

第88回CSEC・第182回DPS 合同研究発表会
2020/03/12

履歴データの数理モデルの 提案と **k** -匿名化に必要な ダミーレコード数推定への応用

伊藤 聡志 (明治大学大学院 先端数理科学研究科)
菊池 浩明 (明治大学 総合数理学部)

研究背景（匿名加工と疑似レコード）

匿名加工：データから個人が識別されることを防ぐために情報を加工する技術

元データ

User ID	Goods
1	Apple
2	Apple
2	Book
3	Book

购买商品から
個人が簡単に識別される

疑似レコード
追加



k -匿名化

加工データ

User ID	Goods
1	Apple
1	Book
2	Apple
2	Book
3	Apple
3	Book

3人の区別がつかない
(3-匿名化)

研究背景（疑似レコードの数）

疑似レコード数は加工データの**有用性評価値**として用いられる

- 原田玲央, 伊藤聡志, 菊池浩明, 「商品の特徴による再識別リスクとクラスタリングを用いた購買履歴データ匿名加工手法の提案」, SCIS-2017

顧客400人, 38,087レコードの購買履歴データをJaccard距離を用いて5-匿名化する際に, 疑似レコードが約**170,000**必要になる

- 匿名加工・再識別コンテストPWSCUP-2018

レコード数の多い上位2顧客の区別をつかなくするために
2,688レコードの追加/削除が必要である

研究概要

問題点

- ・ 実際に加工作るまで疑似レコードの数がわからない！
→ 作成したデータが役に立たないものであることも

研究目的

- ・ **加工をする前に**疑似レコードの(だいたいの)数が知りたい！

解決手法

- ・ 加工前に得られる値(元データの統計量や加工のパラメータ)から疑似レコードの数を見積もるモデルを作る

発表の流れ

1. 基礎定義, 疑似レコード数の求め方
2. なぜ疑似レコード数は加工後でないかわからない?
それを解決するための仮定・問題
3. 履歴データの数理モデルの提案
提案モデルの応用による疑似レコード数の見積もり
4. 提案モデルの分析

発表の流れ

1. 基礎定義, 疑似レコード数の求め方
2. なぜ疑似レコード数は加工後でないとわからない？
それを解決するための仮定・問題
3. 履歴データの数理モデルの提案
提案モデルの応用による疑似レコード数の見積もり
4. 提案モデルの分析

基礎定義

n : 顧客数, m : レコード数, l : 商品全種類数

$I(u_i)$: 顧客 u_i の購買商品集合

c : クラスタ数, Δm : 疑似レコード数

$s_i = |U_i|$: クラスタ U_i の大きさ

$I(U_i)$: U_i の購買商品

(例)右のデータの場合

$n = 3, m = 4, l = 2, I(u_2) = \{\text{Apple}, \text{Book}\}$

$c = 1, \Delta m = 2, s_1 = 3, I(U_1) = \{\text{Apple}, \text{Book}\}$

元データ D

User ID	Goods
u_1	Apple
u_2	Apple
u_2	Book
u_3	Book

加工データ D'

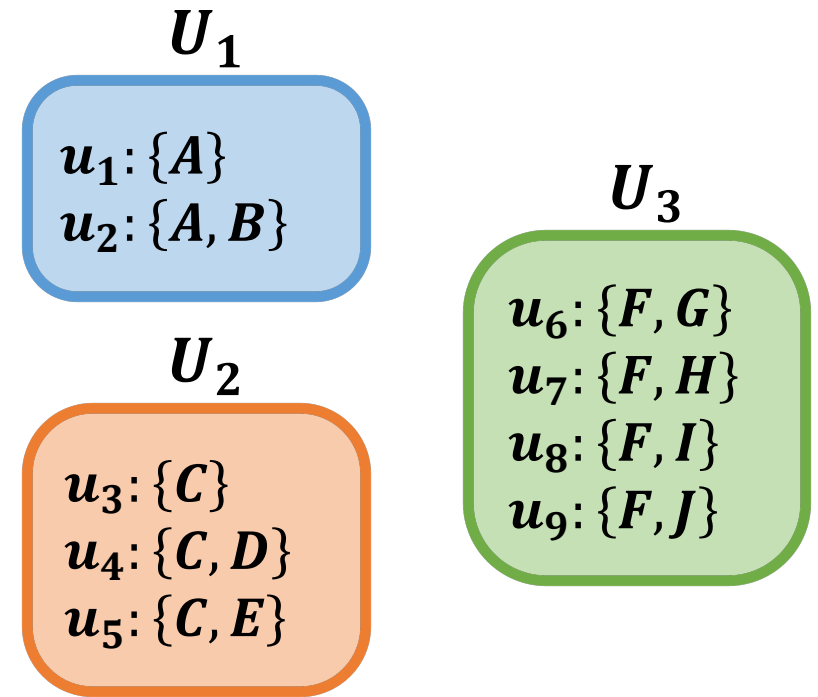
User ID	Goods
u_1	Apple
u_1	Book
u_2	Apple
u_2	Book
u_3	Apple
u_3	Book

疑似レコード数の厳密解

疑似レコード数は $\Delta m = \sum_{i=1}^c s_i |I(U_i)| - \sum_{i=1}^n |I(u_i)|$ で求められる

(例) 右図の3クラスタ内の顧客の区別がつかないようにするとき

$$\begin{aligned}\Delta m &= \sum_{i=1}^3 s_i |I(U_i)| - \sum_{i=1}^9 |I(u_i)| \\ &= s_1 |I(U_1)| + s_2 |I(U_2)| + s_3 |I(U_3)| \\ &\quad |I(u_1)| + \dots + |I(u_9)| \\ &= 2 * 2 + 3 * 3 + 4 * 5 - (16) \\ &= 17 \text{ レコード追加すればよい!}\end{aligned}$$



発表の流れ

1. 基礎定義，疑似レコード数の求め方
- 2. なぜ疑似レコード数は加工後でないとわからない？
それを解決するための仮定・戦略**
3. 履歴データの数理モデルの提案
提案モデルの応用による疑似レコード数の見積もり
4. 提案モデルの分析

Δm は加工しないと求められない！

$$\Delta m = \sum_{i=1}^c s_i |I(U_i)| - \sum_{i=1}^n |I(u_i)|$$

s_i : クラスタ U_i の大きさ (人数)

$|I(U_i)|$: U_i の購買商品の種類数

$|I(u_i)|$: 顧客 u_i の購買商品の種類数

これらの値は実際に加工をしないと手に入らない
しかし、これらを加工前に手に入る数で置き換えれば
加工をしなくても Δm の (おおまかな) 値を計算できる！

3つの仮定と問題

目的を達成するために、本研究では3つの仮定を置く

1. l 種類の値は独立で一様な確率 $1/l$ で生起する (**$1/l$ 仮定**)
2. n 人の顧客の c 個のクラスタの大きさはすべて等しく n/c である (**n/c 仮定**)
3. n 人の各顧客が持つ計 m 個のレコード数はすべて等しく m/n である (**m/n 仮定**)

s_i : n/c 仮定より, $s_i = n/c$

$|I(U_i)|$: m/c レコードあるときの商品種類数

$|I(u_i)|$: m/n レコードあるときの商品種類数

これらの期待値を求めることができれば
 $|I(U_i)|, |I(u_i)|$ を置き換えることができ、
 Δm の期待値を求めることができる！

[問題]

履歴データが x レコードあるとき、商品(全 l 種類)の種類数はいくらか？

クーポンコレクター問題

[問題]

一様に分布する l 種類のクーポンをすべて集めるためには何箱のシリアルを買えばよいか？

[答え]

購入箱数の期待値は $l \ln l + O(l)$ である

今回求めたいものは、 x 箱購入したときに集まっているクーポン(全 l 種類)の種類数の期待値なので、この解は使えない

発表の流れ

1. 基礎定義, 疑似レコード数の求め方
2. なぜ疑似レコード数は加工後でないとわからない？
それを解決するための仮定・問題
- 3. 履歴データの数理モデルの提案
提案モデルの応用による疑似レコード数の見積もり**
4. 提案モデルの分析

求めたいもの

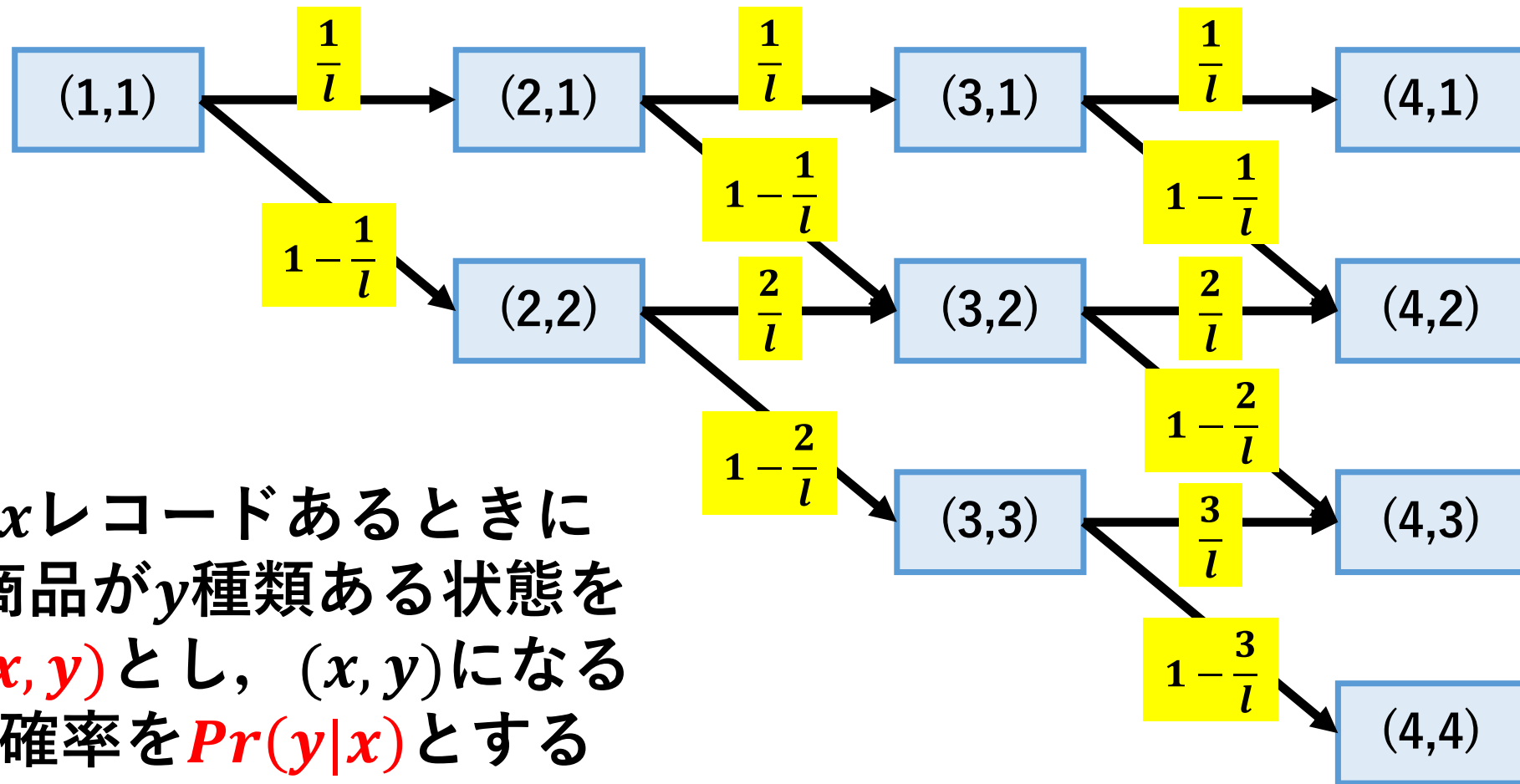
x : レコード数, y : 種類数, l : 全種類数, n : 顧客数
 m : 全レコード数, c : クラスタ数, Δm : 疑似レコード数

提案モデル

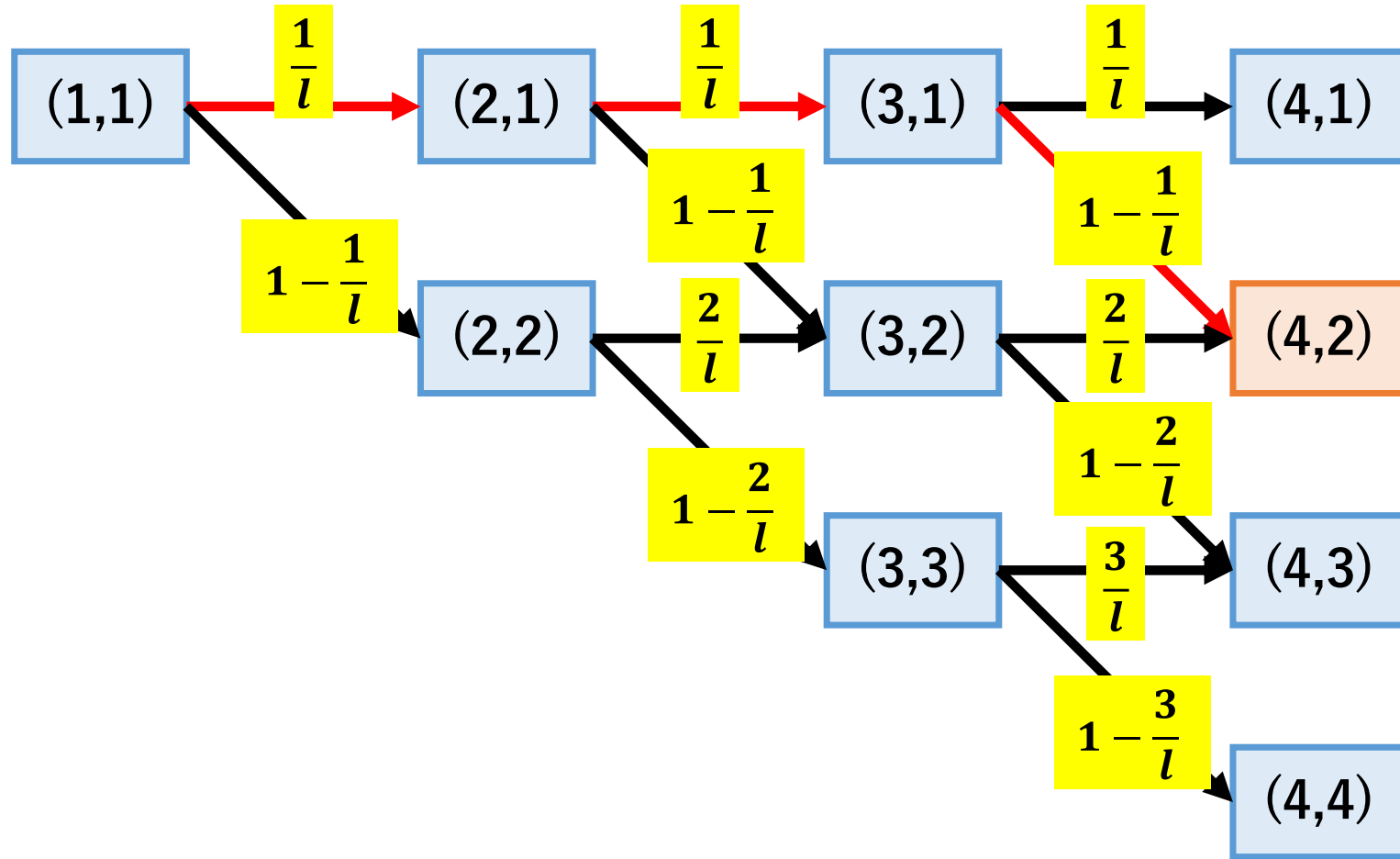
x レコードの履歴データが y 種類の商品(全 l 商品)を持つ確率	$Pr(y x) = ???$
x レコードの履歴データが持つ 商品種類数(全 l 商品)の期待値	$f(x, l) = ???$
加工に必要な疑似レコード数 Δm の期待値	$E(\Delta m) = ???$

提案モデルの応用

レコード数と商品種類数の状態遷移図



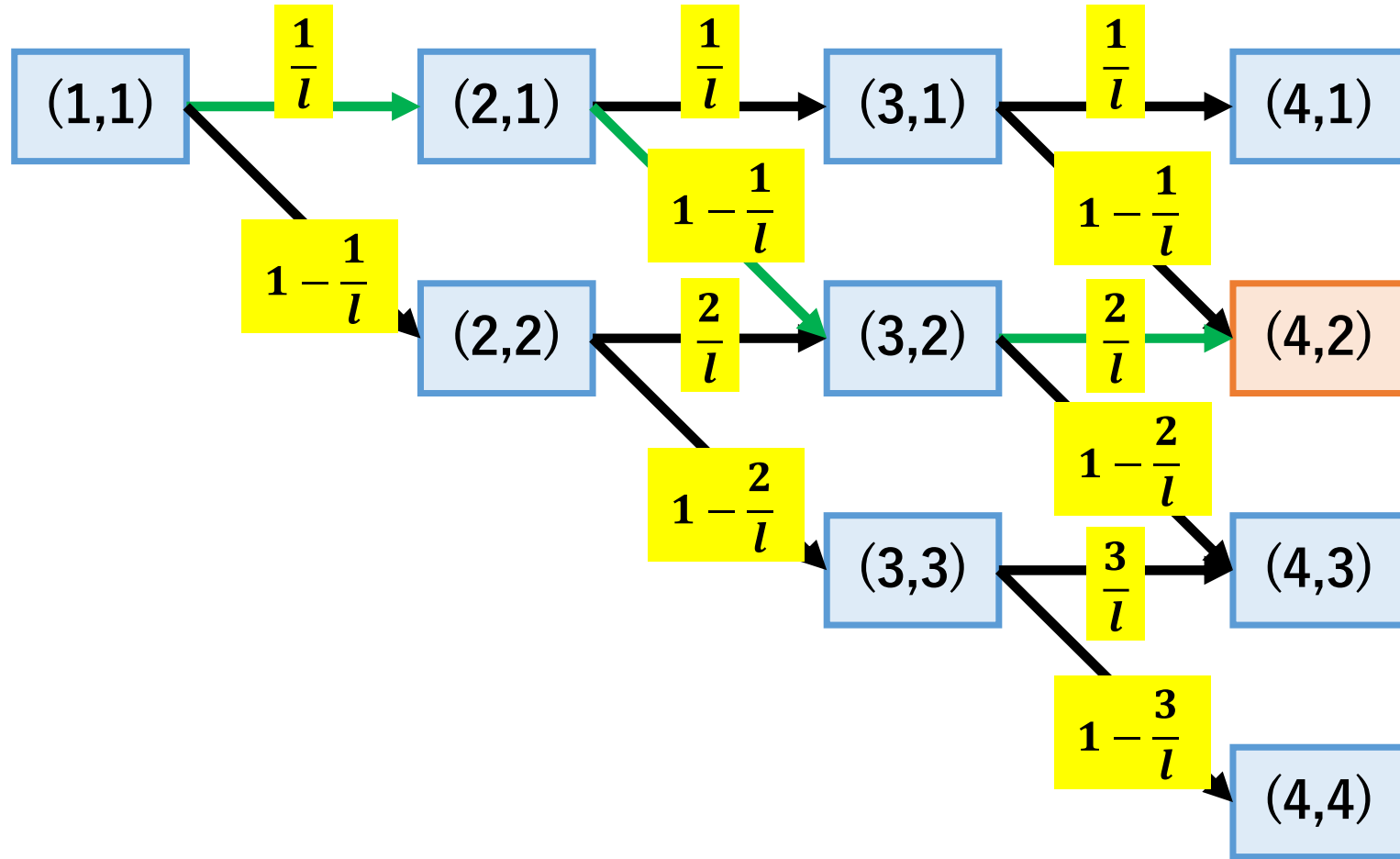
レコード数と商品種類数の状態遷移図



$Pr(2|4)$

$$= \frac{1}{l} * \frac{1}{l} * \left(1 - \frac{1}{l}\right)$$

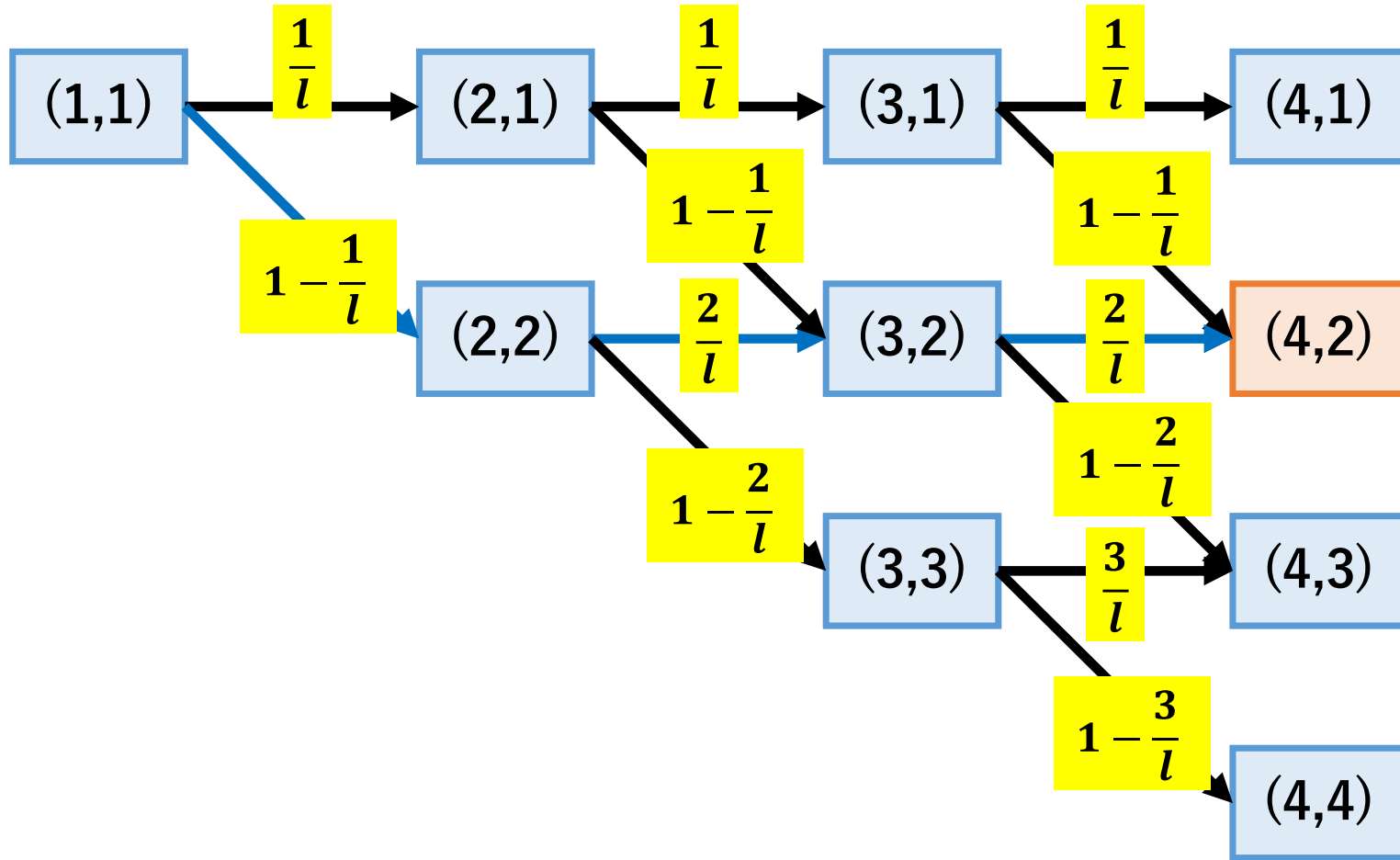
レコード数と商品種類数の状態遷移図



$Pr(2|4)$

$$= \frac{1}{l} * \frac{1}{l} * \left(1 - \frac{1}{l}\right) + \frac{1}{l} * \left(1 - \frac{1}{l}\right) * \frac{2}{l}$$

レコード数と商品種類数の状態遷移図



$$Pr(2|4)$$

$$= \frac{1}{l} * \frac{1}{l} * \left(1 - \frac{1}{l}\right) + \frac{1}{l} * \left(1 - \frac{1}{l}\right) * \frac{2}{l} + \left(1 - \frac{1}{l}\right) * \frac{2}{l} * \frac{2}{l}$$

$$= \frac{7}{l^2} \left(1 - \frac{1}{l}\right)$$

$Pr(\mathbf{y}|\mathbf{x})$ の一般式

右図より,

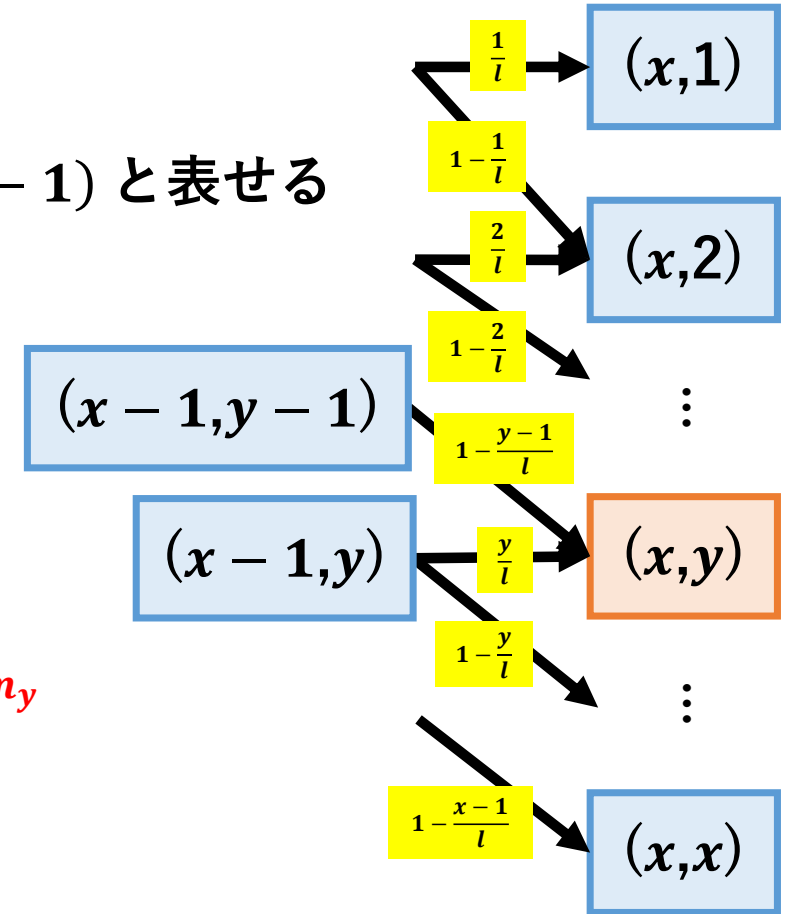
$$Pr(y|x) = \left(1 - \frac{y-1}{l}\right) Pr(y-1|x-1) + \left(\frac{y}{l}\right) Pr(y|x-1) \text{ と表せる}$$

[定理2.1]

l 種類の値をとる x レコードのデータの属性中に y 種類の値がある生じる確率 $Pr(y|x)$ は

$$Pr(\mathbf{y}|\mathbf{x}) = \prod_{j=0}^{y-1} \left(1 - \frac{j}{l}\right) * \sum_{m_1 + \dots + m_y = x-y} \left(\frac{1}{l}\right)^{m_1} \dots \left(\frac{y}{l}\right)^{m_y}$$

である. ただし, $m_1, \dots, m_y \geq 0$, $x \geq y \geq 1$ である.



$f(x, l)$ の一般式 その1

[補助定理2.1]

x レコードの履歴データが持つ値の種類数(全 l 種類)の期待値は

$$f(x, l) = \sum_{i=1}^x \binom{x}{i} \left(-\frac{1}{l}\right)^{i-1} \text{である.}$$

[補助定理2.2]

$$\sum_{i=1}^x \binom{x}{i} \left(-\frac{1}{l}\right)^{i-1} = (-l) \left(1 - \frac{1}{l}\right)^x + l \text{である.}$$

[定理2.2]

$$f(x, l) = (-l) \left(1 - \frac{1}{l}\right)^x + l \text{である.}$$

$f(x, l)$ の一般式 その 2

$$\begin{aligned}\sum_{i=1}^x \binom{x}{i} \left(-\frac{1}{l}\right)^{i-1} &= \sum_{i=1}^x \binom{x}{i} \left(-\frac{1}{l}\right)^i \left(-\frac{1}{l}\right)^{-1} \\&= (-l) \sum_{i=1}^x \binom{x}{i} \left(-\frac{1}{l}\right)^i \\&= (-l) \left\{ \sum_{i=0}^x \binom{x}{i} \left(-\frac{1}{l}\right)^i - \binom{x}{0} \left(-\frac{1}{l}\right)^0 \right\} \\&= (-l) \sum_{i=0}^x \binom{x}{i} \left(-\frac{1}{l}\right)^i \mathbf{1}^{x-i} + l \\&= (-l) \left(\mathbf{1} - \frac{1}{l}\right)^x + l \\&= f(x, l)\end{aligned}$$

二項定理

$$(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}$$

疑似レコード数の期待値 $E(\Delta m)$

$$\Delta m = \sum_{i=1}^c \textcolor{teal}{s}_i |I(\textcolor{brown}{U}_i)| - \sum_{i=1}^n |I(\textcolor{violet}{u}_i)|$$

と3つの仮定より, Δm の期待値は

$$E(\Delta m) = \sum_{i=1}^c \frac{\textcolor{teal}{n}}{\textcolor{teal}{c}} \textcolor{brown}{f}\left(\frac{\textcolor{brown}{m}}{\textcolor{teal}{c}}, l\right) - \sum_{i=1}^n \textcolor{violet}{f}\left(\frac{\textcolor{violet}{m}}{\textcolor{violet}{n}}, l\right) \text{ とかける.}$$

[定理3.1]

履歴データの加工に必要な疑似レコード数の期待値は,

$1/l$ 仮定, n/c 仮定, m/n 仮定のもと,

$$\textcolor{red}{E}(\Delta m) = \textcolor{red}{n}l \left\{ \left(1 - \frac{1}{l}\right)^{\frac{\textcolor{red}{m}}{\textcolor{red}{n}}} - \left(1 - \frac{1}{l}\right)^{\frac{\textcolor{red}{m}}{\textcolor{red}{c}}} \right\} \text{ である.}$$

$$\begin{aligned} n &= 400, m = 38000, \\ l &= 2700, c = 50 \text{ のとき,} \\ E(\Delta m) &= 227658.4 \end{aligned}$$

提案モデルまとめ

x : レコード数, y : 種類数, l : 全商品種類数
 n : 顧客数, m : 全レコード数, c : クラスタ数
 Δm : 疑似レコード数

提案モデル

x レコードの履歴データが y 種類の商品(全 l 商品)を持つ確率	$Pr(y x) = \prod_{j=0}^{y-1} \left(1 - \frac{j}{l}\right) * \sum_{m_1 + \dots + m_y = x-y} \left(\frac{1}{l}\right)^{m_1} \dots \left(\frac{y}{l}\right)^{m_y}$
x レコードの履歴データが持つ 商品種類数(全 l 商品)の期待値	$f(x, l) = (-l) \left(1 - \frac{1}{l}\right)^x + l$
加工に必要な疑似レコード数 Δm の期待値	$E(\Delta m) = nl \left\{ \left(1 - \frac{1}{l}\right)^{\frac{m}{n}} - \left(1 - \frac{1}{l}\right)^{\frac{m}{c}} \right\}$

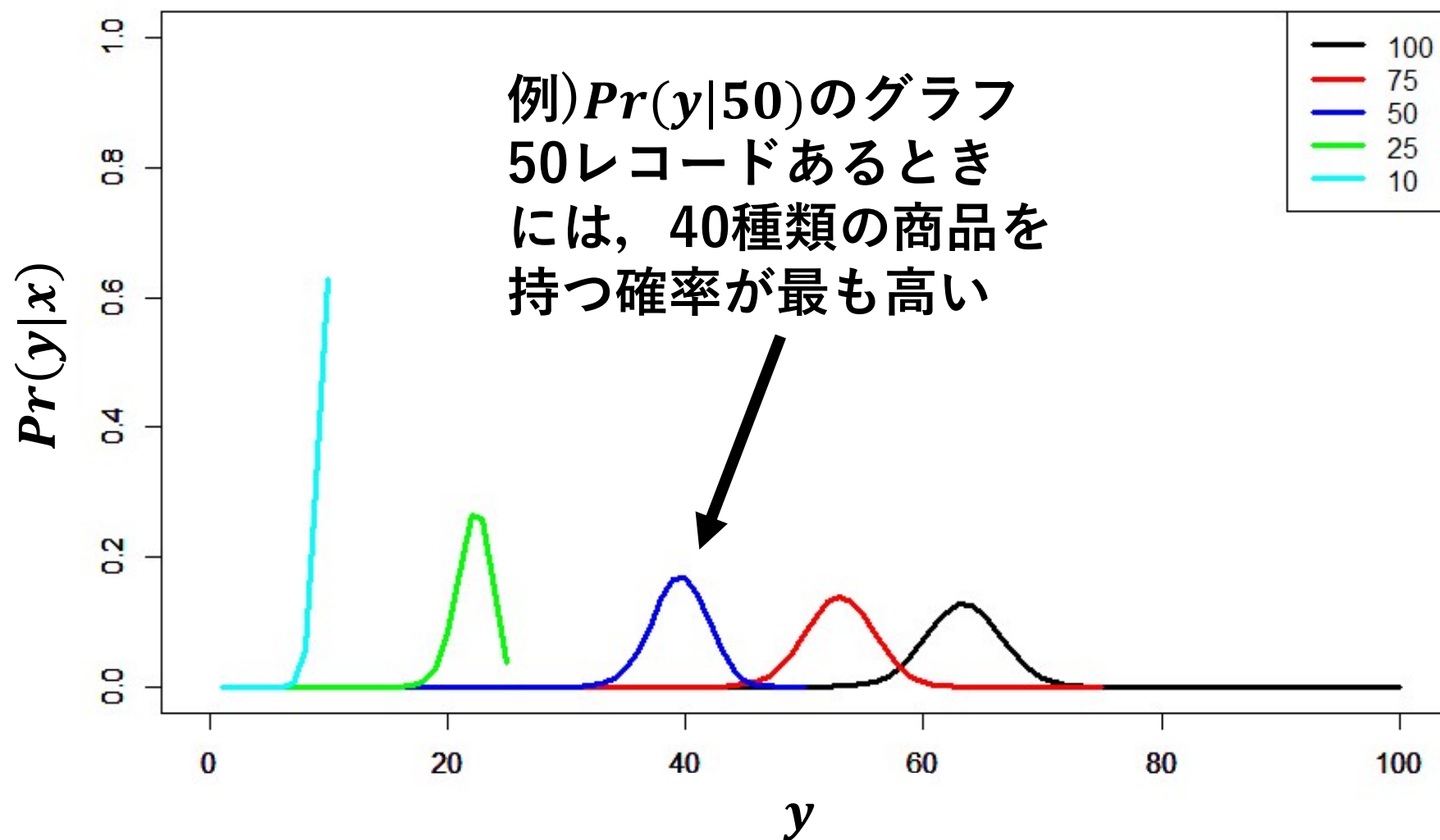
提案モデルの応用

発表の流れ

1. 基礎定義， 疑似レコード数の求め方
2. なぜ疑似レコード数は加工後でないかわからない？
それを解決するための仮定・問題
3. 履歴データの数理モデルの提案
提案モデルの応用による疑似レコード数の見積もり
4. 提案モデルの分析

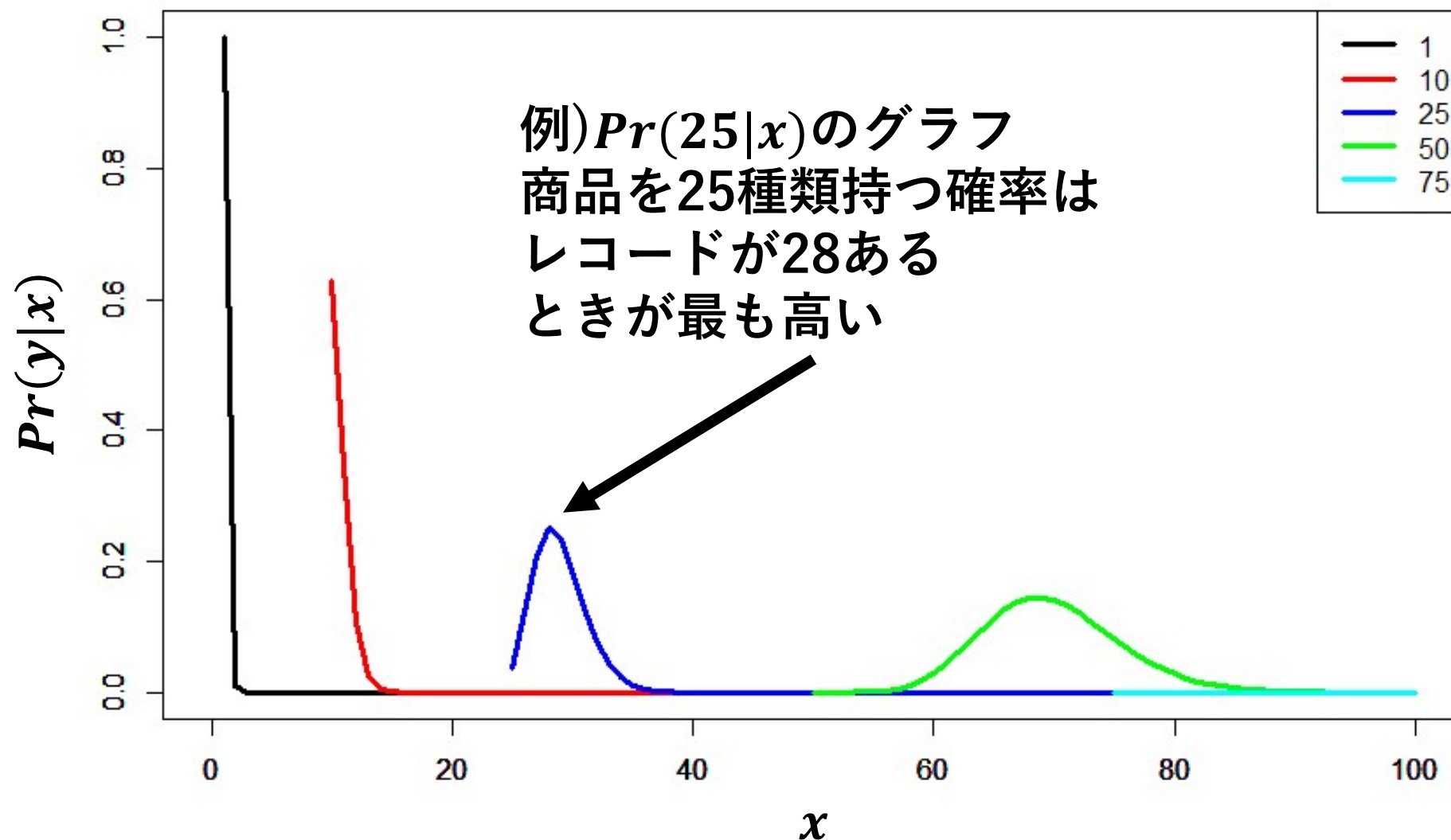
$Pr(y|x)$ の分布 1 ($l = 100$)

レコード数 x

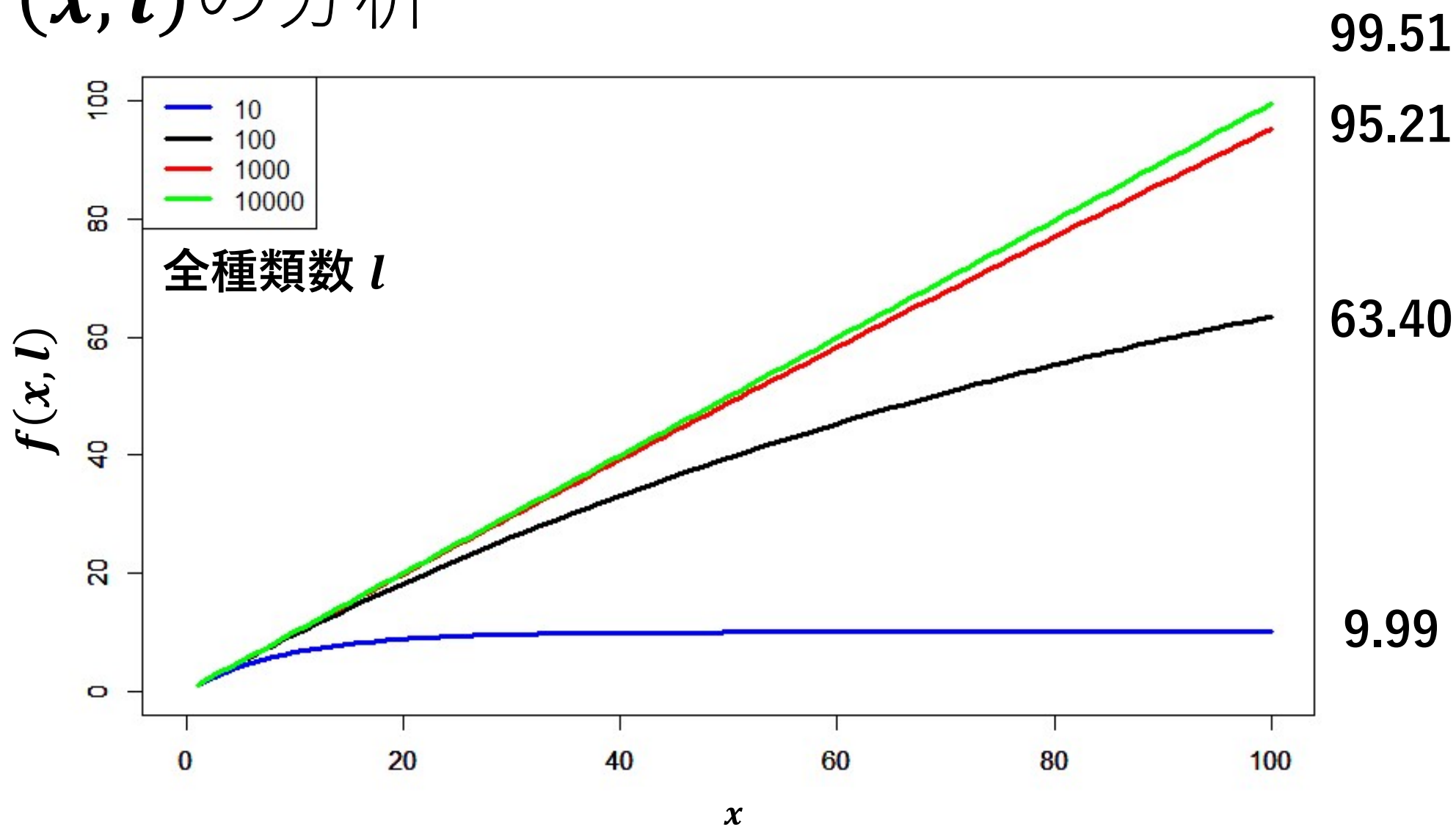


$Pr(y|x)$ の分布 2 ($l = 100$)

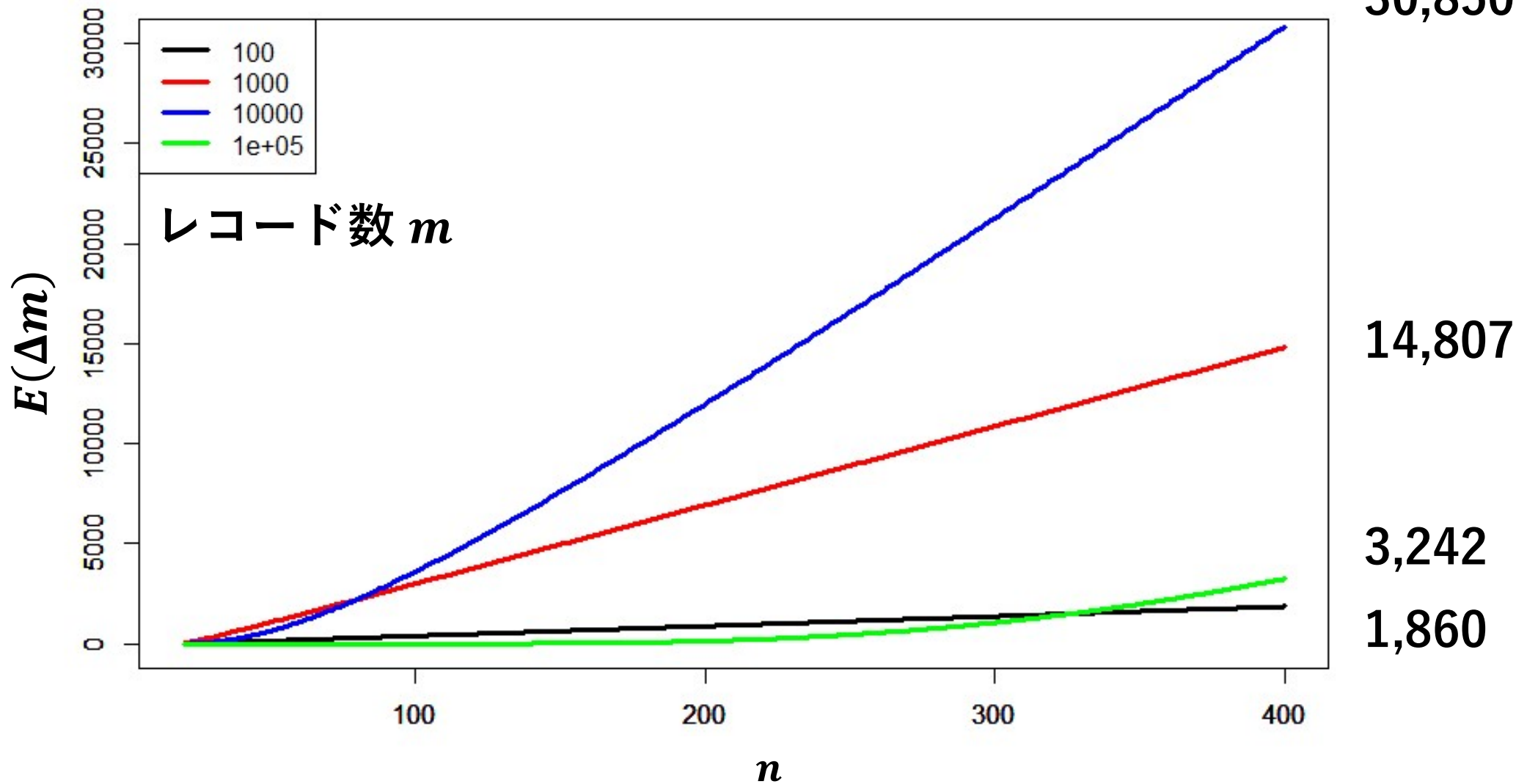
商品種類数 y



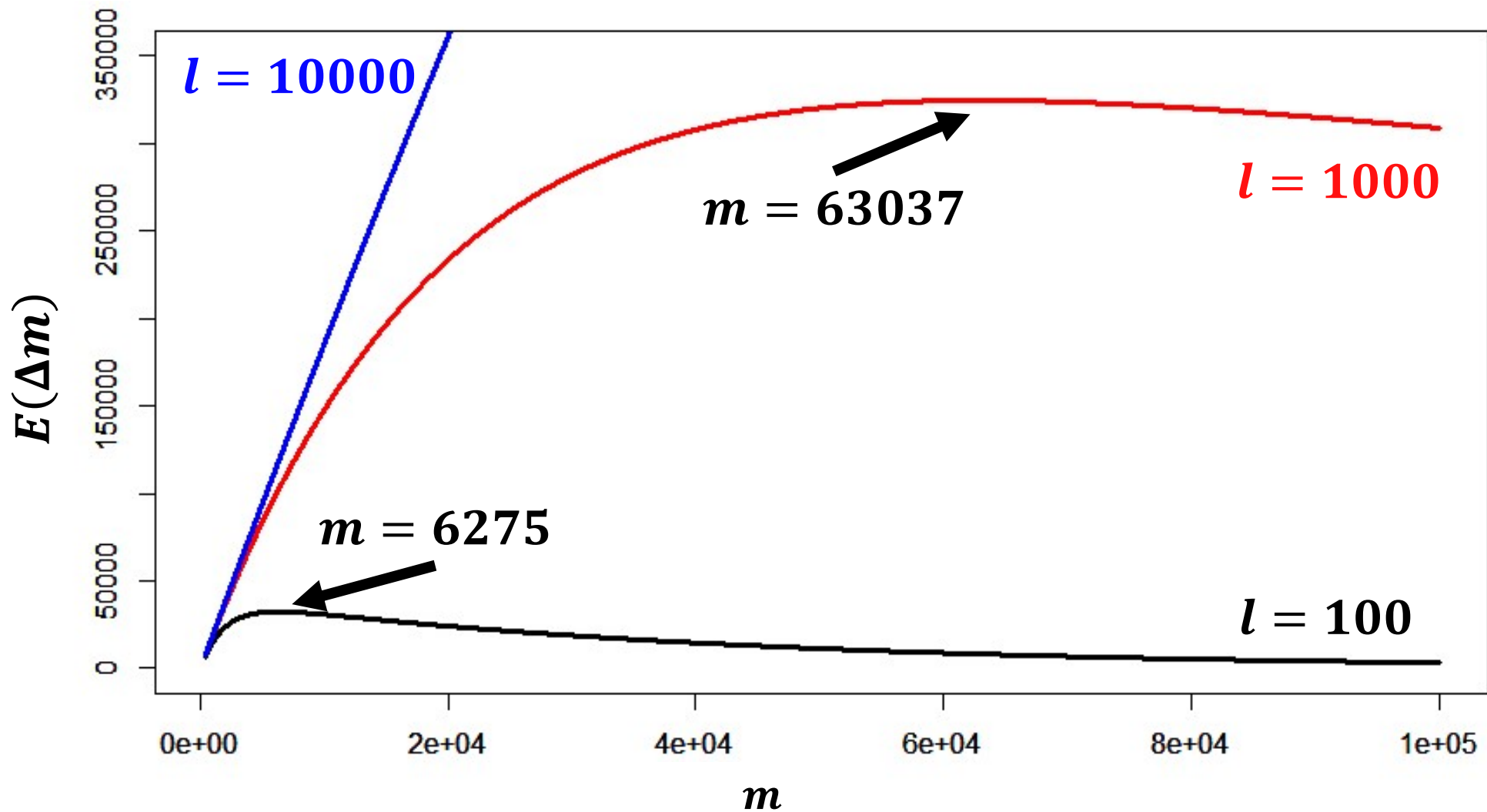
$f(x, l)$ の分析



$E(\Delta m)$ の分析 1 ($l = 100, c = 20$)



$E(\Delta m)$ の分析 2 ($n = 400, c = 20$)

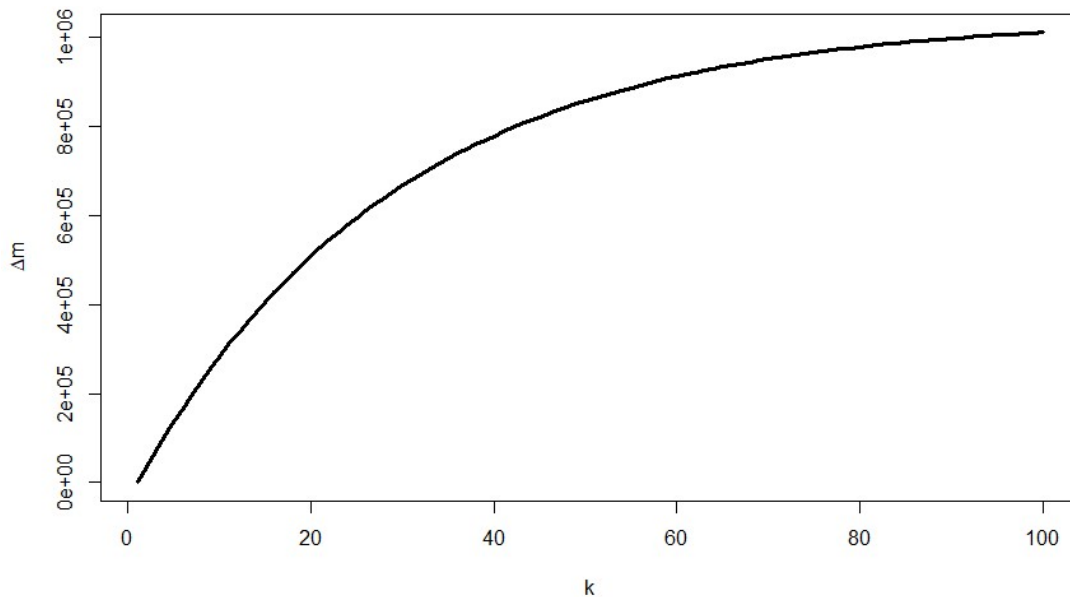


k -匿名化との関係

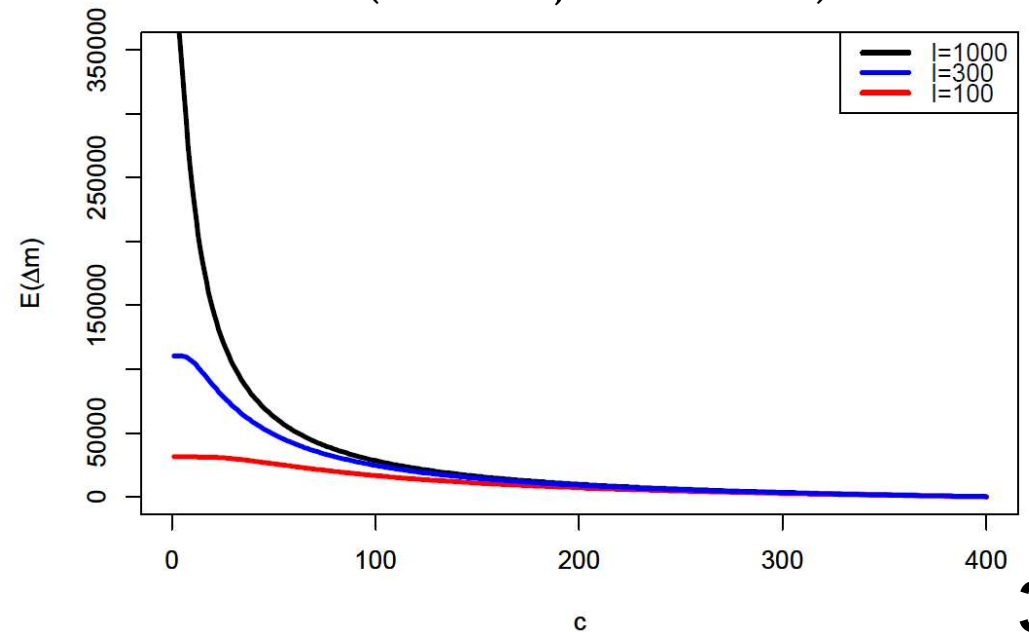
m/c 仮説と m/n 仮説より, $k \leq n/c$ であるため,

$$E(\Delta m) \geq nl \left\{ \left(1 - \frac{1}{l}\right)^{\frac{m}{n}} - \left(1 - \frac{1}{l}\right)^{\frac{km}{n}} \right\} \text{とおくこともできる(系3.1)}$$

k と $E(\Delta m)$ の関係
($n = 400, m = 38000, l = 2700$)



c と Δm の関係
($n = 400, m = 10000$)



まとめ

- データを匿名加工する際に，疑似レコードを追加する必要がある場合があるが，その数は実際に加工しないとわからない
- x レコードの履歴データが y 種類の値を持つ確率 $Pr(y|x)$ や，種類数の期待値 $f(x, l)$ を与える数理モデルを提案した
- 提案モデルを応用することにより，履歴データの k -匿名化に必要な疑似レコード数の期待値 $E(\Delta m)$ を，加工前に手に入る値から求めた
- 提案モデルの改善や他の応用が今後の課題である