가설검정의 원리 II

이 형 기

종근당 의학부 이사

지난 번 호에서는 가설검정의 원리에 대한 앞 부분으로, 귀무가설과 대립가설의 의미, 가설검정에서 발생할 수 있는 두 가지 오류의 종류, p값, 그리고 가설검정의 원칙 등에 대해서 소개하였다. 이번에는 가설검정의 원리에 대한 뒷 부분으로서 가설검정의 오용에 대해 주로 알아보자.

1. 가설검정의 오용에는 어떤 것들이 있는가?

오늘날 통계학적 방법론을 이용하는 제 학문 분야에서, 가설검정처럼 흔히 인용되는 방법도 드물 것이다. 어떤 분야의 전문지(專門誌, specialized journal)를 펼치더라도 '$p<0.05$' 등과 같은 표현을 자주 접할 수 있으며, 심지어는 통계학적으로 의미있는 결과를 얻는 것이 모든 연구의 목적인 것처럼 간주되기도 한다. 일반적으로 검정에는 '가설검정'과 '유의성검정(有意性檢定, significance testing)'의 두 가지가 있으며, 전자는 의사결정(decision making)의 보조적 수단으로 사용될 경우가 많다. 그러나 가설검정이든지 유의성검정이든지 간에 귀무가설과 대립가설을 세우고 유의수준을 결정하여, 이를 p값과 비교한 후 통계적 유의성 여부를 판단한다는 기본 원리는 동일하다. 실제로 가설검정과 유의성검정이 명백히 구분될 수 있는 것은 아니며, 학자들 사이에도 이 둘을 다른 것으로 보아야 하는지에 대해서는 이론(異論)이 분분하다. 그러나 가설검정이 되었든지 아니면 유의성검정이 되었든지 간에 통계적 자료분석방법을 사용하는 연구자들 사이에 검정의 의미에 대한 올바른 이해가 선행되지 않은 채 이를 오용(誤用, misuse)하는 경우가 상당수 있음을 발견하게 된다. 본 절의 목적은 바로 이러한(가설) 검정의 오용에는 어떠한 것들이 있으며, 어떻게 이들을 예방할 수 있는지를 논의하려는 것이다.

지침 1 : 가설검정으로부터 통계적 유의성이 있는 결과가 얻어졌다고 해서 이것이 곧 실제적으로도 의미가 있다는 것을 의미하지는 않는다. 요컨대 실제적 유의성(practical significance)이란 일종의 관점의 문제로서 이는 대개 자료수집 이전에 결정되는 것이 보통이다.

지난 번 전구공장과 동일한 전구수명(단위 : 시간)의 모집단으로부터 이번에는 1,705라는 표본평균을 얻었다고 가정해 보자. 우리가 증명하고자 하는 대립가설이 $H_1 : \mu > 1{,}700$와 같은 단측가설로 주어졌다고 하면, 이 경우 1,705라는 표본평균은 이전에 예로 든 1,756이나 1,736 보다는 대립가설을 지지하는 정도가 확실히 약하다. 이제 추론통계학적 표본표준편차가 S=200이며, 1,705라는 표본평균이 크기가 각각 n=10, 30, 100, 1000, 10000이라는 표본으로부터 얻어졌다고 가정하고 각 경우에 대해 유의수준 0.05에서 $H0 : \mu \leq 1{,}700$이라는 귀무가설을 기각할 수 있는지를 검정하면 표 1을 얻을 수

표 1. 표본의 크기에 따른 귀무가설의 기각 여부

표본의 크기 (=n)	검정통계량의 값[a]	귀무가설의 기각여부[b]
10	0.079	기각하지 못함
30	0.139	기각하지 못함
100	0.25	기각하지 못함
1000	0.791	기각하지 못함
10000	2.5	기각할 수 있음

a : 검정통계량은 n=10과 30인 경우에는 T로, 그 이상은 Z로 구함

b : 기각역 n=10일 때 $T \geq 1.812$

n=30일 때 $T \geq 1.699$

n=100, 1000, 10000일 때 $Z \geq 1.64$

있다.

위의 예에서 표본평균 1,705는 우리가 증명하고자 하는 대립가설의 하한값인 1,700에 비해 단지 "5" 시간이 더 많으므로 표본추출변동을 고려하면 '실제적'인 의미는 없을 수도 있다. 즉, 모집단에서의 산술평균은 실제로 1,700 시간을 넘지 않는데도 우연히 표본평균이 1,700 시간을 넘는 것으로 나왔다는 것이다. 이러한 우연의 정도를 계량화하기 위해 우리는 이미 P값이라는 개념을 도입한 바 있다. 표1로부터 표본의 크기가 1,000인 경우까지는 $p>0.05$로서 유의수준 0.05에서 귀무가설을 기각할 수 없으나, 표본의 크기가 10,000이 되면 $p<0.05$로 귀무가설이 기각됨을 알 수 있다. 이러한 사실은 비록 실제적인 유의성이 없을지라도 표본의 크기가 커지기만 하면 얼마든지 통계적 유의성이 있는 것으로 검정결과가 나타날 수 있음을 말해 준다. 즉, 표본이 어느 수준 이상으로 크기만 하면 아무리 작은 실제적 유의성도 통계적 유의성이 있는 것으로 검정해 낼 수 있으며, 이는 곧 검정력이 커짐을 의미한다. 요컨대 검정력은 표본의 크기에 대한 함수가 되는 것이다.

이러한 실제적 유의성의 문제는, 의학 등에서 새로 개발된 약이 기존의 약물보다 우수하다는 것을 증명하기 위해 대(大)표본으로부터 얻어진 결과에 대해 통계적 검정을 실시하고 이로부터 통계적 유의성을 얻었기 때문에 실제적 유의성 또한 보장받을 수 있다고 주장하는 경우에 흔히 발생한다. 다음의 예를 통해 통계적 유의성과 실제적 유의성이 엄연히 차원을 달리하는 것임을 다시 한 번 확인하기 바란다.

예 1) 기존에 사용되던 A 약물에 비해 새로 개발된 B 약물이 더 우수함을 입증하고자 B 약물 제조회사의 상품홍보부에서는 S 의과대학에 비교실험을 의뢰하였다. S 의과대학 연구팀에서는 20,000 명의 환자 중 10,000 명씩을 각각 무작위로 A, B 두 약물 치료군에 배치하고 사용 후의 임상적 효과를 측정하였다. 그 결과 A 약물에는 6,250명이 호전을 보여 호전율은 62.5%(6,250/10,000)를 나타냈으며, B 약물에는 6,430 명이 호전을 보여 호전율 64.3%를 나타냈다. A와 B 약물에 의한 호전율을 각각 pA, pB라고 한다면 이 경우 귀무가설과 대립가설은 다음과 같다. 즉,

$$H_0 : pA=pB$$
$$H_1 : pA<pB$$

이처럼 모집단에서 추출한 표본의 표본비율을 이용하여 모비율에 대한 통계적 가설을 검정하는 것을 비율검정이라고 한다. 이러한 비율검정의 방법을 이용하면 유의수준 0.05에서 귀무가설을 기각할 수 있다. 즉 B 약물에 의한 호전율은 A 약물에 의한 호전율보다 크다는 대립가설을 유의수준 0.05에서 채택하는 것이다. 문제는 B 약물의 호전율은 64.3%이고 A 약물의 호전율은 62.5%로서 이 둘의 차이란 1.8%에 불과하다는 사실이다. 이것은 100명의 환자 중, B 약물에 임상적 호전을 보이는 환자가 A 약물에 비해 2명이 채 더 안된다는 의미이다. 만일 B 약물의 수가가 A 약물보다 훨씬 비싸다고 가정한다면 이러한 결과를 놓고 자신있게 A 약물대신 B 약물을 선택할 의사가 과연 몇 명이나 되겠는가? 여러 가지 조건을 고려하여 1.8%의 차이가 '실제적 유의성'을 갖고 있지 않다고 판단을 내린다면, 아무리 '통계적 유의성'이 얻어졌다고 해도 이러한 검정결과로부터 기존의 행위나 신념을 바꾸지는 않을 것이다. 요컨대 실제적 유의성은 여러 가지 상황과 조건 등으로부터 판단해야 할 '관점의 문제'이며, 통계적 방법이나 기타 방법 등을 이용하여 이를 계량화하기가 결코 쉽지 않음을 명심해야 한다.

지침 2 : 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.

통계적 방법을 이용하여 자료분석을 실시하는 연구자들 간에 퍼져 있는 잘못된 견해 중 가장 대

표적인 것은, 연구의 주 목적이 마치 통계적 유의성을 얻는 데 있다는 식의 태도이다. 즉, 통계적 유의성이 얻어진 연구결과는 전문학회지 등을 통해 널리 공표할만한 업적이 되고, 통계적 유의성이 얻어지지 않았다면 그것은 연구자의 손 안이나 책상 설합 안에서 사장되어도 좋다는 것이다. 어느 분야라도 좋으니 전문학회지를 들추어 보고 게재된 논문의 몇% 정도가 통계적 유의성이 있는 결과를 발표하고 있는지 확인해 보라. 독자들은 이 수치가 거의 100%에 가깝다는 사실을 발견하게 될 것이다. 이는 모든 연구로부터 항상 통계적 유의성이 있는 결과만이 얻어짐을 말해 주는 것이 아니라, 연구자들이 통계적 유의성이 얻어진 결과만을 주로 발표하고 있고, 심지어 전문잡지의 편집자들도 이러한 견해에 동조하여 통계적 유의성을 얻지 못한 연구는 게재 자체를 꺼리기 때문이다.

통계적 유의성의 추구(追求)를 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. 예를 들어 치료법 B가 기존의 치료법 A에 비해 월등한 효과를 갖고 있는 것으로 통계적 유의성이 얻어졌다면 우리는 두 가지 치료방법의 효과가 동일하다는 귀무가설을 기각할 것이다. 그러나 이러한 귀무가설의 기각이 곧 A라는 치료법을 모두 B라는 치료법으로 바꾸겠다는 결정을 의미하지는 않는다. 지난 번 도입부에서 논의한 것처럼 축적된 자료를 이용하여 통계적 결론을 내리는 것과, 이러한 결론으로부터 기존의 행위나 신념을 바꾸는 것은 전혀 별개의 문제이며 후자는 통계학의 영역 밖에 있다.

연구의 결과, 통계적 유의성이 얻어졌을 때 이로부터 기존의 행위나 신념을 변경하기 위해서는 다음과 같이 통계학외적인 요소들을 고려해야 한다. 첫째, 귀무가설이 현재 얼마나 확고한 지지를 얻고 있는가 하는 것이다. 만일 귀무가설이 일반적으로 통용되는 지식이나 신념이라면 이를 반증하기 위해 더 강한 경험적 증거, 즉 더 작은 유의수준을 필요로 할 것이다. 따라서 유의수준 0.05는 고정된 것이 아니고 경우에 따라 달리 조정할 필요가 있다. 대개의 연구자들이 '$p<0.05$'를 무슨 황금률이나 되는 것처럼 알고 있는데, 이는 분명 잘못된 믿음인 것이다. 요컨대 $\alpha=0.05$가 모든 경우에 통용될 수 있는 것은 아니고, 이는 귀무가설에 대한 현재의 믿음에 따라 신축적으로 바뀌어야만 한다.

둘째, 귀무가설을 기각하고 대립가설을 선택하여 이에 따라 기존의 행위나 신념을 변경할 때 어느 정도의 파급효과가 있을 것인지를 고려해야 한다. 앞에서 예로 든 A와 B 치료법의 경우, 치료법 B로 바꾸기 위해서는 기존의 시설을 모두 새 것으로 교체해야 하고 의료인력들을 모두 재교육해야 한다면 귀무가설에 반하는 더 강한 경험적 증거가 요구되므로 더 작은 유의수준이 필요할 것이다. 이러한 이유들로 전문지 등에 게재되는 논문에는 단순히 '통계학적으로 의미 있다 또는 없다'만을 표기할 것이 아니라 구체적인 p값을 제시하는 것이 바람직하다. 요즈음 시판되고 있는 통계팩키지들은 모두 구체적인 p값을 계산해 주므로, 계산의 어려움은 더 이상 구체적인 p값 제시의 걸림돌이 될 수 없다.

그러나 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. 예를 들어 태양이 지구 주위를 돈다는 귀무가설이 기각되기 위해서는 무려 십 수 세기의 시간이 필요했으며, 수없이 많은 통계적 유의성이 쌓여 오늘날 공전의 개념이 확고부동한 과학적 진리가 되기에 이른 것이다. 앞에서 예로 든 치료법 A, B도 마찬가지인데 단 한 번의 연구결과에서 아무리 작은 P값이 얻어졌고 이로부터 B라는 치료법이 더 우수한 것으로 밝혀졌다고 하더라도 바로 이러한 사실이 '진리'라는 이름을 갖게 되는 것은 아니다. 그러나 만일 여러 번의 실험이나 조사로부터 계속 치료법 A

와 B의 효과가 동일하다는 귀무가설이 기각되었다면, 이로부터 '치료법 B가 치료법 A보다 우수하다'는 방향으로 우리의 신념을 변경할 수 있는 것이다.

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. 따라서 통계적 유의성이 얻어진 결과만을 발표하려는 연구자의 자세나, 통계적 유의성이 입증되지 않은 연구논문의 게재를 꺼리는 전문지 편집자들의 입장은 수정되어야 마땅하다. 보통 통계적 유의성이 입증되지 않은 연구 결과를 '부정적 결과(negative results)'라고 부른다. 이러한 부정적 결과를 얻은 연구 논문을 발표함으로써 기대되는 효과는, 첫째 타 연구자들에게 부정적 결과를 알려 더 이상 불필요한 동(同) 주제의 연구수행을 막을 수 있으며, 둘째 부정적 결과가 새로운 발견으로 이어지는 경우가 종종 있다고 하는 것이다. 후자의 예로 'H_1 : 광속(光速)은 일정하지 않다'라는 대립가설에 대한 통계적 유의성을 입증하지 못함으로써 광속의 불변을 보고했던 미켈슨(Michelson)과 몰레이(Morley)를 꼽을 수 있으며, 이는 과학사적으로 볼 때 20세기 양자역학(量子力學, quantum mechanics)의 첫 장을 여는 도입부 역할을 훌륭히 수행해 냈던 것이다.

지침 3 : 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 것(hypothesis-confirming)이라기 보다, 새로운 가설의 형성을 지지하고 촉진하는 것(hypothesis-generating)으로 이해되어야 한다.

개인용 컴퓨터(personal computer)에서도 쉽게 자료분석을 할 수 있게 해 주는 통계팩키지의 보급은 연구의 활성화를 촉진했다는 긍정적 측면을 갖고 있다. 그러나 자료분석에 대한 연구자의 통계적 소양이 부족한 상태에서 모든 것을 컴퓨터와 통계팩키지에 내 맡김으로써, 연구 주제가 내포하고 있는 과학적 의미에 대한 집중된 지적 탐구를 게을리하게 되고 오히려 시행착오적이 탐색에 의해 연구를 시행하게 했다는 비판을 받고 있음도 또한 사실이다[1]. 통계적 분석방법이나 결과에 대한 해석방법의 기본 원리를 몰라도 컴퓨터는 입력된 자료에 대해 사용자로부터 강요받은 모든 결과를 '유순하게' 내 놓는다. 예 2를 통해 의도하지 않았던 통계적 유의성이 발견되었을 때 대다수의 연구자들이 어떤 자세를 취하는지 알아보도록 하자.

예 2) S 대학병원 종양내과의 박 교수는 간암환자군과 정상인으로부터 각각 100가지 종류의 변수를 측정하고, 이들에 대하여 유의수준 0.05에서 통계적으로 유의한 차이가 있는지를 검정하였다. 그 결과 5개의 변수에서 통계적 유의성이 있음을 발견하고, 통계적 유의성이 발견된 변수만을 문헌에 발표하였다. 자료분석에 대한 박 교수의 입장은 통계학적 방법론을 원용하는 대다수 연구자들의 자세를 대변하고 있다. 그러나 이러한 태도는 근본적으로 두 가지의 문제점을 갖고 있다. 첫째, 5개의 변수에서 통계적으로 유의한 차이가 발견된 것은 우연에 의한 결과이지, 연구자가 사전에 어떤 통계적 가설을 세운 뒤 수집한 자료로부터 가설검정을 거쳐 유도된 것이 아니라는 사실이다. 통계적 유의성을 발견하기 위해 100개나 되는 변

1) Etzioni는 소형컴퓨터 및 팩키지 프로그램이 과학적 연구 수행에 미치는 악영향에 대해 다음과 같은 흥미있는 지적을 한 바 있다. "이제 과학자들은 충분한 사격을 하게 되면 어떤 흥미로운 결과나 상관관계를 맞출 것이라는 가정하에, 어둠 속에서 계속적인 발포만 하고 있도록 하는 유혹을 과거 어느 때 보다 강력하게 받고 있다."(A. Etzioni, Effect of small computers on scientist, Sciene, 189:4197, 1975)

수를 샅샅이 뒤질 수 있는 것은 컴퓨터가 이룩한 자료분석의 개가라고 할 수 있다. 즉, 연구자가 의도하든 의도하지 않았든 간에 컴퓨터는 시키는대로 결과를 출력한다. 이처럼 자료를 여러 가지 통계적 방법으로 샅샅이 뒤져 기대하거나 의도하지 않았던 통계적 유의성을 찾아내려고 하는 것을 자료훑기 비뚤림(data-dredging bias)이라고 부른다.

자료분석에 대한 박 교수의 두 번째 문제점은, 보통 우리가 다중가설의 검정(multiple hypothesis testing)이라고 부르는 것이다[2]. 즉 유의수준이 0.05라면 제 1종의 오류가 5%까지도 일어날 수 있다는 의미이며, 이는 귀무가설이 옳더라도 장기적인 측면에서 100번 중 5번은 통계적 유의성이 있는 결과를 얻을 것으로 기대된다는 뜻이다. 따라서 박 교수가 100개의 변수 중 5개에서 통계적 유의성을 발견한 것은 α=0.05에서 아주 당연한 결과이며, 이것 자체에 어떤 특별한 의미를 부여하기란 어렵다고 하는 사실을 명심해야 한다.

자료를 샅샅이 뒤져 통계적 유의성이 있는 결과를 찾아내려는 것은, 방법상 분명 문제가 많이 있으나, 과학적 타당성이 아주 없지는 않다. 과학사적으로 볼 때도 자료에서 전혀 기대하지 않았던 양상이나 현상 등이 그 후의 재확증을 거쳐 위대한 발견으로 이어진 예가 허다하다. 그러나 이 때 중요한 것은 우연히 발견된 양상이나 현상 등이, 이러한 양상의 검정을 목적으로 하는 추후의 실험이나 통계조사 등을 통해 재확증 또는 재발견되어야 한다는 사실이다. 통계적 가설검정이란 바로 이러한 재확증의 수단에 적합한 방법이라고 말할 수 있다. 따라서 통계적 유의성이 있는 결과를 찾고자 자료를 샅샅이 뒤지는 것은 통계적 가설검정의 기본 목적과는 다름을 명심해야 한다. 요컨대 자료를 샅샅이 뒤져 찾아낸 통계적 유의성으로부터 유도된 가설을 검정하기 위해 동일한 자료를 재 이용하는 것은 엄격한 의미에서 옳지 않다. 의도하거나 기대하지 않았던 통계적 유의성을 발견했을 때, 이를 '가설형성적(假說形成的)'인 측면에서 이해해야 한다는 것은 바로 이러한 상황을 가리키는 것이다.

지침 4 : 가설검정으로 대표되는 통계적 자료 분석 방법은 결코 진리에 도달하는 유일한 방법이 아니다. 아무리 잘 시행된 분석이라 할지라도 완전할 수는 없으며, 자료에는 항상 무엇인가 잘못되었을 가능성이 있음을 명심해야 한다. 요컨대 자료 분석의 성패는 분석 단계가 아닌 자료 수집 단계에서 이미 결정되는 것이다.

가설검정의 오류는 무작위오차만을 대상으로 하며 이는 다른 종류의 오차가 발생하지 않았다는 것을 전제로 했을 때 가능한 것이다. 즉, 자료수집단계에서부터 편의가 없어야만 이후의 공정한 분석이 보장될 수 있음을 의미한다. 따라서 비확률표본추출법으로부터 얻은 자료에 대해 확률론에 근거한 통계적 추론의 방법을 적용하는 것은 원칙적으로 타당하지 않다. 통계학에서 표본추출을 말할 때 확률표본추출만을 의미한다는 것은 바로 이러한 이유에서이다. 그러나 편의표본(convenience sample)이나 판단표본(judgemental sample)에서 얻은 자료에 대해서도 마치 확률표본추출법으로부터 얻은 자료인 것처럼 가정하고 통계적 추론을 실시하는 경우를 흔히 볼 수 있는데, 사회학이나 의학 분야의 자료가 대표적이다. 물론 이러한 사실이 사회학이나 의학 등의 연구논문에서 사용된 자료분석의 방법은 모두 틀렸다는 것을 의미하지는 않는다. 중요한 것은 자료분석의 타당성이 이미 자료수집단계에서부터 결정된다는 사실을 명심하고, 분석의 결과를 해석할 때 전술한 제한점을 고려하는 것이다.

이상 가설검정의 원리 I, II에서 다루어진 내용을 요약하면 다음과 같다.

(1) 연구자나 조사자가 타당성을 입증하고자 하는 통계적 가설을 대립가설, 타당하지 않음을 밝혀 무효화하려는 가설은 귀무가설이라고 부르는데 이

2) 다중가설의 검정에 따라 제 1종의 오류가 증가되는 것을 보정하기 위해 Bonferroni, Tukey, Duncan, Sheffe 등이 고안한 방법이 있으나, 이에 대한 소개는 이 글 수준을 넘는 것이므로 생략한다. 시판되고 있는 통계팩키지에는 이들 중 한두 가지 방법을 이용할 수 있도록 선택사양을 제공한다.

둘은 상호배반이다. 가설검정의 제 절차를 거쳐 해당 가설을 받아들이는 것을 채택, 받아들이지 않는 것은 기각이라고 한다. 보통 가설검정에서는 귀무가설의 측면에서, 귀무가설을 기각할 수 있든가 기각하지 못하든가를 결정하게 된다. 모집단에서는 귀무가설이 참인데도 가설검정의 결과 귀무가설을 기각하는 것을 제 1종의 오류 또는 α 오류라고 부르며, 제 1종의 오류가 발생한 상황을 위양성이라고 한다. 한편 모집단에서는 귀무가설이 참이 아닌데도 귀무가설을 기각한 경우를 제 2종의 오류 또는 β 오류라고 하며, 이러한 상황은 위음성이 된다. 각 오류를 범할 확률을 α, β로 표시하며, 1-β는 검정력으로 이는 모집단에서 귀무가설이 참이 아닐 때 가설검정이 오류의 발생없이 귀무가설을 기각할 수 있게 해 주는 정도를 나타낸다.

(2) 표본으로부터 얻은 통계량의 값이 해당 표본통계량의 분포에서 구해진 표본통계량의 값보다 더 극단적인 값을 갖게 될 누적확률을 p값이라고 하며 이는 일종의 외부확률이 된다. 따라서 p값의 영역내에 있는 실측값들은 해당 표본통계량의 분포에서 비교적 드물게 관찰되는 값들이다. 요컨데 각 표본통계량의 실측값에 대한 p값이란, 임의로 자료를 하나 선택하였을 때 이것의 실측값이 표본통계량의 실측값보다 분포에서 꼬리 쪽으로 치우쳐 있는 값이 될 확률을 의미한다.

(3) 가설검정의 원리는 일단 귀무가설이 옳다는 전제로부터 표본통계량에 대한 분포를 구성하고,이러한 분포로부터 얻은 표본통계량의 p값이 어떤 값 – 유의수준 α – 보다 작을 때 귀무가설을 기각한다는 것이다. 즉, 제 1종의 오류가 발생할 확률을 α만큼 감수하겠다는 것이 가설검정의 원칙이며, α는 이러한 의사결정에 수반되는 위험의 상한선을 확률로 표시한 것이라고 볼 수 있다. 통계학에서는 유의수준을 보통 0.01, 0.05, 0.1 등으로 하며, α 오류를 최소화하면서 적절한 검정력을 보장해 주는 검정법이 좋은 검정법이 된다.

(4) p값의 계산에 사용된 표본통계량을 검정통계량이라고 부르며, 검정통계량이 Z이면 정규검정법, T이면 t 검정법이라고 한다. 정규검정법이나 t 검정법은 모두 모평균에 대한 가설검정에 사용되는 방법이며, 모표준편차를 알고 있는 지의 여부, 표본의 크기 등에 따라 각 검정법을 선택한다. 그러나 표본의 크기가 아주 작지 않다면 어느 검정법을 선택하더라도 큰 차이가 나지 않는다.

(5) p값이 α와 같아지는 검정통계량의 절대값을 임계값이라고 하며, 검정통계량의 절대값이 임계값보다 크면 $p < \alpha$가 되어 귀무가설을 기각할 수 있다. 이 때 귀무가설을 기각할 수 있게 해 주는 검정통계량의 영역을 기각역이라고 한다. 따라서 임의의 검정법에 대해서 기각역은 '|검정통계량| ≤ 임계값' 과 같이 표시할 수 있다. 모수가 어떤 값보다 '크다' 또는 '작다'의 형태로 대립가설이 주어질 때 이를 단측가설이라고 하며, 단측가설의 검정을 단측검정, 기각역은 단측기각역이라고 부른다. 단측기각역은 검정통계량이 양수이면 '검정통계량 ≥ 임계값', 음수이면 '검정통계량 ≤ -임계값'으로 주어진다. 한편 모수가 어떤 값과 '같다' 또는 '다르다'의 형태로 대립가설이 주어질 때 이를 양측가설이라고 부른다. 양측검정에서는 유의수준 α가 양 쪽 꼬리로 $\alpha/2$씩 나누어지므로 $p = \alpha/2$를 만족하는 어떤 값보다 검정통계량의 절대값이 커야만 귀무가설을 기각할 수 있다. 따라서 동일한 유의수준에서는 양측검정의 임계값이 단측검정의 임계값보다 크며 이는 귀무가설을 기각하기가 더 어렵다는 의미가 된다.

(6) 가설검정을 통해 귀무가설을 기각하였을 때 '유의수준 α에서 통계적으로 의미있는' 결과를 얻었다고 말하며, 이를 '통계적 유의성'이란 용어로 표현한다. 그러나 통계적 유의성은 실제적 유의성과는 다른 개념이며, 후자는 관점의 문제로서 통계학의 영역 밖에 위치한다. 연구나 조사의 목적이 통계적 유의성이 있는 결과를 얻기 위한 것이라고 생각하는 것은 잘못이며, 가설을 설정하지 않은 상태에서 우연히 발견된 통계적 유의성으로부터 마치 해당 가설의 통계적 유의성이 증명된 것처럼 취급해서도 안된다. 자료수집에 임의성이 결여되었다면 통계적 추론을 적용하는 것이 논리적으로 타당하지 않다. 따라서 자료분석의 성패 여부는 이미 자료수집 단계에서 결정되는 것이다.