

온라인 종합 대조 실험에 사용되는 통계 이론

가짜연구소 인과추론팀
온라인 통제 실험 연구자로 거듭나기

2024-04-16

조동민



Causal-Lab

흡연은 통계의 필요성을 알려주는 주요 요인 중 하나다.
- 플레처 네벨 -

이분은 왜 이런 말씀을 하셨을까?

Chapter17의 첫 문장

Knebel was married four times from 1935 to 1985. He committed suicide after a long bout with cancer, by taking an overdose of sleeping pills in his home in Honolulu, Hawaii, during 1993.^[1]
He is the source of the quote: "Smoking is one of the leading causes of statistics."

네빌은 미국의 유명한 정치 소설가였음

- 말년에 암으로 고통을 받다가 자살을 택했음.
- Smoking is one of the leading causes of statistics. 이런 말을 남김
- 담배가 암을 유발한다는 것은 통계적으로 검증된 가설임.
- 여기서 “가설 검증”이라는 통계 기법의 주 사용 용도에 대해 생각해볼 수 있다.
- 즉, 통계적으로 검증된 가설이 가지는 강력한 힘(=함부로 무시하면 안됨)
에 대해 말씀하시고 싶었던게 아닐까?

그렇다면, 통계적인 가설 검증은 어떻게 하는 것일까?
그리고, 그 요소들에는 어떤 것들이 있을까?

매주 A/B테스트 스터디를 하고 있는데, 이 스터디 정말로 나한테 도움이 되는걸까?
- 스터디원 조모씨 -



그래 진짜 효과가 있는지 검증해보자!

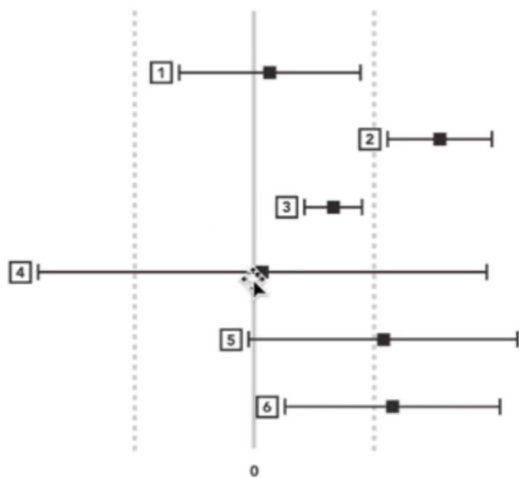


가설 설정

- A/B테스트 스터디에 참여했던 사람들은 A/B테스트 결과 해석 시 통계적 유의도와 실무적 유의도를 모두 고려하여 해석할 것이다. (연구 가설)
- 스터디에 참여했던 사람들과 참여하지 않은 사람들 간에 큰 차이가 없을 것이다. (영가설)

가설 측정 지표 정의(Proxy 지표)

- 면접에서 하단 그래프를 보고 기능 출시/실험 지속에 대한 질문을 받음(제한 시간은 30초)
 - 위 상황에서 실무적/통계적 유의도를 모두 고려하여 답변하면 1
 - 그렇지 않으면 0으로 기록



실험 결과

- A/B테스트 스터디에 참여했던 분석가 10명
 - 1, 1, 1, 1, 1, 1, 1, 1, 0, 0 => 80%가 올바른 답변
- A/B테스트 스터디에 참여하지 않았던 분석가 10명
 - 1, 1, 0, 1, 0, 0, 1, 0, 1, 0 => 50%가 올바른 답변

30%p 차이면 굉장히 의미 있는거 아냐?



30%p 차이면 굉장히 의미 있는거 아냐?라고 단순하게 생각하면 안되는 것이..

- 지금 같은 상황에서는 아래 진술 2 가지 다 가능
 - A. 30%p 차이면 충분히 의미있지!
 - B. 근데, 퇴근하면 집에가서 발 뺀고 누워 자고 싶지
A/B테스트 스테디에 참여했다는 것 자체가 굉장히 특이한 거 아냐?
그래서, 그 사람들이 더 정확한 해석을 한 거 아냐?
- 여기서 B의 주장을 표집오차(Sampling error)라고 한다.
 - 정답률이 잘못 됐다는 의미 보다는 샘플링이 특이하게 됐다 라는 의미

이때 등장하는 개념이 표집분포(Sampling distribution)

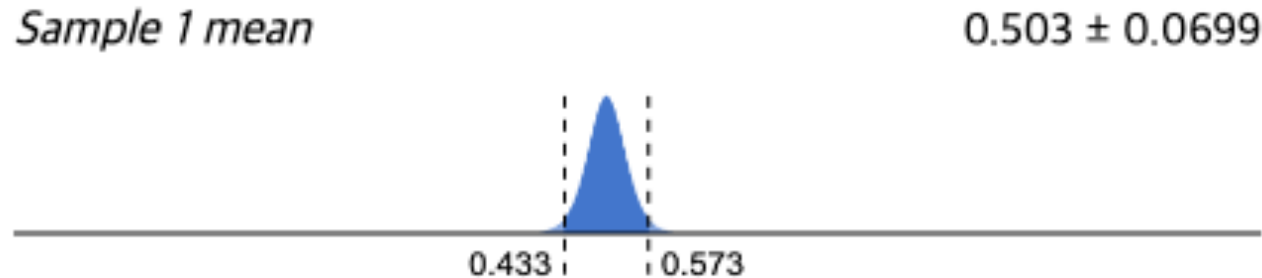


표본 분포(sample distribution)와 표집 분포(sampling distribution)의 차이

- A/B테스트 스터디에 참여했던 분석가 10명
 - 1, 1, 1, 1, 1, 1, 1, 1, 0, 0 (**표본분포**) => 80% (**표본통계치**)
- 50%, 80%, 30%, 50%, ..는 **평균의 표집분포**
 - 즉, **표집분포**는 **표본통계치들의 분포**이며 여기에는 특정 모집단에서 반복적으로 표본을 샘플링했다는 가정이 깔려있음

표집 분포가 있으면, 확률적인 판단을 할 수가 있음 -> 딱 봐도 정답률 80%까지 나오기는 힘들겠는데?

- A/B테스트 스터디에 참여하지 않은 분석가의 정답률 (10명씩 무작위로 10번 추출하여 측정)
 - 50%, 30%, 60%, 50%, 50%, 50%, 55%, 65%, 53%, 40%



가설 검증의 기본 논리

표준화 작업 = 검증 통계치로 만드는 작업

- 평균을 0으로, 분산을 1으로..

$$Z = \frac{X - \mu}{\sigma}$$

왜 표준화를...? 평균/분산이 계속 바뀌면 판단하기 귀찮으니까

- 평균이 0, 분산이 1인 표준 정규 분포로 만들고 아래 표에서 값을 해석
- Z가 1.96 이하일 확률이 97.5% 1.96을 초과할 확률은 2.5% (= p-value)

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

표준정규분포표 = 누적분포함수 $P(Z \leq z)$

근데, 어느 세월에 저렇게 많이 샘플링을 해본담..

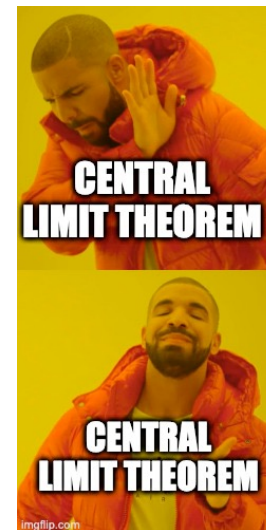


이때 등장하는 개념이 중심극한정리(Center limit theorem)



중심 극한정리를 가장 잘 설명해주는 사진

- 너가 이전에 어떤 분포였는지는 관심 없다.
- 대신에 충분한 표본을 확보해라
- 그리고 표본 통계치만 가져와라
- 그럼, 너(=표본통계치)도 이제 정규분포!



n=29

n=30

중심 극한정리

- 특정 전집에서 무작위(iid, 독립항등분포)로 샘플링 할 때, 샘플링하는 표본의 크기가 크면 클수록
- 표본 평균의 분포는 정규 분포에 가까워지게 됨(전집이 어떤 분포였는지와 무관)
- 표본 평균의 분포에서 평균은 모평균을 그대로 따르고, 분산은 모분산/N을 따르므로 표준화 공식만 변화

$$Z = \frac{X - \mu}{\sigma} \longrightarrow Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

지금까지, 다뤘던 내용들을 토대로 가설 검증의 기본 논리를 아래와 같이 정리해볼 수 있다.

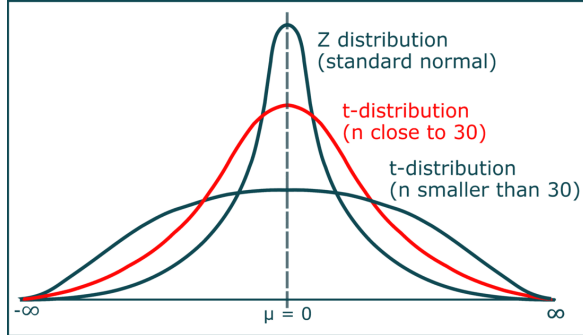
- 1. 연구 가설을 선정해본다. (A/B테스트 스테디의 참여효과)
- 2. 영가설을 설정한다.
 - A/B테스트 스테디에 참여한 분석가와 참여하지 않은 분석가 간에 별 차이가 없을 것이다.
- 3. A/B테스트 스테디에 참여한 분석가들을 sampling한다.
- 4. 영가설이 참이라고 가정했을 때의 표본 평균의 분포를 구한다. (여기서는 차이 없다 = 0 기준)
 - 중심극한정리가 해결해줌
- 5. 4번의 표본 평균의 분포를 토대로 3에서 얻은 평균 혹은 그 이상을 얻을 확률을 계산(검증통계치)
- 6. 확률 분포표에서 검증 통계치 초과일 확률값을 기준으로 영가설 기각(0.05미만, 0.01미만)

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

표준정규분포표 = 검증통계치(=z)의 분포

t통계량을 사용하는 이유

- z통계량을 사용하려면 전집의 분산을 알고 있다는 가정이 있어야 함
 - 전집의 분산(σ^2)을 모르는 경우 표본분산(s^2)으로 전집의 분산을 추정할 수 있으나
- 이 값을 이용하여 z통계량을 구하면 실제보다 크게 추정할 수 있음
- 이런 상황을 대비해서 윌리엄 고셋이라는 분이 스튜던트 t분포라는 것을 고안함
 - 전집의 분산(σ^2) 대신 표본분산(s^2)으로 만든 검증치 분포
- 자유도(표본크기-1)가 커질수록 정규분포에 근사, t통계량을 계산하여 t분포표에 의하여 판단



α df	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587

표본크기가 N만큼 있을 때(N-1) 원하는 유의수준을 달성하기 위해 필요한 t값은?

가장 큰 차이점은?

- 가설의 변화

$$\begin{array}{lcl} H_0: \mu = 6 & \longrightarrow & H_0: \mu_1 = \mu_2 \\ H_a: \mu \neq 6 & & H_a: \mu_1 > \mu_2 \end{array}$$

- 이에 따라 표집 분포에 들어가는 표본 통계량의 변화
 - 평균의 표집분포(기존) => 평균 간 차이의 표집분포(2표본 t검증)

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \longrightarrow t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

두 표본이 독립 -> 공분산 = 0

하지만, 공분산이 0이라고 해서
항상 두 표본이 독립인 것은 아님

불편추정량(unbiased estimator)란?

표본을 뽑아 통계량을 계산하는 이유?

- 모집단의 모수를 추정하기 위해

모수를 잘 추정하는 기준 중 하나가 Unbiasedness(불편성)

- 추정치에 편향이 없음, 수학적으로는 표본의 추정량의 기댓값이 모수와 같아야 한다는 의미

표본분산을 구할 때 $n-1$ 로 나누어 주는 이유?

- 평균의 경우 표본 평균은 불편 추정량으로 다음의 수식을 만족

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$E(\bar{X}) = \mu$$

- 반면, 표본 분산의 경우 분모 n 을 $n-1$ 로 바꾸어주어야 불편 추정량이 됨

$$E(s^2) = \sigma^2$$

$$\begin{aligned} E(s^2) &= E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right) \\ &= \frac{1}{n} E\left(\sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2)\right) \\ &= \frac{1}{n} E\left(\sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n X_i + \sum_{i=1}^n \bar{X}^2\right) \\ &= \frac{1}{n} E\left(\sum_{i=1}^n X_i^2 - 2n\bar{X}^2 + n\bar{X}^2\right) \\ &= \frac{1}{n} \sum_{i=1}^n E(X_i^2) - E(\bar{X}^2) \\ &= \frac{1}{n} \sum_{i=1}^n (\mu^2 + \sigma^2) - \left(\mu^2 + \frac{\sigma^2}{n}\right) \\ &= \frac{n-1}{n} \sigma^2 \end{aligned}$$


후방 카메라의 효과에 대한 실험을 가정해보자.

- 실험 결과, 실험군의 후방 카메라 사용 전/후 운전 실력에는 큰 변화가 있었지만
- 대조군의 운전 실력에는 큰 변화가 없었다고 하자.
- 그런데, 알고 봤더니 대조군에 있던 사람들이 모두 운전 초보자였던 것
- 즉, 여기서 집단간 평균의 차이는
 - $(\text{실험군의 평균} - \text{대조군의 평균}) + \text{선택편향(운전초보자)}$
- 이러한 선택편향을 피하기 위해 무선헌당을 진행하면
- 실험군과 대조군 개별 참여자들 간에 운전 실력 차이는 있을 수 있으나,
- 실험군이 대조군 모두 운전 초보자와 고수가 골고루 섞여 있으므로 집단 간 차이는 없게 된다.
 - = 선택편향 제거
- 즉, 책에서 두 표본이 독립이라는 의미는 무선헌당을 통해 선택편향이 제거된 상태
 - = 실험 처지 외에는 실험군과 대조군 간에 아무런 차이가 없는 상태를 의미

p값에 대한 오해석

p값이 가진 가정

- 무수히 많은 샘플링을 통해 영가설이 참인 분포를 알고 있다는 조건부 가정이 깔려있음.

 p	표본에 나타난 결과 또는 그 이상의 극단적 결과	참인 H_0 , 무선 표집, 그외 다른 모든 가정의 충족
---------------------------------------------------------------------------------------	-------------------------------	--------------------------------------

오류1. p를 영가설이 참일 확률로 생각하는 경우 (= inverse probability)

- 이렇게 생각하는 이유? p값이 영가설 하에서 얻은 자료에 대한 확률이라는 것을 간과하여 발생
- p값만으로는 데이터를 관찰 했을 때 영가설이 참일 확률을 구할 수 없다.
- H_0 가 참일 사전 확률이 추가로 필요

$$\begin{aligned} P(H_0 \text{가 참} | \Delta \text{가 관찰됨}) &= \frac{P(\Delta \text{가 관찰됨} | H_0 \text{가 참}) P(H_0 \text{가 참})}{P(\Delta \text{가 관찰됨})} \\ &= \frac{P(H_0 \text{가 참})}{P(\Delta \text{가 관찰됨})} * P(\Delta \text{가 관찰됨} | H_0 \text{가 참}) \\ &= \frac{P(H_0 \text{가 참})}{P(\Delta \text{가 관찰됨})} * p\text{값} \end{aligned} \quad (17.4)$$

오류2. 1-p를 대립가설이 참일 확률로 생각하는 경우(valid research hypothesis fallacy)

- 오류1과 마찬가지로 p값이 영가설 하에서 얻은 자료에 대한 확률이라는 것을 간과하여 발생
- 1-p는 영가설 하에서 실제 관측된 값보다 결과가 훨씬 덜 극단적일 가능성을 의미

오류3. 1-p가 다른 무작위 표본에서도 동일하게 나타날 것이라 믿는 경우(replicability fallacy)

- p-value값은 영가설이 참이라는 가정 하에 특정 표본에서만 나타난 결과임
- 따라서, 연구 결과를 일반화(=다른 표본에도 적용)하고 싶다면 반드시 메타분석(반복 연구)이 필요함

오류4. p값을 특정 표본 결과값이 우연히 발생할 확률을 나타내는 것이라 생각 (odds against chance fallacy)

- 관찰된 결과값이 무작위(random error)로 발생했다고 바라보는 관점임
 - “극단적인 관찰값”이라는 것을 부정하는 의미
- 하지만, 이미 무수히 많은 샘플링을 통해 영가설이 참인 분포가 있다고 가정을 했음
- 따라서, 우연히라는 표현을 사용하는 것은 적합하지 않음
- 정확한 표현은
 - 영가설이 참이라는 가정 아래에 얻을 확률이 5%미만인 극단적인 결과치를 얻었다.
(=유의미하다.)

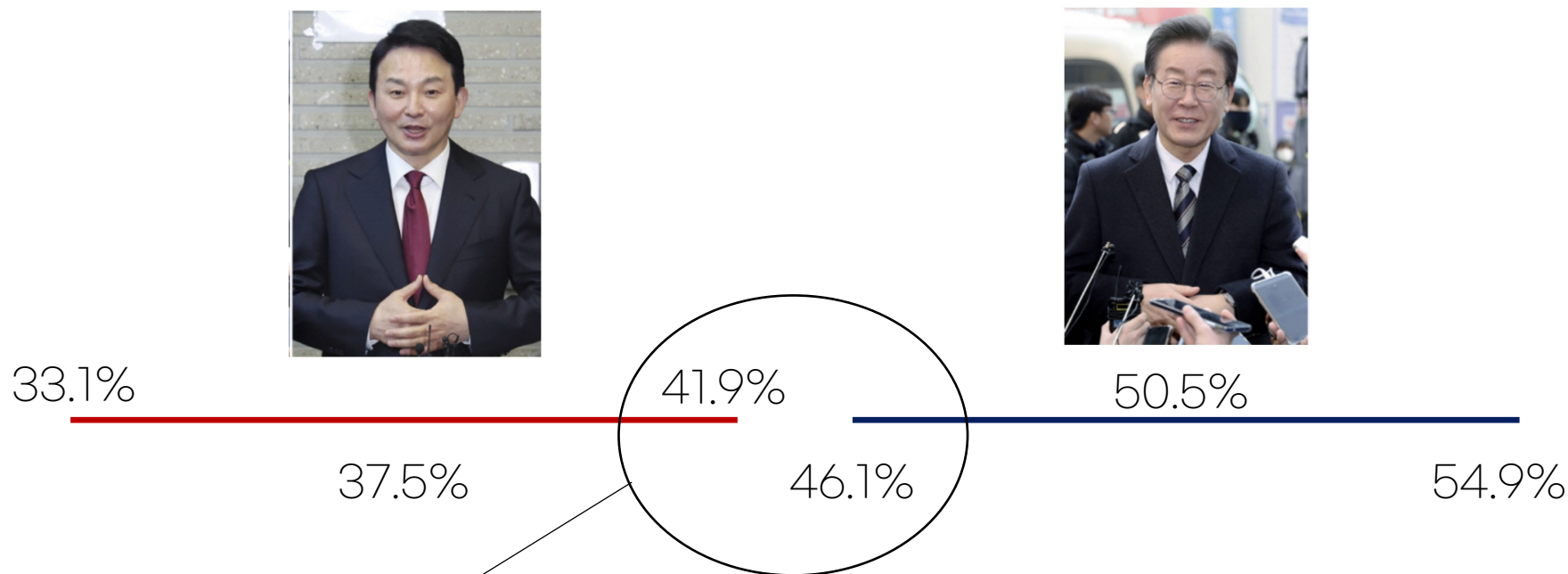
오차범위 밖에서 이겼다는 무슨 의미일까?

‘명룡대전’ 이재명 50.5% 원희룡 37.5%...오차 범위 밖 우세

여론조사 전문업체 리서치앤리서치가 동아일보 의뢰로 지난 24일 인천 계양구에 거주하는 만 18세 이상 성인 507명에게 100% 무선 전화 면접 방식으로 실시한 여론조사 결과 이 대표는 50.5%, 원 후보는 37.5%로 집계된 것으로 28일 나타났다. 두 후보 간 격차는 13%포인트로 오차범위(95% 신뢰 수준에 표본오차 $\pm 4.4\%$ 포인트) 밖에서 이 대표가 우세했다.

- <https://www.munhwa.com/news/view.html?no=2024032901039910018001>

오차범위 밖에서 이겼다는 무슨 의미일까?(본인의 정치 성향과는 무관합니다..)



- 오차 범위 밖이라는 것은 이 구간이 겹치지 않는다는 의미
- (신뢰구간 95%수준에서 = 해당 범위를 벗어나는 득표율을 얻을 확률 5%미만, 그만큼 극단적이다.)
- pvalue가 유의하지 않을 때 신뢰 구간이 0을 포함하는 것도 마찬가지로 원리
 - 2표본 t-test에서 영가설은 두 표본의 평균이 같다. = 다른 한편으로는 평균 간 차이가 없다.
= 평균간 차이가 0이다. = 신뢰 구간에 0을 포함한다.

왜도가 클수록 정규성을 만족하기 위해 필요한 표본의 수가 더 많아짐

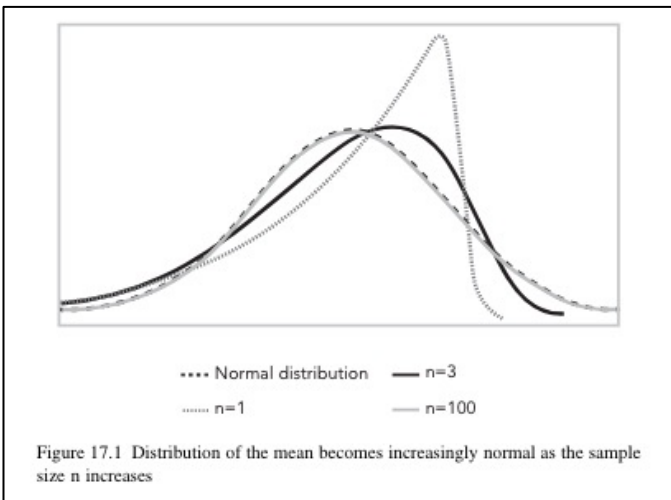
- 여기서 s 는 왜도를 의미함

$$355s^2 + 1$$

따라서, 실험 필요 표본 크기를 위해 변수의 왜도를 줄이는 전략들이 유효할 수 있음

- 예를 들어, 사용자당 수익을 일정 금액으로 제한 하는 경우

N수가 커질수록 정규분포에 근사하는 모습

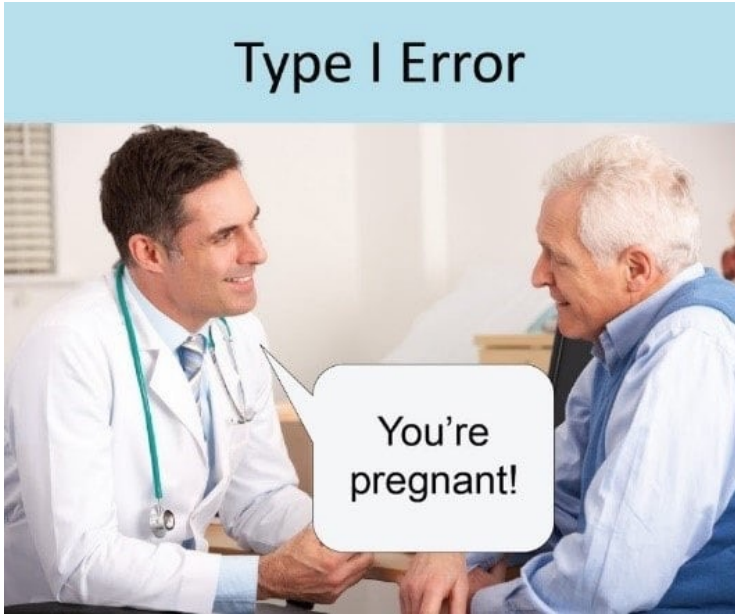


1종, 2종 오류

1종 오류, 2종 오류의 유명한 예시

- 1종 오류 = 기각(임신)할 수 없는데 기각을 해버림(남자는 임신 불가능)
- 2종 오류 = 기각(임신)할 수 있는데 기각을 안함(여자가 실제로 임신)

Type I Error



Type II Error



1종, 2종 오류

중요한 것은 오류 발생 시 의사결정이 미칠 수 있는 영향의 크기

- 1종 오류 = 비가 실제로는 안왔는데, 온다고 해서 우산 챙김 -> 우산 무거움
- 2종 오류 = 비가 실제로는 왔는데, 안온다고 해서 우산 안챙김 -> 비맞음

Type 1 Error



Type 2 Error



1종 오류와 2종 오류의 관계

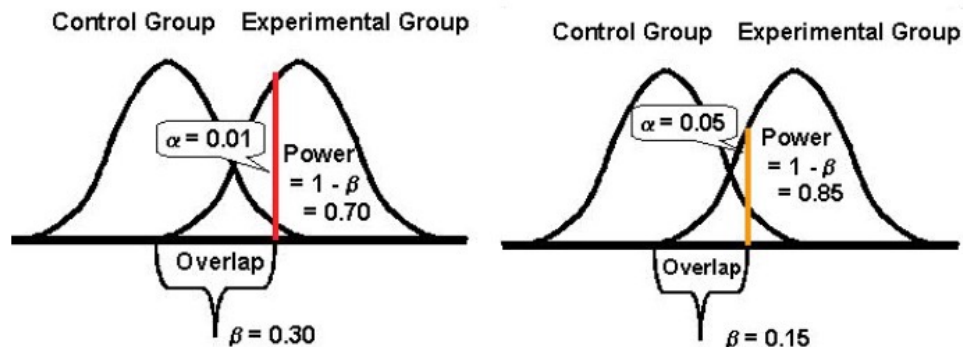
1종 오류, 2종 오류

- 1종 오류는 유의 수준(p-value)의 값을 사용하지만 그 의미는 다름
 - 1종 오류는 영가설이 참일 때, 기각하는 경우
 - 유의 수준은 영가설이 참이라는 가정 하에 극단적인 값을 얻을 확률
 - 경우에 따라 1종 오류에 0.05외에 다른 값을 사용할 수도 있음

결정	실제	
	영가설(H_0)이 참임	영가설(H_0) 거짓임
영가설(H_0) 기각	1종 오류 $p = \alpha$	정확한 결정 $p = 1 - \beta$ (검정력)
영가설(H_0) 기각 못함	정확한 결정 $p = 1 - \alpha$	2종 오류 $p = \beta$

1종 오류, 2종 오류의 관계

- 1종 오류 기준을 낮추면($0.01 \rightarrow 0.05$) \Rightarrow 2종 오류가 줄어든다. \Rightarrow 그리고 검정력이 커진다.



검정력이란?

- 영가설이 실제로 참이 아닐 때, 정확하게 기각할 확률

검정력을 표현하는 방법들

1. 1종오류의 함수

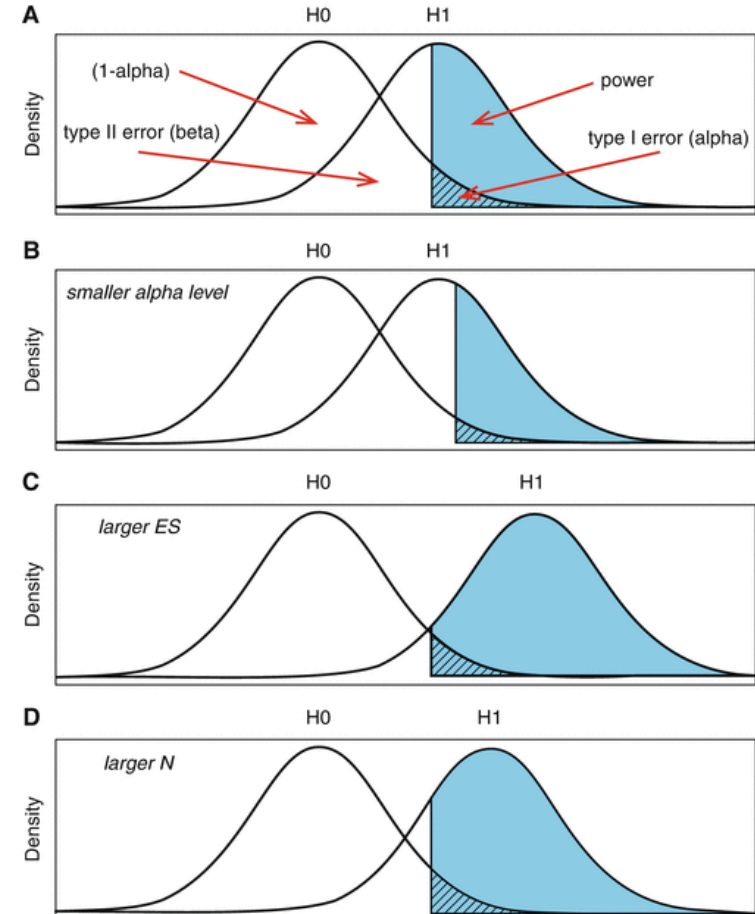
- 1종 오류 기준을 높게 잡으면 (0.05 -> 0.01)
- 검정력이 줄어듦

2. 두 표본 간 평균의 차이 (탐지하고자 하는 효과 크기)의 함수

- 두 표본의 평균 간 차이가 커지면
- 검정력이 늘어남

3. 표본과 분산의 함수

- 표본 평균 분포의 분산 = $\sigma_{\bar{X}}^2 = \sigma^2/n$
- 따라서, n수가 커지면 -> 분산이 줄어듦
-> 겹치는 면적 줄어듦 -> 검정력 늘어남



재현성의 문제를 실제로 경험해 본 적이 있으신지?

- 유의했다고 생각한 가설이 다른 표본에서도 재현이 되었는지
- 혹은, 다른 표본에 테스트 했을 때는 재현 되지 않았는지

3장에서 나오는 편향(트위먼의 법칙, 심슨의 역설, 샘플 비율 불일치)등을 실제로 경험한 적이 있으신지?

수십, 수백개의 지표를 어떤식으로 관리하시는지?

- 책에서는 다중테스트 챕터에서 지표가 100개면 5개는 유의미(=이상해 보일)할 수 있다고 함
 - 해결책으로, 실험에 대한 민감도(=영향을 받을 수 있는지 없는지)에 따라 3단계로 분류하라고 함.
- 어떤 상황에서 어떤 지표가 더 민감하다에 대한 리스팅이 따로 되어 있는지?
- 만약에 한다면 어떻게 해야 할지?

면접에서 A/B테스트 질문을 받는다면 어떤 식으로 받는지/하는지?

- 통계 개념을 물어보는지? 아니면 직접 해보라고 하는지?

행동 과학을 위한 통계학

- <https://m.yes24.com/Goods/Detail/59228690>

크롤리의 통계학 강의

- <https://m.yes24.com/Goods/Detail/32991798>

구조방정식 모형 (원리와 적용)

- <https://m.yes24.com/Goods/Detail/28308606>

A/B테스트 신뢰할 수 있는 온라인 종합실험

- <https://m.yes24.com/Goods/Detail/110044064>

수식없는 신뢰수준 신뢰도 신뢰구간

- https://youtu.be/VizFXel_jnw?si=STWUXwak1-cj4bgG