

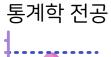


# CTR의 분산 추정 중 빠질 수 있는 함정 (feat. 수리통계학)

방태모 | 지마켓

### 본인 소개







Data Scientist



Data Scientist



글쓰기



<u>블로그</u> 요즘IT





#### 온라인 통제 실험 연구자로 거듭나기

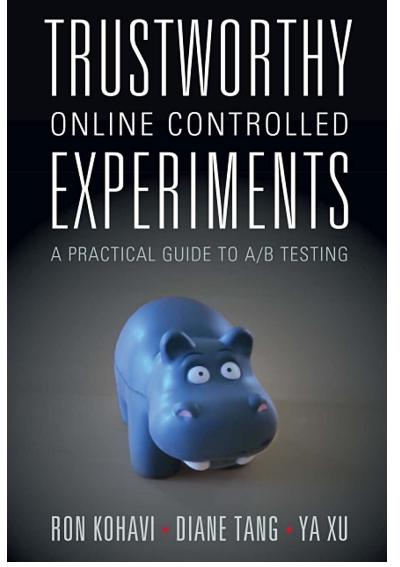






#### <u>깃허브</u>에서 공유 중인 자료

- 스터디 자료
- 실험 관련 질문, 고민 나누기 (issues)



#### 온라인 통제 실험 연구자로 거듭나기



가짜연구소 인과추론 팀 <온라인 통제 실험 연구자로 거듭나기> 프로젝트 스터디 자료 모음

회차	날짜	발표자	주제	발표 자료
1회차	2024-03-13	<u>방태모</u>	상반기 스터디 OT	<u>링크</u>
2회차	2024-03-19	<u>방태모</u>	온라인 통제 실험 소개 및 분석 기초	<u>링크</u>
3회차	2024-03-26	<u>방태모</u>	예제를 통한 A/B 실험의 과정 이해	<u>링크</u>
4회차	2024-04-02	<u>이재호</u>	조직 운영을 위한 지표	<u>링크</u>
5회차	2024-04-09	양유승	실험을 위한 지표와 종합평가기준	<u>링크</u>
6회차	2024-04-16	조동민	온라인 종합 대조 실험에 사용되는 통계 이론	<u>링크</u>
7회차	2024-04-23	<u>지서영</u>	분산 추정 및 민감도 개선: 함정 및 해결책	<u>링크</u>
8회차	2024-04-30	-	쉬어가는 주	-
9회차	2024-05-07	<u>박혜지</u>	A/A 테스트	<u>링크</u>
10회차	2024-05-14	황의림	민감도 향상을 위한 트리거링	<u>링크</u>
11회차	2024-05-21	문정하	샘플 비율 불일치 및 기타 신뢰 관련 가드레일 척도	<u>링크</u>

- ▶ 프로젝트 소개
- <u> 스터디 요약, Discuss, Q&A</u> (Open, Closed 둘 다 참고하세요!)
- ▶ 가짜연구소 인과추론팀 블로그
- ▶ 가짜연구소 블로그

#### References

[1] Kohavi, R., Tang, D., & Xu, Y. (2020). Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing. Cambridge University Press. <a href="https://experimentguide.com/">https://experimentguide.com/</a>

[2] Gupta, S., Ulanova, L., Bhardwaj, S., Dmitriev, P., Raff, P., & Fabijan, A. (2018). The Anatomy of a Large-Scale Experimentation Platform. 2018 IEEE International Conference on Software Architecture (ICSA), 1–109. https://doi.org/10.1109/ICSA.2018.00009

그림 2. <온라인 통제 실험 연구자로 거듭나기 프로젝트> 깃허브



# 오늘 할 이야기

두 지표 간 비율 지표에 관한 분산 추정

# 실험설계에서 알아두어야 할 기초 개념



#### 관찰 단위

• 실험에서 최종 측정값이 나오는 최소 단위

# 온라인 통제 실험의 관찰 단위



퀴즈 1. 온라인 통제 실험의 관찰 단위는?

#### 온라인 통제 실험의 관찰 단위



퀴즈 1. 온라인 통제 실험의 관찰 단위는?

사용자 (고객)

#### 온라인 통제 실험의 관찰 단위

Causal-Lab

대조군, 실험군에 대한 무작위 배정 또한

사용자에 대해 이루어짐

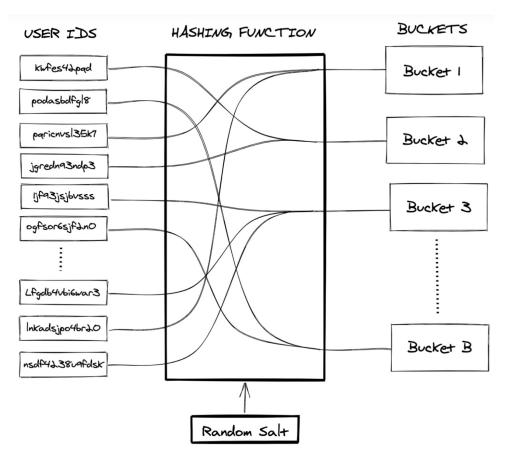
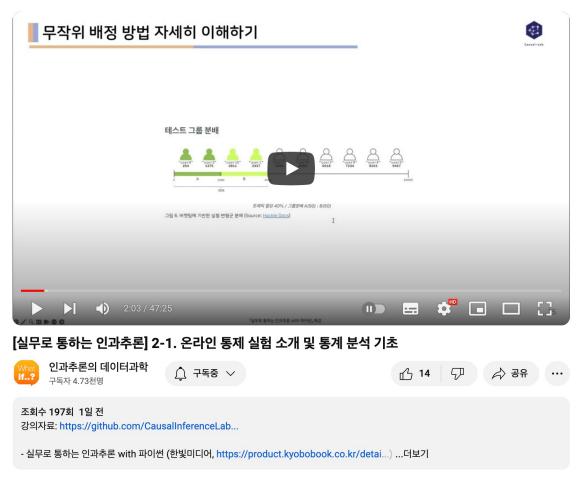


그림 2. 해시 알고리즘을 이용한 무작위화 및 고객 버켓팅 (Source: Spotify, <u>Schultzberg, Kjellin, and Rydberg 2021</u>)

#### 광고 타임



#### 그림 2에 관한 자세한 설명이 궁금하시다면..



실무로 통하는 인과추론 특강 2-1 보러가기

#### 온라인 통제 실험에서의 일반적인 관심사



#### 대조군, 실험군 간 사용자들의 반응 차이가 어떠한가?

- PV
- Click
- SE Conversion (add to cart, buy it now)
- Order
- GMV
- ..



그렇다면, 온라인 실험에서는 어떤 지표를 봐야할까?

• 관찰 단위도 사용자이며, 무작위 배정도 사용자 중심인데..



그렇다면, 온라인 실험에서는 어떤 지표를 봐야할까?

• 관찰 단위도 사용자이며, 무작위 배정도 사용자 중심인데..

먼저, 분석 단위라는 것을 정의해보자

• 분석 단위: 분석에 사용되는 지표의 단위



그렇다면, 온라인 실험에서는 어떤 지표를 봐야할까?

• 관찰 단위도 사용자이며, 무작위 배정도 사용자 중심인데..

먼저, 분석 단위라는 것을 정의해보자

• 분석 단위: 분석에 사용되는 지표의 단위

퀴즈 2. 그렇다면, 온라인 통제 실험에서 분석 단위는 무엇이 되어야할까?



그렇다면, 온라인 실험에서는 어떤 지표를 봐야할까?

• 관찰 단위도 사용자이며, 무작위 배정도 사용자 중심인데..

먼저, 분석 단위라는 것을 정의해보자

• 분석 단위: 분석에 사용되는 지표의 단위

퀴즈 2. 그렇다면, 온라인 통제 실험에서 분석 단위는 무엇이 되어야할까?

사용자

#### 사용자 단위 지표



- PV per visitor
- Click per visitor
- Order per visitor
- Orderer per visitor (Conversion 개념)
- GMV per visitor
- ..



우리 프로덕트의 핵심 지표는 CTR인데요?..

## 관찰단위와 분석단위가 다르면 발생하는 상황



CTR = Sum(Click) / Sum(PV)

→ 분석 단위: PV

→ 온라인 실험의 관찰 단위: 사용자

→ 분석 단위 ≠ 관찰 단위



- 해당 관심 지표의 분산 추정에 주의를 기울이지 않으면 Bias가 생김
- 이에 따라, 통계적 가설 검정 시 잘못된 의사 결정을 할 수 있게됨



관찰 단위와 일치하지 않는 분석 단위를 갖는 지표의 분산 추정에 Bias가 생기는 이유

- 확률표본(random sample)의 i.i.d. 가정이 깨지기 때문
  - Independent and identically distributed (i.i.d.)
  - Independent 가정이 깨짐



다시 CTR로 돌아가보면..

- CTR = Sum(Click) / Sum(PV)
- 관찰 단위는 사용자이나.. CTR의 분석 단위는 PV



다시 CTR로 돌아가보면..

- CTR = Sum(Click) / Sum(PV)
- 관찰 단위는 사용자이나.. CTR의 분석 단위는 PV

퀴즈 3. PV가 독립적이라고 할 수 있는가?



다시 CTR로 돌아가보면..

- CTR = Sum(Click) / Sum(PV)
- 관찰 단위는 사용자이나.. CTR의 분석 단위는 PV

퀴즈 3. PV가 독립적이라고 할 수 있는가?

No, 사용자가 해당 페이지에서 어떤 경험을 했는지에 따라 다음 PV 여부가 결정됨



그럼 이러한 지표들에 대한 분산 추정이 필요할때는 어떻게 해야하나요?



CTR와 같은 지표를 비율 지표(Ratio metric)이라고 칭합니다.

- CTR
- GMV per order

• ..



$$CTR = \frac{\sum Click}{\sum PV}$$



$$CTR = \frac{\sum Click}{\sum PV}$$

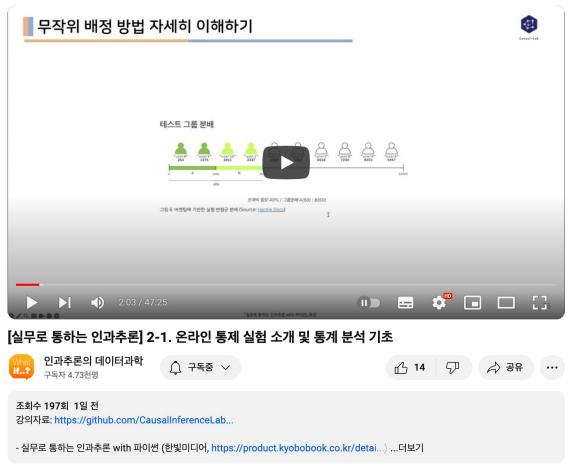
#### 여기서 중요한 포인트

- 사용자로부터 측정되는 모든 지표들은 확률변수(random variable)
- 상수로 취급할 수 있는 값은 고유 사용자의 수 (표본 크기, N)

#### 광고 타임



#### 확률변수, 통계량, 추정량에 관한 개념을 잡고 싶으시다면..



실무로 통하는 인과추론 특강 2-1 보러가기



$$CTR = \frac{\sum Click}{\sum PV}$$

#### 여기서 중요한 포인트

- 사용자로부터 측정되는 모든 지표들은 확률변수(random variable)
- 상수로 취급할 수 있는 값은 고유 사용자의 수 (표본 크기, N)

$$CTR = \frac{\sum Click}{\sum PV} = \frac{\frac{1}{N}\sum Click}{\frac{1}{N}\sum PV} = \frac{\overline{X}}{\overline{Y}} = M_{ctr}$$



$$Var(M_{ctr}) = Var\left(\frac{\overline{X}}{\overline{Y}}\right)$$

두 표본평균의 비(i.e. 비율지표)에 관한 분산 추정이 필요한 상황..

• 사용자 단위 지표의 분산 추정은 굉장히 간단함 (Click per visitor =  $\bar{X}$ )



$$Var(M_{ctr}) = Var\left(\frac{\overline{X}}{\overline{Y}}\right)$$

두 표본평균의 비(i.e. 비율지표)에 관한 분산 추정이 필요한 상황..

• 사용자 단위 지표의 분산 추정은 굉장히 간단함 (Click per visitor =  $\bar{X}$ )

$$Var(\bar{X}) = Var(\frac{1}{N}\sum X_i) = \frac{1}{N^2}Var(\sum X_i) = \frac{1}{N^2} \cdot N \cdot Var(X)$$

$$Var(X) = \hat{\sigma}^2 = \frac{1}{N-1} \sum (X_i - \bar{X})^2$$
이라 하자:

$$\therefore Var(\bar{X}) = \frac{1}{N^2} \cdot N \cdot \hat{\sigma}^2 = \frac{\hat{\sigma}^2}{N}$$



아무튼, 사용자 단위 지표의 분산 추정은 쉬운 것을 알겠고..

그래서, 비율 지표는 두 표본평균의 비에 관한 통계량에 해당하는데.. 이럴 때엔 분산 추정을 어떻게?..

$$Var(M_{ctr}) = Var\left(\frac{\overline{X}}{\overline{Y}}\right)$$



아무튼, 사용자 단위 지표의 분산 추정은 쉬운 것을 알겠고..

그래서, 비율 지표는 두 표본평균의 비에 관한 통계량에 해당하는데.. 이럴 때엔 분산 추정을 어떻게?.

$$Var(M_{ctr}) = Var\left(\frac{\overline{X}}{\overline{Y}}\right)$$

- → Delta Method (*Deng et al., 2018*)
- → 확률변수의 점근분포(asymptotic distribution)를 유도하는데에 사용되는 수리통계학적 방법론
- → 두 표본평균의 비는 새로운 통계량이자 확률변수에 해당
- → 두 표본평균의 비에 관한 분산 추정에도 Delta Method가 사용됨



$$Var(M_{ctr}) = Var\left(\frac{\overline{X}}{\overline{Y}}\right) = \frac{1}{\overline{Y}^2}Var(\overline{X}) + \frac{\overline{X}^2}{\overline{Y}^4}Var(\overline{Y}) - 2\frac{\overline{X}}{\overline{Y}^3}Cov(\overline{X}, \overline{Y})$$



$$Var(M_{ctr}) = Var\left(\frac{\overline{X}}{\overline{Y}}\right) = \frac{1}{\overline{Y}^2}Var(\overline{X}) + \frac{\overline{X}^2}{\overline{Y}^4}Var(\overline{Y}) - 2\frac{\overline{X}}{\overline{Y}^3}Cov(\overline{X}, \overline{Y})$$

$$Var(M_{ctr}) = \frac{\overline{Y}^2 Var(\overline{X}) - 2\overline{X}\overline{Y}Cov(\overline{X}, \overline{Y}) + \overline{X}^2 Var(\overline{Y})}{\overline{Y}^4}$$

Symmetric하게 예쁘게 식이 떨어짐



$$Var(M_{ctr}) = Var\left(\frac{\overline{X}}{\overline{Y}}\right) = \frac{1}{\overline{Y}^2}Var(\overline{X}) + \frac{\overline{X}^2}{\overline{Y}^4}Var(\overline{Y}) - 2\frac{\overline{X}}{\overline{Y}^3}Cov(\overline{X}, \overline{Y})$$

$$Var(M_{ctr}) = \frac{\overline{Y}^2 Var(\overline{X}) - 2\overline{X}\overline{Y}Cov(\overline{X}, \overline{Y}) + \overline{X}^2 Var(\overline{Y})}{\overline{Y}^4}$$

Symmetric하게 예쁘게 식이 떨어짐

여기서 생각해볼만한 재밌는 점은..

$$\bar{X} = \frac{1}{N} \sum Click, \bar{Y} = \frac{1}{N} \sum PV$$

퀴즈 4. 두 지표 각각을 한마디로 정의해본다고 하면, 무엇이 떠오르시나요?

# 비율 지표의 분산 추정



$$Var(M_{ctr}) = Var\left(\frac{\overline{X}}{\overline{Y}}\right) = \frac{1}{\overline{Y}^2}Var(\overline{X}) + \frac{\overline{X}^2}{\overline{Y}^4}Var(\overline{Y}) - 2\frac{\overline{X}}{\overline{Y}^3}Cov(\overline{X}, \overline{Y})$$

$$Var(M_{ctr}) = \frac{\overline{Y}^2 Var(\overline{X}) - 2\overline{X}\overline{Y}Cov(\overline{X}, \overline{Y}) + \overline{X}^2 Var(\overline{Y})}{\overline{Y}^4}$$

Symmetric하게 예쁘게 식이 떨어짐

여기서 생각해볼만한 재밌는 점은..

$$\bar{X} = \frac{1}{N} \sum Click, \bar{Y} = \frac{1}{N} \sum PV$$

퀴즈 4. 두 지표 각각을 한마디로 정의해본다고 하면, 무엇이 떠오르시나요?

 $\bar{X}$ 는 사용자 당 클릭,  $\bar{Y}$ 는 사용자 당 PV (즉, CTR은 사용자 단위의 두 지표 간 비율지표)

#### 비율 지표의 분산 추정



$$Var(M_{ctr}) = \frac{\overline{Y}^2 Var(\overline{X}) - 2\overline{X}\overline{Y}Cov(\overline{X}, \overline{Y}) + \overline{X}^2 Var(\overline{Y})}{\overline{Y}^4}$$

사용자 단위 지표의 표본평균, 표본분산, 공분산 정보만 있으면 됨

- → CTR의 분산 추정이 굉장히 쉬워짐
- → 수리통계학의 아름다움

# 비율 지표의 분산 추정



$$Var(M_{ctr}) = \frac{\overline{Y}^2 Var(\overline{X}) - 2\overline{X}\overline{Y}Cov(\overline{X}, \overline{Y}) + \overline{X}^2 Var(\overline{Y})}{\overline{Y}^4}$$

필요 데이터는 이게 끝 -> 실험 기간 동안 관측된 개별 사용자의 클릭(X), PV(Y) 관측값

- 사용자 당 클릭, 사용자 당 PV 연산
- 사용자 당 클릭, 사용자 당 PV 각각의 표본분산 연산
- 사용자 당 클릭, 사용자 당 PV 간 표본 공분산 연산  $(Cov(\bar{X}, \bar{Y}) = \frac{1}{n}Cov(X, Y))$



보너스. A/B 테스트의 최종 관심인 Delta %에 관한 분산 추정



$$\Delta\% = \frac{M_{trt} - M_{con}}{M_{con}}$$

 $\Delta$ %는 대조군 대비 실험군에서 핵심 관심 지표(M)의 변화율



$$\Delta\% = \frac{M_{trt} - M_{con}}{M_{con}}$$

Δ% 에 대해 통계적 가설 검정을 한다고 해보자.

퀴즈 5. Δ% 값 자체 외에 추가적으로 꼭 필요로 되는 값은?



$$\Delta\% = \frac{M_{trt} - M_{con}}{M_{con}}$$

Δ% 에 대해 통계적 가설 검정을 한다고 해보자.

퀴즈 5. Δ% 값 자체 외에 추가적으로 꼭 필요로 되는 값은?

- → ∆%의 분산
- $\rightarrow$  표본평균의 계산은 쉬움.. 문제는 표본분산인  $Var(\Delta\%)$ 에 관한 추정



$$Var(\Delta\%) = Var\left(\frac{M_{trt} - M_{con}}{M_{con}}\right) = Var\left(\frac{M_{trt}}{M_{con}} - 1\right) = Var\left(\frac{M_{trt}}{M_{con}}\right)$$

또 한 번 비율 지표에 관한 분산 추정 필요



$$Var(\Delta\%) = Var\left(\frac{M_{trt} - M_{con}}{M_{con}}\right) = Var\left(\frac{M_{trt}}{M_{con}} - 1\right) = Var\left(\frac{M_{trt}}{M_{con}}\right)$$

또 한 번 비율 지표에 관한 분산 추정 필요

여기서 실험의 핵심 관심 지표 M을 앞서 구해둔 CTR이라 하자.

앞서 배운 Delta method를 그대로 적용하면:

$$Var\left(\frac{M_{trt}}{M_{con}}\right) = \frac{M_{con}^2 Var(M_{trt}) - 2M_{trt}M_{con}Cov(M_{trt},M_{con}) + {M_{trt}}^2 Var(M_{con})}{M_{con}^4}$$



$$Var(\Delta\%) = Var\left(\frac{M_{trt} - M_{con}}{M_{con}}\right) = Var\left(\frac{M_{trt}}{M_{con}} - 1\right) = Var\left(\frac{M_{trt}}{M_{con}}\right)$$

또 한 번 비율 지표에 관한 분산 추정 필요

여기서 실험의 핵심 관심 지표 M을 앞서 구해둔 CTR이라 하자.

앞서 배운 Delta method를 그대로 적용하면:

$$Var\left(\frac{M_{trt}}{M_{con}}\right) = \frac{M_{con}^2 Var(M_{trt}) - 2M_{trt}M_{con}Cov(M_{trt},M_{con}) + {M_{trt}}^2 Var(M_{con})}{M_{con}^4}$$

여기서 희소식은 공분산 항을 소거할 수 있다는 것!

퀴즈 6. 대조군과 실험군 각각의 관심지표에 관한 공분산은 어째서 소거가 될까?



$$Var(\Delta\%) = Var\left(\frac{M_{trt} - M_{con}}{M_{con}}\right) = Var\left(\frac{M_{trt}}{M_{con}} - 1\right) = Var\left(\frac{M_{trt}}{M_{con}}\right)$$

또 한 번 비율 지표에 관한 분산 추정 필요

여기서 실험의 핵심 관심 지표 M을 앞서 구해둔 CTR이라 하자.

앞서 배운 Delta method를 그대로 적용하면:

$$Var\left(\frac{M_{trt}}{M_{con}}\right) = \frac{M_{con}^2 Var(M_{trt}) - 2M_{trt}M_{con}Cov(M_{trt}, M_{con}) + M_{trt}^2 Var(M_{con})}{M_{con}^4}$$

여기서 희소식은 공분산 항을 소거할 수 있다는 것!

퀴즈 6. 대조군과 실험군 각각의 관심지표에 관한 공분산은 어째서 소거가 될까?

실험군과 대조군은 무작위 배정을 통해 독립적으로 할당되기 때문 (독립 -> 공분산 0, 단 역은 성립 안함)



$$Var\left(\frac{M_{trt}}{M_{con}}\right) = \frac{M_{con}^2 Var(M_{trt}) + M_{trt}^2 Var(M_{con})}{M_{con}^4}$$

결론적으로 필요한 데이터는 여전히 똑같다.



$$Var\left(\frac{M_{trt}}{M_{con}}\right) = \frac{M_{con}^2 Var(M_{trt}) + M_{trt}^2 Var(M_{con})}{M_{con}^4}$$

결론적으로 필요한 데이터는 여전히 똑같다.

실험 기간 동안 관측된 개별 사용자의 클릭(X), PV(Y) 관측값

- 실험군과 대조군 각각의 CTR 연산 ( $M_{trt}$ ,  $M_{con}$ )
- 실험군과 대조군 각각의 CTR의 표본분산 연산  $(Var(M_{trt}), Var(M_{con}))$



$$Var\left(\frac{M_{trt}}{M_{con}}\right) = \frac{M_{con}^2 Var(M_{trt}) + M_{trt}^2 Var(M_{con})}{M_{con}^4}$$

결론적으로 필요한 데이터는 여전히 똑같다.

실험 기간 동안 관측된 개별 사용자의 클릭(X), PV(Y) 관측값

- 실험군과 대조군 각각의 CTR 연산 ( $M_{trt}$ ,  $M_{con}$ )
- 실험군과 대조군 각각의 CTR의 표본분산 연산 ( $Var(M_{trt}), Var(M_{con})$ )

# 수리통계학의 아름다움



# 오늘 한 이야기 요약



- 사용자로부터 측정되는 모든 값은 확률변수
- 실험 기간 동안 관측된 사용자의 개별 관심 행동(e.g. Order, Order Conversion, Click) 데이터만 있다면 사용자 단위 지표 외에도 실험에서 관심을 가질만한 거의 모든 통계량이 계산 가능해짐
  - How? Delta Method 이용
    - CTR, GMV per order 등과 같은 비율 지표에 관한 분산 추정 가능
    - A/B 테스트의 최종 관심인 Δ%에 관한 분산 추정 가능

#### References



[1] Deng, A., Knoblich, U., & Lu, J. (2018). Applying the Delta method in metric analytics: A practical guide with novel ideas (arXiv:1803.06336). arXiv. <a href="http://arxiv.org/abs/1803.06336">http://arxiv.org/abs/1803.06336</a>

[2] Bao, W. (2023, March 28). How to Size For Online Experiments With Ratio Metrics. Expedia Group Technology. <a href="https://medium.com/expedia-group-tech/how-to-size-for-online-experiments-with-ratio-metrics-3d57362f1967">https://medium.com/expedia-group-tech/how-to-size-for-online-experiments-with-ratio-metrics-3d57362f1967</a>

[3] Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press. <a href="https://experimentguide.com/">https://experimentguide.com/</a>

[4] Metric Deltas | Statsig Docs