総　説

生物統計学とデータサイエンス

緒　方　裕　光*

Biostatistics and data science

Hiromitsu OGATA*

Abstract

In recent years, the term "data science" has come into widespread use in various fields, yet its definition remains unclear. Given that data science is the science or discipline that collects and analyzes data to derive useful insights, data science is essential for most scientific research that uses data.

Meanwhile, the academic discipline of statistics has existed for a long time, and statistical methods have been used for data analysis in a variety of fields. In other words, statistics itself already plays a role in data science. However, data science today encompasses a broader meaning, including statistics. Therefore, understanding the relationship between statistics and data science is crucial when considering the future of data science.

This paper summarizes the relationship between statistics and data science and discusses the potential role of biostatistics in future data science, particularly to aid those involved in data science in the field of health science.

In the broad sense of data science, these two research fields share nearly identical goals. It is anticipated that their integration or complementation will yield a variety of results that will be beneficial to society. In particular, the combination of biostatistics with computer science will be of great benefit in fields such as public health, health science, nutrition, and medicine.

Key words: biostatistics, basic statistical concepts, data science, big data, computer science

1．はじめに

近年，様々な分野で「データサイエンス」という言葉が使われるようになってきているが，その定義はまだ曖昧である。データサイエンスが，データを収集・分析して有用な知見を導き出すための科学または学問であるとすれば，データを用いる大部分の科学研究においてデータサイエンスは不可欠である。

一方，古くから統計学という学問が存在しており，様々な分野において統計的方法がデータ解析に用いられている。すなわち，統計学自体がすでにデータサイエンスの一端を担っているといえる。しかし，現在のデータサイエンスは，統計学を含めてより広い意味を包含している。したがって，今後のデータサイエンスのあり方を考える上では，統計学とデータサイエンスの関係を理解しておくことはきわめて重要である。

本稿では，統計学とデータサイエンスとの関係について整理するとともに，とくに健康科学分野でデータサイエンスに関わる人たちの一助となるように，データサイエンスにおいて生物統計学が果たす役割について将来の可能性を述べる。

2．統計学の歴史と基本概念

統計学の起源は，古代エジプトやローマ帝国の時代にまでさかのぼる。古代においては，戦争を行う際の兵士（兵力）の人数調査，税の徴収のための人口・土地の調査などが行われていた。すなわち，統計の始まりは，様々な現状の特徴を数で示すための記述統計であった。その後18世紀頃までは，統計学という学問は存在しておらず，統計が科学の世界において特別重要な役割を果たすことはほとんどなかった。この間，16世紀頃からいわゆる賭博やゲームで勝つための計算に端を発して確率論が議論され始め，19世紀初めには基本的に現在の形の確率論の概念が出来上がった。最初は単に何かの数を数えるだけであった統計は，この確率論と結びつくことによって，急速に数学の一分野である「統計学」として理論的な体系化が進んだ。この頃にはすでに人口統計や保険の分野で統計学的モデルが使われ始めた[1-4]。

* 疫学・生物統計学研究室，女子栄養大学：Laboratory of Epidemiology and Biostatistics, Kagawa Nutrition University

19世紀末から20世紀初めにかけて，K. Pearsonにより「相関係数」「カイ2乗検定」などが提案され，1908年には，W. S. Gossetにより「t検定」（Gossetはビール会社の社員であり，本名で論文発表ができなかったためStudentというペンネームで論文を発表した。以降，t検定は，「スチューデントのt検定」と呼ばれている）が提案され[5]，統計学の応用という観点から大きな進歩を遂げた。さらに，R. A. Fisherは，スチューデントのt検定について数学的理論を発展させて，「有意差検定」「P値」などの概念を確立した。Fisherの一連の方法は，小さいサイズの標本データから，より大きなサイズの母集団の状況を推定するという考え方に基づいており，科学研究の方法に大きな影響をもたらした。Fisherは，その他にも「実験計画法」「分散分析」などを考案し，実験データや観測データにおいて制御不可能な原因により発生するバイアスを取り除くという方法を提示した。Fisherの提案したこれらの方法は，いずれもデータを取り扱う科学の世界では画期的な方法であったといえる[6]。

1930年代には，統計学の利便性または実用性という観点から大きな変化があった。J. NeymanとE. Pearson（前述のK. Pearsonの息子）は，有意差検定においてP値の有意水準（5%または1%など）を設定することによって，「有意差の有無」を判定するという統計的仮説検定の方法を提案した[7]。前述のFisherの有意差検定では，P値は帰無仮説が観測データに適合するかどうかを評価するための1つの指標であり，有意水準や対立仮説という考えはなかった。しかし，P値の有意水準を設定することによって，仮説検定の手順が定式化され，「対立仮説」「検定に必要な標本サイズ」「検出力」「第1種の過誤」「第2種の過誤」などの言葉が生まれた。Fisherは，P値に境界値を設定して帰無仮説が正しいか対立仮説が正しいかという議論をすることは科学にはふさわしくないと反論したが，結局はNeymanとPearsonの考え方はあらゆる科学の世界で広く受け入れられ，現在に至っている。科学研究において有意差があるかどうかを結論づけることは，多くの研究者にとって便利であり，とくに二者択一のような意思決定には非常に有用であったことが，Neymanらの方法が広く受け入れられたことの大きな理由の1つであろう。ただし，有意水準5%や1%という数字に科学的根拠はなく，P値には本質的に境界値は存在しない。P値の利用についてはいくつかの問題点が指摘されており，現在も統計学者の間で議論の余地がある[6, 8, 9]。

上記の長い歴史を経て，現代の統計学の基本的部分についてはほぼ確立されたと言える。この統計学を用いたデータ解析には以下の特徴がある。すなわち，①母集団から無作為（ランダム）に抽出した標本データの解析から母集団の状況を探ることができること，その際，②母集団のデータの分布と標本データ（または標本データから計算される統計量）の分布との関連が重要であること，③データの特性に応じて検出力の高い検定方法を選ぶことができること，などである。これらの考え方は，標本調査を行う限りにおいては将来もその有用性は変わらない。

統計学の応用分野の1つである生物統計学は，人間（生物全般と考えてもよい）に関わるデータを解析する方法論や理論に関する学問である。この生物統計学に関しては，20世紀後半に入って大きな変革があった。第二次世界大戦終了後の1950年頃から，アメリカ，イギリスをはじめとする先進諸国では肺がんによる死亡率が急増したが，その原因は不明であった。1950年代に，イギリスでは医師のR. DollとA. B. Hillにより，最初は症例対照研究[10]，のちにコホート研究が行われ[11]，肺がん死亡率増加の原因が喫煙（戦時中から兵士の喫煙者が増加し，戦後もそのまま喫煙習慣が定着した）によるものであることを突き止めた。この研究は，慢性疾患に疫学的研究方法を応用した世界最初の例となった。しかし，両者の因果関係の有無については，その後20年にわたり激しい論争があり，喫煙が肺がんの原因であることが世間に認められるには長い時間がかかった。Hillはこれらの議論を受けて，因果関係の有無について議論する際の判定基準（時間的前後関係，関連性の強さ，生物学的妥当性，他の知見との整合性など）を提案し[12]，統計的な関連があることは，因果関係が存在することを示唆する要因の1つであることを示した。この頃から，保健学，公衆衛生学，疫学などにおける統計学の応用が進み，現在は，広く人間の健康を取り扱う統計学は「生物統計学」と呼ばれている。

20世紀の終盤までは，多くの科学研究において，統計学を単にデータ解析のツールとみなす考え方が主流にあった。一方で，統計学を数学の一部と考える統計学者は数学的な側面の探求に力を注いできた。しかし，1990年代に入りコンピュータの進展もあって欧米を中心にデータサイエンスという言葉をキーワードとする研究が増えてきた（Fig. 1）。日本でも統計学を「データの科学」としてとらえることが多くなった[13-15]。これらの状況は，データに基づいて意思決定を行うという考え方やエビデンスに基づく医療（EBM）などの概念とも並行して，科学だけでなく様々な実用的分野において科学的データが重視され始めたことを意味している。

3．コンピュータサイエンスの発展

古くから計算をするための簡単な道具（算盤，アバカスなど）が考案されてきたが，現在のコンピュータと言われる装置またはシステムが最初に開発され始めたのは1940年代である。その後1970年代，1980年代以降にいわゆるパーソナルコンピュータ（パソコン）が登場し，統計学分野において，最尤法，ブートストラップ法，マルコフ連鎖モンテカルロ法など，複雑な計算が実行可能になってきた[16]。近年の統計学が応用科学としての性質を強く持つようになった主な理由は，コンピュータの普及と計算技術の進歩により，統計学の専門家でない研究

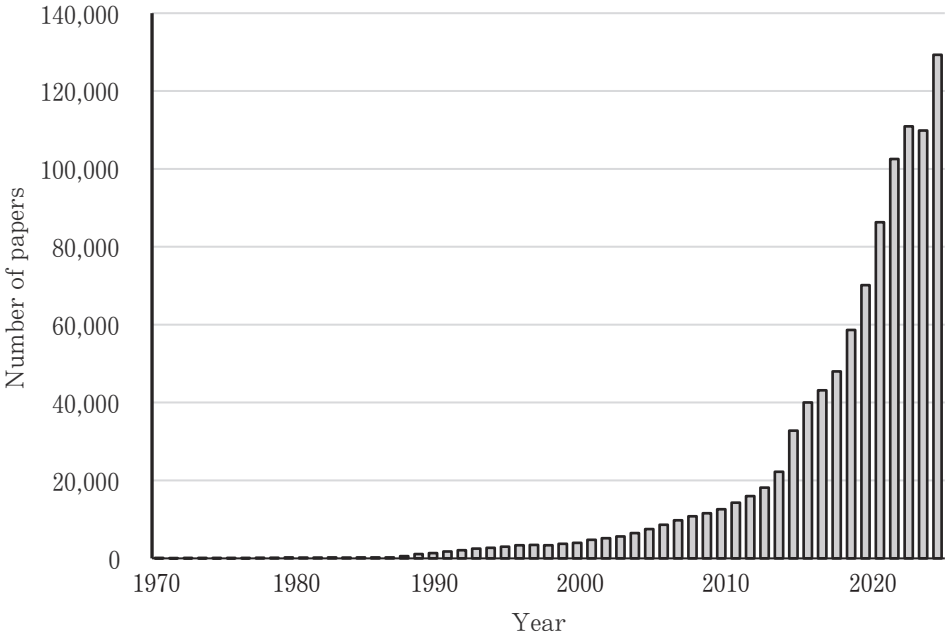

Fig. 1　Frequency distribution of scientific papers that were recorded in PubMed between 1970 and 2024, and their content was related to “data science”

者，実務者が比較的簡単に高度な統計的方法を使ってデータ解析ができるようになったことにある。すなわち，20世紀末までに統計学はすでにデータサイエンスの一部として成立していたといえる。このように，統計学の進歩はコンピュータの発展と密接に関連してきた。一方，1950年代，1960年代の初期のコンピュータの登場とともに，すでに学問として成立し始めていたコンピュータサイエンス（計算機科学）が21世紀に入ってから飛躍的に進歩したことが，データサイエンスの分野に大きな変化をもたらすことになった。

近年のデータサイエンスにおけるコンピュータサイエンスの最も大きな貢献は，機械学習，深層学習，人工知能（AI）などによりビッグデータの取り扱いが可能になったことである。一般に言われるビッグデータの特徴は，①情報量が膨大であること，②非常に速い速度でデータが更新されること，③データの種類が多様であること，の3点である。これらの特徴に加えて，データ自体の信頼性，意思決定における有用性，バイアスの有無などがビッグデータの評価基準として考えられている[17, 18]。21世紀以降のコンピュータサイエンスは，このビッグデータの解析に焦点が当てられ，統計学とある程度共通の目標を持ちながらも互いにほとんど独立した分野として研究が進められてきた。一方で，統計学はコンピュータサイエンスの進歩の速度に追いついておらず，その進歩の速度はコンピュータサイエンスに比べてはるかに遅い印象である。これには，数学的理論を重視する統計学と，現実世界の実用性を重視するコンピュータサイエンスとの違いが影響していると言われている。例えば，1つの新しい統計学的理論が認められるまでには，学術誌への投稿論文が査読を通して公表されるまでを含めて長い時間がかかるのに対して，コンピュータサイエンスの成果は学会発表の講演要旨でも十分に評価されている。

いずれにしても，科学の一分野としてのデータサイエンスには，統計学とコンピュータサイエンスという2つの起源がある。ビッグデータの分析という観点からみると，21世紀以降はデータサイエンスの中心はコンピュータサイエンスに移っている。今日では，データサイエンスのことをコンピュータサイエンスの応用と考える人も多い。しかしながら，データサイエンス自体が多様な要素を含む動的な科学であるとすれば，技術の急速な進歩とともに今後どのような発展を遂げていくかについて確実な予測をすることは困難である。長い歴史を経て強固な数学的理論に裏付けされた統計学は，データサイエンスの中でもいくつかの利点を持っている。また，今後の世界に存在するすべてのデータがビッグデータになるわけではないことを考えれば，統計学が引き続きデータサイエンスの基盤の1つとなることは間違いない。

4．データサイエンスにおける統計学とコンピュータサイエンスの違い

20世紀までは，統計学とコンピュータサイエンスは互いに関連をもちながら発展してきたが，21世紀以降は，統計学が依然として数学的理論を深めることに中心的課題があったのに対して，コンピュータサイエンスの専門家たちは実用的なデータ分析のために独自のツールを開発し始めた。両者は，いずれも近年のデータサイエンスの源流ではあるが，大きな発想の違いがあり，それらがアプローチや焦点の違いに反映されている[2, 19, 20]（Table 1）。以下では，それらの違いについて述べる。

Table 1. Main differences between statistics and computer science-inspired data science

	Statistics	Computer science
Main purposes of data sampling	Inference about population	Prediction, classification, etc.
Main focus	Model development, validating methods, correlations between variables, causation, etc.	Performance on benchmark datasets, application of machine learning, data mining, etc.
Models	Simple model such as general regression models	Algorithm model allowing over-parameterized models
Interpretability	High	Low
Accuracy	Low	High
Data	Well-structured data	Unstructured data

4.1 データ収集の目的

統計学では，明確に定義された母集団から原則として無作為に標本を抽出し，その標本データを分析することによって母集団に関する何らかの合理的推論を行うことを目的としている。したがって，母集団のデータの分布に関する仮定や，そこから抽出した標本データの分布などが重要な問題となる。また，興味のある変数間の関連を見るために，交絡因子やバイアスの影響をなるべく排除することに大きな注意を払っている。一方，コンピュータサイエンスを起源とするデータサイエンスでは，原則として予測に必要なパターンを学習するためにデータを収集するので，その時々で利用可能なデータを取り扱うことが多い。コンピュータサイエンスにおいては，母集団と標本という概念はあまり重要ではないため，データの代表性（推論を一般化できるかどうか）は，用いるデータの種類に依存する。

統計学で用いられる手法は，数学的に説明できることがその妥当性を担保することになるので，多くの統計学者は統計学の理論的側面に焦点を当ててきた。コンピュータサイエンスおいても確率論は重要な数学的基盤の一つではあるが，その重点は，より効率的で信頼性の高いソフトウェアの開発やシステム構築などにある。コンピュータサイエンスを用いたデータサイエンスでは，ある時点で利用可能なデータを利用して予測や分類を行い，その後データが蓄積されていくことによって計算をやり直し，次第に予測や分類の精度を改良していくというアプローチをとる。

4.2 使用されるモデル

統計学では，確率論に基づく統計モデルを想定し，そのモデルが適切であれば，標本データから導かれる推論は母集団に適用される。しかし，標本データがその統計モデルと適合しない場合には，採用したモデルが不適切なのか，または単にサンプルサイズが小さいのか，などのことを検討しなければならない。例えば，回帰モデルのようにパラメータ数が少ない単純なモデルがデータに良くあてはまるようであれば，結果の解釈は比較的容易である。したがって，統計モデルは，データに基づく現象論をメカニズム論に変換する際の「橋渡し」的な役割をもつことができる。さらに，統計モデルを使うことにより，推定値の不確実性を定量的に表現することも可能である。しかし，サンプルサイズ（データの個数）が大きくなり，変数の数が増えるにつれて，単純なモデルのあてはめが困難になる。一般にパラメータ数が多いほどデータへの適合度は良くなるが，パラメータ数が多すぎるモデルでは，現実的な解釈が成り立たなくなる。そのため，統計学ではAICやBICなどの情報量基準がモデル選択の判断材料として用いられることが多い。これらの情報量基準は，パラメータの数とデータへの適合度の両者のバランスを考慮してモデルを選択するという考え方に基づいている。一方，コンピュータサイエンスの分野で用いられるモデルは，広義には人間が経験を通じて学習していく過程をコンピュータで模倣するための計算式や計算方法を示している。これらのモデルは機械学習モデルまたはAIモデルなどと呼ばれる。

4.3 データ解析の観点から見た課題

ビッグデータの取り扱いが可能になった現時点での統計学とコンピュータサイエンスは，データ解析という観点からみるとそれぞれに課題や限界がある。

まず，統計学については，以下のような課題がある。①ビッグデータが全数調査に近いデータであるとすれば，小標本のデータから母集団を推定するという統計学的推論が意味をなさなくなる。②サンプルサイズが膨大で，なおかつ変数の数も非常に多い場合，通常の統計計算（最尤法など）による解が収束しない可能性がある。③統計学では比較的単純な一般線形モデルまたは一般化線形モデルが使われることが多いが，未知の状況を予測するにはこれらのモデルの予測精度はあまり高くない。

一方，コンピュータサイエンスについては，以下のような課題がある（と思われる）。①母集団と標本という関係を重視しないために，使われるデータによっては偏った結果（予測や分類）になる可能性がある。すなわち，データの代表性に関する問題が残る。②ビッグデータを扱う場合，算出するパラメータの数が多く，モデルの過剰適合（過学習）になる可能性がある。このことは，予測が正確になる一方で解釈に結びつけることが難しくなることを意味している。③データに個人情報が多く含まれる可能性が高いため，倫理的な問題をクリアしなければならない。

5．生物統計学の視点から見たデータサイエンスの今後

データサイエンスは，データを用いるあらゆる分野で適用が可能であるが，応用される各分野にはその分野における専門性がある。すなわち，データサイエンスを具体的に各分野で応用するには，その分野の専門知識が不可欠である。したがって，データサイエンスは，統計学とコンピュータサイエンスに加えて各分野の専門性が構成要素となる[1, 2, 16, 19, 20]（Fig. 2）。ここでは，専門分野の 1 つとして，健康科学（保健学，公衆衛生学，栄養学など）に焦点を当てて，この分野で使われる統計学（すなわち生物統計学）の観点から，これまでの議論を踏まえていくつかの方向性を提案する。

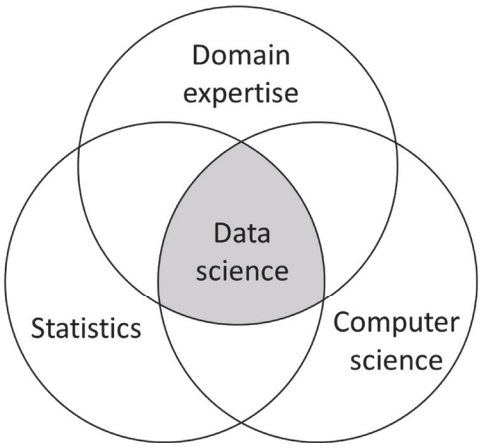

Fig. 2　Components of data science

健康科学の主な目的は，第 1 に，疾患やその他の健康事象の原因を探り，公衆衛生対策などの意思決定に結びつけること，第 2 に，集団のデータに基づいて医療，栄養指導，保健指導など個人に対する適切な対応を見出すことにある。これらの科学的根拠として使われるデータは人間の健康や生活習慣にかかわるものである。その特徴として，人間の健康状況には多種多様な要因が複雑に関係していること，人間に関わる測定値は不確実性を伴うこと，個人間のばらつき（個体差）が非常に大きいこと，などが挙げられる。一方，コンピュータサイエンスの最大の利点は，ビッグデータの解析が可能であることである。この特徴を生かして，生物統計学をデータサイエンスに応用するとすれば以下に挙げるようなテーマが考えられる。いずれのテーマについても，現実的な課題があるものの生物統計学とコンピュータサイエンスが互いの欠点を補完することによって解決の可能性がある。

5.1 因果関係の推論

生物統計学では，現象として変数間の関連の有無を定量的に示すことはできても，（厳密な意味では）因果関係を証明することはできない。疫学研究で 2 つの変数間に因果関係があることを主張する際に満たすべき条件の 1 つが，原因と結果の時間的前後関係である。この条件は，前述の Hill の判定基準の 1 つにも挙げられている（通常，Hill の基準のすべてを満たすことは難しいが，唯一時間的前後関係だけは絶対に必要な条件である）。一般に人間に関わるあらゆるデータは時間に依存している。生物統計学の分野では，コホート研究や生存時間研究などで時間を変数として用いるモデルが様々考案されている。しかし，サンプルサイズの小さい疫学調査では，複数の変数の組み合わせに関して細かい時間的依存性を発見することは困難である。健康事象に関わる時間的変動は，あらゆる健康事象において重要な要素である。これらの影響を見出すことができれば，交絡因子の制御やバイアスの除去につながり，因果関係の推論をより精緻化できる。

ビッグデータのように大量のデータが常に更新される状態にあれば，それらのデータを用いて細かい時間単位で時間的前後関係が明確になる。具体的には，人々の行動，生活，その他の要因が健康状態の変化に影響を与えているかどうか，などを調べることができる。因果関係の推論の精度が上がることによって，予測モデルにもつながる。課題は，交絡因子の影響をどのようにして制御するか，因果関係の推定に用いたモデルの妥当性をどのように評価するか，パラメータの推定方法をどうするか，などである。

5.2 個人差の評価

様々なバックグラウンドを持つ人間には大きな個人差があり，生物統計学におけるデータ解析では，人間の個人差は何らかの統計的分布（正規分布など）に従うと仮定されることが多い。また，個人差は統計的推論における不確実性の原因の 1 つであるが，この不確実性の大きさを定量的に表現できることは生物統計学の利点の 1 つである。しかし，人間に関わるデータの不確実性の原因は無数に存在する。これらの原因をデータから見出すことができれば，集団のデータから得られた知見を個人に適用する際に，より精緻な対応が可能となる。ビッグデータの利点は，複数の多様な種類のデータ，例えば，個人の遺伝情報から個人の生活環境データ，個人の社会経済状況にいたるまで非常に幅広い範囲の個人情報を突合できることにある。これらのデータ解析は，医療における個々の患者の特性に応じた精密医療，公衆衛生や栄養学における個人に対する保健指導や栄養指導，個人の状況に応じた健康危機管理などに活用できる。課題は，どのようなモデルを用いるか，パラメータの推定方法をどうするか，バイアスの除去をどのように行うか，などである。また，データ解析の結果を個人に適用する際には，データ収集以前の段階から個人情報保護や研究倫理の遵守などの問題をクリアすることが必要となる。

5.3 対象者と変数のグループ化

生物統計学の方法の1つに，対象者をその特徴に応じてグループに分ける方法がある。健康科学上のグループ化の意義は，特定のグループが持つ特徴が原因となって何らかの健康事象を誘発している場合に，その特徴と健康事象との関連性を推定できることにある。グループ化は，食生活のように個人によって多様性の大きいデータ（すなわち高次元のデータ）を，データの持つ情報を保持したまま，次元数を減らすことになる。生物統計学では，グループ化の方法として，主成分分析，因子分析，クラスター分析，多次元尺度法などがよく用いられている。これらの方法では対象者をグループ化できるだけでなく，複数の変数についても互いの関連性に基づくグループ化が可能である。ビッグデータのように対象者数も変数の数も非常に多い場合には，データの次元数を減らすことによって高次元のデータを直観的に解釈しやすい形に変換することができる（この変換はデータの可視化にも役立つ）。また，生物統計学におけるグループ化の方法では，一定の値に結果が収束するまで反復計算が行われている。しかし，ビッグデータを用いてグループ化を行う際には，同様の方法では計算結果が収束しない可能性がある。したがって，課題としては，どのようなアルゴリズムを用いてグループ化を行うかという点にある。ただし，この種の課題は，機械学習が得意とするテーマの1つであり，実際にすでにいくつかの解決策は考えられている[2, 21]。

6．結　論

生物統計学を含めた統計学の発展には長い歴史があり，現在に至っている。一方で，コンピュータサイエンスの歴史は統計学に比べれば比較的新しく，とくに21世紀に入ってからの進歩はめざましい。2つの研究分野はいずれも現在のデータサイエンスの起源ではあるが，近年に限って言えば，両者はほとんど無関係に研究が進められている。広義のデータサイエンスにおいては，2つの研究分野はある程度共通の目標を持っており，今後は両者が融合または互いに補完することによって，社会に役立つ様々な成果が得られることが期待される。とくに人間を対象とした生物統計学がコンピュータサイエンスと結びつくことによって，公衆衛生学，保健学，栄養学，医学といった分野で大きな力を発揮するであろう。

文　献

1) Karahan, S., Karaagaoglu, A.E.,: Development of biostatistics: from past to future. *Duzce Med J*, **23**, 234-238 (2021)
2) Kirch, C., Lahiri, S., Binder, H., *et al*.: Challenges and opportunities for statistics in the era of data science. *Harv Data Sci Rev*, **7**, 1-50 (2025)
3) Ocana-Riola, R.: The use of statistics in health sciences: situation analysis and perspective. *Stat Biosci*, **8**, 204-219 (2016)
4) Sprent, P.: Statistics in medical research. *Swiss Med Wkly*, **133**, 522-529 (2003)
5) Student: The probable error of a mean. *Biometrika*, **6**, 1-25 (1908)
6) 緒方裕光: 統計的仮説検定の問題点と今後の課題. 健康教育学会誌, 31, 30-35 (2023)
7) Neyman, J., Pearson, E.S.: On the problem of the most efficient tests of statistical hypotheses. *Philos Trans R Soc Lond Ser A*. **231**, 289-337 (1933)
8) Mayo, D.G., Hand, D.: Statistical significance and its critics: practicing damaging science, or damaging scientific practice? *Synthese*., **200**, 220, 1-33 (2022)
9) Wasserstein, R.L., Schirm, A.L., Lazar, N.A.: Moving to a world beyond "p<0.05". *Am Stat*, **73** (sup1), 1-19 (2019)
10) Doll, R., Hill, A.B.: Smoking and carcinoma of the lung. *Br Med J*, **2 (4797)**, 739-748 (1950)
11) Doll, R., Hill, A.B.: The mortality of doctors in relation to their smoking habits. *Br Med J*, **1 (4877)**:1451-1455 (1954)
12) Hill, A.B.: The Environment and Disease: Association or Causation? *Proc R Soc Med*, **58**, 295-300 (1965)
13) 新村秀一: データ解析からデータ・サイエンスへ. 統計数理, 45, 23-40 (1997)
14) 大隅昇: インターネット調査の適用可能性と限界－データ科学の視点からの考察. 行動計量学, 29, 20-44 (2002)
15) 田栗正章, 岩崎学: コンピュータを活用する新しい統計学. 行動計量学, 19, 34-36 (1992)
16) Wu, H.: Statistics evolution and revolution to meet data science challenges. *Stat Biosci*, Published online 03 September 2024 (2024)
17) Wesson, P., Hswen, Y., Valdes, G., *et al*.: Risks and opportunities to ensure equity in the application of big data research in public health. *Annu Rev Public Health*, **43**, 59-78 (2022)
18) Raita, Y., Camargo, C.A., Jr., Liang L., *et al*.: Big data, data science, and causal inference: a primer for clinicians. *Front Med*, **8**, 1-13 (2021)
19) Hassani, H., Beneki, C., Silva, E.S., *et al*.:The science of statistics versus data science: what is the future? *Technol Forecast Soc Change*, **173**, 121111 (2021)
20) Hewett, T.E., Olsen, G., Atkinson, M.: The use of big data to improve human health: how experience from other industries will shape the future. *Int J Sports Phys Ther*, **16**, 1590-1594 (2021)
21) Wang, J.: Biostatistical challenges in high-dimensional data analysis: strategies and innovations. *Comput Mol Biol*, **14**,163-172 (2024)