# k-匿名が使えない事例 Suicaの乗降履歴はなぜ匿名化できないのか?

菊池浩明
明治大学

# Suicaの中に格納されているデータ

- レコード
  - ID
  - 種別
  - 日付
  - 入駅ID
  - 出駅ID
  - 残高

```
IDm: 01010212360a7e13
端末種:改札機 処理:運賃支払 10/03/26 入:d2/2 出:d0/35 残高:5474 連番:351
端末種:改札機 処理:運賃支払 10/03/26 入:25/8 出:25/7 残高:5764 連番:349
端末種:改札機 処理:運賃支払 10/03/26 入:25/7 出:25/8 残高:5894 連番:347
端末種:改札機 処理:運賃支払 10/03/26 入:d0/35 出:d2/2 残高:6024 連番:345
端末種:車載端末 処理:バス 10/03/24 入:c6a/0 残高:6314 連番:343
端末種:車載端末 処理:バス 10/03/24 入:c6a/0 残高:6424 連番:342
端末種:物販 処理:物販 10/03/22 21:19 入:35/ea 残高:6634 連番:341
端末種:??? 処理:チャージ 10/03/21 入:e0/2e 出:0/0 残高:6912 連番:340
端末種:改札機 処理:運賃支払 10/03/19 入:92/4 出:92/c 残高:1912 連番:339
端末種:改札機 処理:運賃支払 10/03/19 入:92/c 出:92/4 残高:2132 連番:337
端末種:改札機 処理:運賃支払 10/03/17 入:92/a 出:92/c 残高:2352 連番:335
端末種:車載端末 処理:バス 10/03/17 入:c71/0 残高:2522 連番:333
端末種:改札機 処理:運賃支払 10/03/17 入:92/c 出:92/a 残高:2712 連番:332
端末種:車載端末 処理:バス 10/03/14 入:c6a/0 残高:2882 連番:330
端末種:改札機 処理:運賃支払 10/03/13 入:e0/2e 出:e0/32 残高:3092 連番:329
端末種:改札機 処理:運賃支払 10/03/12 入:e0/32 出:e0/2e 残高:3272 連番:328
端末種:車載端末 処理:バス 10/03/12 入:c6a/0 残高:3452 連番:326
端末種:車載端末 処理:バス 10/03/11 入:c6a/0 残高:3662 連番:325
端末種:車載端末 処理:バス 10/03/11 入:c6a/0 残高:3772 連番:324
端末種:車載端末 処理:バス 10/03/10 入:c6a/0 残高:3982 連番:323
```

# 匿名化の処理

- 処理
  - 1. 仮名化（IDを除くだけ）
  - 2. 属性削除（列削除)
  - 3. レコード削除(行削除）
  - 4. 一般化
  - 5. 統計化
- 注意.
  - *k*-匿名化, *l*-多様化などの性質はまだ考えない. 2,3,4の組み合わせはNP完全.
  - 他にも, ノイズデータを加える(摂動化), 確率的にデータを交換する(swap), 統計情報を基にサンプリング(re-sampling), 合成 (synthesis) などもある.

# 匿名化例

列削除

| 氏名 | 仮ID | 日付 | 乗駅 | 降駅 | 残高 |
|---|---|---|---|---|---|
| 菊池 | 3 | 10/14 | 新宿 | 中野 | 1200 |
| 高橋 | 4 | 10/14 | 新宿 | 三鷹 | 840 |
| 佐藤 | 5 | 10/14 | 新宿 | 御茶ノ水 | 600 |
| 菊池 | 3 | 10/15 | 中野 | 御茶ノ水 | 1020 |

駅数
s=4

| 仮ID | 駅1 | 駅2 |
|---|---|---|
| 3 | 新宿 | 中野 |
| 4 | 新宿 | 中野 |
| 5 | 新宿 | 御茶ノ水 |
| 3 | 中野 | 御茶ノ水 |

| 駅3 | 4駅 |
|---|---|
| 中野 | 御茶ノ水 |
| 三鷹 | 新宿 |
| 御茶ノ水 | 新宿 |

ユーザ数
n=3

# 「再識別化」の種類

(1) 特定個人
再識別可能

| | |
|---|---|
| 菊池 | 3 |

| 仮ID | 駅1 | 駅2 |
|---|---|---|
| 3 | 新宿 | 中野 |
| 4 | 新宿 | 中野 |
| 3 | 中野 | 御茶ノ水 |
| 4 | 中野 | 御茶ノ水 |
| 6 | 新宿 | 信濃町 |

(2) 識別非特定
（仮IDが同じ人を
リンク出来る）

(3) 一意識別
（その駅に降りた人が
一人しかいない）

SUICAの案件は(2)を見落としている.
k−匿名性で議論しているのは(3)だけ.

# 基本データ

- 人口
  - 東京 930万，神奈川 448, 千葉395, 埼玉 291, 栃木，群馬
  - n = 42,598,300
    $= 4 \times 10^7$
    (2012年4月1日, Wikipedia)
  - Suica 発行枚数 42,470,000
- 駅数
  - 総数 9,262件
  - 関東地方(東京 930, 神奈川 448, 千葉 395)
    m = 2,497
    $= 2.5 \times 10^3$
    (http://info.jmc.or.jp/ekiensen.html )
  - JR西日本 811駅，Pasmo 1,291駅，JR東海 149, JR西日本 430駅

# 評価1（仮名化の評価）

- 問題
  - 各カードにs個の駅名があるとすると，(全員が)再識別できるsはいくらか.
- 仮定
  - m個の駅乗降は一様に分布する独立事象
- 解
  - $m^s \geqq n$ (全ユーザ数）を解いて，s = 2.237駅.（3駅あれば全利用者が再識別できる）

# 乗降数

- JR, メトロ，私鉄
  - 上位100駅
  - 1日平均乗降数（のべ回数）
- JR
  - 100駅平均 114,697

東日本旅客鉄道株式会社

サイトマップ 検索 えきねっと

JR東日本トップ | 鉄道・駅のご利用案内 | Suica | 企業・IR・採用情報

## 各駅の乗車人員

JR東日本エリア内の1日平均の乗車人員を把握できる駅を掲載しています。
駅名をクリックするとその駅の情報を見ることができます。

※東日本大震災の影響により運転を見合わせていた区間の駅については掲載しておりません。

2012年度 | 2011年度 | 2010年度 | 2009年度 | 2008年度 | 2007年度
2005年度 | 2004年度 | 2003年度 | 2002年度 | 2001年度 | 2000年度

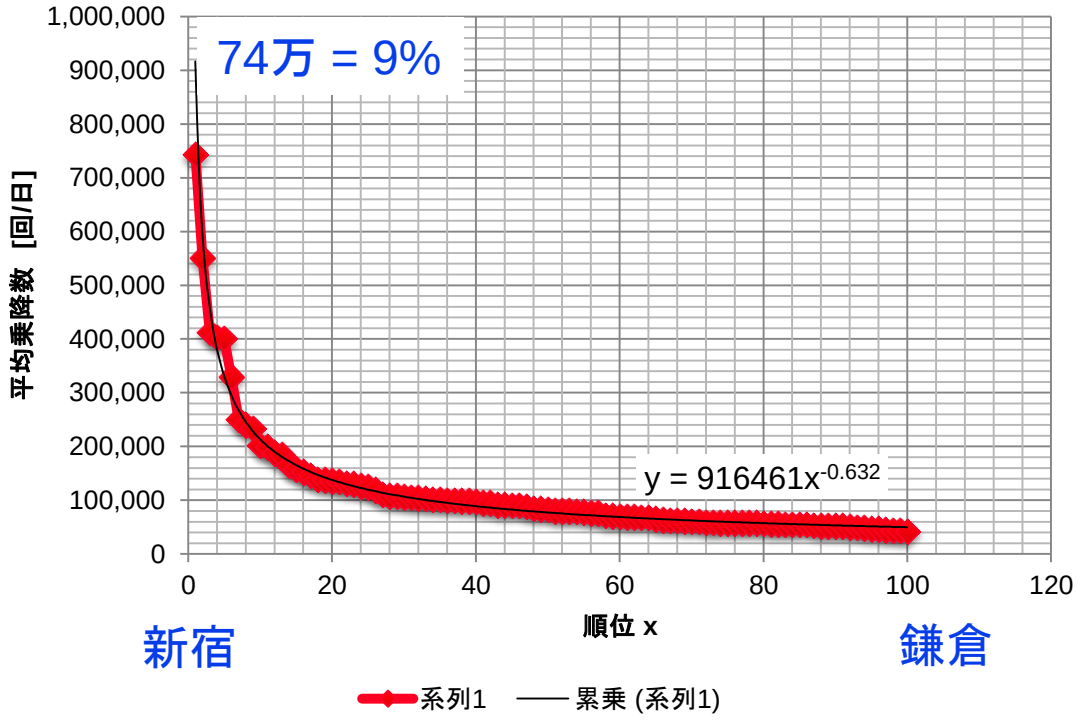
ベスト100
ベスト100以外の駅 (1) (2) (3) (4)
新幹線駅別乗車人員

| 順位 | 駅名 | 1日平均 定期外 | 定期 | 合計 |
|---|---|---|---|---|
| 1 | 新宿 | 346,974 | 395,859 | 742,833 |
| 2 | 池袋 | 229,055 | 321,700 | 550,756 |
| 3 | 渋谷 | 194,407 | 217,602 | 412,009 |
| 4 | 東京 | 189,621 | 212,655 | 402,277 |
| 5 | 横浜 | 154,250 | 246,404 | 400,655 |
| 6 | 品川 | 138,788 | 190,890 | 329,679 |
| 7 | 新橋 | 96,129 | 154,552 | 250,682 |
| 8 | 大宮 | 88,408 | 151,735 | 240,143 |
| 9 | 秋葉原 | 126,683 | 107,503 | 234,187 |
| 10 | 高田馬場 | 77,501 | 124,263 | 201,765 |
| 11 | 北千住 | 50,498 | 148,126 | 198,624 |

| 順位 | 駅名 | 定期外 |
|---|---|---|
| 51 | 浦和 | 27,715 |
| 52 | 仙台 | 42,006 |
| 53 | 武蔵溝ノ口 | 28,098 |
| 54 | 川口 | 26,236 |
| 55 | 登戸 | 25,279 |
| 56 | 鶴見 | 25,461 |
| 57 | 巣鴨 | 30,586 |
| 58 | 原宿 | 47,354 |
| 59 | 新小岩 | 23,309 |
| 60 | 代々木 | 36,224 |
| 61 | 舞浜 | 39,871 |

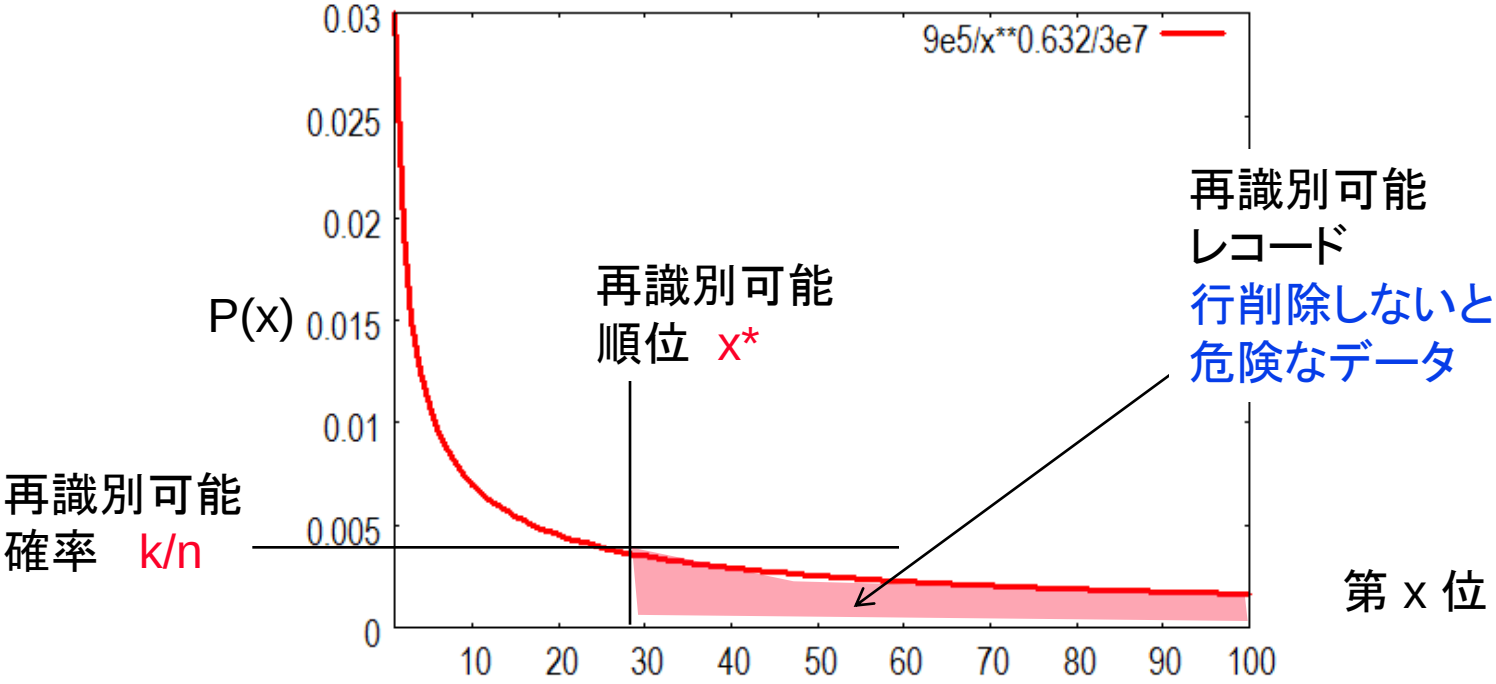
# 数理モデル

- Zipf則
  - ❑乗降数 $y = f(x) = 9x10^5/x^{0.632}$ [回/日]

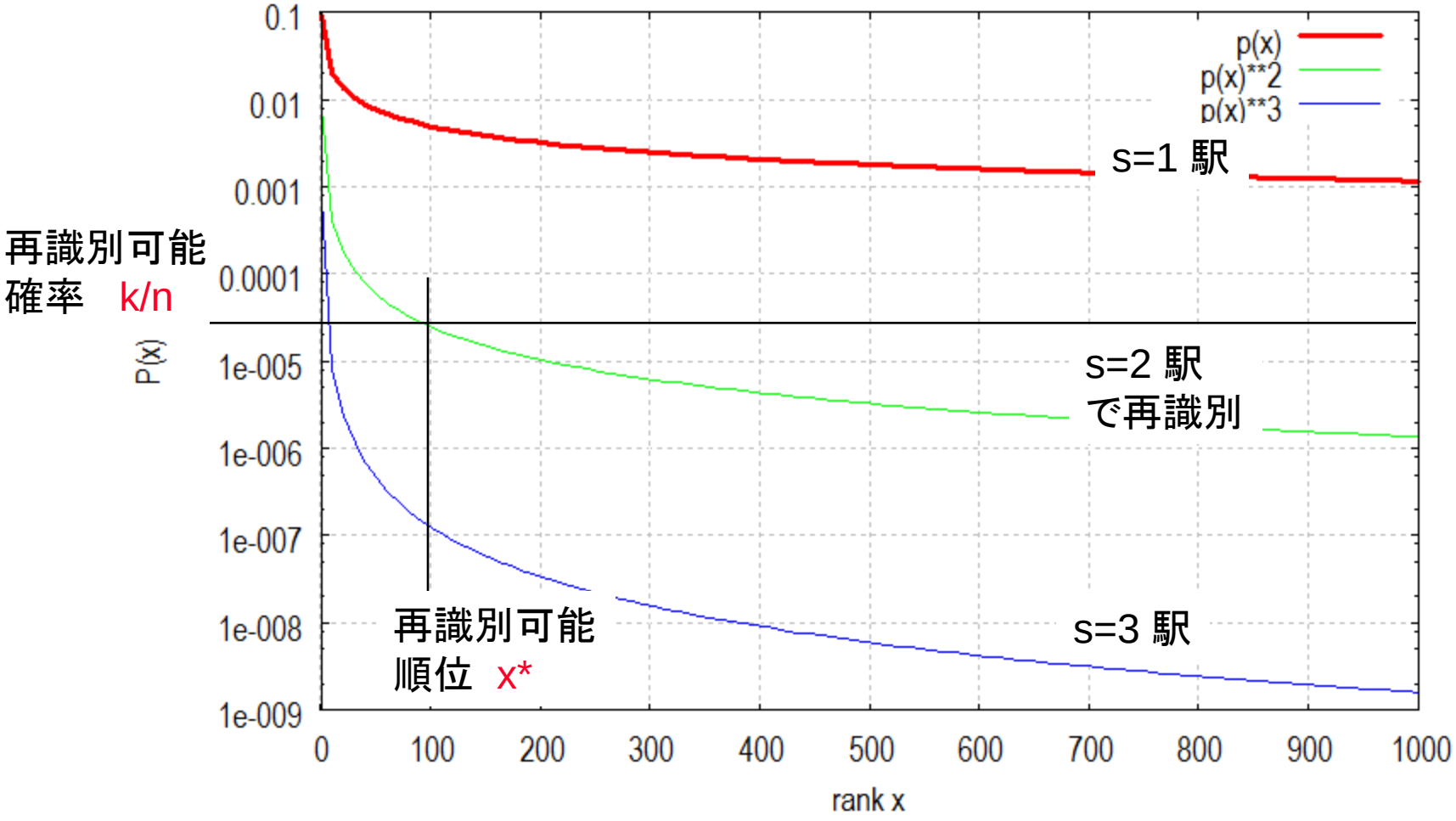


# 評価2 (k-匿名化の評価)

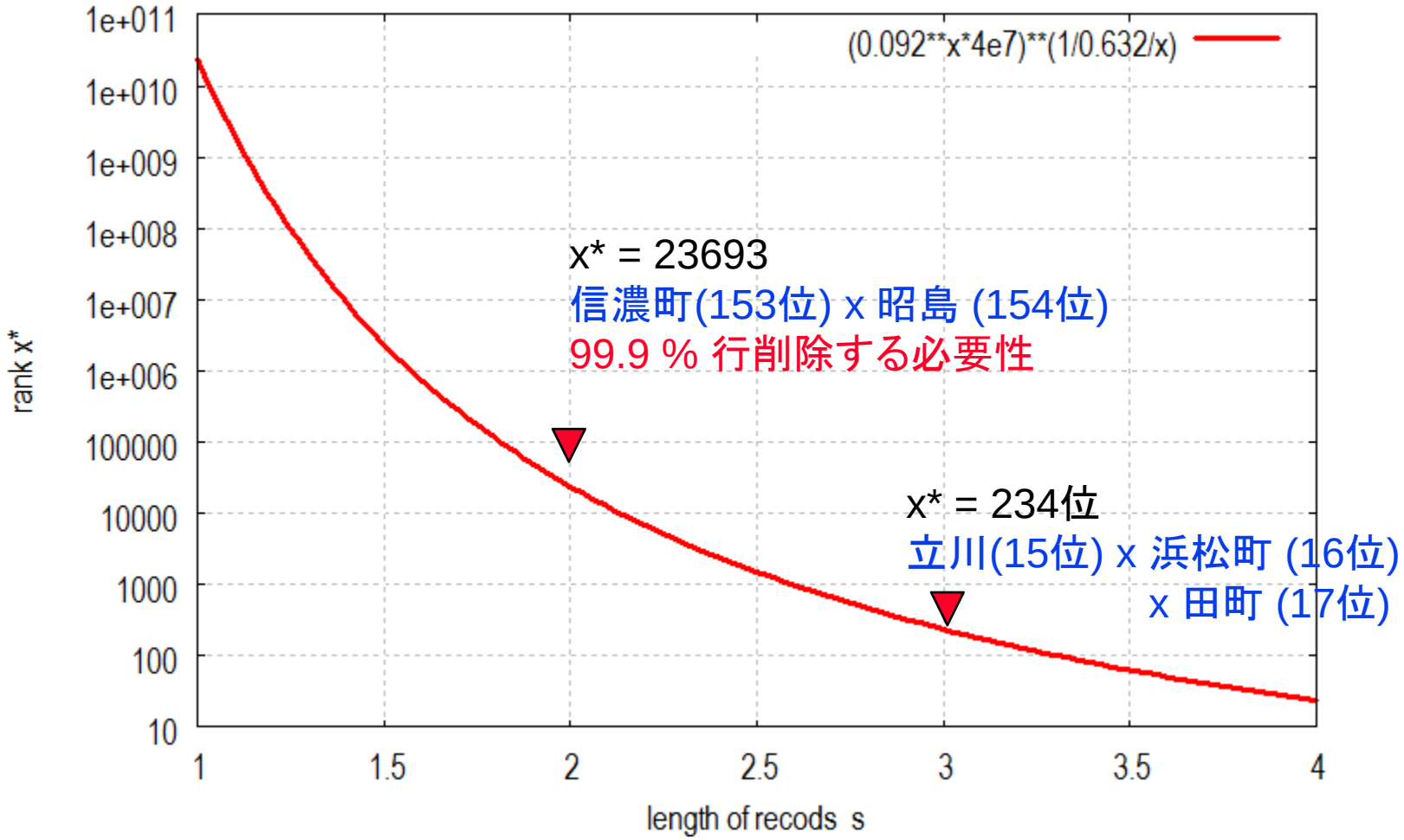
- 駅の確率分布 (一様分布仮定を外す)
  - N = $\int_1^m 9 \cdot 10^5 / x^{0.632}\, dx$ = [2.5 $x^{0.4}$]$^m$=3・$10^7$
  - p(x) = $f$(x)/N = x位の駅が選ばれる確率



# s>1の場合


再識別可能
確率 k/n
P(x)
0.1
0.01
0.001
0.0001
1e-005
1e-006
1e-007
1e-008
1e-009
0
100
200
300
400
500
600
700
800
900
1000
rank x
p(x)
p(x)**2
p(x)**3
s=1 駅
s=2 駅
で再識別
s=3 駅
再識別可能
順位 x*

# sに対する再識別順位x*



# 結論

- 列削除（仮名化)のみの匿名化では，3駅(=s)で再識別可能である.
  - 3回の利用ごとに仮名IDを付け替える必要.
- 行削除をしてk=2-匿名性を保証しても，s=2の時に99.9%のレコードを削除する必要がある.

# 再識別確率となる順位x*

- s=1 の時 （単一の駅から攻撃）
  - k = 同じ駅列を持つ利用者数
    = 1 = 存在的再識別可能（一人でも特定できる）
  - 1 = n p(x*) (kの期待値）を満たす順位

$x^* = (n\ 9e^5/\ N)^{10/6} = 5\ e^9 >> m$

(そんなに駅名はないので，特定不能= 安全）

- s = 2 (2駅分の情報から攻撃）
  - $p'(x) = p(x)^2$（駅選択が独立の仮定）
  - $x^* = (n/N\ 9e^5)^{1/0.63\ s} = 23698$

# 匿名化

- 列削除
  - 日付，降りる駅，金額
  - ID, 駅1, 駅2, …, 駅s
  - 例）

    ID=1, 新宿，池袋，新宿

    ID=2, 新宿，渋谷
    - » レコードをつなぎ，駅名の重複を許す
    - » ID当りの長さsは可変とする