

データサイエンス入門

—データ取得・可視化・分析の全体像がわかる—

上田雅夫・後藤正幸

担当編集から

本書は、大学の「データサイエンス」や「統計学」の講義で教科書として使われることを目的として作られましたが、ビジネスや意思決定の場面で、データを活用したい方に向けたデータサイエンスの入門書にもなっています。統計学と機械学習の手法・モデルをいずれも一冊で紹介しているところ、またデータの可視化手法に紙幅を割いているところが特徴的です。

データサイエンスと法学は、近年ますますかわりが強くなってきているように思います。ひとつには、裁判例などの法に関連するテキストデータを、機械学習などの計量的手法によって分析する研究があります。こういった分析では、裁判の結果をデータから予測したり、法や裁判例の構造を明らかにしたりすることを試みているようです。また、機械学習などの技術が使われるAIの実社会への応用について、法学的観点から分析する研究分野もあり、実務面でもとても注目されているかと思えます。

法学とのかかわりに限らず、一般社会でもデータサイエンスの考え方は役に立つ場面が多いです。少しでも気になった方はぜひ一読を！（岡山）

Point

第5章「データの可視化手法」では、目的に適したさまざまな可視化手法が具体例とともに紹介されています。

第5章 データの可視化手法

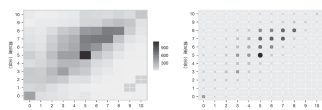


図 5.15 散点図の改良

にデータが集中しているのかをグラフで表現することで回避することができます。

——グラフに別な情報を付加する

データをグラフにし、手前にあるデータの情報をグラフを用いて伝える際、棒グラフや折れ線グラフに別の情報を付加することで、分析者の意図を早く正しく伝えることができます。図 5.16 は図 5.12 でも用いた内閣府の調査における各府の満足度の各選択肢に対する回答の結果を相対頻度（比率）で表したものです。この結果を見ると 5 点以上の回答者が多く、満足度の高い人の比率が高いことが理解できます。

このグラフに、累積人数の折れ線グラフを追加したのが、図 5.16 の右のグラフになります（グラフ中の点線は、累積人数の比率が 50% を示しています）。このグラフからは、左のグラフから読み取るよりも容易に、「選択肢 6 で累積人数の比率が 50% を超過する」「選択肢 8 以降は、回答人数が顕著に減る」といった内容を理解することができます。

基本となるグラフに情報を追加する事例は、他にもあります。先ほど、箱ヒゲ図の課題を指摘し、その課題を克服するようなヴァイオリン・プロットというグラフがあると説明しました。同じようなことは、図 5.17 の右図にあるように、箱ヒゲ図にデータの分布を点として重ねることで表現することができます。

先の項で示した工夫は、すでにあるデータを別の形で表現するものです。累積人数の比率は、棒グラフのそれぞれの値を足し上げたものでしかありません。

124

2 データの構造から考える可視化

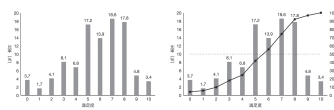


図 5.16 満足度に関する各選択肢の相対頻度

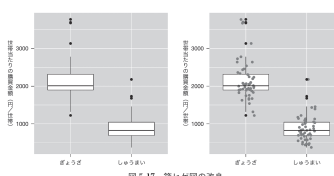


図 5.17 箱ヒゲ図の改良

ん、時には、グラフに別の情報を付加することで、そのグラフで表現されている内容をより深く理解できるようになることがあります。例えば、図 5.18 は都市別のしゅうまいの世帯当たりの購買金額を示したのですが、左のグラフからは、都市によって金額に差があることしか理解できません。そこで、左のグラフに平均値を示す線（黒い実線）を加えたのが、右上のグラフです。このグラフから、平均よりも多い都市、少ない都市は同じ程度であることが理解できます。図 5.18 の右下のグラフは、±1 標準偏差の値を示す線（破線）を加えたものです。この結果より、しゅうまいの世帯当たりの購買金額は、平均よりも上振れている都市が多いことが理解できます。

このように、グラフに何らかの線を引くことは、珍しいことではありません。工場などの製造現場で利用されている、管理図という時系列チャートにも

125

レベル	用途	対象
初級	学習 教養 実務	学部 一般

2022年12月発売／278頁／定価2420円（税込）

A5判／並製

詳細を見る



BOOK INFORMATION

詳細は、小社ウェブサイトの本書のページをご覧ください。

