

유의도 검증의 결과해석에 대한 오류분석*

조 남 국

육군사관학교 심리학처

본 논문은 Neyman과 Pearson의 검증이론에 기초를 둔 유의도 검증(significance test)의 결과 해석에서 흔히 발생하는 오류들을 열거하고 분석하였다. 그러한 오류들이 발생하게 되는 근본적인 이유는 α 수준에 대한 그릇된 이해와 가설검증의 판단과정에서 α 수준에 의한 일련의 결정과 그 의미에 대한 맹신때문이라고 할 수 있다. 이러한 오류들을 올바르게 이해하기 위해서는 가설 검증 절차에 대한 근본적인 분석을 필요로 한다. 그러므로 본 논문은, 형식 논리학을 빌려서, 경험적인 연구에서 흔히 사용되는 가설들과 유의도 검증의 논리적 구조 분석을 시도하였다. 필자는 그 분석결과로서, 올바른 α 해석에 대한 제안과 더불어, α 에만 치중하는 연구 관행의 미흡성을 지적하고 효과크기 및 β 에 대한 관심을 촉구하였다.

오늘날 심리학이 경험적인 과학이라는 주장에 대해서 이의를 제기할 사람은 아무도 없을 것이다. 그러나 그런 것을 너무 강조해 온 나머지, 심리학 분야의 일각에서는 일부 잘못된 인식들이 후진들에게 전수되어온 경향이 있었다. 그 대표적인 예 중의 하나로는, 꼭 실험을 해야지만 논문으로 인정된 시기가 있었다는 것을 손꼽을 수 있다. 이는 마치 실험방법이 경험적인 과학을 대표하기라도 하는 것 같은 인상을 주는데 그 내용은 물론 참(眞)이 아니다. 어떤 특정한 주제 혹은 주제 영역이 있을 때, 그것은 이론적 혹은 경험적으로 다루어질 수 있으며, 그 가운데에서도 경험적인 방법은 다양한 형태들을 띠고 있다. 물론 그러한 여러가지 형태의 방법들에는 각각 그 장·단점에 따른 차이와 특성이 있게 마련이다. 그것은 실험방법이라고 해서 예외일 수 없으며 그 역시 나름대로의 한계가 있다. 실험이 다른 방법

에 비하여 변수에 대한 조작과 통제가 보다 체계적일 수 있다고 해서, 그렇다고 어떤 종류의 문제를 다룰 때나 막론하고 항상 최고의 방법이 될 수는 없는 것이다. 그러므로 경험적인 연구에서 중요한 것은, 하나의 특정한 연구방법에만 연연하기 보다는, 상정된 문제의 해결을 위하여 우리가 적용하려는 연구 방법이나 논리가 실제로 가용한 수단중에서 최선의 것인가를 먼저 생각해 보는 것이다.

이러한 문제와 관련하여, 유사한 범주에 속하면서도 그 심각성을 더욱 돋보이게 하는 또 하나의 경향을 들 수 있다. 그것은 방금 언급한 것처럼 실험만이 주로 강조되어 온 것 외에, 그 실험결과 역시 통계적으로 최소한 유의미(significant)한 결과로 나타날 때만 학술지에 발표될 가치가 있다고 인정한다는 관념이다. 따라서 이러한 논리를 일관성있게 적용해 나가면 극단적으로는, 연구결과의 통계적 유의미성 여부가 학위 논문으로서의 가치여부를 결정하는 척도로서까지 등장할 수도 있게 된

* 본 논문을 읽고 조연해 주신 이만영 선생님께 감사드립니다.

다는 것이다.

이러한 오류는 이미 1960년대 초부터, "Journal of Experimental Psychology"의 게재 논문에 대한 채택기준을 제시한 Melton(1962)의 정신에서 여실히 반영되고 있었던 사항이다(Bredenkamp, 1972). 그렇지만 대체로 1960년 이후, Melton을 전후하여, 구미지역에서는 유의도 검증(significance test : 이하 ST)에 대한 활발한 논의가 전개되었다(Nunnally, 1960; Rozeboom, 1960; Bakan, 1966; Lykken, 1968; Bredenkamp, 1972). 이에 따라 ST에 관한 이해가 증진된 것은 사실이지만, 그러나 ST와 관련되는 연구의 실제(practice)에서는 아직도 잘못된 관행을 떨쳐 버리지 못하는 경향이 많이 있다고 가끔 지적되기도 한다(Haagen & Seifert, 1979). 그러므로 사실상 ST에 대한 논의는 이 시점에서 특별히 더 새로울 것은 없겠으나, 그에 대한 잘못된 결과 해석이 가져다 줄 수 있는 역기능의 중대성을 고려하여 본 논문에서 다시 한번 생각을 정리하여 보고자 하는 것이다.

문제의 제1

일반적으로 경험적인 연구에서 사용되고 있는 가설검정은 그 나름대로의 검증논리를 갖추고 있으며, 또 그러한 검증의 대상이 되고 있는 가설은, 그 또한 논리적인 구조에 따른 특성을 내포하고 있다. 이러한 논리들이 외면 당한 채 무작정하게 시행되는 "유의미한(significant) 결과에 대한 추구와 해석"은 무의미할 따름이다. 나중에도 언급되겠지만, 2종 오류의 확률 β 에 대한 실질적인 고려가 없는 상태에서 1종 오류의 확률 α 에만 국한된 연구결과의 해석은 무리를 낳을 수 밖에 없다. 그러므로 예를 들면 ω^2 과 같은 보다 실질적인 지표, 즉 효과크기(effect size)나 검증력(power : $1 - \beta$) 등을 고려하지 않은 연구들이, 단지 그 결과가 통계적으로 유의미했다고 해서, 소위 "좋은" 연구로서의 가치가 인정될 수는 없다는 것이다(Cohen, 1969; Bredenkamp, 1972; Hays, 1974; Haagen & Seifert, 1979; Witte, 1980; Schulz, Muthig & Koeppler, 1981; Stelzl, 1982; Hager & Westermann, 1983; Bortz, 1984). 그 뿐만아니라, 하나의 가설과 관련하여 유의미한 결론이 도출된 논문들

의 양(量)이 아무리 많이 축적되었다고 하더라도, 단지 그 사실 자체만으로 "그 가설이 참(眞)"이라고 단정할 수 있는 확실한 明證(evidence)을 가져다 주는 것은 아니다(Hedges & Olkin, 1985). 또 이와 같은 맥락에서 볼 때, 반복연구(replication)에서 설사 significant(s.)와 non-significant(n.s.)한 2개의 상반된(?) 연구결과들이 나타났다고 할지라도, 그 결과들은 연구자가 검증하고 싶은 가설의 궁극적인 진위여부 판단에 아무런 도움을 줄 수도 없는 것이다. 그 이유 중의 하나는, 결과 해석이 단지 α 에만 국한되어 고려되고 있는 상황이므로, 그러한 결과들만 가지고는 그들 중 어느 하나를 분명히 선호할 수 있을 정도로, 각 연구들간의 근본적인 질적 차이를 나타낼 수 있는 기준이 제공되고 있지 않기 때문이다.

표 1은 ST의 α 에 대한 그릇된 이해를 근거로 하여 연구결과들이 잘못 종합될 수 있는 psycho-logic의 가상적인 일 예를 제시해 본 것이다.¹⁾ 이 표에는 연구의 유의미한 결과와 무의미한 결과의 단순한 2가지 판단만 구분되어 예시되고 있지만, 만약 추가적으로 매우 유의미한(very significant) 연구결과들까지 포함되어 표가 확대되었다고 하더라도, 그 논리는 시종 같을 것이다. 그것이 어떻든간에, 표 1에서 S**과 같은 psycho-logic에 의한 판단이 내포하고 있는 문제의 의미는, 성공적으로 반복된 유의미한 결과가 마치 주장되고 있는 가설이 참(眞)이라는 "明證" 또는 "確信"을 주기라도 한 것처럼 논리가 비약되거나 혼동되고 있다는 것이다. 이러한 오류의 성격은 가설의 진위여부와 확신의 중대여부라는 2가지 측면의 내용을 함께 내포하고 있으며, 이들 모두가 α 와 직접 혹은 간접적으로 관련되는 사항들이다. 그런데 가설의 진위판단과 그 판단에 대한 신뢰에 관련된 이러한 문제들은 각각 다음과 같이

1) 필자가 생각하는 psycho-logic이란 $logic \subset psycho-logic$ 의 관계로서 일반적인 형식 논리(logic)가 논리의 모순을 허용하지 않는 반면에, 심리적 논리(psycho-logic)는 logic과 같은 올바른 논리의 전개는 물론 모순된 논리까지도 인지차원에서 실재할 수 있다는 의미에서 logic보다는 포괄적인 개념으로 사용되었다(\subset 는 집합 이론에서 내포의 뜻임).

표 1. 반복연구의 결과들이 잘못 종합될 수 있는 psycho-logic의 일 예²⁾

가능성	연구 1	연구 2	연구 1 \wedge 연구 2 ³⁾ (심리적 판단의 예)	선호순서(예)	비 고
1	s.	s.	S**	1	반복연구 성공
2	n.s.	n.s.	NS**	4	반복연구 성공
3	s.	n.s.	S	3	반복연구 실패
4	n.s.	s.	S*	2	반복연구 실패

(1) 가설검증의 논리적 구조

(2) 유의도 검증의 결정요인(α , 검증력, 효과크기 및 N)

의 관점에서 집중적으로 다루어질 수 있다. 그리고 이에 대한 세부적인 고찰은 ST가 경험적인 연구에서 차지하는 비중과 그 검증방법의 특성에 대한 전반적인 이해를 높이는 데 기여할 수 있을 것이다.

본 논문은 제기될 수 있는 이러한 문제들 중에서 전자의 문제에 국한하여 논의하고자 한다. 다만 논의를 시작하기 전에 한가지 언급해 두어야 할 사항이 있다. 원칙적으로 가설 검증의 논리적 구조를 다루기에 앞서 선행되어야 하는 결정사항은 가설 검증방법의 선택이다. 뒤에서 다루어지겠지만, 가설검증은 결국 확실적인 판단일 수 밖에 없는데, 확률에 대한 다양한 이론들이 있듯이 대응되는 가설검증의 이론들도 다양하게 존재하고 있다(Hager & Westermann, 1983; Witte, 1980). 그러나 필자는 앞에서 반복연구 결과의 예를 설명하면서 이미 은연중에 하나의 특정한 가설검증 방법을 선택하고 있다. 그것은 Neyman과 Pearson의 가설검증 이론인데, 그 이론을 채택하게 된 이유는 그것이 일반적으로 사회과학분야의 경험적 연구에서 주로 실시되고 있는

보편적인 방법임에도 불구하고, 아직도 그 방법의 적용 과정에서 많은 오류가 발생되고 있기 때문이다. 본 논문은 그 중에서도 특히 결과 해석에서 자주 범하는 오류의 경향과 문제점들을 지적하고 ST에 대한 올바른 인식을 심어 보려고 한다. 그에 앞서 우선 ST의 결과가 일반적으로 어떻게 잘못 사용되며 해석되고 있는가에 대한 실태들을 알아보기로 하자.

유의도 검증의 잘못된 결과 해석

유의도 검증 방법의 적용에서 흔히 나타나고 있는 오류들의 상당 부분은 그 결과들에 대한 잘못된 취급과 해석이 차지하고 있다. 그 근본적인 이유는 대체적으로 유의미성(significance)과 유효적절성(relevance)의 개념 혼동, “유의미성(significant)=효과적(effective)”이라는 잘못 이해된 등식 및 ST에 대한 일반적인 오해에서 연유하는 α 에 대한 맹신과 과대평가 때문이라고 할 수 있다. 그렇게해서 발생하는 오류의 구체적인 내용들을 정리하여 열거하면 다음과 같이

(오류 1): α 수준에서 유의미하다는 것은, 대립가설(H_1)이 $1-\alpha$ %의 확률을 가지고 참(眞)이라는 것을 증명하는 것이다. 따라서 α 가 작아질수록, H_1 이 참이라는 확률은 더욱 커진다.

(오류 2): α 가 작을수록, 동일한 실험의 반복연구에서, 관찰되었던 효과가 반복되어 입증될 수 있는 가능성이 커진다.

(오류 3): 연구 결과가 유의미하다는 것은, 그 결과를 해석할 수 있다는 것을 의미하고, 더 나아가서는 모집단에 일반화시킬 수 있다는 것을 의미한다.

(오류 4): 연구 결과가 유의미하다는 것의 의미는,

2) 여기에서 소문자 s는 ST의 결과, 대문자 S는 유의미하다고 느끼는 주관적인 심리 상태, 그리고 *는 확신의 강도를 의미함.

3) 본 논문에서 사용되고 있는 논리기호는 다음과 같이

\bigwedge : 普遍量詞(universal quantifier) \wedge : 連言(conjunction)
 \bigvee : 存在量詞(existential quantifier) \vee : 選言(disjunction)
 \neg : 否定(negation)
 \rightarrow : 含蓄(implication)
 \equiv : 同値(equivalent)

의 의미를 띄고 있다(v.Kutschera & Breilkopf, 1971; Stegmüller, 1969 참고).

그 연구에서 독립변수가 종속변수에 미치는 커다란 효과(effect)를 발견했다는 것이다.

(오류 5) : 결과가 유의미하지 않은 연구는 연구로서의 가치가 없다. 따라서 발표될 가치조차도 상실된다.

(오류 6) : α % 수준에서 검증을 시행한다는 것은, 영가설(H_0)이 α %만큼 잘못 기각되고, 따라서 그 만큼의 잘못된 결과를 초래하는 것이므로, 결국 유의도 검증 방법은 신뢰할 수 없다.

등으로 집약될 수 있다(Bakan, 1966; Witte, 1980; Patzelt, 1985 참고).

이러한 오류들에 대한 비판은 별로 어려운 것이 아니다. 이들 오류의 전부는 아니더라도 대부분의 경우, 그 맹점이 어디에 있는가를 가려내어 비판하기 위해서는, 우리들이 흔히 알고 있는 기본적인 통계지식의 응용만으로도 충분하다.

우선 오류 1에서 주장하고 있는

$1 - \alpha \geq H_1$ 이 참일 수 있는 확률 $\wedge H_1$ 채택

의 등식은, 어떤 통계책에서나 흔히 접할 수 있는 표 2를 참고해 볼 때, 잘못되었다는 것이 너무나 자명해진다. 즉 $1 - \alpha$ 란 H_0 가 참(眞)일 때 H_0 를 채택하는 확률일 뿐, 오류 1이 주장하는 내용처럼 H_1 이 참(眞)일 때 H_1 을 채택하는 확률은 아니다. 그것은 실제로는 $1 - \beta$, 즉 검증력(power)이기 때문이다.

다음으로, α 가 작을수록 반복연구의 성공확률이 커

표 2. 통계적 가설의 판단과 관련된 확률

	H_0 眞	H_1 眞
H_0 채택	$1 - \alpha$	β
H_1 채택	α	$1 - \beta$

진다는 주장을 담은 오류 2는, 간단히 그림 1을 참고하면 그 문제성을 쉽게 파악할 수 있을 것이다. 동일한 실험의 반복연구이므로 H_0 과 H_1 의 모집단 분포가 일정하다는 조건을 전제로 해야 하는데, 이때 α 가 작아질수록 β 는 커지게 된다. 그러나 β 란 H_1 이 참(眞)일 때 H_0 을 채택하는 확률이다(표 2 참고). 따라서 α 가 작아질수록 반복연구에서 성공할 수 있는 확률은 그 만큼 작아지게 되는 것이다.

다음, 오류 3은, 결과가 유의미할 때만 “해석”과 “일반화”가 가능해진다는 주장인데, 이는 대체적으로 연구가설이 대립가설(H_1)에 상응한다는 통상적인 관념으로부터 발생하는 편견인 것 같다. 그러나 이러한 편견은 근본적으로 ST의 이론내에서 차지하는 H_0 과 H_1 의 논리적 관계, 그리고 모집단과 표집간의 관계 등에 대한 이해 부족에서부터 연유하는 것이다. 그 뿐만 아니라 “결과를 일반화시킬 수 있다”는 내용은 또 다른 각도에서도 그 문제의 심각성이 나타나고 있다. 즉 하나의 연구결과를 β 에 대한 일체의 고려 없이 α 만으로 판단하고 해석하려는 경향인데, 그때 어떤 문제가 발생될 수 있

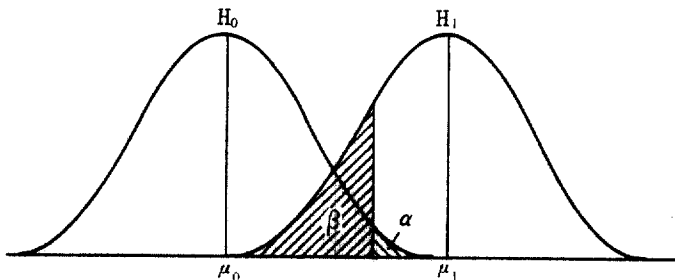


그림 1. α 와 β 의 관계

는지 그림 1을 참고하면서 하나의 예를 생각해 보기로 하자. 우리가 만약 각각 $n=45$ 명인 통제 및 실험 집단을 구성하여 연구를 실시하였다고 가정하였을 때, 각 집단의 측정기대치 간의 차이가 $|\mu_0 - \mu_1| = 0.5$ 이고, 추정된 표준편차치는 2집단에 동일하게 $s=1$ 이었으며, 유의미 수준은 $\alpha=0.01$ (일방적)로서 매우 유의미한 결과를 얻었다고 하자. 이때 오류 3의 주장대로라면 1%수준에서 유의미하기 때문에 소위 “일반화”를 할 수 있어야 하겠지만, 이는 전적으로 잘못된 것이다. 그 이유는 바로 위에서 든 예의 경우에 $\beta=0.50$ 이기 때문이다. 이러한 예는 그림 1에서 $\alpha=0.01$ 과 $\beta=0.50$ 가 되는 특정한 경우인데, 이때 β 가 50%라는 의미는 $1-\beta$, 즉 검증력이 50%라는 것이다. 그렇다면 그러한 연구는 힘들어서 할 필요도 없이, 연구가설을 반영하는 H_1 의 채택여부를 결정하기 위하여 동전을 던지는 것과 똑같은 정도로 전혀 무의미한 것이 되고 만다. 따라서 어떤 연구의 결과가 “매우 유의미”했다고 해서, 무조건 그 대립가설(H_1)의 채택에 대한 보다 더 뚜렷한 확신이 설 수 있는 것은 아니며, 또 그렇게 애매한 결과에 의존하여 마치 연구가설이 참(眞)인 것처럼 “결과를 일반화” 시켜서는 더욱 안될 것이다(Stelzl, 1982 참고).

다음에 오류 4가 시사하고 있는 내용, 즉 효과(effect)와 “유의미하다”는 결론의 동일시가 잘못된 인식이라는 것은, Bredenkamp(1972)가 자세한 예를 들었던 것처럼, 하나의 검증치 공식을 관찰해 보면 쉽게 알 수 있는 것이다. 동일한 피험자수를 갖춘 2개의 독립집단에서 평균치를 비교하는 t -검증 공식,⁴⁾

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s} \sqrt{\frac{n}{2}}$$

$$\text{이때, } s = \sqrt{(s_1^2 + s_2^2)/2}$$

n : 각집단의 피험자 수

를 보면, 효과(effect)를 반영하는 $\bar{X}_1 - \bar{X}_2$ 가 아무리 작아도 피험자수 n 이 충분히 커지면 검증치 t 는 유의미할 수 있는 반면에, $\bar{X}_1 - \bar{X}_2$ 가 아무리 커도 n 이 너무 작으면 검

증치 t 는 유의미하지 않을 수도 있다. 극단적으로 표현한다면, 어떤 연구자에게 유의미한 결과의 획득만이 목적이라고 할 때, n 의 크기를 늘리기만 하면 얼마든지 원하는 결과를 추구할 수 있다는 것이다. 따라서 오류 4의 명제는 문제점이 있으며 그 주장의 논리적인裏面, 즉 H_1 을 기각하고 H_0 을 채택한다는 것의 의미는 효과(effect)를 발견하지 못한 것이라는 주장도 역시 오류가 된다.

오류 5는 앞의 오류들에 대한 분석으로 이미 충분한 설명이 되었으리라고 믿는다. 그러나 오류 5에서 특히 혼동을 일으키고 있는 문제는 “유의미성(significant)=유효적절성(relevant)”이라는 잘못된 인식이다. 이는 하나의 연구결과가 연구자들로부터 주목될만한 가치를 획득한다는 것이 꼭 유의미한 결과에만 국한되는 것은 아니라는 사실을 무시하고 있는 것이다. 일반적으로 높은 비중을 둘 수 있는 유효적절한 연구결과란, 그 결과로부터 이론적이거나 실제적인 측면에서 실용적인 결론을 도출해 낼 수 있을 때를 말한다. 그러므로 유의미한 결론이라고 해서 꼭 유효적절(relevant)할 수는 없으며, 또 동일한 논리의 맥락에서 볼 때 유의미하지 않은 결론이었다고 해서 무조건 그 연구가 유효적절성의 관점에서 가치를 상실한다고 볼 수는 없는 것이다(Patzelt, 1985).

마지막으로 오류 6은, 앞의 오류들이 보인 α 에 대한 맹신 내지 과대평가와는 반대의 입장에서, ST 전반에 걸친 편견에 의한 과소평가를 나타내는 것이다. 그러나 이러한 편견 역시, 앞서 언급된 오류들과 더불어, ST에 대한 근본적인 이해가 부족해서 나타나는 현상이므로, 본 논문이 다음에 기술하고자 하는 검증의 논리에 대한 이해를 근거하여 볼 때 잘못된 견해이다.

이상으로 경험적인 연구에서 흔히 범하기 쉬운 ST 결과해석의 오류들을 살펴보았다. 그러나 위에서 제시된 각종 오류들의 이면에는 공통적으로 흐르고 있는 하나의 맥락이 있다. 그것은 α , β 등의 문제를 떠나서, ST가 마치 진리를 계산해 낼 수 있는 강력한 도구처럼 인식되고 있다는 것이다. 이는 경험적 과학에서의 ST의 좌표와 역할에 대한 이해부족에서 오는 것이므로, 근본적인 이해 증진을 위해서는 ST에 대한 구조적인

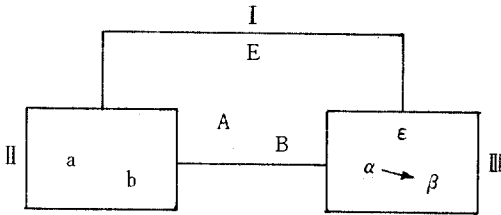
4) 이 공식에서의 s 는 표준편차로서 표 1에서 제시된 s 혹은 S 와는 무관함.

분석을 필요로 한다. 그런데 ST의 도입 목적이 궁극적으로 이론이나 가설의 검증에 있기 때문에, 바로 이러한 원천적인 내용부터, 일부 과학 철학의 개념들을 빌려, 간단히 고찰해 보기로 한다.

반증의 이론

일반적으로 모든 경험적인 연구는 이론에서부터 출발하여 가설을 도출하고, 그 가설을 검증하는 절차를 밟는다. 이때 이론이란, 하나의 특정한 대상 영역이나 현실의 일부분 속에서 현상들간의 관계를 재조립해 보려는 시도라고 할 수 있다(Friedrichs, 1973). 그리고 가설은 $A \rightarrow B$ 의 논리적 구조, 즉 “A라는 현상이 있으면, 꼭 B라는 현상이 뒤따른다(if A, then B)”라는 형태를 띠고 있다.

이러한 이론과 가설이 경험적인 연구에서 차지하는 좌표를 Popper(1963)는 그림 2와 같이 제시하고 있다. 우선 경험적 연구에서 상정되는 영역들, 즉 실재(Ding an Sich)영역, 관찰영역 및 명제체계 영역을 구분하고, 전체영역의 테두리내에서 각 영역들 간에 상응되는 변수들을 조명하는 것이다. 이때 이론(ϵ)과 가설($\alpha \rightarrow \beta$)



- I : 기본적인 실체들의 전체집합(실재영역)
- II : 관찰 가능한 현상들의 전체집합(관찰영역)
- III : 서술언어 및 상징적 표현의 전체집합(명제체계영역)
- a, b : 각각 현상
- A, B : 각 현상뒤에 상응하여 존재하는 실체
- α, β : a, b의 서술 혹은 표현
- E : A, B간의 실제적 관계
- ϵ : 서술된 이론

그림 2. 경험적 연구의 대상영역에서 이론과 가설의 좌표 (Popper, 1963; Friedrichs, 1973)

은 실재영역으로부터 명제체계로의 일종의 寫像(mapping)이라고도 볼 수 있다. 일반적으로 그림 2가 시사하는 바는, 경험적인 세계에서 실재영역은 직접적인 접근이 불가능하고, 현상에 대한 관찰영역을 통해서만 간접적인 접근이 가능해진다는 것이다. 또 이러한 간접경험에 의하여 형성되는 실재영역의 실체들에 대한 이론(ϵ)은 실제로 관찰될 수 있는 것이 아니며, 다만 α 에서 β 를 도출하는데 도움이 될 수 있는 내용을 담은 명제체계를 뜻하는 것이다. 결국 우리가 경험적인 학문을 한다는 것은, 명제체계를 빌려서 관찰을 통해 실재영역에 대한 인식을 확대시키자는 것이다. 그런데 이렇게 현상을 설명하는 명제체계는 다양할 수 있다. 따라서 우리에게 중요시 되는 문제는, 상호 경쟁관계에 있는 여러 명제체계 혹은 이론들 중에서 어떠한 명제체계 내지 이론이 현상을 서술하고 설명하는데 보다 적합한가의 여부를 결정해야 하는 것이다. 환언하면, ϵ 에서 도출되는 α, β 간의 관계, 즉 $\alpha \rightarrow \beta$ 에 대한 명제의 진위여부를 결정하는 검증기준이 요구된다는 것이다. 그러므로 이러한 이론(ϵ)과 가설검증의 사이에는 경험을 통한 올바른 인식의 획득을 돕기 위하여 적절한 방법론적인 규칙의 도입을 필요로 한다. 바로 이 부분이 소박한 경험론(naive empiricism)에서부터 논리실증주의(logical empiricism)를 거쳐 Popper의 비판적 합리주의(critical rationalism)에 이르기까지 근대의 과학 철학이 학문의 발전에 기여하기 위하여 노력을 집중시켜온 분야이며, 논리적인 분석과 실용적인 제안들을 통하여 인식의 확대를 추구해온 영역이다.

Popper를 전환점으로, 그 이전까지 가설검증의 판단 기준으로 제시되었던 입증(verification)은 반증(falsification)으로 대체되었다. 그 이유는 입증에 의한 가설검증이 귀납법(induction)의 문제로 한계에 부딪칠 뿐만 아니라, 논리적으로도 명쾌한 해답을 주지 못했기 때문이다. 예를 들어 하나의 가설 혹은 명제 A, 즉 “모든 백조는 희다”(Allsatz)에서 명제 B, 즉 “백조 a는 희다”(Basissatz)를 논리적으로 도출할 수 있는데, A와 B 사이에는 논리적으로 다음과 같은

$$A \rightarrow B \equiv \neg B \rightarrow \neg A$$

$$\neq B \rightarrow A$$

표 3. 입증과 반증의 진리표

A	B	A→B	(A→B)∧B	(A→B)∧¬B	입증: (A→B)∧B → A	반증: (A→B)∧¬B → ¬A
T	T	T	T	F	T	T
T	F	F	F	F	T	T
F	T	T	T	F	F	T
F	F	T	F	T	T	T

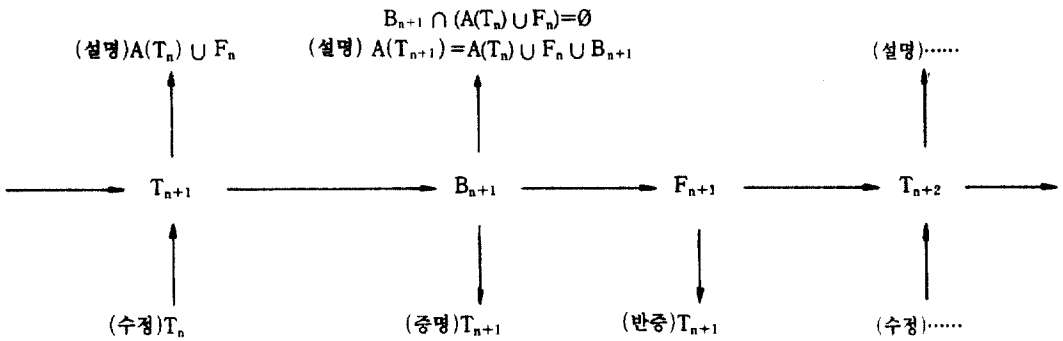
관계가 성립된다. 즉, B가 확인 내지 입증되었을 때에는 A가 참일 수도 있고, 거짓일 수도 있기 때문에 A가 입증될 수 없는 반면에, B가 반증되었을 때에는 A가 함께 반증될 수 있다는 것이다. 이는 표 3에서도 볼 수 있듯이 $(A \rightarrow B) \wedge B \rightarrow A$ 가 항상 참이 아닌 논리적 추론인데 반하여, $(A \rightarrow B) \wedge \neg B \rightarrow \neg A$ 는 항상 참인 논리적 추론으로서 恒眞命題(tautology)이다(modus tollens). 그런 의미에서 입증과 반증은 비대칭적인 성격을 띄고 있다고도 할 수 있다(Westermann, 1987b).

이러한 반증의 개념을 학문적인 이론의 발전을 위하여 도입한 Popper의 기본 정신은 그림 3으로 잘 나타낼 수 있다. 즉 하나의 이론 T_{n+1} 가 그 이론의 적용으로서 설명될 수 있는 관찰영역(B)까지는 유용한 이론으로 존재하다가, 반증되는 관찰영역(F)에 당면하면 새로운 이론 T_{n+2} 로 확대되어야 한다는 것이다. 이때 이 새로운 이론은 그전의 이론 T_{n+1} 로 설명할 수 있었던 내용은 물론, 설명할 수 없었던 F_{n+1} 까지도 포함할 수 있어야 하는데, 이 역시 마찬가지로 F_{n+2} 에 당면하게 되면 반증될 수 있다. 그러므로 Popper에 의하면 과학은 이

러한 점진적인 이론의 확대에 의해서만 발전할 수 있다고 한다. 따라서 경험적 연구에서 모든 가설은 반증가능성(falsifiability)을 지녀야 함은 물론, 그들의 증명이란 모두 임시적인 성격(vorläufiger Charakter)을 띠 수밖에 없다는 것이다(Popper, 1963). 물론 학문적인 이론이 항상 이렇게 이상적으로만 확대, 전개될 수 있는가 하는 것은 중요한 비판점이 될 수 있으나, Popper의 반증 논리는 전반적으로 현대과학의 방법론에 기여한 바 크다. 그리고 이러한 내용을 다시 그림 2와 연관시켜 볼 때, 최소한 이 시점에서 우리가 인식할 수 있는 것은, 실제, 관찰 및 명제체계 영역의 울타리내에서 확실한 것, 즉 참(眞)에 도달하거나 참에 대한 확신을 갖게 된다는 것이, 의외로 그렇게 단순한 문제가 아니라는 것이다.

가설의 논리적 구조와 검증

앞절에서 다룬 반증(falsification)에 대한 논의는 “모든 백조는 희다”와 같은 특정한 형태의 가설에 국한시



T : 이론 A : 응용영역 B : T로 설명된 관찰영역 F : T로 반증된 관찰영역

그림 3. Popper의 반증 schema

켜서 이루어졌다. 그러나 여기에서 우리가 꼭 염두에 두어야 할 사항은, 경험적인 연구에서 다루어질 수 있는 가설의 형태가 다양하다는 사실이다. 그러므로 입증(verification)과 반증(falsification)의 관점에서, 이렇게 다양한 가설들이 어떠한 속성을 띠고 있는지 살펴볼 필요가 있다.

이에 대한 개관은 표 4에 제시되었다(Bredenkamp, 1972, 참고). 가설의 형태는 대체적으로 “모든 ...이다”로 시작되는 보편양사(universal quantifier, Allsatz), “최소한 ...있다”로 표현되는 존재양사(existential quantifier, Existenzsatz) 및 확률적인 내용을 포함하고 있는 확률명제(Wahrscheinlichkeitsaussage)로 나뉘어 진다. 그리고 각 명제들은 또 다시 무한성을 띠고 있는 명제와 시간과 공간의 제한을 받는 유한성의 명제로 구분된다. 여기에서 특징적인 것은, 유한성을 띤 모집단에 대한 명제는 원칙적으로 모두가 입증과 반증이 함께 될 수 있다는 것이다. 그 이유는 입증이 되든 반증이 되든, 유한한 모집단에 속하는 개체는 연구자가 원할 경우 실제로 모두 확인이 가능하기 때문이다. 그러나 무한성을 띠고 있는 보편양사, 즉 “모든 백조는 희다”의 경우는 다르다. 아무리 많은 백조들을 찾아서 희다는 것이 입증되었다(verified)해도, 아직까지 확인되지 않은, 그리고 무한성의 특징때문에 앞으로도 확인될 수 없는 백조

들 중에는 검은 백조가 있을 수 있다는 의문의 여지를 항상 남겨 놓고 있기 때문에, 궁극적으로는 그러한 명제가 입증될 수 없다는 논리이다. 반면에 그러한 명제가 틀렸다는 것을 반증하는 것은 간단하다. 최소한 하나의 검은 백조를 발견하고나면 그 다음의 백조들은 더 이상 확인할 필요가 없어지기 때문에, 그 명제의 논리적 구조는 반증(falsification)에 의해서만 증명 가능하게 된다. 이는 다음과 같은 논리적인 관계

$$\begin{aligned} \neg \{ \bigwedge_x (S_x \rightarrow W_x) \} &\equiv \bigvee_x \{ \neg (S_x \rightarrow W_x) \} \\ &\equiv \bigvee_x \{ \neg (\neg S_x \vee W_x) \} \\ &\equiv \bigvee_x (S_x \wedge \neg W_x) \end{aligned}$$

에서도 알 수 있듯이 $\bigvee_x (S_x \wedge \neg W_x)$, 즉 백조라고 불리면서 동시에 희지 않은 x를 최소한 하나만 확인할 수 있으면, 그것은 $\neg \{ \bigwedge_x (S_x \rightarrow W_x) \}$, 즉 “모든 백조는 희다”를 부정하는 명제와 논리적인 동치(equivalent)가 되므로 반증이 완료되는 것이다. 같은 논리적 맥락에서 볼 때, 무한성을 띠고 있는 존재양사는 입증은 가능하나(verifiable), 반증은 불가능하다. 그리고 무한성을 띤 확률명제는, 어느 경우이건 확률의 최종적인 확인이 불가능해지기 때문에 입증도, 반증도 할 수 없게 된다.

그러면 경험적인 연구에서 주로 사용되고 있는 가설

표 4. 가설의 형태와 가능한 검증방법

구분	모집단		예	기호 표시	검증방법	
	무한	유한			verifiable	falsifiable
보편양사 (universal quantifier)	○		모든 백조는 희다.	$\bigwedge_x (S_x \rightarrow W_x)$	×	○
		○	특정 장소 S에서 특정시간 t에 있는, 모든 백조는 희다.	$\bigwedge_x (S_{xst} \rightarrow W_{xst})$	○	○
존재 양사 (existential quantifier)	○		최소한 하나의 흰 백조가 있다.	$\bigvee_x (S_x \wedge W_x)$	○	×
		○	특정장소 s에서 특정시간 t에, 최소한 하나의 흰 백조가 있다.	$\bigvee_x (S_{xst} \wedge W_{xst})$	○	○
확률 명제	○		동전을 던져 숫자가 나올 확률은 1/2이다.	$P(\text{동전 숫자})=1/2$	×	×
		○	특정 장소 s에서 특정시간 t에, 백조가 있을 확률은 0이다.	$P(\text{백조}/s,t)=0$	○	○

의 형태는 무엇일까? 사회과학분야, 그 중에서도 특히 심리학에서는 이론이나 관찰에서부터 도출하는 가설을 통상 무한 보편양사의 형태로 채택하고 있다. 그 이유는 하나의 가설이 보편성을 띄고 있지 못한다면, 설사 그 가설이 증명되었다고 해도 현실적으로 확인된 자료 수준 이상의 것을 주장할 수는 없기 때문이다. 그러나 이렇게 보편성을 띤 가설로부터의 출발은, 경험적인 연구의 과정에서 아무리 늦어도 모집단을 결정하는 시기에는, 유한 보편양사의 형태로 제한이 가해진다. 따라서 원칙적으로 이 때의 가설들은 모두 입증과 반증이 가능해진다. 그러나 일상적인 경험적 연구의 과정에서는, 모집단이 아무리 유한성을 띄고 있더라도 그 모집단의 자료들을 모두 수집하고 분석할 수 있는 경우는 극히 드물다. 그러므로 모집단으로부터 무선표집을 뽑고, 그 표집에서 획득된 자료를 통하여 거꾸로 모집단의 모수치(parameter)를 추정하게 된다. 이러한 연구절차가 갖는 의미는, 유한 보편양사의 형태를 띄고 있는 모집단에 대한 가설을, 마치 무한 보편양사이거나 한 것처럼 다루는 것과 같다. 그래서 결국은, 표 4에서도 보았듯이, 표집의 자료만 가지고 모집단에 대한 가설을 검증할 수 있는 방법이란 반증(falsification)밖에 없다는 것을 알 수 있다. 그리고 그 반증의 과정에서 ST의 실시를 위한 통계적 가설이 도입되는데, 이는 항상 관찰된 무선변인의 확률분포에 대한 명제를 내포하고 있기 때문에, 결국은 확률적인 명제로 취급될 수 밖에 없다. 따라서 통계적 가설은, 상술한 바와 같은 유의도 검증과 무선표집의 문제점들에서 드러난 바와 같은 그 검증 방법상의 절차때문에, 입증할 수도 반증할 수도 없는 논리적 구조를 지니고 있다(Bredenkamp, 1972).

Neyman과 Pearson의 검증이론

가설검증에는 다양한 이론들이 존재하지만, 흔히 보급되어 있는 Neyman과 Pearson의 고전적 검증이론은 통상 그림 4와 같은 절차를 밟아 시행되고 있다. 그 전반적인 절차의 흐름은, 이론 및 가설검증이라는 과학의 목표를 달성하기 위하여 경험적인 자료와 수학적인 모델의 도움을 통하여 합리적인 결정을 내리려는 노력이

라고 요약될 수 있다. 그 내용은 기본적인 이론에서부터 시작하여 우리에게 잘 알려져 있는 ST에 이르기까지 다양한 개념들을 내포하고 있다. 여기에서 다루어지는 가설들(연구가설, 조작적 및 통계적 가설)은 모두 유한성을 띄고 있는데, 그러한 조건 하에서 연구가설은 대체적으로 보편양사(Allsatz)의 형태를 띄고 있으면서, 가설의 유효영역, 즉 모집단을 규정하고 있는 반면, 조작적 가설에서는 연구 대상 변수들의 조작적 정의와 표집의 선정까지도 구체적으로 언급되어야 하기 때문에 보다 제한된 영역에서의 보편양사 형태를 띄게 된다. 이러한 조작적 가설은 구체적인 연구계획이 수립된 후에야 성립될 수 있는 것이며, 따라서 경험적 연구의 구체적인 시발점이기도 하다. 그 다음의 통계적 가설은 일반적으로 상관계수, 회귀계수, 평균, 분포 및 그 지표등 다양한 형태로 제시되고 있는데, 그 내용은 해당 연구에서 사용되는 통계방법에 의하여 규정된다. 이 가설은 엄밀히 따지면 상호 보완적인 2개의 가설로 구성되어 있다. 하나는 소위 상위가설(Oberhypothese)로서 연구대상이 되는 변수의 분포형태(예: 정상분포) 혹은 특성(예: 변수 상호간의 독립성)등을 가정하는 것이다. 또 다른 하나는 협의의 통계적 가설로서 지표가설(Parameterhypothese)이라고 하며, 앞서 가정된 분포의 지표들(예: 평균, 편차 등)을 제시하고 있다(Groeben & Westmeyer, 1975). 이러한 통계적 가설은 관찰된 무선 변인의 확률 분포에 대한 가정과 더불어 항상 분포 가설 및 확률적인 명제를 내포하고 있다. 그리고 이러한 가설의 검증은 일반적으로 결정론적(deterministic)인 성격을 띤 보편양사가 검증되는 경우와는 다르다. 즉, 그 가설 자체 하나만 고립된 상태에서 검증되는 것이 아니라, 필히 다른 통계적 가설들에 대비시켜 상대적으로만 검증이 가능토록 되어 있다. 우리가 흔히 사용하고 있는 영가설(H_0)과 대립가설(H_1)이 바로 그것인데, 결국 ST는 이렇게 상호 경쟁적인 입장에 있는 가설들 간에 합리적인 비교평가를 통하여 그 중 하나의 가설 혹은 가설군을 채택하고, 다른 한쪽을 기각하는 것을 목표로 삼고 있다.

일단 여기까지는 가설검증을 위한 연구계획의 단계로서 연구가설과 경험을 연결시키는 개략적인 구상을 나

타내는 것이다. 그 다음에는 경험적인 단계로 들어가서, 우선 자료를 무선표집계획에 의하여 수집한 후 통계적인 지표(예 : 평균차, 상관계수, 2변량의 대비등)를 계산하게 되는데, 이 지표는 가설과 관련되는 모든 중

요한 정보를 포함하고 있다. 이때 대두될 수 있는 의문은, 표집에서 특정한 지표로 나타난 현상이 우연에 의한 것인지, 아니면 모집단의 특성이 표집에서도 대표적으로 반영된 것인지의 여부로서, 하나의 결정을 필요로

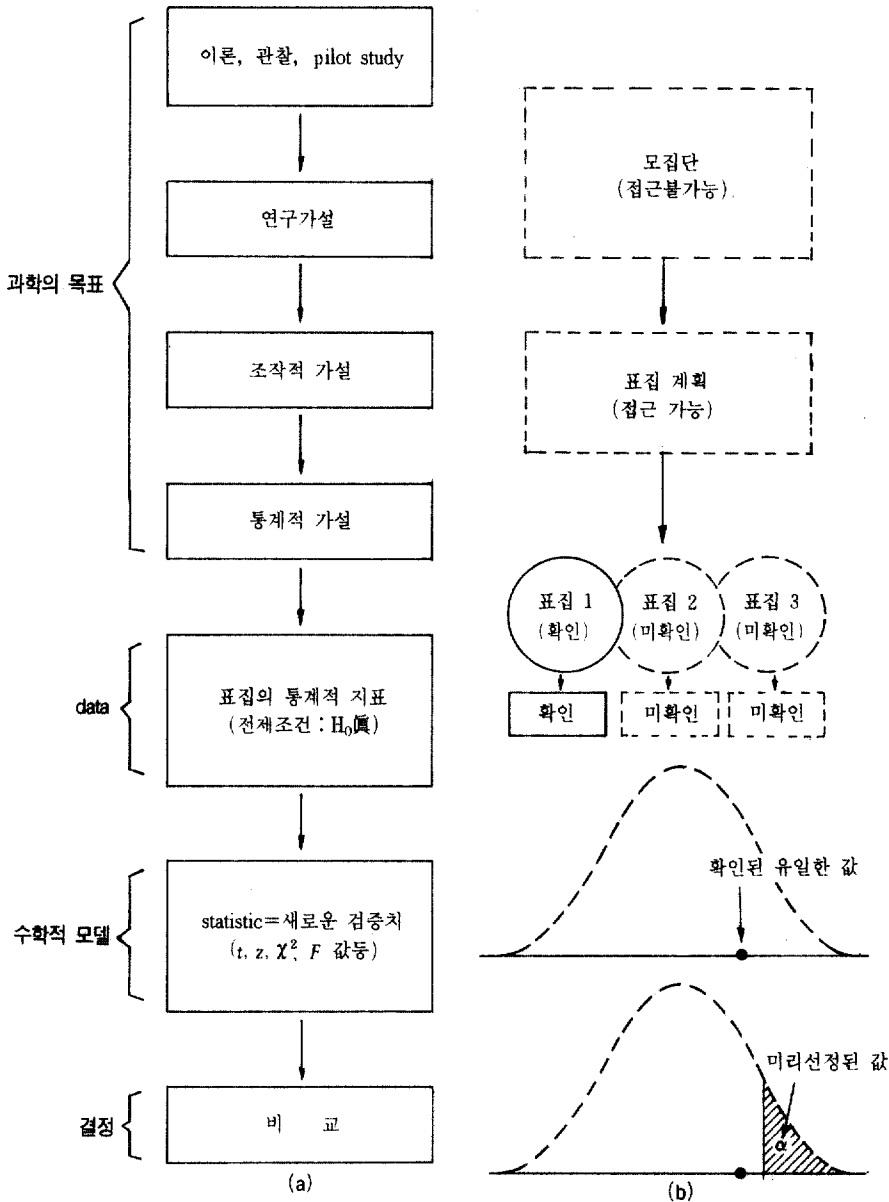


그림 4. 유의도 검증의 절차 : (a)내용 (b)가정된 부분(점선)과 시행된 부분(실선)

한다. 그런데, 표집 결과로 획득된 통계적 지표는, 해당되는 무선표집의 특성에 따라 결정되었기 때문에, 똑같은 표집계획에 의하여 연구를 반복한다고 하더라도 동일한 통계적 지표를 얻기란 어렵게 된다. 따라서 이러한 통계적 지표는 하나의 무선변인처럼 취급되는데, 이때 표집으로부터 획득된 지표가 극단적인 것인지 혹은 대표적인 것인지에 대한 판단은 특정 무선변인 “통계적 지표”의 분포(sampling distribution에서 density function 혹은 probability function)를 알고 있을 때에만 가능하다. 그러나 그림 4의 (b)에서도 표시되었듯이 우리가 선택한 표집은 단 하나이므로, 표집에 의하여 무선 변인 “통계적 지표”의 분포를 알 수 있는 방법이 없다. 다만 우리가 모집단의 모수치(parameter)들을 미리 알고 있다면, 당연히 그 분포를 알 수 있게 되는데, 그런 상황이란 실제로는 있을 수 없기 때문에, 바로 이 부분에서 ST를 도입하게 되는 것이다.

근본적으로 모집단의 모수치에 대해서 우리가 언급할 수 있는 길은 추측 밖에는 없다. 그렇기 때문에 진정한 모수치가 정확하게 무엇이라고 얘기할 수는 없더라도 하나의 우회적인 방법을 선택할 수는 있다. 즉, 일반적으로 영가설(H_0) 혹은 영가설이 주장하는 모집단의 특성이 옳다고 가정하고, 그러한 특정 조건하에서 표집 지표가 어떤 분포를 나타내는가를 확인해 본다는 것이다. 바로 이같은 문제는 수학적인 과제로서, 흔히 사용되고 있는 통계적 지표에 대한 모든 해답이 이미 제시되어 있는 상태이다. 그 내용은, 통계적 지표와 관련되는 보조 가정들이 충족된다는 조건하에서, 모든 종류의 표집 지표의 분포들은 잘 알려져 있는 소수의 수학적 분포들로 환원될 수 있다는 것이다. 이 때 수학적인 전환(transformation)을 통하여 새로운 지표 혹은 검증치(예: t -, z -, χ^2 -, F -값등)를 얻게 되는데, 이러한 지표들의 분포는 어떤 통계책에서나 확인될 수 있다.

이렇게 본다면 결국 ST란, 경험적으로 획득되고 수학적으로 전환된 지표와 그에 상응하는 지표분포에서 미리 선정된 α %에 해당되는 지표값과의 비교라고 집약될 수 있다. 그런데 이러한 유의도(significance)여부와 관련되어 다루어지는 명제는, 어디까지나 H_0 이 옳다는 것을 전제로한 조건적인 확률명제이다. 그 확률은

H_0 이 참(眞)이라고 가정될 때, 통계적 지표를 통하여 서술된 표집결과와 확률을 말하는 것으로서, 예를 들면 계산된 확률이 미리 선정된 α 보다 작을 경우 H_0 을 기각할 수 있다는 것이다. 여기에서 간과할 수 없는 하나의 중요한 사실이 있다. 그것은 지금까지 ST의 일환으로 언급된 확률이라는 것이, 검증하고 싶은 연구가설 그 자체에 대한 확률을 나타내는 것이 아니라, 다만 통계적 지표의 확률, 그것도 H_0 이 참일 경우에 대해서만 국한되고 있는 조건확률이었다는 점이다. 따라서 하나의 유의미한 결과가 나타났다고 할지라도, 그것은 연구가설이 참이라는 확률을 의미하는 것이 아니다. 연구가설 그 자체는 하나의 명제로서 참이 아니면 거짓일 수밖에 없다. 그런데 이와 병행하여 ST의 결과 역시 참일 수도 있고 거짓일 수도 있기 때문에 가설검증에 대한 판단은 당연히 위험 부담을 안게 된다. 우리는 이와 관련된 사항을 앞의 표 2에서 확인한 바 있다.

지금까지 우리는 H_0 이 모집단에서 참(眞)일 것이라는 가정하에 논리를 전개하여 왔다. 그러나 만약 우리의 가정이 틀렸었다면 어떻게 해야 할까? 예를 들면, 모집단에서 H_0 가 아닌 어느 특정한 H_1 의 값이 참일 경우를 생각할 수도 있는데, 그때 우리는 그 H_1 의 값을 기준으로 한 새로운 확률 분포를 설정하여야 한다. 다음에는, 그 새로운 분포를 중심으로 H_0 의 경우에서 전개되었던 것과 같은 맥락의 사고와 논리가 H_0 과 H_1 의 위치만 바뀌어서 적용된다. 그런 의미에서 볼 때 H_0 과 H_1 은 상호 대칭의 입장에서 있다고도 할 수 있다. 다만 그러한 경우들의 검증치에 대한 분포를 우리가 흔히 사용하고 있는 통계책에서는 찾아볼 수 없을 뿐이다.

유의도 검증의 논리적 구조

McGuigan(1978)도 똑같은 오류를 범하고 있지만, 가설검증에서 우리가 흔히 범하기 쉬운 논리적인 오류가 있다. 그것은 가설 “A이면 B이다($A \rightarrow B$)”가 있을 때 $A \wedge B$ 의 확인, 즉 A와 B가 함께 입증(verification)될 수 있다면 그 가설이 참(眞)일 것이라고 추론하는 논리이다(Bredenkamp, 1972). 그러나 우리는 앞에서

$$A \rightarrow B \equiv \neg (A \wedge \neg B)$$

의 논리적 구조를 통하여 $A \wedge B$, 즉 A의 입증과 B의 반증이 함께 확인되어야만 가설의 반증이 성립되고, 기타의 경우에는 가설 $A \rightarrow B$ 가 참(眞)이라는 것을 증명할 수 있는 가능성이 없다는 것을 알았다.

이제 우리는 이러한 개념을 확대시켜서 명제 혹은 가설들간의 함축(implication)관계를 고찰해 보기로 하자. 앞절에서 언급되었던 가설검증 절차상의 여러 가설들은 각각

- 이론(T) → 연구 가설(WH)
- 연구 가설(WH) → 조작적 가설(OH)
- 조작적 가설(OH) → 통계적 가설(SH)

의 형태로서 논리적으로 도출되어 내포관계를 형성하고 있다. 이때 내포되는 가설은 유효분야가 좁아지게 되는데, 그 유효분야가 좁아질수록 가설의 일반화 정도 역시 감소하게 된다. 그런데 Popper(1966)에 의하면 일반화 정도가 낮은 가설의 반증(falsification)은, 일반화 정도가 높은 가설도 함께 반증하게 된다. 그러므로 Bredenkamp(1972)는 표 5와 같이 논리적으로 점차 범위를 좁혀가며 도출 가능한 연구 가설, 조작적 가설 및 통계적 가설을 예로 들고, 그들 간의 내포관계에 따른 가설검증의 가능성을 따져 보았다. 만약 우리가, 어떤 연구 가설이 참(眞)이라는 것을 확실히 알고 있다면, 거기서부터 논리적으로 도출되는 OH와 SH는 도출상의 논리적인 비약이 없는한 참이어야 한다. 즉 WH가 참일 때, 자연히 OH는 참이며, 또 OH가 참일 때 SH도 역시 참이 된다. 그러나 우리가 경험적인 연구에서 적용하는 ST에서는, 거꾸로 통계적 가설(ST)의 검증을 통하여

궁극적으로 연구 가설(WH)내지 그 가설이 도출된 이론(T)을 증명하려는 것을 목표로 삼고 있다. 따라서 우리는 연구 가설이나 이론의 진위 여부를 모르는 채 통계적인 추론을 해야만 하는 문제를 안고 있다. 이때 이들 가설들간의 논리적인 관계를 고찰해 볼 필요가 있는데, 그들간에 존재할 수 있는 모든 가능한 관계를 표 6과 같은 진리표가 제시하고 있다. 그러나 표 6에 대한 고찰에 의하면, 우리는 이러한 가능성들이 모두 우리의 관심 대상이 될 수는 없다는 것을 쉽게 알 수 있다.

통계적 추론을 위한 ST의 논리에서는, 우선 그 가설들간의 논리적 구성이 $WH \rightarrow OH$ 가 참이어야 하고, 또 $OH \rightarrow SH$ 가 참이어야 하기 때문에 $(WH \rightarrow OH) \wedge (OH \rightarrow SH)$ 가 참인 가능성만 고려되어야 하는데, 그러자면 표 6에서는 위에서부터 넷째줄까지만 우리의 고려대상이 될 수 밖에 없다. 그 중에서도 특히 넷째줄이 시사하는 것은

$$\{(OH \rightarrow SH) \wedge \neg SH\} \rightarrow \neg OH,$$

그리고 $\{(WH \rightarrow OH) \wedge \neg OH\} \rightarrow \neg WH$

로서, $\neg SH$ 가 확인됨으로서 OH 및 WH가 각각 반증되는 경우이다. 또 첫째줄은 이상적인 경우로서 WH, OH 및 SH가 각각 참인 경우를 나타내고 있다. 그러나 반대로 SH가 참이라고 하더라도 OH, 더 나아가서 WH가 꼭 참이라는 논리는 성립되지 않는다. 표 6의 둘째 및 셋째 줄이 그러한 가능성을 시사하고 있기 때문이다. 그러므로 이 시점에서 다시 한번 분명히 하고 넘어가야 할 사항은, 통계적인 가설검증 그 자체가 직접적으로 연구 가설의 증명여부를 의미할 수 없으며, 또 결정할 수도

표 5. Bredenkamp(1972)가 예로 든 가설들

구 분	내 용	설 명
연구가설 (WH)	$\bigwedge (i) \{Y_i = a_i X_i + b_i\}$	정의된 모든 개인 i에 해당되는 것은; 변수 X와 Y간에 선형관계가 존재하는데 a와 b의 개인차가 허용된다.
조작적 가설 (OH)	$\bigwedge n(i) \{Y_i = a_i X_i + b_i\}$	정의된 개인 i에 상응하는 최소한의 n명이 있는데 그들에 해당되는 것은; 변수 X와 Y간에 선형관계가 존재하는데 a와 b의 개인차가 허용된다.
통계적 가설 (SH)	$\bar{Y} = aX + b$	변수 Y의 평균치와 변수 X간에는 선형관계가 존재한다: 선형관계로서의 평균치 경향

표 6. Bredenkamp(1972)가 제시한 가설들간의 진리표

WH	OH	SH	WH→OH	OH→SH	(WH→OH)∧(OH→SH)
T	T	T	T	T	T
F	T	T	T	T	T
F	F	T	T	T	T
F	F	F	T	T	T
F	T	F	T	F	F
T	T	F	T	F	F
T	F	T	F	T	F
T	F	F	F	T	F

없다는 것이다. 이것은 논리적으로도 뒷받침될 수 없는 내용이다.

다음에는 ST결과의 해석상의 문제를 논리적 구조를 통하여 알아보기로 하자. 앞절에서 우리는 통계적 가설이 상위 가설과 지표 가설로 구성되고 있음을 보았고, 또 지표 가설은 H_0 과 H_1 , 2개의 가설 혹은 가설군으로 나뉘어짐을 보았다. 이때 ST란, 표집에 의하여 실제로 획득된 자료 분포가 H_0 이 옳다는 가정하에 나타나는 자

료분포로부터 얼마나 차이가 나는가를 알아보는 것이다. 그러한 목적달성을 위하여 추진되고 있는 ST의 논리를 그림 5의 (a)를 참고하여 다시 살펴보기로 하자. 모집단에서 H_0 이 옳다는 가정(엄밀히 말하자면, 분포형태에 관한 가정도 추가)을 할 수 있다면, 검증치(teststatistic)는 $1-\alpha$ 의 확률을 가지고 A 영역에 놓여지게 되는데, 그와 반대로 검증치가 만약 그 영역 밖($\neg A$)에 놓여진다면 α 의 오류 확률을 가지고 H_0 이 기각된

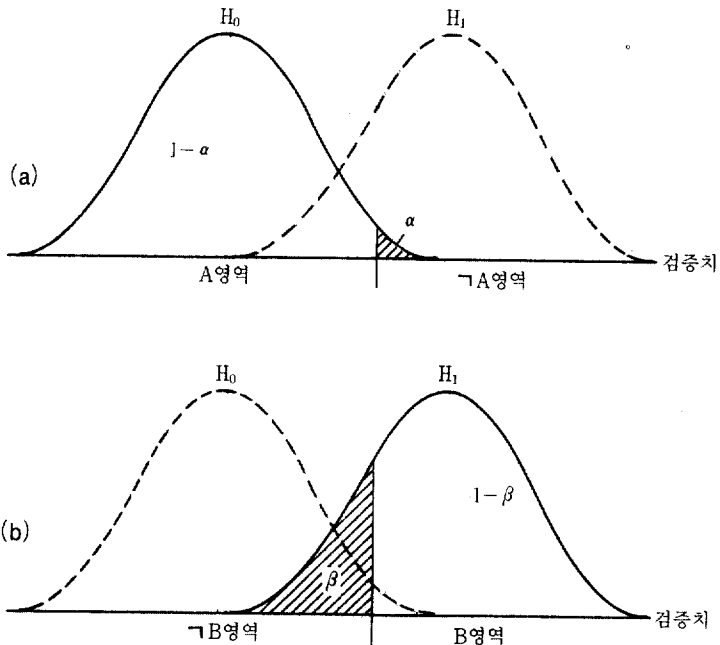


그림 5. 모집단에서 각각 H_0 (a)와 H_1 (b)이 참이라는 가정 하에 생각할 수 있는 통계적 지표의 영역

다. 이를 우리는

$$\{(H_0 \xrightarrow{1-\alpha} A) \wedge \neg A\} \xrightarrow{\bar{\alpha}} \neg H_0$$

라고 표시해 보자.⁵⁾ 이러한 논리는 정당한 것으로서, 여기까지는 ST결과의 해석상의 문제는 대두되지 않는다. 그러나 논리의 비약은 그 다음 단계에서부터 시작된다. 사실상 현 단계까지 한가지 분명해진 사실이 있다면, 그것은 다만, 우리가 H_0 을 α %만큼의 오류확률을 가지고 기각시켰다는 사실일 뿐이다. 그리고 이때 우리는 H_0 과 H_1 이 상호 대립적인 관계($\neg H_0 \equiv H_1$, $\neg H_1 \equiv H_0$)에 있다는 것을 알기 때문에, 기각된 H_0 대신 H_1 을 채택하는 것일 뿐이다. 그렇다고 해서 H_0 에 관련된 확률들마저 H_1 에 일반적으로 전이(transfer)될 수는 없다. 그 이유를 들기 전에, 우선 H_0 과 H_1 의 구성 형태를 다시 생각해 보자. H_0 는, 예를 들어, 모집단의 모수치가 $\mu = 0$ (양방적인 경우)의 단순가설이거나 혹은 $\mu \leq 0$ (일방적인 경우)의 복합가설로 구성되어 있다. 그런데 전자의 경우와 같이 후자의 경우에서도, 대립가설(H_1)에 대응하기 위하여 영가설(H_0)을 대표할 수 있는 값은 μ_i 의 극단치인 $\mu = 0$ 이 된다. 따라서 어떤 경우이건, 영가설에서는 모집단의 모수치가 $\mu = 0$ 이라고 가정할 때의 표집분포만을 생각하게 되는데, H_0 의 기각이란, 획득된 경험적인 자료가 최소한도 이러한 가정에 부합될 확률이 작다는 것을 말해 줄 뿐이다. 그러나, 그대신 채택된 대립가설(H_1)도 역시 여러 단순가설로 구성된 복합가설로서, H_0 만 기각된 상태로는, 그 중 어떤 가설이 옳은 것인지도 모를 뿐만 아니라, 또 어느 특정한 단순가설 H_1 의 진위여부를 가정할 때의 확률은 α 대신 β 로 서 대치되기 때문이다.

이러한 과정에서 발생하는 논리적인 오류나 비약은, 원래 H_0 이 옳다는 가정하에서 확인된 표집분포에 내재되어 있는 $p(A/H_0) + p(\neg A/H_0) = 1$ 의 관계를 무시하고, 마치 $p(\neg A/H_0) + p(\neg A/H_1) = 1$ 의 관계가 성립되고 있는 것처럼 다루는 데에 그 근본적인 원인이 있다.

이 잘못된 관계식은 정리하면

$$\begin{aligned} p(\neg A/H_1) &= 1 - p(A/H_0) \\ &= 1 - \alpha \end{aligned}$$

와 같이 되는데, 여기에서 좌측 항은 $1 - \beta$ 이기 때문에 결국 $1 - \beta \neq 1 - \alpha$ 가 성립됨을 알 수 있다. 바로 그렇기 때문에 H_1 의 채택이 옳다는 통계적 추론의 결과에 대하여 $1 - \alpha$ 만큼의 신뢰를 부여하는 것은 오류이며, 한 걸음 더 나아가서 마치 절대적인 결과인 것처럼 확신하려는 경향 역시 오류이다. 이러한 논리의 비약들은 다음과 같은 논리식⁶⁾

$$\begin{aligned} &\{(H_0 \xrightarrow{1-\alpha} A) \wedge \neg A\} \xrightarrow{1-\alpha} H_1 \\ \text{혹은 } &\{(H_0 \xrightarrow{1-\alpha} A) \wedge \neg A\} \xrightarrow{\prime} H_1 \end{aligned}$$

들을 통하여 나타낼 수 있으며, 방금 언급된 내용들을 고려할 때, 그 논리적인 추론들이 참(眞)이 될 수 없다는 것을 우리는 한 눈에 알 수 있다. 그리고 물론 이들은 바로 위에서 제시된 ST의 논리식과는 同値(equivalent)가 될 수 없다는 것도 알 수 있다. 다시 말하면, 일정한 오류 확률 α 를 가지고 H_0 을 기각한다는 사실은, ST의 논리에서 볼 때 H_1 에 대한 $1 - \alpha$ 만큼의 확신을 의미하는 것이 아니기 때문에, 무작정 α 가 작을수록 좋다고 말할 수는 없으며, 더 나아가서 H_1 이 마치 참(眞)인 진리로서 밝혀지거나 한 것처럼 취급되어서도 안된다는 것이다.

이와 똑같은 논리로, 이번에는 모집단에서 H_1 이 옳다는 가정을 하고 그림 5의 (b)를 참고하면 다음과 같은 논리식들

$$\begin{aligned} &\{ \{(H_1 \xrightarrow{1-\beta} B) \wedge \neg B\} \xrightarrow{\beta} \neg H_1 \} \\ &\neq \{ \{(H_1 \xrightarrow{1-\beta} B) \wedge \neg B\} \xrightarrow{1-\beta} H_0 \} \\ &\neq \{ \{(H_1 \xrightarrow{1-\beta} B) \wedge \neg B\} \xrightarrow{\prime} H_0 \} \end{aligned}$$

으로 표시될 수 있다. 이러한 논리 역시 마찬가지로인데

5) 여기에서 $\xrightarrow{1-\alpha}$ 는 $(1 - \alpha)$ %만큼의 확신을, $\xrightarrow{\bar{\alpha}}$ 는 α %만큼의 오류 확률을 나타내는 함꼐관계를 의미함.

6) 여기에서 $\xrightarrow{\prime}$ 표시는 H_1 의 논리적인 추론이 옳다는 주관적 확신을 의미함.

표현하자면 다음과 같다: 일정한 확률 β 를 가지고 H_1 을 기각한다는 사실은, $1-\beta$ 만큼의 H_0 에 대한 확신을 의미하는 것이 아니고, 또 H_0 의 채택이 절대적인 확신을 보장하지도 않는다. 그러므로 이와 같이 ST의 논리가, 앞에서 다루어진 Neyman과 Pearson의 검증이론에서도 보았듯이, 확률적인 명제의 범위내에서 벗어날 수 없다는 사실은, 통계적 가설(SH)의 검증에 의하여 하나의 이론(T)이나 연구 가설(WH)의 진위여부를 밝히는데 도움이 되지 못한다는 주장을 다시 한번 지지해 주는 것이다.

결과에 대한 올바른 해석과 대안

이제까지 앞절에서 다루어진 논리식에서는, 주로 상호 내포와 외연의 관계에 있는 가설 혹은 명제등의 논리적 관계들만을 고찰해 보았다. 그러나 하나의 이론 혹은 가설의 검증에 있어서 흔히 사용되는 ST를 통하여 통계적 추론(statistical inference)을 할 수 있기까지는 수 많은 보조 가설들이 가정부분으로서 도입되고 있다. 그렇다면 예측된 하나의 현상이 반증됨으로 해서, 그 예측을 가능케 했던 가설이 반증될 수 있는 가능성은 더욱 희박해진다. 즉

$$[\{(WH \wedge ZH_s) \rightarrow E\} \wedge \neg E] \rightarrow \neg (WH \wedge ZH_s)$$

에서 연구 가설(WH)과 그에 상응하는 수많은 보조 가설군(ZH_s)이 있을 때, WH로서 예측되어야 하는 현상 E가 반증되면, WH만 반증되는 것이 아니라 WH와 ZH_s 가 한꺼번에 반증된다. 그러나 이것은

$$\neg (WH \wedge ZH_s) \equiv \neg WH \vee \neg ZH_s$$

에서 보는 바와 같이 논리적으로 연구 가설(WH)만이 반증되거나, 혹은 보조 가설군(ZH_s)만이 반증되거나, 혹은 모두 다 반증될 수 있는 3가지 가능성을 안고 있다. 따라서 예측된 현상 E의 반증($\neg E$)이 꼭 연구 가설(WH)을 반증한 것이라고 볼 수 없게 된다는 것이다. 이러한 논리적 구조는, 본 논문의 맥락에서 본다면, 가

설이 검증되어야 한다는 연구자들의 기대와 그들이 실제 도달 할 수 있는 현실간의 간격을 더욱 심화시켜주는 또 하나의 중요한 요소가 된다. 그러나 다른 한편, 이러한 논리는 간혹 연구자의 입장에서 자신의 “가설보호”라는 논리로 적용되기도 한다. 즉 E가 반증될 경우에 연구자 자신의 가설이 유효하지 않다는 것을 인정하기 이전에, 우선 수많은 보조 가설들 중의 일부에서 그 오류를 찾아내려는 경향도 있다는 것이다.

이러한 문제 이외에도, 앞절에서 보았듯이, ST결과에 대한 해석 문제는 많은 오류를 발생시키고 있다. 그 오류들의 성격은 대부분 ST가 할 수 있는 것 이상으로 의미를 부여하려 하거나, 논리의 비약이라는데 초점을 두고 있다. 그리고 이러한 추론 통계(inference statistic)에 대한 과대 평가의 배후에는, 앞서 언급했듯이, 진리를 “계산해낼 수 있다”고 믿는 엄청난 착각이 자리 잡고 있는 것이다. ST의 근본적인 문제는, 제한된 피험자 혹은 응답자의 자료를 가지고 무한성을 띤 보편양사(Allsatz)형태의 연구 가설을 검증한다는 것 자체에 있으며, 그것은 논리적으로 많은 문제를 안고 있다.

그러므로 일부 학자들은 아예 모집단에 대한 가정을 전혀 할 필요가 없는 새로운 방식의 ST방법(예: randomization-test)을 모색하거나(Edgington, 1969), 혹은 종래의 ST를 경험적인 요소가 내포되지 않은 하나의 보조 이론 정도로만 파악하려는 극단적인 경향도 있다(Westermann, 1987a). 그럴 경우에 ST는 가설검증을 통한 “진리의 발견”에 기여한다기보다는, 일종의 결정 규칙(decision rule)으로 축소 해석될 수 밖에는 없다(Hager, 1987).

여하튼 간에, 종래 흔히 사용되어 온 ST의 논리적 구조를 고찰해 보고 난 우리로서는, 그 결과 해석에서도 보다 신중을 기하지 않을 수 없을 것이다. 즉, “연구 결과가 α %수준에서 유의미(significant)하였다”라는 검증 결과에 대한 해석은 “그에 해당되는 연구가설이 하나의 ST를 α 수준에서 견디어 낼 수 있었다”라고 표현하는 것이 최적이며, 따라서 그 이상도 그 이하도 될 수 없다는 것이다. 앞의 절들에서 고찰하였듯이, 하나의 주장이 담긴 가설은, 우리가 그것을 반증할 수 있을 때까지 유효하다고 취급하는 것이지, 그 가설이 꼭 참

(眞)이라고 할 수 있는 근거는 논리적으로도 찾아볼 수 없다는 것이다(그림 3과 그림 2참고). 그리고 그러한 소위 “유효 취급”마저도 ST의 논리적 구조 때문에 확률적인 명제로 밖에 다루어질 수 없으므로, 그만큼 신뢰도가 약화되는 것이다. 확률적인 명제라는 측면에서 볼 때 ST 결과의 의미는 다음과 같다: α 수준에서의 H_0 의 기각은 반증(falsification)을 의미하는 것이 아니며, 또 H_1 의 채택 역시 입증(verification)을 의미하는 것이 아니다. 그 이유는 어느 경우에서나 판단의 오류가 발생할 확률이 0이 아니기 때문이다.

따라서 참(眞)인 법칙 혹은 진리를 추구하는데 있어서 ST를 판단기준의 척도로 삼을 수는 없다. 우리가 할 수 있는 것은 다만 ST의 논리를 잘 파악하고, 그 결과가 절대적일 수 없다는 것을 인식하면서, 그에 알맞게 신중히 결과를 다루는 수 밖에 없다. 또한, 앞에서든 시사한 바 있으나, α 수준만 가지고 모든 것을 결정(?) 짓고 해석하려는 자세는 ST의 한쪽 면만 생각하는 것으로서 절대적으로 미흡하다. 그 이유는 본래 ST를 결정하는 조건들로서 (1) α 수준, (2) 검증력($1-\beta$), (3) 효과크기(예: ω^2 , f 등), 그리고 (4) 선정된 검증치(teststatistic)에 따르는 표집분포의 변산(이는 N에 의존적)이 있는데(Hager & Westermann, 1983), 이들 모두가 하나같이 중요한 역할을 담당하고 있기 때문이다. 그러므로 소위 α 수준의 해석 여부에서부터 야기되는 ST의 문제를 앞으로는 보다 다른 각도에서 조명해야 할 필요가 절실해진다.

결 론

가설검증의 다양한 방법들 중에서도 사회과학 분야에서 흔히 사용되고 있는 방법은 Neyman과 Pearson의 전통적인 유의도 검증법(ST)이다. 그 방법적용의 관행에서 자주 나타나고 있는 현상으로 지적될 수 있는 α 수준에 대한 해석 오류가 내포하고 있는 문제점은 심각한 것이다. 그러나 해석의 오류 그 자체도 문제이지만 그러한 오류가 단순한 해석의 수준을 넘어서서 연구결과에 대한 구체적인 조치가 뒤따르는 관행으로까지 확대될 수 있는 가능성을 고려한다면 그 심각성의 비중은

더욱 커질 것이다.

우리는 본 논문을 통하여 가설검증이라는 맥락속에서 다양한 개념들을 고찰해 보았다. 우선 경험적 연구의 대상영역 내에서 이론과 가설의 좌표를 생각해 보고 Popper의 반증개념을 도입하면서, 경험적 연구에서 원칙적으로 대두될 수 있는 가설들의 논리적 구조와 검증 방법을 생각해 보았다. 또 ST의 절차에 대한 개관을 제시하고, 그 논리적인 구조들을 반증의 논리에 입각하여 분석해 보았다.

그 분석결과로서, 우리는 ST의 결과가 결코 절대적인 성격을 띠지 않으며, 또 본래 목표하는 바 가설이나 이론의 증명과는 상당한 거리를 두고 있음을 보았다. 그러므로 ST의 유의미한 결과 여부가 그 결과를 도출케 한 경험적인 연구자체의 질적 판단을 위한 직접적인 척도나 기준이 될 수 없을 뿐만 아니라, 그 결과의 해석에 있어서도 결국은 본질적인 ST의 논리에 합당한 해석 이상의 의미가 부여되어서도 않된다는 것을 알게 되었다. 이와 같은 맥락에서 Bredenkamp와 Feger(1970)는 이미 오래전부터 독일의 사회심리학 학술지 “Zeitschrift für Sozialpsychologie”에서 이색적(?)인 논문 게재의 기준을 제시한 바 있다. 그들은 논문 기고자들에게 연구의 내적 및 외적 타당도를 고려하되 유의미하지 않은 결과들도 게재해 줄 것을 촉구하였다. 또 그들은 모든 연구에서 최소한도 효과크기(effect size)를 제시해 줄 것을 요구하였고, 가능한 한 검증력($1-\beta$)에 대한 실질적인 고려가 이루어져야 함을 강조하였다. 그 중에서 특히 최소한의 요구사항인 효과크기는 α 수준에만 국한된 무의미한 유의도 검증의 반복을 탈피하여, 설명된 변량의 비중을 나타내 주는 지표 활용을 통한 “실질적 유의도(practical significance)”의 추구라고 이해되어야 한다(Bredenkamp, 1970 참고).

필자는 바로 이러한 제안들이 ST의 논리에 충실하면서도 경험적인 연구 결과에 대한 신뢰도를 높이는 한편, 연구의 질을 향상시킬 수 있는 최선의 방법이라고 생각한다. 그런 의미에서 우리 심리학계도 주로 α 에만 국한시켜 관심을 보여 온 협의의 ST를 넘어서서 앞으로는 광의의 ST를 추구해야 한다고 권고하고 싶다.

참고문헌

- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66, 423-437.
- Bortz, J. (1984). *Lehrbuch der empirischen Forschung für Sozialwissenschaftler*. Berlin: Springer.
- Bredenkamp, J. (1970). Über Maße der praktischen Signifikanz. *Zeitschrift für Psychologie*, 177, 310-318.
- Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung*. Frankfurt am Main: Akademische Verlagsgesellschaft.
- Bredenkamp, J., & Feger, H. (1970). Kriterien für die Entscheidung über die Aufnahme empirischer Arbeiten in die Zeitschrift für Sozialpsychologie. *Zeitschrift für Sozialpsychologie*, 1, 43-47.
- Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- Edgington, E. S. (1969). *Statistical inference: The distribution-free approach*. New York: McGraw-Hill.
- Friedrichs, J. (1973). *Methoden empirischer Sozialforschung*. Reinbek: Rowohlt.
- Groeben, N., & Westmeyer, H. (1975). *Kriterien psychologischer Forschung*. München: Juventa.
- Haagen, K., & Seifert, H. G. (1979). *Methoden der Statistik für Psychologen*. Band II. Stuttgart: Kohlhammer.
- Hager, W. (1987). Grundlagen einer Versuchsplanung zur Prüfung empirischer Hypothesen der Psychologie. In G. Lüer(Ed.), *Allgemeine experimentelle Psychologie*. (pp.43-264). Stuttgart: Fischer.
- Hager, W., & Westermann, R. (1983). Planung und Auswertung von Experimenten. In J. Bredenkamp, & H. Feger(Eds.), *Hypothesenprüfung*. Enzyklopädie der Psychologie, Serie Forschungsmethoden der Psychologie, Band 5. (pp.24-238). Göttingen: Hogrefe.
- Hays, W. L. (1974). *Statistics for the social sciences*(2nd ed.). London: Holt, Rinehart & Winston.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for metaanalysis*. Orlando: Academic Press.
- Kutschera, F. von, & Breitkopf, A. (1971). *Einführung in die moderne Logik*. Freiburg: Alber Verlag.
- Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, 70, 151-159.
- McGuigan, F. J. (1978). *Experimental psychology*(3rd ed.). Englewood Cliffs, N. J.: Prentice-Hall.
- Melton, A. W. (1962). Editorial. *Journal of Experimental Psychology*, 64, 553-557.
- Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, 20, 641-650.
- Patzelt, W. J. (1985). *Einführung in die sozialwissenschaftliche Statistik*. München: Oldenbourg.
- Popper, K. R. (1972). *Conjectures and refutations: The growth of scientific knowledge*. London: Routledge & Kegan Paul.
- Popper, K. R. (1973). *Logik der Forschung*(5th ed.). Tübingen: Mohr.
- Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, 57, 416-428.
- Schulz, Th., Muthig, K.-P., & Koepler, K. (1981). *Theorie, Experiment und Versuchsplanung in der Psychologie*. Stuttgart: Kohlhammer.
- Stegmüller, W. (1969). *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie*. Band 1: Wissenschaftliche Erklärung und Begründung. Berlin: Springer.
- Stelzl, I. (1982). *Fehler und Fallen der Statistik*. Bern: Huber.
- Westermann, R. (1987a). *Strukturalistische Theorienkonzeption und empirische Forschung in der Psychologie*. Berlin: Springer.
- Westermann, R. (1987b). Wissenschaftstheoretische Grundlagen der experimentellen Psychologie. In G. Lüer(Ed.), *Allgemeine experimentelle Psychologie*. (pp.5-42). Stuttgart: Fischer.
- Witte, E. H. (1980). *Signifikanztest und statistische Inferenz*. Stuttgart: Enke.

원고 초본 접수: 1988.10.4
최종 수정본 접수: 1988.12.2

**Analysis of misusing tendency in interpretation of
 α -level of significance test**

Nam-Guk Cho

Korea Military Academy

This paper lists and analyzes the errors that can be often produced by the misunderstanding and blind faith in α level of significance test(N-P theory). To understand such errors properly, a fundamental analysis of the hypothesis testing procedure is needed. Therefore, this paper, using formal logic, attempts a logical structural analysis of hypotheses and tests of significance commonly used in empirical research. Along with a suggestion of a proper interpretation of α level, the result points out the inadequacy of the customary practice in research which set too much weight on α , and calls for further interest in the effect size and the type II error β .