k-匿名化されたデータの安全性評価について

山田　道洋　菊池浩明
（明治大学）

PWS Cup 2018概要

- 共通の**データセット**を加工
  - 購買履歴である Online Retail Data Set [1] から抜粋
  - 5属性(顧客ID，購入日，商品ID，単価，数量)
  - 1000人をランダムサンプリング
    - 予備戦: 1000人，81776レコード
    - 本戦: 1000人，89278レコード
- 評価方法 (よい加工の決め方)
  - 十分な**安全性**がある (十分に再識別しにくい)
  - **有用性**がより高い (元のデータにより近い)

[1] UCI Machine Learning Repository: Online Retail Data Set.
https://archive.ics.uci.edu/ml/datasets/Online+Retail

特徴: 安全性評価 「撃墜」

- 安全性: 仮名とIDの対応の**当てにくさ**で評価

  当てやすい証拠となる再識別が1つでも存在

  ➔「撃墜」 **安全でない**と判断

公開加工データ(A')

PID	Date
A3	[1/10, 1/30]
C5	[1/10, 1/30]
A3	1/20

再識別(F')

PID	ID	
A3	**Alice**	✔
C5	**Bob**	

PID	ID	
A3	**Alice**	✔
C5	**Chris**	✔

PID	ID	
A3	**Bob**	
C5	**Alice**	

撃墜

PID	Date
A3	[1/10, 1/30]
C5	[1/10, 1/30]
A3	[illegible]

安全

PID	ID	
A3	**Alice**	✔
C5	**Bob**	

PID	ID	
A3	**Alice**	✔
C5	**Bob**	

PID	ID	
A3	**Bob**	
C5	**Chris**	✔

本戦の結果 (再識別)

*: 再提出ペナルティ-0.1点
**: 書式エラー

ID	チーム名	01	02	03	05	06	07	08	09	10	11	12	13	14	撃墜
01	鋼鉄の錬金術師	-	0(2/7)	0(4/7)	0(5/7)	1(7/7)	0(2/7)	0(4/7)	1(7/7)	0(6/7)	0(5/7)	1(7/7)	0(6/7)	0(2/7)	3
02	**paddy**	1(8/8)	-	0(0/8)	0(4/8)	1(8/8)	0(0/8)	0(4/8)	1(8/8)	0(4/8)	0(6/8)	1(8/8)	0(4/8)	0(6/8)	**4**
03	ほぼぼっち	0(4/8)	0(1/7)	-	0(4/8)	1(7/7)	0(5/7)	0(4/8)	1(8/8)	0(4/8)	0(2/8)	1(7/7)	0(4/8)	0(2/8)	3
05	Unhackable Anonymity	0(3/7)	0(1/7)	0(3/7)	-	1(7/7)	0(3/7)	0(5/7)	1(7/7)	0(3/7)	0(4/7)	1(7/7)	0(3/7)	0(3/7)	3
06	westlab	0(2/7)	0(3/7)	0(5/7)	0(5/7)	-	0(1/7)	0(4/7)	1(7/7)	0(5/7)	0(4/7)	1(7/7)	1(7/7)	0(5/7)	3
07	unicorn	**	**	0(2/8)	0(4/8)	0(0/7)	-	0(2/8)	1(7/7)	1(8/8)	0(2/8)	**	0(6/8)	**	2
08	ステテコ西垣2	0(9/11)	0(2/7)	0(2/11)	0(4/11)	1(184/184)	0(4/7)	-	1(338/338)	0(2/11)	0(3/11)	1(480/480)	0(7/11)	0(7/11)	2.9*
09	筑波大学	0(3/7)	0(1/7)	0(4/7)	0(2/7)	0(1/7)	0(2/7)	0(3/7)	-	0(5/7)	0(3/7)	1(7/7)	0(3/7)	0(1/7)	1
10	**MaeData**	0(2/7)	0(3/7)	0(3/7)	0(4/7)	1(7/7)	0(3/7)	0(4/7)	1(7/7)	-	1(7/7)	1(7/7)	0(5/7)	0(3/7)	**4**
11	匿名係長 NINJA HATTORI	0(0/8)	0(1/7)	0(4/8)	0(4/8)	1(184/184)	0(5/7)	0(4/8)	1(338/338)	0(4/8)	-	1(480/480)	0(4/8)	0(4/8)	3
12	X-面（仮）	0(4/7)	0(2/7)	1(7/7)	0(0/7)	0(2/7)	0(6/7)	0(0/7)	0(3/7)	0(3/7)	0(2/7)	-	0(3/7)	0(1/7)	1
13	あの一に	-	-	-	0(16/22)	-	-	-	**	**	-	-	-	**	0
14	先駆け	0(2/7)	0(1/7)	0(5/7)	0(0/8)	1(21/21)	0(4/7)	0(2/7)	1(21/21)	0(5/7)	0(2/7)	-	0(5/7)	-	2

本戦の結果

- 観察
  - 上位チームのほとんどがk=2の匿名化
  - ほとんどのチームが，m=7個の仮名を選んで再識別試行した.
  - チーム「ステテコ西垣」は予備戦で再識別された

PWS Cup 2018 の反省

再識別され

た

2-匿
7

だから安心
だ！

Satoshi Ito 20:36
冷静になって考えてみると，2人見捨てた2-匿名化データの，残りの5人を識別される確率って1/32じゃなくて1/8ですね

T. Mori 20:37
俺もデータ見てて気づいた

Satoshi Ito 20:37
2人組のどっちかを識別すりゃあいいわけだからね

2人ペアの
どちらかを
識別できれば
よいので
実際の確率は

$1/2^4 = 1/16$

$1/2^3 = 1/8$

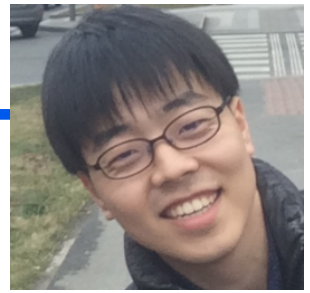
この勘違いのせいで
危険なデータを
提出してしまい
撃墜された

二つの仮説

- 伊藤説
  - 2-匿名化されていれば，m=7仮名識別される確率は，

    $(1/2)^7 = 1/128$

  - 1,2個無加工で放置しても安全だろう



- 山田説
  - 2-匿名化されていても，7仮名は3対しかない．各対について再識別される確率が1/2なので，全体では

    $(1/2)^3 = 1/8$

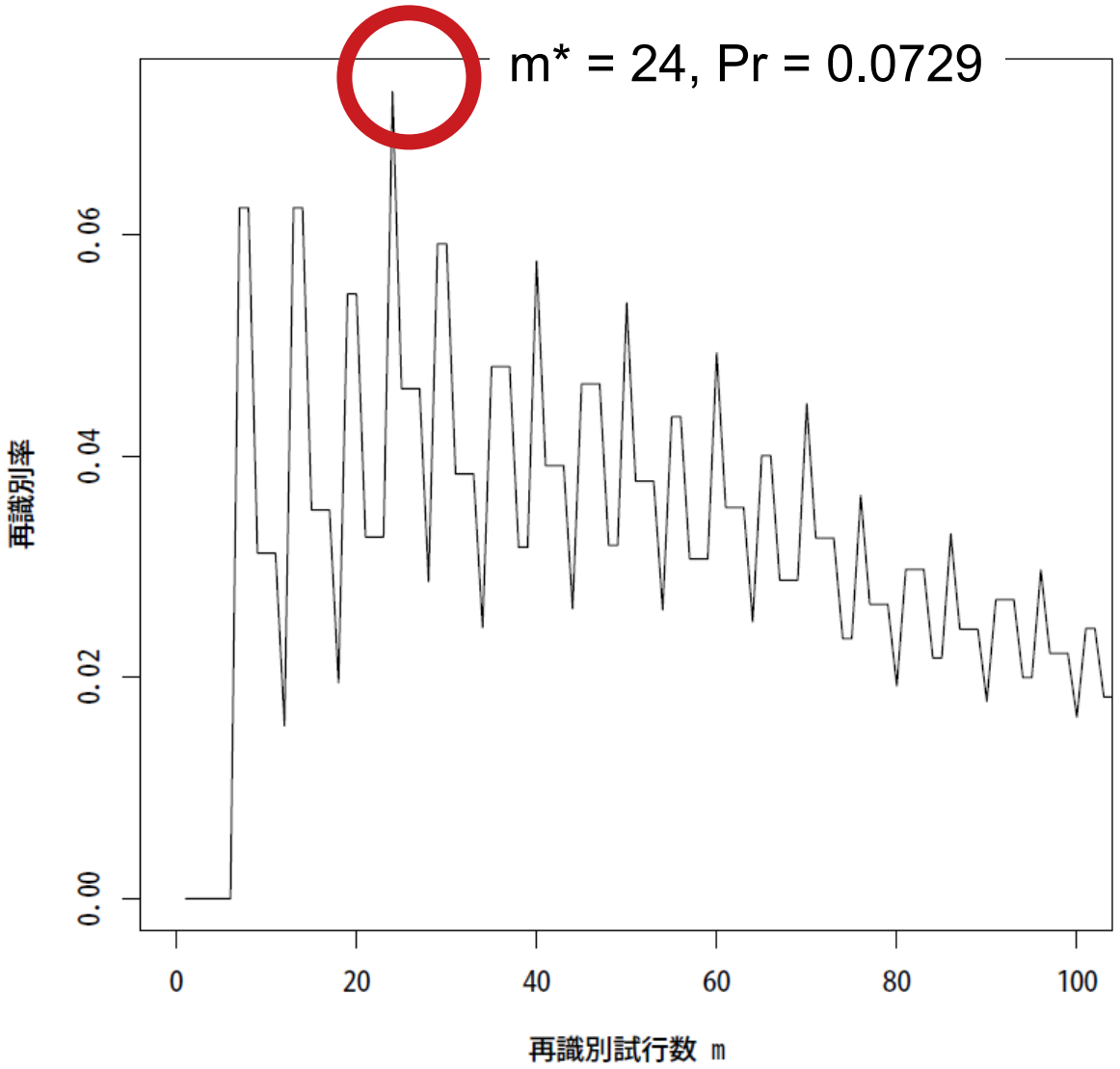
  - 全部加工しないと危ない

疑問

- Q1. Cup 2018の最適な仮名試行数mはいくらか?
- Q2. k-匿名化されたデータの正確な再識別率はいくらか？
  - 伊藤説は本当に誤りだったのか？
- Q3. k=2の時の「撃墜」の妥当な仮名数はいくらだったのか？

A1. 最適な再識別試行数m*

m	s	再識別率
7	7	0.0625
8	8	0.0625
9	9	0.0312
10	10	0.0312
11	10	0.0312
12	11	0.0156
13	11	0.0625
14	12	0.0625
15	13	0.0351
16	13	0.0351
17	14	0.0351
18	15	0.0195
19	15	0.0546
20	16	0.0546
21	17	0.0327
22	17	0.0327
23	18	0.0327
24	18	0.0729
25	19	0.0461
26	20	0.0461
27	20	0.0461
28	21	0.0286
29	21	0.0592
30	22	0.0592
31	23	0.0384



解析

- m仮名試行する時の再識別確率

$$\text{再識別率} = \sum_{i=B}^{A} \left(\frac{1}{k!}\right)^{i} \left(1 - \frac{1}{k!}\right)^{A-i} {}_{A}C_{i}$$

- 例) m = 24の時，安全性基準数b = 18,
  A=12(24/2), B=9(18/9), Pr(24)= 0.0729

A2. k-匿名化の安全性

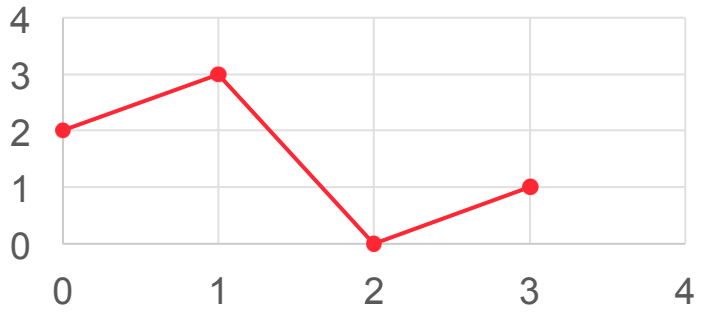
- 問題
  - k個の要素からなる仮名化データが，再識別される仮名の平均数を求めよ.
- 解
  - k個の仮名が再識別される期待値は1である.

$$\sum_{x=0}^{k} x \cdot \frac{f_k(x)}{k!} = \sum_{x=0}^{k} x \cdot Pr(\text{不動点が } x \text{ 個}) = 1$$

不動点数xと置換の数 $f_k(x)$

- 不動点
  - k個の置換の中で，不動な要素
  - k=3の例

	x	$f_k(x)$
1,**2**,**3**	3	1
1,3,2	1	3
2,1,**3**	1	
3,**2**,1	1	
2,3,1	0	2
3,1,2	0	
計		6 (=3!)



不動点数の理論値

- 不動点をx持つ置換の数 $f_k(x)$

$$f_k(x) = \binom{k}{x} a_{k-x}$$

  - 攪乱順列（k個の置換で不動点が0のものの数）

$$a_k = (k-1)(a_{k-1} + a_{k-2})$$

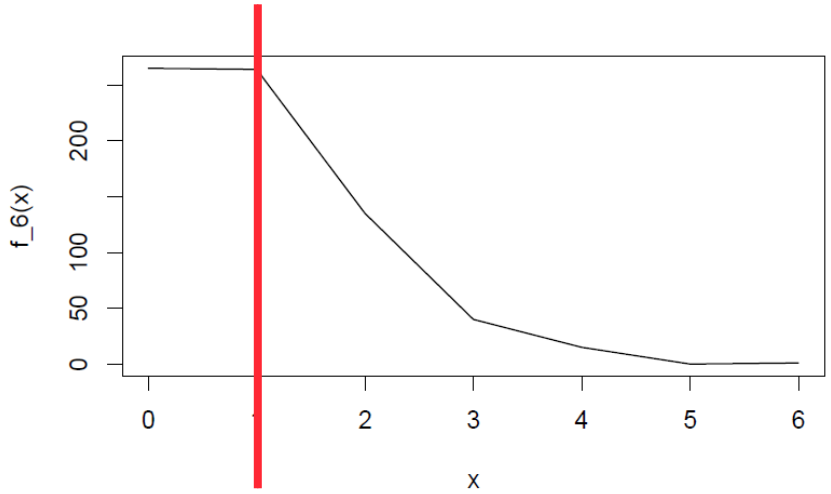
表 2　長さ k の順列の不動点数 $f_k(x)$

$x\backslash k$	2	3	4	5	6	7
7						1
6					1	0
5				1	0	21
4			1	0	15	70
3		1	0	10	40	315
2	1	0	6	20	135	924
1	0	3	8	45	264	1855
0	1	2	9	44	265	1854
sum	2	6	24	120	720	5040

不動点数の期待値

- （モンモール問題）
  - $f_k(x)$を不動点をk個持つn個の置換の数とする時

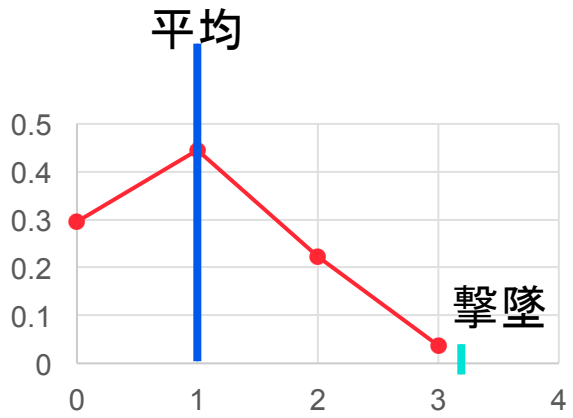
$$\sum_{x=0}^{k} x \cdot f_k(x) = k!$$



k=6の時,
不動点の期待値は1

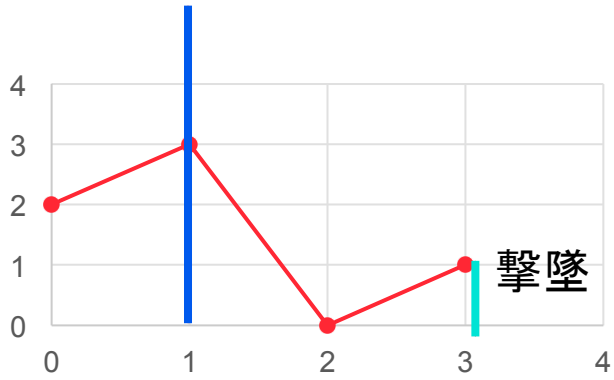
伊藤説は本当に誤りだったか？

- 伊藤説
  - p=1/kの二項分布



  - 期待値 E(B(m,x)) = m/p = 1

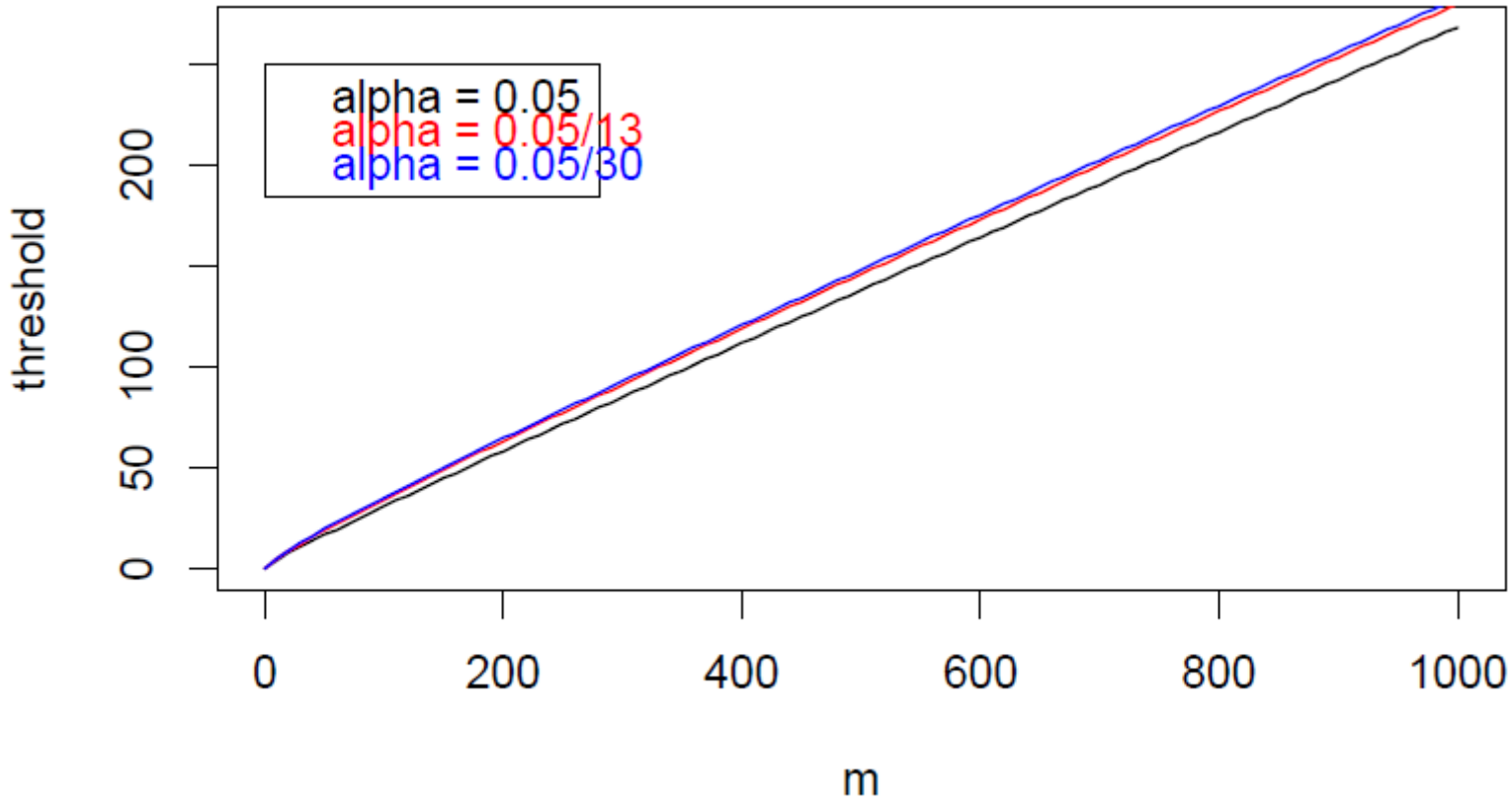
- 不動点
  - k個の要素の置換の不動点数



  - 期待値 E(f_k(x))= 1

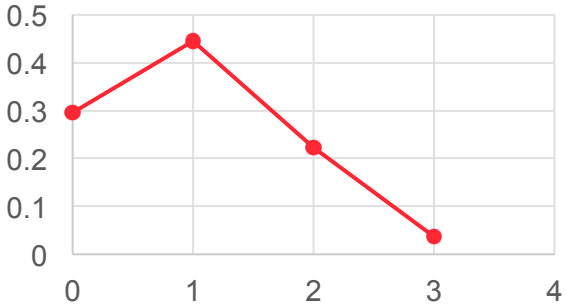
A3. 妥当な撃墜の試行数

- 試行数m についての棄却域

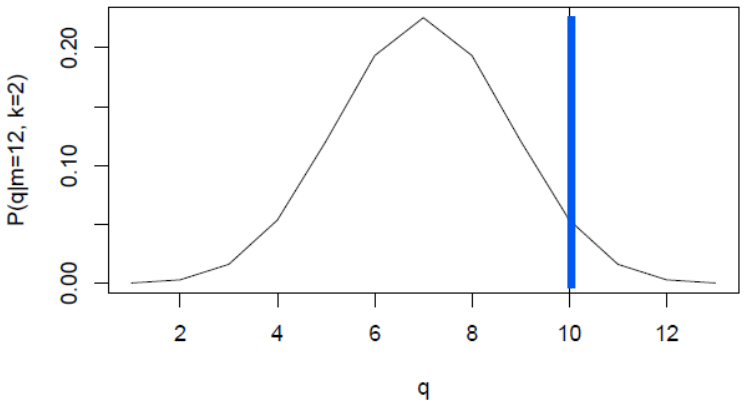


二つの確率分布

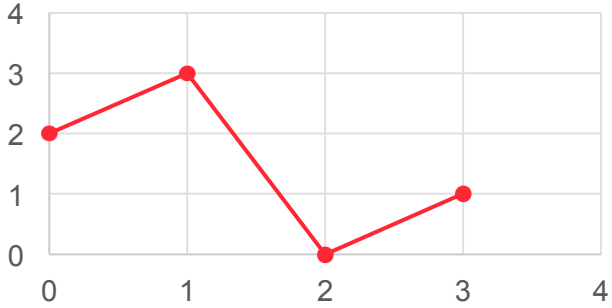
- 二項分布 (p=1/3)
  - k=3のクラスタ内



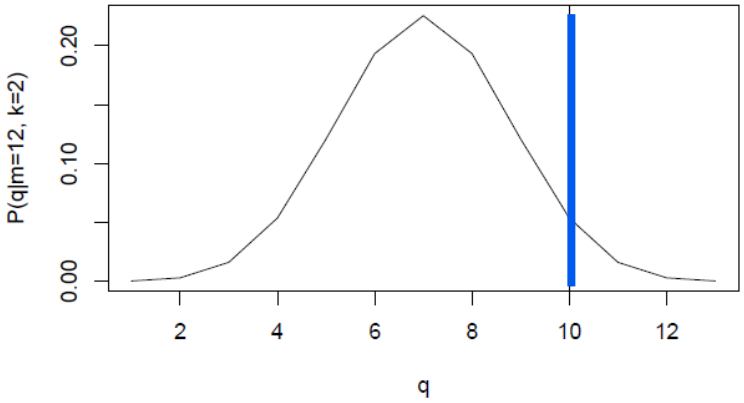
  - m/k個のクラスタの和



- 不動点の分布
  - クラスタ内



  - m/k個のクラスタの和



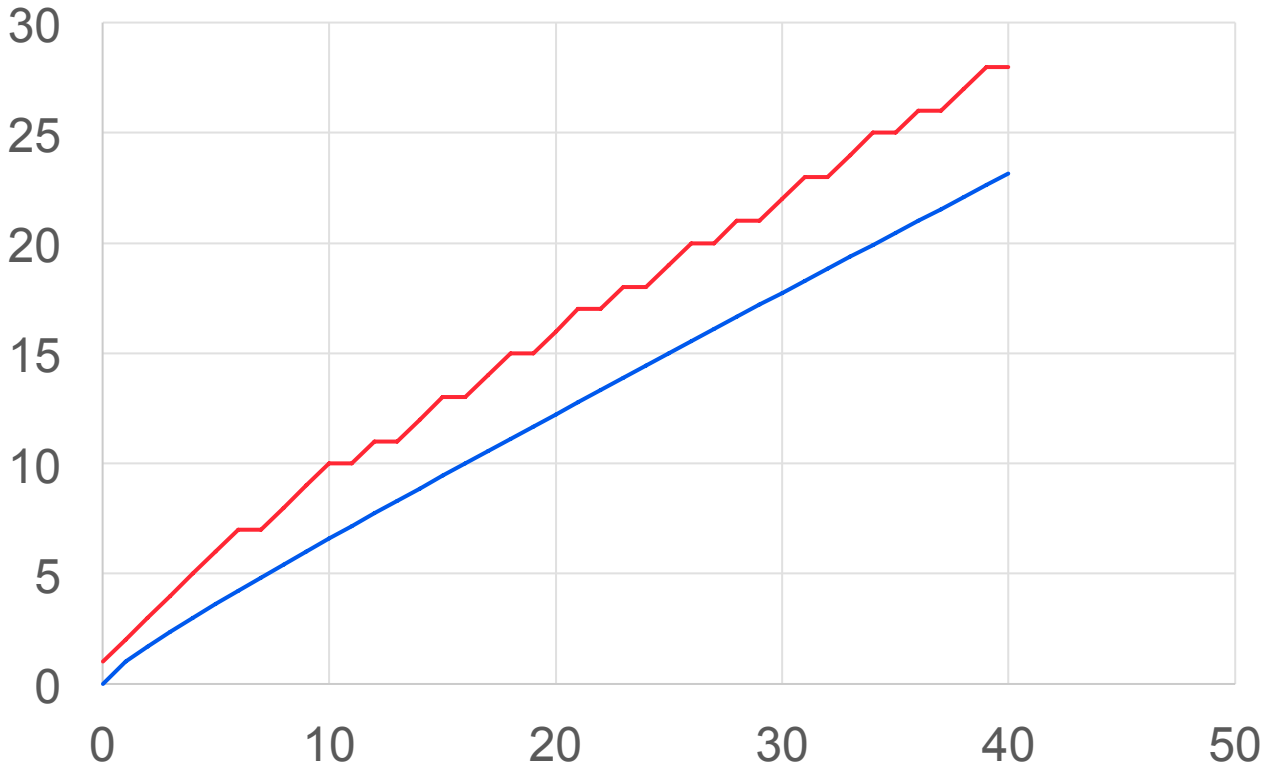
統計的仮説検定

- 二項検定
  - 帰無仮説 H0: k匿名化されたデータは安全である.（再識別は偶然に起きた）
  - 対立仮説 H1: k匿名化されたデータは安全でない（有意な再識別が起きた）
  - 有意水準α = 0.05とする. 帰無仮説H0の元で, q個の仮名が再識別される確率は, p=1/3のサイズm/kの二項分布B(m/k, 1/k)に従う

$$\mu + \sigma = \frac{m}{k^2} + \sqrt{\frac{k-1}{k^2} m}$$

Cup 2018との比較

- 「撃墜」の閾値の比較



結論

- CUP 2018の最適な試行数m*は24であった.
- k-匿名化されたデータの正確な再識別率は攪乱順列で算出される．任意のkについて，m=k回試行した時の期待値は1である.
- k=2の時の「撃墜」の仮名数はO(m+√m)に比例する.
- 伊藤説（２項分布）の期待値は正しかった