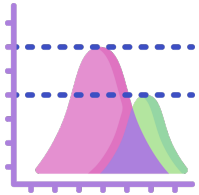
실험에 기반한 당신의 가설 검정 결과를 신뢰할 수 없는 이유

방태모 | G마켓

본인 소개

데이터야.놀자2024

통계학 전공



Data Scientist

가천대 길병원
G-ABC

Data Scientist

Gmarket

글쓰기

[블로그](#)
[요즘IT](#)

Taemo Bang

Data Scientist @ Gmarket | Experimentation | Causal Inference | Time Series

작년에 데놀에서 했던 이야기

데이터야.놀자 2024

데이터야놀자 유튜브 발표 영상

오늘 할 이야기의 주된 키워드?

검정력 (Power)

이 발표를 준비한 이유

우리는 검정력에 대해 얼마나 이해하고 있는가?

- 1종 오류에 관한 관심에 비해, 검정력에 관한 관심은 미약

이 발표를 준비한 이유

우리는 검정력에 대해 얼마나 이해하고 있는가?

- 1종 오류에 관한 관심에 비해, 검정력에 관한 관심은 미약

	귀무가설이 참	**귀무가설이 거짓**
귀무가설 기각	False positive (1종 오류, α)	True positive (검정력, $1-\beta$)
귀무가설 기각 실패	True negative ($1-\alpha$)	False negative (2종 오류, β)

표 1. 통계적 가설 검정 결과

이 발표를 준비한 이유

우리는 검정력에 대해 얼마나 이해하고 있는가?

- 1종 오류에 관한 관심에 비해, 검정력에 관한 관심은 미약

	귀무가설이 참	**귀무가설이 거짓**
귀무가설 기각	False positive (1종 오류, α)	True positive (검정력, $1-\beta$)
귀무가설 기각 실패	True negative ($1-\alpha$)	False negative (2종 오류, β)

표 1. 통계적 가설 검정 결과

이 발표를 준비한 이유

우리는 검정력에 대해 얼마나 이해하고 있는가?

- 1종 오류에 관한 관심에 비해, 검정력에 관한 관심은 미약

	귀무가설이 참	**귀무가설이 거짓**
귀무가설 기각	False positive (1종 오류, α)	True positive (검정력, $1-\beta$)
귀무가설 기각 실패	True negative ($1-\alpha$)	False negative (2종 오류, β)

표 1. 통계적 가설 검정 결과

이 발표를 준비한 이유

우리는 검정력에 대해 얼마나 이해하고 있는가?

- 1종 오류에 관한 관심에 비해, 검정력에 관한 관심은 미약

	귀무가설이 참	**귀무가설이 거짓**
귀무가설 기각	False positive (1종 오류, α)	True positive (검정력, $1-\beta$)
귀무가설 기각 실패	True negative ($1-\alpha$)	False negative (2종 오류, β)

표 1. 통계적 가설 검정 결과

이 발표를 준비한 이유

우리는 검정력에 대해 얼마나 이해하고 있는가?

- 1종 오류에 관한 관심에 비해, 검정력에 관한 관심은 미약

	귀무가설이 참	**귀무가설이 거짓**
귀무가설 기각	False positive (1종 오류, α)	True positive (검정력, $1-\beta$)
귀무가설 기각 실패	True negative ($1-\alpha$)	False negative (2종 오류, β)

표 1. 통계적 가설 검정 결과

이 발표를 준비한 이유

우리는 검정력에 대해 얼마나 이해하고 있는가?

- 사후 검정력 분석에 관한 잘못된 이해
  - 의학 논문 리뷰에서도 간혹 요구
  - 이미 관측된 실험 데이터를 기반으로 한 사후 검정력 분석은 무의미함

이 발표를 준비한 이유

3가지를 해소하고자 합니다.

1. 검정력이란 무엇이고 대체 왜 중요한가? 검정력 분석은 또 무엇인가?
2. 검정력이 낮은 실험에서는 어떤 현상이 발생하는가?
3. 사후 검정력 분석은 우리를 어떤 함정에 빠뜨리는가?

본 발표를 통해 여러분들이 느꼈으면 하는 2가지

- “실험에서 검정력 분석은 선택이 아닌 필수겠구나”
- “검정력 분석을 수행하지 않은 실험의 경우 이런 위험이 있겠구나”

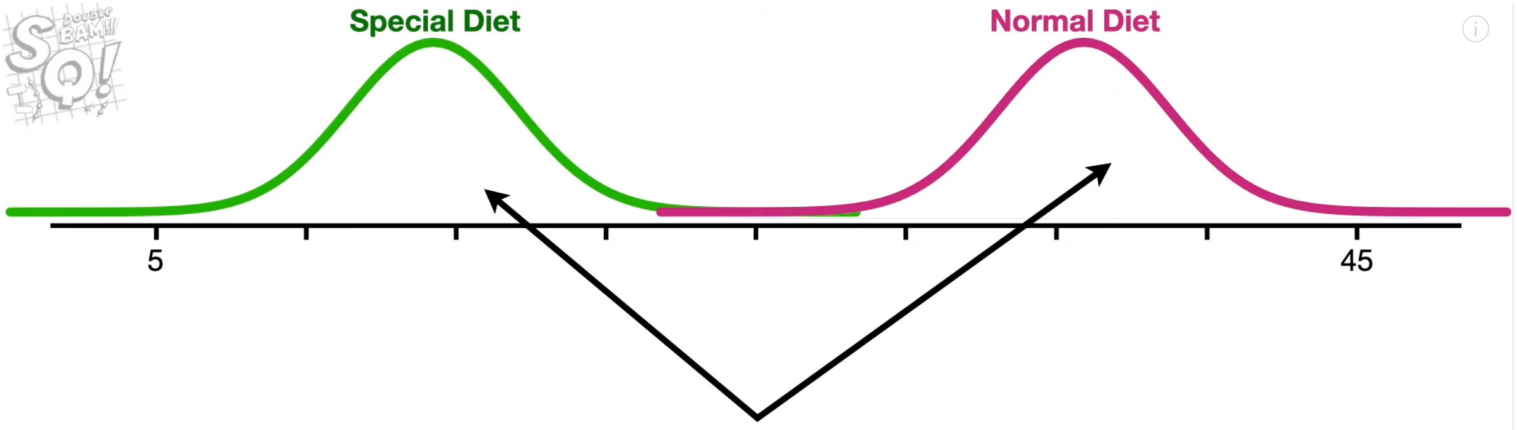
1 검정력

검정력이란?

검정력(Power, $1 - \beta$)

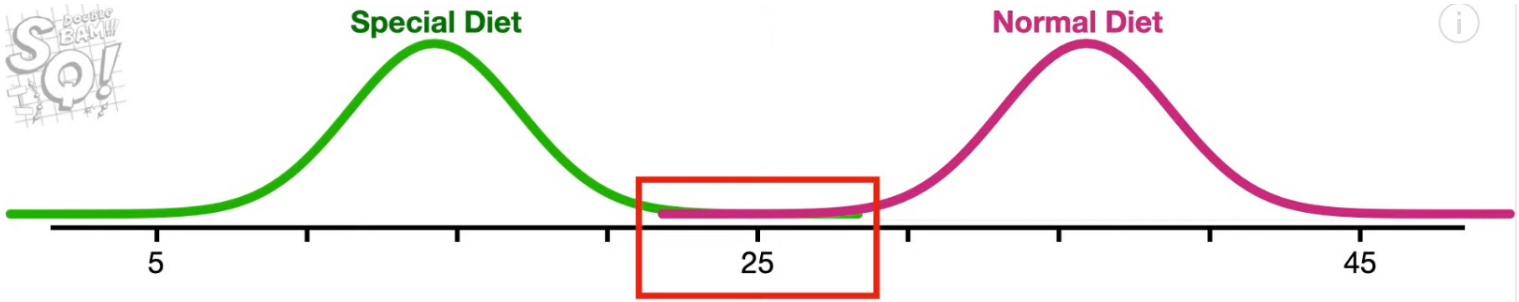
- 실험 변형군 간 의미있는 차이가 있음을 **올바르게** 탐지할 확률
  - (학술적 표현) 귀무가설(H_0)을 올바르게 기각시킬 확률

검정력의 의미



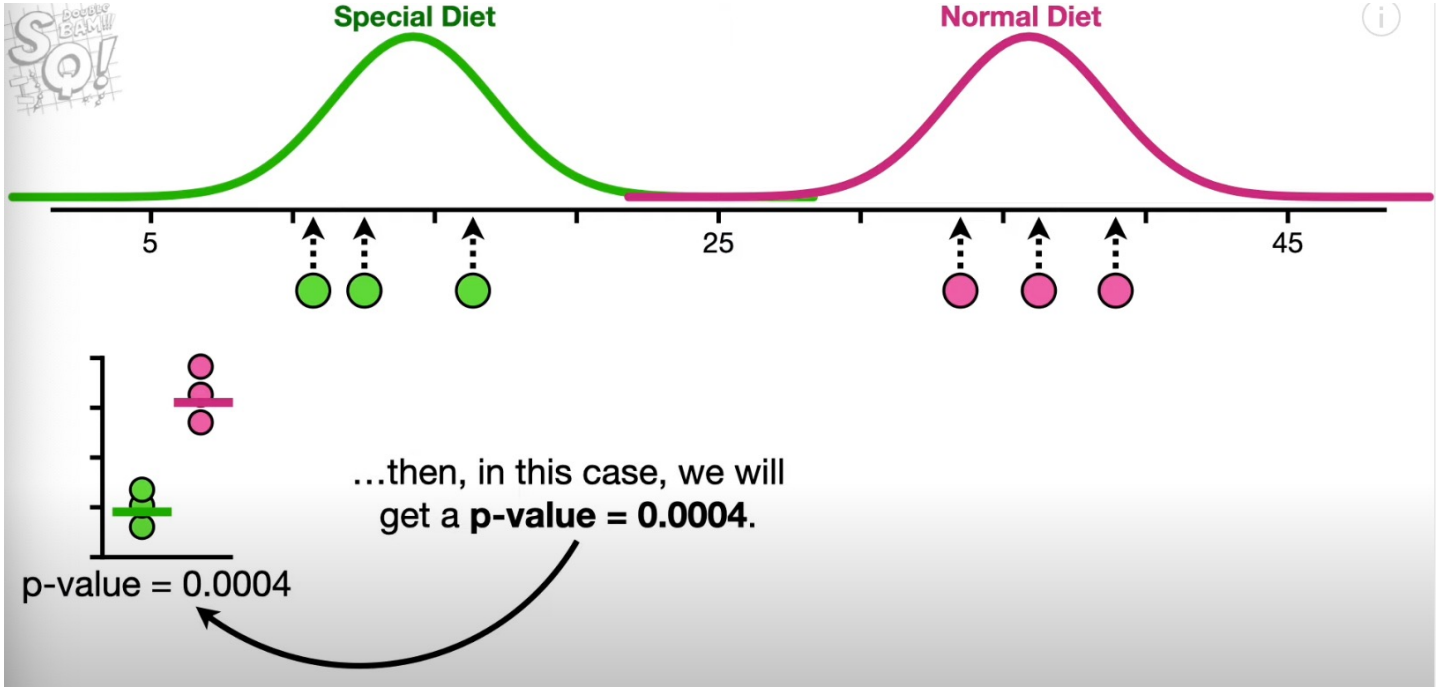
Source: Statquest, Statistical Power, Cleary Explained!

검정력의 의미



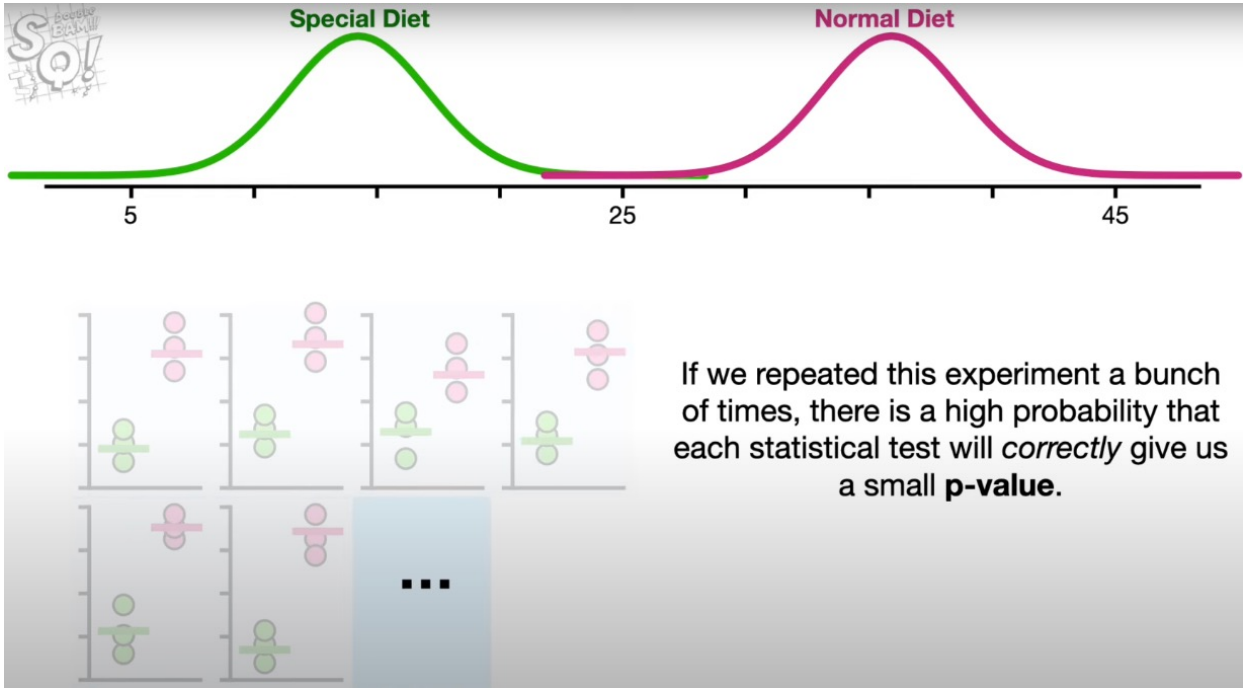
Source: Statquest, Statistical Power, Cleary Explained!

검정력의 의미



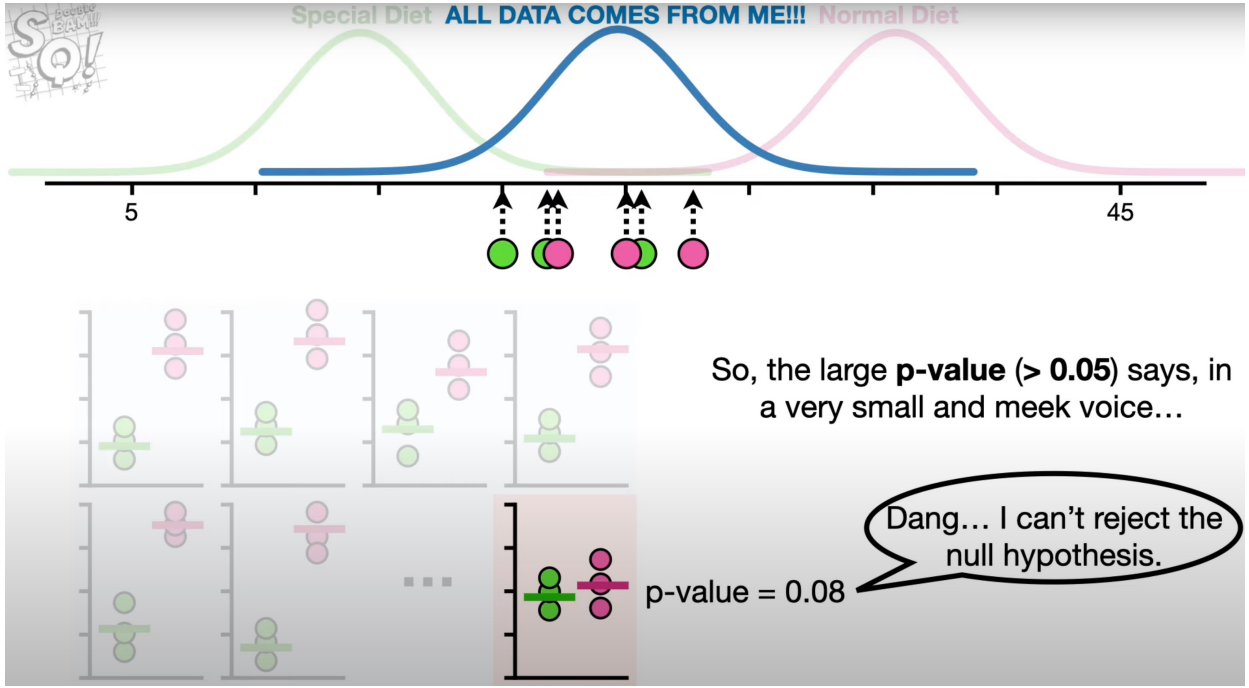
Source: Statquest, Statistical Power, Cleary Explained!

검정력의 의미



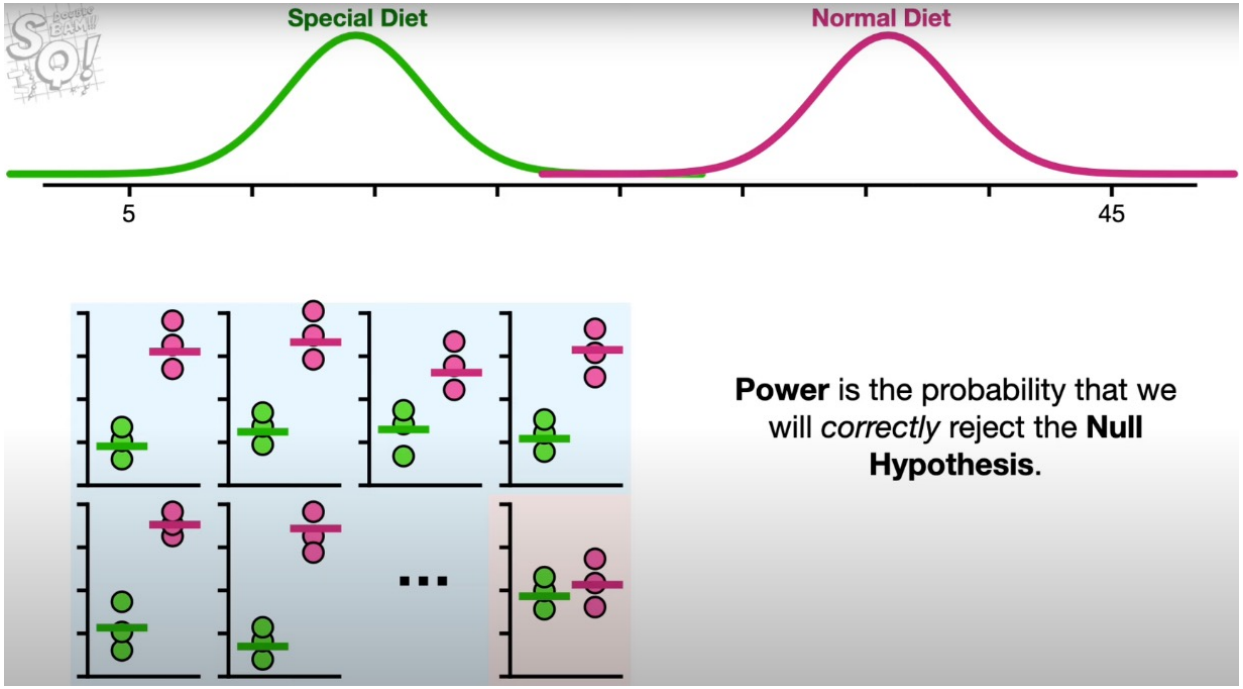
Source: Statquest, Statistical Power, Cleary Explained!

검정력의 의미



Source: Statquest, Statistical Power, Cleary Explained!

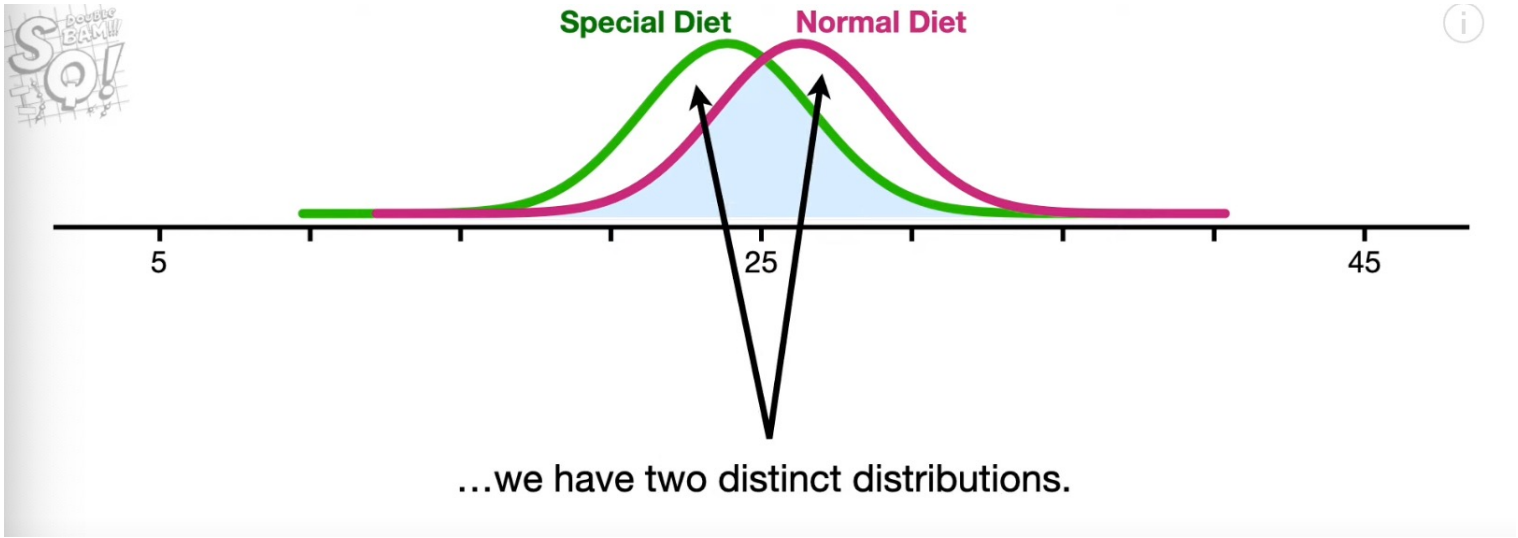
검정력의 의미



Source: Statquest, Statistical Power, Cleary Explained!

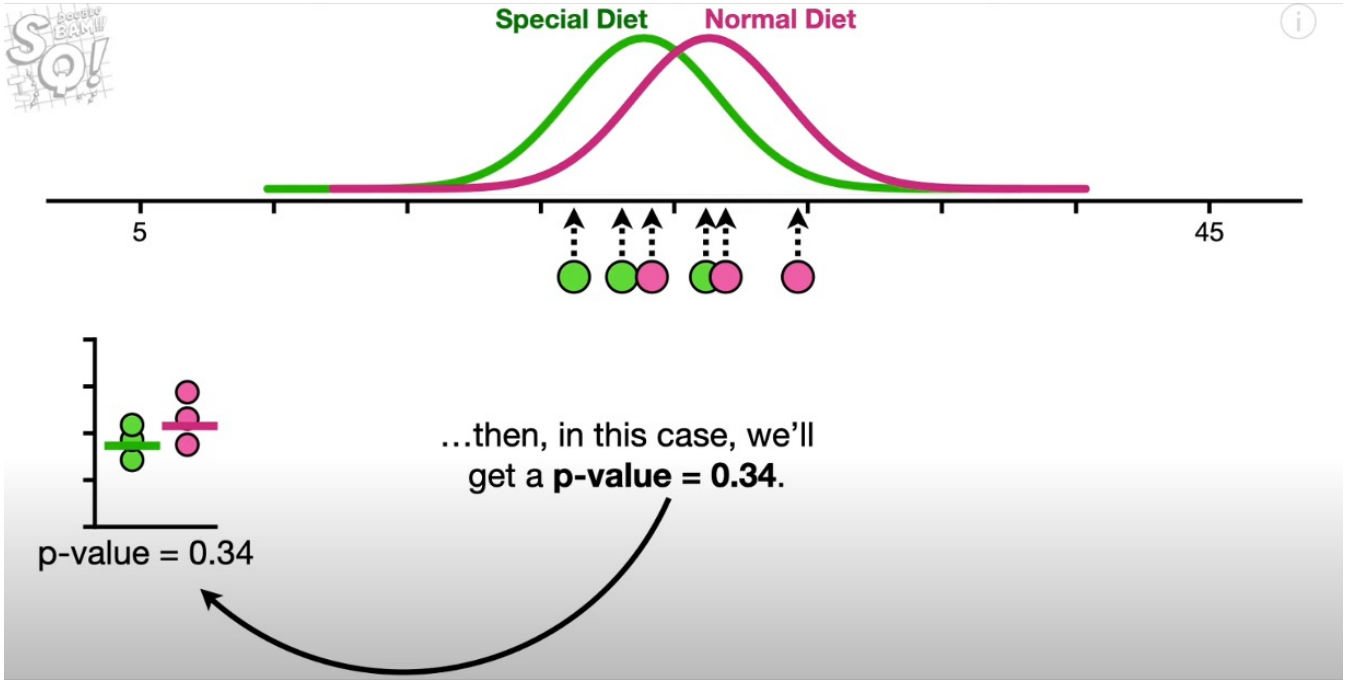
검정력의 의미

거리가 가까운 경우를 추가 예시로 들어보면 어떻게 될까?



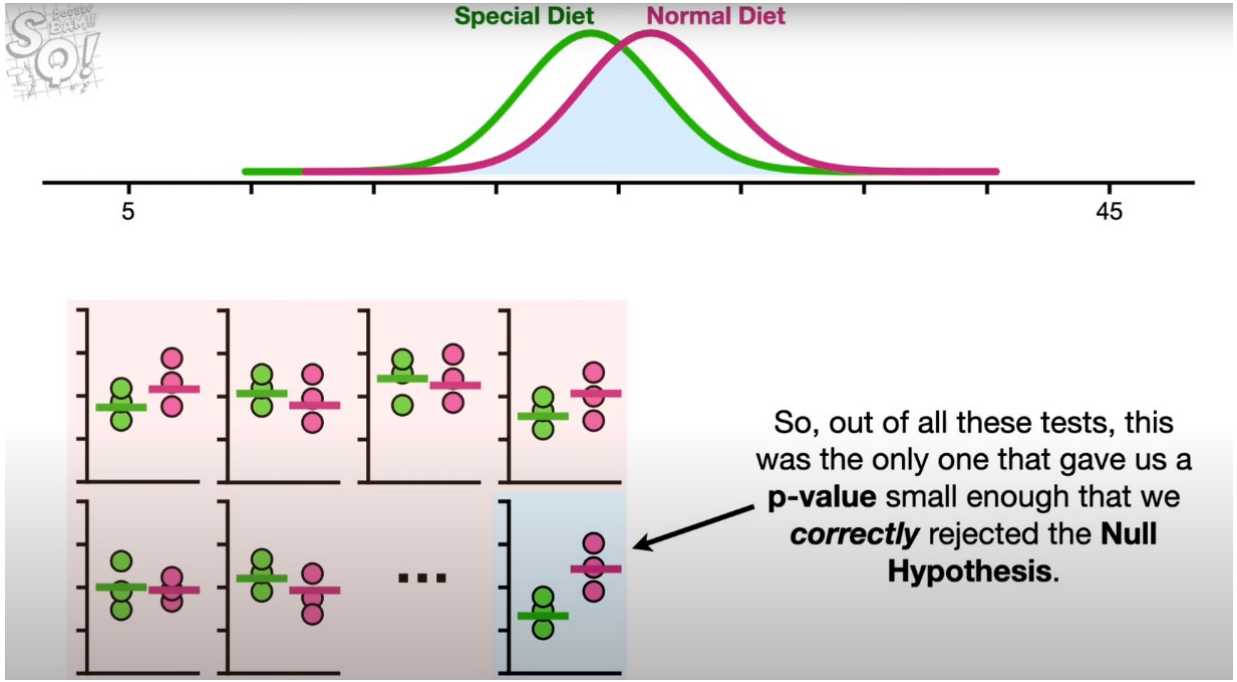
Source: Statquest, Statistical Power, Cleary Explained!

검정력의 의미



Source: Statquest, Statistical Power, Clearly Explained!

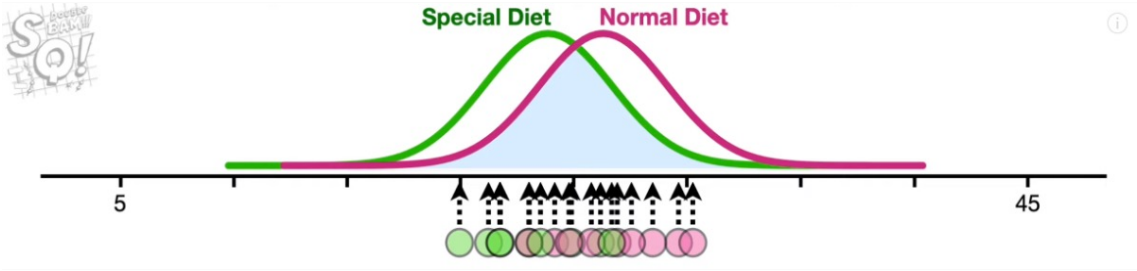
검정력의 의미



Source: Statquest, Statistical Power, Cleary Explained!

검정력을 높이려면 어떻게 해야할까?

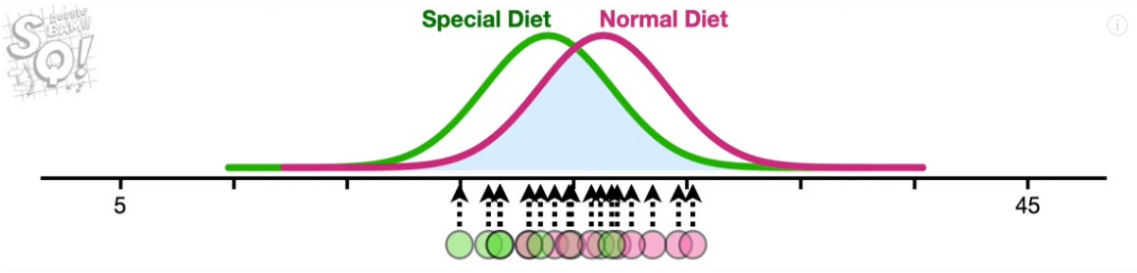
검정력을 높이는 방법



Source: Statquest, Statistical Power, Cleary Explained!

퀴즈 1.
이렇게 검정력이 낮은 상황은
어떻게 극복할 수 있을까?

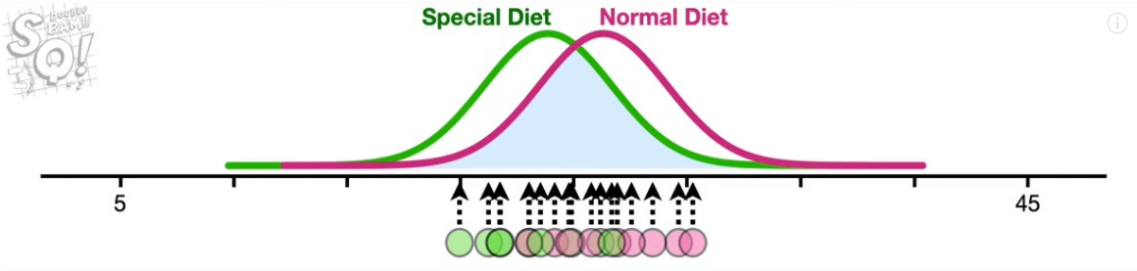
검정력을 높이는 방법



Source: Statquest, Statistical Power, Cleary Explained!

퀴즈 1.
이렇게 검정력이 낮은 상황은
어떻게 극복할 수 있을까?
표본 크기를 늘리자!

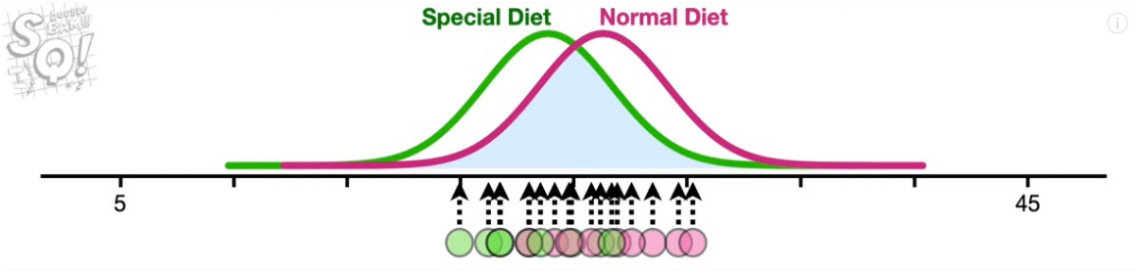
검정력을 높이는 방법



Source: Statquest, Statistical Power, Cleary Explained!

퀴즈 2.
그래서, 얼마나 늘리면 충분한가요?..
(산업 표준 검정력 = 80%)

검정력을 높이는 방법



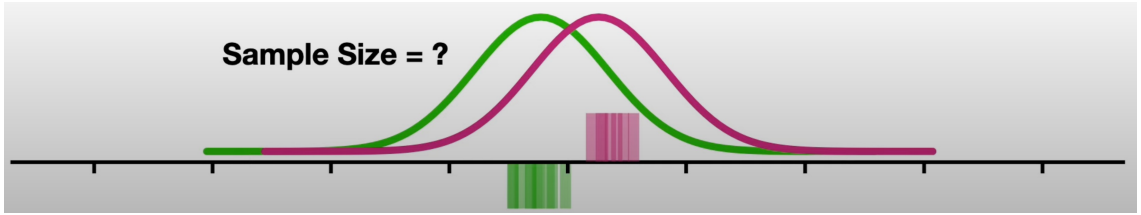
Source: Statquest, Statistical Power, Cleary Explained!

퀴즈 2.
그래서, 얼마나 늘리면 충분한가요?..
(산업 표준 검정력 = 80%)
검정력 분석의 모티베이션

2 검정력 분석

검정력 분석이란?

특정 수준의 검정력을 갖기 위해 얼마나 많은 표본을 수집해야하는지 알려주는 분석 방법론



Source: Statquest, Power Analysis, Cleary Explained!

검정력 분석 수식 정의

대조군, 실험군 균등비율 및 양측 검정이라는 가정 하에 검정력 분석 식은 다음과 같이 정의 됨

$$n = \frac{2\sigma^2\left(Z_{1-\beta} + Z_{1-\frac{\alpha}{2}}\right)^2}{\delta^2}$$

자세한 유도 과정은 Kohavi et al., 2022 참고

검정력 분석 수식 정의

대조군, 실험군 균등비율 및 양측 검정이라는 가정 하에 검정력 분석 식은 다음과 같이 정의 됨

$$n = \frac{2\sigma^2\left(Z_{1-\beta} + Z_{1-\frac{\alpha}{2}}\right)^2}{\delta^2}$$

여기서

- n: 각 변형군에서 확보해야 하는 최소 표본 크기
- σ^2: 수행할 실험의 대조군에 관해 구해진 관심 지표의 분산
- δ: 최소 검출 가능 효과 (MDE, Minimum detectable effect)
- $1-\beta$: 검정력
- α: 유의수준 (1종 오류의 최대 허용한계)

검정력 분석 수식 정의

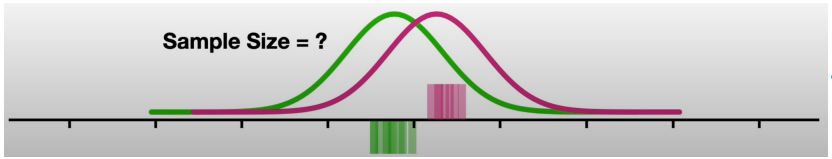
대조군, 실험군 균등비율 및 양측 검정이라는 가정 하에 검정력 분석 식은 다음과 같이 정의 됨

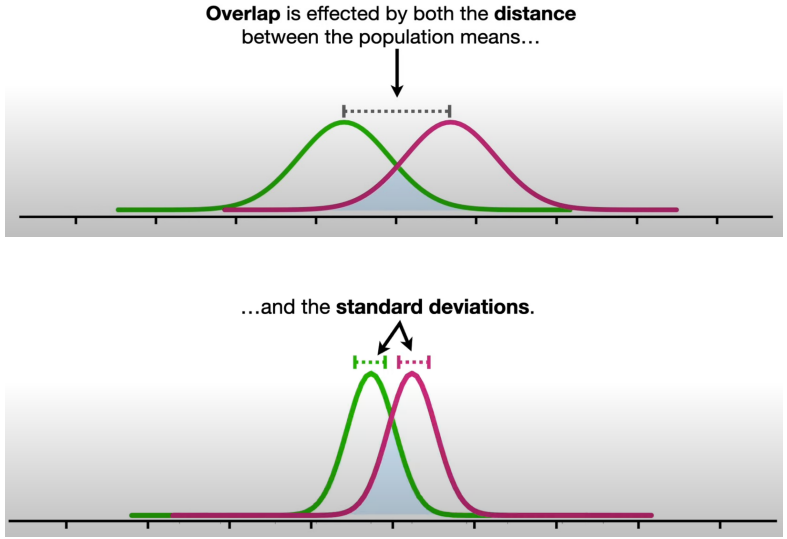
$$n = \frac{2\sigma^2\left(Z_{1-\beta} + Z_{1-\frac{\alpha}{2}}\right)^2}{\delta^2}$$

여기서

- n: 각 변형군에서 확보해야 하는 최소 표본 크기
- σ^2: 수행할 실험의 대조군에 관해 구해진 관심 지표의 분산
- δ: 최소 검출 가능 효과 (MDE, Minimum detectable effect)
- $1-\beta$: 검정력
- α: 유의수준 (1종 오류의 최대 허용한계)

그림으로 살펴보는 검정력 분석





Source: Statquest, Power Analysis, Cleary Explained!

산업 표준의 검정력 분석

여기서 산업 표준인 검정력$(1-\beta)$ 80%, 유의수준(α) 5%를 대입하면 다음과 같이 식이 간단해짐

$$n = \frac{16\sigma^2}{\delta^2}$$

산업 표준의 검정력 분석

$$n = \frac{16\sigma^2}{\delta^2}$$

여기서 포인트는

- 검출해내고 싶은 효과(δ)가 얼마인지에 따라 필요한 표본 크기는 달라진다는 것
- 수행할 실험의 대조군에 관해 구해진 관심 지표의 분산(σ^2)에 따라 필요한 표본 크기는 달라진다는 것

검정력 분석 계산 예시

$$n = \frac{16\sigma^2}{\delta^2}$$

다음의 상황을 가정

- 구매 전환율에 관한 10%의 상대적 변화를 한 번의 실험에서 검출하고 싶음
- 이때 실험에서 대조군으로 사용될 그룹의 구매 전환율은 3.7%였음

검정력 분석 계산 예시

$$n = \frac{16\sigma^2}{\delta^2}$$

다음의 상황을 가정

- 구매 전환율에 관한 10%의 상대적 변화를 한 번의 실험에서 검출하고 싶음
- 이때 실험에서 대조군으로 사용될 그룹의 구매 전환율은 3.7%였음

$$\sigma^2 = p * (1 - p) = 3.7\% * (1 - 3.7\%) = 3.563\%$$

$$\delta = 3.7\% * 10\% = 0.37\%$$

검정력 분석 계산 예시

$$n = \frac{16\sigma^2}{\delta^2}$$

다음의 상황을 가정

- 구매 전환율에 관한 10%의 상대적 변화를 한 번의 실험에서 검출하고 싶음
- 이때 실험에서 대조군으로 사용될 그룹의 구매 전환율은 3.7%였음

$$\sigma^2 = p * (1 - p) = 3.7\% * (1 - 3.7\%) = 3.563\%$$

$$\delta = 3.7\% * 10\% = 0.37\%$$

이에 따라 산업 표준 하의 검정력 분석에 기반해 계산된 두 그룹 각각에 필요한 표본 크기는:

$$n = \frac{16\sigma^2}{\delta^2} = \frac{16 * 3.563\%}{(0.37\%)^2} = 41{,}642$$

검정력 분석 계산 예시

$$n = \frac{16\sigma^2}{\delta^2} = \frac{16 * 3.563\%}{(0.37\%)^2} = 41{,}642$$

즉, 본 예제에서 귀무가설을 올바르게 기각시킬 확률(=검정력)을 **최소 80% 이상 보장**하기 위해 필요한 각 실험 변형군의 최소 표본 수는 41,642명

3 검정력이 낮은 실험에서 발생하는 현상

검정력이 낮은 실험에서 발생하는 오류

검정력이 낮은 실험의 검정 결과는 신뢰할 수 없음 (Gelman et al. 2014)

2종 오류(β)와 또 다른 2가지 형태의 오류율이 높아지기 때문

- Type S(Sign) error rate
- Type M error rate (Exaggeration ratio)

검정력이 낮은 실험에서 발생하는 오류

Type S error rate

- 추정 자체에서 효과의 방향을 반대로 잘못 추정해버리는 경우
- 예를 들어 새롭게 개발한 신약이 질병을 악화시키는 효과가 있는데,
- 통계 분석 결과 그 약이 질병을 개선한다고 잘못 결론을 내리는 경우

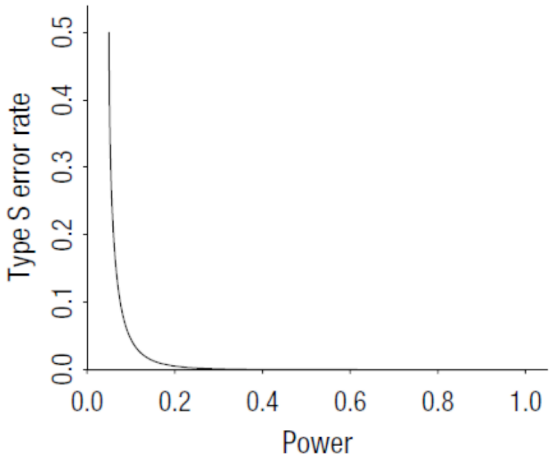


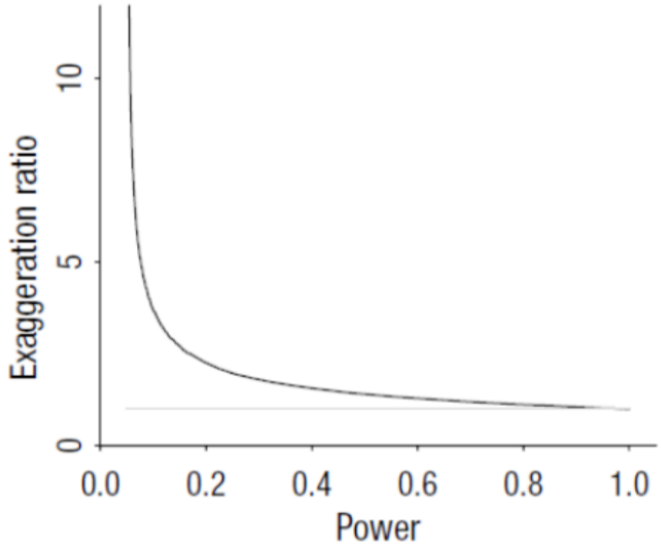
Figure 1: Type S (sign) error of the treatment effect as a function of statistical power (Gelman and Carlin 2014)

Source: Kohavi et al., 2022

검정력이 낮은 실험에서 발생하는 오류

Type M error rate

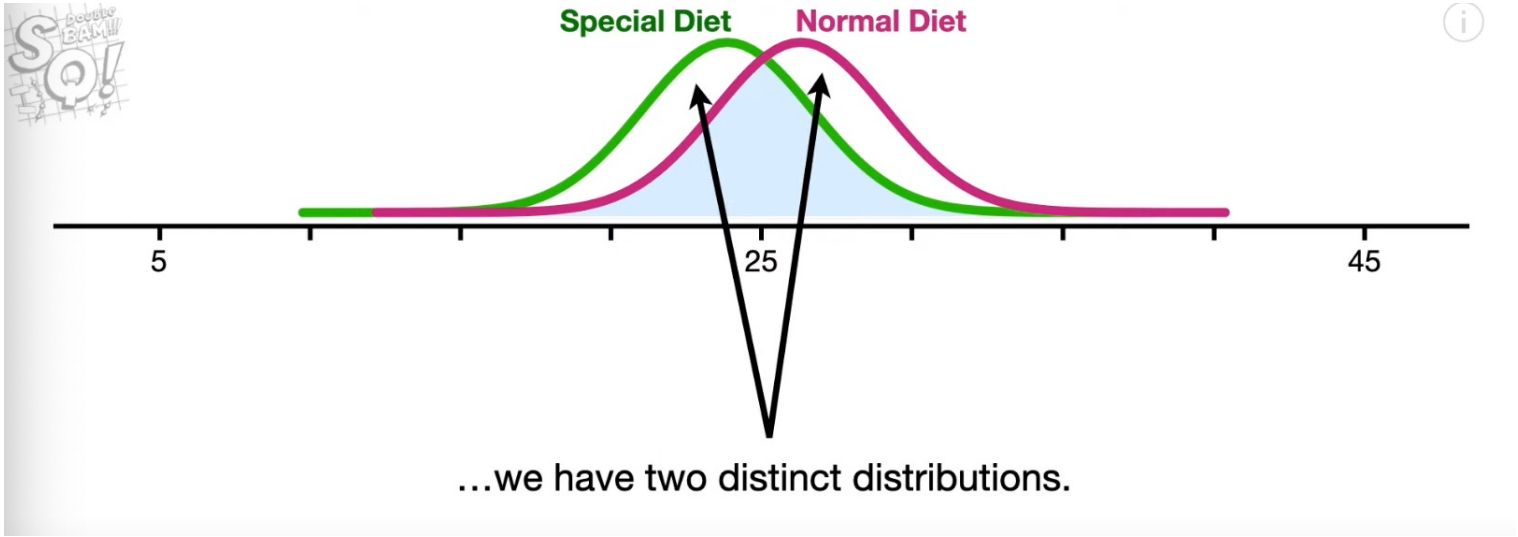
- 효과의 중요도가 과대평가될 수 있음



Source: Kohavi et al., 2022

그림으로 이해하면 쉬워요

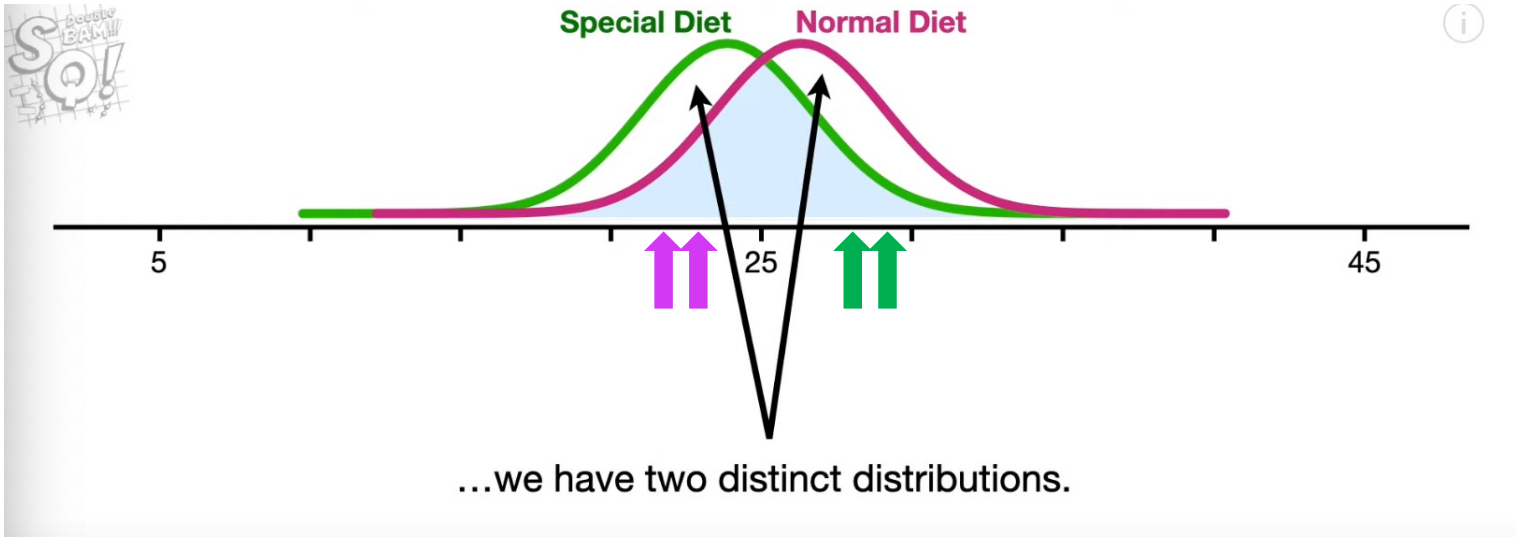
설명의 편의상 각 변형군으로부터 표본을 2마리씩 채취한다고 하자



Source: Statquest, Statistical Power, Cleary Explained!

그림으로 이해하면 쉬워요

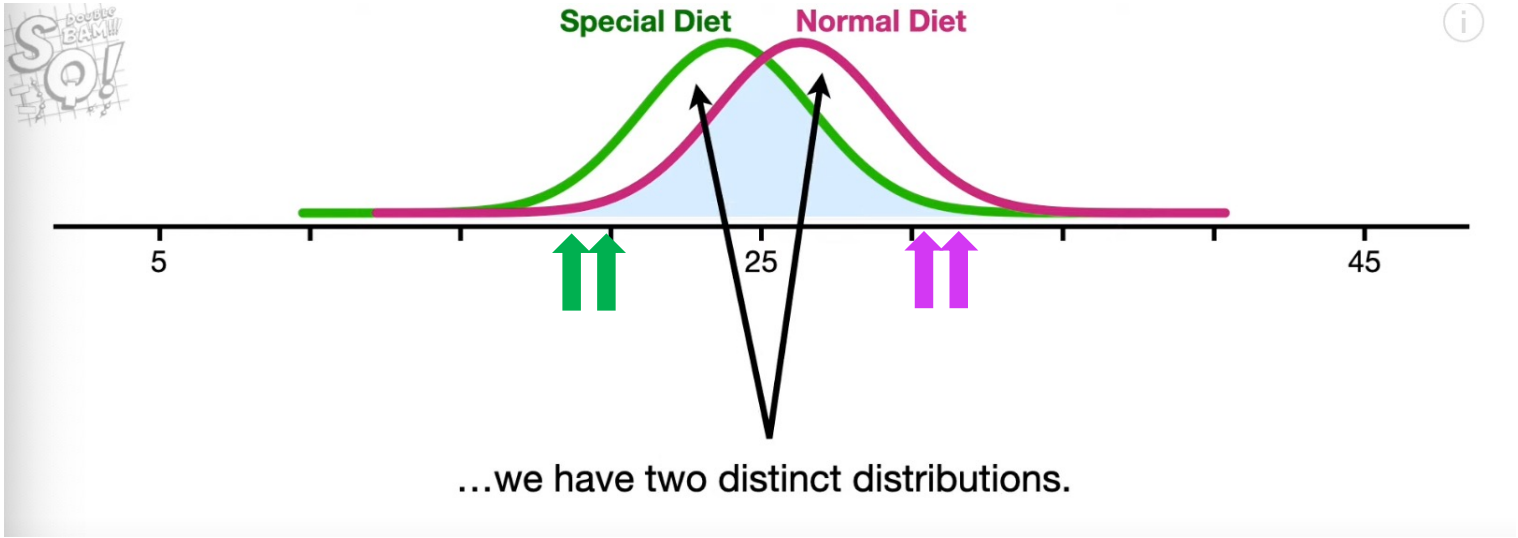
Type S error rate



Source: Statquest, Statistical Power, Cleary Explained!

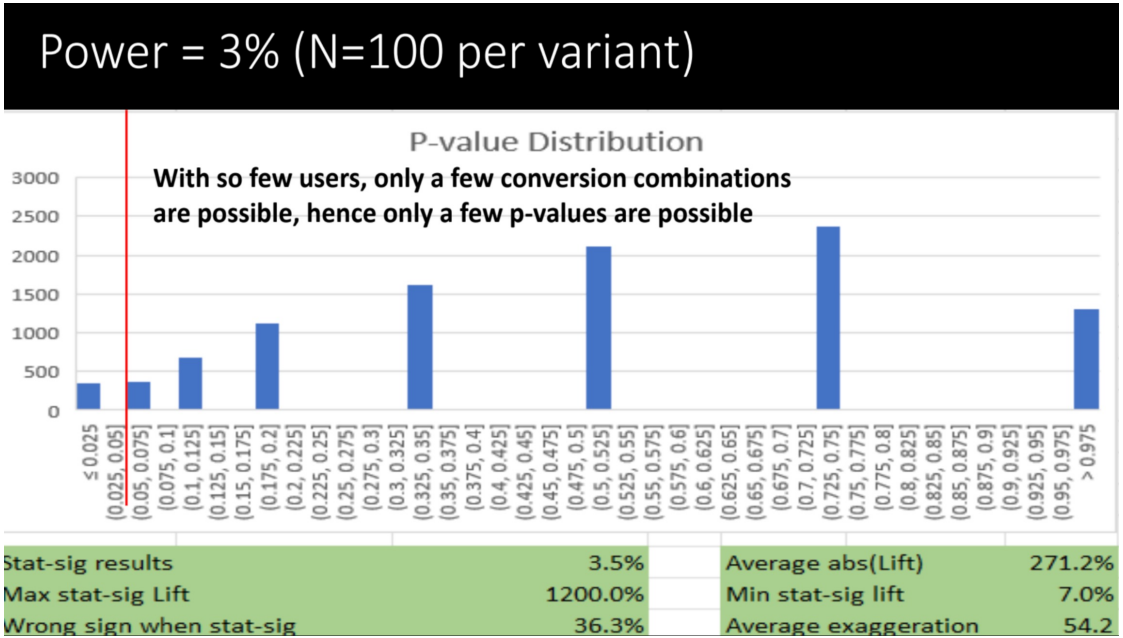
그림으로 이해하면 쉬워요

Type M error rate



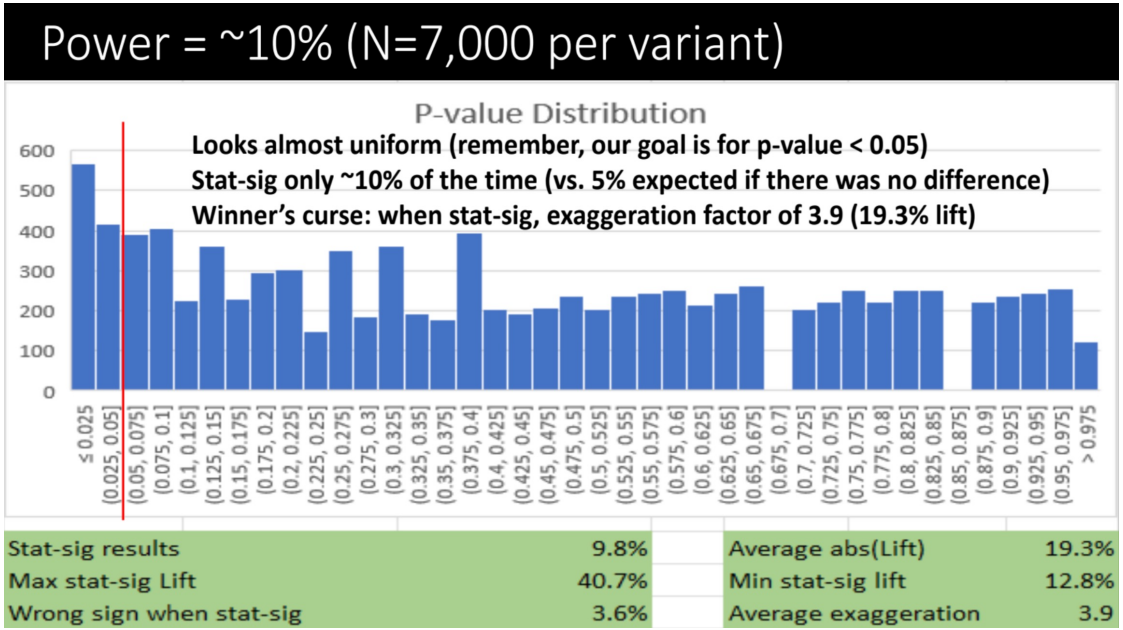
Source: Statquest, Statistical Power, Cleary Explained!

시뮬레이션 스터디 사례



Source: Kohavi, Practical defaults for A/B Testing, 2022

시뮬레이션 스터디 사례

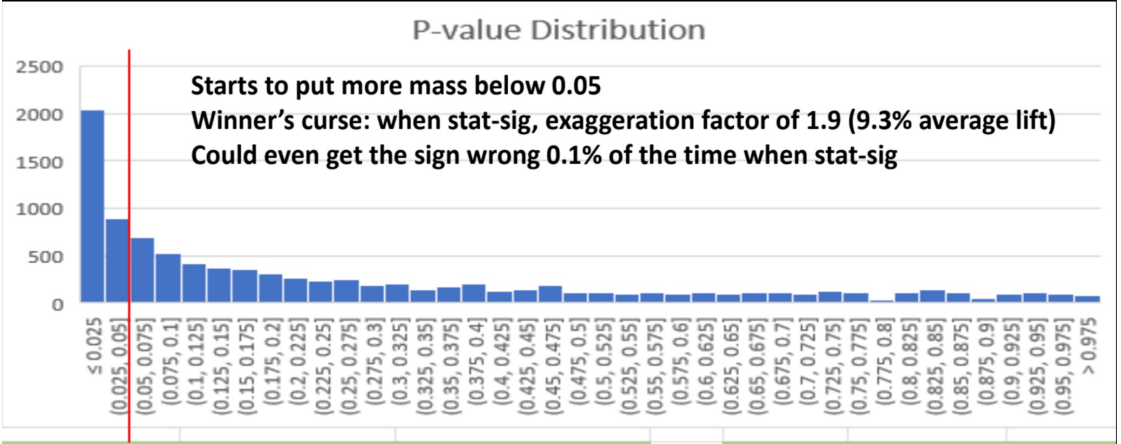


Stat-sig results	9.8%	Average abs(Lift)	19.3%
Max stat-sig Lift	40.7%	Min stat-sig lift	12.8%
Wrong sign when stat-sig	3.6%	Average exaggeration	3.9

Source: Kohavi, Practical defaults for A/B Testing, 2022

시뮬레이션 스터디 사례

Power = ~30% (N=32,000 per variant)

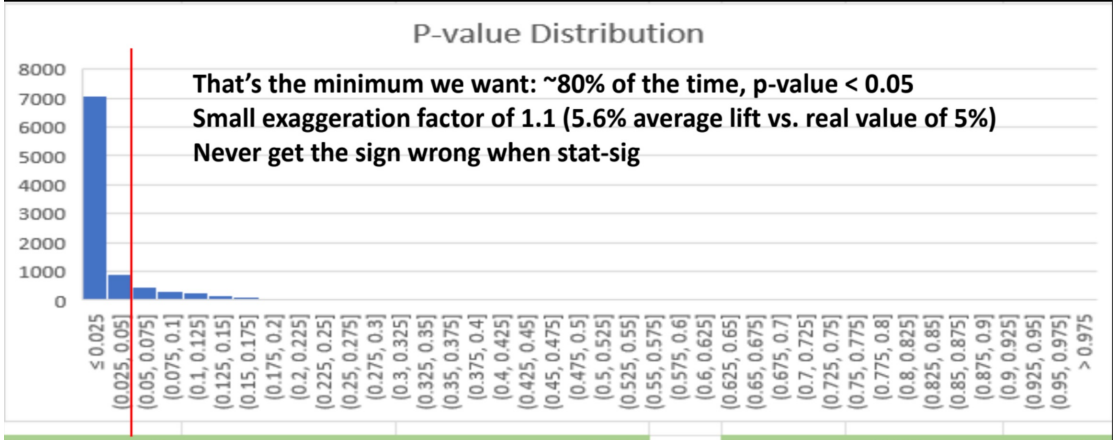


Stat-sig results	29.5%	Average abs(Lift)	9.3%
Max stat-sig Lift	20.2%	Min stat-sig lift	6.6%
Wrong sign when stat-sig	0.1%	Average exaggeration	1.9

Source: Kohavi, Practical defaults for A/B Testing, 2022

시뮬레이션 스터디 사례

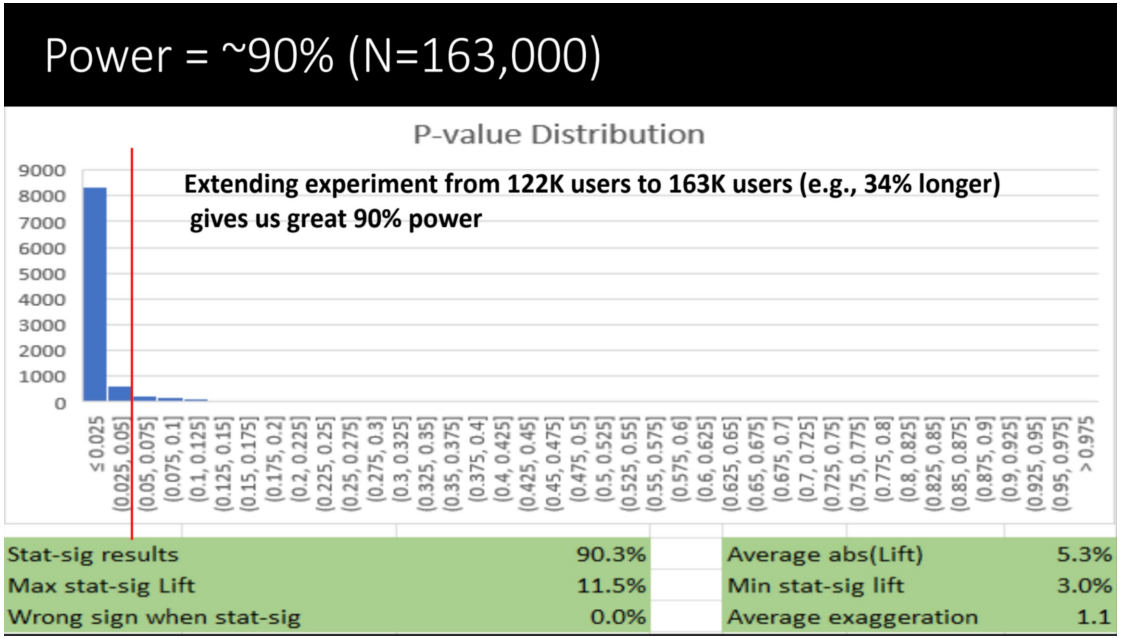
Power = ~80% (N=122,000)



Stat-sig results	79.6%	Average abs(Lift)	5.6%
Max stat-sig Lift	12.5%	Min stat-sig lift	3.4%
Wrong sign when stat-sig	0.0%	Average exaggeration	1.1

Source: Kohavi, Practical defaults for A/B Testing, 2022

시뮬레이션 스터디 사례



Stat-sig results	90.3%	Average abs(Lift)	5.3%
Max stat-sig Lift	11.5%	Min stat-sig lift	3.0%
Wrong sign when stat-sig	0.0%	Average exaggeration	1.1

Source: Kohavi, Practical defaults for A/B Testing, 2022

즉 정리하면,

검정력이 낮은 실험의 검정 결과는
신뢰할 수 없습니다.

그래서, 검정력 분석에 기반해
실험을 설계하는 것은 굉장히 중요합니다.

앞선 이야기 한 정상적인 형태의 검정력 분석을
사전 검정력 분석이라 정의하겠습니다.

다음으로는 우리를 함정에 빠지게 만드는
사후 검정력 분석에 대해 이야기하고자 합니다.

4 사후 검정력 분석의 함정

사후 검정력 분석을 수행하는 흔한 상황

사후 검정력 분석은 언제 수행하게 될까요?

사후 검정력 분석을 수행하는 흔한 상황

사후 검정력 분석은 언제 수행하게 될까요?

➔ 흔히 검정력 분석에 기반해 설계되지 않은 실험을 진행한 경우에 수행하게 됩니다.

➔ 특히, 귀무가설을 기각시키지 못한 결정을 내렸을때 말이죠.

➔ 이런 의심을 하는거죠. “검정력이 너무 낮아서 귀무가설을 기각시키지 못한 것 아니야?”

사후 검정력 분석을 수행하는 흔한 상황

반대로, 귀무가설을 기각시킨 경우 검정력 따위는 신경쓰지 않고 넘어가곤 합니다.

사후 검정력 분석을 수행하는 흔한 상황

반대로, 귀무가설을 기각시킨 경우 검정력 따위는 신경쓰지 않고 넘어가곤 합니다.

➔ 내가 원하는 결과를 얻었기 때문이죠.

➔ 그러나, 귀무가설을 기각시킨 상황에도 검정력이 얼마인지는 중요합니다.

➔ Type M/S error에 의해 과장된 또는 반대의 효과를 보이는 경우가 발생할 수 있기 때문이죠.

사후 검정력 분석을 수행하는 흔한 상황

사전 검정력 분석은 앞서 보았듯 수행할 실험의 대조군에 관해 구해진 관심 지표의 분산(σ^2)과 MDE(δ)에 의존합니다.

$$n = \frac{2\sigma^2\left(Z_{1-\beta} + Z_{1-\frac{\alpha}{2}}\right)^2}{\delta^2}$$

사후 검정력 분석을 수행하는 흔한 상황

사전 검정력 분석은 앞서 보았듯 수행할 실험의 대조군에 관해 구해진 관심 지표의 분산(σ^2)과 MDE(δ)에 의존합니다.

$$n = \frac{2\sigma^2\left(Z_{1-\beta} + Z_{1-\frac{\alpha}{2}}\right)^2}{\delta^2}$$

문제는 대조군의 분산에 관한 정보가 없는 경우입니다.

➔ 그래서, 우리는 자연스럽게 이런 사고를 하게됩니다.

➔ 실험에서 관측한 데이터로 검정력 분석을 하면 되지 않을까?

➔ 사후 검정력 분석이라는 끔찍한 괴물이 탄생합니다.

사후 검정력 분석이란?

사전 검정력 분석과 사후 검정력 분석은 다음과 같이 정리 됩니다.

- 사전 검정력 분석 (Priori power analysis)
  - 실험 수행 전 해당 실험의 대조군에 관해 구해진 관심 지표의 분산(σ^2)과 MDE(δ)에 의존
- 사후 검정력 분석 (Post-hoc power analysis)
  - 실험 수행 후 관측된 데이터에 기반한 p-value에 의존

사후 검정력 분석이란?

사전 검정력 분석과 사후 검정력 분석은 다음과 같이 정리 됩니다.

- 사전 검정력 분석 (Priori power analysis)
  - 실험 수행 전 해당 실험의 대조군에 관해 구해진 관심 지표의 분산(σ^2)과 MDE(δ)에 의존
- 사후 검정력 분석 (Post-hoc power analysis)
  - 실험 수행 후 관측된 데이터에 기반한 p-value에 의존

설명의 편의상 둘을 다른 용어로 구분지어 말하고 있으나,

후자는 완전히 잘못된 형태의 검정력 분석

사후 검정력 분석의 함정

실험 데이터로 검정력을 계산하면 어떤 문제가 발생하나요?

$$n = \frac{2\sigma^2\left(Z_{1-\beta} + Z_{1-\frac{\alpha}{2}}\right)^2}{\delta^2}$$

사전 검정력 분석 식으로부터 사후 검정력 분석 식을 유도할 수 있습니다.

$$1 - \beta = \Phi\left(\frac{\delta}{SE} - Z_{1-\frac{\alpha}{2}}\right),$$
$$where\ \Phi\ is\ a\ cdf\ of\ standard\ normal\ distribution$$

이처럼 사후 검정력 분석은 실험에서 관측한 p-value에 의존하게됨

$$\therefore 1 - \beta = \Phi\left(Z_{1-\frac{pval}{2}} - Z_{1-\frac{\alpha}{2}}\right)$$

자세한 유도 과정은 슬라이드 맨 뒷 장의 Appendix 참고

사후 검정력 분석의 함정

P-value에 의존하는게 무슨 문제를 발생시키는가?

사후 검정력 분석의 함정

P-value에 의존하는게 무슨 문제를 발생시키는가?

➔ 앞선 사전 검정력 분석 예제를 떠올려보자

$$n = \frac{16\sigma^2}{\delta^2} = \frac{16 * 3.563\%}{(0.37\%)^2} = 41{,}642$$

➔ 80% 검정력을 보장하기 위해 필요한 각 그룹의 표본 수는 41,642명 이었음

사후 검정력 분석의 함정

- 사후 검정력 분석의 함정을 보여주기 위해 변형군 당 80명의 표본 크기만을 확보해 실험 진행
- 실제 논문에서 제시한 예시에 해당 (Kohavi et al., 2022)

사후 검정력 분석의 함정

- 사후 검정력 분석의 함정을 보여주기 위해 변형군 당 80명의 표본 크기만을 확보해 실험 진행
- 실제 논문에서 제시한 예시에 해당 (Kohavi et al., 2022)
- 해당 예제에서 실험 데이터로부터 계산된 관측된 p-value는 0.009
- 이에 따라 사후 검정력 분석 식에 의해 계산된 검정력은 74%에 달함

$$1 - \beta = \Phi\left(Z_{1-\frac{0.009}{2}} - Z_{1-\frac{0.05}{2}}\right) = 74\%$$

사후 검정력 분석의 함정

그러나, 변형군 당 80개의 표본만 확보한 경우 사전 검정력 분석을 해보면...

사후 검정력 분석의 함정

그러나, 변형군 당 80개의 표본만 확보한 경우 검정력 분석을 해보면...

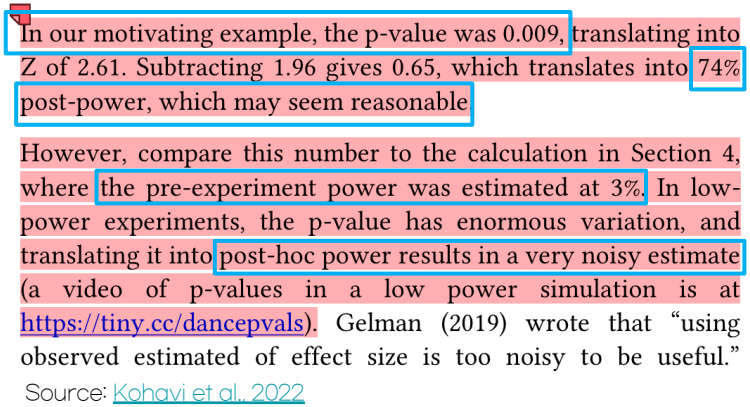
$$\sigma^2 = p * (1-p) = 3.7\% * (1-3.7\%) = 3.563\%$$
$$\delta = 3.7\% * 10\% = 0.37\%$$

$$1-\beta = \Phi\left(\frac{\delta}{SE} - Z_{1-\frac{\alpha}{2}}\right)$$
$$= \Phi\left(0.37\%/\sqrt{\frac{2\sigma^2}{n}} - 1.96\right)$$
$$= \Phi\left(0.37\%/\sqrt{\frac{2 * 3.563\%}{80}} - 1.96\right)$$
$$= 3.3\%$$

실제 해당 실험의 검정력은 3.3% 불과하다는 것을 알게 되죠.

사후 검정력 분석의 함정

논문에서는 이렇게 이야기 합니다.



In our motivating example, the p-value was 0.009, translating into Z of 2.61. Subtracting 1.96 gives 0.65, which translates into 74% post-power, which may seem reasonable.

However, compare this number to the calculation in Section 4, where the pre-experiment power was estimated at 3%. In low-power experiments, the p-value has enormous variation, and translating it into post-hoc power results in a very noisy estimate (a video of p-values in a low power simulation is at https://tiny.cc/dancepvals). Gelman (2019) wrote that "using observed estimated of effect size is too noisy to be useful."

Source: Kohavi et al., 2022

"실험에서 관측된 effect size를 기반으로 검정력 분석을 수행하는 것은 잡음이 굉장히 커서, 전혀 쓸모가 없다."

- Andrew Gelman (2019)

사후 검정력 분석의 함정

사후 검정력 분석은 무의미합니다.

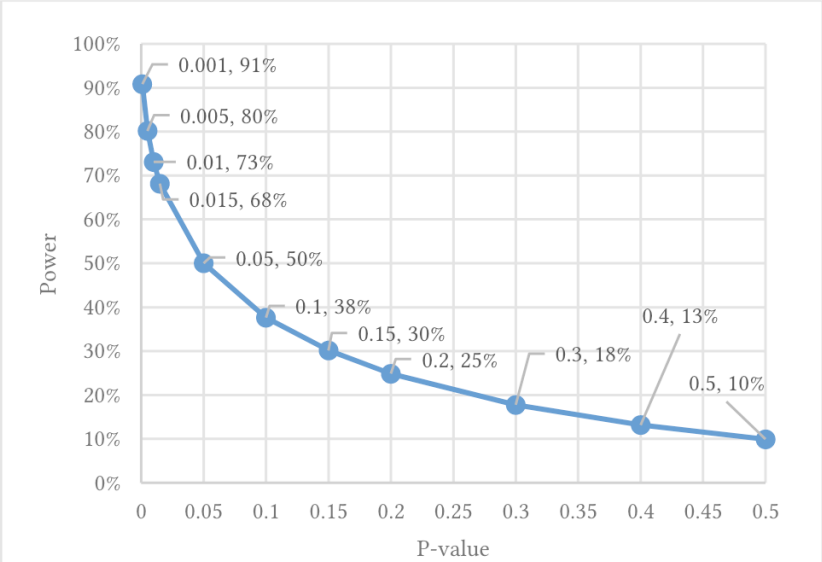


Figure 3: post-hoc power is determined by p-value

Source: Kohavi et al., 2022

사후 검정력 분석의 함정

앞서 수식으로 복잡하게 설명했지만.. 생각해보면 당연한 결과입니다.

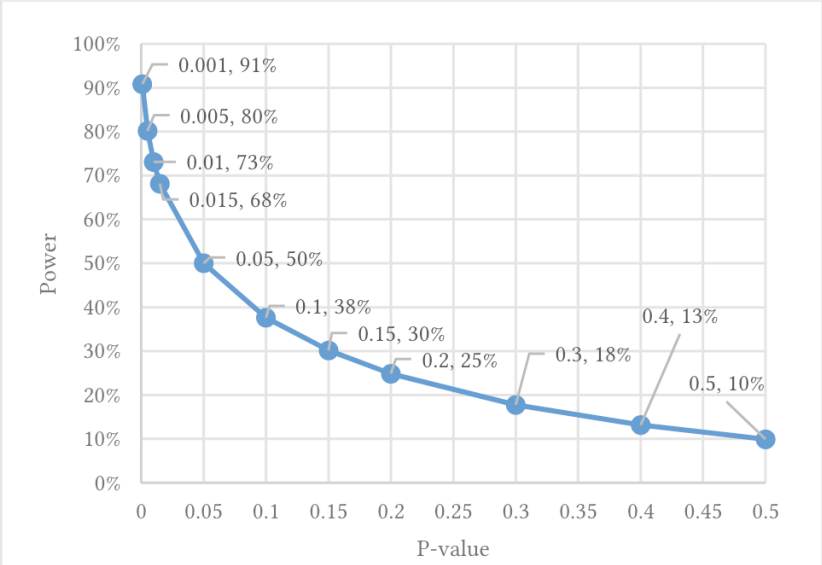


Figure 3: post-hoc power is determined by p-value

Source: Kohavi et al., 2022

사후 검정력 분석의 함정

충분한 표본을 확보하지 못한 실험에서 관측된 p-value에 기반해 검정력 분석을 수행한다?..

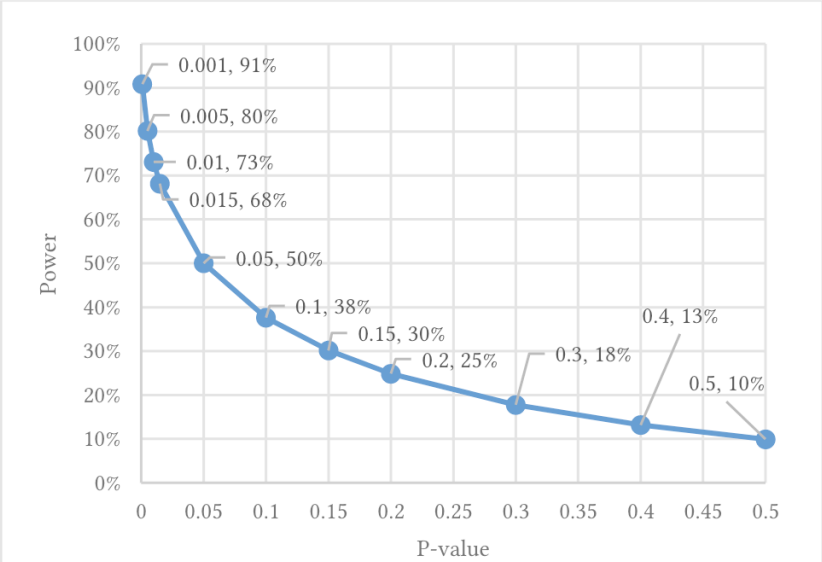


Figure 3: post-hoc power is determined by p-value

Source: Kohavi et al., 2022

사후 검정력 분석의 함정

Type S/M error로 오염된 잘못된 값을 기반으로 검정력 분석을 하는 행위라고 할 수 있죠.

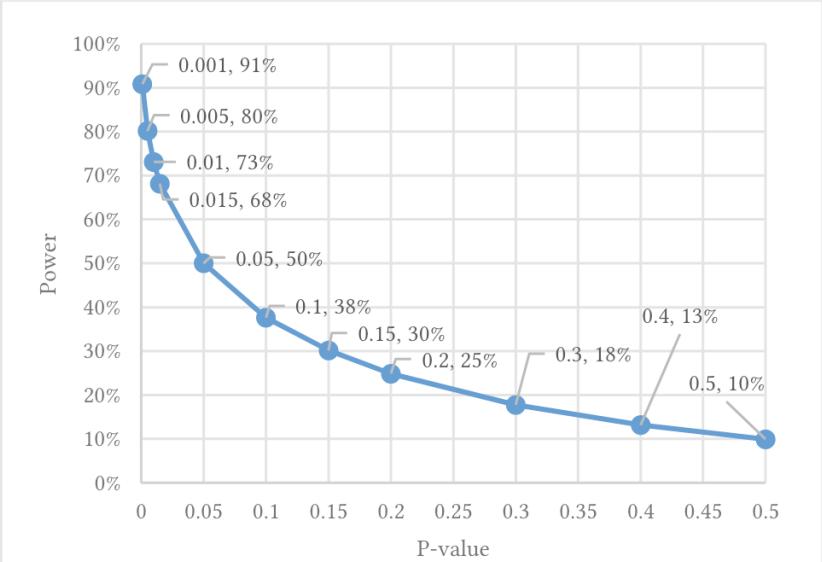


Figure 3: post-hoc power is determined by p-value

Source: Kohavi et al., 2022

그럼, 과거 데이터가 없는 경우..
정상적인 검정력 분석이 불가능한데 어떡하나요?

2가지 해결책이 있습니다.

- 과거 관련 연구 찾아보기
- 파일럿 스터디

다행히! 온라인 서비스의 경우 특수한 경우를 제외하고 검정력 분석이 가능합니다.

기존 시스템에서 활동하고 있는 고객들의
행동, 거래 데이터를 적재하고 있기 때문이죠.

오늘 배운 이야기 요약

검정력 분석을 통해 우리는

"충분한 표본을 확보하여 검정력과 유의수준을 밸런스있게 통제함으로써,
신뢰할 수 있는 통계적 가설 검정 결과를 제공할 수 있다."

2가지 포인트

- 실험 전 검정력 분석은 선택이 아닌 필수
- 사후 검정력 분석의 함정에 빠지지 말자

(번외) 온라인 실험에서의 검정력 분석

온라인 실험에서의 검정력 분석

검정력 분석에서는 관심지표의 표준편차와 MDE에 의해 표본 크기가 결정되죠.

$$n = \frac{16\sigma^2}{\delta^2}$$

퀴즈) 온라인 실험에서의 표본 크기는 2가지에 의존하는데, 이 2가지는 무엇일까요?

온라인 실험에서의 검정력 분석

검정력 분석에서는 관심지표의 표준편차와 MDE에 의해 표본 크기가 결정되죠.

$$n = \frac{16\sigma^2}{\delta^2}$$

퀴즈) 온라인 실험에서의 표본 크기는 2가지에 의존하는데, 이 2가지는 무엇일까요?

1. 트래픽
2. 실험 기간

온라인 실험에서의 검정력 분석

검정력 분석에서는 관심지표의 표준편차와 MDE에 의해 표본 크기가 결정되죠.

$$n = \frac{16\sigma^2}{\delta^2}$$

퀴즈) 온라인 실험에서의 표본 크기는 2가지에 의존하는데, 이 2가지는 무엇일까요?

1. 트래픽
2. 실험 기간

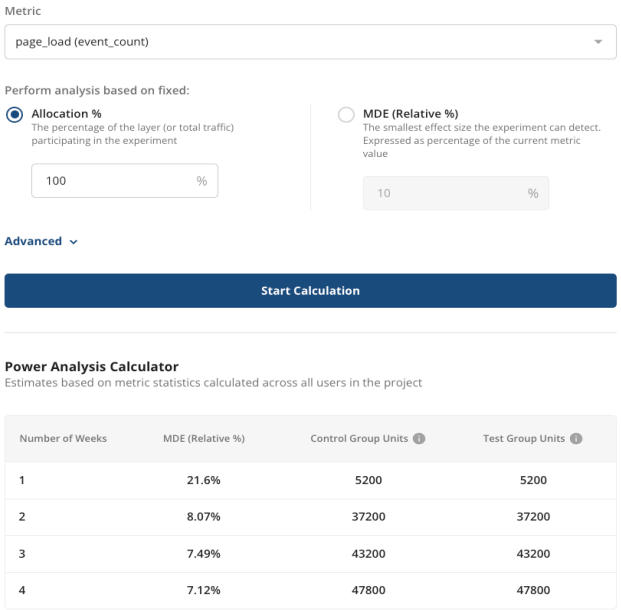
즉, 우리는 검정력 분석으로 각 온라인 실험의 최적 트래픽과 실험 기간을 결정할 수 있습니다.

검정력 분석 제공 방식

데이터야.놀자2024

트래픽 입력

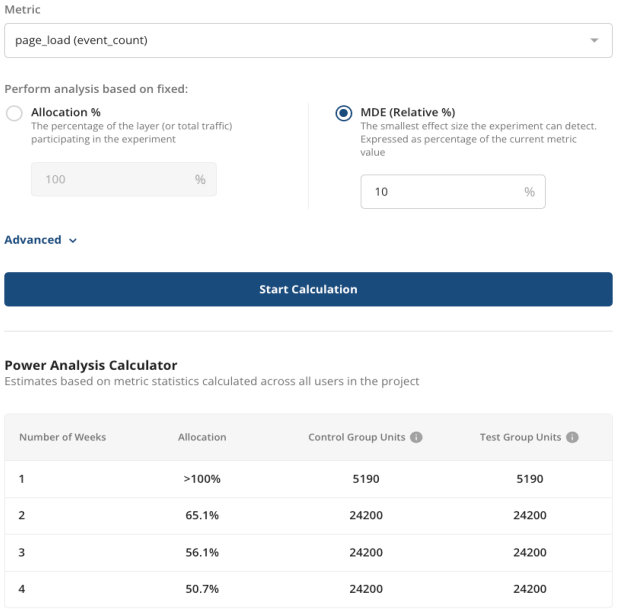
- 실험 기간에 따른 Relative MDE 계산



Source: Stagsig Docs, Power Analysis

Relative MDE 입력

- 실험 기간에 따른 필요 트래픽 계산



Source: Stagsig Docs, Power Analysis

References

References

[1] StatQuest with Josh Starmer (Director). (2020). *Statistical Power, Clearly Explained!!!* https://www.youtube.com/watch?v=Rsc5znwR5FA

[2] Kohavi, R., Deng, A., & Vermeer, L. (2022). A/B Testing Intuition Busters: Common Misunderstandings in Online Controlled Experiments. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3168-3177. https://doi.org/10.1145/3534678.3539160

[3] Kohavi, R. (2022, November). *Practical Defaults For A/B Testing*. Google Docs. https://drive.google.com/file/d/18jukd0M4PgHpBKC_uDFQREMylYSS3qiL/view?usp=sharing&usp=embed_facebook

[4] Gelman, A. (2019). Don't Calculate Post-hoc Power Using Observed Estimate of Effect Size. *Annals of Surgery*, *269*(1), e9. https://doi.org/10.1097/SLA.0000000000002908

[5] Gelman, A., & LOKEN, E. (2014). *The Statistical Crisis in Science*. American Scientist. https://www.americanscientist.org/article/the-statistical-crisis-in-science

Appendix

Appendix

사전 검정력 분석 식으로부터 사후 검정력 분석 식 유도

$$n = \frac{2\sigma^2\left(Z_{1-\beta} + Z_{1-\frac{\alpha}{2}}\right)^2}{\delta^2}$$

$$\frac{2\sigma^2}{n} = \frac{\delta^2}{\left(Z_{1-\beta} + Z_{1-\frac{\alpha}{2}}\right)^2}$$

$$SE = \frac{\delta}{Z_{1-\beta} + Z_{1-\frac{\alpha}{2}}}$$

$$SE * \left(Z_{1-\beta} + Z_{1-\frac{\alpha}{2}}\right) = \delta$$

$$Z_{1-\beta} = \frac{\delta}{SE} - Z_{1-\frac{\alpha}{2}}$$

Appendix

사전 검정력 분석 식으로부터 사후 검정력 분석 식 유도

$$Z_{1-\beta} = \frac{\delta}{SE} - Z_{1-\frac{\alpha}{2}}$$

$$1 - \beta = \Phi\left(\frac{\delta}{SE} - Z_{1-\frac{\alpha}{2}}\right),$$

$$where\ \Phi\ is\ a\ cdf\ of\ standard\ normal\ distribution$$

$$\therefore 1 - \beta = \Phi\left(Z_{1-\frac{pval}{2}} - Z_{1-\frac{\alpha}{2}}\right)$$

Reference

- Kohavi, R., Deng, A., & Vermeer, L. (2022). A/B Testing Intuition Busters: Common Misunderstandings in Online Controlled Experiments. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3168-3177. https://doi.org/10.1145/3534678.3539160