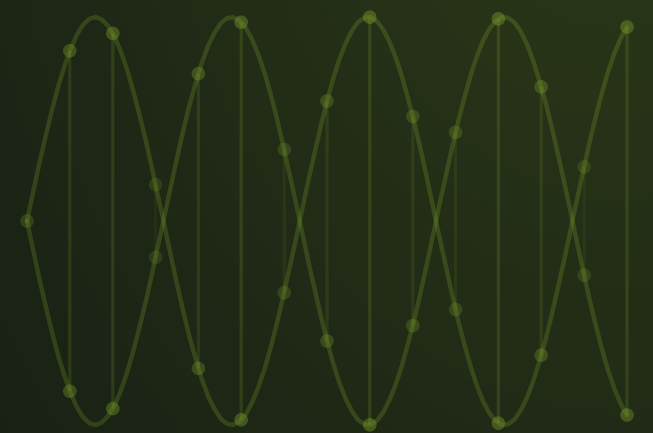


레슨: p-값, 그리고 유전학이 0.000000005를 쓰는 이유



확률과 통계

수학

유전학

만 16세+

심화

학생 한 명당 동전 하나로 하는 무료 통계 Lesson입니다. 학급이 확실히 일어나지 않는 일에 대해 600건의 연구를 수행하고, 약 38건의 발견을 얻고, 거기서부터 유전학이 왜 0.05가 아니라 0.000000005를 필요로 하는지 직접 끌어냅니다.

자유롭게 읽고, 쓰고, 고쳐 쓰실 수 있습니다. 계정도, 구매도, 수집되는 정보도 없습니다.

전체 레슨과 그림, 갱신본: <https://www.mygenelog.com/ko/updates/lesson-the-threshold>

이 자료는, 실제 유전학 데이터로 만든, 바로 쓸 수 있는 통계 Lesson입니다. 어느 학교·칼리지·대학에서나 자유롭게 쓰고 고치실 수 있습니다. 저희에게서 살 것도, 계정도, 누구에게서든 수집하는 데이터도 없습니다.

- 수준: 만 16세부터 대학원까지. 활동 자체는 어느 수준에서나 동일하고, 끝의 질문이 단계별로 나뉘어 있습니다. 마지막 묶음은 실제로 어렵습니다.
- 시간: 45-60분. 활동 자체는 15분쯤 걸립니다.
- 준비물: 학생 한 명당 동전 하나, 또는 휴대전화. 칠판. 마지막 부분에 계산기. 그것이 전부입니다 — 주문할 것도 쓸 돈도 없습니다.
- 다루는 내용: p-값, 다중검정(multiple testing) 문제, 집단별 오류율, 본페로니 교정(Bonferroni correction), 유전체 전체 유의성이 왜 5×10^{-8} 인지, 엄격한 기준을 넘는 유일한 방법이 왜 데이터를 더 모으는 것뿐인지, 그리고 과학의 관행이 실제로 어떻게 정해지는지.

여기서 시작합니다: 166개 중 6개

2002년, 유전학에는 입 밖에 내고 싶지 않은 문제가 하나 있었습니다.

연구자들은 십 년 동안 어떤 질환의 그럴듯한 용의자로 보이는 유전자를 고르고, 그 유전자의 변이가 환자에게 더 흔한지 건강한 사람과 비교하고, 답이 통상적인 기준인 $p < 0.05$ 아래로 나오면 발표했습니다. 그렇게 보고된 양성 연관이 600건이 넘었습니다.

그러다 네 명의 연구자가 앉아서 그중 무엇이 살아남았는지 확인했습니다. 독립적인 연구진이 세 번 이상 검증한 연관 166건 가운데, 일관되게 재현된 것은 6건이었습니다.

6건. 166건 중에서.

아무도 속이지 않았습시다. 그 연구들은 전부 통계학 입문 수업이 가르치는 바로 그 기준을, 올바르게 적용했습니다. 문제는 기준 쪽이었고, 앞으로 한 시간은 그 이유에 대한 것입니다.

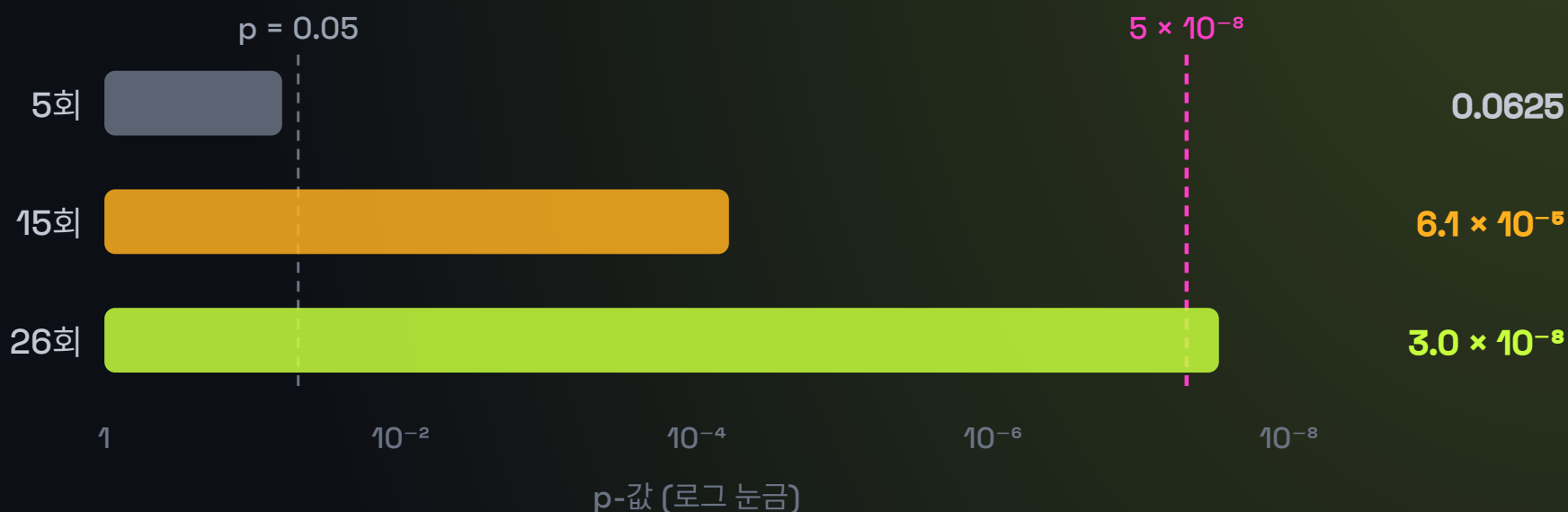
이렇게 하세요

1. 동전을 던지기 전에 먼저 예측을 받으세요. 이렇게 물으십시오. “유전체에서 100만 개의 위치를 검사한다고 합시다. 그리고 그중 어느 것도 이 질환에 아무 영향이 없다고 합시다. 아무 데서도 아무 일이 일어나지 않고 있습니다. $p < 0.05$ 로 유의하게 나오는 것은 몇 개일까요?” 세 명에게 답을 받아 칠판에 적으십시오. 어느 것이 가깝다고는 말하지 마십시오.
2. 연구 한 건을 정의합니다. 연구 한 건은 동전 5회 던지기입니다. 다섯 번이 모두 같은 면 — 앞면 다섯 또는 뒷면 다섯 — 이면 그 결과를 “유의하다”고 부릅니다. 확률은 알려 주지 말고 학생들과 함께 구하십시오. $2 \div 25 = 2/32 = 0.0625$, 통상적인 0.05와 비슷합니다.
3. 무엇이 대조군인지 소리 내어 말하십시오. 이 실험 전체가 대조 조건입니다. 동전은 공정하고, 아무것도 그것에 영향을 주지 않으며, 찾을 것이 확실히 없습니다. 학생들이 지금부터 하게 될 모든 “발견”은 설계상 거짓입니다 — 그것이 결함이 아니라 설계입니다. 이 한 가지가 수업 전체를 지탱하므로 여기서 한 번 멈출 값어치가 있습니다.
4. 각자 20건씩 수행합니다. 학생마다 다섯 번 던져 앞/뒤를 적고, 다섯이 모두 같으면 표시하고, 이것을 스무 번 반복합니다. 10분쯤 걸리고 시끄럽습니다. 시끄럽게 두십시오.
5. 학급 합계를 칠판에 적습니다. 숫자는 둘뿐입니다. 학급 전체가 수행한 연구 건수(학생 수 $\times$ 20), 그리고 그중 “유의하게” 나온 건수. 30명 학급이라면 600건을 수행하고, $600 \div 16 = 37 \sim 38$ 건의 발견을 얻게 됩니다. 아무것도 아닌 것에 대한 발견을 말입니다.
6. 이제 칠판의 예측으로 돌아갑니다. 100만 번 검사에 $p < 0.05$ 면 거짓 발견은 50,000건입니다. 그렇게 큰 수를 답하는 사람은 거의 없고, 학생들은 방금 그 축소판을 자기 손으로 만들어 냈습니다.
7. 고쳐 보고, 고치는 데 무엇이 드는지 보십시오. 600번 검사하면서 거짓 발견이 하나도 통과하지 않을 것을 95% 확신하고 싶다면, 기준을 검사 횟수로 나눕니다. $0.05 \div 600 = 8.3 \times 10^{-5}$. 학생들에게 물으십시오. 5회 던지기로 그 값에 닿을 수 있습니까? 닿을 수 없습니다 — 5회로 만들 수 있는 가장 작은 p-값이 0.0625이기 때문입니다. 어떤 운도 너무 작은 연구를 구해 주지 못합니다.
8. 실제 숫자로 마무리합니다. 그러면 몇 번 던져야 할까요? $2 \div 2n \leq 8.3 \times 10^{-5}$ 를 풀면 $n = 15$ 입니다. 그러면 실제 유전학이 쓰는 기준인 5×10^{-8} 는? $n = 26$ 입니다. 25회는 5.96×10^{-8} 여서 모자랍니다.

이 수업에서는 누구의 몸에 대한 것도 수집하지 않으며, 수집해서도 안 됩니다. 이 활동이 동전을 쓰는 것은 의도된 것입니다. 학생 본인의 유전 정보나 건강, 가족에 대해 묻지 마십시오 — 심화 활동으로도, 숙제로도, 익명으로도 안 됩니다. 숫자가 필요하면 교실을 세십시오. 명단은 절대 만들지 마십시오. 학생이 자기 유전 정보를 스스로 이야기한다면 그것은 그 학생의 일이지 학급의 데이터가 아닙니다.

학생들이 방금 증명한 것

같은 면이 연달아 나올 때 닿을 수 있는 가장 작은 p-값 →



눈대중이 아니라 정확한 값입니다. 같은 면이 n번 연달아 나올 확률은 $2 \div 2^n$ 입니다. 5회는 0.0625여서 0.05 선을 아슬아슬하게 넘지 못합니다. 5×10^{-8} 를 넘으려면 26회가 필요합니다 — 25회는 5.96×10^{-8} 로 모자랍니다. 그 21회의 차이가 곧 “데이터를 더 모은다”는 말의 뜻입니다.

연구 한 건의 기준은 연구 천 건의 기준이 아니고, 애초에 그런 적도 없었습니다. $p < 0.05$ 가 실제로 약속하는 것은 이것입니다. 아무 일도 일어나지 않고 있다면, 이 정도로 강한 결과는 스무 번에 한 번쯤 나온다. 한 번 검사한다면 감수할 만한 위험입니다. 100만 번 검사한다면 그것은 5만 개의 틀린 답을 보장하는 것입니다.

이것이 다중검정 문제의 전부이고, 학생들은 그것을 설명으로 들은 것이 아니라 처음부터 직접 만들어 냈습니다.

5×10^{-8} 는 어디에서 왔나

그래서 유전학에는 숫자가 필요했습니다. 2008년, 두 연구진이 완전히 다른 방식으로 그것을 찾아 냈습니다. 그리고 이 부분은 학생들이 거의 듣지 못하는 이야기입니다.

한쪽은 세었습니다. 유전체에서 가까이 있는 위치들은 함께 유전되기 때문에, 검사한 위치 300만 개는 독립적인 검정 300만 회와는 거리가 멉니다. 국제 HapMap 자료를 바탕으로, 이들은 실제 검정 부담을 유럽계에서 약 100만 회의 독립 검정으로 추정했고 — 아프리카계에서는 그 두 배로 보았습니다. 아프리카계 유전체가 변이를 더 많이 지니고, 그 변이가 덩어리로 함께 움직이는 정도는 덜하기 때문입니다. 0.05를 100만으로 나누면 정확히 5×10^{-8} 가 나옵니다.

다른 쪽은 측정했습니다. 이들은 웰컴트러스트 환자-대조군 컨소시엄의 실제 유전형 자료를 가져와, 질환 표지를 무작위로 뒤섞어 어떤 신호가 나오든 잡음일 수밖에 없게 만든 뒤, 그 잡음이 얼마나 강해지는지를 기록하고, 표지 밀도를 높여 가며 답이 더 움직이지 않을 때까지 반복했습니다. 외삽으로 얻은 기준값은 7.2×10^{-8} 였습니다.

실제 영국 유전형 자료에 순열 검정으로 측정 (2008)

7.2

0.05 ÷ 독립 검정 1,000,000회 (유럽계) — 관행이 된 값

5.0

0.05 ÷ 독립 검정 2,000,000회 (아프리카계) — 관행이 아닌 값

2.5

기준값, 10^{-8} 단위 — 막대가 길수록 넘기 쉬운 기준입니다

같은 질문을 같은 해에 세 가지 방식으로 물었습니다. 학계는 가운데 값을 골라 모두에게 똑같이 적용합니다 — 이 수업 끝에 나오는 논쟁이 바로 그것입니다.

하나는 이론이고 하나는 실측인 두 방법이, 1.5배 안쪽으로 서로 맞았습니다. 학계는 5×10^{-8} 로 반올림했고 그 뒤로 계속 그것을 씁니다.

학생들에게 분명히 말해 줄 값어치가 있습니다. 과학의 관행이란 안에서 보면 이렇게 생겼습니다. 누가 내려준 것도, 증명된 것도 아니고 — 두 번 따져 본 뒤, 각자 자기 것을 고르는 것보다 합의하는 편이 낫기 때문에 합의된 것입니다.

실제 데이터: 그 기준이 실제로 막는 것

이 사이트는 공개된 GWAS Catalog에서 유전 연관 정보를 자동으로 수집하고, 5×10^{-8} 보다 약한 것은 받지 않습니다. 그 필터가 무슨 일을 했는지, 자체 로그에서 센 숫자입니다.

2026년 9월 8일에 센 것입니다. 목록은 그 뒤로 자랐지만 이 숫자들은 그대로 둡니다. 수집기가 한 번 돌아간 결과를 기술한 것이라, 오늘 값으로 바꾸면 그 값에서 뽑아 그린 비율과 맞지 않게 됩니다.

연관 기록 3,510건, 필터를 통과한 뒤의 모습

47.5%

44.4%

- 282건 공개 — 8.0%, 그리고 전부 5×10^{-8} 를 넘었습니다
- 1,669건 거절 — p-값이 5×10^{-8} 보다 약해서
- 1,559건 거절 — 대립유전자 불명확, 유전자 미확인, 이미 수록됨

추정이 아니라 이 사이트의 수집기 로그에서 직접 센 숫자입니다. 살펴본 것의 절반 가까이가 p-값 하나 때문에 거절되었습니다.

살펴본 연관 기록 3,510건 가운데 1,669건 — 47.5% — 이 p-값 하나 때문에 거절되었습니다. 또 1,559건은 다른 이유로 거절되었습니다. 대립유전자 자료가 깨끗하지 않았거나, 유전자가 지정되지 않았거나, 이미 여기 있는 기록이었습니다. 282건이 페이지가 되었습니다.

절반 가까이입니다. 거절된 기록 하나 하나는 공개 데이터베이스에 들어 있는, 발표되고 동료 심사를 거친 결과입니다. 그런데도 이 사이트는 그것을 실지 않습니다. 학생들이 방금 끌어낸 그 기준이, 그것은 동전이 다섯 번 연달아 앞면이 나온 것일 수도 있다고 말하기 때문입니다.

이 사이트가 실고 있는 변이 653건은 서로 다른 논문 182편에서 왔고, 그 하나하나가 모두 5×10^{-8} 를 넘었습니다.

아직 끝나지 않은 논쟁

위의 그림을 다시 보여 주고 학생들끼리 싸우게 두십시오. 연구자들이 지금도 정확히 그러고 있기 때문입니다.

기준은 모두에게 적용되는 하나의 숫자인데, 검정 부담은 모두에게 같지 않습니다. 유럽계 연구가 정말로 독립 검정 100만 회를 하고 있고 아프리카계 연구가 200만 회를 하고 있다면, 5×10^{-8} 는 한쪽에는 맞는 기준이고 다른 쪽에는 너무 느슨한 기준입니다. 아프리카계 연구에 더 엄격한 2.5×10^{-8} 를 쓰는 것이 더 옳겠지만, 그것은 유전학이 이미 가장 적게 연구해 온 집단의 발견을 발표하기 더 어렵게 만들기도 합니다. 여기에는 순수하게 통계적인 답이 없습니다.

그리고 이 숫자는 지금 교체되고 있는 기술을 기준으로 계산되었습니다. 독립 검정 100만 회는 흔한 변이를 읽는 유전형 칩에 대한 이야기입니다. 전장유전체 시퀀싱은 드문 변이까지, 훨씬 더 많이 읽고, 그래서 기준을 5×10^{-9} 이하로 내려야 한다는 주장이 있습니다. 다른 쪽은 안전할 만큼 엄격한 기준은 실제 효과를 어떤 표본 크기로도 찾을 수 없게 만들 만큼 엄격하다고 답합니다.

두 주장 모두 살아 있습니다. 그림과 20분이 있는 학급은, 지금 논쟁하는 사람들과 똑같은 정보를 가지고 있는 것입니다.

질문

몸풀기

1. 학급은 확실히 일어나지 않고 있는 일에 대해 여러 건의 연구를 수행했습니다. 시작하기 전이라면 “발견”이 몇 건 나오리라고 예상했겠습니까? 실제로는 몇 건이 나왔습니까?
2. 동전이 공정한 것이 왜 중요했습니까? 그중 하나가 한쪽으로 기울어져 있었다면 이 수업의 무엇이 잘못되었겠습니까?
3. 어떤 유전자가 수학을 잘하는 것과 연관되었다고, $p < 0.05$ 로 신문에 났습니다. 가장 먼저 물어야 할 질문은 무엇입니까?

본문제

1. 한 번 검사할 때는 관측은 기준이 100만 번 검사할 때는 왜 쓸모없어지는지, 12살이 이해할 수 있는 한 문장으로 설명하십시오.
2. 학급은 기준을 $0.05 \div (\text{연구 건수})$ 로 고쳤습니다. 누가 아무리 운이 좋았더라도 학급의 발견이 하나도 남김없이 사라진 이유는 무엇입니까?
3. 26회가 5×10^{-8} 를 넘는 가장 작은 횟수임을 보이고, “더 많이 던진다”는 것이 실제 유전 연구에서 무엇에 해당하는지 말하십시오.
4. 2008년의 두 연구진 중 한쪽은 기준을 계산했고 다른 쪽은 측정했습니다. 각각의 장점은 무엇이며, 둘이 거의 일치한 것이 왜 안심할 만한 일입니까?

더 나아가기

1. 본페로니 교정은 검정들이 서로 독립이라고 가정합니다. 염색체에서 가까이 있는 위치들은 함께 유전되므로 독립이 아닙니다. 그러면 이 교정은 너무 엄격해집니까, 너무 느슨해집니까? 그리고 2008년의 연구진은 왜 표지 개수를 그냥 세지 않고 독립 검정 횟수를 추정해야 했습니까?
2. 엄격한 기준은 거짓 발견을 줄이고 놓치는 진짜 발견을 늘립니다. 두 종류의 오류는 각각 누가 대가를 치릅니까? 흔한 질환에 대한 연구와 희귀 질환에 대한 연구에서 답이 같습니까?
3. 올바른 부담이 유럽계에서 100만 회, 아프리카계에서 200만 회인데 학계가 양쪽 모두에 5×10^{-8} 를 쓴다면 어떤 일이 일어납니까? 이제 반대편에서 주장해 보십시오. 이미 어떤 집단을 덜 연구하는 분야가, 그 집단의 결과를 발표하기까지 더 어렵게 만든다면 어떻게 됩니까?
4. 2002년에는 166건 중 6건만 재현되었습니다. 처음부터 기준이 5×10^{-8} 였다고 가정해 보십시오. 나머지 160건이 전부 사라졌겠습니까, 아니면 진짜 발견도 함께 사라졌겠습니까?

교사용 메모 — 각 질문이 향하는 곳

1. 기댓값은 (연구 건수 ÷ 16)입니다. 관측값은 그것과 맞지 않을 것이고, 그 차이 자체가 1분쯤 다를 값어치가 있습니다. 거짓 발견율 위에 표본 변동이 얹혀 있고, 둘이 지금 교실 안에 동시에 있습니다.
2. 기울어진 동전이었다면 실제로 찾을 것이 있었다는 뜻이고, 발견이 전부 거짓은 아니게 됩니다. 이 수업은 귀무가설이 참이라는 것을 안다는 데 기대고 있습니다 — 실제 연구가 결코 알 수 없는 바로 그것이고, 이 지점은 답을 내지 말고 남겨 두십시오.
3. “몇 개를 검사했습니까?” 나머지는 전부 여기서 따라옵니다.
4. 기술적인 답이 아니라 강한 답을 기대하십시오. “스무 번에 한 번 일어나는 일은 100만 번 해 보면 반드시 일어난다.”
5. 교정이 기준을 그 설계가 물리적으로 낼 수 있는 최선의 결과보다 아래로 옮겼기 때문입니다. 학생들이 자주 뒤섞는 두 가지를 여기서 갈라 줍니다. 증거의 강도는 연구의 크기에 의해 위가 막혀 있고, 운은 그 한계를 넘지 못합니다.
6. $2/2^n \leq 5 \times 10^{-8}$ 이라면 $2^n \geq 4 \times 10^7$ 이어야 하므로 $n = 26$ 입니다 ($2^{26} = 67,108,864$, 곧 2.98×10^{-8}). 더 많이 던진다는 것은 참여자를 더 모은다는 뜻입니다. 도달 가능한 가장 작은 p-값을 움직이는 지렛대는 표본 크기뿐입니다.
7. 계산은 일반화되고 값이 싸며, 측정은 상관 구조를 포함한 실제 데이터의 지저분함을 담습니다. 가정이 서로 다른 두 방법이 일치하는 것은 어느 한쪽보다 값어치가 큼니다 — 이 사이트의 백반증 페이지와 혈압 페이지가 쓰는 것과 같은 논리입니다.
8. 너무 엄격해집니다. 검정들이 상관되어 있으면 실질 검정 수가 표지 개수보다 적으므로, 표지 개수로 나누면 과하게 교정됩니다. 그래서 독립 검정 수를 추정한 것입니다. 잘하는 학생은 이것이 기준을 기술의 성질이 아니라 집단의 성질로 만든다는 것을 알아챌 수 있습니다 — 그것이 10번 질문입니다.
9. 정답이 하나 있는 질문이 아닙니다. 실제로 누가 피해를 보는지 밀어붙이십시오. 거짓 발견은 연구비를 낭비하고 환자들에게까지 갈 수 있으며, 놓친 발견은 치료를 십 년 늦출 수 있습니다. 희귀 질환에서는 놓치는 쪽이 대개 더 나쁜 오류이고, 그래서 기준을 조정하는 분야도 있습니다.
10. 가장 어렵고 가장 좋은 질문입니다. 통계에는 깔끔한 답이 있고 그 결과에는 없습니다. 대신 정리해

주지 마십시오. 아무도 꺼내지 않으면, 검정 부담이 더 큰 집단이 애초에 연구 자체가 가장 적은 집단이라는 점을 짚어 주십시오.

11. 진짜 발견도 분명히 함께 사라졌을 것입니다 — 재현된 6건 중 몇 개는 그 크기의 연구에서 유전체 전체 기준을 넘지 못했을 것입니다. 정직한 답은, 학계가 발견의 수를 신뢰도와 맞바꾸었고 그 거래가 그럴 만했다고 판단했다는 것입니다.

더 나아가기

- 부모가 물려주지 않은 대립유전자 — 300만 명을 본 연구, 그리고 아주 큰 표본이 무엇을 사 주는가. 더 나아가기 질문의 후속 읽기로 좋습니다.
- 데이터 페이지 — 실시간 집계. 변이가 몇 건이고 그 뒤에 논문이 몇 편인지 포함해서.
- 이 데이터는 어디에서 오는가 — 어떤 연관 정보가 여기서 페이지가 되기까지 통과해야 하는 필터 전부. p-값은 그중 첫 번째일 뿐입니다.
- C반응단백 — 유의성 다음에 오는 질문의 실제 사례입니다. 연관은 진짜인데, 그러면 그것이 원인인가?

자유롭게 쓰십시오. 인쇄하고, 복사하고, 잘라 쓰고, 학습지에 옮기고, 번역하고, 질문을 바꾸십시오. 허락받을 것도 낼 돈도 없습니다. 출처를 밝히신다면 mygenelog.com이면 충분합니다.

수업에 써 보시고 잘 안 되는 부분이 있으면 알려 주십시오 — 저희에게는 감사 인사보다 그 편이 값집니다.