参考資料：データサイエンスとは何か？ 『事例で学ぶ！あたらしいデータサイエンスの教科書』から第 1 章を引用公開

019/12/24 07:00

様々な領域のデータを収集できるようになった今、データサイエンスに基づく意思決定や社会課題の解決に期待が寄せられています。その手法を学ぶにあたっては、そもそもデータサイエンスがどういうものなのかを知っておかなければなりません。横浜市立大学データサイエンス学部の初代学部長である岩崎学さんが解説する『事例で学ぶ！あたらしいデータサイエンスの教科書』（翔泳社）から、「第 1 章 データサイエンスとは」を抜粋して紹介します。

本記事は『事例で学ぶ！あたらしいデータサイエンスの教科書』の「第 1 章 データサイエンスとは」からの抜粋です。掲載にあたり、一部を編集しています。

データサイエンスとは何でしょうか。

その解答は十分に確立しているとはいえませんし、個人ごと組織ごとに違った答えを持っているかもしれません。しかし大まかには、

データサイエンス＝（統計学＋情報科学）×社会展開

といえるのではないでしょうか。データを扱う学問である統計学に加え、実際にデータを処理するための情報科学をその基盤とし、様々な社会課題の解決への展開につなげるのがその使命です。

本書では、社会展開を念頭に置いた上で、主として統計学の視点からデータサイエンスについての様々な側面を取り上げて論じます。

本章ではプロローグとして、これまでの統計的データ解析について概観したのち、それが現在のデータサイエンスではどのように変貌しつつあるかを見ていきます。

1.1 これまでの統計的データ解析の流れ

まず、本節でこれまでのデータ解析の流れを確認し、それを元にして次の 1.2 で現在のデータサイエンスの特徴について論じます。

1.1.1 これまでのデータ解析の手順

これまでのデータ解析の一連の手順を図 1 にまとめます。

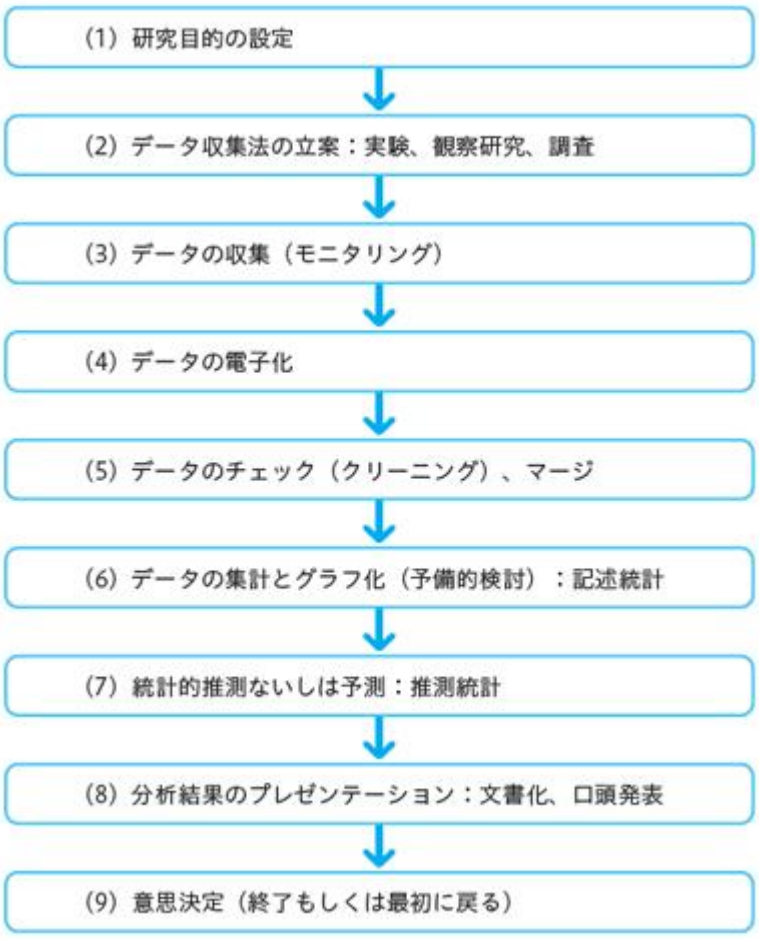

図 1 統計的データ解析の流れ

図 1 について、若干の説明を以下に加えます。

（1）研究目的の設定

極めて当然ですが、まずは研究目的が明確に認識されている必要があります。単なる現状把握なのか、近未来の予測なのか、あるいは人為的な介入による変化をもたらすための方策を提供するのか。目的に応じてデータの取り方は変わってきますし、分析の方法論および結果の提示の仕方も影響を受けます。

（2）データ収集法の立案：実験、観察研究、調査

研究目的が認識されたら、それを実現するためのデータ取得の計画を立てる必要があります。データの取得にはコストがかかりますから、研究目的を確実に実行できることを前提に、なるべく効率的なデータの取得法を工夫しなければなりません。

統計学はこれまで、データ取得法の方法論を発展させてきました。研究目的が調査であれば「標本調査法」が、処置効果の立証などであれば「実験計画法」がデータ取得の方法論を与えてくれます。大学における統計学の授業では近年、これらの内容が講義されることが少なくなってきていますが、

"garbage in, garbage out"

の言葉があるように、データの質が悪ければ、よい分析結果は望むべくもありません。統計的データ解析で最も重要なものはデータを集める方法論であるとは、統計学の大御所のご託宣です。

（3）データの収集（モニタリング）

よいデータを得ることがデータ分析のイロハのイですが、黙っていてはよいデータは得ることはできません。ここにある程度のコストをかけなくてはなりません。

例えば新薬開発の臨床試験では、各製薬メーカーはモニターという職種の部隊を抱えていて、かなりの人数の人たちがよいデータを取るための業務に携わっています。これはどの分野でも同様で、よいデータを取るための方策なくして質のよいデータは決して得られないと知るべきです。

（4）データの電子化

現在では、データ分析を紙と鉛筆および電卓で行う人はいません。データは必ずコンピュータに入力した上で分析にかける必要があります。

しかし以前には、データは調査票やアンケート用紙などの紙媒体で提供されるのが一般的でしたので、それを電子化する必要がありました。現在ではほぼ死語となったキーパンチャーのようなデータ入力の専門家もいました。

現在でもデータ入力は極めて重要な仕事で、その後の分析を見据えた上でのデータの準備が必要です。

（5）データのチェック（クリーニング）、マージ

データは、多くの場合というよりほとんどすべての場合、そのままでは分析にかけることはできません。分析のための整形が必要ですし、データの欠損や異常値の存在など多くの問題を解決せねばなりません。また、分析が１つのデータセットのみで完結することは稀で、複数のデータセットの結合（マージ）が必要となります。その際には、データのマッチングを含めた地道な作業が必要となります。

実際にデータ分析を行うとすぐにわかりますが、ここの部分でのエネルギーの消費はかなりの量に上ります。人によってはデータ分析の 7～8 割の労力がこの段階でかかるといわれることもありますが、これは決して誇張ではありません。

（6）データの集計とグラフ化（予備的検討）：記述統計

本格的な分析の前に、データの全体像を把握しておく必要があります。ここで有用な方法論が、いわゆる記述統計的手法です。データは多くの場合数字の羅列ですので、それを見やすくするためのデータの集計は欠かせません。また、データのグラフ表示による視覚化も重要な手立てです。この段階だけで、分析の目的が達成されることも多くあるでしょう。

（7）統計的推測ないしは予測：推測統計

統計的な推定や検定などのいわゆる推測統計的手法は、データの素性を的確に捉え、近未来の予測や新しい知見を得るために必要となります。

大学などにおける統計学の授業ではここの部分が主として講義されます。数学的な扱いが主となり、難解な数式展開などが含まれたりしますので、とっつきづらい面は否めませんが、統計手法の数理的な側面の理解はデータの分析によって妥当な結論を導くために必要不可欠です。

（8）分析結果のプレゼンテーション：文書化、口頭発表

データを分析したらその結果を何らかの形で示さなくてはなりません。文書化および口頭での発表が必要となります。その際に重要なのは、分析結果を過不足なく客観的に伝える姿勢です。データの持つ情報を十分に捉えきれないのでは分析者として失格ですし、逆に結果をことさらに誇張するのも慎まなくてはなりません。データ分析の結果はその後にデータで証明されます。

例えば新薬の開発で薬の効果をことさらに強調し過ぎても、その薬が実際に患者さんに投与されれば、その有効性はデータとして返ってきます。新商品に関するアンケート調査の結果を誇大に強調しても、実際にそれを販売すれば売上高がその成否を証明してくれます。

（9）意思決定（終了もしくは最初に戻る）

データの分析結果は、それを得ることだけが目的であることはないでしょう。それに基づいた何らかの意思決定がなされなくてはなりません。もし意思決定に至らないのであれば、さらにデータを取り直すなどの算段が必要となり、このリストの最初に戻ります。

1.2　データサイエンスの特徴

ここでは、統計的データ解析の流れがどのように変化してきているのか、あるいは変わってはいけないものは何であるのかを議論します。

1.2.1　統計的データ解析の現状と変化の方向

1.1 の統計的データ解析の流れは、いかにコンピュータが発達し人工知能（AI）がもてはやされようとも、また統計学がデータサイエンスに取って代わられようとしても陳腐化するものではなく、やはり押さえておかなければいけない真理を含んでいます。

普遍的な価値を持つ原則を押さえることにより、そこからの乖離の程度を測りながら現代の複雑なデータの分析を行う必要があります。以下では、前節で提示した統計的データ解析の流れが現状どのように変化しつつあるかを見ていきます。

（1）研究目的

世の中にはデータがあふれています。そのままにしておいたのでは宝の持ち腐れ、何とかしなければというのは誰もが思うことです。しかし、目的がなくては何のしようもありません。初めは目的が不明確であったとしても、最終的なゴールを早く見つけ出さなくてはなりません。

特にデータの量が膨大になり、その質もまちまちである現在、データのハンドリングには思ったより長い時間がかかるようにもなっています。分析のツールやシステムを導入すればすべてが解決する、というのは全くの幻想です。目的があいまいなままいたずらに時間を浪費する愚を犯してはなりません。

（2）データ収集法の立案

データは、それを集める時代から、集まっているあるいは集まってくる時代へと変わってきました。特に各種センサーの発達により日々刻々とデータが自動的に蓄積され、SNS のように我々一人ひとりがデータの入力源となって、せっせとデータを蓄積しつつあります。それに伴い昨今、データを集める方法論がおろそかになっているという危惧があります。どのようなデータ収集法が理想であるのかの知識を持った上で、現在あるデータがいかにして取られたのか、それは理想的な収集法に比べどこに不備があるのかを認識することが重要です。

（3）データの収集

データが自動的に集まってくる昨今ですが、どのようにして収集がなされているのかのモニタリングはやはり重要です。データの背景に関する知識は、適切なデータの分析法の選択と実行のために必要不可欠です。

（4）データの電子化

数値に限らず、昨今ではテキスト、画像、音声そして動画などが電子化され、電子データとして入手が可能になっています。コンピュータの記憶媒体の大容量化と通信速度の飛躍的向上がそれを後押ししています。以前の、データが紙で提供されていた時代とは様変わりしました。情報科学の技術革新の賜物といえるでしょう。

（5）データのチェック

この段階は、現在でもやはり手間暇がかかります。人手で行うにはデータの量が膨大過ぎるからでしょう。ここをいかに自動化し人手を煩わさないようにできるかが、迅速なデータ分析のポイントです。異常値の検出やデータの欠損への対応などの自動化は、データの分析の一連の流れを加速させる上で極めて有効な手段となります。さらなる研究が待たれる分野です。

（6）データの集計とグラフ化

この段階のテクノロジーの進展には目を見張るものがあります。超大量のデータの迅速な集計、美しい動画を交えた洗練されたデータの可視化などを実現化する様々なツールが提供されています。最近のデータは大量であるが故にその大まかな特徴を的確に捉える必要があり、そのためにはこの種の可視化ツールは大いに有用です。この段階で必要にして十分な情報が得られることも多いでしょう。また、その後の分析法の選択にも示唆を与えてくれます。

しかし注意すべきは、きれいなグラフィックスが得られただけで満足してしまいかねないことです。だからどうした、結局どうなるの、といった疑問に的確に答えるためには、やはり「次の段階」が必要でしょう。

（7）統計的推測ないしは予測

ビッグデータの扱いなどでは、推測統計の精細な議論は必要でないかもしれません。しかし推測統計の元となる、「現象をモデル化してデータの分布の関数形を定め、そこに含まれる未知パラメータをデータから推定した上でその推定値の精度の情報も提供する」という考え方や哲学は、やはり必要不可欠といわざるを得ません。

しかし予測に関しては、深層学習（ディープラーニング）に代表される機械学習の諸手法が、これまでの古典的な統計学のいわば型にはまった統計手法の限界を超え、極めて柔軟なモデルに基づいた、精度のよい予測値を与えることができるようになりました。予測の方法は面目を一新したといっても過言ではありません。

とはいえ、その予測方法は中身がブラックボックス化していて、なぜ予測が当たるのかについての知見をもたらしてくれない、という課題があります。単なる予測を超え、当該現象における因果関係の確立による人間の介入法にまで到達できるかがカギとなります。

（8）分析結果のプレゼンテーション

プレゼンテーションツールも非常な進化を遂げつつあります。これまでの単なる数表やごく簡単なグラフだけでなく、目を見張るようなビジュアルによるダイナミックなプレゼンテーションが現実のものとなっています。しかしそれらに幻惑されてはならず、やはり中身が重要です。

もちろん、中身の充実があった上での説得力のあるプレゼンテーションツールの適用は、望ましいものでしょう。

（9）意思決定　データ分析の結果は、意思決定の役に立たなければ何の意味もありません。そのことを肝に銘じて新時代のデータ解析を行う必要があります。

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊以上