Journal of Biotechnology and Bioindustry
Vol.13, December, 2025. 30~45
https://doi.org/10.37503/jbb.2025.13.30

Be Biologist: 아무도 알려주지 않는 바이올로지스트가 되기 위한 기본 5

김 종 은*

충청북도 증평군 대학로 61 한국교통대학교 보건생명대학 식품생명학부 식품공학전공 27909

Be Biologist: Basic Knowledge for Biologist Which No One Told 5

Jong-Eun Kim*

Department of Food Engineering, Korea National University of Transportation, Jeungpyeong 27909, Korea

ABSTRACT

Biological research increasingly relies on high-throughput and high-dimensional measurements, making experimental design and statistical analysis inseparable determinants of credibility. A frequent failure mode in early-stage research is treating statistics as an end-stage formality rather than a design principle, which inflates false positives, reduces power, and weakens reproducibility. This article presents a practical, design-to-analysis framework for novice experimental biologists. Key elements include rigorous control-group construction (negative, vehicle, sham, and positive controls), bias prevention through randomization and blinding, clear separation of technical versus biological replication, and proactive management of batch effects. Guidance is provided on sample-size planning and data quality control as prerequisites for valid inference, along with interpretation strategies that account for interaction terms and hierarchical or correlated data structures. For mean comparisons, the logic of multiple testing is formalized via family-wise error rate control, and decision rules are outlined for selecting post-hoc procedures aligned with the comparison objective (all-pairs versus control-based contrasts). Finally, a reproducible workflow using open-source R/Python tools in cloud notebook environments (e.g., Google Colab) is described, including how large language models can lower the barrier to script-based analysis without replacing statistical reasoning. Emphasis is placed on reporting effect sizes and confidence intervals to support biologically meaningful conclusions beyond statistical significance.

Key words : biologist, graduated school student, statistics, newbe

Ⅰ. 서 론

생명과학 연구의 핵심 과제는 복잡한 생명 현상 속에서 원인과 결과의 관계를 분리하고, 그 인과적 연결을 검증 가능한 증거로 제시하는 데 있다. 그러나 생체 시스템은 물리·화학적 시스템과 달리 항상성 유지와 적응 반응을 포함한 다층적 조절 메커니즘 위에서 작동하며, 동일한 조건을 설정했다고 하더라도 개체 또는 세포 집단의 이질성, 미세 환경의 차이, 시간에 따른 상태 변화가 필연적으로 개입한다(1). 따라서 관찰된 변화가 독립변수 조작에 의해 유발된 효과인지, 우연적 변동이나 측정 과정의 오차가 만들어낸 결과인지 구별하는 것은 연구의 출발점이자 결론의 신뢰도를 좌우하는 핵심 절차다. 이때 통계학적 추론은 단순히 결과를 정리하는 기술이 아니라, 연구 질문을 검정 가능한 형태로 구조화하고 불확실성 하에서 결론이 갖는 정당성을 정량화하는 연구 방법론으로 기능한다(1, 2).

연구 과정에서 문헌 고찰은 가설의 근거와 연구의 위치를 제공하고, 실험 수행은 생물학적 현상을 관찰 가능한 데이터로 전환하며, 연구 노트는 조작과 조건의 세부를 기록하여 재현을 가능하게 한다. 데이터 시각화는 패턴과 이상치를 드러내고 후속 분석의 방향을 제시한다(2). 그러나 이러한 요소들은 실험 디자인과 분석 계획이라는 기반 위에서만 설득력 있는 지식으로 수렴한다. 대조군의 구성과 비교 구조가 적절하지 않으면, 관찰된 변화는 독립변수의 효과가 아니라 용매, 조작 스트레스, 배치 차이, 기기 상태와 같은 교란 요인의 반영일 수 있다. 무작위 배정이 부실하면 특정 처리군에만 특정 조건이 집중되어 처리 효과처럼 보일 수 있고, 블라인딩이 부족하면 측정·분석 단계에서 기대에 따른 편향이 개입할 수 있다. 반복 단위가 명확하지 않거나 생물학적 반복과 기술적 반복이 혼동되면 의사반복으로 인해

* Jekim14@ut.ac.kr

유의성이 과대평가될 위험이 크다. 배치 효과가 처리군과 얽히면 처리 효과와 날짜 · 로트 · 실험자 효과를 분리할 수 없게 되어, 통계적 유의성의 유무와 무관하게 해석 가능성이 훼손된다. 결국 통계 분석은 실험이 끝난 뒤 적용하는 장식적 절차가 아니라, 설계 단계에서부터 무엇을 통제하고 무엇을 비교할지 결정하는 논리 체계와 결합할 때 의미가 있다(2, 3).

초보 연구자에게 흔한 문제는 통계를 실험 종료 후의 형식적 검증으로 오인하는 데서 발생한다. 표본 크기 산정을 생략한 채 관행적으로 반복수를 정하거나, 음성 대조군을 단일화하여 기준점이 불명확한 상태로 실험을 진행한 뒤, 기대한 유의확률이 나오지 않으면 분석 방법을 바꾸거나 일부 데이터를 제외하는 방식으로 대응하는 사례가 반복된다. 이러한 접근은 연구 효율을 저해할 뿐 아니라, 결론이 특정 선택에 과도하게 의존하도록 만들어 결과의 안정성을 낮춘다(4). 또한 실험 조건과 분석 절차가 충분히 기록·공유되지 않으면 동일한 연구자가 재분석해도 결과가 달라지거나, 다른 연구자가 동일 설계를 재현했을 때 결론이 반복되지 않을 가능성이 커진다. 재현성 위기는 단순히 동일한 수치가 재현되지 않는 문제를 넘어, 후속 연구의 방향 설정을 왜곡하고 자원을 낭비하며 학술 커뮤니케이션의 신뢰를 약화하는 구조적 문제로 이어질 수 있다(5).

현대 생물학 연구는 고처리량 스크리닝, 정량 이미징, 오믹스 기반 분석 등으로 인해 데이터의 규모와 복잡성이 동시에 증가하고 있다. 이 환경에서는 단순한 평균 비교만으로 연구 질문에 답하기 어려우며, 데이터 생성 과정의 계층 구조와 상관 구조를 반영하는 모델링, 다중 비교와 다중 검정에 대한 오류 통제, 효과 크기와 신뢰구간을 기반으로 한 해석이 요구된다. 더 나아가 분석 과정의 선택을 투명하게 기록하고, 동일 데이터와 동일 절차에서 동일 결과를 재생산할 수 있는 재현 가능한 워크플로가 필수 조건으로 자리 잡고 있다. 이러한 요구에 대응하기 위해 본고는 대조군 설계와 비교 가능성 확보 원칙을 정교화하고, 분산분석 이후 사후 검정 선택의 이론적 근거와 적용 기준을 연구 목적 중심으로 정리한다. 또한 오픈소스 기반의 R을 활용하여 데이터 처리, 통계 검정, 시각화, 결과 보고, 환경 고정과 협업까지 연결되는 재현 가능한 분석 체계를 제안함으로써, 연구 설계와 분석이 분리되지 않는 통합적 연구 역량의 필요성을 논의하고자 한다.

II. 본 론

1. 실험 디자인의 내적 타당성: 대조군의 정교한 설정

과학적 실험, 특히 복잡계인 생명 현상을 다루는 바이오 연구의 본질은 엄밀한 비교를 통해 인과관계를 규명하는 데 있다. 연구자가 독립변수를 조작했을 때 관찰되는 종속변수의 변화가 오직 그 독립변수에 기인한 것임을 입증하기 위해서는, 독립변수 이외의 모든 잠재적 교란 요인이 통제된 비교 집단이 필수적으로 요구된다. 이러한 맥락에서 대조군은 단순히 결과의 차이를 보여주기 위한 기준선 역할에 그치지 않는다. 대조군은 실험 시스템 내에 존재할 수 있는 미지의 교란 요인을 탐지하고, 측정 도구의 안정성과 민감도를 보증하며, 해석 가능한 결론의 논리적 범위를 규정하는 실험의 나침반과 같다. 만약 대조군 설정이 부실하여 내적 타당성이 확보되지 않는다면, 이후 수행되는 그 어떠한 정교한 통계 분석이나 p-값의 유의성도 과학적 사실을 담보할 수 없다. 따라서 훌륭한 연구 데이터는 화려한 실험 기법이 아닌, 치밀하게 설계된 대조군에서 비롯된다는 사실을 명심해야 한다(1, 2).

실험 디자인에서 가장 기초적이면서도 빈번하게 오류가 발생하는 부분은 음성 대조군의 설정이다. 많은 초보 연구자가 음성 대조군을 단일한 그룹으로 설정하곤 하지만, 실험의 해상도를 높이기 위해서는 이를 무처리 대조군과 용매 대조군으로 엄격히 세분화해야 한다. 우선, 무처리 대조군(null control 또는 intact control)은 어떠한 인위적 조작도 가하지 않은 자연 상태의 생체 또는 세포 집단을 의미한다. 이 그룹은 실험 환경 자체가 유발하는 '배경 잡음(background noise)'을 평가하는 데 핵심적인 역할을 한다. 연구자가 인지하지 못하는 사이, 인큐베이터 개폐 시 발생하는 CO_2 농도와 온도의 미세한 변동, 배지 교체 과정에서의 물리적 전단력, 현미경 관찰을 위한 광 노출, 플레이트 이동 중의 진동 등은 살아있는 세포의 대사 및 신호 전달 체계에 영향을 미칠 수 있다. 무처리 대조군은 이러한 환경적 변수 하에서 시스템이 보이는 기저 반응(basal level)을 보여주며, 만약 이 그룹의 데이터 변동 폭이 크다면 약물 처리 효과를 논하기에 앞서 실험 환경의 안정성을 재점검해야 함을 시사한다(3).

무처리 대조군과 구별되어야 할 필수적인 비교 집단은 용매 대조군(vehicle control)이다. 대부분의 생리활성 물질이나 신약 후보 물질은 소수성 성질을 띠어 물에 잘 녹지 않으므로, 이를 용해하기 위해 DMSO(dimethyl sulfoxide), 에탄올, 메탄올 등의 유기 용매를 사용하게 된다. 문제는 이러한 용매들이 생물학적으로 불활성이지 않다는 점이다. 예를 들어,

DMSO는 저농도에서도 세포막의 유동성을 증가시켜 물질 투과를 비정상적으로 촉진하거나, 특정 신호 전달 경로를 활성화하고 산화 스트레스를 유발할 수 있다. 심지어 고농도에서는 세포 분화를 유도하거나 세포 주기 정지를 일으키기도 한다. 따라서 약물 처리군과 비교해야 할 통계적 기준점은 자연 상태의 무처리 대조군이 아니라, 반드시 약물과 동일한 농도의 용매가 포함된 용매 대조군이어야 한다. 이를 통해 관찰된 효과가 약물 고유의 작용인지, 용매의 독성이나 부작용인지를 명확히 분리해낼 수 있다(6).

용매 대조군 설정 시 가장 주의해야 할 기술적 원칙은 '최종 용매 농도의 동일성' 유지이다. 예를 들어, 약물의 농도 의존적 효과를 확인하기 위해 저농도, 중농도, 고농도 처리군을 설정할 때, 고농도 약물을 제조하기 위해 더 많은 용매가 들어가는 경우가 발생할 수 있다. 이때 모든 처리군과 대조군에서의 용매 최종 농도(예: 0.1%)를 엄격하게 일치시키지 않으면, 고농도 처리군에서 나타난 세포 사멸이나 독성이 약물에 의한 것인지 증가한 용매 농도에 의한 것인지 구분할 수 없게 된다. 이는 데이터 해석을 불가능하게 만들 뿐만 아니라, 논문 심사 과정에서 실험 디자인의 치명적인 결함으로 지적받을 수 있다. 따라서 부족한 용매만큼을 추가하여 모든 웰이 정확히 동일한 용매 환경에 노출되도록 하는 것은 실험의 내적 타당성을 확보하기 위한 타협할 수 없는 원칙이다.

약물 처리가 아닌 물리적, 기계적 조작이 포함되는 실험에서는 조작 대조군(sham control)의 도입이 필수적이다. 생체 내 실험(*in vivo*)이나 유전자 도입 실험 등에서 연구자가 가하는 행위 자체가 시스템에 강력한 스트레스로 작용하여 결과를 왜곡할 수 있기 때문이다. 예를 들어, 뇌의 특정 부위에 전기 자극을 주는 동물 실험을 수행할 때, 두개골을 절개하고 전극을 삽입하는 외과적 수술 과정은 실험 동물에게 심각한 염증 반응과 통증 스트레스를 유발한다. 이때 전기 자극의 순수한 효과만을 분리해내기 위해서는, 대조군 동물에게도 동일하게 마취와 수술, 전극 삽입 과정을 수행하되 전류만 흘리지 않는 Sham Operation을 실시해야 한다. 세포 실험(*in vitro*)에서도 마찬가지로, 유전자 과발현을 위해 플라스미드 DNA를 트랜스펙션할 때, 타겟 유전자가 없는 공벡터(empty vector)를 동일한 조건으로 처리한 Mock Transfection 그룹을 설정해야 한다. 이를 통해 관찰된 표현형의 변화가 타겟 유전자의 기능인지, 아니면 외부 물질 유입과 리포좀 시약에 의한 세포 스트레스 반응인지를 명확히 구분할 수 있다.

실험 시스템의 유효성을 검증하기 위해 양성 대조군(positive control)을 반드시 포함해야 한다. 초보 연구자들은 자신의 타겟 물질에만 집중하여 양성 대조군을 생략하는 경향이 있으나, 이는 실험 결과가 예상과 다르게 나왔을 때 그 원인을 규명할 수 없게 만드는 위험한 선택이다. 만약 신약 후보 물질을 처리했을 때 아무런 효과가 관찰되지 않았다면, 이것이 후보 물질의 효능 부재 때문인지, 아니면 세포 상태 불량, 시약 변질, 기기 오작동 등 실험 시스템의 실패(technical failure) 때문인지 판단할 근거가 필요하다. 이때 이미 효능이 입증된 표준 약물(standard drug)을 양성 대조군으로 사용하여 확실한 반응을 확인해야만, 후보 물질의 '효과 없음'이라는 결과도 과학적 팩트로 인정받을 수 있다. 또한, 양성 대조군은 연구 결과의 외적 타당성을 높이는 잣대가 된다. 단순히 효과가 있다는 사실을 넘어, 기존의 표준 치료제와 비교하여 동등하거나 우수한 효능을 보였다는 상대적 평가는 연구 결과의 가치를 극대화하고 독자를 설득하는 강력한 근거가 된다.

2. 편향 방지 전략: 무작위화와 블라인딩의 체계적 적용

실험 디자인에서 대조군을 아무리 정교하게 설정하였다고 하더라도, 실험 대상(subject)이 각 군(group)에 배정되는 과정에서 편향(bias)이 개입된다면 실험의 비교 가능성(comparability)은 근본적으로 붕괴한다. 과학적 인과 추론의 핵심 전제는 '다른 모든 조건이 동일할 때(*Ceteris paribus*), 오직 독립변수의 조작만이 종속변수의 변화를 유발한다'는 것이다. 그러나 배정 과정에서의 편향은 각 군 간의 기저 특성(baseline characteristics)을 이질적으로 만들며, 결과적으로 관찰된 효과가 독립변수에 의한 것인지, 아니면 애초에 존재했던 집단 간의 특성 차이에 의한 것인지를 구분할 수 없는 교란(confounding) 상태를 초래한다. 이를 방지하고 내적 타당성을 확보하기 위한 핵심적인 방법론적 장치가 바로 무작위화(randomization)와 블라인딩(blinding)이다(7, 8).

1) 무작위화(randomization): 교란 변수의 확률적 통제

무작위화는 연구자가 인지하고 있는 '알려진 교란 요인'뿐만 아니라, 미처 파악하지 못한 '알려지지 않은 잠재적 교란 요인'까지도 각 처리군에 균등하게 분산시키는 가장 강력한 통계적 도구이다. 생물학적 시스템은 본질적으로 다변량적이며, 연구자가 통제할 수 없는 수많은 확률적 변수가 존재한다. 무작위화는 이러한 변수들이 특정 군에 편중되는 것을 확률적으로 방지함으로써, 처리 효과 추정치(estimate of treatment effect)의 비편향성(unbiasedness)을 보장한다.

특히 세포 생물학이나 고속 스크리닝(high-throughput screening) 연구에서 널리 사용되는 96웰 플레이트(96-well plate) 실험

에서 무작위화의 중요성은 빈번히 간과된다. 다수의 연구자가 실험의 편의성을 위해 A행에는 대조군, B행에는 저농도, C행에는 고농도 약물을 배치하는 순차적 블록 배치를 관행적으로 사용한다. 그러나 플레이트의 가장자리 웰(edge well)은 중앙부 웰에 비해 배지의 증발 속도가 빨라 삼투압과 약물 농도가 비정상적으로 상승하는 '가장자리 효과(edge effect)'가 발생할 가능성이 매우 높다. 또한, 인큐베이터 내부의 공기 순환에 따른 미세한 온도 구배(thermal gradient)나 실험자가 파이펫팅을 수행하는 순서와 시간에 따른 세포의 스트레스 차이와 같은 '위치 효과(position effect)'는 데이터에 심각한 왜곡을 초래할 수 있다. 만약 특정 처리군을 특정 행이나 열에 집중 배치한다면, 위치 효과가 처리 효과와 혼재되어 위양성(false positive) 또는 위음성(false negative) 결론을 유도하게 된다. 따라서 난수표를 이용한 완전 무작위 배치(completely randomized design)나, 플레이트 구역을 구획화하여 무작위화하는 블록 무작위화(block randomization)를 적용해야 한다. 아울러, 물리적 변동성이 큰 가장자리 웰은 데이터 수집에서 배제하고 배지만을 채워 완충 구역(buffer zone)으로 활용하는 것이 증발에 의한 편향을 최소화하는 기본적인 설계 전략이다.

동물 실험(*in vivo*)에서도 무작위화는 필수적이다. 실험 동물의 체중, 월령, 성별, 유전적 배경뿐만 아니라, 사육 케이지의 위치(랙의 상단 대 하단, 조명과의 거리), 사육 밀도 등은 모두 결과에 영향을 미칠 수 있는 잠재적 교란 요인이다. 단순히 케이지에서 잡히는 순서대로 대조군과 실험군에 배정할 경우, 행동이 둔하거나 체중이 많이 나가는 개체, 혹은 사람을 덜 피하는 개체가 특정 군에 집중될 위험이 있다. 이를 방지하기 위해 층화 무작위화(stratified randomization)를 적용하여 체중이나 성별과 같은 주요 공변량(covariate)이 각 군에 균형 있게 분포되도록 강제해야 한다. 이는 군 간의 기저 변동성을 줄여 오차 분산(error variance)을 감소시키고, 결과적으로 통계적 검정력(statistical power)을 향상시키는 효율적인 전략이다.

2) 블라인딩(Blinding): 측정 및 분석 편향의 차단

무작위화가 실험 시작 전의 선택 편향(selection bias)을 통제한다면, 블라인딩은 실험 수행 및 데이터 분석 단계에서 발생할 수 있는 관찰자 편향(observer bias)과 확인 편향(confirmation bias)을 차단한다. 생명과학 실험 데이터는 기계적으로 산출되는 객관적인 수치로만 구성되지 않는다. 현미경 이미지 분석에서 관심 영역(ROI)을 설정하거나, 조직병리 슬라이드에서 염증 정도를 점수화하거나, 유세포 분석(flow cytometry)에서 게이팅(gating) 경계를 설정하는 과정에는 필연적으로 연구자의 주관적 판단이 개입된다.

연구자가 특정 샘플이 약물 처리군임을 인지하고 있는 상태에서 데이터를 분석할 경우, 자신도 모르게 연구 가설을 지지하는 방향으로 데이터를 해석하려는 무의식적 편향(unconscious bias)이 발생할 수 있다. 예를 들어, 처리군 샘플에서는 미세한 변화도 유의미한 신호로 해석하려 하고, 대조군에서는 동일한 변화를 배경 잡음(noise)으로 치부할 가능성이 존재한다. 이러한 편향을 방지하기 위해 최소한 데이터 파일의 라벨을 암호화하여(예: Sample A, Sample B) 어떤 처리가 되었는지 모르는 상태에서 분석을 수행하는 단일 맹검(single blind)을 적용해야 한다. 이상적으로는 실험 수행자와 데이터 분석자를 분리하거나, 제3자가 라벨링을 변경하여 전달하는 이중 맹검(double blind) 방식을 도입해야 한다. 특히 소규모 실험일수록 연구자 개인의 기대가 결과에 미치는 영향력이 상대적으로 커질 수 있으므로, 가능한 모든 단계에서 엄격한 블라인딩을 도입하여 데이터의 객관성을 확보해야 한다.

3. 반복의 정의와 배치 효과의 통제

실험의 재현성(reproducibility)과 신뢰도(reliability)를 확보하기 위해 반복(replication)은 필수적인 요소이다. 그러나 단순히 표본 수(n)를 늘리는 것만이 능사가 아니며, 그 반복이 통계적으로 어떤 의미를 갖는지 명확히 정의하는 것이 선행되어야 한다.

1) 반복수의 정의: 기술적 반복 대 생물학적 반복

반복수는 '통계적 독립성이 확보된 반복 단위'의 개수로 정의되어야 한다. 이를 혼동하여 발생하는 가장 치명적인 통계적 오류가 바로 의사반복(pseudoreplication)이다. 동일한 샘플(예: 한 마리 쥐의 혈액, 하나의 세포 배양 디시)에서 시료를 여러 번 채취하여 측정한 것은 기술적 반복(technical replicate)에 해당한다. 이는 파이펫팅 오차나 기기 측정 오차를 줄여 측정값의 정밀도(precision)를 높이는 데 기여하지만, 생물학적 변동성을 대변하지 못하므로 독립적인 표본수(n)를 증가시키는 근거가 될 수 없다. 만약 기술적 반복을 n=3으로 간주하여 통계 분석을 수행할 경우, 자유도(degrees of freedom)가 인위적으로 부풀려져 표준오차(standard error)가 과소추정되고, 결과적으로 *p*-값이 실제보다 낮게 산출되는(유의하게 나오는) 제1종 오류(Type I Error)를 범하게 된다(9, 10).

연구 결론의 일반화 가능성은 오직 생물학적 반복(biological replicate)에 의해 정당화된다. 생물학적 반복은 서로 독립적

으로 준비된 샘플 또는 개체를 의미한다. 세포 실험에서 하나의 모(母) 배양 플라스크에서 얻은 세포를 96웰 플레이트의 3개 웰에 나누어 분주한 것은 동일한 생물학적 기원을 가지므로 생물학적 반복이 아닌 기술적 반복에 불과하다. 진정한 생물학적 반복을 확보하기 위해서는, 서로 다른 날에 동결 보존된 세포를 새롭게 해동하여 배양을 시작하거나(independent culture batch), 완전히 독립적으로 준비된 배양 배치를 반복 단위로 설정해야 한다. 동물 실험의 경우 서로 다른 개체가 생물학적 반복 단위가 된다. 논문 작성 시 n수를 제시할 때는 이것이 기술적 반복인지 생물학적 반복인지를 명확히 명시해야 하며, 모든 통계적 가설 검정은 반드시 생물학적 반복 수를 기준으로 수행되어야 한다.

2) 배치 효과(batch effects): 실험의 현실적 변동성 관리

이상적인 실험 환경과 달리, 현실의 연구는 시공간적 제약을 받는다. 따라서 실험은 불가피하게 여러 번에 나누어(batch) 수행되는데, 이 과정에서 발생하는 체계적 오차가 바로 배치 효과(batch effects)이다. 실험 수행 날짜에 따른 미세한 온습도 변화, 시약의 제조 로트(Lot) 변경에 따른 활성 차이, 실험 기기의 램프 강도 저하, 세포의 계대 수(passage number) 증가, 심지어 실험 수행자의 컨디션 변화까지 모든 요소가 배치 효과를 유발하는 변수가 된다(11).

가장 치명적인 실험 설계 오류는 배치 효과와 처리 효과(treatment effect)가 교락(confounding)되는 것이다. 예를 들어, '첫째 날은 대조군만 3번 실험하고, 둘째 날은 약물 처리군만 3번 실험'하는 방식이다. 이 경우 관찰된 데이터의 차이가 약물 처리에 의한 것인지, 아니면 날짜가 변경되면서 발생한 환경적 변화에 의한 것인지 분리해낼 수 없다. 따라서 '각 배치 내에 모든 처리군을 포함시키는 것'이 실험 설계의 철칙이다. 즉, 하루에 대조군과 처리군을 적어도 하나씩은 세트로 구성하여 실험해야 한다. 이를 완전 확률화 블록 설계(randomized complete block design)라고 하며, 통계 분석 시 배치를 '블록(block)' 변수로 모델에 포함하여 분석하면(예: Two-way ANOVA without interaction), 배치 간의 변동을 수학적으로 보정하여 순수한 처리 효과만을 더 정확하게 추정할 수 있다. 만약 물리적 한계로 배치가 분리될 수밖에 없다면, 최소한 각 처리군이 여러 배치에 균형적으로 분산되도록 균형 설계(balanced design)를 계획해야 한다(11, 12).

4. 표본 크기 산정과 데이터 품질 관리

1) 표본 크기(sample size) 산정: 관행이 아닌 연구 질문에 의한 결정

"n수는 관행적으로 3번이면 충분하다"는 인식은 통계적으로 타당하지 않다. 표본 크기는 단순한 관행이 아니라, 연구자가 답하고자 하는 '연구 질문'과 '탐지하고자 하는 효과의 크기'에 의해 결정되어야 한다. 과학적으로 타당한 표본 크기를 산정하기 위해서는 실험 수행 전에 검정력 분석(power analysis)이 필수적으로 선행되어야 한다. 이를 위해서는 ① 유의수준(α, 통상 0.05), ② 목표 검정력(1-β, 통상 0.8 이상), ③ 데이터의 변동성(표준편차 추정치), 그리고 ④ 탐지하고자 하는 의미 있는 최소 효과 크기(effect size)를 정의해야 한다(13).

지나치게 작은 표본수는 실제 효과가 존재함에도 불구하고 통계적으로 이를 탐지하지 못하는 위음성(Type II Error)의 위험을 높여, 투입된 연구 자원과 시간을 낭비하게 만든다. 반대로 과도하게 큰 표본수는 생물학적으로 의미가 없는 미미한 차이(biological insignificance)를 통계적으로만 유의하게(statistical significance) 만들어 해석의 오류를 유발할 수 있으며, 불필요하게 많은 실험 동물을 희생시키는 윤리적 문제를 초래한다. 여기서 핵심은 '효과 크기'와 '유의성(*p*-value)'을 명확히 구분하는 것이다. *p*-값이 0.001이라 하여 해당 약물의 효능이 생물학적으로 매우 강력하다는 것을 의미하지는 않는다. 이는 단지 관찰된 결과가 귀무가설 하에서 우연히 발생할 확률이 매우 낮다는 것을 의미할 뿐이다. 효과의 크기(예: 세포 사멸률 10% 증가 대 50% 증가)가 생물학적으로 유의미한지는 연구자의 전문적 지식에 의해 판단되어야 하며, 표본 크기 산정은 이 '의미 있는 차이'를 통계적으로 입증하기 위해 필요한 최소한의 데이터 양을 계산하는 과정이다. 이는 실험 종료 후 사후 분석으로 보완할 수 없는 영역이므로, 반드시 실험 설계 단계에서 수행되어야 한다(14).

2) 데이터 품질 관리(QC): 설계의 완성

데이터 수집 단계에서의 품질 관리(QC)는 실험 디자인의 연장선상에 있다. 살아있는 생명체를 다루는 실험에서는 필연적으로 결측치(missing value)와 이상치(outlier)가 발생한다. 실험 동물이 마취 사고로 사망하거나, 세포 배양 플레이트가 미생물에 오염되거나, 기기 오류로 데이터가 손실되는 일은 연구 현장에서 빈번히 발생한다. 중요한 것은 이러한 비정상 데이터를 처리하는 원칙을 '사전에' 정의하고 연구 노트에 명문화하는 것이다. 데이터를 모두 확인한 후에 "이 값은 경향성에서 벗어나니 제외하자"라고 결정하는 것은 데이터 조작(data manipulation) 또는 P-해킹(P-hacking)의 범주에 속할 수 있다. 오염, 시료의 건조, 약물 투여 실패, 기기 에러 등 객관적이고 타당한 제외 기준(exclusion criteria)을 실험 전에 마련하고, 이를 연구 노트에 상세히 기록해야 한

다. 또한 이 기준은 대조군과 실험군에 편향 없이 동일하게 적용되어야 한다. 예를 들어, 약물 처리군에서만 세포 사멸이 많이 발생하여 데이터를 제외했다면, 이는 약물의 독성을 의도적으로 은폐하는 결과를 초래한다. 결측치 처리 방식 또한 결과의 타당성에 영향을 미친다. 무작정 평균값으로 대체하거나(mean imputation) 결측이 발생한 샘플 전체를 삭제하는 것(listwise deletion)은 편향을 유발할 수 있다. 결측이 발생한 메커니즘(완전 무작위 결측인지, 특정 조건에서만 발생하는 비무작위 결측인지)을 고려하여 적절한 통계적 처리 방법을 선택해야 하며, 이 모든 과정은 투명하게 기술되어야 한다. 철저한 데이터 품질 관리는 데이터의 노이즈(noise)를 줄이고 신호(signal)를 명확히 하여, 설계된 실험의 내적 타당성을 최종적으로 완성하는 단계이다.

3) 복합 설계와 데이터 구조의 반영: 상호작용과 계층성

생물학적 시스템은 단일 요인의 변화가 결과로 직결되는 선형적 장치가 아니라, 다수의 요인이 동시에 작동하며 서로의 효과를 조절하는 복합적 네트워크로 이해해야 한다. 유전적 배경, 환경 요인, 약물 농도와 노출 시간, 영양 상태, 스트레스 자극과 같은 변수들은 독립적으로 존재하는 것이 아니라 상호 의존적으로 얽혀 최종 표현형을 형성한다. 이러한 특성 때문에 생물학 연구에서 전통적으로 사용되던 단일 요인 순차 실험은 중요한 정보를 누락할 위험이 있으며, 실제 현상을 더 정확하게 포착하기 위해 다요인 설계가 필요해진다. 다요인 설계는 효율적으로 여러 요인의 효과를 동시에 검증할 수 있을 뿐 아니라, 요인 간 결합이 만들어내는 비선형적 변화, 즉 조건 의존적 효과를 직접 평가할 수 있다는 점에서 생물학적 기전 탐색에 적합하다. 다만 설계가 복합해질수록 데이터는 단순한 독립 관측치의 집합이 아니라, 상호작용과 계층적 생성 구조를 포함하는 형태로 생성되므로, 연구자는 실험 단계에서부터 분석 모델이 반영해야 할 데이터 구조를 명확히 정의해야 한다.

5. 복합 설계와 데이터 구조의 반영

1) 상호작용: 주효과 중심 해석의 한계

다요인 설계에서 흔히 발생하는 해석상의 문제는 각 요인의 주효과만을 근거로 결론을 서술하는 것이다. 주효과는 다른 요인의 수준을 평균화한 상태에서 특정 요인이 결과에 미치는 평균적 영향을 의미한다. 그러나 생물학적 시스템에서 두 요인이 결합할 때 나타나는 효과는 단순한 합으로 설명되지 않는 경우가 많다. 즉, 한 요인의 효과가 다른 요인의 수준에 따라 달라지는 현상이 빈번하며, 이를 통계적으로 상호작용으로 정의한다. 상호작용이 존재하는 상황에서 주효과만을 해석하면, 실제로는 특정 조건에서만 나타나는 효과를 일반화하거나, 반대로 조건 의존적으로 크게 나타나는 변화를 평균 처리로 상쇄시켜 핵심 기전 정보를 놓칠 수 있다. 예를 들어 항암제 처리 여부와 특정 유전자의 과발현 여부를 동시에 조작하고 세포 사멸을 관찰한다고 가정하자. 만약 해당 유전자가 항암제의 대사를 억제하거나 세포 내 축적을 증가시키는 기전을 가진다면, 항암제의 효과는 유전자 상태에 의해 증폭될 수 있다. 이 경우 항암제의 효과는 유전자가 과발현된 조건에서만 뚜렷하게 나타나고, 유전자가 과발현되지 않은 조건에서는 미미할 수 있다. 이러한 관계는 항암제의 효과가 유전자 조건에 의존한다는 의미이며, 상호작용이 존재하는 전형적인 형태이다. 그럼에도 주효과만을 보고 항암제의 평균 효과를 제시하면, 조건에 따라 효과가 크게 달라지는 사실이 희석되어 생물학적 해석의 타당성이 약화된다. 또한 유전자의 주효과가 평균적으로 유의하지 않다고 결론 내릴 경우, 실제로는 항암제 존재하에서만 유전자의 영향이 나타나는 중요한 기전적 단서를 간과하게 된다. 따라서 복합 설계에서는 요인 간 구조를 반영하는 모형을 적용해야 하며, 대표적으로 이원 분산분석 또는 일반화된 선형모형에서 상호작용 항을 포함한 형태로 분석한다. 상호작용 항이 유의한 경우에는 주효과에 대한 해석을 우선적으로 확정하기보다, 효과가 어떤 조건에서 나타나는지 구체화하는 절차가 필요하다. 이를 위해 단순 주효과 분석 또는 사후 비교를 통해 각 요인 수준 조합에서의 차이를 평가하며, 결과 서술 역시 조건을 명시하는 방식으로 제한하여 기술해야 한다. 예컨대 항암제의 효과가 특정 유전자 과발현 조건에서만 유의하다는 형태로 결론을 제시하면, 효과의 존재 여부뿐 아니라 효과가 발현되는 맥락까지 함께 제시할 수 있다. 이러한 접근은 복잡한 생물학적 현상을 과도하게 단순화하지 않으면서, 관찰된 데이터가 지지하는 범위 내에서 해석을 제시한다는 점에서 과학적 엄밀성을 높인다.

2) 계층 및 상관 구조: 독립성 가정의 위배와 모델의 선택

통계적 추론의 많은 절차는 관측치 간 독립성을 주요 가정으로 둔다. 그러나 실제 생물학 실험에서 생성되는 데이터는 실험 단위와 관측 단위가 일치하지 않는 경우가 많고, 그 결과 데이터는 계층 구조 또는 상관 구조를 내재한다. 이러한 구조를 무시하고 독립성을 가정한 분석을 적용하면, 표준오차가 과소추정되고 검정이 과도하게 유의하게 나타나는 문제가 발생할 수 있다. 특히 세포 실험과 동물 실험은 대표적으로 내포 설계와 반복 측정을 포함하는 경우가 많아, 분석 단계에서 데이터의 생성 구조를 정확히 반영하는

것이 필수적이다.

세포 기반 실험은 전형적인 계층 구조를 갖는다. 다수의 세포가 하나의 웰에 배치되고, 여러 웰이 하나의 플레이트에 포함되며, 플레이트는 실험 일자나 배치에 의해 묶인다. 같은 웰에 존재하는 세포는 동일한 배지, 동일한 처리 조건, 유사한 미세환경을 공유하므로 서로 독립 표본으로 보기 어렵다. 그런데도 개별 세포를 독립 표본으로 간주해 분석하면, 실제로는 웰 또는 플레이트 수준에서 발생하는 변동을 표본 수 증가로 착각하게 되어 유효 표본 크기가 과대평가된다. 이는 표준오차를 인위적으로 감소시키고, 결과적으로 p값이 실제보다 작게 산출되는 경향을 만들 수 있다. 따라서 이와 같은 내포 구조에서는 분석의 단위를 명확히 설정해야 하며, 웰 평균을 사용해 독립 단위를 웰로 정의하거나, 더 일반적으로는 계층 구조를 무작위 효과로 반영하는 혼합 모형을 적용하는 것이 적절하다.

반복 측정 또한 독립성 가정을 흔히 위배한다. 동일한 개체에서 시간에 따라 반복적으로 측정된 혈당, 동일한 세포군에서 시간 경과에 따라 측정된 신호 강도, 동일 개체에서 여러 조건을 교차 적용한 반응값은 측정값 간에 강한 상관이 존재한다. 이는 개인 내 변동이 개인 간 변동보다 작고, 연속 시점의 측정값이 서로 유사하게 움직이는 경향을 의미한다. 이런 상황에서 단순한 t검정이나 일반 분산분석을 그대로 적용하면, 상관으로 인해 유효 정보량이 줄어드는 사실이 반영되지 않아 표준오차가 과소추정될 수 있다. 반복측정 분산분석은 이러한 구조를 일부 반영할 수 있으나, 결측이 존재하거나 측정 간격이 불규칙한 경우, 또는 개인별 기저 수준이 크게 다른 경우에는 적용이 제한될 수 있다.

이러한 이유로 생물학 연구에서는 고정 효과와 무작위 효과를 함께 포함하는 선형 혼합 모형이 중요한 대안이 된다. 혼합 모형은 처리 조건, 유전자 조작, 시간과 같은 실험자가 관심을 두는 요인을 고정 효과로 모델링하면서, 배치, 플레이트, 웰, 개체와 같은 군집 또는 반복 단위를 무작위 효과로 포함해 상관과 계층 구조를 동시에 반영할 수 있다. 이를 통해 처리 효과에 대한 추정은 계층적 변동을 고려한 형태로 산출되며, 결과적으로 추정의 불확실성이 보다 현실적인 수준으로 평가된다. 또한 혼합 모형은 불균형 자료나 일부 결측이 있는 자료에도 비교적 유연하게 적용 가능하여 실제 실험 데이터의 특성에 부합한다(15, 16).

결국 복합 설계의 핵심은 요인을 많이 넣는 것 자체가 아니라, 요인 간 상호작용과 데이터의 생성 구조를 분석 모형에 정합적으로 포함시키는 데 있다. 상호작용을 고려하지 않으면 효과가 나타나는 조건을 규명할 수 없고, 계층 및 상관 구조를 무시하면 유의성 판단이 왜곡될 수 있다. 따라서 실험 설계 단계에서부터 실험 단위와 관측 단위를 구분하고, 요인 간 결합이 의미를 가질 가능성을 사전에 가정하며, 분석 단계에서는 해당 구조를 반영할 수 있는 적절한 모형을 선택해야 한다. 이러한 접근은 단순한 통계적 유의성의 확보를 넘어, 생물학적 기전 해석의 정확성과 재현성 향상에 직접적으로 기여한다.

6. Student t-test와 one-way ANOVA: 등분산 중심의 평균 비교와 사후검정 선택, 가족오류율 포함

생물학 실험에서 통계적 추론의 가장 기본 과제는 서로 다른 조건에서 관측된 연속형 결과변수의 평균 차이를 검정하는 것이다. 처리 농도, 노출 시간, 유전자형, 배양 조건처럼 하나의 요인이 여러 수준을 가질 때, 평균 비교는 연구 결론의 중심 근거가 된다. 비교 집단 수가 두 개인지, 세 개 이상인지에 따라 표준 절차가 달라지며, 특히 다수 비교가 포함될 때는 제1종 오류의 누적을 관리하는 통제 전략이 필수적이다. 실험실 기반 데이터는 동일 장비, 동일 프로토콜, 동일 분석 파이프라인에서 생성되는 경우가 많아 등분산 가정이 실무적으로 자주 타당해진다. 본 절은 등분산을 기본값으로 설정한 상태에서 Student t-test와 one-way ANOVA의 연결, 그리고 사후검정으로서 Tukey HSD, Dunnett, Duncan, Fisher's LSD, Bonferroni, Scheffé, Games-Howell의 선택 논리를 체계적으로 정리한다. 본문에서는 따옴표 기호를 사용하지 않는다.

1) 두 집단 비교의 출발점: Student t-test

두 집단 평균 비교의 표준 도구는 Student t-test이다. 귀무가설은 두 집단의 모평균이 동일하다는 것이며, 관측된 평균 차이를 표준오차로 정규화하여 유의확률을 계산한다. 적용 전제는 관측의 독립성, 잔차의 근사적 정규성, 그리고 등분산성이다. 생물학 실험에서 등분산이 비교적 잘 성립하는 이유는 데이터 생성 과정이 표준화되어 있기 때문이다. 동일한 세포주, 동일 배지, 동일 배양 조건, 동일 장비 설정, 동일 정량 알고리즘이 반복될수록 오차의 발생 기전이 집단 간 유사해지고 분산 구조가 안정된다. 또한 검출한계 근처의 저신호 영역이나 포화 영역을 피하도록 조건을 설계하는 관행은 평균 수준에 따라 분산이 급변하는 구간을 줄여 등분산성을 강화한다(17).

다만 t-test에서 가장 취약한 전제는 독립성이다. 동일 개체에서 반복 측정된 값이나 동일 배치에서 파생된 기술 반복을 독립 표본처럼 취급하면 표준오차가 과소추정되어 p값이 과도하게 작아질 수 있다. 따라서 독립 단위를 개체,

독립 배양, 독립 실험일 등으로 명확히 정의하고, 비교는 그 단위에서 수행되어야 한다. 또한 *p*값만 제시하는 보고는 생물학적 해석에 충분하지 않다. 평균 차이, 신뢰구간, 표준화 효과크기(Cohen d 등)를 함께 보고하면 차이의 크기와 불확실성이 동시에 제시되어 기전적 논의와 연결이 용이해진다.

2) 다중 비교와 가족오류율: 왜 사후검정이 필요한가

세 집단 이상에서 모든 쌍에 대해 *t*-test를 반복 수행하면 다중 비교 문제가 발생한다. 이 문제의 본질은 검정 횟수가 증가할수록 제1종 오류가 누적되어 거짓 양성이 포함될 확률이 높아진다는 점이다. 이를 연구 질문 단위에서 정량화하는 대표 지표가 가족오류율(family-wise error rate, FWER)이다. 가족오류율은 한 분석 단위에서 수행되는 여러 가설 검정 집합 가운데, 적어도 한 번이라도 제1종 오류가 발생할 확률을 의미한다. 개별 비교마다 유의수준 0.05를 유지하더라도, 비교가 여러 번 수행되면 전체 비교 집합에서 최소 1개의 거짓 양성이 포함될 위험이 커진다.

독립 검정이라는 단순 가정에서 비교 횟수를 m, 각 검정의 유의수준을 α로 두면, 적어도 한 번 거짓 양성이 발생할 확률은 $1 - (1 - \alpha)^m$으로 증가한다. 예를 들어 α가 0.05이고 비교가 10회면 $1 - 0.95^{10}$은 약 0.40 수준이며, 비교가 20회면 이 확률은 더 커진다. 실제 데이터에서는 비교들이 완전히 독립이 아닐 수 있으나, 비교 횟수 증가가 오류 누적 위험을 높인다는 구조적 결론은 유지된다. 생물학 실험에서는 처리 수준이 많거나, 서로 다른 시간대와 농도 조합이 동시에 평가되거나, 여러 지표가 함께 분석되는 경우가 흔하므로, 가족오류율을 통제하지 않은 결과는 재현성 저하와 과도한 결론 도출로 이어질 수 있다(18). 가족오류율 통제는 보수적 목표이기도 하다. 거짓 양성을 강하게 억제하는 대신 일부 진짜 차이를 놓칠 가능성, 즉 검정력 저하가 발생할 수 있다. 따라서 어떤 수준의 오류 통제가 필요한지는 연구 단계와 연구 질문에 의존한다. 그러나 다수의 처리군 비교 결과를 논문 결론으로 제시하는 일반적 상황에서는 최소한 가족오류율을 어떤 방식으로 관리했는지 명시하는 것이 설득력과 심사 친화성을 높인다.

3) 세 집단 이상 비교의 표준 관문: One-way ANOVA

요인 수준이 세 개 이상일 때 평균 차이를 평가하는 기본 모형은 one-way ANOVA이다. 귀무가설은 모든 집단 평균이 동일하다는 것이며, 전체 변동을 집단 간 변동과 집단 내 변동으로 분해한 뒤 그 비율을 F 통계량으로 평가한다. ANOVA의 유의성은 적어도 한 집단 평균이 다르다는 사실을 시사하지만, 어떤 집단 간 차인지까지는 알려주지 않는다. 따라서 ANOVA 이후에는 사후검정이 필요하다(19). 등분산 조건에서 ANOVA는 가장 안정적으로 작동한다. 집단 내 변동이 공통 분산 구조를 공유한다는 가정이 성립하면, 집단 내 평균제곱은 공통 오차분산의 추정치로 해석되며, 이후 사후검정의 임계값과 신뢰구간 산출도 일관되게 이루어진다. 실험실 기반 데이터에서 등분산이 흔한 이유는 *t*-test에서 언급한 표준화 조건이 그대로 유지되기 때문이다. 따라서 기본 전략은 등분산을 전제로 분석을 설계하되, 잔차 도표와 간단한 분산 점검에서 위반이 명확할 때만 예외적 대안을 적용하는 것이다.

ANOVA 결과 보고에서는 F값과 자유도, *p*값뿐 아니라 효과크기(η^2 또는 ω^2)를 함께 제시하는 것이 바람직하다. 효과크기는 요인이 결과 변수 변동의 어느 정도를 설명하는지 정량화하며, 통계적 유의성과 생물학적 의미의 연결을 강화한다. 또한 각 집단 평균과 신뢰구간을 제시하면 차이의 크기와 불확실성까지 제시되어 해석의 과학성이 높아진다.

4) 사후검정 선택의 1차 원리: 비교 목표를 먼저 고정

사후검정 선택은 방법의 유명도보다 비교 목표의 정의에 의해 결정되어야 한다(Table 1). 생물학 실험에서 비교 목표는 대체로 두 유형이다.첫째, 모든 쌍 비교가 필요한 경우다. 여러 처리군 간 관계를 전반적으로 기술하거나, 농도 단계 간 평균 차이를 모두 제시해야 하는 연구에서는 모든 집단 쌍을 동시에 추론해야 한다. 둘째, 대조군 대비 비교가 핵심인 경우다. Vehicle 또는 untreated 같은 대조군을 기준으로 각 처리군의 변화만 확인하면 충분한 경우가 많다. 이때 처리군끼리의 비교는 연구 질문의 중심이 아니다. 이 두 목표는 비교 집합 자체가 다르므로, 동일한 사후검정을 적용하는 것은 논리적으로 비효율적일 수 있다. 모든 쌍 비교 목표에는 Tukey 계열이, 대조군 대비 목표에는 Dunnett가 가장 정합적이다.

5) Tukey HSD와 Tukey Kramer: 모든 쌍 비교에서의 표준적 선택

Tukey HSD는 모든 집단 쌍의 평균차를 동시에 평가하면서 가족오류율을 통제하도록 설계된 대표적 사후검정이다. Tukey 계열의 핵심 전제는 표본수가 동일해야 한다는 조건이 아니라 등분산 가정이다. 다만 균형설계, 즉 각 집단 표본수가 동일한 경우에는 Tukey HSD의 유도와 계산이 가장 단순하고 성능도 안정적으로 나타나므로 전형적 적용 사례로 널리 소개된다. 표본수가 불균형한 경우에도 모든 쌍 비

Table 1. 일원분산분석 후 사후검정 방법의 선택 기준과 특성

방법	언제 쓰는가	비교 범위	등분산 가정	FWER 통제	장점	주의점	실험 예시
Tukey HSD	처리군이 여러 개이고 모든 처리군끼리 차이를 다 보고해야 할 때	모든 쌍 비교 (각 집단-각 집단)	필요	강하게 통제	모든 쌍 비교의 표준, 해석이 직관적, 과도한 거짓 양성 억제	표본수 불균형이면 일반 Tukey보다 Tukey-Kramer가 더 적절할 수 있음. 이분산이 크면 적합성 저하	4개 농도(0, 1, 5, 10)에서 모든 농도 간 차이를 보고
Tukey Kramer	표본수가 집단마다 다르지만 모든 쌍 비교가 필요한 경우	모든 쌍 비교	필요	강하게 통제	Tukey를 불균형 표본수에 맞게 확장한 표준 방법	등분산이 핵심 전제. 이분산이 크면 다른 방법 고려	어떤 농도군은 n=5, 어떤 군은 n=8이지만 모든 농도 간 차이 비교
Dunnett	대조군 대비 효과만이 핵심일 때	대조군 vs 각 처리군	필요	강하게 통제	불필요한 비교를 줄여 검정력 유리, 생물학 실험 질문과 잘 맞음	처리군끼리의 차이는 직접 제공하지 않음	Vehicle 대조군과 3개 약물 처리군을 각각 비교
Fisher's LSD	비교 쌍이 아주 적고 사실상 계획된 비교에 가깝거나 집단 수가 작을 때만 제한적으로	보통 모든 쌍 비교처럼 사용되기도 함	필요	약하거나 상황 의존	계산과 적용이 단순, 민감하게 차이를 잡아냄	비교 수가 늘면 거짓 양성 급증 가능. 심사에서 문제 될 수 있어 근거 제시 필요	집단이 3개이고 비교 목표가 사실상 1~2개 쌍으로 제한된 경우
Duncan	결과를 군집화(letter)로 쉽게 보여주고 싶을 때 관행적으로 사용되기도 함	단계적 다중비교 (군집화)	필요	약한 편 (덜 보수적)	그룹핑 표현이 쉬움	거짓 양성 위험이 상대적으로 큼. 엄격한 연구에서는 권장 어려움	여러 처리군을 letters로 묶어 차이 그룹을 시각적으로 제시
Bonferroni	비교가 소수이고 미리 정해진 가설 비교를 수행할 때	사용자가 정한 비교만	조건부 (대체로 필요)	매우 강하게 통제(보수적)	가장 이해하기 쉬움, 심사에서 설명이 명확	비교가 많으면 너무 보수적이라 유의성 놓칠 수 있음	사전에 정한 3개 비교만 수행(대조군 vs 고농도, WT vs KO 등)
Scheffé	단순 쌍 비교가 아니라 다양한 선형 대비(contrast)까지 폭넓게 보호해야 할 때	모든 대비 (쌍 비교 포함)	필요	매우 강하게 통제	대비가 많고 탐색적 비교가 넓을 때 안전	매우 보수적이라 검정력 낮음. 표본수 작으면 차이를 잡기 어려움	여러 처리군의 조합 대비(예: 저농도 평균 vs 고농도 평균)까지 고려
Games-Howell	이분산이 명확하거나 표본수 불균형과 분산 차이가 함께 큰 경우	모든 쌍 비교	불필요	비교적 안정적으로 통제(근사)	등분산이 깨진 상황에서 실무적으로 강건	등분산이 대체로 성립하면 굳이 기본값으로 쓸 필요는 적음. 구현 옵션 확인	고농도군에서 변동이 커져 등분산이 명확히 깨진 상태에서 모든 농도 비교

교는 가능하며, 이때는 표본수 차이를 반영하도록 일반화된 Tukey Kramer 방식이 표준적으로 사용된다. 실제 실무에서는 소프트웨어가 Tukey로만 표기하더라도 내부적으로 Tukey Kramer를 적용하는 경우가 있어, 출력과 옵션을 확인하는 것이 안전하다(20, 21).

Tukey 계열의 장점은 비교 쌍이 많아도 비교 집합 전체에서 거짓 양성이 최소 한 번 발생할 확률을 통제한다는 점이다. 이는 다수 처리군을 포함하는 생물학 실험에서 결과의 신뢰도를 유지하는 데 유리하다. 또한 평균차 신뢰구간을 함께 산출하면 차이의 존재 여부뿐 아니라 차이의 크기와 불확실성을 동시에 제시할 수 있어, 효과의 생물학적 의미를 정량적으로 해석하는 데 도움이 된다. 등분산이 타당하고 모든 쌍 비교가 연구 질문의 중심이면 Tukey 계열이 기본값으로 기능한다(20).

6) Dunnett: 대조군 대비 비교의 최적화된 선택

Dunnett 검정은 대조군과 각 처리군의 비교에 특화되어 있다. 비교 집합을 대조군 대비로 제한한 상태에서 가족오류율을 통제하므로, 불필요한 비교를 줄여 검정력을 상대적으로 확보하는 데 유리하다. 생물학 실험에서 가장 흔한 질문 구조가 대조군 대비 변화이므로, 대조군이 명확하고 연구 결론이 대조군 대비 효과에 기반한다면 Dunnett는 방법 선택의 근거가 매우 강하다. 반대로 처리군 상호 간 비교가 결론의 핵심이라면 Dunnett만으로 충분하지 않으며, 목적에 따라 Tukey 계열과의 병용 또는 연구 질문의 재정의가 필요하다(22).

7) Bonferroni: 단순하고 강한 가족오류율 통제, 계획된 비교에 적합

Bonferroni 보정은 유의수준을 비교 횟수로 조정하여 가족오류율을 통제하는 가장 직관적인 방법이다. 장점은 단순성과 강한 보수성이다. 특히 비교가 소수의 계획된 가설로 제한될 때 적합하다. 예를 들어 세 집단 이상이 존재하더라도 연구 질문이 특정 두세 쌍 비교로 명확히 제한된다면, Bonferroni는 방법 선택과 보고가 명료하다. 그러나 비교 횟수가 많아질수록 매우 보수적으로 작동해 검정력이 감소할 수 있으므로, 모든 쌍 비교 목적에는 Bonferroni보다 Tukey 계열이 더 자연스럽고 효율적인 경우가 많다(23).

8) Scheffé: 모든 대비를 폭넓게 보호하는 매우 보수적 절차

Scheffé 방법은 단순한 쌍 비교뿐 아니라 가능한 모든 선형 대비에 대해 가족오류율을 통제하도록 설계된 절차다. 이는 사후적으로 다양한 대비를 고려해야 하는 연구에서 안전성을 제공한다. 그러나 그 대가로 보수성이 매우 크며, 표본수 제약이 흔한 생물학 실험에서는 검정력이 낮아질 수 있다. 따라서 Scheffé는 대비의 범위가 넓고, 여러 형태의 대비를 폭넓게 보호해야 한다는 연구 목적이 명확할 때 선택하는 것이 합리적이다(24).

9) Fisher's LSD: 제한적 조건에서만 신중한 적용이 필요

Fisher's LSD는 전통적으로 ANOVA가 유의하다는 전제하에서 개별 쌍 비교를 수행하는 방식으로 알려져 있다. 그러나 집단 수가 늘거나 비교가 많아질수록 가족오류율 통제가 충분하지 않을 수 있어 거짓 양성 위험이 커진다. 따라서 Fisher's LSD는 비교 쌍이 매우 제한적이며 사실상 계획된 비교에 가까운 상황에서만 보조적으로 고려하는 것이 안전하다. 다수 처리군을 포함한 분석에서 Fisher's LSD를 일반적 사후검정처럼 적용하면 통계적 엄밀성에 대한 비판 가능성이 커질 수 있다(1).

10) Duncan: 결과 제시는 용이하나 거짓 양성 위험을 고려해야 함

Duncan 다중범위검정은 처리군을 군집화하여 결과를 시각적으로 제시하기 쉬운 장점이 있다. 그러나 오류 통제 관점에서 덜 보수적으로 작동하는 경향이 있어, 비교가 많을수록 거짓 양성이 포함될 위험이 증가할 수 있다. 심사 환경에서는 Duncan보다 Tukey, Dunnett, Bonferroni, Scheffé 같은 명확한 가족오류율 통제 절차를 요구하는 경우가 많으므로, Duncan을 사용할 때는 선택 근거와 해석의 제한을 명확히 기술해야 한다(25).

11) 이분산이 확인되는 예외 상황과 Games-Howell의 제한적 역할

등분산이 실험실 데이터에서 흔하더라도, 특정 조건에서는 이분산이 관찰될 수 있다. 예를 들어 강한 처리에서 반응이 양극화되어 변동이 커지는 경우, 독성으로 일부 샘플이 불안정해지는 경우, 검출한계 또는 포화 영역에서 상대오차가 달라지는 경우, 표본수 불균형과 분산 차이가 동시에 큰 경우가 이에 해당한다. 이러한 상황에서 등분산을 전제로 한 사후검정을 그대로 적용하면 오류율이 왜곡될 수 있으므로, 이분산을 허용하는 사후검정이 필요해진다. Games-Howell은 등분산을 가정하지 않으면서도 다중 쌍 비교를 수행하는 방법으로 실무적 대안이 된다. 다만 이는 등분산 중심의 표준 흐름에서 벗어나는 예외적 선택지이며, 적용 근거는 가정 점검 결과로 명확히 제시되어야 한다(26).

12) 종합적 권고: 등분산 기본값에서의 선택 규칙

등분산이 타당한 생물학 실험 환경에서 사후검정 선택은 비교 목표에 따라 단순화할 수 있다. 모든 쌍 비교가 핵심이면 Tukey HSD 또는 표본수 불균형 시 Tukey Kramer가 기본값이다. 대조군 대비가 핵심이면 Dunnett가 1차 선택이다. 비교가 소수의 계획된 가설로 제한되면 Bonferroni가 명료하다. 다양한 대비를 폭넓게 보호해야 한다면 Scheffé를 고려하되 검정력 저하를 감수해야 한다. Fisher's LSD와 Duncan은 거짓 양성 위험을 고려하여 제한적 조건에서만 신중히 사용해야 한다. 이분산이 명확히 확인될 때만 Games-Howell을 예외적 대안으로 적용한다.

이와 같은 정합적 선택 논리는 단순히 p값을 생산하는 절차가 아니라, 연구 질문 단위에서 오류를 통제하면서 해석

가능한 결론을 도출하기 위한 방법론적 설계로 기능한다. 특히 가족오류율에 대한 명시적 고려는 다수 처리군 비교 결과의 재현성과 신뢰도를 높이는 핵심 요소이며, 생물학 연구에서 통계적 근거가 기전적 주장과 결합되는 지점에서 필수적 논리 장치가 된다.

7. 고가 통계 소프트웨어에서 Python·R 기반 분석으로의 전환과 Google Colab 활용, 그리고 LLM이 낮춘 학습 장벽

생물학 연구 현장에서는 통계 분석이 실험 설계와 결과 해석의 핵심 단계로 자리 잡았지만, 실제 수행 방식은 오랫동안 상용 통계 소프트웨어에 크게 의존해 왔다. 이러한 도구들은 그래픽 사용자 인터페이스 기반의 조작 편의성을 제공하는 반면, 라이선스 비용이 높고 기관 단위의 유지 관리가 필요하며, 분석 과정의 세부 설정이 클릭 중심으로 분산되어 재현성과 감사 추적성이 약해질 수 있다. 특히 다중 비교, 혼합효과 모형, 대규모 오믹스 데이터의 반복 분석처럼 분석 파이프라인이 길어질수록, 동일한 절차를 정확히 반복하고 공유하는 능력이 연구 품질을 좌우하게 된다. 이 지점에서 Python과 R 기반의 오픈소스 통계 도구는 비용 절감 이상의 의미를 갖는다. 분석이 코드로 기록되기 때문에 전처리부터 검정, 사후검정, 시각화, 보고까지 전 과정을 동일한 규격으로 재실행할 수 있고, 연구실 내부 표준 운영 절차로 고정하기가 쉽다(27, 28).

Python과 R은 생물학 연구에서 가장 널리 사용되는 분석 생태계를 형성하고 있다. R은 통계 모델링과 실험 설계, 다중 비교, 생물통계 패키지의 축적이 두텁고, Python은 데이터 전처리, 자동화, 기계학습 및 파이프라인 운영에 강점을 가진다. 예를 들어 R에서는 분산분석과 사후검정, 선형모형, 일반화선형모형, 혼합모형을 비교적 자연스럽게 구현할 수 있으며, Python에서는 SciPy, statsmodels 등으로 고전 통계 검정을 수행하고, pandas 기반으로 반복 분석을 자동화하기가 용이하다. 중요한 점은 특정 언어의 우열이 아니라, 연구 질문에 맞는 분석 흐름을 코드 기반으로 고정하고 재현 가능하게 운용할 수 있다는 것이다. 결과적으로 상용 소프트웨어가 제공하던 기능의 상당 부분은 오픈소스 조합으로 대체 가능하며, 오히려 분석 자동화와 문서화 측면에서는 더 높은 확장성을 확보할 수 있다(29, 30).

이 전환을 실제로 가능하게 만드는 실행 환경으로 Google Colab이 유용하다. Colab은 웹 브라우저만으로 노트북 기반 실행 환경을 제공하며, 별도 설치 없이도 표준 Python 분석을 즉시 시작할 수 있다. 실험실에서 자주 발생하는 환경 문제, 예를 들어 운영체제 차이, 패키지 버전 충돌, 개인 컴퓨터 의존성은 코드 공유와 재현성의 주요 장애물인데, Colab은 클라우드 런타임을 통해 이러한 변동을 크게 줄인다. 또한 Google Drive와 연동하여 원자료, 전처리 산출물, 그림 파일, 결과 테이블을 체계적으로 저장할 수 있고, 노트북 자체가 분석 기록이자 공유 가능한 실험 노트로 기능한다. R의 경우 기본 런타임은 Python이지만, 추가 설정을 통해 R 실행을 구성할 수 있으며, 이를 통해 동일한 협업 환경에서 R 기반 통계까지 통합 운용하는 전략도 가능하다(31).

Colab 기반 분석 흐름의 핵심 장점은 분석을 실행하고 저장하는 과정이 하나의 구조로 결합된다는 점이다. 일반적으로 다음의 단계로 표준화할 수 있다. 첫째, 데이터 입력 단계에서 원자료 파일을 Drive에 저장하고, 노트북에서 이를 불러오는 경로를 명시한다. 둘째, 전처리 단계에서 결측 처리, 이상치 규칙, 정규화 또는 변환을 코드로 기록해 동일 규칙이 반복되게 한다. 셋째, 통계 분석 단계에서는 선택한 검정과 가정 점검 절차를 코드로 고정한다. 넷째, 결과 보고 단계에서 그림과 표를 파일로 저장하고, 최종 결과를 CSV, XLSX, PDF, HTML 등으로 내보낸다. 이때 노트북은 분석 내역을 시간 순서로 보존하며, 어떤 설정으로 어떤 결과가 나왔는지를 추적하는 감사 추적성을 확보한다. 상용 GUI 중심 분석에서는 클릭 경로가 기록되지 않거나, 동일 분석을 반복할 때 설정 차이가 은밀하게 발생하기 쉽지만, 코드 기반 환경에서는 동일 입력과 동일 코드가 동일 출력을 생성하는 구조가 기본값이 된다(32, 33).

그러나 Python과 R 기반 분석이 실험 생물학자에게 항상 쉽게 느껴졌던 것은 아니다. 과거에는 통계적 개념을 이해하는 것과 별개로, 프로그래밍 문법, 데이터 구조, 패키지 의존성, 오류 메시지 해석이라는 추가 장벽이 존재했다. 특히 분석이 간헐적으로 필요한 연구자에게는 학습 투자 대비 체감 효용이 낮아 보일 수 있었고, 이로 인해 고가 상용 소프트웨어의 클릭 기반 워크플로가 선호되기도 했다. 최근에는 대규모 언어모델의 확산으로 이 장벽이 실질적으로 낮아지고 있다. ChatGPT를 포함한 LLM은 사용자가 가진 통계적 의도를 코드로 변환하는 과정에서 보조 도구로 기능할 수 있다. 예를 들어 실험 설계와 데이터 구조를 설명하면, 필요한 전처리 코드, 적절한 검정 선택의 분기, 사후검정 옵션, 결과 시각화 템플릿을 빠르게 제시할 수 있다. 또한 오류가 발생했을 때 에러 메시지의 의미를 해석하고 수정 방향을 제안함으로써 디버깅 시간을 단축한다. 결과적으로 프로그래밍을 전업 역량으로 갖추지 않은 연구자도, 분석을 수행하는 데 필요한 최소 수준의 코딩을 현실적으로 확보할 수 있게 된다(34, 35).

다만 LLM의 도움은 학습을 대체하는 것이 아니라 학습의 비용을 줄이는 방향으로 이해하는 것이 적절하다. 통계 분석은 선택한 검정의 전제, 데이터 생성 구조, 다중 비교 통제 목표, 효과 크기와 신뢰구간 해석을 포함한 논리 체계 위에서 정당화된다. LLM이 제공하는 코드는 유용한 출발점이 될 수 있지만, 그대로 복사해 실행하는 방식은 분석의 타당성을 보장하지 않는다. 따라서 다음의 원칙이 권장된다. 분석 목적과 가설, 독립 단위 정의, 반복 구조를 먼저 명시하고, 그에 맞는 모델을 선택한다. 가정 점검 결과와 예외 처리 기준을 코드에 포함한다. 패키지 버전과 난수 시드, 입력 데이터 버전을 기록한다. 핵심 결과는 다른 방법이나 다른 패키지로 교차 확인하여 견고성을 점검한다. 이러한 절차를 갖추면 LLM은 통계적 판단을 대신하는 존재가 아니라, 연구자가 내린 판단을 구현하고 문서화하는 생산성 도구로 자리 잡는다.

Colab 환경에서의 저장과 공유는 협업과 교육 측면에서도 이점이 크다. 연구실 내부에서 분석 템플릿 노트북을 표준으로 만들어 두면, 신규 구성원은 동일 구조를 재사용하면서 실험별 차이만 최소 수정하여 분석을 수행할 수 있다. 또한 노트북에 분석 목적, 전처리 규칙, 검정 선택 이유를 서술형으로 함께 기록하면, 방법 절 작성과 내부 재검토가 동시에 쉬워진다. 교육 관점에서도 통계 개념을 설명한 뒤 곧바로 동일 노트북에서 예제 데이터를 실행해 보는 방식은 이해를 빠르게 한다. 상용 소프트웨어에서 기능을 메뉴 위치로 암기해야 했던 학습은, 코드 기반 환경에서는 논리 구조와 데이터 구조를 중심으로 학습이 재구성된다.

정리하면, 고가 상용 통계 소프트웨어 중심 워크플로는 비용과 재현성의 한계를 갖는 반면, Python과 R 기반 오픈소스 도구는 코드 기반 재현성과 자동화를 통해 생물학 연구의 요구에 더 잘 부합할 수 있다. Google Colab은 설치와 환경 관리 부담을 낮추고, 실행과 저장, 공유를 하나의 플랫폼에서 통합해 이러한 전환을 현실화한다. 과거에는 프로그래밍 장벽이 전환의 핵심 장애였으나, ChatGPT 같은 LLM이 의도에서 코드로의 변환, 오류 해결, 템플릿 제공을 지원하면서 학습 부담이 크게 감소했다. 엄밀한 통계적 판단과 검증 책임은 연구자에게 남아 있지만, 구현과 문서화의 난이도는 분명히 낮아졌다. 따라서 생물학 연구에서도 코드 기반 분석을 표준 운영 체계로 받아들이는 것이 충분히 가능하며, 이는 단순한 도구 교체가 아니라 재현성과 투명성을 강화하는 연구 방법론의 진화로 이해될 수 있다.

8. 결과의 해석: 유의성을 넘어

통계적 유의성은 연구 결과를 평가하는 데 유용한 지표이지만, 과학적 결론을 구성하는 단일 기준이 될 수는 없다. *p* 값은 귀무가설이 참이라는 가정 하에서 관측된 통계량과 같거나 더 극단적인 값을 얻을 확률을 의미하며, 데이터가 우연 변동만으로 설명되기 어려운 수준인지 판단하는 도구로 기능한다. 그러나 *p*값은 효과의 크기나 방향, 생물학적 중요도, 그리고 연구 결과가 다른 조건과 집단에서도 유지될 가능성에 대한 정보를 직접 제공하지 않는다. 생명과학 데이터는 변동성이 크고 생성 구조가 복잡하며, 측정 오차와 배치 효과가 빈번하게 개입하므로, 유의성 중심의 이분법적 해석은 실제 현상을 과장하거나 반대로 중요한 신호를 조기에 배제하는 오류를 낳기 쉽다. 결과 해석의 목적은 유의성 자체를 확인하는 데 그치지 않고, 효과의 크기와 불확실성을 정량화하여 결론의 범위와 강도를 규정하는 데 있어야 한다(36).

*p*값 해석에서 가장 구조적인 문제는 표본수와의 강한 결합이다. 표본수가 증가하면 표준오차가 감소하여, 매우 작은 차이라도 통계적으로 유의해질 수 있다. 이 경우 효과가 검출되었다는 사실만으로 생물학적 의미를 과장하는 결론이 형성될 수 있다. 반대로 표본수가 제한적인 실험에서는 실제로 의미 있는 효과가 존재하더라도 변동성에 의해 유의성을 확보하지 못하는 사례가 흔하다. 이러한 상황에서 유의하지 않음을 효과 부재로 단정하면, 검정력 부족으로 인한 위음성 결과를 근거로 기전적 가능성을 배제하는 오류가 발생한다. 따라서 *p*값은 효과의 존재 여부를 판정하는 단일 관문으로 취급되기보다, 효과 추정의 불확실성과 결합해 해석되어야 한다. 특히 연구 질문이 단지 차이가 있는가에 머무르지 않고 어느 정도의 차이가 있는가, 그 차이가 의미 있는가, 어떤 조건에서 재현되는가로 확장될 때 *p*값만으로는 결론을 정당화할 수 없다.

이러한 한계를 보완하는 핵심 개념이 효과 크기와 신뢰구간이다. 효과 크기는 집단 간 차이 또는 변수 간 관계의 크기를 정량화한다. 평균 차이, 중앙값 차이, 비율 차이, 회귀계수, 오즈비, 위험비, 상관계수, 표준화된 평균 차이 등은 모두 효과 크기의 형태이며, 적절한 지표는 데이터 유형과 연구 질문에 의해 결정된다. 표준화 효과 크기는 서로 다른 척도의 결과를 비교하거나 문헌 간 결과를 통합할 때 유용하지만, 생물학적 의미를 직접 해석해야 하는 상황에서는 원 단위 효과 크기(예: 농도 단위 증가량, 생존일수 변화, 발현량 변화)가 더 직접적인 정보를 제공한다. 효과 크기를 보고할 때 중요한 점은 통계적 유의성과의 분리다. 유의한 효과가 항상 큰 효과를 의미하지 않으며, 유의하지 않은 결과가 효과가 없음을 의미하지도 않는다. 효과 크기는 연구가 목표로 하는 최소 의미 효과 크기와 연결되어 평가되어야

하며, 그 기준은 생물학적 기능 변화, 임상적 유용성, 공정 성능, 비용 대비 효익 같은 실질적 판단 기준과 결부될 필요가 있다(13, 37).

신뢰구간은 효과 크기 추정치의 불확실성을 제공한다. 예를 들어 평균 차이에 대한 95% 신뢰구간은 반복 표본 추출을 가정할 때 동일한 절차로 계산된 구간이 참값을 포함하는 비율이 95%가 되도록 구성된 범위를 의미한다. 신뢰구간이 좁다면 추정의 정밀도가 높고, 결과가 안정적으로 재현될 가능성이 상대적으로 높을 수 있다. 반대로 신뢰구간이 넓다면 표본수 부족, 높은 생물학적 변동성, 측정 오차, 배치 효과, 또는 모델 부적합을 시사하며, 유의성 여부와 무관하게 결론의 강도를 낮춰 해석하는 것이 과학적으로 타당하다. 특히 유의하지 않은 결과에서 신뢰구간은 해석을 크게 개선한다. 신뢰구간이 의미 있는 효과 크기 범위를 포함한다면, 유의하지 않다는 사실은 불확실성이 크다는 의미에 가깝다. 이때 결론은 효과 부재가 아니라 추가 데이터가 필요하다는 방향으로 구성되어야 한다. 반대로 신뢰구간이 실질적으로 의미 있는 범위를 대부분 배제한다면, 효과가 있더라도 크기가 작아 실용적 의미가 제한적일 가능성을 논리적으로 제시할 수 있다. 이런 방식은 결과를 이분법적으로 분류하지 않고 데이터가 허용하는 결론의 범위를 명시적으로 제한한다(38).

효과 크기와 신뢰구간 중심 해석은 후속 연구 설계의 정교화에도 직접적으로 기여한다. 관찰된 효과 크기와 변동성은 표본 크기 산정과 검정력 분석의 핵심 입력값이 된다. 또한 신뢰구간의 폭이 큰 결과는 반복 실험을 통해 분산을 줄이거나, 배치 효과를 설계적으로 통제하거나, 측정 방법의 정밀도를 개선해야 함을 시사한다. 특정 배치에서만 효과가 나타나거나 효과 방향이 배치마다 달라지는 경우는 배치-처리 상호작용 또는 숨겨진 교란 요인의 존재를 의심하게 하며, 단순한 유의성 보고보다 훨씬 생산적인 후속 질문을 생성한다. 여러 실험에서 효과 방향이 일관되지만, 개별 실험의 유의성이 불안정한 경우에는, 효과 크기와 신뢰구간을 기반으로 결과를 통합하는 메타분석적 관점이 유용하다. 이 접근은 연구 축적을 유의성의 획득 여부가 아니라 효과의 크기와 불확실성의 감소라는 관점에서 재구성한다.

결과 해석은 데이터 유형과 생성 구조를 반영한 모델 선택과 결합하여야 한다. 평균 비교는 연속형 데이터에서 직관적이지만, 이분형 결과(생존 여부, 양성 여부), 계수형 결과(CFU, 세포 수), 비율형 결과(퍼센트, 점유율), 시간경과 측정(반복측정), 계층 구조 데이터(플레이트 내 웰, 동물 개체 내 반복, 실험일/로트 단위 묶음)에서는 단순 평균 비교가 데이터 구조를 왜곡할 수 있다. 이 경우 회귀 기반 모델은 효과를 더 해석 가능한 형태로 제공하고 공변량을 포함해 교란 요인을 조정할 수 있다. 예를 들어 로지스틱 회귀는 이분형 결과에서 오즈비 형태의 효과 크기를 제공하며, 포아송 또는 음이항 회귀는 계수형 데이터의 발생률 비를 해석할 수 있게 만든다. 생존분석은 사건 발생 시간에 대한 위험비를 제공하여 시간 정보를 손실 없이 활용하게 한다. 혼합 효과 모형은 배치나 개체 수준의 랜덤 효과를 포함해 계층 구조와 상관 구조를 반영함으로써 표준오차의 과소추정을 방지한다. 이러한 모델링은 단순히 복잡도를 증가시키는 선택이 아니라, 데이터 생성 과정을 분석 단계에서 재현함으로써 결론의 타당성을 강화하는 선택으로 이해되어야 한다. 또한 회귀 기반 접근은 상호작용의 평가를 체계화한다. 생물학적 효과는 특정 조건에서만 나타나는 경우가 많다. 예컨대 약물 효과가 특정 시간점에서만 나타나거나, 스트레스 조건에서만 활성화되거나, 특정 유전형에서만 발현될 수 있다. 상호작용이 존재하는 데 주 효과만 보고 결론을 내리면, 중요한 조건의존적 기전 정보를 놓칠 가능성이 크다. 상호작용을 포함한 모델은 효과가 나타나는 조건을 명시적으로 제시할 수 있으며, 결과 해석을 평균적 결론이 아니라 조건부 결론으로 전환한다. 이는 재현성 향상에도 기여한다. 후속 연구자는 어떤 조건에서 효과가 관찰되는지 명확히 알 수 있고, 재현 실패가 단지 우연이 아니라 조건 차이에서 비롯되었는지 평가할 수 있다(15, 16).

보고 방식 역시 유의성을 넘어서는 해석을 실질적으로 뒷받침해야 한다. p 값, 효과 크기, 신뢰구간을 함께 제시하는 것은 최소 요건이며, 가능하다면 개별 관측치 분포를 보여주는 시각화를 통해 변동성과 표본수를 투명하게 드러내는 것이 바람직하다. 평균과 표준오차만 제시하는 방식은 데이터의 분포 형태, 이상치, 다봉성, 표본수 차이를 숨길 수 있어 해석을 왜곡할 위험이 있다. 또한 다중 비교가 포함된 분석에서는 어떤 오류율을 통제했는지, 어떤 비교 집합이 정의되었는지, 어떤 사후 보정이 적용되었는지 명확히 기술해야 한다. 오믹스 분석에서는 FDR 기준과 전처리 절차가 결과에 큰 영향을 미치므로, 정규화, 배치 교정, 필터링 기준, 모델링 방법을 재현 가능한 수준으로 기록해야 한다. 이러한 정보가 포함될 때, 효과 크기와 신뢰구간이 제공하는 과학적 의미가 실제로 독자에게 전달된다(37, 38).

결론적으로 결과 해석은 유의성의 확보가 아니라 효과의 크기와 불확실성을 정량화하고, 데이터 생성 구조에 부합하는 모델을 통해 결론의 범위와 조건을 규정하는 과정으로 재정의되어야 한다. 통계적 결론은 우연만으로 설명되기 어려운 수준의 신호가 존재하는지 평가하고, 과학적 결론은 그 신호가 얼마나 크며 어떤 조건에서 의미를 가지는지 제

시한다. 이 두 층위를 혼동하면 유의성이 과학적 의미를 대체하는 오류가 반복된다. 따라서 p 값은 해석의 일부로 제한하고, 효과 크기와 신뢰구간, 모델 기반 추정, 조건부 해석을 결합해 결과를 기술하는 것이 생명과학 연구의 신뢰성과 재현 가능성을 동시에 강화하는 실무적 기준이 된다.

Ⅲ. 결 론

실험 디자인과 통계 분석은 생물학 연구의 신뢰성을 지탱하는 두 개의 기둥이며, 이 둘의 결합이 연구 결과를 단순한 관찰에서 검증 가능한 과학적 증거로 전환한다. 먼저 실험 디자인의 단계에서 대조군의 설정, 무작위 배정(randomization), 블라인딩(blinding), 반복 측정과 독립 반복의 구분, 그리고 교란 요인(confounder)의 통제는 필수적이다. 특히 생물학적 변동성이 큰 시스템에서는 “측정 횟수”를 늘리는 것과 “독립 표본 수”를 늘리는 것이 동일한 정보량을 제공하지 않으므로, 연구 질문의 단위(unit of analysis)를 명확히 규정하고 그 단위에 맞춘 표본 크기 산정이 이루어져야 한다. 표본 크기는 관행적 숫자가 아니라 효과 크기(effect size), 변동성(variance), 유의수준과 검정력(power)을 기반으로 사전에 계획되어야 하며, 이는 연구 결과의 재현성과 해석 가능성을 좌우하는 핵심 조건이다.

통계 분석은 실험 종료 후의 ‘정리 단계’가 아니라, 설계 단계에서부터 내재화되어야 하는 사고 체계이다. 데이터 분포와 측정 척도, 실험 구조(독립군/반복측정/계층 구조)에 부합하는 모형과 검정을 선택하지 않으면, 동일한 데이터라도 결론이 왜곡될 수 있다. 정규성 · 등분산성 가정 점검, 이상치 처리 원칙의 사전 정의, 다중비교에 따른 오류율 통제, 그리고 단순 p-value 중심 보고를 넘어 신뢰구간과 효과 크기의 병행 제시는 현대 생물통계 보고에서 사실상 표준으로 간주한다. 또한 결측치가 발생하는 실제 실험 조건을 고려할 때, 결측의 메커니즘을 평가하고 적절한 처리 전략을 마련하는 것 역시 분석의 타당성을 구성한다. 결국 타당한 검정 선택은 유의성을 찾는 기술이 아니라, 데이터가 지지하는 범위 내에서 생물학적 결론을 엄밀하게 제한하고 구조화하는 과정이다(36, 37).

이러한 흐름에서 오픈소스 소프트웨어인 Python과 R의 도입은 분석 역량의 확장에 그치지 않고, 연구의 투명성과 재현성을 제도화하는 전환점이 된다. 스크립트 기반 분석은 데이터 불러오기, 전처리, 통계 모델링, 시각화, 결과 표 작성까지의 전 과정을 코드로 기록하여 분석의 감사추적(audit trail)을 가능하게 한다. 더 나아가 R Markdown · Quarto와 같은 리포트 도구를 활용하면 코드와 해석, 그림과 표가 하나의 문서에서 동기화되어 결과 보고의 일관성이 강화된다. 패키지 버전 고정, 시드(seed) 설정, 분석 파이프라인 표준화, 그리고 원자료 · 메타데이터 · 코드의 공개는 동료 검증을 용이하게 하고 연구 공동체의 누적성을 높인다. 특히 생물학 연구가 점점 복합 요인 설계와 고차원 데이터를 다루는 방향으로 이동하는 상황에서, R은 선형모형을 넘어 혼합효과모형, 베이지안 접근, 다변량 분석 등 다양한 방법론을 비교 · 검증하는 플랫폼으로 기능할 수 있다.

연구자는 데이터를 생산하는 테크니션에 머물러서는 안 되며, 연구 질문을 검증 가능한 형태로 구조화하고 결과를 논리적으로 해석하여 지식을 창출하는 주체가 되어야 한다. 이를 위해 통계적 사고는 실험의 사후 처리가 아니라 실험의 시작점이라는 원칙이 반복적으로 강조될 필요가 있다. 사전 가설과 분석 계획의 명시, 설계-분석-보고의 일관성 확보, 그리고 재현 가능한 도구 체계의 구축은 신진 연구자에게 가장 중요한 연구 역량으로 연결된다. 본고에서 논의한 실험 디자인의 원칙과 통계적 방법론, 그리고 R 기반 분석 워크플로의 정착이 개별 연구의 신뢰도를 높이는 데 그치지 않고, 바이오 분야 전반의 연구 문화를 한 단계 성숙시키는 기반으로 작동하기를 기대한다.

사 사

본 과제(결과물)는 2025년도 교육부 및 충청북도의 재원으로 충북RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE) 글로컬대학30의 결과입니다(2025-RISE-11-004).

참고문헌

1. Fisher RA. (1935) The design of experiments. Oliver and Boyd. 1~252.
2. Montgomery DC. (2017) Design and analysis of experiments. 9th ed. Wiley. 1~730.
3. Landis SC, Amara SG, Asadullah K, et al. (2012) A call for transparent reporting to optimize the predictive value of preclinical research. Nature. 490, 187~91.
4. Ioannidis JPA. (2005) Why most published research findings are false. PLoS Med. 2, e124.
5. Begley CG, Ellis LM. (2012) Drug development: Raise standards for preclinical cancer research. Nature. 483, 531~3.

6. Tunçer S, Banerjee S. (2018) Low dose dimethyl sulfoxide driven gross molecular changes have biological significance. Sci Rep. 8, 14860.
7. Schulz KF, Altman DG, Moher D; CONSORT Group. (2010) CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. BMJ. 340, c332.
8. Percie du Sert N, Hurst V, Ahluwalia A, et al. (2020) The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. PLoS Biol. 18, e3000410.
9. Vaux DL, Fidler F, Cumming G. (2012) Replicates and repeats—what is the difference and is it significant? EMBO Rep. 13, 291~6.
10. Hurlbert SH. (1984) Pseudoreplication and the design of ecological field experiments. Ecol Monogr. 54, 187~211.
11. Leek JT, Scharpf RB, Bravo HC, et al. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 11, 733~9.
12. Johnson WE, Li C, Rabinovic A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 8, 118~27.
13. Cohen J. (1988) Statistical power analysis for the behavioral sciences. 2nd ed. Lawrence Erlbaum Associates. 1~567.
14. Button KS, Ioannidis JPA, Mokrysz C, et al. (2013) Power failure: Why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 14, 365~76.
15. Gelman A, Hill J. (2006) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press. 1~625.
16. Pinheiro JC, Bates DM. (2000) Mixed-effects models in S and S-PLUS. Springer. 1~528.
17. Student. (1908) The probable error of a mean. Biometrika. 6, 1~25.
18. Holm S. (1979) A simple sequentially rejective multiple test procedure. Scand J Stat. 6, 65~70.
19. Fisher RA. (1925) Statistical methods for research workers. Oliver and Boyd. 1~356.
20. Tukey JW. (1949) Comparing individual means in the analysis of variance. Biometrics. 5, 99~114.
21. Kramer CY. (1956) Extension of multiple range tests to group means with unequal numbers of replications. Biometrics. 12, 307~10.
22. Dunnett CW. (1955) A multiple comparison procedure for comparing several treatments with a control. J Am Stat Assoc. 50, 1096~121.
23. Dunn OJ. (1961) Multiple comparisons among means. J Am Stat Assoc. 56, 52~64.
24. Scheffé H. (1953) A method for judging all contrasts in the analysis of variance. Biometrika. 40, 87~104.
25. Duncan DB. (1955) Multiple range and multiple F tests. Biometrics. 11, 1~42.
26. Games PA, Howell JF. (1976) Pairwise multiple comparison procedures with unequal n's and/or variances: A Monte Carlo study. J Educ Behav Stat. 1, 113~25.
27. Sandve GK, Nekrutenko A, Taylor J, Hovig E. (2013) Ten simple rules for reproducible computational research. PLoS Comput Biol. 9, e1003285.
28. Kluyver T, Ragan-Kelley B, Pérez F, et al. (2016) Jupyter notebooks—A publishing format for reproducible computational workflows. In: Positioning and Power in Academic Publishing. IOS Press. 87~90.
29. Wickham H, Averick M, Bryan J, et al. (2019) Welcome to the tidyverse. J Open Source Softw. 4, 1686.
30. R Core Team. (2024) R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/
31. Google. (2024) Colaboratory FAQ. https://research.google.com/colaboratory/faq.html
32. Xie Y, Allaire JJ, Grolemund G. (2018) R Markdown: The definitive guide. CRC Press. https://bookdown.org/yihui/rmarkdown/
33. Posit. (2024) Quarto. https://quarto.org/
34. Chen M, Tworek J, Jun H, et al. (2021) Evaluating large language models trained on code. arXiv:2107.03374. https://arxiv.org/abs/2107.03374
35. Noy S, Zhang W. (2023) Experimental evidence on the productivity effects of generative AI. Science. 381, 187~92.

36. Wasserstein RL, Lazar NA. (2016) The ASA's statement on p-values: Context, process, and purpose. Am Stat. 70, 129~33.

37. Lakens D. (2013) Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. Front Psychol. 4, 863.

38. Gardner MJ, Altman DG. (1986) Confidence intervals rather than P values: Estimation rather than hypothesis testing. BMJ. 292, 746~50.

Received Nov. 23, 2025, Revised Dec. 22, 2025, Accepted Dec. 28, 2025