있는 일련의 시사와 암시로 구성되어 있다[55,56]." 라카토스 연구방법론은 반증을 시도하는 변칙사례의 도전에 직면하여 부정적 발견법이 견고한 핵을 보존하고 긍정적 발견법과 부정적 발견법에 의해 보호대를 수정하고 보완하는 과정으로 이해할 수 있다[54]. 연구프로그램이 진보적인지 또는 퇴보적인지 평가할 수 있는 가장 중요한 척도는 신선한 예측이다. 이론의 정합성을 유지하면서 신선한 예측의 일부가 입증될 경우 진보적인 연구프로그램으로 평가되는 반면에 이론의 정합성을 상실하거나 신선한

예측을 입증하지 못한 경우 퇴행적인 연구프로그램으로 평가된다[55,56]. 만약 연구프로그램이 변칙 사례를 극복하지 못한다면, 이론을 구성하는 핵심 원리들이 반증될 수 있다. 하지만 반증이 일어날 경우 이론이 즉각 폐기된다고 주장하는 포퍼의 반증주의와 다르게 라카토스의 연구프로그램 방법론에서는 이론을 즉각적으로 포기하지 않는다[56].

7. 베이지안 통계학(Bayesian statistics)을 이용한 부비동염의 변증(辨證) 예시

이 장에서는 베이지안 통계학이 실제로 한방이비인후과학 분야의 변증에 있어서 어떻게 활용이 될 수 있는지 부비동염 환자를 예로 들어서 설명을 하고자 한다. 부비동염은 한방이비인후과학 분야에서 다루는 대표적인 질환이며 적절한 치료를 위한 진단의 단계로서 '변증(辨證)'은 매우 중요하다. 여기에서 소개한 부비동염의 증상 및 변증은 『한의안이비인후과학』[58]의 내용을 기반으로 하였고 베이지안 통계학을 이용한 계산법의 기본 구조는 와쿠이 요시키(涌井良幸)와 와쿠이 사다미(涌井貞美)가 소개한 예시[59]를 기반으로 하였다.

먼저 부비동염의 변증을 위하여 다음과 같이 세 가지의 설정을 하였다. 첫째, 환자는 부비동염으로 진단이 되었으나 변증에 있어서 肺經熱盛인지 脾虛肺弱인지 구별이 힘든 상태이다. 둘째, 환자의 차트에서 '두중통'과 '후각저하'라는 단어가 1회씩 검색되었다. 셋째, 기존에 해당 진료실을 방문했던 환자들 중에서 肺經熱盛과 脾虛肺弱으로 변증된 환자의 비율은 8:2였다.

부비동염의 변증을 위하여 증상을 표현하는 4개의 단어인 '두중통', '미열', '천식', '후각저하'에 대하여 주목하였으며, 이들 단어들은 다음에 제시된 확률로 각각 肺經熱盛(H_1)과 脾虛肺弱(H_2)으로 분류되어 있다고 가정하였다(Table 1).

이와 같은 가정을 전제로 하여 베이즈 정리의 계산법을 사용하여 계산하여 보면 다음과 같다. 먼저 1단계는 원인 H와 검출단어 D로 모델화를 한다(Table 2, 3). 또한 그로부터 '우도' $P(D|H_1)$, $P(D|H_2)$를 산출한다. '우도' $P(D|H_1)$, $P(D|H_2)$는 검출되는 '증상을 표현하는 단어'의 출현확률이다(Table 4). 2단계는 사전확률인 $P(H_1)$, $P(H_2)$를 설정하는 단계이다. 기존에 방문했던 부비동염 환자들 중에서 肺經熱盛(H_1)과 脾虛肺弱(H_2)으로 변증된 환자의 비율이 8:2라고 가정하였으므로 이것을 사전확률에 넣는다(Table 5). 3단계는 두 표(Table 4, 5)의 데이터를 대입해서 사후확률을 산출하는 단계이다. 처음에 '두중통'이라는 단어가 검출되었으므로 이 단어가 나타났을 때 肺經熱盛(H_1)으로 변증될 확률을 베이즈 정리로 계산을 하면 다음과 같다.

$$P(H_1|D_1) = \frac{P(D|H_1)\,P(H_1)}{P(D_1)} = \frac{0.7\times0.8}{P(D_1)} \quad \cdots (1)$$

$$P(H_1|D_4) = \frac{P(D4|H_1)\,P(H_1)}{P(D4)} = \frac{0.07\times0.7\times0.8}{P(D_1)P(D_4)} = \frac{0.0392}{P(D_1)P(D_4)} \quad \cdots (2)$$

Table 1. Probability of Sinusitis by Differentiation of Syndromes

Symptom	H_1 (Lung Meridian Excessive Heat Syndrome)	H_2 (Spleen Deficiency and Lung Weakness Syndrome)
Headache with Heavy Sense	0.7	0.2
Slight Fever	0.6	0.3
Asthma	0.01	0.8
Hyposmia	0.07	0.6

2회째에는 '후각저하'라는 단어가 검출되었으므로 그 단어가 나타났을 때 肺經熱盛(H_1)으로 변증될 확률을 베이즈 정리로 계산한다. 이때 베이즈 갱신을 이용하고 사전확률에는 위의 [식 (1)]의 값을 이용한다.

마찬가지로 脾虛肺弱(H_2)으로 변증될 확률도 계산해 보면 다음과 같다.

$$P(H_2|D_4) = \frac{0.6\times0.2\times0.2}{P(D_1)P(D_4)} = \frac{0.024}{P(D_1)P(D_4)} \cdots (3)$$

[식(2)]와 [식(3)]을 비교해보면 [식(2)]의 확률이 더 큼을 알 수 있다. 따라서 이 환자는 肺經熱盛(H_1)으로 변증된다는 결론을 내릴 수 있다.

Ⅲ. 고 찰

앞서 서론에서 언급하였듯이, 이[5]의 연구에서는 귀무가설 유의성 검정의 문제점을 지적하였으며 귀무가설 유의성 검정을 보완하거나 대체할 수 있는 몇 가지 대안을 제시하였는데 그 내용을 살펴보면 첫째는 효과 크기(effect size), 둘째는 신뢰 구간(confidence interval), 셋째는 베이지안 통계(Bayesian statistics)이다. 이들 대안에 대하여 하나하나 살펴보고자 한다.

첫 번째로 제시된 대안은 효과 크기(effect size)이

Table 2. Modeling of Cause 'H'

Cause	Meaning
H_1	Lung Meridian Excessive Heat Syndrome
H_2	Spleen Deficiency and Lung Weakness Syndrome

Table 3. Modeling of 'D' that Conserned with Symptoms

Data	Meaning
D_1	Appearance of 'Headache with Heavy Sense'
D_2	Appearance of 'Slight Fever'
D_3	Appearance of 'Asthma'
D_4	Appearance of 'Hyposmia'

Table 4. Appearance Probability of the Words that Showing Sinusitis Symptoms

Word	$P(D\|H_1)$	$P(D\|H_2)$
D_1(Headache with Heavy Sense)	0.7	0.2
D_2(Slight Fever)	0.6	0.3
D_3(Asthma)	0.01	0.8
D_4(Hyposmia)	0.07	0.6

Table 5. Prior Probability of Sinusitis Outpatiens by Differentiation of Syndromes

Prior Probability	$P(H_1)$	$P(H_2)$
Probability	0.8	0.2

다. 효과 크기는 독립변수와 종속변수 간 연관성의 강도를 나타내는 지표이다. 실험군의 평균과 대조군의 평균 사이의 차이를 효과 크기라고 할 수 있으나, 임의적 척도를 사용한 연구에서처럼 변수의 측정치 자체가 내재적 의미를 가지고 있지 않거나 메타분석 연구에서처럼 상이한 척도를 사용한 여러 연구들의 결과를 종합하여야 할 때에는 표준화된 효과 크기를 사용한다[5,60,61]. P값과 달리, 효과 크기는 표본 크기에 민감하지 않다는 이점이 있다. 효과 크기는 표본 수에 따라 증가하거나 감소하지 않는 안정적인 수치이다[5,60,61]. 효과 크기는 기술통계량이지 추론통계량이 아니기 때문에, 효과 크기 자체는 변수들 간의 연관성이 우연에 의한 것일 가능성을 알려주지 않는다. 즉, 효과 크기는 표본 내 효과의 크기를 드러낼 뿐이며 모집단에 이 값이 존재할 가능성에 대한 정보는 제공하지 않는다. P값 대신에 효과 크기를 사용하기보다는 p값을 보완하기 위해 효과 크기를 사용하는 것이 바람직하다[5].

두 번째로 제시된 대안은 신뢰 구간(confidence interval)이다. 신뢰구간을 사용하면 단지 p값이 0.05보다 작다는 이유로 효과가 없다고 단정하는 것을 피하고 연구가설을 지지하는 증거를 찾을 가능성이 커진다. 신뢰구간은 p값보다 훨씬 더 많은 정보를 제공하며 검정보다 추정(estimate)이 우월함을 보여준다. 신뢰구간은 구간의 폭을 통해 추정의 정밀도 또는 신뢰도를 나타내며, 결과가 통계적으로 유의하기보다 실제적으로 유의한지 여부를 더 쉽게 볼 수 있다[5]. P값, 신뢰구간의 어느 쪽을 사용해도 상관없지만, 신뢰구간을 이용하면 치료효과의 크기와 연구의 신뢰성(유의차의 정도)을 동시에 알 수 있다. 또 임상적 판단은 p=0.05를 경계로 흑백을 가리는 것이 아니므로, p값보다도 신뢰구간을 이용하여 오차범위를 포함하여 판단을 내리는 방법이 불확실성을 동반하는 임상에 적합하다. 또 신뢰구간의 폭이 너무 넓은 경우에는 재현성이 부족하거나 검정법 선택이 잘못되었을 가능성이 시사된다. 최근 임상연구 보고에서는 신뢰구간과 p값을 모두 기술하도록 하는 경우가 흔히 있다[26].

세 번째로 제시된 대안은 베이지안 통계(Bayesian statistics)이다. 베이지안 통계는 귀무가설과 대립가설을 구분하지 않으며, 동시에 검정하고자 하는 가설의 수가 세 개 이상이어도 무관하다. 또한 가설들 사이에 서로 내포되어야 한다는 조건도 필요치 않다. 다수의 가설을 포함하는 복잡한 연구모형의 분석에 사용할 수 있으며, 현대 사회가 생산해 내는 막대한 양의 상호 연관된 이질적인 자료들을 연결하여 분석할 수 있는 능력이 있다[5]. 베이지안 통계는 사전정보가 없을 때에도 적응적 연구를 설계하고 시행하는데 유용하다. 연구 시작 후에도 중간 분석, 표본 크기의 변화, 표집 방법의 변화와 같이 계획되지 않았던 작업이나 연구계획의 수정이 가능하다. 가능도 원리(likelihood principle)에 고착함으로써 적응적 연구의 설계와 시행에서 유연성을 제공할 수 있다[5].

본 연구에서는 이에 더하여 라카토스의 연구프로그램을 소개하고자 한다. '과학적 연구프로그램의 방법론'으로 불리우는 라카토스의 과학방법론은 '반증'을 중심으로 이루어지는 포퍼의 과학철학에 대한 쿤의 비판을 적극적으로 수용한다. 그리고 라카토스의 과학방법론은 과학의 역사적 성격을 받아들이면서, 쿤의 과학사 논의를 통해서 형성된 비합리주의와 상대주의를 극복하려는 시도였다[62]. 라카토스는 연구프로그램의 핵심과 보호대를 구분하고, 부정적 연구지침과 긍정적 연구지침을 구분함으로써, 포퍼의 한계를 넘어서서 이론이 단순한 관찰명제에 의해서 반증되고 폐기된다는 소박한 반증주의의 난점을 극복하고자 하였고, 또 쿤의 주장처럼 숱하게 많은 반증의 사례에도 불구하고 대부분의 반박을 무시하고, 지속적인 연구가 진행된다는 점을 설명하였다[51,55].

라카토스에 의하면 연구프로그램은 네 가지 형태로 진행될 수 있다. 첫째, 변칙사례에 대한 입증에 실패하는 경우이다. 둘째, 변칙사례를 입증사례로 전환하였지만 신선한 예측을 제공하지 못한 경우이다. 셋째, 변칙사례를 입증사례로 포섭할 뿐만 아니라 동시에 신선한 예측을 제공하는 경우이다. 이러한 경우 연구프로그램

은 '이론적으로 진보적'이라고 평가된다. 넷째, 신선한 예측의 일부가 경험적으로 입증되는 경우다. 이 경우 연구프로그램은 '경험적으로 진보적'이라고 평가된다[56,63].

본론의 6장에서 이미 소개하였듯이, 본 연구에서는 이[5]의 연구와는 달리 베이지안 통계와 라카토스의 연구프로그램의 두 가지를 중요한 대안으로 인식하며, 이는 귀무가설 유의성검정의 문제점의 본질이 '이분법적 사고'를 중시하는 데에 있다고 보기 때문이다. 네이만-피어슨의 연구에서와 같이, 귀무가설과 대립가설의 서로 대립되는 두 가지 가설을 놓고 하나의 가설을 선택하는 과정은, 방법적으로는 매우 간단하지만 흑백논리에 빠질 우려가 있다. 특히 귀류법은 모순된 전제를 부정하는 방법이기 때문에 귀무가설은 이미 '부정을 전제로 한 가설'이라는 문제점을 안고 있다. 그리고 이와 같은 문제점들의 기반에는 '이분법적 사고'가 있다.

이분법적 사고는 우리들이 어떤 것을 분명히 인식하는데 도움을 준다. 우리가 무엇을 정확히 알았다는 것은 그 무엇에 해당하는 것과 그에 해당하지 않는 것을 이분법으로 구분할 수 있기 때문에 가능하다. 우리들이 시험을 칠 때 정답을 찾아가는 것도 이분법을 사용하기 때문에 가능하다[64]. 이분법은 편리함을 위한 가정(假定)이라고 보아야 한다. 그런데 진정한 실재가 실제로도 이분법으로 분리되어 있다고 보는 것은 흑백오류보다 더욱 심한 오류이다. 화이트헤드(A. N. Whitehead)는 이것을 '잘못 놓여진 구체성의 오류'(fallacy of misplaced concreateness)라고 불렀다[64].

일반적으로 말해서 인식론은 우리의 지식이나 믿음을 다루는 분야로 간주된다. 이와 관련해서 기존 인식론은 우리의 지식 혹은 믿음을 전부 아니면 전무(all or nothing)의 문제로 간주해 왔으며, 그런 이가(二價)적 모형 속에서 지식과 믿음의 인식적 정당성을 탐구해 온 경향이 있었다. 이에 비해, 베이즈주의 인식론은 우리의 믿음을 정도의 문제(matter of degree)로 간주하고 있으며, 그런 믿음의 정도가 합리적이기 위해서 갖추어야 할 조건들을 탐구한다. 이런 베이즈주의 인식론의 기본적인 탐구 방향은 공시적(synchronic) 요소와 통시적(diachronic) 요소로 구분하여 명료해질 수 있다[65]. 라카토스에 의하면 경쟁관계에 있는 연구프로그램의 우위는 반증에 의하여 즉각적으로 판정되지 않는다. 즉, '즉각적 합리성은 종언'을 고해야 한다. 연구프로그램을 순식간에 무너뜨릴 수 있다는 의미의 결정적인 실험은 존재하지 않는다. 일련의 연계된 이론들이 일거에 반증되기는 어렵기 때문이다[51].

이에 더하여, 피셔가 창안한 'p값'이 이분법적 사고에 기초한 의사결정의 수단으로 사용될 수 있다는 점에 대해서도 주의를 기울여야 한다.

통계적 유의성의 추구를 연구 목적이라고 생각하는 많은 연구자들의 무의식에는, 마치 통계적 유의성이 얻어지면 귀무가설로 대변되는 기존의 신념이나 행동을 쉽게 변경할 수 있다는 전제가 깔려 있는 것처럼 보인다. 이것은 통계적 가설검정이 일종의 의사결정의 성격을 띠고 있기 때문에 일견 타당한 것처럼 보인다. 즉, 귀무가설에 반하는 경험적 증거의 강약은 p값으로 요약할 수 있고, 이러한 p값이 유의수준 α보다 작으면 귀무가설을 기각하는 일종의 '결정'을 내리는 것이다[38].

가설검정으로부터 통계적 유의성이 있는 결과가 얻어졌다고 해서 이것이 곧 실제적으로도 의미가 있다는 것을 의미하지는 않는다[38]. 단 한 번의 아주 작은 p값 또는 단 한 번의 커다란 반증이 곧바로 과학적 신념이나 행위 수정의 결과로 이어지는 것이 아니라는 사실을 명심하면, 통계적 유의성이 있는 결과를 얻는 것이 결코 연구의 목적이 될 수는 없으며, 통계적 유의성이 없는 결과가 얻어졌다고 해서 실망할 필요도 없음을 쉽게 수긍할 수 있다[38]. 이미 피셔나 네이만 같은 학자들이 통계학적 검정법을 개발하고 얼마 되지 않아서부터 이 방법의 문제점을 지적하고 지나친 남용을 경고하는 연구들이 여럿 나왔지만 상황은 별로 개선되지 않았다. 그러자 2015년 초 미국에서 발행되는 어느 심리학 학술지(Basic and Applied Social Psychology)[66]에서는 아예 p값이 들어있는 논문은 싣지 않겠다는 극단적인 폭탄선언까지 발표했다. 또한 2016년 미국통계학회는

이례적으로 p값을 잘못 적용하는 경우에 대한 명확한 가이드라인을 제시했다[67]. 2017년에는 다양한 분야의 연구자 72명의 이름으로 '새로운 발견을 했다고 주장하는데 필요한 p값의 기준을 0.05에서 0.005로 내려야 한다'라고 주장하는 논문도 등장했다[68]. 통계적 방법을 이용할 때의 오류를 줄이기 위해서라도 기본적인 원리를 이해할 필요가 있겠다[13].

한의학계에서도 이미 네이만과 피어슨이 정립한 귀무가설의 유의성검정에 기반한 임상연구방법론이 많이 사용되어 왔다. 빈도주의자 통계학의 귀무가설검정을 이용한 임상연구방법은 한 번 유의한 결과가 나오면 반복연구가 어려우며[5], 따라서 각 연구들의 결과가 일회적이고 독립적이다. 물론 이전 연구들로부터 얻은 정보를 연구 설계의 단계에서 사용할 수는 있지만[5], 연구의 본질상 이전 연구의 결과와 새로운 연구의 결과는 직접적인 연계가 성립되지 않는다. 그러나 베이지안 통계학에서는 이전 연구들의 결과와 새로운 연구의 결과를 연속적인 자료의 흐름으로 간주하며[5], 연구의 대상자가 방문하는 횟수가 늘어남에 따라 처음부터 축적된 정보가 계속 활용되어 누적이 된다. 임상의 현장에서는 동일한 장소에 동일한 환자가 반복적으로 방문하는 일이 많음을 고려했을 때, 연구대상자의 계속적인 방문에 따른 정보의 누적과 그에 따른 판단의 갱신이 필요할 것이며, 이러한 측면에서 빈도주의자 통계학의 귀무가설검정을 이용한 임상연구방법보다는 베이지안 통계학을 활용한 임상연구방법이 실제 임상에 더 적합할 것으로 생각된다.

특히 한의학은 그 이론 안에서 이미 질병을 일회적인 것이 아닌 시간에 따른 변화의 흐름으로 파악하였으며, 과거의 질병이 현재에 영향을 미치는 시간에 따른 정보의 축적에 대하여 논하고 있다. 예를 들면 傷寒論의 太陽病에서 '兩經이나 三經의 증상이 동시에 출현하는 것을 合病이라 하고, 한 經의 증상이 완전히 消除되지 않았는데 또 다른 한 經의 증상이 나타나는 것을 幷病이라 하며, 誤治 후에 그 證候性質이 전환되는 것을 變證이라고 한다'고 하였는데[69], 이는 베이지안 통계학에서의 '새로운 자료가 얻어질 때마다 가설이 갱신(update)되는 모습'[5]과 유사하다고 볼 수 있을 것이다.

이와 같은 이론적 관점과 본론에서 제시한 예시와 같이, 한의학에서의 데이터를 해석하는데 있어서 베이지안 통계학을 이용한 방법이 충분히 사용될 수 있다. 물론 본론에서 제시된 질환인 부비동염의 변증 과정에서 이렇게 증상만으로 명료하게 구별하기에는 부족함이 있으며, 영상 자료를 비롯한 다른 진단법들을 이용하여 종합적으로 판단하는 것이 합리적임은 분명하다. 그러나, 이와 같이 베이지안 통계학을 진단에 활용하는 것은 임상연구방법론에 있어서 새로운 의미가 있다고 생각된다.

본 논문에서는 베이지안 통계학을 이용한 한방이비인후과 질환에 대한 진단법을 소개하였는데, 라카토스의 연구프로그램(Lakatos methodology of scientific research programmes)은 소개하지 못한 아쉬움이 있다. 라카토스의 연구프로그램의 경우 일련의 과정을 설명하는데 있어서 베이지안 통계학을 이용한 방법보다 과정이 매우 복잡하고 많은 지면이 요구되므로 추후 별도의 논문으로 자세하게 발표하고자 한다. 검정을 통한 이론의 증명은 중요한 일이지만, 이 과정에서 발생하는 이분법적 사고의 극복을 위한 노력 역시 매우 중요할 것이다. 무엇보다도, 연구의 목적이 빠른 가설의 선택과 의사의 결정이 아니라, 임상에 도움이 될 수 있는 '실재의 추구'라는 점을 생각하고 연구를 진행하는 것이 중요하다고 생각된다.

Ⅳ. 결 론

본 연구에서는 현재 임상연구의 과정에서 사용되고 있는 귀무가설과 대립가설을 이용한 검정법의 문제점과 그에 대한 대안에 대하여 고찰해보았으며 그 과정을 통하여 다음과 같은 결론을 얻었다.

첫째, 현재 임상연구에 사용되는 가설검정의 방법은 피셔의 유의성검정과 네이만-피어슨의 가설검정을 20세기의 심리학자들이 임의로 혼합시켜서 만든

것이다.

둘째, 피셔는 유의성을 판단하기 위하여 창안한 p값을 여러 가지 크기로 유연하게 사용하였고 상황에 맞게 판단했음에 반하여 후대의 학자들은 p값이 0.05인 경우를 중시하여 이를 귀무가설을 기각하고 유의성을 검정하는 기준으로 삼았다.

셋째, 귀무가설은 증명되거나 확립될 수는 없지만 반증은 가능하며, 이는 실제로 증명하고자 하는 명제를 직접 증명하기 어려울 때 사용되는 귀류법과 논리적 구조가 유사하다.

넷째, 귀무가설과 대립가설을 이용한 가설검정법에서 두 가설은 상호배반적이고 둘 중 하나를 선택해야만 하는 조건이 전제되어 있으며, 이러한 점에서 이분법적 사고에 빠질 수 있는 단점이 있다.

다섯째, 임상연구방법론에서 이분법적 사고를 극복하고 실재를 추구하는 것이 중요하며 따라서 그 보완 및 대체의 방법론으로 베이지안 통계학과 라카토스의 연구프로그램이 대안이 될 수 있다.

V. 감사의 글

본 연구는 2018년도 부산대학교병원 임상연구비 지원으로 이루어 졌음.

ORCID

Seung-Pyo Nam
(https://orcid.org/0000-0003-4455-3478)

Jae-Min Bae
(https://orcid.org/0000-0002-1266-0386)

Kang Kwon
(https://orcid.org/0000-0002-7250-2603)

References

1. Folks JJ. Ideas of Statistics. New York:John Wiley&Sons. 1981.
2. Park SH. The Past and Present of Statistical Study, and the Role and Vision of Data Science in the 4th Industrial Revolution Era. The National Academy Sciences(Natural Science). 2017;56(2):53-82.
3. Stacey Plichta Kellar, Elizabeth A. Kelvin. Statistical Methods for Health Care Research the 6th Edition. Paju:Koonja Publisher. 2017:72-3.
4. Jo HC. Misinterpretation about Hypothesis Setting and Hypothesis Tests. The Review of Industrial economics and Manegement. 1999.
5. Lee KH. Review on problems with null hypothesis significance testing in dental research and its alternatives. The Journal of the Korean academy of Pediatric dentistry. 2013;40(3):223-32.
6. Lee HK. Principles of Hypothesis Test Ⅰ. Pediatric Infection and Vaccine. 1997;4(1):183-192.
7. Jo JK. The Biometry-Mendelian Controversy in the History of Statistics. Communications of the Korean Statistical Society. 2008;15(3):303-24.
8. Pearson K. On the fundamental conceptions of biology. Biometrika. 1902;1:320-44.
9. David salsbrug. The lady tasting tea : How Statistics Revolutionized Science in the Twentieth century. Paju:Jayu academy. 2012:20,22,29,82,114-16,124-7,129-31.
10. Heo MH. History of statistics colloquium. Seoul:Freedom academy. 1991:9-5.
11. Lee KH. History of Probability and Statistics. J of Elementary Mathematics Educationb in

Korea. 1997;1(1):53-65.
12. Fisher RA. The Design of Experiments. 1st ed. 1935. [The 9th edition published by Macmillan. New York, in 1971.]
13. Jo JK. Statistics, catch big data. Seoul: hankookmunhwasa publisher. 2017:222-24.
14. Jo JK. A Study on the History of Statistics in the Early Twentieth Century Focused on Statistical Tests and Psychology. Journal for History of Mathematics. 2013;26(4):277-99.
15. Gigerenzer G. Probabilistic Thinking and the Fight against Subjectivity, The Probabilistic Revolution. Vol. 1: Ideas in History (edited by L. Krüger, L. Daston, M. Heidelberger). MIT Press. 1987:11-33.
16. Gigerenzer G. Rationality for Mortals. How People Cope with Uncertainty. Oxford University Press. 2008.
17. Gigerenzer G, Murray DJ. Cognition as Intuitive Statistics. L. Erlbaum Associates. 1987.
18. Jo JK. History of the Error and the Normal Distribution in the Mid Nineteenth Century. Communications of the Korean Statistical Society. 2008:15(5):737-52.
19. An YO, Yu GY, Park BJ, et al. Epidemiology : The principles and applications. Seoul: Seoul National Univ. Publisher. 2015:254.
20. John Gribbin, Mary Gribbin. Science : a History in 100 Experiments. Seoul: Yeamoonarchive. 2017:15.
21. Oh KU. Study on Testing Hypothesis. The Korean Economic Review. 1989;17(1):45-59.
22. Lee JY. Medical statistics at a glance 3rd edition. Seoul:Epublic Publisher. 2010:50-2.
23. Compilation committee of clinical research methodology. Clinical research applications to practice. Seoul:Hyeonmun publisher. 2019:104.
24. Agresti A, Finlay B. Statistical methods for the social sciences. Upper Saddle River, NJ:Prentice-Hall. 1997.
25. SB Hulley, SR Cummings, WS Browner, DG Grady, TB Newmann. Designing Clinical Research the 4th Edition. Seoul:Koonja Publisher. 2015:58.
26. Noto Hiroshi. Clinical statistics ready for daily care. Seoul:Daehanuihagseojeog. 2009:105.
27. Cheong HB. Popper's Philosophy of Science and Methods of Value Inquiry. Research in Social Studies Education. 2017;24(1):1-10.
28. Popper, K. R. Conjectures and Refutations: The Growth of Scientific Knowledge, New York & Evanston: Harper & Row Publishers. 1963.
29. Cheong HB. A Study of an Inquiry Method of Knowledge in Social Studies Education Based on Popper's Searchlight Theory of Knowledge. Research in Social Studies Education. 2016;23(1):25-40.
30. Jang D. Kuhn & Popper : There is something special about Science. Seoul:Gimyeongsa. 2014:73-5,164.
31. Fisher, R. The Design of Experiments, 9th ed. New York:Macmillan. 1971[1935].
32. S. Nassir Ghaemi. A Clinician's Guide to Statistics and Epidemiology in Mental Health. Seoul:Hwangsogeoleumacademy. 2015:98,99,105,247,249,265,266,269.
33. Jo HM. Secondary school mathematics teachers' understanding of proof by

contradiction in evaluating students' proofs. Graduation school of Seoul National Univ. Master's degree. 2009:20.
34. Hwang JY, Shin BM. An Analysis of Teacher's Knowledge about Reductio Ad Absurdum - Focused on 'Subject Matter Knowledge' and 'Knowledge of Students' Understanding -. J. Korean Soc. Math. Ed. Ser. A. 2016;55(1):91-106.
35. Lee GD, Hong GJ. A study on improvement of introductions and Applications of 'Proof by contradiction' in testbooks. J Korea Society Educational Studies in Mathematics. School Mathematics. 2016;18(4):839-56.
36. S. G. Krantz, An Episodic History of Mathematics: Mathematical Culture through Problem Solving, Maa, 2010.
37. Hyun JS, Park CJ. Proof by Contradiction of Creative Research's Contradiction Resolution and the Representation with the Butterfly Diagram. Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology. 2016;6(4):407-15.
38. Lee HK. Principles of Hypothesis Test II. Pediatric Infection and Vaccine. 1997; 4(2):314-19.
39. Gelman A, Stern H. The difference between 'significant' and 'not significant' is not itself statistically significant. Am Statistician. 2006;60(4):28-31.
40. Royall RM. The effect of sample size on the meaning of significance tests. Am Statistician. 1986;40(4):313-15.
41. Hand DJ. Data mining: statistics and more?. Am Statistician. 1998;52(2):112-8.
42. Wang S, Campbell B. Mr. Bayes goes to Washington. Science. 2013;339:758-9.
43. Efron B. Why isn't everyone a Bayesian(with discussion)?. Am Statist. 1986;40(1):1-11.
44. Nurminen M, Mutanen P. Exact Bayesian analysis of two proportions. Scand J Stat. 1987;14(1)67-77.
45. Diaconis P, Freedman D. On the consistency of Bayes estimate (with discussion). Ann Math Stat. 1986;14(1):1-67.
46. Rhee YE. Bayesianism : The journey from rationality to objectivity. Seoul: Hangugmunhwasa. 2015.
47. Cheon HD. Is Science a Bayesian Machine? : A Critical Review of Rhee(2015)'s Bayesianism. The Korean Journal for the Philosophy of Science. 2016;19(3):87-107.
48. Yeo YS. Luc Bovens and Stephen Hartmann 『Bayesian Epistemology』 - Formal epistemology and Bayesianism. Philosophy and Reality. 2006;3:262-71.
49. Yeo YS. Bayesianism and Eliminative Inductivism. Korean Journal of Logic. 2004;7(2):121-46.
50. Yeo YS. Bayesian Prior Probability and Scientific Objectivity. Philosophy inquiry. 2007;22:147-72.
51. Lee YC. From paradigm to reality : from Constructivism Science to Realistic Science realism. Journal of Governmental Studies. 2010;16(1):155-79.
52. Jha, Stefania R. The Bid to Transcend Popper, and the Lakatos-Polanyi Connection. Perspectives on Science. 2006:14(3):318-46.
53. Kim JS. Lakatos' Methodology of Scientific Research Program and The Rationality of

Science. Journal of the Society of Philosophical Studies. 1992;20:151-81.
54. Yang KE. Development of International Relations Theory Analyzed by Lacatos Research Methodology. Social Science Education Research. 2014;16:90-9.
55. Lakatos I. The Methodology of Scientific Research Programmes. Seoul:acanet. 2002: 73,86,88-94,121.
56. Park IJ. The Development of Modern Evolutionary Theories from the Perspective of Lakatos' Research Program. Korea National University of Education. Master's Thesis. 2018:1-52.
57. Chalmers AF. What is this thing called Science?. Seoul:Seogwangsa. 2003:190.
58. The Society of Korean Medicine Ophthalmoliogy, Otolaryngology & Dermatology. Korean Medicine Ophthalmoliogy, Otolaryngology. Paju: Globoolks. 2019:185-7.
59. Yoshiyuki Wakui, Sadami Wakui. Statistical diagram. Seoul:Seongandang. 2018:134-5.
60. Rosenthal R. Effect size estimation, significance testing, and the file-drawer problem. J Parapsychol. 1992;56:57-8.
61. Vaughan GM, Corballis MC. Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. Psychol Bulletin. 1969;72(3):204-13.
62. Park EJ. 'The Heuristic' and 'The Methodology of Scientific Research Programmes' and 'The Philosophical Reconstruction of History of Science'. Philosophy of Science. 2002;8:185-99.
63. Jang D. There is something special about Science. Seoul:Gimyeongsa. 2008:164-165.
64. Lee TH. Overcoming dichotomous thinking for understanding Lao tse's Tao Re Ching. The Korean Journal of East West Science. 2015;18(1):1-13.
65. Rhee YE, Park IH. Bayesian Epistemology. Korean Journal for the Philosophy of Science. 2015;18(2):1-14.
66. Basic Appl. Soc. Psych. 2015;37:1-2. Available from : [URL] https://www.tandfonline.com/doi/full/10.1080/01973533.2015.1012991.
67. Future & Science in Hangyeorye Newspaper company [cited 2019 April 5]. Available from : [URL] http://m.hani.co.kr/arti/science/science_general/888909.html
68. Science on [cited 2017 August 16]. Available from : [URL] http://scienceon.hani.co.kr/540289
69. Moon JJ, An GS, Kim SH, Eom HS, Ji GY, Kim JB. Sanghanron Jeonghae. Seoul: Kyunghee University Press. 1996:20.