Tencent 腾讯 DataFun.

AB实验中的指标协方差及其应用

熊涛 腾讯 后台开发工程师

个人简介

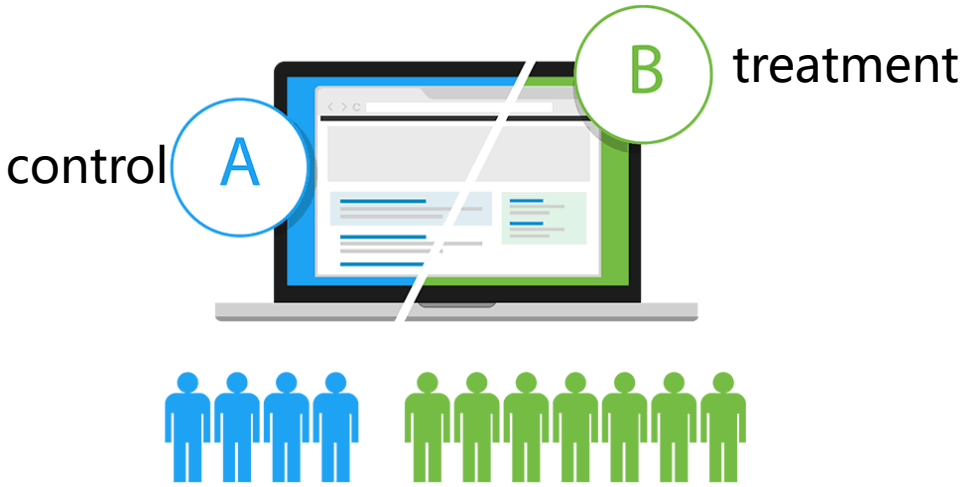
- 2017硕士毕业于浙江大学，校招加入腾讯
- 2018年开始，参与微信AB实验平台后台开发，以及统计模型相关研发
- 目前负责微信AB实验平台的相关研发

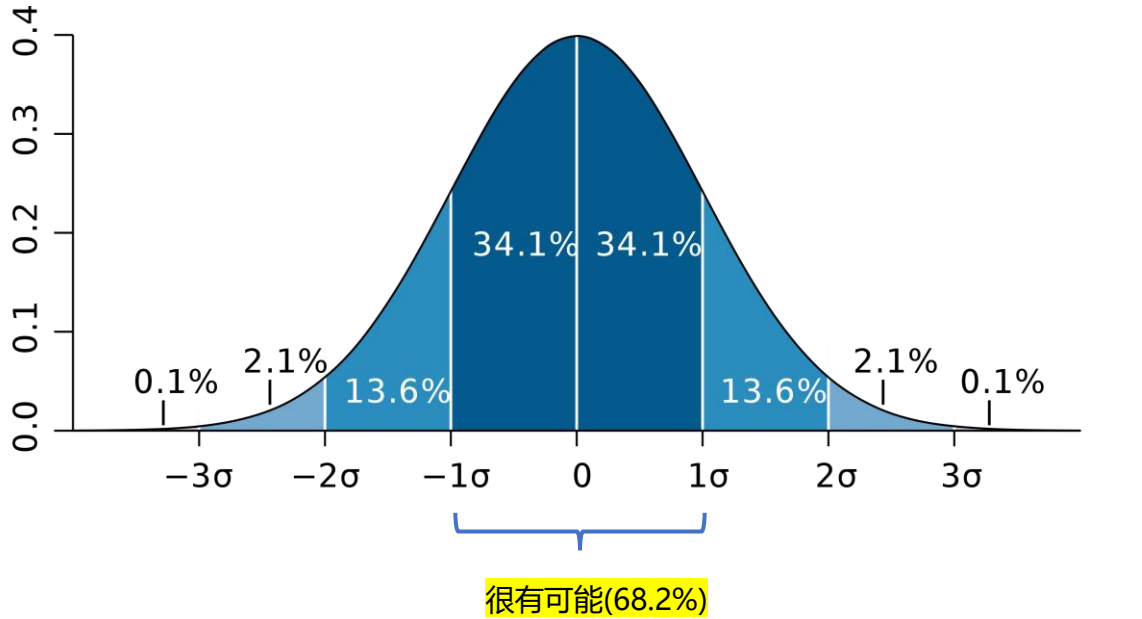
目录 CONTENT

Tencent 腾讯 DataFun.

1 AB实验简介

随机对照实验(AB Testing)





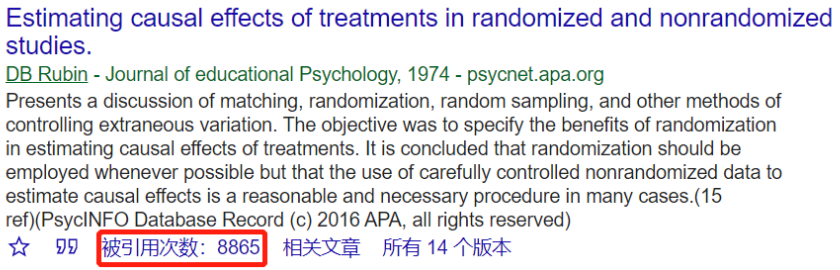
假设treatment不会影响人均停留时长：

$$\bar{Y}_t - \bar{Y}_c \sim Norm(0, Var[\bar{Y}_t - \bar{Y}_c])$$

$$\sigma^2 = Var[\bar{Y}_t - \bar{Y}_c] = Var[\bar{Y}_t] + Var[\bar{Y}_c]$$

部分用户的均值为什么能代表全体用户的均值?

潜在结果模型[1][2]：随机实验的数学框架


Estimating causal effects of treatments in randomized and nonrandomized studies.

DB Rubin - Journal of educational Psychology, 1974 - psycnet.apa.org

Presents a discussion of matching, randomization, random sampling, and other methods of controlling extraneous variation. The objective was to specify the benefits of randomization in estimating causal effects of treatments. It is concluded that randomization should be employed whenever possible but that the use of carefully controlled nonrandomized data to estimate causal effects is a reasonable and necessary procedure in many cases.(15 ref)(PsycINFO Database Record (c) 2016 APA, all rights reserved)

☆ 99 被引用次数：8865 相关文章 所有 14 个版本

subject	$Y_t(u)$	$Y_c(u)$	$Y_t(u) - Y_c(u)$
Joe	130	135	−5

STUVA：Stable unit treatment value assumption 不满足的例子

subject	Joe = c, Mary = t	Joe = t, Mary = t	Joe = c, Mary = c	Joe = t, Mary = c
Joe	140	130	125	120

[1] https://en.wikipedia.org/wiki/Rubin_causal_model
[2] Rubin D B. Estimating causal effects of treatments in randomized and nonrandomized studies[J]. Journal of educational Psychology, 1974, 6

因果推断（实验推断）的根本问题

subject	$Y_t(u)$	$Y_c(u)$	$Y_t(u) - Y_c(u)$
Joe	130	135	−5
Mary	130	145	−15
Sally	130	145	−15
Bob	140	150	−10
James	145	140	+5
MEAN	135	143	−8

subject	$Y_t(u)$	$Y_c(u)$	$Y_t(u) - Y_c(u)$
Joe	130	?	?
Mary	?	125	?
Sally	100	?	?
Bob	?	130	?
James	?	120	?
MEAN	115	125	−10

随机抽样下，部分可以代表总体

比如，对于每个用户，随机扔一枚均匀硬币，决定用户是否分配到实验组，扔硬币的结果用一个Indicator来表示：

$$I(user)$$

I变量和{Y_t,Y_c}独立，因此：

$$E[Y|I=1] = E[Y_t|I=1] = E[Yt]$$

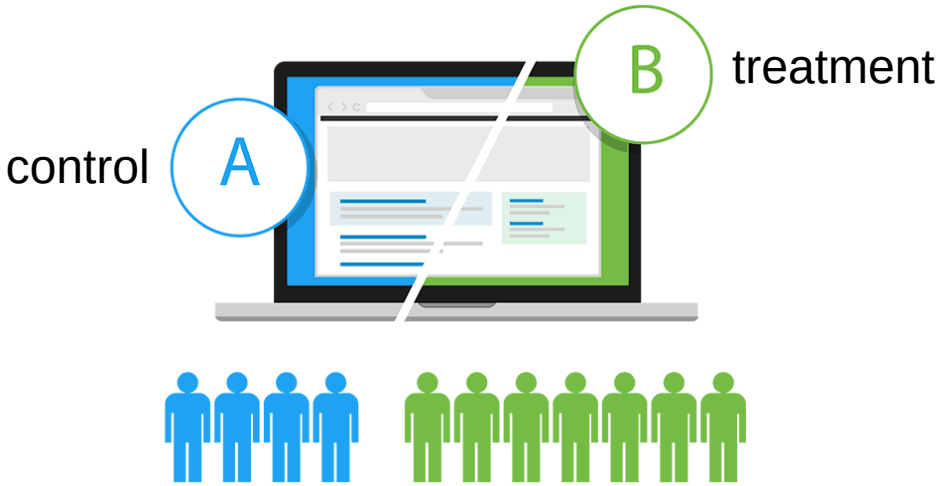
随机抽样下，实验组用户的选取，和他们的潜在的结果{Y_t, Y_c}无关

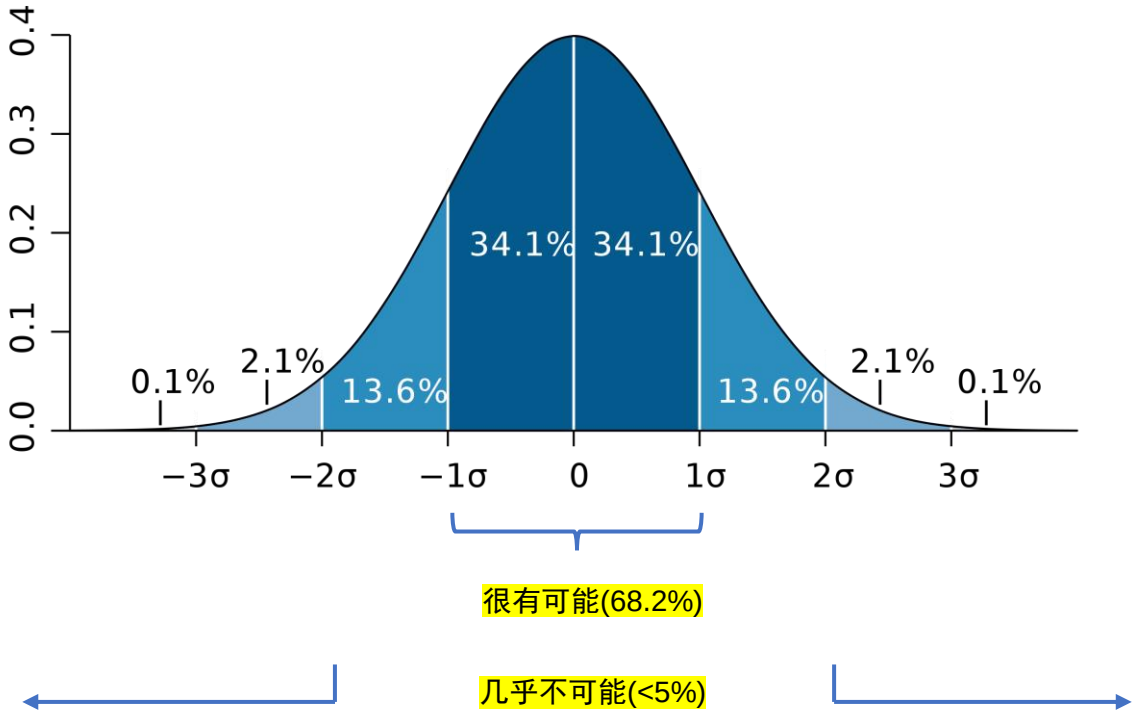
	subject	$Y_t(u)$	$Y_c(u)$	$Y_t(u) - Y_c(u)$
$I(Joe) = 1$	Joe	130	?	?
$I(Mary) = 0$	Mary	?	125	?
$I(Sally) = 1$	Sally	100	?	?
$I(Bob) = 0$	Bob	?	130	?
$I(James) = 0$	James	?	120	?
	MEAN	115	125	−10

Tencent 腾讯 DataFun.

2 指标协方差的应用场景

假设检验中的方差估计





$$\bar{Y}_t - \bar{Y}_c \sim Norm(0, Var[\bar{Y}_t - \bar{Y}_c])$$

$$\sigma^2 = Var[\bar{Y}_t - \bar{Y}_c] = Var[\bar{Y}_t] + Var[\bar{Y}_c]$$

$$Var[\bar{Y}_t] = Cov[\bar{Y}_t, \bar{Y}_t]$$

方差消减

- CUPED[1]是一种普适性很广，效果优越的方差消减方案，其原理为：
  - 对于任意指标M，使用实验开始前的历史数据进行回溯得到回溯指标P
  - 构造新的统计量Y=M+βP， β是一个待定的常数
  - 原指标M的AB两组的差异为：$delta\_m=M_B-M_A$
  - 新统计量X的AB两组的差异为：$delta\_x= X_B-X_A=M_B-M_A+ \beta(P_B-P_A)$
  - E(delta_x)=E(delta_m)
  - 对于任意一个组，$Var(X)=Var(M)+\beta^2Var(P)-2\beta Cov(M,P)$，在$\beta=$**Cov(M,P)**$/Var(P)$时取到最小值$Var(M)(1-\rho^2)$

[1] A. Deng, Y. Xu, R. Kohavi, and T. Walker, “Improving the sensitivity of online controlled experiments by utilizing pre-experiment data,” in Proceedings of the sixth ACM international conference on Web search and data mining, 2013, pp. 123–132.

持续观测

- 随着实验的进行，我们不断地观测到数据点，记为$M_1,M_2,\ldots,M_n$，在观测新数据的同时做推断（停止 or 继续）
- 一个典型的模型：Bayes Factor[1]

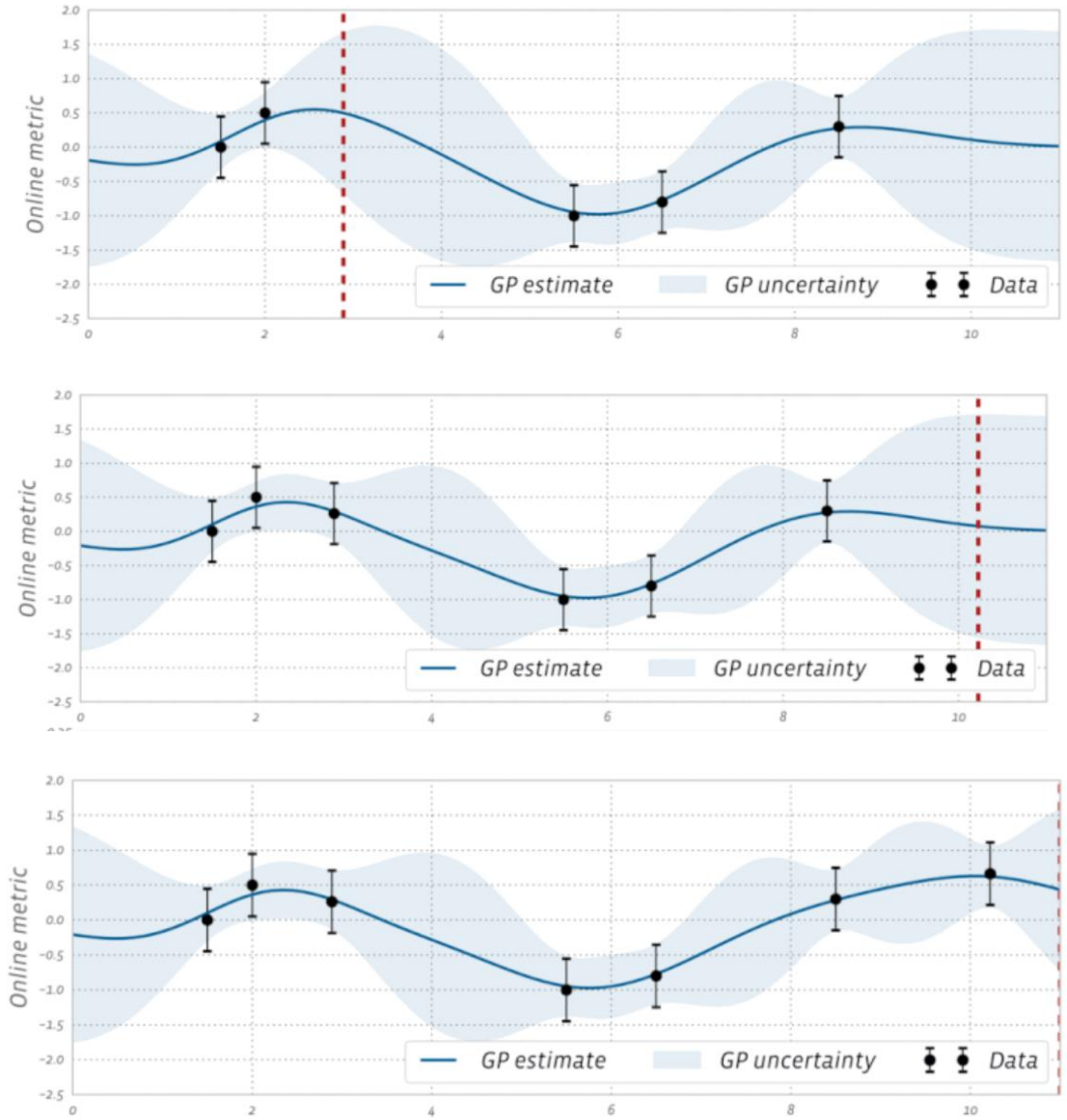
$$\frac{P(H_1|Data)}{P(H_0|Data)} = \frac{P(H_1)}{P(H_0)} \times \frac{P(Data|H_1)}{P(Data|H_0)}$$

- 似然的计算：
  - 对于指标M，假设每天观测一次，每次观测得到当天的指标数据，记为$M_1,M_2,M_3\ldots$
  - 由于大多数场景下按照用户分流，观测到的数据点会有相关性
  - $(M_1,M_2,M_3\ldots)$渐进意义上是一个多维正态分布
  - 利用多维正态分布进行似然计算，需要估计$(M_1,M_2,M_3\ldots)$的**协方差矩阵**

[1] Deng A, Lu J, Chen S. Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing[C]//2016 IEEE international conference on data science and advanced analytics (DSAA). IEEE, 2016: 243-252

贝叶斯优化

- 贝叶斯优化[1]：
  - 给定一个目标指标obj(x)，对参数x进行优化，使得obj(x)尽可能最优
  - 模型会给出参数空间每个x下，目标指标的预测结果
  - 不断迭代，每次迭代会选择下一个需要探索的点
  - 对被探索的点收集数据，更新模型的预测结果
- 现实场景中，被优化的指标可能是一个由多个指标复合组成的复杂指标
  - obj(x)=a*f(x)+b*g(x)
  - 估计方差时需要考虑指标之间的协方差
  - $Var(obj)=a^2Var(f(x))+b^2Var(g(x))+$**$2ab*Cov[f(x),g(x)]$**



[1] Letham, B., Karrer, B., Ottoni, G., & Bakshy, E. (2019). Constrained Bayesian optimization with noisy experiments.

FDR control under dependence

- 单个指标的检验，犯第一类错误的概率为5%
- 针对多个指标的多次检验，存在至少一个第一类错误的概率膨胀，可能远大于5%
- FDR control用于控制多次检验下的False Discovery Rate，常用的有BH方法[1]
- 近年有方法[2]考虑了同时检验的多个指标之间的相关性，可以相比BH方法提高power，该方法需要对指标的协方差矩阵进行估计

[1] https://en.wikipedia.org/wiki/False_discovery_rate#Benjamini%E2%80%93Hochberg_procedure
[2] Fithian, W., & Lei, L. (2022). Conditional calibration for false discovery rate control under dependence. *The Annals of Statistics*, *50*(6), 3091-3118.

Tencent 腾讯 DataFun.

3 指标协方差的严格定义

实验组指标的定义

对于任意实验组，定义我们观测到的某个人均指标为：

$$A = \frac{\sum_u I(u)\,Y(u)}{\sum_u I(u)}$$

	subject	$Y_t(u)$	$Y_c(u)$	$Y_t(u) - Y_c(u)$
$I(Joe) = 1$	Joe	130	?	?
$I(Mary) = 0$	Mary	?	125	?
$I(Sally) = 1$	Sally	100	?	?
$I(Bob) = 0$	Bob	?	130	?
$I(James) = 0$	James	?	120	?
	MEAN	115	125	−10

简单指标

- $S = \sum_u I(u)\,Y(u)$（实验组总停留时长、总曝光数、总命中人数）
- Indicator随机变量$I(u)$是独立同分布的随机变量，在一些正则化条件满足的情况下，可以应用Lindeberg CLT[1]：
- $S \sim Norm(p\sum Y, p(1-p)\sum Y^2)$ p是实验组的流量比例

Lindeberg CLT [edit]

Main article: Lindeberg's condition

In the same setting and with the same notation as above, the Lyapunov condition can be replaced with the following weaker one (from Lindeberg in 1920).

Suppose that for every $\varepsilon > 0$

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E}\left[(X_i - \mu_i)^2 \cdot \mathbf{1}_{\{X_i : |X_i - \mu_i| > \varepsilon s_n\}}\right] = 0$$

where $\mathbf{1}_{\{\ldots\}}$ is the indicator function. Then the distribution of the standardized sums

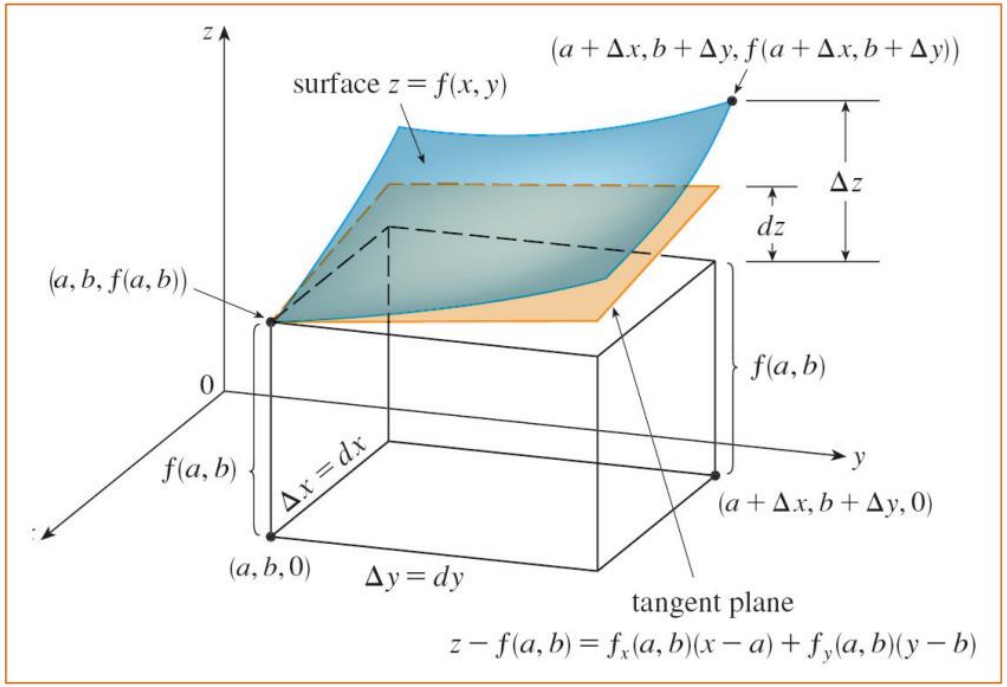
$$\frac{1}{s_n} \sum_{i=1}^{n} (X_i - \mu_i)$$

converges towards the standard normal distribution $\mathcal{N}(0,1)$.

[1] https://en.wikipedia.org/wiki/Central_limit_theorem

从简单指标到复杂指标

- $A = \frac{\sum_u I(u)Y(u)}{\sum_u I(u)} := \frac{S_1}{S_2}$
- 对简单指标除以总体个数N，使其收敛
  - $\frac{S}{N} \sim Norm\left(pE[Y], p(1-p)\frac{E[Y^2]}{N^2}\right)$
  - $\lim_{N\to\infty} S_1/N = a$
  - $\lim_{N\to\infty} S_2/N = b$
- $A = \frac{S_1/N}{S_2/N} := \frac{a+\Delta x}{b+\Delta y} = f(a+\Delta x, b+\Delta y)$
  - 其中$f(x,y) = x/y$
- $A \approx \frac{a}{b} + fx(a,b)\Delta x + fy(a,b)\Delta y$
- $A \approx \frac{a}{b} + \frac{1}{b}\left(\frac{S_1}{N} - a\right) - \frac{a}{b^2}\left(\frac{S_2}{N} - b\right) = k_1 S_1 + k_2 S_2 + k_3$
- A是渐进正态的，而且渐进分布满足$E[A] = \frac{E[S_1]}{E[S_2]} = \frac{pE[Y]}{p} = E[Y]$
- 这是Delta Method[1]的基本原理



[1] https://en.wikipedia.org/wiki/Delta_method

指标的协方差估计

- 对于实验组的任意两个指标
- 如何估计它们的协方差$Cov[m_1, m_2]$？
- $m_1 = k_1S_1 + k_2S_2 + k_3 \ldots$
- $m_2 = k_4S_4 + k_5S_5 + k_6 \ldots$

$$\operatorname{cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \operatorname{cov}(X_i, Y_j)$$

- 我们只需要估计形如$Cov[kS, k'S'] = kk'$ $\boldsymbol{Cov[S, S']}$的协方差项

4 指标协方差的高效估计

朴素方法：以估计人均指标协方差为例

- 从总体中进行小流量抽样，近似看成有放回抽取，样本iid
- 用样本协方差估计总体协方差
- $q = \frac{1}{n-1}\sum(Xi - \bar{X})(Yi - \bar{Y})$ 用于估计$Cov[X, Y]$
- $Cov[\bar{X}, \bar{Y}] = \frac{Cov[X,Y]}{n}$

Subject	停留时长X	曝光feed数Y
Joe	15min	43个
Marry	10min	28个
Sally	30min	100个
Bob	5min	15个
James	22min	60个

朴素方法：data missing问题

- 某些情况存在data missing问题
- 直接用两天都来的用户计算样本协方差？
  - Biased！

Subject	停留时长	昨天的停留时长
Joe	？（用户没来）	13min
Marry	10min	3min
Sally	30min	20min
Bob	5min	？（用户没来）
James	22min	15min

数据增广(Data Augmentation[1])

- 没来的用户直接补0值
- 引入两个Indicator表示用户有没有来
  - I_X
  - I_Y
- $Cov[\bar{X}, \bar{Y}] = Cov\left[\frac{\sum X}{\sum I_X}, \frac{\sum Y}{\sum I_Y}\right]$
- 应用Delta Method

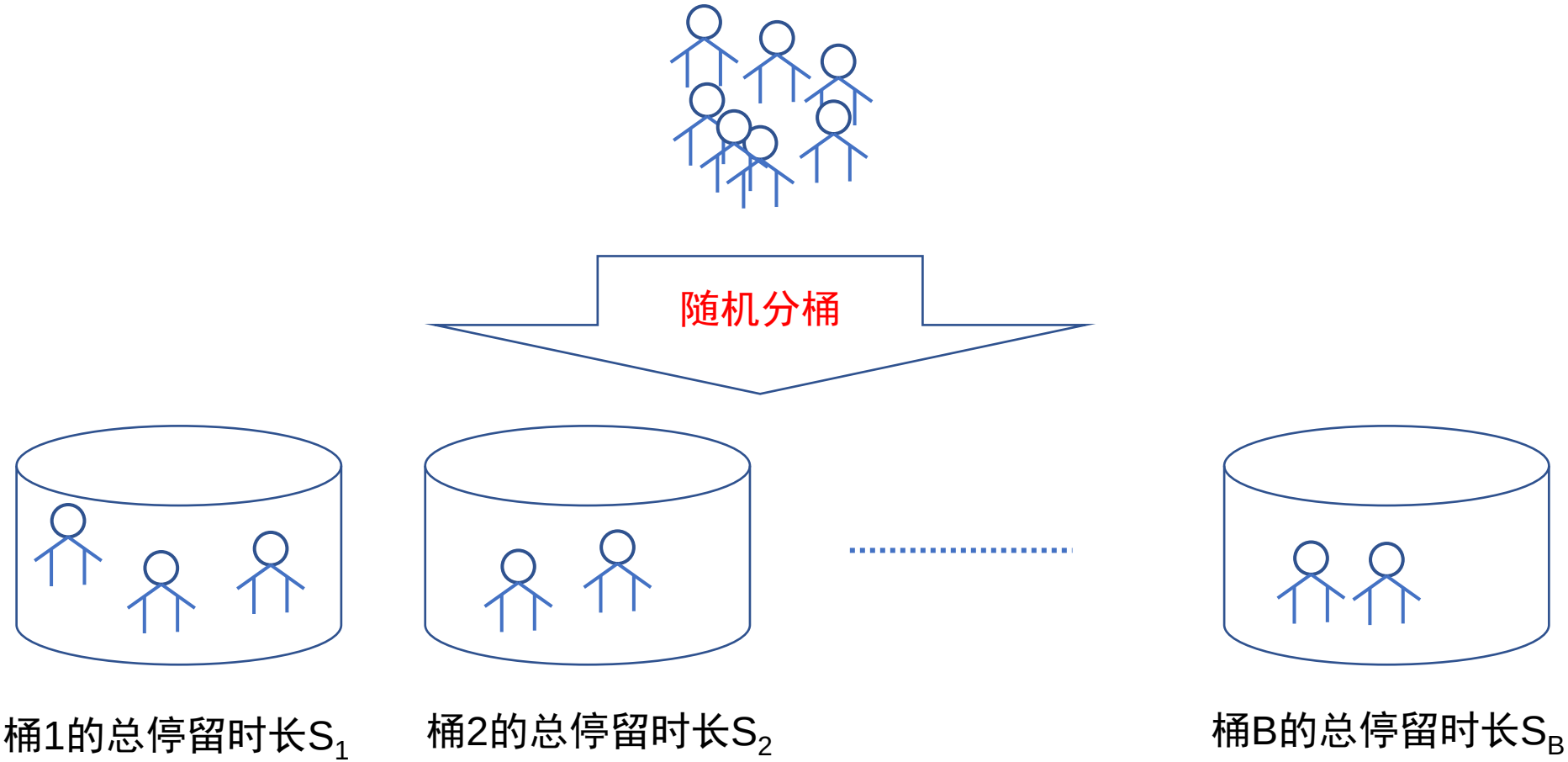
Subject	停留时长X	昨天的停留时长Y
Joe	0（用户没来）	13min
Marry	10min	3min
Sally	30min	20min
Bob	5min	0（用户没来）
James	22min	15min

[1] Deng A, Knoblich U, Lu J. Applying the Delta method in metric analytics: A practical guide with novel ideas[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 233-242.

用户级别特征拼接的性能开销问题

- 上面的方法依赖用户级别的指标数据拼接
- 实验组数量多，平均命中的用户量大
- 指标数量多，口径复杂，来源于大量不同的数据源
- 总体来说，性能开销比较大

高效估计协方差：随机分桶



使用分桶后的数据估计协方差

- 引入B个表示随机分桶结果的Indicator: $I_b(u)$
- 每个桶的简单指标：$S_b = \sum_u I_b(u) I(u) Y(u)$
- 基于桶数据进行协方差计算：
  - $K(S, S') = \frac{1}{B-1} \sum_{b=1}^{B} [(Sb - \bar{S})(S'_b - \bar{S'})]$ 其中$\bar{S}$和$\bar{S'}$表示桶均值
- 可以证明[1]：$E[B(1-p)K(S, S')] = Cov[S, S']$

- 对于实验组的两个指标
- 如何估计它们的协方差$Cov[m_1, m_2]$？
- $m_1 = k_1 S_1 + k_2 S_2 + k_3 \ldots$
- $m_2 = k_4 S_4 + k_5 S_5 + k_6 \ldots$

$$\operatorname{cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} \operatorname{cov}(X_i, Y_j)$$

- 我们只需要估计形如$Cov[kS, k'S'] = kk'$ $\boldsymbol{Cov[S, S']}$的协方差项

[1] https://arxiv.org/pdf/2108.02668.pdf

一个实际的例子：clickhouse指标性能优化

- 指标明细数据按天存储在clickhouse中
- 计算多天累计数据，并进行假设检验

```
select
sum(y) as sum_y,
stddevPop(y) as stddev
from
(
    select uid, sum(y) as y
    from metric_detail
    where ds BETWEEN 2.11 and 2.12
    group by uid
)
```

日期(ds)	用户(uid)	停留时长(y)
2.11	Joe	15min
2.11	Marry	10min
2.11	Sally	30min
2.11	Bob	5min
2.11	James	22min

分桶方案去掉了按uid的group by，转而用桶号group by，性能更好[1]

```
select bucketid, sum(y) as y
from metric_detail
where ds BETWEEN 2.11 and 2.12
group by (murmurHash3_32(uid) % 100) as bucketid
```

日期(ds)	用户(uid)	停留时长(y)
2.12	Joe	12min
2.12	Sally	20min
2.12	Bob	8min

https://clickhouse.com/docs/en/sql-reference/statements/select/group-by/#group-by-optimization-depending-on-table-sorting-key

Tencent 腾讯 DataFun.

5 一些测试数据

测试数据：协方差估计

- 表一展示了不同协方差估计算法的效果
- 表二展示了实验抽样比例对数据增广方法的影响

TABLE I: Simulation Results: Ground truth, average, standard deviation and the time for calculation of covariance in millisecond.

Method	Ground Truth	Avg.	SD.	Time
Naive	8.825	21.085	1.708	135
Data Augmentation	8.825	9.790	0.685	1219
Bucketing of $B = 100$	8.825	8.792	1.684	155
Bucketing of $B = 200$	8.825	8.807	1.269	298
Bucketing of $B = 500$	8.825	8.801	0.931	733
Bucketing of $B = 1000$	8.825	8.802	0.790	1458

TABLE II: Simulation Results: Data augmentation method with different sampling ratios. The results indicate that the upward bias decreases when the ratio decreases.

Ratio	Ground Truth	Avg.	SD.
0.2	3.862	4.798	0.222
0.1	8.639	9.572	0.689
0.05	18.081	19.065	1.977
0.01	94.121	94.164	22.596

测试数据：方差消减

- 桶数越多，相关性越强，β的估计越准

TABLE III: Simulation Results: Relative error of the optimal β estimated using our method compared to the optimal value of β in theory.

Relative Error	B=50	B=100	B=200	B=500	B=1000
ρ=0.3	0.1708	0.1197	0.0817	0.0534	0.0410
ρ=0.5	0.0866	0.0574	0.0413	0.0278	0.0201
ρ=0.6	0.0630	0.0437	0.0301	0.0189	0.0142
ρ=0.8	0.0252	0.0174	0.0127	0.0080	0.0059

测试数据：持续观测

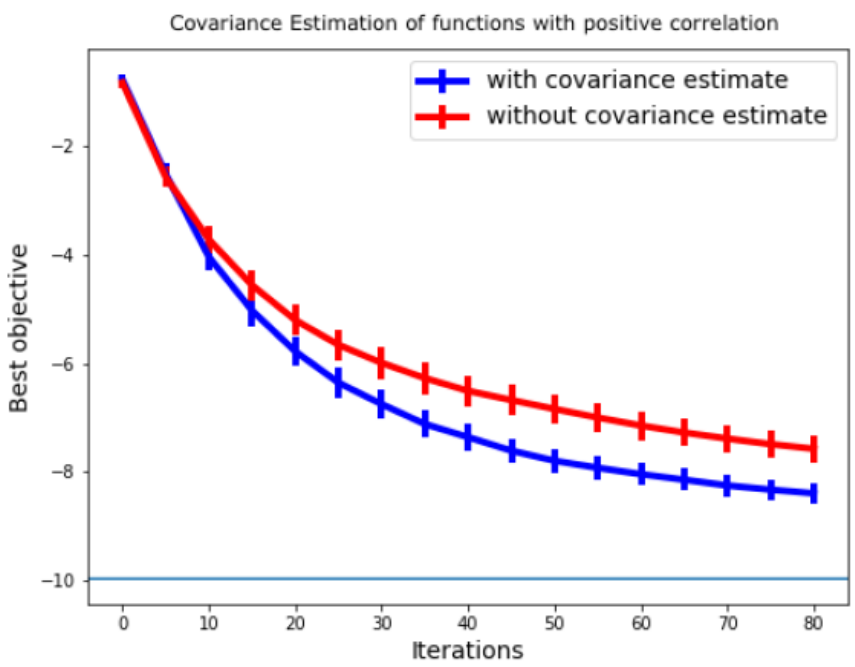
- 通过多元高斯分布生成non-iid的数据，对数据用Bayes Factor进行持续观测，统计FDR
- 不考虑相关性，FDR膨胀79%
- 用真实的协方差矩阵计算似然，FDR得到控制
- 用分桶方法估计协方差矩阵，FDR基本得到控制

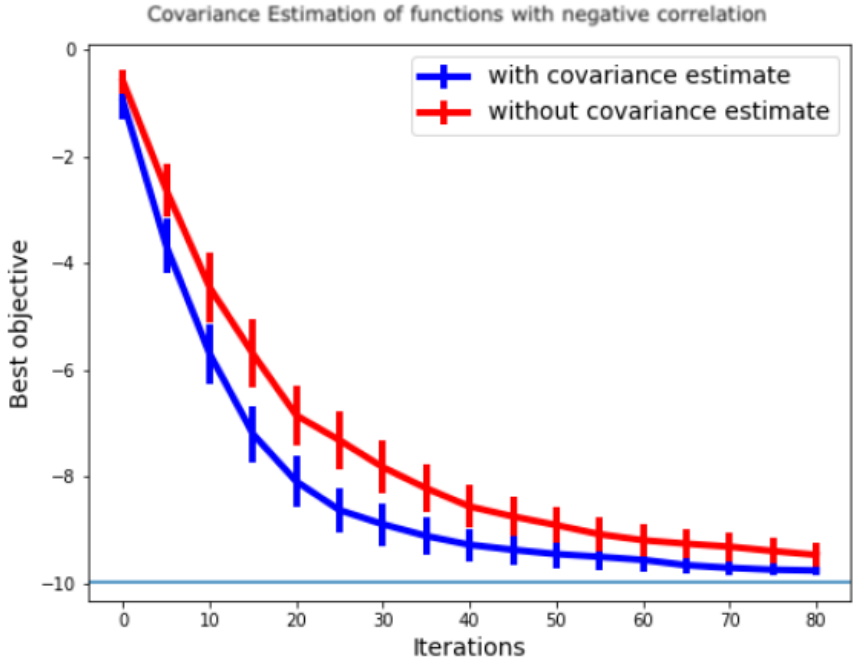
TABLE IV: Simulation Results: FDR is the false discovery rate which should be bounded by FDR ≤ 0.1 in our setting, and Power is the ratio of true rejections to the number of tests where H_1 is true.

Method	FDR	Power
Non-Covariance	0.179	0.827
True Corvariance	0.079	0.693
Estimated Corvariance ($B = 300$)	0.103	0.742
Estimated Corvariance ($B = 200$)	0.115	0.753

测试数据：贝叶斯优化

- 两个指标f和g，对obj(x)=a*f(x)+b*g(x)进行贝叶斯优化
- 分别模拟具有正相关性和负相关性两种场景





复现测试结果

- 以上所有测试的代码均可在如下仓库找到并复现
  - https://github.com/xt2357/covariancesimulation

Tencent 腾讯 DataFun.

6 总结

总结

- 指标协方差是指标相关性的度量，能够广泛应用于各种场景
- 基于用户粒度计算的协方差性能开销较大
- 分桶协方差方法可以在性能开销和精度之间进行权衡：
  - 每个指标可以灵活地独立计算分桶数据备用
  - 需要评估相关性时，取出分桶数据，基于桶进行协方差估计
  - 桶数越大，精度越高，性能越差
  - 针对不同的场景，选择一个合适的桶数，达到性能开销和精度之间的平衡