http://dx.doi.org/10.5933/JKAPD.2013.40.3.223 ISSN (print) 1226-8496 ISSN (online) 2288-3819

Review on Problems with Null Hypothesis Significance Testing in Dental Research and Its Alternatives

Kwang-Hee Lee

Department of Pediatric Dentistry, College of Dentistry, Wonkwang University

Abstract

There are many problems in evaluating study results by *p* value in null hypothesis testing for dental research. It is a logical fallacy to conclude that the null hypothesis is true when the it is not rejected. There are much serious misunderstanding about *p* value, and researchers should be cautious about interpreting *p* value in writing papers. As alternatives to complement or replace the null hypothesis significance testing, effect size, confidence interval, and Bayesian statistics are introduced.

Key words : Null hypothesis, Significance testing, *p* value, Effect size, Confidence interval, Bayesian statistics

Ⅰ. 서 론

2011년 3월, 미국 대법원은 오랜 심의 끝에 임상시험의 결과가 통계적으로 유의(significant)하지 않더라도 여전히 중요(important)할 수 있다고 판결하였다. 제약회사 Matrixx Initiatives는 처방전 없이 살 수 있는 약인 Zicam이 후각상실을 일으키는 부작용이 있지만 그 발생빈도가 통계적 유의수준에 미치지 않았기 때문에 이 부작용을 알릴 이유가 없었다고 주장하였으나 대법원은 그 주장을 기각하였다[1]. 'Matrixx Initiatives사 대 Siracusano'로 알려진 이 사건은 통계학자들의 논평과 함께 언론에 보도되었다[2].

치의학 학술지에 게재되는 대부분의 논문들은 특정 *p*값을 기준으로 연구결과를 판단하여 '유의하다' 또는 '유의하지 않다'라는 결론을 내리고 있다. 그리고 *p*값이 0.05보다 큰 결과가 나온 연구는 성공하지 못한 연구로 간주되어 학술지에 제출하거나 게재하지 않는 경향이 있다. 이렇게 *p*값을 기준으로 하여 기계적으로 연구를 평가하는 관행이 객관적이고 과학적인 방법으로 간주되고 있다.

귀무가설 또는 영가설의 유의성 검정(null hypothesis significance test)은 Fisher의 유의성 검정과 Neyman과 Pearson의 가설검정이 혼합된 것이다. Fisher의 *p*값은 귀무가설을 기각하기 위한 기준으로 고안되었으며, Neyman과 Pearson은 연구가설과 귀무가설 간의 판단 기준으로서 1종 오류의 확률인 α를 사용하였다. 이 두 방법은 철학적 배경이 다르나, 점차 하나의 체계로 진화하였다. 그러나, 귀무가설검정은 결코 완전하거나 최선의 방법이 아니며, 귀무가설검정에 심각한 결함이 있다는 사실은 수십년 전부터 많은 통계학자들에 의해 지적되어 왔다[3-7].

이 논문에서는 치의학 연구에서 사용되고 있는 귀무가설 유의성 검정에 내포된 문제점을 살펴보고, 귀무가설검정을 보완하거나 대체할 수 있는 대안을 찾아보고자 한다.

Ⅱ. 귀무가설 유의성 검정의 문제점

1. 논리적 오류

1) 후건 긍정의 오류

귀무가설은 연구가설(대립가설, alternative hypothesis)의 반대가 되는 가설로서, 실제로 연구에서 알고자 하는 효과가 없다고 가정하는 가설이다. 유의성 검정에서 연구가설을 검정하

Corresponding author : Kwang-Hee Lee
Department of Pediatric Dentistry, College of Dentistry, Wonkwang University, 460 Iksandaero, Iksan, 570-749, Seoul
Tel: +82-63-850-2955 / Fax: +82-63-851-5324 / E-mail: kwhlee@wonkwang.ac.kr
Received July 10, 2013 / Revised July 16, 2013 / Accepted July 24, 2013

※ This research was supported by Wonkwang University in 2013.

지 않고 연구가설의 반대인 귀무가설을 검정을 하는 것은 후건 긍정의 오류를 피하기 위함이다. 하지만, 귀무가설이 기각되지 않은 경우에 귀무가설이 옳다고 해석하는 것도 후건 긍정의 오류이다.

삼단논법(syllogism)의 대전제, 소전제, 결론에서, 조건 명제 'p이면 q이다.' 에서 p를 전건, q를 후건이라고 할 때, 긍정논법(modus ponens)은

> p이면 q이다.
> p이다. (전건의 긍정)
> 그러므로 q이다. (후건의 긍정)

이며 타당하다. 부정논법(modus tollens)은

> p이면 q이다.
> q가 아니다. (후건의 부정)
> 그러므로 p가 아니다. (전건의 부정)

이며 타당하다. 그러나,

> p이면 q이다.
> p가 아니다. (전건의 부정)
> 그러므로 q가 아니다. (후건의 부정)

는 타당하지 않으며, 전건 부정 추리에서 후건 부정을 타당한 결론으로 받아들이는 것을 전건 부정의 오류(fallacy of negating the antecedent)라고 한다. 또한,

> p이면 q이다.
> q이다. (후건의 긍정)
> 그러므로 p이다. (전건의 긍정)

도 타당하지 않으며, 후건 긍정 추리에서 전건 긍정을 타당한 결론으로 받아들이는 것을 후건 긍정의 오류(fallacy of affirming the consequent)라고 한다[8].

이것을 귀무가설 유의성 검정에 적용시켜 보면,

> 연구가설이 옳다면 이 자료가 발생한다.
> 이 자료가 발생하였다. (후건의 긍정)
> 그러므로 연구가설은 옳다. (전건의 긍정)

는 타당하지 않으며 후건 긍정의 오류이다. 따라서, 후건 긍정의 오류를 피하기 위하여 연구가설을 검증하는 대신에 연구가설의 반대인 귀무가설을 검정한다.

귀무가설이 기각될 때 귀무가설이 옳지 않다고 해석하는 것은 다음과 같이 논리적으로 타당한 부정논법이다.

> 귀무가설이 옳다면 이 자료는 발생하지 않는다.
> 이 자료가 발생하였다. (후건의 부정)
> 그러므로 귀무가설은 옳지 않다. (전건의 부정)

그러나, 귀무가설이 기각되지 않았을 때 귀무가설이 옳다고 해석하는 것은 다음과 같이 후건 긍정의 오류이다.

> 귀무가설이 옳다면 이 자료는 발생하지 않는다.
> 이 자료가 발생하지 않았다. (후건의 긍정)
> 그러므로 귀무가설은 옳다. (전건의 긍정)

2) 확률적 표현이 포함된 삼단논법 추리의 오류

양자역학에서 입자를 확률 분포로 나타내듯이, 현대 과학에서는 세계를 고정된 실체가 아니라 관찰된 값의 확률 분포로 본다. 현대 통계학이 탄생시킨 귀무가설검정도 확률 분포에 근거하고 있다. 귀무가설검정의 타당한 추리인 '귀무가설이 옳다면 이 자료는 발생하지 않는다 - 이 자료가 발생하였다 - 그러므로 귀무가설은 옳지 않다' 를 확률적으로 표현하면

> 귀무가설이 옳다면 이 자료는 발생할 확률이 매우 낮다.
> 이 자료가 발생하였다. (후건의 부정)
> 그러므로 귀무가설이 옳지 않을 확률이 매우 높다. (전건의 부정)

가 된다.

이것이 타당하지 않으며 오류를 내포하고 있다는 것을 설명하기 위해 통계학자들이 자주 인용하는 예들은 다음과 같다[5,9].

> 어떤 사람이 미국인이면 아마 국회의원이 아닐 것이다.
> 이 사람은 국회의원이다. (후건의 부정)
> 그러므로 이 사람은 아마 미국인이 아닐 것이다. (전건의 부정)

> 주사위를 던지면 아마 1보다 큰 수가 나타날 것이다.
> 1이 나타났다. (후건의 부정)
> 그러므로 아마 주사위를 던지지 않았을 것이다. (전건의 부정)

위의 두 예는 '후건 부정 - 전건 부정' 의 타당한 형식이지만 결론이 옳지 않은 것이 확실하다. 여기서 '아마 —일 것이다.' 는 '—일 확률이 매우 높다.' , '아마 —가 아닐 것이다.' 는 '—일 확률이 매우 낮다.' 는 확률적 표현이다. 이것은 위의 귀무가설검정의 확률적 표현과 형식이 정확히 일치한다. 그렇다면 확률 분포에 근거한 귀무가설검정은 논리적 오류를 필연적으로 내포하게 된다.

이와 관련하여, 확률적 표현을 포함하는 삼단논법 추리가 배중률을 위반하고 있다는 설명이 있다[10]. 배중률(principle of excluded middle, 排中律)은 'A는 A가 아니고 비(非)A도 아닌 어떤 것일 수는 없다' 로 표현된다. 즉, A와 그의 부정 비(非)A사이에 제3의 중간적인 존재를 인정하지 않으며, 명제는

참과 거짓 두 가지만 있을 수 있다고 해석된다. 삼단논법의 수학적 증명은 결정적 결론에 이르게 하지만, 확률 분포에 근거한 귀무가설검정은 불확실성을 내포한 확률적 결론에 이르게 하므로, 귀무가설검정의 논리는 삼단논법에 해당하지 않는다고 볼 수도 있다.

2. *p*값에 대한 오해[11-16]

□ 오해 - "*p*값이 0.05이면 귀무가설이 옳을 확률이 5%이고 연구가설이 옳을 확률이 95%이다."

이것은 *p*값에 대한 오해 중에서 가장 널리 퍼져 있고 가장 유해한 것으로 지적되어 왔다[17,18]. 이 오해는 귀무가설이 옳다는 가정 하에서 관찰된 자료의 확률 $P(D|H_0)$을, 관찰된 자료에 근거하여 귀무가설이 옳을 확률 $P(H_0|D)$로 혼동한 것이다. 〔여기서 D는 자료(data), H_0은 귀무가설(null hypothesis), '|'는 '주어진(given)'의 뜻으로 $P(A|B)$는 B가 주어졌을 때 A의 조건부확률(conditional probability)을 가리킨다.〕

*p*값, 즉 $P(D|H_0)$는 '귀무가설이 옳고 연구 모형이 적합하며 표집이 무작위로 행해졌을 때, 관찰된 자료 및 그보다 극단적인 자료가 우연에 의해 발생할 확률'이다. *p*값은 귀무가설이 옳다는 전제 하에 계산된 것이므로 귀무가설이 옳을 확률이 될 수 없고 '1 - *p*값'을 연구가설이 옳을 확률로 할 수도 없다.

귀무가설검정을 사용하는 빈도주의자 통계(frequentist statistics)에서는 가설에 확률을 붙이지 않으며, 붙일 수도 없다. 그러므로 *p*값으로 가설의 확률에 관해 추리할 수 없다. 가설의 확률을 계산하는 유일한 방법은 베이지안 통계(Bayesian statistics)의 베이즈 정리(Bayes theorem)를 사용하는 것이다.

베이지안 통계를 사용한 계산에 따르면, *p*값 0.05, 0.01, 0.001에 해당하는 귀무가설의 사후확률 $P(H_0|D)$는 표본 수가 50일 때 각각 0.52, 0.22, 0.034이고 표본 수가 100일 때 각각 0.60, 0.27, 0.045로서, p값과 $P(H_0|D)$ 간의 불일치가 뚜렷하게 나타났다. 또한, *p*값 0.05가 산출된 자료의 경우, 귀무가설의 사후확률이 최소 0.30이었다[19].

따라서, *p*값은 귀무가설의 사후확률과 일치하지 않고 귀무가설에 대항하는 근거를 과장하는, 즉 통계적으로 유의한 결과를 얻기가 쉬운 측정치이므로, 0.05와 같이 *p*값이 아주 작지 않은 연구는 그 타당성이 의심되며, 정밀한 가설을 검정하는 경우에는 *p*값의 공식적 사용이 권장되지 않는다[20-22].

□ 오해 - "귀무가설이 기각되었을 경우, $p=0.05$는 1종 오류의 확률이 5%임을 의미한다."

1종 오류는 효과가 없을 때 효과가 있다는 결론을 내리는 오류, 즉 가양성(false positive)의 확률이다. 귀무가설이 옳지 않아 기각되었다면, 즉 효과가 있다면, 1종 오류 α를 생성할 수 없고, 효과가 있을 때 효과가 없다는 결론을 내리는 2종 오류 β, 즉 가음성(false negative)의 확률만 문제가 된다. 이 경우에는 '1-β' 인 통계검정력(power of statistics)이 더 중요하다. 통계검정력은 연구가설이 옳을 때 연구가설을 채택할 확률이다.

귀무가설검정에는 두 개의 아주 다른 유의성 측정법이 함께 들어 있다. 하나는 Fisher의 *p*값으로, 귀무가설에 대항하는 근거 강도를 추론하는 지표이고, 자료에 근거한 무작위 변수이며, 개별 연구에 적용될 수 있다. 다른 하나는 Neyman-Pearson 가설검정의 α수준으로, 이 검정은 1종 오류 α(귀무가설을 잘못 기각할 확률)와 2종 오류 β(귀무가설을 잘못 채택할 확률)를 최소화하는 것과 관계가 있다. α는 귀무가설을 기각하거나 채택하기 위한 기준이고 근거의 측정 수단이 아니며, 사전에 선택된 고정 값으로서 무작위 변수가 아니고, 동일 집단으로부터 반복해서 무작위로 표집하는 경우에 적용되며 단일 실험에는 적용되지 않는다[23].

이렇게 양립할 수 없는 *p*와 α가 혼합되면서 통계적 유의성이 정확히 무엇을 의미하는지에 대해 큰 혼동이 있게 되었고 *p*값이 1종 오류의 확률로 잘못 해석되고 있다. *p*와 α의 함수적 관계에 대해 계산한 연구[24]에 따르면, $p=0.05$는 $\alpha=0.289$로, $p=0.01$은 $\alpha=0.111$로 산출되었다. 여기에서도 *p*값은 귀무가설을 기각하기에 충분하지 않은 것으로 나타났다.

□ 오해 - "유의수준을 $p=0.05$로 설정하였을 경우, 1종 오류의 확률은 5%일 것이다."

이것은 귀무가설 기각 이후가 아니라 실험을 하기 전에 1종 오류의 확률을 예측하는 것을 의미한다. 하지만, 1종 오류의 확률은 귀무가설이 옳을 사전 확률(prior probability)에 의존한다.

(1) 실험 전에 귀무가설이 옳다는 것을 알고 있다면, 귀무가설의 기각이 옳지 않을 확률인 1종 오류의 확률이 실제로 5%이다.
(2) 실험 전에 귀무가설이 옳지 않다는 것을 알고 있다면, 1종 오류의 확률은 0이다. 치아우식증에 대한 불소의 효과와 같이 이미 효과가 입증된 치료법에 관한 실험을 한다고 할 때, 귀무가설이 옳지 않은 것을 이미 알고 있으므로 귀무가설의 기각이 옳지 않을 확률인 1종 오류의 확률은 0이다. 이 경우에는 귀무가설의 기각이 모두 옳기 때문에 *p*값이 얼마인지는 문제가 되지 않는다.
(3) 실험 전에 귀무가설이 옳은지 옳지 않은지 확실히 알지 못한다면, 1종 오류의 확률은 0에서 5% 사이에 있다.

□ 오해 - "유의한(significant) 결과는 중요한(important) 결과이다."

통계적 유의성은 실제적 유의성이 아니다. 귀무가설검정은 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 *p*값이 직접적인 정보를 제공하지 않는다는 사실을 간과하는 데에서 기인한다[25].

통계적 유의성과 실제적 중요성 사이에 네 가지 경우가 있을 수 있다. 실제로 중요하지 않은 효과가 통계적으로도 유의하지 않은 경우와 실제로 중요한 효과가 통계적으로도 유의한 경우는 문제가 없다. 그러나 실제로 중요한 효과가 통계적으로 유의하지 않은 경우와 실제로 중요하지 않은 효과가 통계적으로 유의한 경우는 문제가 된다.

통계적으로 유의한 결과가 나온 연구들이 주로 학술지에 제출되고 게재되고 있으나, 국제의학저널편집자위원회(International Committee of Medical Journal Editors, ICMJE)는 부정적 연구를 게재할 의무(obligation to publish negative studies)를 언급하고, 통계적 유의성이 결여되었다는 이유로 연구결과의 제출이나 게재를 못하는 것이 출판편향(publication bias)의 중요한 원인이라고 하였다[26].

유의한 결과가 중요한 결과가 아닌 이유 중 하나는 *p*값이 표본 수에 의존하기 때문이다. 귀무가설 기각의 가능성은 표본이 작을 때에는 작고 표본이 클 때에는 크며, 표본이 매우 클 때에는 어떤 결과도 유의하게 나타난다. 큰 효과 크기를 가진 작은 표본 연구는 작은 효과 크기를 가진 큰 표본 연구와 동일한 *p*값을 산출할 수 있다[27]. 오늘날의 대규모 자료(big data)를 사용하는 연구에서는 거의 모든 귀무가설에서 *p*값이 아주 작게 나타날 수 있다[28]. 그렇기 때문에 *p*값은 고정된, 객관적 의미를 가지고 있지 않다.

*p*값이 표본 크기의 함수라는 이 사실은 다음의 세 가지 오해와도 관련이 있다.

□ 오해 - "*p*값이 작을수록 효과가 크다."

귀무가설검정에서는 단순히 표본의 수를 늘림으로써 *p*값을 작게 만들 수 있다. t값을 구하는 공식은 다음과 같다.

$$t = \frac{\overline{X_1} - \overline{X_2}}{s_{\overline{X_1} - \overline{X_2}}}, \qquad s_{\overline{X_1} - \overline{X_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

여기서 $\overline{X}$는 표본 평균, s는 표본 표준편차, n은 표본 크기이다. 그러므로 t값은 (1) 평균의 차이, 즉 효과 크기와 (2) 표본 크기에 비례한다. t값이 클수록 *p*값이 작아지므로, *p*값은 (1) 효과 크기와 (2) 표본 크기에 반비례한다. 표본이 충분히 크다면 아무리 작은 효과라도 귀무가설을 기각할 수 있다. 따라서, *p*값은 표본이 클수록 그 유용성이 감소한다.

□ 오해 - "*p*값이 0.05보다 큰 연구와 0.05보다 작은 연구는 그 결과가 상충된다."

동일한 귀무가설을 사용한 두 연구의 결과에서 효과 크기가 동일하여도 표본 크기에 따라 *p*값이 유의하거나 유의하지 않을 수 있다.

다음 Table 1의 예를 보면, A군과 B군의 표본 수가 각각 10명인 경우에는 평균이 각각 5.50, 7.50이고 *p*값이 0.157로서 평균의 차이 2.00가 유의하지 않으나, 측정치 당 표본 수를 2배로 늘리면 평균은 변화가 없으나 *p*값이 0.038이 되어 동일한 차이가 유의한 것으로 바뀐다(Table 1).

□ 오해 - "*p*값이 동일한 연구들은 귀무가설에 대해 동일한 근거를 제공한다."

동일한 귀무가설을 사용한 두 연구의 결과에서 효과 크기가 달라도 표본 크기에 따라 *p*값이 같을 수 있다.

위의 Table 1의 예에서 A군과 B군의 표본 수가 각각 10명인 경우에 B군의 측정치가 모두 1.00씩 높을 경우에는 평균이 각각 5.50, 8,50이 되고 *p*값은 0.040가 되어, 표본 수가 각각 20명인 경우보다 평균의 차이가 50% 더 큼에도 불구하고 *p*값은 거의 동일하다(Table 2). 이것은 위에 언급된 '큰 효과 크기를 가진 작은 표본 연구는 작은 효과 크기를 가진 큰 표본 연구와 동일한 *p*값을 산출할 수 있다.' 의 예이다.

□ 오해 - "유의한 결과는 신뢰할만한(reliable) 결과이다."

신뢰도(reliability)는 반복 연구에서 동일한 결과를 얻을 확률을 의미하나, 유의성 검정에 의한 신뢰도는 단일 연구가 아니라 일련의 연구에서 실시할 때 의미를 가진다. *p*값은 '반복 연구가 동일한 결론을 산출하지 않을 확률' 이 아니다.

지식의 축적적 발전을 촉진하는 것은 개별 연구의 *p*값이 아니라 이전 연구의 결과를 체계적으로 반복하고 확대하는 것이다. *p*값을 고안한 Fisher 자신도 단일 연구의 유의한 결과는 단

Table 1. Difference in *p* value according to the number of samples

	N	Mean ± SD	*p* value
Group A	10	5.50 ± 3.03	0.157
Group B	10	7.50 ± 3.03	
Group A : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10			
Group B : 3, 4, 5, 6, 7, 8, 9, 10, 11, 12			
	N	Mean ± SD	*p* value
Group A	20	5.50 ± 2.95	0.038
Group B	20	7.50 ± 2.95	
Group A : 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9, 10, 10			
Group B : 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9, 10, 10, 11, 11, 12, 12			

Table 2. Difference in effect size with same *p* value

	N	Mean ± SD	*p* value
Group A	20	5.50 ± 2.95	0.038
Group B	20	7.50 ± 2.95	
Group A : 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9, 10, 10			
Group B : 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9, 10, 10, 11, 11, 12, 12			
	N	Mean ± SD	*p* value
Group A	10	5.50 ± 3.03	0.04
Group B	10	8.50 ± 3.03	
Group A : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10			
Group B : 4, 5, 6, 7, 8, 9, 10, 11, 12, 13			

지 일시적으로 신뢰할 수 있다고 하였으며 연구의 반복을 첫째 원칙으로 강조하였다[29,30].

□ 오해 - "유의성은 인과 관계(causality)의 증거이다."

인과 관계에 대한 가설 자체가 옳지 않거나 다중 인과 관계(multiple causality)가 존재할 경우에는 모집단의 변수 사이의 인과 관계는 유의수준과 상관 없이 존재하지 않을 수 있다.

□ 오해 - "*p*값의 유의성 여부에 근거하여 과학적 결론을 내려야 한다."

우리가 연구에서 알고자 하는 것은 효과의 크기이나, *p*값은 효과의 크기를 직접 알려주지 않기 때문에 *p*값에 의존하여 결론을 내리는 것은 효과 크기가 중요하지 않다고 주장하는 것과 마찬가지이다. 또한, 과학적 결론은 반복 연구에서 동일한 결과가 나올 확률, 즉 신뢰도에 기초해 있으므로, 신뢰도와 상관이 없는 *p*값을 근거로 과학적 결론을 내릴 수는 없다.

고정된 유의수준을 사용하면 $p=0.049$인 연구와 $p=0.051$인 연구 사이에 무의미한 구별을 하게 된다. 가설검정 자체가 근사적인 가정들에 근거한 것이기 때문에 그러한 미소한 차이로 평가하는 것은 타당하지 않다. *p*값은 가설의 연역적 타당성, 선행 연구의 결과들과 같은, 가설에 관한 다른 모든 증거들과 더불어 평가해야 한다.

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.

Ⅲ. 귀무가설 유의성 검정을 보완하거나 대체할 수 있는 대안들

1. 효과 크기(effect size)[31-33]

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. 상관관계 연구에서는 상관 계수 r로, 두 집단의 실험일 경우에는 d로, 분산 분석일 경우에는 f로 나타낸다. 여기서 d는 평균의 차이를 공통의 표준 편차로 나눈 값이다.

*p*값과 달리, 효과 크기는 표본 크기에 민감하지 않다는 이점이 있다. 효과 크기는 표본 수에 따라 증가하거나 감소하지 않는 안정적인 수치이다. 표본 크기의 증가로 새로운 효과의 발견이나 효과 크기의 증가를 기대할 수 있으나, 효과 크기가 표본 크기의 함수는 아니다.

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. *p*값 대신에 효과 크기를 사용하기보다는 *p*값을 보완하기 위해 효과 크기를 사용하는 것이 바람직하다.

2. 신뢰구간(信賴區間, confidence interval)

국제의학저널편집자위원회(ICMJE)는 연구 결론을 검정하기 위해 귀무가설에만 의존하는 것에 대해 경고하고 신뢰구간의 보조적 사용을 제안하였다[26]. 영어 출판물에서 *p*값의 배타적 사용은 1995-1996년의 41%에서 2005-2006년의 21%로 감소하였고 신뢰구간의 사용은 시간이 지나면서 증가하였다[34].

신뢰구간을 사용하면 단지 *p*값이 0.05보다 작다는 이유로 효과가 없다고 단정하는 것을 피하고 연구가설을 지지하는 증거를 찾을 가능성이 커진다. 사실 *p*값이 유의하지 않다는 것이 의미하는 것은 관찰된 효과에 대한 설명으로서 우연이나 표집 오차 등을 배제할 수 없다는 것이 전부이다. 신뢰구간은 *p*값보다 훨씬 더 많은 정보를 제공하며 검정보다 추정(estimate)이 우월함을 보여준다[35].

신뢰구간은 구간의 폭을 통해 추정의 정밀도 또는 신뢰도를 나타내며, 결과가 통계적으로 유의하기보다 실제적으로 유의한지 여부를 더 쉽게 볼 수 있다. 귀무 값을 포함하지 않는 95% 신뢰구간은 가설을 0.05 수준에서 기각하는 것과 동등하다[36].

신뢰구간은 연구의 반복과 확대에서 중추적 역할을 한다. 비슷한 연구들에서 점 추정(point estimate) 주위에 겹치는 신뢰구간을 반복 연구의 성공 기준으로 할 수 있다. 충분히 겹치는 신뢰구간은 동일한 모집단의 추정을 시사한다[37].

3. 베이지안 통계(Bayesian statistics)

베이지안 통계는 Bayes가 발견한 정리에 기초한 통계로서, Fisher와 Neyman-Pearson의 빈도주의자 통계(frequentist statistics)와 구별된다. 베이지안 통계는 빈도주의자 통계에 기초한 가설검정 방법인 귀무가설 유의성 검정의 문제점을 보완하거나 대체할 수 있는 대안으로 부상하고 있다.

1) 베이즈 정리(Bayes' theorem)

> 어떤 질병의 유병률이 1%이고 검사의 민감도가 80%인 경우에, 검사 결과가 양성으로 나왔을 때 실제로 그 질병에 걸려 있을 확률은 얼마인가?

이 문제의 답을 구하는 과정에서 베이즈 정리가 자연스럽게 유도된다. 우리가 알고자 하는 것은 검사 결과가 양성인 경우

중에서 실제로 질병에 걸려 있을 확률인 P(질병|양성)이며, 이것을 수식으로 표현하면 다음과 같다.

$$P(\text{질병}|\text{양성}) = \frac{P(\text{질병} \cap \text{양성})}{P(\text{양성})}$$

P(질병∩양성)은 질병에 걸려 있으면서 동시에 검사 결과도 양성일 확률, P(양성)은 검사 결과가 양성일 확률이다. 한편, 질병에 걸려 있는 경우 중에서 검사 결과가 양성으로 나올 확률, 즉 검사의 민감도(sensitivity)는 다음과 같다.

$$P(\text{질병}|\text{양성}) = \frac{P(\text{질병} \cap \text{양성})}{P(\text{질병})}$$

여기서 P(질병)은 질병에 걸려 있을 확률, 즉 유병률이다. 위 문제의 답이 80%라고 생각하는 것은 문제에서 주어진 민감도 P(양성|질병)와 우리가 알고자 하는 답인 P(질병|양성)을 혼동한 것이다. 위의 두 식을 연결하면 다음과 같다.

$$P(\text{질병}|\text{양성})P(\text{양성}) = P(\text{질병} \cap \text{양성})$$
$$P(\text{양성}|\text{질병})P(\text{질병}) = P(\text{질병} \cap \text{양성})$$
$$P(\text{질병}|\text{양성})P(\text{양성}) = P(\text{양성}|\text{질병})P(\text{질병})$$
$$P(\text{질병}|\text{양성}) = \frac{P(\text{양성}|\text{질병})P(\text{질병})}{P(\text{양성})}$$

위의 마지막 식이 바로 '베이즈 정리'에 해당한다.

P(양성)을 계산하려면 P(양성|질병)에 P(질병)을 곱한 값에, P(양성|건강)에 P(건강)을 곱한 값을 더하면 된다. P(양성|건강)은 건강한 경우 중에서 검사 결과가 양성으로 나올 확률인 가양성률(false positive)로서 '1 - 특이도'(specificity, 건강한 경우 중에서 검사 결과가 음성으로 나올 확률)이고, P(건강)은 건강할 확률로서 유병률의 반대이므로 '1 - P(질병)'이다. 특이도를 민감도와 같이 80%라고 가정하면,

$$P(\text{양성}) = P(\text{양성}|\text{질병})P(\text{질병}) + P(\text{양성}|\text{건강})P(\text{건강})$$
$$P(\text{질병}|\text{양성}) = \frac{P(\text{양성}|\text{질병})P(\text{질병})}{P(\text{양성}|\text{질병})P(\text{질병}) + P(\text{양성}|\text{건강})P(\text{건강})}$$
$$P(\text{질병}|\text{양성}) = \frac{0.8 \times 0.01}{0.8 \times 0.01 + 0.2 \times 0.99} \fallingdotseq 0.039$$

따라서, 위 문제에 대한 답은 '4%에 미치지 못한다'이다. 위 내용을 요약하면, 베이즈 정리는 조건부확률로부터 다음과 같이 유도된다.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$
$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$
$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

베이즈 정리를 가설검정과 관련된 형태로 나타내면 다음과 같다.

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

여기서 P(H|D)는 사후 확률(posterior probability), 즉 D가 관찰된 후 H의 확률, P(D|H)는 H의 조건에서 D를 관찰할 확률(가능도 또는 우도, likelihood), P(H)는 H의 사전 확률(prior probability), 즉 D가 관찰되기 전 H의 확률, P(D)는 D의 주변확률(marginal probability)이다.

2) 베이지안 통계의 장점

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. 지금까지 얻은 정보를 하나하나 다시 계산하는 것이 아니라 최신 정보만 개정하면 결과적으로 같은 수치를 얻을 수 있다는 편리성이 있다. 최근의 미국 대통령 선거에서 한 설문조사 사이트는 누적되는 설문조사 결과를 이어지는 설문조사 결과와 통합 분석하는 베이지안 통계 방법을 사용하여 정확한 예측에 성공할 수 있었다[38].

7년간의 임상시험 끝에 한 연구팀이 A약이 B약보다 0.05 유의수준에서 더 효과가 크다는 결론을 내렸다. 팀의 리더는 왜 그렇게 오래 걸렸느냐는 질문을 받고 "결과가 0.05 수준에 도달한 것이 이번이 처음이었습니다"라고 대답하였다. 미국식품의약국(FDA) 담당자는 연구의 중간 결과들을 볼 때 5% 오류율의 주장을 인정할 수 없다며 빈도주의자 통계의 관점에서 승인을 거부하였다. 만일 이 연구팀이 베이지안 통계를 사용하였다면 연구의 중간 결과가 축적적으로 통합되기 때문에 7년간의 임상시험이 필요하지 않았을 수도 있다. 사전 정보가 많은 경우에는 신약 개발을 위한 임상시험의 규모와 기간을 줄이는 것이 정당화될 수 있다[39,40].

베이지안 통계는 사전정보가 없을 때에도 적응적 연구를 설계하고 시행하는 데 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.

베이지안 통계는 귀무가설과 대립가설을 구분하지 않으며, 동시에 검정하고자 하는 가설의 수가 세 개 이상이어도 무관하

다. 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. 또한, 전문가들의 다양한 사전 신념들이 동일 자료 상황에서 다수의 사전 분포들에 의해 자연스럽게 반영될 수 있다[41].

빈도주의자 통계에서는 많은 측정을 한 후에 결론을 내릴 수 있기 때문에 대량생산체제에 적합한 방법이나, 베이지안 통계에서는 단 한 번의 시행이라도 그 결과를 살려서 추정치를 구할 수 있기 때문에 현대 사회의 다품종 소량생산체제 및 개성과 환경을 중요시하는 추세와 부합한다.

베이지안 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. 어린이가 지식을 습득하는 방식이 베이지안 통계의 방식과 같기 때문에, 인공 지능의 한 분야로 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 기계 학습(machine learning) 방법은 베이지안 통계의 원리에 따라 만들어졌다.

3) 베이지안 통계의 가설검정 방식이 일상생활에서 사용되는 예

> 같은 종류의 음식을 파는 A, B 두 음식점이 있고 A 음식점이 요리를 더 맛있게 한다는 사전정보가 있을 때, 어느 음식점이 요리를 더 맛있게 하는지 알아내는 방법은 무엇인가?

베이지안 통계의 가설검정 방식은 우리가 일상생활에서 상품과 서비스를 구입할 때 판단하는 과정과 비슷하다. 같은 종류의 음식을 파는 A, B 두 음식점이 있다고 하자. A 음식점이 요리를 더 맛있게 한다는 사전정보가 있는 경우에는 일단 그리로 가게 된다. A 음식점에서 식사를 해 본 결과, 맛에 만족한 경우에는 A 음식점이 요리를 더 맛있게 한다는 사전정보의 타당성이 사후정보에 의해 강화된다. B 음식점에서 식사를 해 본 결과, A 음식점보다 더 맛있는 경우에는 사전정보를 수정하여 B 음식점이 요리를 더 맛있게 한다는 사후정보를 획득하게 된다. 각 음식점을 방문하는 횟수가 늘어남에 따라 처음부터 축적된 정보가 전부 활용되어 점점 더 정확한 판단을 할 수 있게 된다. 실제로는 두 음식점을 몇 번씩만 방문해 보아도 어느 음식점이 요리를 더 맛있게 하는지 확실하게 알 수 있다.

이 경우에 빈도주의자 통계의 귀무가설검정을 한다고 하면 A, B 음식점에 무작위로 추출된 두 집단을 각각 보내어 요리의 맛을 보게 한 후 두 집단의 평가결과를 비교하게 된다. 통계적으로 유의한 차이를 얻기 위해서는 요리를 맛보는 사람의 수가 많아야 하고 그에 따른 비용이 증가하며, 1회 평가로는 항상 유의수준에 따른 오류의 가능성을 내포하고 있고, 두 음식점 요리에 맛의 차이가 있는 경우에는 둘 중 한 음식점에서 식사를 하는 사람들은 맛없는 요리를 먹게 된다. 요리가 아닌 의료기술이나 의약품인 경우에는 윤리적인 문제가 부각되며, 한 번 유의한 결과가 나오면 반복 연구가 어려워진다.

4) 베이지안 통계의 단점

베이지안 통계는 사전 확률을 사용하는 것이 특징이고 사전 확률은 본질적으로 불확실한 것일 수 밖에 없으나, 사전 확률에 연구자의 주관성(subjectivity)이 개입될 수 있다는 점이 가장 큰 비판의 대상이 되어 왔다. 과학은 객관성(objectivity)을 생명으로 하며, 과학자들은 주관성을 극도로 싫어한다. 빈도주의자 통계에서는 객관성을 매우 중요시한 결과로, p값을 제외하고 가설과 연관된 다른 근거들을 포기하는 대가를 치루었다. 근거중심 치의학(evidence-based dentistry)에서도 전문가 견해(expert opinion)의 근거 강도를 가장 낮게 평가한다.

이러한 비판에 대응하여 베이지안 통계에서는 사전정보가 없거나 불확실할 경우에는 무정보 사전확률(uninformative prior)을 사용한다. 위 예에서 만일 A 음식점이 요리를 더 맛있게 한다는 사전정보가 없는 경우에는 A, B 두 음식점에 똑같이 50%의 확률을 부여하고 시작한다.

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[42]. 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[43,44].

5) 베이지안 통계를 위한 소프트웨어

Markov Chain Monte Carlo (MCMC) 같은 특별 컴퓨터 알고리즘이 자료의 분석, 모델의 가정 점검, 설계 단계에서 사전 확률의 사정, 다양한 결과의 확률 사정을 위한 시뮬레이션 수행 등을 위해 사용된다. 베이지안 분석을 위한 소프트웨어 프로그램 WinBUGS는 http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml에서 다운 받을 수 있다. 또한, R package도 베이지안 통계를 위한 공개 소프트웨어이다. 베이지안 분석을 위한 국제협회(International Society for Bayesian Analysis)의 홈페이지 http://www.bayesian.org에서 추가 정보를 얻을 수 있다.

6) 치의학 연구에서 베이지안 통계의 사용

치의학 연구에서 베이지안 통계의 사용이 증가하고 있다. 베이지안 모델을 사용한 우식의 구강내 공간적 대칭성에 관한 연구[45], 치주학에서 베이지안 네트워크 메타분석 연구[46], 디지털 방사선촬영에서 노이즈를 제거하기 위한 베이지안 알고리즘에

대한 연구[47] 등이 최근에 보고된 연구들이다. 현재의 추세로 볼 때, 앞으로 임상 및 역학 연구에서 베이지안 통계의 사용이 불가피할 것으로 전망되며 이에 대한 대비가 필요하다고 사료된다[48].

Ⅳ. 요 약

치의학 연구에서 사용되는 귀무가설 유의성 검정에서 *p*값을 기준으로 연구의 결과를 평가하는 것은 많은 문제점을 내포하고 있다. 귀무가설이 기각되지 않은 경우에 귀무가설이 옳다는 결론을 내리는 것은 논리적 오류이다. *p*값에 대한 중대한 오해가 많이 있으며 연구자는 논문을 작성할 때 *p*값의 해석에 신중해야 한다. 귀무가설검정을 보완하거나 대체할 수 있는 대안으로서, 효과 크기, 신뢰구간, 베이지안 통계 등이 있다.

References

1. Seaman JE, Allen IE : Not significant, but Important? Know the pitfalls of p-values and formal hypothesis tests. Quality Progress, 2011 August. Available from URL : http://asq.org/quality-progress/2011/08/statistics-roundtable/not-significant-but-important.html (Accessed on July 8, 2013)
2. Matrixx Initiatives, Inc. v. Siracusano. Available from URL: http://en.wikipedia.org/wiki/Matrixx_Initiatives,_Inc._v._Siracusano (Accessed on July 8, 2013)
3. Meehl PE : Theory-testing in psychology and physics: a methodological paradox. *Philosophy Sci*, 34:103-115, 1967.
4. Meehl PE : Theoretical risks and tabular asterisks: sir Karl, sir Ronald, and the slow progress of soft psychology. *J Consult Clin Psychol*, 46:806-834, 1978.
5. Cohen J : The earth is round (*p*<.05). *Am Psychol*, 49:997-1003, 1994.
6. Schmidt FL, Hunter JE : Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In Harlow LA, Mulaik SA, Steiger JH (eds.) : What if there were no significance tests? *Mahwah, NJ, Lawrence Erlbaum Associates*, 37-64, 1997.
7. NHST problems. Available from URL: http://www.faculty.biol.ttu.edu/strauss/stats/LectureNotes/20_NHSTProblems.pdf (Accessed on July 8, 2013)
8. Fallacy of affirming the consequent. Available from URL: http://terms.naver.com/entry.nhn?cid=1137&docId=275047&mobile&categoryId=1137 (Accessed on July 8, 2013)
9. Pollard P, Richardson JTE : On the probability of making type I errors. *Psychol Bull*, 102:159-163, 1987.
10. Reese HW : Problems of statistical inference. *Mex J Behav Anal*, 25:39-68, 1999.
11. Goodman S : A dirty dozen: twelve p-value misconceptions. *Semin Hematol*, 45:135-140, 2008.
12. Hubbard R, Lindsay RM : Why p values are not a useful measure of evidence in statistical significance sesting. *Theory Psychol*, 18:69-88, 2008.
13. Sterne JAC, Smith GD : Sifting the evidence - what's wrong with significance tests? *BMJ(Clin res)*, 322:226-231, 2001.
14. Johnson, DH : The insignificance of statistical significance testing. *J Wildlife Manag*, 63:763-772, 1999.
15. Nurminen M : Statistical significance - a misconstrued notion in medical research. *Scand J Work Environ Health*, 23:232-235, 1997.
16. Schervish MJ : P values: what they are and what they are not. *Am Stat*, 50:203-206, 1996.
17. Carver RP : The case against statistical significance testing. *Harvard Educat Review*, 48:378-399, 1978.
18. Nickerson RS : Null hypothesis statistical testing: a review of an old and continuing controversy. *Psychol Methods*, 5:241-301, 2000.
19. Berger JO, Sellke T : Testing a point null hypothesis: the irreconcilability of p values and evidence (with comments). *J Am Stat Assoc*, 82:112-139, 1987.
20. Berger JO, Delampady M : Testing precise hypotheses (with comments). *Stat Science*, 2:317-352, 1987.
21. Nester MR : An applied statistician's creed. *Statistician*, 45:401-410, 1996.
22. Berger JO, Berry DA : Statistical analysis and the illusion of objectivity. *Am Scientist*, 76:159-165, 1988.
23. Hubbard, R : Alphabet soup: blurring the distinctions between p's and α's in psychological research. Theory Psychol, 14:295-327, 2004. 24. Sellke T, Bayarri MJ, Berger JO : Calibration of p values for testing precise null hypotheses. *Am Statistician*, 55:62-71, 2001.
25. Gelman A, Stern H : The difference between 'significant' and 'not significant' is not itself statistically significant. *Am Statistician*, 60:328-331, 2006.
26. International committee of medical journal editors : Uniform requirements for manuscripts submitted to

biomedical journals. Available from URL: http://www.icmje.org/manuscript_1prepare.html (Assessed on June 27, 2013)
27. Royall RM : The effect of sample size on the meaning of significance tests. *Am Statistician*, 40:313-315, 1986.
28. Hand DJ : Data mining: statistics and more? *Am Statistician*, 52:112.118, 1998.
29. Fisher RA : The design of experiments (8th ed.). Edinburgh, Oliver & Boyd, 1966.
30. Fisher BJ : R.A. Fisher: The life of a scientist. New York, Wiley, 1978.
31. Denis DJ : Alternatives to null hypothesis significance testing. Theory & Science, 4(1), 2003. Available from URL: http://theoryandscience.icaap.org/content/vol4.1/02_denis.html (Accessed on July 8, 2013)
32. Rosenthal R : Effect size estimation, significance testing, and the file-drawer problem. *J Parapsychol*, 56:57-58, 1992.
33. Vaughan GM, Corballis MC : Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychol Bulletin*, 72:204-213, 1969.
34. Silva-Aycaguer LC, Suarez-Gil P, Fernandez-Somoano A : Null hypothesis significance test in health sciences research (1995-2006): statistical analysis and interpretation. *BMC Med Res Methodol*, 10:44, 2010.
35. Schmidt FL : Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychol Methods*, 1:115-129, 1996.
36. Cumming G, Finch S : Inference by eye: confidence intervals and how to read pictures of data. *Am Psychol*, 60:170-180, 2005.
37. Schenker N, Gentleman JF : On judging the significance of differences by examining the overlap between confidence intervals. *Am Statistician*, 55:182-186, 2001.
38. Wang S, Campbell B : Mr. Bayes goes to Washington. *Science*, 339:758-759, 2013.
39. Efron B : Bayes' Theorem in the twenty-first century. *Science*, 340:1177-1178, 2013.
40. FDA : Guidance for the use of Bayesian statistics in medical device clinical trials. Available from URL : http://www.fda.gov/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm071072.htm (Accessed on July 8, 2013)
41. Lilford RJ, Braunholtz D : The statistical basis of public policy: a paradigm shift is overdue. *Br Med J*, 313:603-607, 1996.
42. Efron B : Why isn't everyone a Bayesian (with discussion)? *Am Statist*, 40:1-11, 1986.
43. Nurminen M, Mutanen P : Exact Bayesian analysis of two proportions. *Scand J Stat*, 14:67-77, 1987.
44. Diaconis P, Freedman D : On the consistency of Bayes estimate (with discussion). *Ann Math Stat*, 14:1-67, 1986.
45. Zhang Y, Todem D, Kim K, Lesaffre E : Bayesian latent variable models for spatially correlated tooth-level binary data in caries research. *Stat Modelling*, 11:25-47, 2011.
46. Tu YK, Needleman I, Chambrone L, *et al.* : A Bayesian network meta-analysis on comparisons of enamel matrix derivatives, guided tissue regeneration and their combination therapies. *J Clin Periodontol*, 39:303-314, 2012.
47. Frosio I, Olivieri C, Lucchese M, *et al.* : Bayesian denoising in digital radiography: a comparison in the dental field. *Comput Med Imaging Graph*, 37:28-39, 2013.
48. Freedman L : Bayesian statistical methods. A natural way to assess clinical evidence (editorial). *Br Med J*, 313:569-570, 1996.

국문초록

치의학 연구에서 귀무가설 유의성 검정의 문제점과 대안에 관한 고찰

이광희

원광대학교 치과대학 소아치과학교실

치의학 연구에서 사용되는 귀무가설 유의성 검정에서 p값을 기준으로 연구의 결과를 평가하는 것은 많은 문제점을 내포하고 있다. 귀무가설이 기각되지 않은 경우에 귀무가설이 옳다는 결론을 내리는 것은 논리적 오류이다. p값에 대한 중대한 오해가 많이 있으며 연구자는 논문을 작성할 때 p값의 해석에 신중해야 한다. 귀무가설검정을 보완하거나 대체할 수 있는 대안으로서, 효과 크기, 신뢰구간, 베이지안 통계 등이 있다.

주요어: 귀무가설, 유의성 검정, p값, 효과 크기, 신뢰구간, 베이지안 통계