Vol.2022-DPS-190 No.43
Vol.2022-CSEC-96 No.43
2022/3/11

オンライン教育データ解析支援システムの開発に向けた匿名化アルゴリズムに関する妥当な設定値の検討

高木 理[1]　浜元信州[1]　竹房あつ子[2]　横山重俊[2]　合田憲人[2]

概要：本研究では，オンラインによる教育活動を支援する学習管理システム（LMS）上のデータ解析を安全に行うことを目的として，匿名化の対象となるデータに対してノイズを付加しながらデータの統計処理を行うアルゴリズム（本稿では，このようなアルゴリズムも匿名化アルゴリズムと見做す）に着目し，その設定値（パラメータ）の妥当性を吟味するための基準を提案する．本稿の後半において，提案された基準を満たす設定が実際に出来るのかどうかを検討するために，実在の大規模データを用いて匿名化アルゴリズムによる統計処理を行い，基準を満たす設定値がどのようなものになるのかを検討する．さらに，従来の設定の仕方では基準を満たす設定が出来ない場合の対処法についても議論する．なお，本研究では，大規模な LMS データの取得が困難であったため，実在の商品レビューデータで代用して議論を進める．

キーワード：データの匿名化，学習管理システム（LMS），教育データ解析，差分プライバシー，ラプラスメカニズム

Study on Reasonable Parameters of Anonymization Algorithms toward Development of Assistant Systems of Online Educational Data Analysis

Abstract: We propose a criteria of reasonable parameters for an algorithm that aggregates target data with anonymizing the data for safe analysis of data on online educational systems, which are called learning management system (LMS). We also investigate a set of the parameters that satisfy the proposed criteria by statistical processing with an anonymization algorithm and large scale real data. Furthermore, we discuss an approach to address the case in which we can set no parameter that satisfies the proposed criteria with usual way to set parameters. Since we couldn't obtain real LMS data for this investigation, we use real review data of commodities.

Keywords: anonymization, learning management system (LMS), educational data analysis, differential privacy, Laplace mechanism

1. はじめに

新型コロナウィルスの全世界的な蔓延により，学校教育のオンライン化が急速に普及しつつある．それに伴い，インターネット上のシステムを用いた教育活動が欠かせないものとなりつつある．知識や技術の習得には，それを身につけるための演習や習得度を計るための試験をこまめに行うことが重要だが，情報システムは演習や試験を行うことの手間を削減し，かつ，演習や試験のやり方の自由度を高めている．そのため，新型コロナウィルスの問題が解消されたとしても，教育活動や学習活動における情報システムの利用は，今後ますます普及するものと思われる．本原稿では，このような情報システムを学習管理システム（LMS）と呼ぶ．

本研究では，LMS 上のデータ解析を安全に行うための手法として，データの匿名化の一種である「データの攪乱」（つまり，ノイズの付加）を伴うデータの統計処理アルゴリズムに注目し，その匿名化（攪乱）の度合いに関する設定値を分かりやすい形で定めるための基準について議論する．

[1] 群馬大学
Gunma University, Gunma, Maebashi 371–8510fs, Japan
[2] 国立情報学研究所
National Institute of Informatics

1.1 本研究の背景

学習管理システムを有効に利用するために，システムのユーザによる利用履歴データを分析し，教育活動に活かす取組がなされている [9], [10], [4], [8]．業務システムなどの情報システムは，そのシステム上のデータを解析することにより顧客の動向を把握し，業務の改善や新たなビジネスの開発等を行ってきた．同様に，今後，LMS が教育現場に普及することを考えると，そのデータの分析に基づく教育改善も盛んに行われることが期待できる．一方，LMS 上のデータ分析に基づく教育改善が盛んになれば，データを利用・共有・流通する規模が拡大し，それに伴って，データの情報流出や悪用など，利用上のリスクも高まっていくことが予想される．さらに，LMS 上のデータの利用が大規模化すれば，そのステークホルダーや価値も大規模化・多様化していくため，データの情報提供者（LMS の利用者）を始めとするステークホルダーに対して，データの管理・運用における安全性を確保しつつ，ステークホルダーからの同意や協力を得る必要がある．データの匿名化は，データの暗号化などの他の情報セキュリティの問題と比較すると，その目的や効果を説明することが難しい．この点でも，データを匿名化する意義を，データの提供者や利用者に対して分かりやすく説明することは重要である．よって，LMS 上のデータ分析を安全性を保ちつつ社会全体に普及させるためには，データの情報提供者のプライバシー保全を始めとした安全性を保証するだけでなく，ステークホルダーへの説明責任を果たしていくことが重要であると考える．

本論文の著者達は，大学内の LMS のデータおよび LMS が稼働している計算機のログデータをクラウドシステムに集積させ，集積されたデータを包括的に分析するための基盤環境の開発を行っているこの基盤環境の開発において，データを以下に安全に管理・利用するかは重要な課題である．浜元ら [10] は，データをクラウドシステムに集積させる際の安全性を確保するために，データの匿名化や仮名化を行うための具体的な方法論について議論している．クラウドシステム上のデータの分析を行う際に，複数の組織にまたがる教育データの利用やその知識の共有がなされることになるため，利用者が扱うデータの匿名化は重要な課題である．本論文の著者達は最初の段階では k-匿名性に基づくデータの非特定化（曖昧化，仮名化，切り落とし）を施したデータを利用者間で共有する機能について吟味していたが，時系列データの場合，データの非特定化のみでは k-匿名性と十分な有用性を両立させたデータを得ることは困難であることが分かってきた．そのため，分析者に匿名化された LMS 関連のデータを提供する代わりに，クラウドシステム上で LMS 関連のデータを直接確認することなく統計データを得るための LMS 上のデータ分析支援システムの開発を目指している．LMS 上のデータの代わりに統計データを提供することは，分析者のデータの扱いに関する自由度を制限することになるが，プライバシー保護の観点での安全性がより高まることが期待される．しかしながら，統計データを提供するという状況においては，プライバシー保護の観点での安全性を高めつつ，k-匿名化を始めとする特定可能性や識別可能性のような従来の安全性基準では評価・保証することが難しい．

上記の状況において，非特定化とは別の匿名化手法であるデータの攪乱，つまり，統計データの生成過程において，データにノイズを付加する手法は有用である．また，攪乱手法に関するよく知られた安全性基準として差分プライバシー [1] が挙げられる．データの攪乱を伴う統計処理アルゴリズム（このような確率アルゴリズムは，差分プライバシーの理論においては「メカニズム」と呼ばれる）は，多くの場合，攪乱の度合いを決定するパラメータを持つ．差分プライバシーは，与えられたメカニズムに対し，そのパラメータに基づいて，メカニズムの安全性を定量化することを可能にしている．しかしながら，その値の解釈は，従来の安全性の指標である k-匿名性などの基準に基づくものと比較して，容易とは言えない．そのため，当該の統計処理に対して，適切なパラメータを設定することは自明なことではない．

1.2 本原稿の目的および構成

本原稿の目的は，データの攪乱を伴う統計処理アルゴリズムに対して，攪乱の度合いに関するパラメータの基準を提案することである．本原稿で提案する基準は，LMS 上のデータ解析という特定の状況を踏まえたものであるため，匿名化パラメータの基準に関する提案内容の妥当性の説明に際しては，教育システム上のデータの利用という観点からの説明を試みる．第 3 節において，匿名化パラメータ基準の適用例として，統計処理アルゴリズムとしてラプラスメカニズムに注目し，ラプラスメカニズムのパラメータをどうやって決めるのかを，実際のデータを用いた実験を交えつつ説明する．第 4 節において本稿のまとめを述べる．

2. パラメータの設定基準の提案

この節において，匿名化パラメータ基準について説明する．まず，本稿の著者達が考える LMS 上のデータ解析における分析対象あるいは分析結果の多くが，1 次元あるいは多次元の度数分布表あるいはそれらをヒストグラムなどのグラフで視覚化したものであることを踏まえ，本稿では度数分布表に焦点を当てることにする．さらに，匿名化アルゴリズムとして，個票データに対してデータにノイズを与えつつ度数分布表を生成するアルゴリズムを想定する．上記の想定に加えて，上記の匿名化アルゴリズム $\mathcal{A}$ は，匿名化の度合い，つまり，ノイズの大きさに関するパラーメータ p を持つものとする．

次に，個票データのような元のデータ D および匿名化パラーメータ p を持つ匿名化アルゴリズム $\mathcal{A}_p$ が与えられているとして，$\mathcal{A}_p$ の匿名化パラーメータ p に対して以下の性質を定義する．

D に対してノイズを加えずに得られる（本来の）度数分布表 F，および，D に対してノイズを加えつつ得られる度数分布表 $\mathcal{F}(:=\mathcal{A}_p(D))$ を考える．このとき，数値 $0 \leq \alpha, \beta \leq 1$ について，p が以下の条件 (1) および (2) を満たすとき，p は (α, β)-基準を満たすと言う．

(1) F において度数が最大となる任意の階級 $C_{\max}$ に対して，$C_{\max}$ が $\mathcal{F}$ においては度数最大の階級でない確率は，α 以下である．

(2) F において度数が最小となる任意の階級 $C_{\min}$ に対して，$C_{\min}$ が $\mathcal{F}$ においては度数最小の階級でない確率は，β 以上である．

ここで，本論における提案として，匿名化のパラメータ p として，$(0.05, 0.05)$-基準を満たすもの（つまり，$\alpha = \beta = 0.05$ を満たすもの）を提案したい．以下において，この提案の根拠について説明する．

まず，パラメータの基準設定において，度数最大および度数最小の階級に注目した理由について述べる．LMS 上のデータなどの教育データの分析において，度数最大の階級に関する知見は，最も多数の学生の傾向を示す場合が多く，このような知見は，研究成果としても，最も頻繁に他者と共有される．一方，度数最小の階級に関する知見は，一部の例外的な学生の傾向に相当することが多く，そのような知見は当該学生の管理や指導のような直接の教育業務に係るものであり，研究などの二次的な理由で広く共有されるような知見ではない場合が多い．さらに，度数最小の階級に関する知見は，データの特定や識別に結びつきやすいという傾向を持つ．よって，情報の匿名化が必要となるような状況，つまり，直接の教育業務とは異なる理由により，他者と知識を共有するような状況においては，度数最小の階級に関する知見は，あまり正確に把握されないことが望ましいと考える．

次に，$\alpha = 0.05$ とする理由は，α が，「$C_{\max}$ は F における度数最大の階級である」という仮定を考えたとき，あたかもそれを統計的検定における帰無仮説とし，α をこの仮説に対する第 1 種の過誤の確率，あるいは，ある種の有意水準と見做した場合，$\alpha = 0.05$ という設定は妥当なものとして広く受け入れられやすいと考えたためである．さらに，$\beta = \alpha = 0.05$ とする理由は，攻撃者が度数最小の階級として $C_{\min}$ に注目していることを想定し，分析者による $C_{\max}$ に対する主張と同じ有意水準による主張を行うことを否定することを意図しているためである．

3. 提案基準の適用実験

前節において，LMS 上のデータなどの教育データに対して匿名化アルゴリズムを適用する際のパラメータを設定するために，(α, β)-基準という概念を定義し，$(0.05, 0.05)$-基準を満たすパラメータを妥当な匿名化のパラメータとすることを提案した．ところで，$(0.05, 0.05)$-基準を満たすパラメータの有無，あるいは，そのようなパラメータの集合 $\mathcal{P}$ がどのようなものになるのかは，処理の対象となるデータ D や匿名化アルゴリズム $\mathcal{A}$ に依存する．現時点では $\mathcal{P}$ に関する理論の構築は未だ出来ていないため，$\mathcal{P}$ がどのようなものになるのかや，$\mathcal{P}$ が空集合になる場合等の対処の仕方について，実在のデータを使用して具体的に例示する．

3.1 実験に用いるデータおよび匿名化アルゴリズム

本原稿の主旨を踏まえると，本実験で LMS 上のデータを用いるのが望ましい．しかしながら，実在の LMS 関連のデータで一定の規模を持つものを本研究に使用することは，手続き上困難であった．そこで，LMS 関連のデータの代用として，国立情報学研究所の情報研究データリポジトリ [6] において，楽天グループ株式会社が提供している，商品レビューデータ [5] の内の 1 年分 (約 1500 万件）を用いる．この商品レビューデータには商品や評価内容等の情報が記録されているが，今回の実験では商品レビューの 1 時間毎の分布状態に着目した度数分布表を生成することに焦点を当て，レビューデータの登録年月日および時刻のみを用いることとし，1 日ごとに 1 つのデータセットでまとめられているものとする．（以下，このデータセットを「商品レビュー時刻データ」と呼ぶ．）

一方，匿名化アルゴリズムとしては，ラプラス分布に従ってランダムに生成されるノイズを付加しながら，商品レビュー時刻データ D から，1 時間ごとのレビューの件数からなる度数分布表を作成するアルゴリズム（差分プライバシーにおける，ラプラスメカニズムに基づく度数分布表生成アルゴリズム）を考える．簡単のため，このアルゴリズムも「匿名化アルゴリズム」と呼ぶことにする．差分プライバシーの理論においては，ラプラスメカニズムは最もよく知られたメカニズムの一つであり，ラプラスメカニズムにおける匿名性の度合い（ノイズの強さ）を決めるパラメータは，問い合わせ（統計処理）の種類によって定められる敏感度 S と差分プライバシー理論における安全性の指標であるイプシロンパラメータ ϵ との比 S/ϵ によって表現される．実際，（平均が 0 の）ラプラス分布におけるスケールパラメータとして，固定された統計処理 Q に関する敏感度 S_Q およびイプシロンパラメータ ϵ による比 S_Q/ϵ を取った場合の（Q を組み込んだ）ラプラスメカニズムは ϵ-差分プライバシーを満たす [2]．一方，度数分布表の作成，つまり，集計処理に対して，拘束なし隣接条件を仮定する場合，つまり，ある階級の度数を 1 つ増加（減少）させるごとに，別の階級の度数を 1 つ減少（増加）させるという条件を考えない場合は，集計処理に対する敏感度は 1 にな

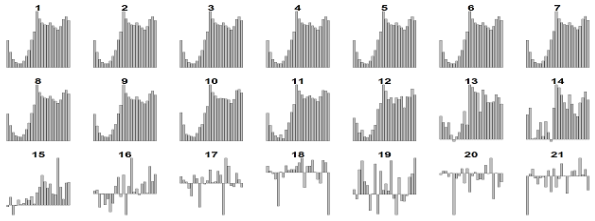

図 1 ある一つの対象データに対して，異なる強さの匿名化の設定値に基づいて生成されたノイズ入り度数分布表（ヒストグラムとして表現）

る [7]．よって，本節の実験では，敏感度を 1 として，イプシロンパラメータの逆数 $1/\epsilon$ のみを匿名化パラメータ p として考え，議論を進める．なお，上記のアルゴリズムの実装に際して，Google LLC の“Google's Defferntial Privacy Package”[3] を利用する．

3.2 実験結果

まず，365 日分の商品レビュー時刻データ D_i（$i = 1, ..., 365$）の各 D_i に対して度数分布表を生成する匿名化アルゴリズム $\mathcal{A}$ を用いて，異なる強さのノイズを加えた度数分布表（以下，「ノイズ入り度数分布表」と呼ぶ）を作成した．その際，各生成対象データに対して，匿名化パラメータを，$1/(4 \times \log 3)$ を初期値とし，順次倍増させていく．その結果の例として，ある日のデータ D_i に対するノイズ入り度数分布表を図 1 に例示する．

各ヒストグラムは，ある 1 日の商品レビューの登録件数を 1 時間毎に集計した数を表す．ここでは，度数分布表とそれをグラフ化したヒストグラムをしばしば同じものとして扱う．図 1 の 1 番目（左上）のヒストグラムは，ノイズを全く入れずに生成したヒストグラムを表す．その後，匿名化パラメータ p を倍増させながらノイズを加えつつ生成したヒストグラムを横並びに並べている．匿名化パラメータを指数関数的に増加させているため，14 番目のヒストグラム（上から 2 段目の一番右）を境にして，形を大きく崩していっていることが分かる．なお，簡単のため，今回の実験では，ノイズの付与に際して度数の非負制約を課していない．今後，商品レビュー時刻データ D_i に対して，ノイズを加えずに作成した度数分布表（=ヒストグラム）を F_i^0 で表し，匿名化パラメータ p によるノイズ入り度数分布表を F_i^p で表す．

次に，365 日分の各データ D_i および $\mathcal{A}$ に対して，$(0.05, 0.05)$-基準を満たすパラメータの集まりがどのようなものになるのかを確認するために，各データ D_i およびパラメータ p に対して，$F_i^p := \mathcal{A}_p(D_i)$ を 1000 回生成し，F_i^p における度数最大の階級（時間帯）が F_i^0 における度数最大の時間帯と異なる割合を計算した．その結果の内の最初の一か月分を図 4 に示す．

図 4 の中の各グラフは，1 つの商品レビュー時刻データ D_i に対して，匿名化パラメータ p を指数関数的に増加させていったとき，ノイズ入り度数分布表 F_i^p における度数最大の時間帯がノイズ無しの度数分布表 F_i^0 における度数最大の時間帯と異なる割合を黒の折れ線で示し，ノイズ入り度数分布表 F_i^p における度数最小の時間帯がノイズ無しの度数分布表 F_i^0 における度数最小の時間帯と異なる割合を赤の折れ線で示したものである．これらの異なる割合を，それぞれ，度数最大の時間帯および度数最小の時間帯に対する推定失敗率と呼ぶことにする．グラフの中の，緑の横線は失敗率の閾値 0.05 を表し，黒の縦線（破線）は度数最大の時間帯に対する失敗率が 5%を超えないパラメータの最大値（指数値）を表し，赤の縦線（点線）は度数最小の時間帯に対する失敗率が 5%未満にならないパラメータの最小値（指数値）を表す．(0.05,0.05)-基準を満たすパラメーターは，度数最大の時間帯に対する失敗率が 5%以下で，かつ，度数最小の時間帯に対する失敗率が 5%以上となることを満たす．よって，各グラフにおいて，赤の点線が黒の破線よりも左側にある，あるいは，一致しているとき，赤の点線と黒の破線の間の区間が (0.05,0.05)-基準を満たすパラメータの集まりということになる．（赤と黒の線が重なっているときは一点集合となる．）この黒の破線の位置，つまり，各データ D_i に対して，ノイズ無しの度数分布表 F_i^0 における度数最大の階級（今回の例では時間帯）に関する推定失敗率が 5%を超えないパラメータの最大値を「D_i あるいは F_i^0 に対する (0.05,0.05)-基準の上限」と呼び，$p_i^{\max}$ で記す．さらに，F_i^0 における度数最小の階級に関する推定失敗率が 5%以上となるパラメータの最小値を「D_i あるいは F_i^0 に対する (0.05,0.05)-基準の下限」と呼び，$p_i^{\min}$ で記す．

3.3 (0.05, 0.05)-基準を満たすパラメータの集合のサイズ

本稿の主張に基づいて (0.05,0.05)-基準を満たすパラメータを適度な設定値と見做すという立場を取るならば，先の部分節で述べた (0.05,0.05)-基準を満たすパラメータの集合 $\mathcal{P}$ が小さいほど，パラメータを自然に設定することが出来る．しかし，$\mathcal{P}$ が空集合になってしまうような場合，例えば，図 4 において，赤の点線が黒の破線よりも右側にできてしまうような場合，何らかの対処が必要になる．そこで，次の節において，$\mathcal{P} = \emptyset$ となる場合の対処法について述べる．しかしその前に，$\mathcal{P} = \emptyset$ となる場合がどのくらいあるのかを検討するために，1 年分の商品レビュー時刻データ $\{D_i | i \leq 365\}$ に対して，各 D_i に対する (0.05, 0.05)-基準の上限と下限の差，つまり，$p_i^{\max} - p_i^{\min}$ の大きさに関するヒストグラムを図 3 に示す．

図 3 の横軸の値（$-10 \sim 10$）は，$p_i^{\max} - p_i^{\min}$ を表す．前節で述べた通り，この値が 0 のときは，(0.05, 0.05)-基準を満たすパラメータが（今回の実験におけるパラメータ

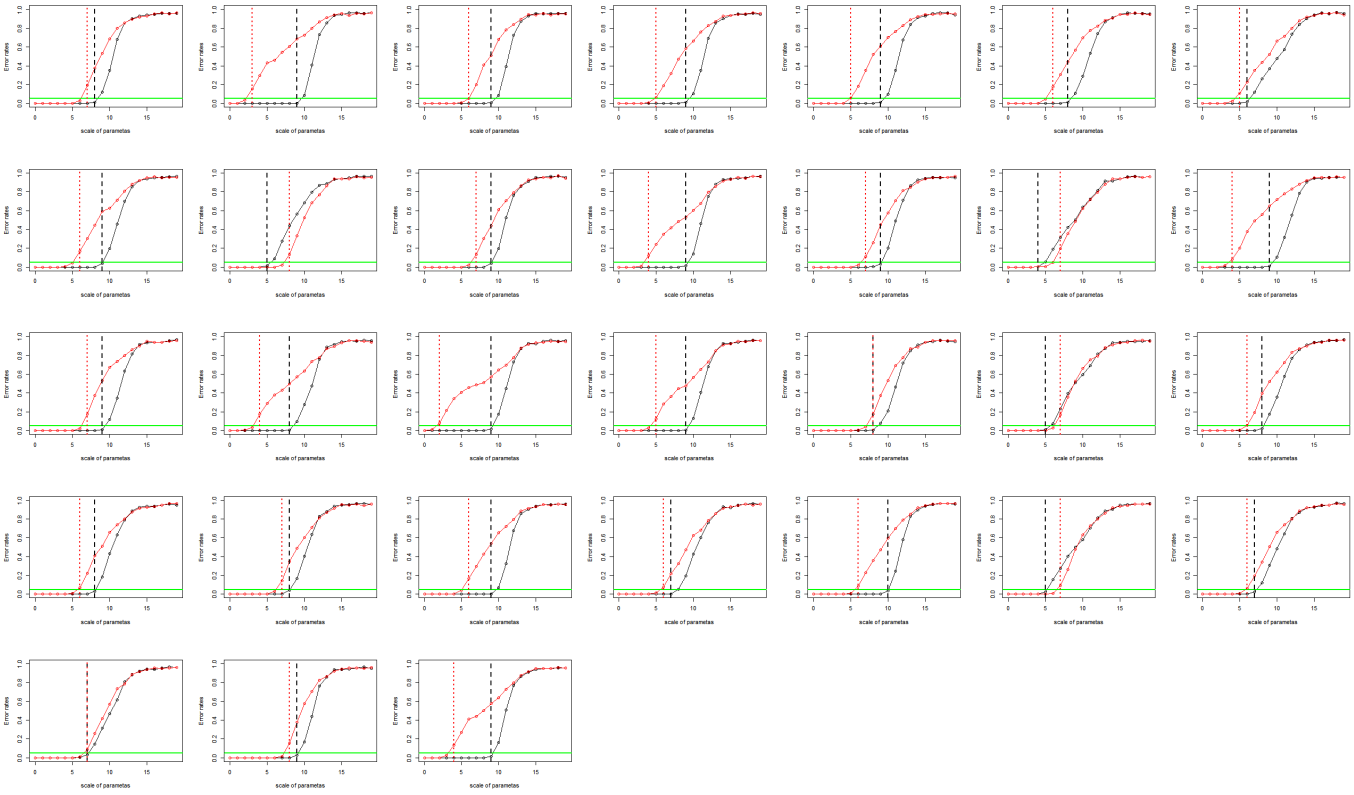

図 2 各商品レビュー時刻データ D_i に対して，それぞれ 20 段階で匿名化を強めていった際の，度数最大の時間帯が F_i^0 のものと異なる割合（黒の折れ線）と，度数最小の時間帯が F_i^0 のものと異なる割合（赤の折れ線）を示したグラフを 1 か月分並べた図式

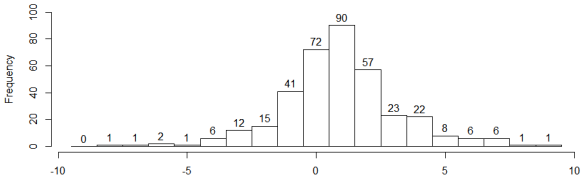

図 3 $p_i^{\max} - p_i^{\min}$ の大きさに関するヒストグラム

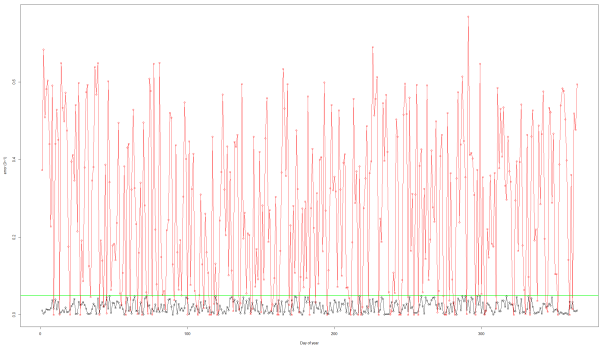

図 4 度数最大の時間帯の推定失敗率が 5%以上となる最小のパラメータによってノイズ入りのヒストグラムを作成した場合の度数最大の時間帯の推定失敗率（黒の折れ線グラフ）および度数最小の時間帯の推定失敗率（赤の折れ線グラフ）

の設定の仕方に関して）ただ一つに定まっていることを意味し，負のときは，（今回の実験におけるパラメータの設定の仕方に関して）設定値が無いことを意味する．図 3 によると，今回の 365 件のケースでは，219 件（60%）のケースでパラメータが 1~3 個あり，その内の 72 件（20%）のケースでパラメータが一つに決まる一方で，76 件（21.6%）のケースでパラメータが一つも無い状況になっている．

この点をより詳しく見るために，1 年分の商品レビュー時刻データセット $\{D_i | i = 0, ..., 365\}$ の各 D_i に対して，$p_i^{\min}$ によってノイズ入りのヒストグラムを作成した場合の，度数最大および度数最小の時間帯の推定失敗率をグラフ化したものを図 4 に示す．図 4 において，度数最大の時間帯の推定失敗率が黒の折れ線グラフによって，度数最小の時間帯の推定失敗率が赤の折れ線グラフによって表されている．緑の横線は 5%を示している．約 8 割のケースにおいて，赤の折れ線が黒の折れ線を上回っているが，やはり，21.6%のケースにおいて，赤の折れ線が黒の折れ線を下回っている．

3.4 異なる強さのパラメータを局所多段的に適用することによる対処法

(0.05, 0.05)-基準を満たすパラメータが一つもない場合の対処法として，1 つの対象データに対して 1 つの設定値

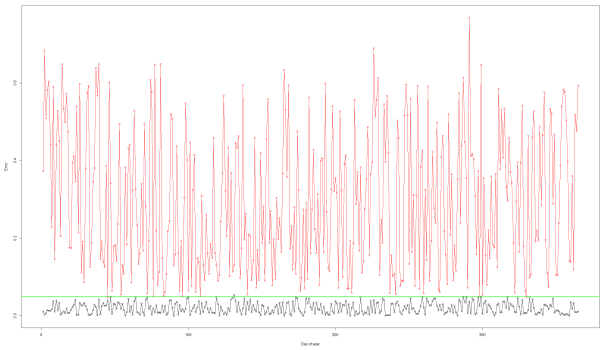

図 5 (0.05, 0.05)-基準を満たすパラメータが一つも無かった D_i に対して局所多段的な設定値の適用を行った後で，図 4 と同様のグラフを生成した結果

によって匿名化度数分布表を生成する代わりに，特定の階級に対して異なる強さを持ったパラメータに基づく匿名化を施す方法を以下に提案する．この手法を大まかに述べると，匿名化の必要性が比較的低い，度数の比較的大きな階級に対しては弱いノイズを加え，匿名化の必要性が比較的高い，度数の比較的小さな階級に対しては強いノイズを加えることである．より具体的には，

(1) （ノイズ無しの度数分布表における）度数最大の階級（に含まれるデータ）には (0.05, 0.05)-基準の上限 $p_{\max}$ でアルゴリズムを適用する．
(2) 度数最小の階級には (0.05, 0.05)-基準の下限 $p_{\min}$ でアルゴリズムを適用する．
(3) それ以外の階級については，ノイズ無しの度数分布表における度数による順序に応じて，$p_{\max}$ 以上かつ $p_{\min}$ 以下の設定値をシグモイド曲線による重みづけに基づいて与える．

このような手法を，「匿名化アルゴリズムに対する局所多段的な設定方法」と呼ぶことにする．(0.05, 0.05)-基準を満たすパラメータが一つもない場合では，$p_{\max} < p_{\min}$ が成り立っている点に注意されたい．

今回の実験において，(0.05, 0.05)-基準を満たすパラメータが一つも無かった 76 件の D_i に対して，局所多段的な設定値の適用を行ったところ，76 件の D_i の殆どすべてのケースにおいて，(0.05, 0.05)-基準を満たすパラメータを持たせることができるようになった（図 5）．

3.5 (0.05,0.05)-基準を満たす設定値による匿名化

前節において，365 日分の商品レビュー時刻データ D_i ($i = 1, ..., 365$) に対する (0.05,0.05)-基準を満たすパラメータの集合 $\mathcal{P}_i$ を考え，$\mathcal{P}_i = \emptyset$ となるような D_i に対して，局所多段的な設定により (0.05,0.05)-基準を満たすノイズ入り度数分布表を生成した．ここで，各 $\mathcal{P}_i$ について，$\mathcal{P}_i = \emptyset$ のときは，局所多段的な設定により生成されるノイズ入り度数分布表を，$\mathcal{P}_i \neq \emptyset$ のときは，以下の方針で選択されるノイズ入り度数分布表を考える．

- 正確性を重視した度数分布表: $\mathcal{P}_i$ の内で最小のパラメータを用いて生成される，つまり，最も弱いノイズの入った度数分布表
- 安全性を重視した度数分布表: $\mathcal{P}_i$ の内で最大のパラメータを用いて生成される，つまり，最もの強いノイズの入った度数分布表

ここで，今回の商品レビューデータの内のある 30 日分のデータ D_i ($i = 1, ..., 31$) に対する，正確性を重視した度数分布表の集合と，安全性を重視した度数分布表の集合を，それぞれ図 6 および図 7 によって示す．

図 6 および図 7 において，灰色のヒストグラムはノイズの無い度数分布表を表し，水色のヒストグラムと緑色のヒストグラムはそれぞれ左隣のヒストグラムと同じデータによって生成されたノイズ入りの度数分布表を表す．図 6 の図 7 におけるノイズの無い度数分布表は，それぞれ同じものである．

水色のヒストグラムも緑色のヒストグラムも，(0.05,0.05)-基準を満たすパラメータによって生成されている．しかし，水色のヒストグラムは度数最小の階級（時間帯）に関する推定失敗率が 5%以上になるものの内で最もノイズが弱くなるパラメータであるため，どのヒストグラムも，度数最小の部分を含め，殆ど同じに見える．一方，緑色のヒストグラムは度数最大の時間帯に関する推定失敗率が 5%以下になるものの内で最もノイズが強くなるパラメータであるため，（図の画質を落としているため分かりにくいかもしれないが）水色のヒストグラムと比較すると，形状の違いが比較的大きい．実際，度数最小の時間帯を比較すると，水色のヒストグラムの内で度数最小の時間帯が変化したものは 30 件中 1 件であるのに対して，緑色のヒストグラムの内で度数最小の時間帯が変化したものは 30 件中 9 件であった．一方，度数最大の時間帯に関しては，どちらの場合も推定失敗率が 5%以下であり，実際，ノイズの無いヒストグラムと比較して時間帯が変化した件数は，どちらの場合も 0 件であった．（今回の実験では，パラメータを指数関数的に増加させながら決めていたので，度数最大の時間帯に関する推定失敗率が 5%を大きく下回り得る点に注意されたい．）

少し一般的な話になるが，正確性と安全性のどちらを軸にしてパラメータを選択するか？については，データの管理者が，自分達が所有するデータを匿名化する際の目的によって決める，という方針が挙げられる．一般に，データの匿名化の目的として，以下のものが挙げられる．

(1) データ利用の目的が明確であり，その目的を果たすことは出来るが，それ以外の目的の達成が阻害されるようなデータにするための匿名化
(2) データ利用の目的は明確でないが，避けるべきこと（リスク）が想定されており，それらのリスクが起きる可能性を低減させるための匿名化

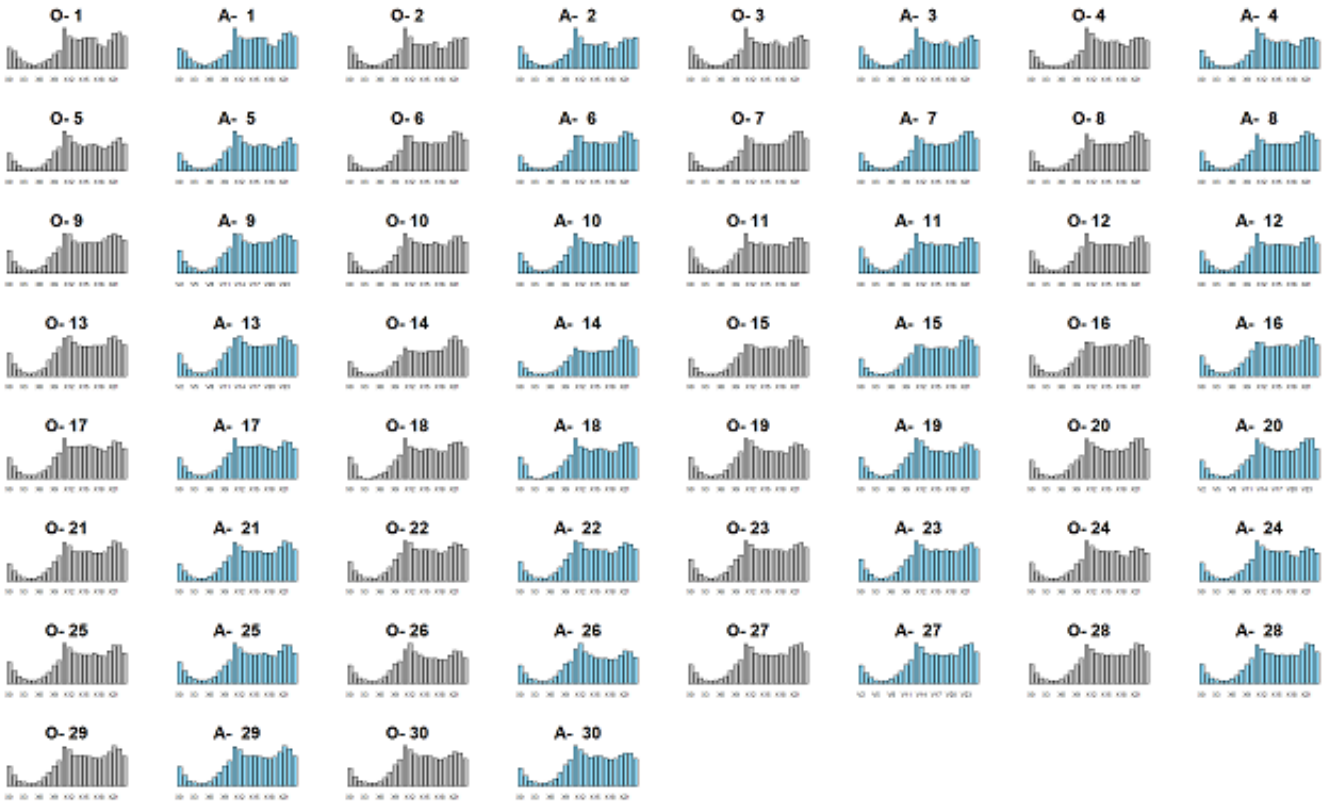

図 6 ノイズ無しの度数分布表（灰色）および正確性を重視したノイズ入り度数分布表（水色）

図 7 ノイズ無しの度数分布表（灰色）および安全性を重視したノイズ入り度数分布表（緑色）

一般に，データ利用の目的と比較して，データを利用することや他者とデータ分析の結果を共有する際のリスクは明確ではない．よって，目的が明確ならば，その目的を達成するための必要最小限のデータを提供する，という意味で，

安全性を重視したパラメータを選択するのが自然であると思われる．一方，避けるべきリスクが明確に定められているような場合では，そのようなリスクを十分に避けられるような範囲で，最も便宜性の高いデータを提供するという方針も考えられる．そのような場合では，正確性を重視したパラメータの設定の仕方が適用できると考えれる．

4. おわりに

本研究では，オンラインによる教育活動を支援する学習管理システム（LMS）のデータ解析を安全に行うことを目的として，ノイズを加えつつ統計処理を行うアルゴリズムに着目し，その設定値の妥当性を吟味するための基準として，(0.05, 0.05)-基準を提案した．また，(0.05, 0.05)-基準を満たすパラメータの集合がどうなるのかを検討するために，実在の大規模データおよびラプラスメカニズムをベースとした度数分布表生成プログラムを用いた実験を行った．さらに，(0.05, 0.05)-基準を満たすパラメータが一つも無い場合の対処法として，局所多段的なパラメータの設定方法を提案した．

今回の分析結果や提案方法については，理論的な基盤が構築されているわけではなく，あくまでも，一連の実在データを用いて検討した結果に過ぎない．そのため，(0.05, 0.05)-基準そのものの妥当性や，(0.05, 0.05)-基準を満たす設定値が一つも無い場合に対する理論的に保証された手法の構築は，本研究における重要な課題である．

また，(0.05, 0.05)-基準を満たすパラメータの計算や，局所多段的なパラメータの設定は，度数最大あるいは度数最小の階級の推定失敗率を計算しながら実行しており，本研究ではこれらの作業の効率性を考慮しなかった．そのため，これらの作業を効率的に行うためのアルゴリズムの開発も望まれる．

謝辞 本研究では，国立情報学研究所の情報学研究データリポジトリにより，楽天グループ株式会社から提供された「楽天データセット」（`https://rit.rakuten.com/data_release/`）を利用した．ここに記して感謝する．本研究は，2021 年度国立情報学研究所公募型共同研究（21FA05）の助成を受けて行われた．

参考文献

[1] Dwork, C.: Differential Privacy, *Automata, Languages and Programming* (Bugliesi, M., Preneel, B., Sassone, V. and Wegener, I., eds.), Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 1–12 (2006).
[2] Dwork, C. and Roth, A.: The Algorithmic Foundations of Differential Privacy., *Foundations and Trends in Theoretical Computer Science*, Vol. 9, No. 3-4, pp. 211–407 (2014).
[3] LLC, G.: Google's Differential Privacy Libraries, Google LLC (online), available from ⟨https://github.com/google/differential-privacy⟩ (accessed 2022-02-14).
[4] 島田敬士：大学教育における学習分析の活用事例，情報処理学会論文誌教育とコンピュータ（TCE），Vol. 6, No. 2, pp. 16–24 (2020).
[5] 国立情報学研究所：楽天データセット，楽天グループ株式会社（オンライン），DOI: info:doi/10.32130/idr.2.1（参照 2022-02-14）.
[6] 国立情報学研究所：情報学研究データリポジトリ，国立情報学研究所（オンライン），入手先 ⟨https://www.nii.ac.jp/dsc/idr/⟩（参照 2022-02-14）.
[7] 寺田雅之，山口高康，本郷節之：匿名化個票開示への差分プライバシの適用，情報処理学会論文誌，Vol. 58, No. 9, pp. 1483–1500（オンライン），入手先 ⟨https://ci.nii.ac.jp/naid/170000148916/⟩ (2017).
[8] 大西淑雅，山口真之介，近藤秀樹，西野和典：ネットワークログを用いた学習活動の把握の提案，技術報告 1，九州工業大学学習教育センター, 九州工業大学学習教育センター, 九州工業大学学習教育センター, 九州工業大学教養教育院 (2019).
[9] 浜元信州，横山重俊，竹房あつ子，合田憲人：端末特定のためのログ解析クラウド環境の構築，技術報告 29，群馬大学総合情報メディアセンター, 群馬大学総合情報メディアセンター／国立情報学研究所, 国立情報学研究所, 国立情報学研究所 (2019).
[10] 浜元信州，横山重俊，竹房あつ子，合田憲人：クラウドを利用した Moodle のログ解析環境の実装，技術報告 23，群馬大学総合情報メディアセンター, 群馬大学総合情報メディアセンター／国立情報学研究所, 国立情報学研究所, 国立情報学研究所 (2021).