과학철학 20권 3호 (2017) pp. 31-54

메이요의 엄격한 시험에 대한 비판

정 동 욱†

이 논문은 데보라 메이요의 ‘엄격한 시험’ 개념을 비판적으로 검토한다. 메이요는 ‘엄격한 시험’을 통해 방법론적 이론 미결정성의 위협을 막아내고 증거 판단을 위한 적절한 기준과 정량적 분석을 제공하고자 했다. 그러나 ‘엄격한 시험’을 일관되게 적용할 경우 방법론적 이론 미결정성의 위협은 간편하게 해결될 수 없으며, 비현실적인 대안가설에 의한 증거의 무력화 시도를 방어하기 어렵다. 또한 표본 자료에 근거한 신뢰구간 가설의 평가에서도 메이요의 엄격한 시험은 증거에 대한 깔끔한 정량적 분석을 제공해주지 못한다.

【주요어】 엄격한 시험, 증거, 데보라 메이요, 예측주의, 신뢰구간, 방법론적 이론 미결정성

† 서울대학교 과학사 및 과학철학 협동과정, zolaist@gmail.com

1. 도입

데보라 메이요(Deborah G. Mayo)에 따르면,[1] 가설과 일치하는 자료는 그것이 가설에 대한 '엄격한 시험'을 제공할 때에만 가설에 대한 좋은 증거를 제공한다. 메이요의 '엄격한 시험'이란 가설 H가 거짓일 경우 통과하기 어려운 시험을 말한다. 메이요는 가설과 증거 사이의 논리적 관계뿐 아니라 시험 절차를 증거 판단의 핵심 요소로 끌어들인 '엄격한 시험' 개념이 방법론적 이론 미결정성의 위협을 막아내고, 다양한 상황에서 증거 판단을 위한 적절한 기준을 제공할 것으로 기대했다. 특히 그는 가설 구성에 사용된 자료가 다시 가설의 증거로 사용될 수 있는지의 문제에 대해서도 '엄격한 시험'이 그에 대한 올바른 기준을 제공할 수 있을 것이라고 생각했다. 로널드 기어리(Giere 1983)와 존 워럴(Worrall 1989) 등의 사용-참신성 예측주의자는 가설 구성에 사용된 자료는 그 자료를 이용해 구성된 가설의 증거가 될 수 없다고 주장했지만, 메이요는 '엄격한 시험'만 통과한다면 가설 구성에 사용된 자료도 가설의 증거가 될 수 있다고 주장했다.

이 논문에서 나는 메이요가 엄격한 시험의 장점을 소개하기 위해 스스로 제시했던 사례들을 검토함으로써, 메이요가 엄격한 시험을 적용하는 방식이 일관적이지 않음을 보일 것이다. 즉 방법론적 이론 미결정성의 위협을 극복하기 위해 엄격한 시험을 적용한 방식을 일관되게 적용할 경우, 가설 구성에 사용된 자료는 그 자료를 이용해 구성된 가설에 아무런 증거도 제공할 수 없게 된다. 반대로 가설 구성에 사용된 자료가 그 가설에 증거를 제공할 수 있도록 엄격한 시험을 적용할 경우, 엄격한 시험은 방법론적 이론 미결정성의 위협으로부터 증거를 구제해주지 못한다. 또한 나는 메이요의 대표적인 사례인 신뢰구간 추정 절차조차도 엄격한 시험으로 일관되게 분석되는 데 상당한 어려움이 있음을 보일 것이다. 특히 그가 정의한 방식의 시험의 엄격성은 우리가 신뢰구간을 신뢰하는 근거조차 제대로 해명해주지 못한다는 것이

---

[1] Mayo (1996), p.180.

드러나게 될 것이다.

2. 메이요의 엄격한 시험과 시험의 엄격성

메이요에 따르면,[2] 가설과 일치하는 자료는 그것이 가설에 대한 엄격한 시험을 제공할 때에만 가설에 대한 좋은 증거를 제공한다. 메이요의 '엄격한 시험'이란 가설 *H*가 거짓일 경우 통과하기 어려운 시험을 말한다. 이를 보다 엄밀하게 정의하기 위해 그는 '통과/탈락'만 있는 간단한 시험에 대한 엄격성 기준(Severity Criteria, SC)을 아래와 같이 정의한다.

(SC) *H*가 거짓일 경우, 시험 절차 *T*가 *H*를 탈락시킬 확률이 매우 높다.

이에 따르면, 시험 *T*의 엄격성 S(*T*)는 아래와 같이 정량적으로 계산될 수 있다.[3]

S(*T*) = P(test *T* fails *H* | *H* is false) = 1 – P(test *T* passes *H* | *H* is false)

메이요는 이를 다음의 사례를 통해 설명한다. 프림과 커피를 탈 때 그 순서에 따라 맛이 다르다고 주장하는 사람 X가 있다. X가 정말 둘을 구분할 수 있는지를 확인하는 시험은 다음과 같다.

시험 가설 *H* : 맞출 확률 $p>0.5$ [귀무가설 H_0 : 맞출 확률 $p=0.5$]
시험 절차 *T* : 100번 중 60번 이상 맞춘다면, 가설 *H*를 통과시킨다.

위의 시험에 대한 엄격성 S(T)는 다음과 같이 계산된다.

---

2) Ibid., p.180.
3) Ibid., pp. 192-4. 메이요는 엄격성의 값을 지칭하는 기호를 따로 사용하지 않았지만, 여기서는 편의를 위해 S(T)라는 기호를 사용했다.

$$S(T) = 1 - P(\text{test } T \text{ passes } H \mid H \text{ is false[즉, } H_0 \text{ is true]})$$
$$= 1 - 0.03 = 0.97 \ [2SD = 0.1]$$

위의 계산에는 약간의 트릭이 숨어 있다. 가설 $H(p>0.5)$가 거짓이라는 것과 귀무가설 $H_0(p=0.5)$이 참이라는 것은 동치가 아니기 때문이다. 이러한 의문은 어떻게 해소될 수 있을까? 우선 우리는 H_0을 H의 유일한 대안가설로 취급할 수 있다. 만약 가설 H가 거짓이어서 X가 맛을 구분하지 못할 경우, 그에 대한 유일한 정량적 해석은 $p=0.5$라는 것이다. 즉 $p<0.5$라는 것은 비현실적이다. 이러한 해석은 충분히 가능한 해석이긴 하지만, 대안가설이 분명히 여럿으로 보이는 시험 상황에서는 적용되기 어렵다.

보다 일반적인 시험 상황에 대비하여, 우리는 귀무가설 H_0를 H의 유일한 대안가설로 취급하는 대신에 H의 대표적 대안가설로 간주할 수도 있다. 즉 가설 H가 거짓이라는 것은 $p=0.5$ 또는 $p=0.4$ 또는 $p=0.3$ 등등을 모두 포함하는 $H_f(p \leq 0.5)$가 되어야 하지만, H_0는 H_f의 대표로 간주될 수 있다는 것이다.

일단 H_f를 정식화하면 다음과 같다.

$$H \text{ is false} \equiv H_f \text{ is true}$$
$$\equiv H_0 \text{ is true} \vee H_1 \text{ is true} \vee H_2 \text{ is true} \vee \cdots \vee H_n \text{ is true}$$

이를 고려하여 $S(T)$를 다시 계산할 경우 이는 다음과 같은 복잡한 수식이 된다.

$$S(T) = P(T \text{ fails } H \mid H \text{ is false})$$
$$= P(T \text{ fails } H \mid H_f \text{ is true})$$
$$= \sum_{k=0}^{n} P(T \text{ fails } H \mid H_k \text{ is true}) \frac{P(H_k \text{ is true})}{P(H_f \text{ is true})}$$

즉 그 값은 $P(T \text{ fails } H \mid H_k \text{ is true})$들의 가중치 평균이다. 그러나 가중치항($P(H_k \text{ is true})/P(H_f \text{ is true})$)에 포함된 H가 거짓일 확률 $P(H_f$

is true) 또는 특정 대안 가설 H_k가 참일 확률 P(H_k is true)의 값은 어떻게 구할 수 있는가? 베이즈주의 입증 이론은 (주관적인) 사전확률 부여를 통해 여기에 어떤 값이든 만들어낼 수는 있으나, 메이요는 그러한 주관적인 사전확률 부여에 의존하지 않고자 한다. 오히려 메이요는 자신의 엄격성 기준이 그러한 의문스러운 사전확률을 필요로 하지 않기 때문에 장점을 가진 것처럼 내세웠다. 그는 가설로부터 어떤 일이 벌어질지에 대한 확률을 구하는 것은 신빙성이 있어도, 자료로부터 가설의 믿음의 정도를 구하는 것은 신빙성이 없다고 생각했다. 그러나 위의 경우처럼 수많은 대안가설이 가능할 경우엔 그가 정의한 엄격성 값을 구하기 위해 각종 가설이 참이거나 거짓일 확률을 알아야 하는 난처함이 발생한다. 그렇다면 0.97이라는 값을 구한 그의 계산은 도대체 무엇일까?

메이요가 허용하는 확률 하에서도, 우리는 특정한 대안가설 H_k가 참이라면 시험절차 T가 시험 가설 H를 탈락시킬 확률 P(T fails H | H_k is true)는 계산할 수 있으며, 우리는 이 확률값을 특정 대안가설 H_k에 비추어 가설 H를 시험하는 절차 T의 엄격성 $S_k(T)$로 간주할 수 있다. 그러면 엄격한 시험이란 시험 가설 H와 양립 불가능한 대안가설 중 어떤 가설 H_k가 참이더라도 시험절차 T가 H를 탈락시킬 확률이 '언제나' 높은 시험이라고 재정의될 수 있으며, 시험의 엄격성은 특정 대안가설에 비춘 시험의 엄격성 $S_k(T)$ 중 최솟값으로 얘기할 수 있다.

$$S(T) = \text{MIN}[S_1(T),\ S_2(T),\ \ldots\ S_k(T)\ \ldots\ ,\ S_n(T)]$$

다행히, 모든 대안가설 H_k를 다 따져볼 수 있는 경우도 있고, 하나만 따져보면 자동적으로 나머지의 엄격성은 그보다 크다는 것을 알 수 있는 경우도 있다. X가 맛을 구분할 수 있다는 가설을 시험하는 위의 사례에서는, 귀무가설로 제시되었던 H_0(p=0.5)가 참일 때 가설 H를 탈락시킬 확률이 다른 나머지 대안가설들이 참일 때 가설 H를 탈락시킬 확률보다 작다. 따라서 우리는 H가 거짓이라면 어떤 대안가설이 참이든 시험절차 T가 가설 H를 탈락시킬 확률이 최소한 0.97이 된다고 말

할 수 있으며, 따라서 위의 시험 절차가 매우 엄격한 시험이라고 말할 수 있게 된다.

이러한 정식화는 메이요 본인의 정식화와 사실상 똑같은 정식화이다. 그는 다음과 같이 말한다. “그것[엄격성 기준]은 엄격성이 **각각의 대안에 대하여** 높을 것을 요구한다. 달리 말해, 이러한 각 대안가설에 대한 최소의 엄격성이 높아야(최대의 오류 확률이 낮아야) 하며, 그것을 계산하는 데 사전확률 부여는 필요하지 않다(강조는 원저자).”[4]

문제는 이러한 방식의 엄격성 계산이 다른 종류의 시험에도 적용될 수 있는가 하는 것이다. 메이요 본인은 모든 종류의 가설에 대해 이러한 방식의 엄격한 시험을 구성할 수 있다고 주장하진 않았다. 다만 그는 최소한 통계적 가설에 대해서는 이러한 엄격한 시험이 잘 적용되며, 아마도 꽤 많은 다른 과학적 가설에서도 엄격성 기준이 올바른 시험 절차와 그렇지 못한 시험 절차를 구분해주는 유용한 기능을 할 수 있을 것으로 기대했다.[5]

사실 메이요는 엄격한 시험을 통해 두 가지를 동시에 추구했다. 첫째, 그는 엄격한 시험 개념을 이용하면 방법론적 이론 미결정성 논변을 반박할 수 있다고 생각했다(Mayo 1996, 6장). 둘째, 그는 가설 구성에 사용된 자료가 다시 증거로도 사용될 수 있는지의 문제에 대해 엄격한 시험 개념이 적절한 기준을 제공할 수 있을 것이라고 생각했다(Mayo 1996, 8장). 그러나 ‘엄격한 시험’ 개념을 통해 두 가지 목표를 함께 달성하는 것이 쉽지 않음이 아래의 논의를 통해 드러날 것이다.

3. 방법론적 이론 미결정성과 사용-참신성 예측주의

메이요는 이론 미결정성 논제의 위협을 ‘엄격한 시험’을 통해 빠져나가려고 한다. 그는 방법론적 미결정성 논제를 시험과 연결지어 다음과 같이 재정의한다.[6]

---

4) Ibid., p. 196.

5) Ibid., pp. 195-196.

방법론적 이론 미결정성(MUD) : 가설 H에 대한 좋은 시험(혹은 좋은 뒷받침)으로 간주되는 모든 증거는 *H*의 일부 경쟁 가설에도 똑같이 좋은 시험(혹은 좋은 뒷받침)으로 간주될 것이다.

이러한 방법론적 이론 미결정성이 옳다면, 어떠한 증거가 특정한 가설을 뒷받침한다는 판단이 불가능하게 된다. 그 가설에 대한 증거가 있다면, 그것을 설명해주는 대안 가설은 하나 이상 존재하기 마련이고, 따라서 그것은 그 대안 가설에게도 동일한 증거를 제공해줄 것이기 때문이다. 메이요는 어떠한 증거에 대해서도 시험 중인 가설 이외의 대안 가설이 하나 이상 존재한다는 논제를 부정하지 않는다. 그러나 그는 그 증거를 설명할 수 있는 가설이 둘 이상이라고 하더라도, 그 가설들 사이에서 증거의 차별화가 가능하다고 생각한다. 두 개의 가설이 동일한 자료를 설명할 수 있다고 해서, 두 가설이 동일한 종류의 시험을 통과하는 것은 아니기 때문이다. 그에 따르면, 어떤 가설은 엄격한 시험을 통과하지만, 다른 가설은 엄격하지 못한 시험만을 통과하는 것이 가능하다. 그는 아래와 같이 주장한다.[7]

1. 증거 *e*를 함축하거나 그와 맞는 *H*의 대안가설의 존재는 *H*가 *e*와 결부된 엄격한 시험을 통과하는 것을 막지 않는다.
2. 시험의 엄격성을 계산하는 것은 가설에 확률을 부여하는 것을 요구하지 않는다.
3. 증거 *e*를 함축하거나 그와 맞는 대안가설이 항상 존재한다는 것을 허용하더라도, *e*에 의해 똑같이 엄격히 시험되는 대안가설이 항상 존재하진 않는다.

이를 뒷받침하기 위해 메이요는 다음과 같은 사례를 제시한다.[8] 어떤 동전이 공정하다(p=0.5)는 가설 H_0을 여러 차례의 동전 던지기 시행을 통해 시험하는 상황을 상상해 보자. 이때 우리는 어떤 시행 결과 *e*를 얻더라도, P[*e*|*G*(*e*)]=1이 성립하는 가설 *G*(*e*)를 아래와 같이 만들

6) Ibid., p. 176.
7) Ibid., p. 175.
8) Ibid., pp. 201-3.

어낼 수 있다.

$G(e)$: 앞면이 나온 각 시행에서 앞면이 나올 확률 p는 1이고, 나머지는 0이다.

어떠한 결과 e를 얻더라도, 가설 $G(e)$로부터 e가 나올 확률은 1이지만, H_0로부터 특정한 e가 나올 확률은 그보다 작다. 가설과 증거 사이의 논리적 관계만 본다면, $G(e)$는 e를 산출할 가능성이 H_0보다 높은 가설이 되고, 따라서 e는 H_0을 뒷받침하는 데 실패하게 될 것이다. 메이요는 '엄격한 시험' 개념을 사용하면 이러한 반직관적인 증거 판단을 피할 수 있다고 주장한다.

메이요에 따르면, 동전 던지기 시행 결과로부터 사후적으로 구성된 가설 $G(e)$는 동전 던지기 시행 결과와 비교하는 시험에서 탈락할 확률이 0이다. 이는 "심지어 $G(e)$가 거짓이고 H_0가 참이더라도(즉, 동전이 '공정'하더라도)" 성립하며, 따라서 이 시험 절차의 엄격성은 0이다. 그는 유리 겔러의 이름을 따서 $G(e)$처럼 자료와는 매우 잘 맞아떨어지지만 전혀 엄격하지 않은 시험 과정을 거친 가설을 '겔러식 가설(gellerized hypothesis)'이라고 불렀다.[9]

결국 메이요는 엄격한 시험이라는 기준을 사용하여 겔러식 대안가설을 제거할 수 있으므로, 방법론적 이론 미결정성의 위협으로부터 벗어날 수 있다고 주장한다. 그에 따르면, 자료를 설명할 수 있는 대안가설이 무궁무진하더라도, 모든 대안가설이 엄격한 시험을 통과하는 것은 아니다. 특히 자료와 완전히 맞아떨어지도록 구성된 가설은 그 자료에 의한 시험에서 탈락될 가능성이 전혀 없으므로, 그러한 가설은 엄격한 시험을 통과한 것이 아니며 그 자료에 의해 어떠한 증거도 얻지 못한다. 그런데 이러한 주장은 기어리나 워럴과 같은 사용-참신성 예측주의자들의 주장을 연상시킨다.

기어리와 워럴 등의 사용-참신성 예측주의에 따르면, 이미 알려진 자료를 설명하기 위해 가설이 제시된 경우 그 자료는 그 가설을 뒷받

9) Ibid., p. 202.

침하는 증거가 될 수 없으며, 가설을 구성하는 데 사용되지 않은 가설의 참신한 귀결(사용-참신한 예측, use-novel prediction)만이 가설의 시험지가 되어 가설을 뒷받침하는 증거를 제공할 수 있다(Giere 1983; Worrall 1989). 기어리는 이러한 사용-참신성 예측주의를 옹호하기 위해 메이요의 '엄격한 시험' 개념과 흡사한 '적절한 시험(Appropriate Test)' 개념을 제시한 바 있다.[10]

적절한 시험 : 참인 가설은 참으로 수용할 뿐 아니라 거짓인 가설은 거짓으로 거부하도록 이끌 적절히 높은 확률을 가진 절차

기어리에 따르면, 과학자가 어떤 자료를 설명하기 위해 가설을 만들었다면(선택했다면), 그렇게 만들어진(혹은 선택된) 가설이 그 자료를 설명하지 못할 가능성은 없다. 즉 그 가설이 참이든 거짓이든 상관없이, 그 가설은 무조건 그 자료를 설명하는 데 성공할 것이다. 따라서 가설을 선택하는 데 사용된 자료와의 일치 여부를 통한 시험은 "거짓인 가설을 거짓으로 거부할" 확률이 0이므로 적절한 시험이 될 수 없다.[11] 다른 방식으로 말하자면, 그 자료는 가설이 참이 아니더라도 가설의 예측과 일치했을 가능성이 매우 높기 때문에, 그 자료는 좋은 증거로 간주될 수 없다. 이를 위해 가설의 시험에 사용되는 가설의 경험적 귀결은 자료와의 일치 여부가 사전에 정해져 있지 않은 것이어야 하며, 따라서 완전히 새로운 예측이거나 적어도 가설을 선택하는 데 사용되지 않은 것, 즉 '사용-참신한 예측'이어야 한다.[12] 워럴 역시 사용-참신한 예측이 증거의 필요조건임을 옹호하기 위해 기어리와 비슷한 '진정한 시험'에 의존한 논변을 제시했다.

---

10) Giere (1983), p.278.

11) Ibid., p. 278.

12) 이와 유사한 논변은 기어리, 비클, 몰딘의 '모형 개발'과 '결정적 실험'의 증거적 차이에 대한 논의에서도 찾아볼 수 있다(기어리, 비클, 몰딘 2008, pp. 95-97.).

> “이론 T가 실제로는 자료 e의 기초 위에 묶여 있었다면 … e에 대한 조사는 T에 대한 진정한 시험(real test)을 구성하지 않는다. … 그것은 T의 잠재적 반증자가 전혀 아니었다. 왜냐하면 T는 그것의 구성 방법 때문에 e에 의해 기술된 사실로부터 전혀 위협을 받지 않았기 때문이다.” (Worrall 1989, pp. 148-9)

사용-참신성 예측주의를 옹호하는 기어리와 워럴의 논변들은 메이요의 엄격한 시험 논변과 매우 유사하지만, 메이요는 가설 구성에 사용된 자료가 무조건 그 가설에 증거를 제공하지 못하는 것은 아니라고 주장한다.[13] 메이요는 오히려 기어리와 워럴이 엄격한 시험 개념을 잘못 적용했다고 생각한다. 예를 들어, 신뢰구간에 대한 가설을 얻는 과정은 주어진 자료로부터 사후적으로 구성되는 과정으로서, 어떤 자료 x가 주어지든 정해진 규칙에 따라 무조건 신뢰구간 가설 $H(x)$를 산출하겠지만, 그럼에도 메이요는 그렇게 얻어진 가설 $H(x)$가 그 자료 x에 의거한 엄격한 시험을 통과할 수 있다고 생각했다.[14]

2014년의 논문에서 메이요는 자신의 논변을 ‘가설 구성 절차’에 초점을 맞추어 다음과 같이 재구성한다(Mayo 2014). 가설을 구성하는 데 자료가 사용된 경우, 그 가설 구성 절차 R의 엄격성은 “주어진 자료 생성 방식에 비추어”, H가 거짓이라면 가설 구성 절차 R이 H를 산출하지 않았을 확률로 정의된다. 수식으로 말하자면 다음과 같다.

$$\begin{aligned} \mathrm{S}(R) &= \mathrm{P}(R \text{ would not output } H \mid H \text{ is false}) \\ &= 1 - \mathrm{P}(R \text{ would output } H \mid H \text{ is false}) \end{aligned}$$

예를 들어, 표본 자료로부터 신뢰구간 가설을 구하는 다음의 절차 R_{CI}는 다음과 같이 정의된다.

R_{CI} : 모집단으로부터 임의 표집된 표본 x에 대해 다음의 가설 $H(x)$

---

13) Mayo (1996), p. 203, 8장.

14) Ibid., pp. 271-4.

를 세운다.

$H(x)$: f-2SD$\leq p\leq f$+2SD (단, p는 모집단의 비율, f는 표본 빈도, SD는 표준 편차)

메이요에 따르면, 이러한 가설 구성 절차 R_{CI}는 엄격한 시험을 제공한다. 이 가설 구성 절차 R_{CI}는 $H(x)$가 거짓이라면, 즉 모집단의 비율이 그 구간 내에 포함되어 있지 않았더라면, $H(x)$와 같은 가설을 산출하기 매우 어려웠을 것이다. 왜냐하면 임의 표집이라는 자료 생성 방식에 비추어볼 때, 모집단의 비율로부터 그렇게 많이 벗어난 표본 빈도를 가진 표본이 만들어지기는 매우 어렵기 때문이다. 메이요는 다음과 같이 추론한다(Mayo 2014).

$P(f\text{-}2SD\leq p\leq f+2SD) = 0.95$

$P(R \text{ outputs } H(x) \mid H(x) \text{ is false}) = 0.05$

특정한 시행 결과에 대해 따질 경우에는 다음과 같이 계산한다.

$P(R \text{ would output } H(x_0) \mid H(x_0) \text{ is false}) \leq 0.05$

그에 따르면, 이러한 결과는 신뢰구간 가설을 산출하는 절차 R_{CI}가 0.95 이상의 높은 엄격성을 가진 시험이 될 수 있다는 것을 말해준다. 이를 통해 메이요는 자료를 이용하여 가설을 구성하더라도 무조건 그 자료가 가설에 증거를 제공하지 못하는 것은 아니며, '엄격한 시험'이 그 둘을 구분하는 기준이 될 수 있다고 주장한다. 동전 던지기 시행 결과로부터 겔러식 가설을 구성하는 절차와 표본 자료로부터 모집단의 비율에 대한 신뢰구간 가설을 구성하는 절차를 비교한 메이요의 분석 결과는 메이요의 주장을 지지해주는 것 같지만, 여기에는 다음과 같은 문제들이 숨어 있다. 첫째, 메이요는 두 사례에 대해 일관된 방식으로 엄격성을 판단하지 않았다. 둘째, $G(e)$와 같은 겔러식 가설을 대안가설로 가정할 경우, 앞서의 커피 맛 구분 가설에 대한 시험도 엄격하지

못한 시험이 될 수 있다. 셋째, 엄밀하게 계산할 경우 신뢰구간 가설 구성 절차의 엄격성 값은 0.95가 아니다. 다음 절부터는 이에 대해 차례대로 살펴보도록 할 것이다.

4. 엄격성 판정의 비일관성

메이요에 따르면, 표본 자료 x로부터 구성된 신뢰구간 가설 $H(x)$는 0.95의 엄격성을 가진 시험을 통과한다. 그러나 $H(x)$가 거짓일 경우 신뢰구간 가설 구성 절차 R_{CI}가 $H(x)$를 산출할(또는 산출했을) 확률이 0.05(또는 그 이하)라는 계산은 어떻게 가능한가? 우선 표본 자료 x가 고정될 경우, $H(x)$가 거짓이더라도 가설 구성 절차 R_{CI}는 무조건 $H(x)$를 산출하게 되어 있다. 그렇다면 그 확률은 1이라고 해야 하지 않는가? 이러한 계산 결과를 수용할 경우, 표본 자료를 이용한 신뢰구간 가설의 시험 절차는 전혀 엄격하지 않은 시험이 되며, 결국 신뢰구간 가설은 표본 자료로부터 아무런 증거도 얻지 못한다는 결론이 도출된다.

이러한 결론을 피하려면, 우리는 자료가 아니라 **자료 생성 방식**이 고정된 것으로 보아야 한다. 즉 $H(x)$가 거짓인데도 주어진 자료 생성 방식에 의해 $H(x)$로 추론되는 자료가 만들어질 확률을 따져야 하는 것이다. 이는 “만약 똑같은 임의 표집 방식으로 새로 표본을 구해본다면”과 같은 상상을 요구한다. 물론 이번의 표본 자료 x_0에 대해 $H(x_0)$를 산출했듯이, 새로 표본 자료를 구하더라도 새 표본 자료 x_1에 대해 $H(x_1)$을 산출할 것이다. 규칙 R_{CI}는 어떠한 표본 자료 x_k가 나오든 그에 맞춰 무조건 $H(x_k)$를 만들어낼 것이다. 따라서 그 $H(x_k)$가 거짓임에도 규칙 R_{CI}가 $H(x_k)$를 산출할 확률은 역시 1인 것처럼 보인다. 그렇다면 도대체 R_{CI}는 어떻게 $H(x)$를 산출할 확률이 1보다 작을 수 있는가? 그 계산은 바로 이번에 생성된 특정한 표본 x_0를 통해 만들어진 신뢰구간 $H(x_0)$가 만약 거짓이라면 새로운 표본을 구할 경우 새 표본 x_1은 $H(x_0)$를 산출하지 않을 가능성이 높다는 임의 표집 방식에 대한 수학

적 분석에 의존한다.[15]

그러나 엄격한 시험을 신뢰구간 사례에 적용한 위의 방식을 동전 던지기 사례에 일관되게 적용하면, 앞서 겔러식 가설로 분석된 $G(e)$ 역시도 엄격한 시험을 통과할 수 있다. 동전 던지기의 시행 결과로부터 3절의 겔러식 가설 $G(e)$를 구성하는 가설 구성 절차를 R_G라고 해보자. 다음과 같은 특정한 자료 e_0=<h,t,t,h>를 얻은 경우, 가설 $G(e_0)$는 첫 번째와 네 번째는 앞면이 나올 확률이 1이었고, 두 번째와 세 번째는 앞면이 나올 확률이 0이었다고 말한다. $G(e_0)$가 거짓이고 만약 동전이 앞면이 나올 확률이 매번 0.5라는 대안가설 H_0가 참이라면, 똑같은 자료 생성 방식에 의해 자료를 새로 구할 경우, 새롭게 구한 자료 e_1은 $G(e_0)$를 산출하지 않을 가능성이 15/16로 매우 높다.[16] 그렇다면 R_G도 엄격한 가설 구성 절차로 보아야 하는 것이 아닌가? 만약 R_G가 어떠한 자료 e_k에 대해서도 무조건 $G(e_k)$를 산출한다는 이유로 R_G의 엄격성이 0이라고 말한다면, 신뢰구간 가설 구성 절차인 R_{CI} 역시 똑같은 처지에 놓이게 된다. 반대로 신뢰구간 가설 구성 절차 R_{CI}가 엄격한 시험을 구성한다면, 동전 던지기 결과로부터 가설을 구성하는 절차 R_G 역시 엄격한 시험을 구성하게 되며, 이를 수용하면 엄격한 시험을 통해 간편하게 방법론적 이론 미결정성을 극복할 수 있다는 메이요의 논변은 다소 힘을 잃게 된다.

두 개의 가설 구성 절차 R_G와 R_{CI}가 무언가 다르다면, 그것은 무엇에서 비롯된 것일까? 한 가지 가능성은 $G(e_0)$가 동전만에 대한 가설이 아니라 동전과 시점을 모두 포함한 가설이라는 데서 찾을 수 있다. 새로운 자료 생성을 통해 $G(e_0)$를 산출하지 않는 e_1이 새로 나온다고 해도, 그것은 새 시점에 대한 가설 $G(e_1)$을 산출할 뿐, $G(e_0)$에 아무런 영향을 주지 않는다. 즉 동전 가설 $G(e_1)$과 $G(e_0)$는 서로 다르더라도 양립가능하다. 반면 신뢰구간 가설 $H(e_0)$와 $H(e_1)$은 서로 다를 경우 양

15) 3절에서 메이요는 가설 구성 절차 또는 가설 시험 절차의 엄격성을 “주어진 자료 생성 방식에 비추어” 계산되는 것으로 정의했기 때문에, 이러한 해석은 메이요가 제안한 방식을 그대로 따른 것이다.

16) 이는 메이요에 대한 Iseda (1999)의 비판과 동일한 주장을 담고 있다.

립불가능하다. 이러한 차이가 두 가설 구성 절차에 대한 차이를 가져왔을 가능성이 있다. 그러나 이런 해석 방식에서 $G(e_0)$는 동전 던지기 결과 앞면이 나올 확률이 항상 0.5라는 가설 H_0과도 양립가능할 수 있다. 특정 시점을 고려하지 않았을 때에는 앞면이 나올 확률이 0.5이면서도, 특정 시점의 수많은 변수를 고려하면 그 확률이 0이나 1로 확정될 수 있기 때문이다. 그런데 만약 두 가설이 양립가능하다는 해석을 받아들일 경우, H_0를 대안가설로 가정하여 R_G를 엄격하지 않은 겔러화된 시험의 사례로 제시한 메이요의 분석은 의미를 잃게 된다. 이러한 해석 하에서 R_G의 엄격성을 따지기 위해서는 다른 종류의 대안가설을 가정해야 할 텐데 도대체 어떤 대안가설이 가능하겠는가?

5. 겔러식 대안가설에 의한 증거의 무력화

메이요에 따르면, 시험의 엄격성에 대한 판정은 항상 특정한 대안가설에 비추어 이루어진다. 그런데 메이요가 언급한 겔러식 가설을 대안가설로 고려하면 시험의 엄격성은 어떻게 판정될까? 커피 맛을 구분하는 능력에 대한 시험은 p=0.5라는 대표 대안가설에 비추어 엄격하게 시험될 수 있다고 앞서 분석한 바 있다. 그러나 매 시행마다 확률이 달라지는 $G(e)$를 대표 대안가설로 채택하여 분석할 경우엔 다음과 같은 이상한 결과가 도출된다.

e = <*s,f,s,f,s,s,s,f,s,f,s,s,s,f,f,f,s,s,s,*⋯,*s*> (단, 100번 중 success의 횟수 70번)

H : 매 시행에서의 성공 확률 p>0.5

$G(e)$: p equals 1 on those trials that were successes, 0 on the others.

S(T) = P(test T fails $H|G(e)$ is true) = 1 – P(test T passes $H|G(e)$ is true) = 0

이러한 결과는 다음과 같이 정리된다. 첫째, 성공 횟수가 60번이 넘는 자료 e에 대해서, H 외에 그것을 설명할 수 있는 대안가설이 존재한다. 둘째, 대안가설 $G(e)$가 참일 경우 시험절차 T가 가설 H를 통과시킬 확률이 1이고, 시험의 엄격성은 개별 대안가설에 비춘 개별 엄격성의 최솟값으로 정의했으므로, 위의 시험 절차의 엄격성은 0이 된다.

이러한 결론을 방지하는 방법은 $G(e)$와 같은 대안가설을 원천적으로 배제하는 것이다. 첫 번째 선택지는 $G(e)$가 엄격한 시험을 통과할 수 없기 때문에 대안가설로서 고려되어서는 안 된다고 말하는 것이다. 엄격한 시험을 통해 방법론적 이론 미결정성의 위협을 벗어날 수 있다고 주장할 때, 메이요는 바로 이 선택지를 암묵적으로 받아들인 것으로 보인다. 그러나 엄격하게 시험될 수 없는 가설이라고 해서 대안가설로 배제될 이유는 '엄격한 시험' 개념 자체에서 찾아볼 수 없다. 더구나 4절의 분석은 $G(e)$도 엄격한 시험을 통과할 수 있음을 보여주고 있다.

두 번째 선택지는 $G(e)$가 애초에 비현실적이므로 대안가설로서 고려되어서는 안 된다고 말하는 것이다. 그러나 어떠한 가설이 비현실적이기 때문에 배제되어야 한다는 판단은 엄격한 시험이라는 기준을 통해 이루어질 수 없다. 만약 이를 받아들이게 되면, 증거 판단을 위해 엄격한 시험 이외의 기준도 필요하다는 것을 수용해야 하는데, 이는 메이요가 수용하기 어려울 것이다.

마지막 선택지는 $G(e)$가 H와 양립 불가능한 가설이 아니므로 대안가설로서 배제해야 한다고 말하는 것이다. 이는 4절에서의 분석에 의해 어느 정도 타당성이 있다.

각 선택지는 각각의 장단점이 있다. 그러나 어떤 선택지를 따르든, 우리는 표준적인 통계적 시험 절차조차 $G(e)$와 같은 종류의 대안가설을 배제함으로써만 엄격한 시험으로 간주될 수 있다는 것에 주목할 필요가 있다. 이는 우리에게 중요한 시사점을 준다. '엄격한 시험'을 이용한 증거 개념에서는, 통상적인 관점에서는 가설에 대한 좋은 증거로 간주될 수 있는 자료가 그에 대한 겔러식 대안가설에 비추어 보는 순간 더 이상 좋은 증거로 간주될 수 없게 될 수 있다는 것이다. 예를

들어, 다윈주의 진화론을 뒷받침한다고 여겨지는 상당수의 자료들이 그에 대한 겔러식 대안가설인 세련된 버전의 개별 창조 가설에 의해서도 설명될 수 있다고 가정해 보자. 엄격한 시험 개념에 따르면, 그 자료들은 더 이상 다윈주의 진화론의 증거가 될 수 없을 것이다. 엄격한 시험 개념은 겔러식 대안가설에 의해 증거가 무력화되는 것을 어떻게 막을 수 있겠는가?

메이요의 접근 방식의 약점은 다음의 사례를 통해 극명하게 드러난다. 2015년, 출처를 알 수 없는 활자인 '증도가자(證道歌字)'에 남아 있는 먹에 대해 탄소 연대 측정법을 사용하여 그 연대를 추정한 결과, 그 활자가 고려시대의 활자임을 입증됐다는 연구 결과가 보도되었다(경향신문 2015.2.8; 한겨레 2015.2.11). 메이요의 방식으로 탄소 연대 측정 자료 *e*가 "증도가자가 고려시대의 활자"라는 가설 *H*를 뒷받침하는지를 판단하기 위해서는 다음과 같은 대안 가설을 생각해 보아야 한다.

H′ : 교묘하게 조작된 활자에 고려시대 먹을 묻혔다.

이러한 가능성은 신문 기사에도 지적이 되어 있었지만, 고고학 분야의 비전문가인 나는 그러한 비판이 과도하다는 생각이 들었다. 조작된 활자에 고려시대 먹을 묻히는 조작을 하려면 일단 고려시대의 먹이 현존해야 한다. 그러나 인터넷을 이용한 검색 결과 현존하는 '고려 먹'은 찾을 수 없었고, 따라서 대안가설 *H*′은 현실성이 없어 보였다. *H*′의 가능성을 기각하고 나니 탄소연대 측정 자료를 설명할 수 있는 다른 대안가설은 없어 보였다. 따라서 그 자료는 조사 중인 증도가자가 고려시대의 진품이라는 가설에 좋은 증거를 제공한다고 판단하는 것이 적절해 보였다.

그러나 몇 달 후 그 활자들이 국과수의 연구에 의해 조작임이 밝혀졌다는 기사가 나왔다(동아일보 2015.10.27). 그 기사에는 "수백 년 된 먹을 중국이나 국내에서도 손쉽게 구할 수 있다"라는 지적이 함께 적혀 있었다. 나는 고려시대의 먹임을 보여주는 탄소 연대 측정 자료가

나오기 위해서는 정말로 '고려의 먹'을 묻혀야 한다고 생각했지만, 동일한 측정 결과가 나오기 위해서는 고려시대와 같은 시기의 중국 먹이어도 되는 것이었다. 고고학 전문가들은 그런 먹을 쉽게 구할 수 있다는 것을 알고 있었기에 탄소 연대 측정 결과가 좋은 증거를 제공할 수 없다는 판단을 처음부터 하고 있었던 반면, 그러한 상식이 없었던 나는 그 자료의 증거력을 과대평가하는 우를 범한 것이다.

이 사례가 시사하는 바는 무엇인가? 나의 증거 판단은 가설 H의 대안가설 H'이 참일 경우 H가 e에 의한 시험을 통과할 확률에만 의존한 것이 아니었다. 그것에만 의존했다면, 탄소 연대 측정 결과 e을 이용한 시험은 H가 거짓이고 H'이 참이더라도 H를 통과시킬 확률이 1이므로, e는 가설 H에 대해 아무런 증거도 제공할 수 없었다. 그러나 나는 H'의 현실성도 고려했다. 즉 $P(H')$을 고려한 것이다. 그러한 고려를 통해 $P(e|H') \fallingdotseq 1$인 대안가설 H'이 하나 있다는 것만으로 증거를 무로 돌리는 것을 막고자 했던 것이다. 다만 나의 잘못은 $P(H')$의 값을 과소평가했다는 데 있었으며, 반대로 전문가들의 이점은 바로 $P(H')$의 값을 적절하게 판단할 수 있었다는 데 있었다.

이러한 문제점은 귀납적 추론의 정당성을 부정하는 입장이 가지고 있는 일반적인 약점과 일맥상통한다. 그 대표에 해당하는 반증주의의 경우, $H_1\&H_2\&H_3$의 가설로 이루어진 이론이 반증될 경우 그중 어떤 가설이 반증된 것인지 판단할 근거가 없다. 만약 H_1이 문제라는 주장을 하기 위해서는 나머지 가설 H_2와 H_3은 참이거나 참일 확률이 높다는 것을 전제할 필요가 있다. 그러나 귀납 추론의 사용을 거부하는 일관된 반증주의자라면 그러한 판단을 내릴 수 없다. 그러나 이러한 판단은 과학 활동에서 너무나 일반적으로 활용되며, 예측 과정에서 부정하기 어려운 단단한 가정을 사용하는 것은 과학의 중요한 덕목으로 인정받고 있다. 반증주의의 정신을 계승한 메이요는 가설에 확률을 부여하는 것 자체를 거부함으로써, 증거 판단에서 사용할 수 있는 손쉬운 방법 하나를 포기한 것이다.

6. 신뢰구간 가설의 엄격성과 신뢰도

메이요는 [*f*±2SD] 신뢰구간 가설의 엄격성을 손쉽게 0.95로 계산했다. 그러나 0.95는 그 가설의 신뢰도일 뿐, 그 가설이 거짓일 경우 시험이 그 가설을 탈락시킬 확률인 엄격성 값 P(test *T* fails *H* | *H* is false)와는 아무런 관련이 없다. 먼저 신뢰도가 어떻게 구해질 수 있는지부터 살펴보자.

모집단의 비율 *p*가 주어질 때, 임의 표집에 의해 생성되는 표본의 특성에 관한 수학적 분석으로부터 다음이 곧장 도출된다.

$$P(p\text{-}2SD \leq f \leq p\text{+}2SD) = 0.95$$

이는 모집단의 비율 *p*가 주어졌을 때, 일정한 크기의 표본을 반복해서 생성할 때 각 표본 빈도 *f*가 어떤 분포로 나타날지를 수학적으로 분석함으로써 나온 결과이다. 그리고 이로부터 우리는 모집단의 비율 *p*가 주어졌을 때, 그 모집단으로부터 임의 표집된 표본을 반복해서 생성할 때 그 실제 비율 *p*가 표본 빈도 [*f*±2SD] 사이에 있을 확률을 아래와 같이 구할 수 있다.

$$P(f\text{-}2SD \leq p \leq f\text{+}2SD) = 0.95$$

이 값이 바로 신뢰도이다. 이 신뢰도란 특정한 비율 *p*가 주어진 고정된 모집단에서 무한히 반복적으로 임의 표집을 할 때, 그 *p*가 [*f*±2SD]의 구간 내에 포함되는 빈도를 뜻한다. 이 신뢰도 공식은 "모집단의 비율 *p*가 신뢰구간 [*f*±2SD] 내에 있다"는 가설이 참일 확률처럼 표현되어 있긴 하지만, 신뢰도는 그 가설이 참일 확률을 나타내지 않는다는 데 주의해야 한다.[17] 그럼에도 우리는 이 신뢰도가 메이요가 말하는 시험의 엄격성과 같은 것도 아니란 점에 주의해야 한다.

---

17) 5절에서 언급했듯이, 메이요와 같은 빈도주의자들은 가설이 참일 확률을 전혀 도입하지 않으며, '신뢰도'는 빈도주의 통계학에서 사용되는 개념이다.

그렇다면 이 시험의 엄격성은 어떻게 계산될 수 있을까? 시험의 엄격성을 구하는 데는 사전 확률이나 가설에 대한 확률을 구할 필요가 없고, 대표 대안가설만 고려하면 된다. 그럼에도 그 계산은 생각만큼 쉽게 이루어지지 않는다. 신뢰구간 가설의 시험 절차에 대해서는 여러 해석이 가능하나, 그중 가장 그럴듯한 해석만 제시하자면 신뢰구간 가설의 시험은 다음의 두 가지 시험의 합성으로 생각할 수 있다.

시험 가설 H_1 : $p \geq f_e - 2SD$ [귀무 가설 H_0 : $p \fallingdotseq f_e - 2SD$]
시험 절차 T_1 : 표본빈도 f가 f_e 이상이면, 가설 H_1을 통과시킨다.
$S(T_1)$ = 1 - P(test T_1 passes $H_1|H_1$ is false [즉, H_0 is true])
= 1-0.025 = 0.975

시험 가설 H_2 : $p \leq f_e+2SD$ [귀무 가설 H_0' : $p \fallingdotseq f_e+2SD$]
시험 절차 T_2 : 표본빈도 f가 f_e 이하이면, 가설 H_2를 통과시킨다.
$S(T_2)$ = 1 - P(test T_2 passes $H_2|H_2$ is false [즉, H_0' is true])
= 1-0.025 = 0.975

각각의 시험 절차는 커피 맛 구분 능력에 대한 시험과 매우 유사하며 따라서 매우 전형적인 엄격한 시험으로 보인다. 그러나 실제로 신뢰구간 가설은 H_1&H_2이기 때문에, 두 시험을 **모두** 통과해야만 한다. 이 경우 가설이 거짓임에도 시험을 통과할 확률은 P($f=f_e$ | H_0 is true or H_0' is true)로 계산될 수 있고, 그 값은 0.05가 아니며 0.025보다도 훨씬 작다(표본의 크기가 100일 경우, 이 값은 대략 0.011이며, 표본의 크기가 커질수록 더욱 작아진다). 그렇다면 이 시험은 0.99…의 아주 높은 엄격성을 가진 시험이라는 뜻이 되는 것일까?

더 이상한 문제는 가설이 참일 때에도 시험을 통과할 확률이 무척 작다는 데 있다. 심지어 모집단의 비율이 정확히 $p=f_e$일 때조차 그것의 표본 빈도가 $f=f_e$로 나올 확률은 매우 작다(표본의 크기가 100일 경우 그 값은 대략 0.080이며, 표본의 크기가 커질수록 더욱 작아진다). 물론 그 값은 P($f=f_e$ | H_0 is true or H_0' is true)보다는 클 것이다. 그렇

다면 그저 그 점에 만족해야 할까?

이러한 상황은 세 가지 점에서 만족스럽지 못하다. 첫째, 신뢰구간 가설 구성 절차가 평범한 커피맛 구분 능력 시험 절차보다 훨씬 엄격하다는 결론은 상식적으로 받아들일 수 없다. 두 시험은 거의 비슷한 시험처럼 보이기 때문이다. 둘째, 가설이 참일 경우에도 시험을 통과할 확률이 작지만 가설이 거짓일 경우 시험을 통과할 확률보다는 크다는 점에 만족하려면 왜 그것을 받아들여야 하는지에 대한 근거가 필요한데, P(test T fails H | H is false)의 값만을 사용하는 메이요의 엄격한 시험 개념은 그런 근거를 제공하지 않는다. 셋째, 시험의 엄격성 개념은 신뢰도 개념에 의존하는 것보다 신뢰구간 가설 구성 과정의 신뢰성에 대한 이해를 높여주지 않는다. 메이요의 엄격성 개념을 일관되게 적용할 경우, 그 값은 동일한 신뢰도에도 표본의 크기에 따라 달라지게 된다.

그렇다면 혹시 신뢰구간 가설이 표본 자료로부터 증거를 얻는 것은 아예 시험과 관련이 없는 것은 아닐까? 그러나 메이요는 모든 증거를 시험과 관련짓고 있으며, 신뢰구간 추정 방법이 전형적인 통계적 시험과 수학적으로 호환될 수 있다는 점을 강조한다. 표본 자료로부터 추정된 신뢰구간 내의 값들은 바로 통계적 시험을 통과했을 값들로 이루어져 있다는 것이다.[18] 그러나 이를 시험으로 해석하더라도, 어떤 가설이 그 시험을 통과하느냐는 메이요가 제시한 절대적인 엄격성 값에 의해 걸러지지 못하며 가설과 대안가설이 각각 자료를 산출할 확률값의 상대적인 비교에 의존한다.[19]

반면 베이즈주의자는 시험이라는 개념에 의존하지 않고도 신뢰구간 가설을 정당화할 수 있을 것이다. 무차별의 원리에 근거하여 모집단의 비율에 대한 사전 확률 분포를 가정하면 95% 신뢰구간 가설이 참일 확률인 신임도(credence)가 신뢰도(confidence level) 0.95와 비슷한 값

18) Mayo (1996), pp. 273-4.

19) 이러한 해결책은 각 가설들에 의한 가능도(likelihood)를 비교하는 헬만(Hellman 1997)의 베이즈주의적 '엄격한 시험' 개념과 연결된다. 이영의(2004)를 참고할 것.

으로 계산되기 때문이다. 만약 가설이 참일 확률을 직접 알 수 있다면, 굳이 가설의 추정 과정을 복잡하게 시험으로 해석할 필요가 어디에 있겠는가?

7. 결론

메이요는 엄격성 개념이 방법론적 이론 미결정성의 위협을 막아내고 적절한 증거 판단을 가능하게 해줄 것을 기대했다. 그러나 지금까지의 검토는 그러한 기대가 충족되지 못함을 보여준다. 엄격한 시험 개념을 일관되게 적용할 경우 방법론적 이론 미결정성의 위협은 간편하게 해결될 수 없으며, 비현실적인 대안가설에 의한 증거의 무력화 시도를 방어하기 어렵다. 또한 가장 기본적이라고 할 수 있는 신뢰구간 가설의 평가에 대해서도 엄격성 개념은 깔끔한 정량적인 분석을 보여주지 못하며, 평범한 신뢰도 개념보다 나은 분석을 보여주지 못하고 있다.

엄격한 시험 개념을 유지하면서 위의 문제에 대응할 수 있을까? 겔러식 대안가설에 의해 엄격한 시험이 무력화되지 않기 위해, 즉 겔러식 대안가설을 상상할 때마다 시험의 엄격성이 0으로 판정되는 것을 막기 위해, 가설에 대한 엄격한 시험은 특정한 집합의 대안가설에 비추어서만 정의될 수 있는 것으로 제한해야 할 것이다. 즉 어떤 가설에 대한 전면적인 엄격한 시험은 존재하지 않는다. 다만 어떤 가설에 대해 어떤 시험은 특정한 집합의 대안가설에 비추어서만 엄격한 시험을 제공할 뿐이다. 이 경우 증거 판단 역시 항상 특정한 집합의 대안가설에 비추어서만 이루어지게 될 것이다. 또한 표준적인 신뢰구간 추정 절차를 일정한 엄격성 값을 가진 시험으로 분석하기 위해, 엄격성 평가는 대안가설이 참일 경우 가설이 시험을 통과할 확률과 가설이 참일 경우 가설이 시험을 통과할 확률의 상대적 비교를 반영해야 한다.

참고문헌

경향신문 (2015), ‘고려 ‘증도가자’ 현존 세계최고 금속활자’ (2015.02.08).

기어리, 로널드 N., 비클, 존, 몰딘, 로버트 F. (2008),『과학적 추론의 이해』제5판, 조인래, 이영의, 남현 옮김. 서울: 소화.

동아일보 (2015), ‘[단독]“증도가자는 가짜... 最古활자 아니다’ (2015.10.27.).

이영의 (2004),「헬만과 메이요의 엄격한 시험 개념」,『과학철학』7권 1호, pp. 109-128.

한겨레 (2015), ‘‘증도가자’가 세계 최고 금속활자?…문화재 학계 진위 논란 재가열’ (2015.2.11.).

Giere, R. N. (1983), “Testing Theoretical Hypotheses”, in Earman, J. (ed.) (1983), *Testing Scientific Theories, Minnesota Studies in the Philosophy of Science*, Vol. X, Minneapolis: University of Minnesota Press: pp. 269-298.

Hellman, G. (1997), “Bayes and Beyond”, *Philosophy of Science* 64: pp. 191-221.

Iseda, T. (1999), “Use-Novelty, Severity and a Systematic Neglect of Relevant Alternatives”, *Philosophy of Science* 66: pp. 403-413.

Mayo, D. G. (1996), *Error and the Growth of Experimental Knowledge*, Chicago and London: The University of Chicago Press.

_____ (2014), “Some Surprising Facts about (the Problem of) Surprising Facts (from the Dusseldorf Conference, February 2011)”, *Studies in History and Philosophy of Science* 45: pp. 79-86.

Worrall, J. (1989), “Fresnel, Poisson and the White Spot: The Role of Successful Predictions in the Acceptance of Scientific Theories”, in Gooding, D., Pinch, T. and Schaffer, S.(eds.), *The Uses of*

Experiment, Cambridge: Cambridge University Press: pp. 135-157.

논문 투고일	2017. 06. 30.
심사 완료일	2017. 07. 19.
게재 확정일	2017. 11. 16.

Critical Analysis of Mayo's Severe Test

Dongwook Jung

I examine Deborah G. Mayo's 'severe test' concept. She argues that 'severe test' overcomes the methodological underdetermination and provides the criterion and quantitative analysis of evidence. But my examination shows that her 'severe test', if applied consistently, cannot protect evidence from the methodological underdetermination and provide uniform quantitative analysis of evidence, even in cases of evaluating simple statistical hypotheses by sample data.

Keywords: severe test, evidence, Deborah G. Mayo, use-novel predictivism, confidence interval, methodological underdetermination