
ブートストラップサンプリング法とカルバック・ライブラー情報量を用いたクリーンルーム環境の類似性評価

Similarity assessment of cleanroom environments using the Bootstrap sampling method and Kullback–Leibler divergence

中澤 昌高（株式会社 八光）

Masataka Nakazawa (Hakko CO.,ltd.)

Abstract

本研究では、相関分析が困難なデータの類似性評価をするために、ブートストラップサンプリング法とカルバック・ライブラー情報量を組み合わせた新手法を提案する。まず、源データから復元抽出により多数のサンプル平均を生成し、それらを正規分布とみなす。次に、分布間距離をカルバック・ライブラー情報量 (KL) で評価し、KL 値の小さいペアを高い類似度、大きいペアを低い類似度と見なす。このアプローチにより、データ間の分布構造そのものを定量的に比較でき、相関分析よりも柔軟な評価が可能となる。クリーンルーム環境で取得した実データを対象に、カルバック・ライブラー情報量を計算し、クリーンルーム環境が「構造的に似ている」「似ていない」を階層的に識別できることを確認した。

In this study, we propose a new method that combines Bootstrap sampling and the Kullback–Leibler divergence to assess similarity between datasets for which correlation analysis is difficult. First, we generate many sample means from the original data by resampling with replacement and treat those means as normal distributions. Next, we evaluate distances between distributions using the Kullback–Leibler divergence, regarding pairs with small KL values as highly similar and pairs with large KL values as dissimilar. This approach allows quantitative comparison of the distributional structure between datasets and provides a more flexible assessment than correlation analysis. Using real data collected in a cleanroom environment, we computed Kullback–Leibler divergences and confirmed that the method can hierarchically discriminate whether cleanroom environments are “structurally similar” or “not similar.”

1. はじめに

同一の計測手法によって複数の異なる場所から、統計的独立に収集がされたデータがあるものとする。これらの複数の異なる場所にかんして、何かしらの傾向や特徴を評価したい場面があるとする。このような場合、複数の異なる場所の線形関係を相関分析により評価することもデータアプローチとしては関心が高いものと考えられる[1]。相関係数が 0 に近いときには相関がほとんどないというような判断となるが、線形の関係性がほとんどないというだけであるため、出力が

なされた結果に関する解釈には注意が必要であるうえに、複数の分析の対象となっているデータの収集は同じタイミングで行われる必要がある。

相関係数による評価方法はコンピュータソフトウェアの充実もあり扱いが簡便であるが、データの収集のタイミングを合わせなくてはならないなどの制約がある。そして、いうまでもなく、ある程度の量のデータ数が必要な事や十分にデータ数が揃えられないような場面では、外れ値に相当するような数値の影響を受ける可能性が高いことを考慮する必要がある。

本研究では、サンプリングされたデータ数が少ない場合や、比較したいデータ間に線形関係があるかが不明な場合にあっても、ブートストラップサンプリング法により源データから復元抽出を多数回繰り返して平均値を求めたうえで、平均値の確率密度関数が正規分布に収束するという性質を利用して、比較する対象のデータを正規分布に収束させたのちに、カルバック・ライブラー報量により分布間距離を評価した。カルバック・ライブラー情報量が小さければ類似度が高く、逆に高い場合には類似度が低いものと判断することができる。すなわち、もとのデータの構造に依存することなく正規分布の分布間距離によりデータの持つ類似性を評価するといった新しい方法を提案している。

本稿で示しているデータは、筆者の所属している組織の医療機器を製造しているクリーンルームにて収集がなされた粗大塵埃濃度のデータである。なお、クリーンルームや粗大塵埃といった言葉に聞きなじみのない読者に向けて、簡単な解説を付録に示した。必要に応じて参考にさせていただきたい。

2. 概況

まずデータのサンプリング方法について説明する。クリーンルームの粗大塵埃濃度の計測場所は、筆者の所属している組織が定めている手順に従って、クリーンルームの中で粗大塵埃濃度が最大の場所を計測点と定める。計測機器はパーティクルカウンター（日本カノマックス株式会社 KC-31）を使い、1分間に 28.3L の空気を吸引してその累積値を塵埃濃度とする。単位は個/ft³（ft³は立方フィートと読む）である。計測の頻度は毎月1回、月初に行うが、計測をする時刻はランダムである。本研究に用いたデータは、2021年1月から2025年3月までの合計51個である。計測をしているクリーンルームは全部で7部屋あることから、今回の研究では全7部屋を対象にしている。クリーンルームの部屋番号

は本稿では x1 から x7 とラベルをして、本文中の図表では ID として示している。

次にデータの要約統計量を示す。

x1 から x7 までの基本統計量を Table.1 に示し、ヒストグラムを Fig.1 に示した。そして、Table.2 には相関係数行列を示した。

Table.1 には ID、サンプリングしたデータ数を count と示した。母標準偏差は std、平均値は mean と示した。分位点は最小値、下側四分位、中央値、上側四分位および最大値をそれぞれ、min、25%、50%、75% および max として示した。

母標準偏差を母分散に置き換えてみると、いずれの ID も平均値よりも高いことがわかる。すなわち、過分散の状況にあることが読み取れた。この影響は平均値と中央値を比較すると一目瞭然であり、大きな値の影響がデータのひずみを大きくさせていることが読み取れる。

次にヒストグラムは左上から右に向かって ID が x1 から x7 の順に並んでいる。データ数が 51 個と少ないことも影響しているために、横軸の粗大塵埃濃度の階級と縦軸の頻度に対してはヒストグラムに隙間がみられることや形状が滑らかではない状況にあり、このヒストグラムの形状からは特定の確率分布関数を推定することは困難な状況にあると考えられた。

相関係数行列を見ると、最も高い相関係数でも ID が x4 と x7 の 0.29 にとどまっており、各クリーンルームの相関係数は低い状況にあった。これらの結果から次のことが読み取れる。まず過分散の状況にあることから、データには複数の外れ値が存在している可能性がある。次にデータ数が少ないため外れ値の影響を受けている可能性がある。最後に各データには明瞭な線形関係が認められない。したがって相関分析による評価は難しいと考えられる状況にあるといえた。

Table.1 Descriptive value for cleanrooms

ID	count	mean	std	Percentail				
				min	25%	50%	75%	max
x1	51	7.1	5.7	0	3	6	10	25
x2	51	13.1	8.0	0	8	10.7	16.4	39
x3	51	8.9	5.1	0	6	8	11.4	23
x4	51	6.0	5.7	0	2.5	4	7	25
x5	51	6.2	4.2	0	3.2	5.3	8.9	18
x6	51	5.2	3.6	0	2.9	5	6.3	18
x7	51	6.3	4.4	0	2.9	5	8.3	20

note) std : standard deviation

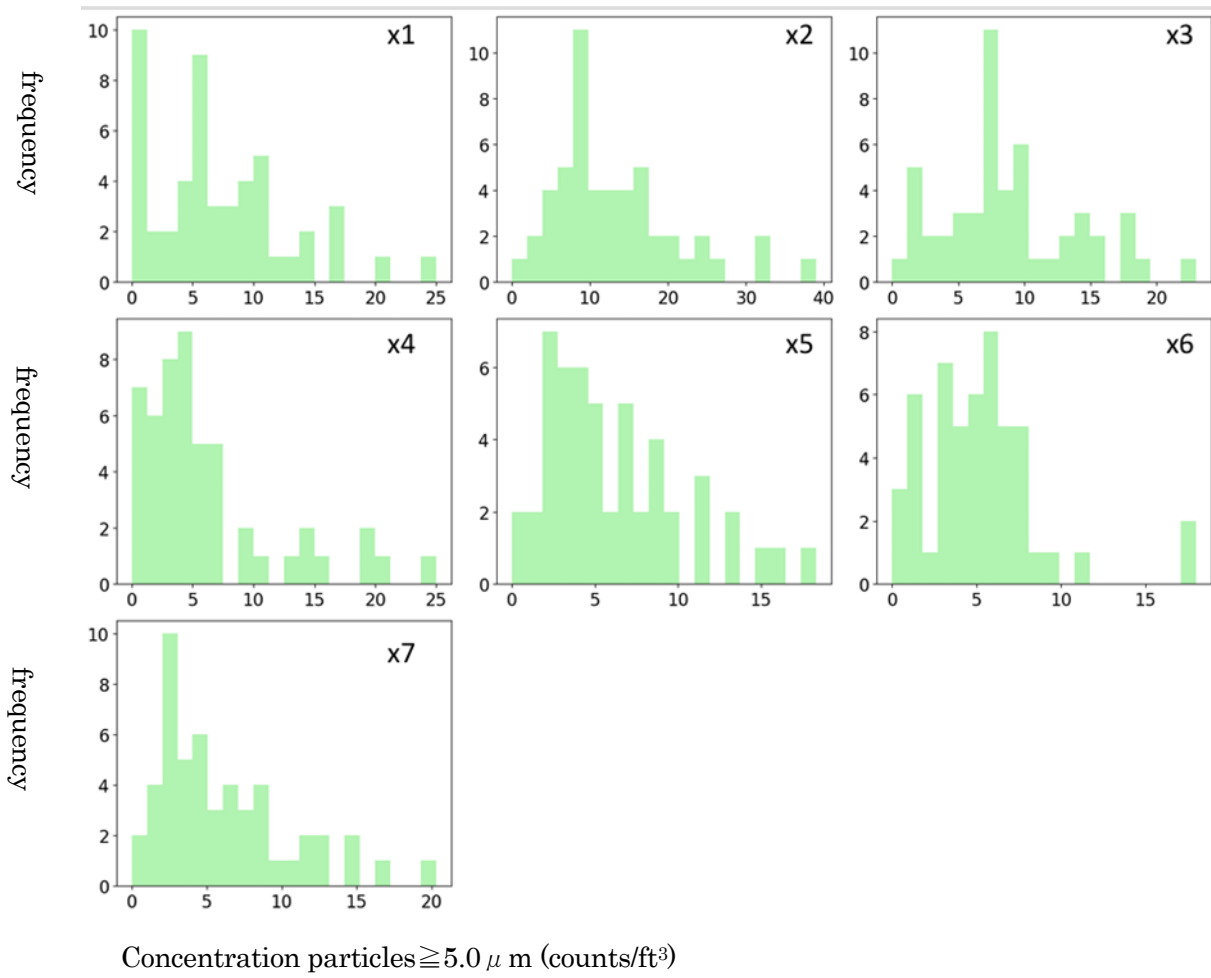


Fig.1 Histogram of cleanrooms

Table.2 Correlation for cleanrooms

ID	x1	x2	x3	x4	x5	x6	x7
x1	—	-0.02	0.05	0.27	0.14	-0.05	0.25
x2	—	—	-0.05	-0.04	0.09	0.16	0.26
x3	—	—	—	0.14	0.25	0.01	0.08
x4	—	—	—	—	-0.03	0.16	0.29
x5	—	—	—	—	—	-0.01	0.21
x6	—	—	—	—	—	—	0.23
x7	—	—	—	—	—	—	—

3. 分析方法の理論的背景

本稿で提案する分析方法の概要は、x1 から x7 のクリーンルームごとに、元のデータからブートストラップサンプリングを多数回反復して大量のデータセットを作る。そのデータセットごとに平均値を求めて平均値の数値とする。差標本数が十分に大きい場合では平均値の数値は中心極限定理と大数の法則により平均が μ 、分散が σ^2/n の正規分布に収束するため、この性質を利用して x1 から x7 のクリーンルームごとの正規分布のカルバック・ライブラー情報量を総当たりの求め。カルバック・ライブラー情報量が小さくなった組は類似性が高いと判断でき、カルバック・ライブラー情報量が大きくなった組では類似性は低いと判断できる。したがって、相関係数のような線形関係による比較ではなく、正規分布の確率密度関数による構造的な類似性の比較を行っているといった点が大きな特徴である。本節では、本研究で用いたブートストラップサンプリング方法とカルバック・ライブラー情報量の理論的な部分を簡単に示す[3]。

ブートストラップサンプリング法により、以下に示す流れで平均値の数値を生成した。

1. 整数 n' をひとつ定める。特に断らない限り $n'=n$ である。また、十分に大きな反復回数 B （本研究では $B=70000$ とした）を定める。
2. $\{1, 2, 3, \dots, n\}$ から等確率（すなわち $1/n$ ）でランダムに整数を選ぶ。これを n' 回繰り返して得られた整数列を $i_1, i_2, i_3, \dots, i_{n'}$ とする。同じ整数が複数回選ばれてもよい。

3. 得られた整数を添え字とする要素をデータの集合 X から取り出して、次式で与える。

$$X^* = \{x_1^*, x_2^*, \dots, x_{n'}^*\}$$

4. 復元抽出により得られた標本の総和をとり、サンプル数 n で割ることでブートストラップ平均を算出する。

$$\bar{X}^* = \frac{\sum_{i=1}^n X^*}{n}$$

5. 上記 2 から 4 を B 回繰り返して得られた B 個のデータを

$$\bar{X}^{*1}, \dots, \bar{X}^{*B}$$

とする。

カルバック・ライブラー情報量は、2 つの確率密度関数の乖離度（ダイバージェンス）を測る尺度である。カルバック・ライブラー情報量は以下のように定義がなされる[4]。

$$KL(g, f) = \int g(x) \log \frac{g(x)}{f(x)} dx$$

カルバック・ライブラー情報量の性質は次のように示される。

$$(1) KL(g, f) \geq 0 \quad (2) g = f \Leftrightarrow KL(g, f) = 0$$

すなわち、2 つの比較する確率密度関数の $f(x)$ と $g(x)$ が似ているほどカルバック・ライブラー情報量は小さくなる傾向にある。なお、より安定した正規分

布を得るため、次の操作を行っている。まずブートストラップサンプリングを行う前に、データの外れ値を除去した。順序統計量に基づく外れ値の除外方法⁵⁾に従えば、 $IQR \pm (1.5 \times Q1 \text{ または } Q3)$ ように示される。なお、 IQR 、 $Q1$ および $Q3$ はそれぞれ四分位範囲、下側四分位および上側四分位である。

次に、原数値を対数変換したのちに 70000 個のブートストラップサンプリングデータを生成してそれぞれの平均値を求めた。対数変換をした理由はクリーンルームにおける塵埃濃度を計測する際の測定誤差やクリーンルーム中の気流の影響を受けることで、データのばらつきは対数正規分布に従うことが経験則から考えられたため、原数値の場合と原数値を対数変換した場合、どちらがより正規分布に近いのかを評価した結果、原数値を対数変換した場合のほうが、より正規分布に近い状況となったためである。

これらの得られたデータをヒストグラムにより可視化したうえで $Q-Q$ プロットにより正規分布に従っているのかを評価した。次に、ブートストラップサンプリングによるデータの安定性を評価するために、母分散は未知であるという条件の下で、母平均と母分散の区間推定を有意水準 5%により実施した。なお、ブートストラップ後の標本数が 70000 個と非常に大きいために、母平均の区間推定には t 分布ではなく正規分布を用いて、 z 値は 1.96 を採用した。最後に、カルバック・ライブラー情報量を求めて、カルバック・ライブラー情報量の小さい順に 7 個のクリーンルームの組を選び出して正規分布の乖離の具合を確認した。

4. 結果

カルバック・ライブラー情報量により、正規分布の乖離の程度を評価するためにまず、ブートストラップサンプリングにより得られた平均値が正規分布に従っていることを確認する目的で、ヒストグラムにより各クリーンルームの平均値の分布を表したものを Fig.2 に示す。元のデータは外れ値の影響やデータ数の少なさの影響により明瞭なヒストグラムとはならなかったが、大量の平均値をブートストラップサンプリングにより生成することによって、安定したヒストグラムの形状を得ることができた。中心極限定理に従い、このヒストグラムの形状は正規分布の形状をしていることが読み取れる結果となった。

このヒストグラムでは、横軸は対数変換をした粗大塵埃濃度であり、縦軸は頻度である。左上から順に ID が x1 から x7 の順に並んでいる。

Fig.3 では Fig.2 で得られたヒストグラムが正規分布に従っていることを $Q-Q$ プロットにより評価をしたものである。横軸に正規分布の理論分位点が表示されており、縦軸にはブートストラップサンプリングから得られた平均値の分位点の値が表示されている。 $Q-Q$ プロットを見ると、ブートストラップサンプリングから求めた平均値の数値は、理論値の赤い対角線上にそのほとんどの多くが乗っているため、平均値の数値は正規分布に収束しているものと判断した。以上の結果を踏まえて、カルバック・ライブラー情報量により正規分布の乖離の程度を評価することができると判断した。

Table.3 にはブートストラップサンプリングを行った x1 から x7 までの測定場所のそれぞれの標本平均と標本分散の結果と母平均と母分散の 95%区間推定を実施した結果を示した。標本平均および標本分散の値の多くは 95%区間推定値の範囲内に収まっており、加えて母分散の区間推定の幅は x1 から x7 までのすべてにおいて非常に狭く、推定の精度が高いことが読み取れた。この背景はブートストラップサンプリングにより 70000 個の大量のデータを用いているため、漸近収束が現れていることを示すものであり、ブートストラップサンプリングの効果がいずれの推定値を非常に安定させたものと考えられた。なお、x6 の標本分散は区間推定値からわずかに外れているがこれは、母分散の推定には χ^2 乗分布が使われていることに起因する。すなわち、 χ^2 乗分布は右に裾の長い確率密度を持つ構造をしていることと、区間推定の幅が非常に狭いこともあり、標本平均とのわずかな差が生じたものと考えられるが、この結果は誤差の範囲とみることができる。

Fig.4 はカルバック・ライブラー情報量に基づく正規分布の分布間距離を示したものである。rank1 と示された左上の図ではカルバック・ライブラー情報量は 0.0067 であり、最も類似性の高い組である。x5 と x7 の正規分布はほとんど一緒といってもいいくらいに重なり合っている。その右隣にある rank2 と示した図ではカルバック・ライブラー情報量は 0.2720 であった。これは 7 個のうち 2 番目に小さかったことを示している。カルバック・ライブラー情報量が最小だったヒストグラムの重なり具合と比較してみる

と、ヒストグラムがややずれていることが読み取れた。以下同様に、カルバック・ライブラー情報量が大きくなるほど2つの正規分布の重なりが乖離していく様子が読み取れた。

この結果からカルバック・ライブラー情報量が最も少ないx5とx7では同じ管理方法が適用できうるだろうが、カルバック・ライブラー情報量が最も大きいものとなったx5とx6ではそれぞれのヒストグラムが重なり合っているところもありつつも乖離が

進んでいる状況にあるため、共通した管理方法に加えて、それぞれで異なった管理方法を設ける必要があることを示唆するものであった。

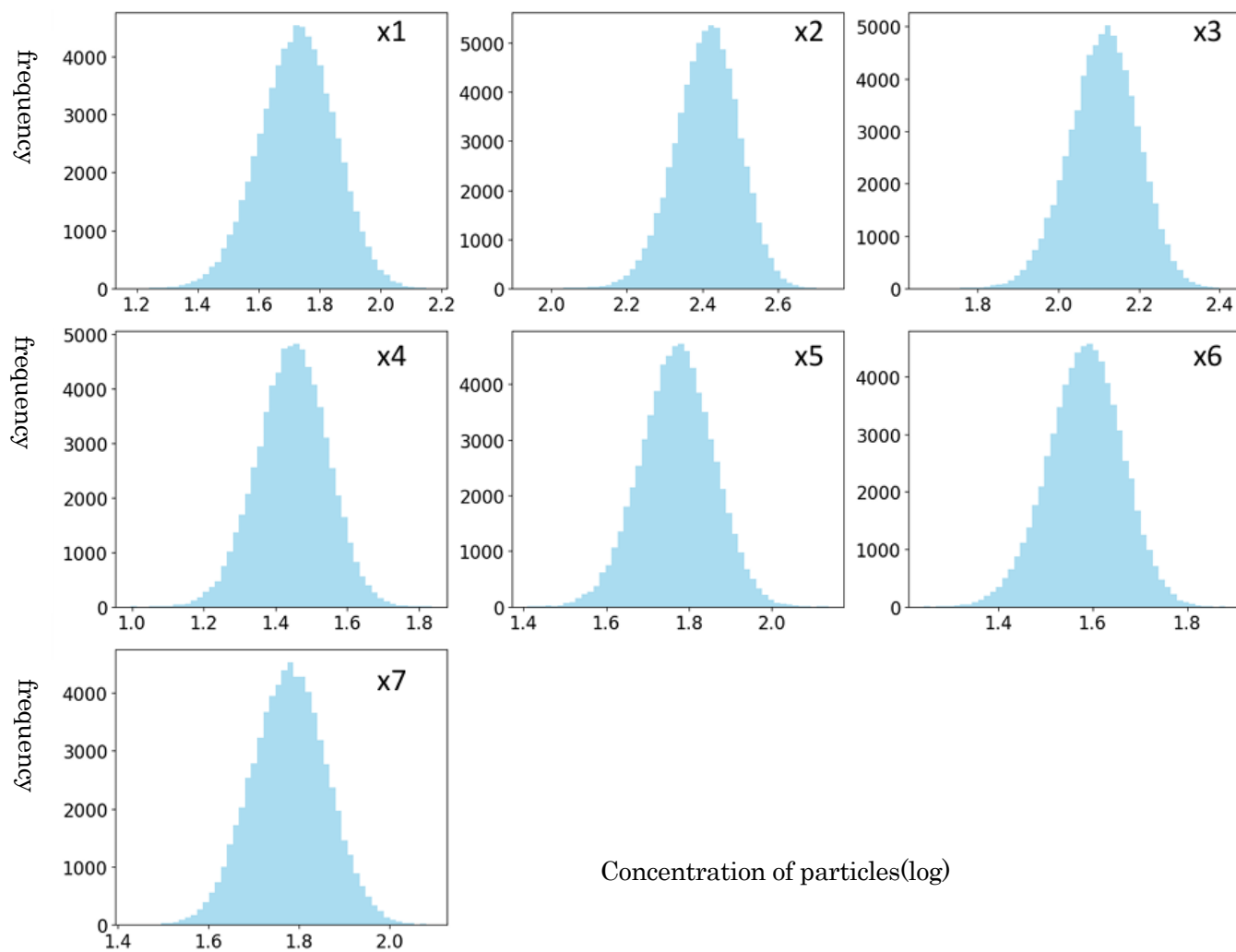


Fig.2 Histogram from Bootstrap sampling

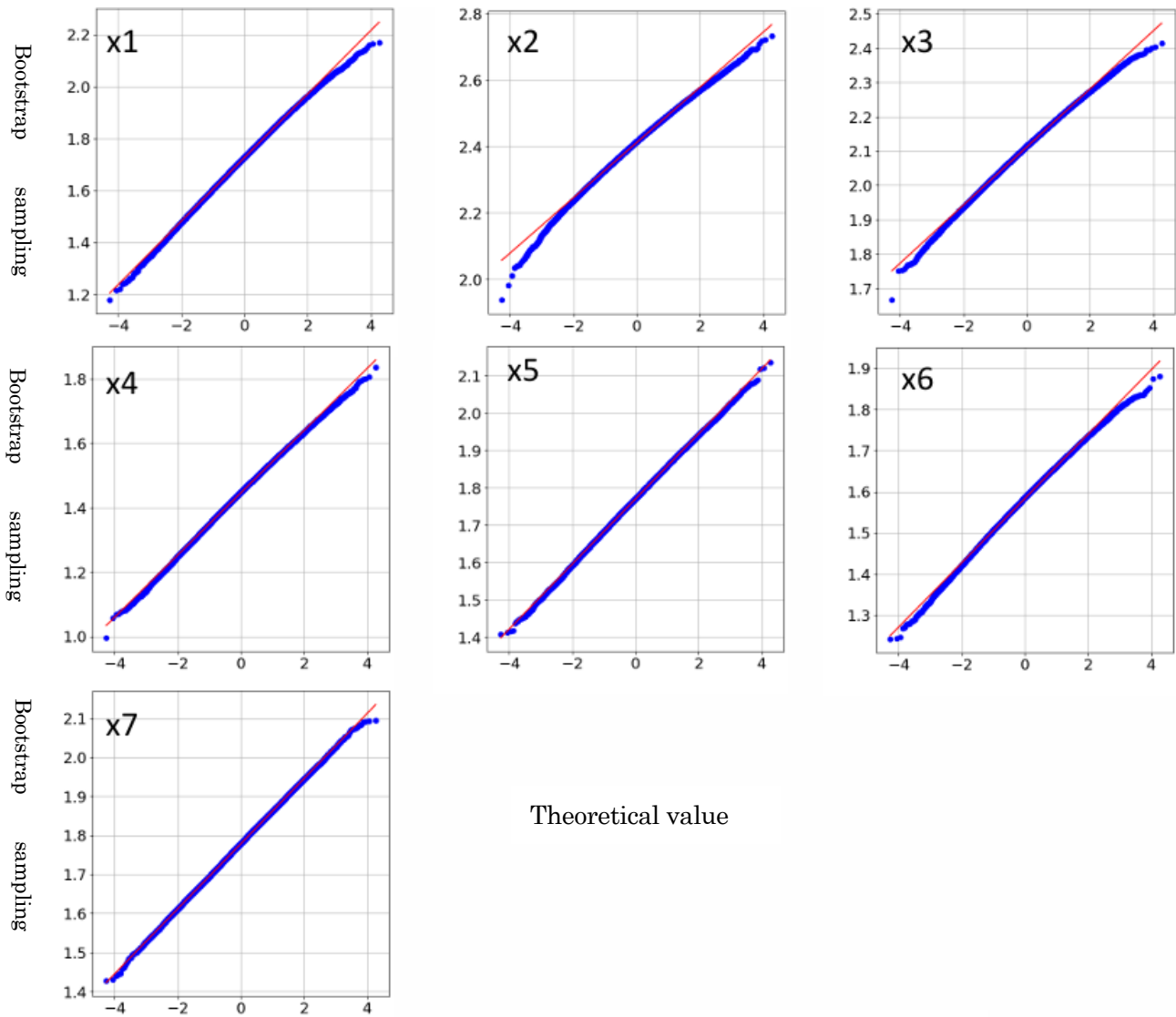


Fig.3 Q-Q plots from Bootstrap sampling

Table.3 Estimated parameters

	Location parameter		Shape parameter	
	Sample	95%Confidence Value	Sample	95%Confidence Value
x1	1.7273	(1.7263~1.7282)	0.1221	(0.1212~0.1225)
x2	2.4114	(2.4107~2.4120)	0.0833	(0.0775~0.0837)
x3	2.112	(2.1113~2.1126)	0.0846	(0.0837~0.0849)
x4	1.4467	(1.4459~1.4474)	0.0965	(0.0959~0.0970)
x5	1.7711	(1.7704~1.7717)	0.0871	(0.0866~0.0872)
x6	1.5838	(1.5832~1.5843)	0.0782	(0.0775~0.0781)
x7	1.7792	(1.7785~1.7797)	0.0833	(0.0825~0.0837)

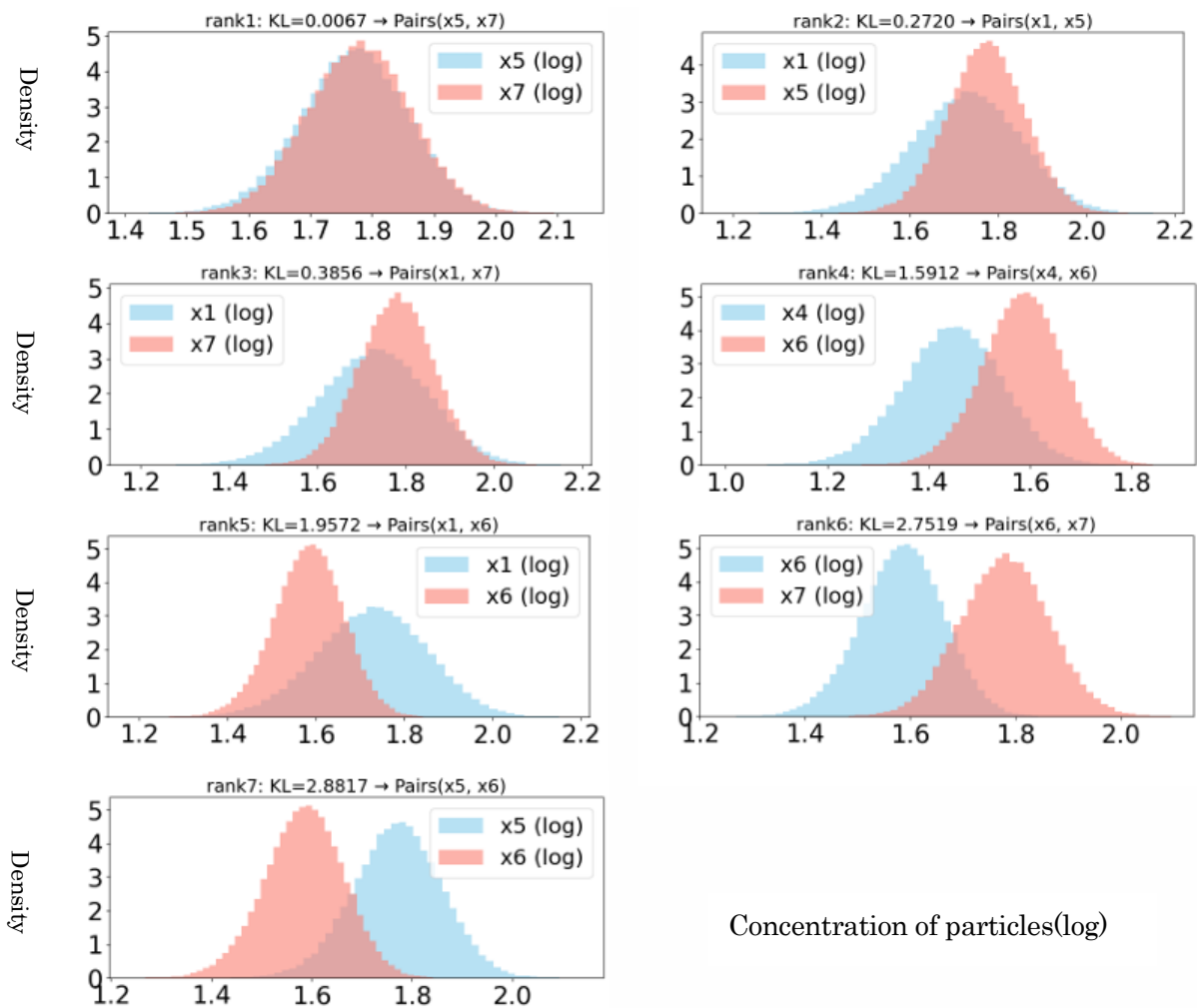


Fig.4 Compare Kullback–Libler divergence

5. 考察

本研究で示した手法は、まず統計解析をするうえで、必ずしも十分な量とは言えないデータ数をもつ計測値を、ブートストラップサンプリングにより、多数回にわたってデータを復元抽出することで、大量のデータセットを構築して平均値を求めたのち、平均値が正規分布に収束するといった性質を利用して、カルバック・ライブラー情報量により正規分布の構造的な類似性を数量的に評価することを試みたものである。

本来のブートストラップサンプリング法の使用目的としては、手元のデータから複数回復元抽出をすることで、より安定した信頼区間やその確率分布のパラメータを推定することにあるため、本研究で示したような、元のデータから平均値を生成してきて正規分布に収束させる。という方法は本来の方法と比較すると方向性が少し異なると考えられる。つま

り、本来であれば元のデータに対してブートストラップサンプリング法を使ってヒストグラムの形を安定させるといった方法が考えられる。しかし本研究で示したような、元のデータの持つ確率分布を正規分布に置き換えて評価をするメリットとしては、実務をするうえでは正規分布がよく知られた確率分布であるため、統計学の専門家ではないような実務家に対しても扱いやすいと考えたためである。一般的には、過分散のデータは離散数であれば負の2項分布、連続数ならば対数正規分布といったような確率分布が想定されるが、それぞれの確率分布の持つ位置母数や形状母数を表す数式は正規分布に比べると複雑な形をしており、直観的な理解が難しいことが考えられる。また、比較する複数の対象が必ずしも同じ確率分布の構造であるとも限らない。このような場合、カルバック・ライブラー情報量による結果の解釈が難しくなるといったことも考えられる。こ

のように、正規分布のみによって評価ができるといったことは、あるひとつの組織の異なる複数の場所の比較がたやすくなることに加えて、同じ産業分野に属している異なる組織との直接の比較が叶うことにある。このことは我が国の産業のさらなる発展を考えた際には有益であると考えられる。

本研究で示した一連の手法は、単にデータ間の線形関係の比較にとどまらず、データの構造というより深い視点での評価ができるため、そこから読み取れた結果の解釈も自然と深いものになることが期待でき、データの持っている意味合いをより強く引き出して意思決定に導くことが期待できるものである。

本研究から得られた課題であるが、本研究ではデータ数が約 50 個あり、このデータをもとにしてブートストラップサンプリングをしている。本研究で示した手法を実際の場面で扱うとなると、どの程度のデータ数が最低限必要なのかという事については、データをサンプリングするためにかかる時間的および人的資源の観点からも重要なテーマである。今回の研究ではその点については十分に掘り下げることができなかったため今後の課題であろう。また、外れ値の影響を抑えるために、外れ値を除去したが、このことは元のデータの持っている情報を一部失っていることを意味する。場合によっては外れ値がデータ構造を説明するうえで重要な意味を持つことも考えられる。安定したブートストラップサンプリングの結果を得るための、いわばトレードオフの関係にあるためより精緻な理論とするためには、外れ値を除去しなかった場合での解析を行い、推定の精度やカルバック・ライブラー情報量がどの程度変化をするのかといったことを確認する必要があるほか、今回の研究で示したような数理科学の枠組みに加えて、空調工学関連の専門家も交えた重層的な議論が必要となるであろう。

6. 結論

ブートストラップサンプリング法により、分析の対象となっているデータ数の少なさといった課題を克服し、大量の平均値を求めることにより、もとのデータの持っている確率分布に依存せずに、正規分布のカルバック・ライブラー情報量から分析対象の類似性を評価することができ、相関分析よりもより柔軟な分析が可能であることを示した。

付録.

クリーンルームに存在している粗大塵埃と呼ばれる微粒子は、粒径が $5\mu\text{m}$ 以上の粒子のことを指す^[6]。粗大塵埃は空中浮遊微生物の媒体として作用することが知られており^[7]、医療機器や医薬品、食品加工の生産現場といった、微生物による製品への汚染が問題となるような産業では重要な管理対象である。よって、定期的に粗大塵埃濃度を監視することや管理することは、製品の品質を管理する観点から極めて重要である。複数のクリーンルームを有しているような大規模な生産施設であればクリーンルームを効率的かつ効果的に管理することは実務上重要な事柄である。

このようにクリーンルームを管理することはクリーンルームを有している組織の意思決定にも影響を与えることが考えられることもあり、適切なデータ分析を行う必要がもの、筆者の知る限り我が国においては、粗大塵埃濃度に着目して複数のクリーンルームの類似性を評価したという研究成果はほとんどないと言える。

参考文献

- [1] 椿広計『ビジネス統計へのアプローチ』(朝倉書店, 2006)
- [2] International Organization for Standardization 『ISO14644-1:2015 Clean rooms and associated controlled environment -part1-』(ISO, 2015)
- [3] 国友直人, 山本拓監修『21 世紀の統計科学Ⅲ 数理・計算の統計科学』(東京大学出版会, 2008)
- [4] 藤澤洋徳『現代基礎数理 13 確率と統計』(朝倉書店, 2006)
- [5] 実務教育研究所『エクセルによるデータ解析コーステキスト 1』(実務教育研究所, 2005)
- [6] 日本規格協会『クリーンルームのおはなし』(コロナ社, 2009)
- [7] 公益社団法人日本空気清浄協会『空気清浄士(基礎コース) 認定セミナー 第 5 回』(公益社団法人空気清浄協会, 2024)

著者紹介

中澤昌高 Masataka Nakazawa
株式会社八光 メディカル事業部

東北大学大学院生命科学研究科修了
修士(生命科学)