

データ分析の基礎知識

滋賀大学データサイエンス学部 学部長 竹村 彰通
滋賀大学データサイエンス学部 教授 高田 聖治

本稿では、まずデータ分析が重要となってきた社会的背景を説明し、その後データの基本的な集計（1変数及び2変数）及び公的データの使い方について説明する。

1 データ分析の重要性と基本的な集計（1変数）

公的統計などのデータは、地方自治体が施策を立案し、その効果を評価する際の根拠として重要なものである。最近では情報通信機器の発展にともないさまざまなデータが入手できるようになったため「データの時代」とよばれている。そしてデータ分析のスキルは自治体の職員にとってもますます重要なものになってきている。政策立案を「経験と勘」ではなくデータに基づいて行うことを「エビデンスに基づく政策立案」（EBPM, Evidence Based Policy Making）とよぶ。最近の政府の統計改革の動きの中でEBPMが強く求められるようになってきている。

企業活動においても、データの収集と有効な利用が競争力のもとと考えられている。特に、インターネット上の検索サービスや通信販売においては、ネットから得られる情報がいわば「原材料」であり、その情報を如何に加工して有用なサービスを展開するかが、それらの企業のいわば本業である。最近では、グーグルやアマゾンのように、インターネット上の情報を大量に収集しそれに基づいてさまざまなサービスを展開してユーザを獲得する企業が急速な成長をとげ、世界的に独占的な位置を確保する状況になっている。それらの企業があまりにも大きな影響力を持つこと

に対する懸念も大きくなっている。

データの時代とよばれるようになったのは、得られるデータの量が増大したことのみならず、例えばソーシャルネットワークサービス（SNS）における書き込みなどのテキストデータや、写真投稿サイトに投稿される画像データなど、非数値的で多様なデータが得られるようになったことも影響している。画像データからその画像に何が写っているかを判別する画像認識の技術も発展してきた。テキスト情報については、人々の興味や関心进行分析して商品開発につなげるような使い方もされている。このような多量かつ多様なデータは「ビッグデータ」とよばれている。ビッグデータを分析する手法がデータサイエンスである。またデータサイエンスを実践する人材をデータサイエンティストとよぶ。

スマートフォンの普及はデータの時代の到来を象徴するものである。スマートフォンが登場したのは2007年であるが、今では世界的に普及し、日本でも20代のスマートフォンの個人所有率は95%程度となっている。人々はスマートフォンによりネットワークと常時接続されている。そして、人々の検索履歴、購買履歴、移動履歴がビッグデータとしてネットワークに蓄積されている。通信方式について述べると、現在のLTE方式など4Gとよばれる通信方式に代わって、今後5Gとよばれる通信方式が普及すると見込まれている。5Gでは通信速度が4Gの100倍となり、例えば自動運転の際の自動車からの大量のデータ通信にも耐え得ると考えられている。このようにスマートフォン経由や自動運転車から蓄積され

るデータの量は今後ますます増えていくものと予想される。

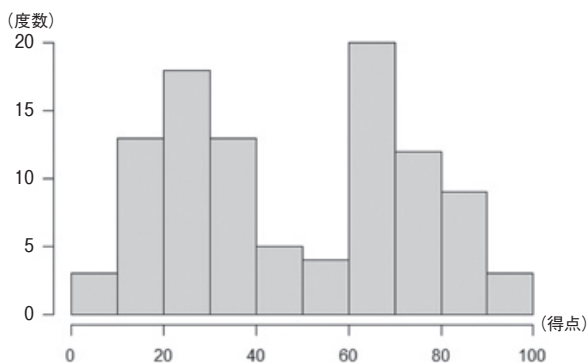
このような中でデータが「21世紀の石油」とであると言われるようになってきた。すでに述べたように、この資源を独占的に活用する企業が現れてきているため、日本においてもデータを流通させる仕組み作りや、データを分析する人材育成が急務となっている。残念ながら現状では、日本はデータ流通を管理しきれないまま、海外企業に活用されてしまっている状況にある。

この状況を改善するため、日本政府もデータ流通を後押しするための法的な整備を進めている。改正個人情報保護法、官民データ活用推進基本法、次世代医療基盤法などである。さらに統計制度の改革や、データ分析にかかわる人材育成にも力を入れている。

以上、データ分析の重要性の社会的背景について述べてきたが、以下ではデータの基本的な集計について説明する。まず1変数データを扱う。

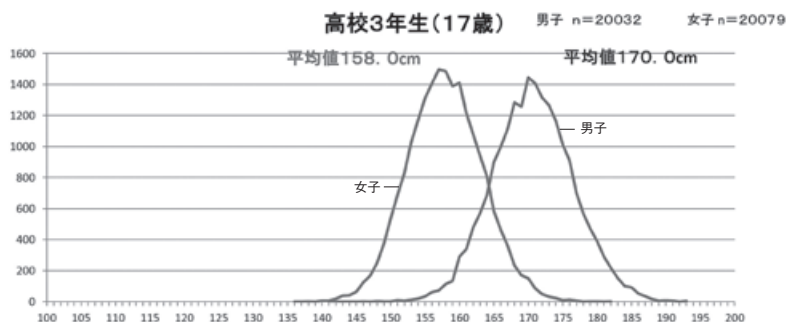
1変数データの集計の基本はヒストグラムの作成である。ヒストグラム自体の説明は省略するが、ヒストグラムを描いた時その形状に注意を払う必要がある。特に図1のように山が2つあるヒストグラムの場合は、調査の対象となる集団（「母集団」）が2つの異なる部分母集団から構成されている可能性がある。

図1 ふた山のヒストグラムの例



身長ヒストグラムでも、図2の男女別のヒストグラムから想像されるように、仮に男

図2 男女別の生徒の身長ヒストグラム
（「平成24年度学校保健統計調査」より）



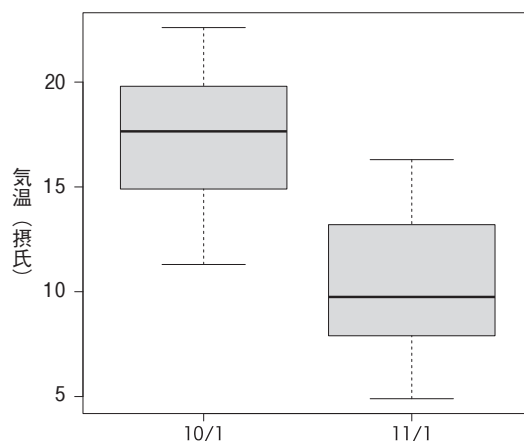
女と一緒にヒストグラムに描いたとするとふた山のヒストグラムが得られる。

母集団が性質の異なる複数の部分母集団からなる場合には、母集団を部分母集団に分割して調べるのが重要である。このことを母集団の「層別」という。性別、地域別、年齢階級別、所得階級別など、層別してデータを眺めることは、データを理解するための基本的な手法である。層別の重要性については散布図やクロス集計表にもふれながら以下で説明する。

ヒストグラムの簡易的な表示として「箱ひげ図」が高等学校の数学Iで教えられるようになり、広く用いられている。箱ひげ図は、データの散らばり具合を、箱とひげを使って表した図であり、データの中央値、最小値、最大値、第1四分位点、第3四分位点の位置を一度に表示できる。第1四分位点と第3四分位点により箱を描き、その箱の中に中央値を線で示す。また最小値、最大値へはひげを伸ばす。

図3は彦根市の10月1日と11月1日の1988年から2017年の30年分の最低気温を箱ひげ図

図3 彦根市の10月と11月の最低気温の箱ひげ図

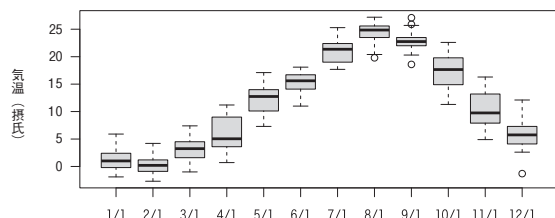


で表したものである。このように箱ひげ図は簡易的な図であり、複数の箱ひげ図の比較も容易である。

箱ひげ図はデータの可視化のすぐれた方法であり、自治体の作成する文書でも活用するとよい。

なお、箱ひげ図には一定のルール（「テューキーの方式」）で外れ値を小さな○で表すこともよく行われる。テューキーの方式で、彦根市の各月の1日について、30年分の最低気温を12個の箱ひげ図で比較したものが図4であり、箱ひげ図の表示の有用性が理解できる。

図4 彦根市の各月の最低気温の箱ひげ図



2 データの基本的な集計方法 ～2変数データと相関～

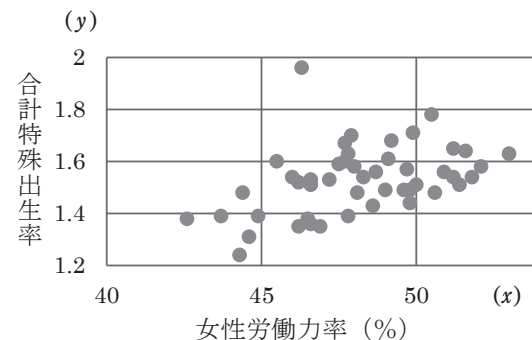
データ分析を実際の政策に活用するためには、複数の種類のデータを組み合わせて、その間の関係を分析することが必要になる。例えば、市の人口の推移をグラフにしてみることは、現状の把握及び将来予測に役立つかもしれないが、それだけでは人口減少に対する対策を立てることは困難である。人口に影響しそうなデータを集めてきて、それらと人口減少との関係を分析し、操作可能な変数であればそれを操作することによって将来の人口減少を食い止めることができる。以下では、変数間の関係を分析するための基本的な手法をいくつか紹介する。

(1) 散布図

散布図は、2つの変数の値を平面上に図示することによって、関係性を可視化（ビジュアライズ）したものである。一つ一つの変数を見ていただけではなかなか分からない関係性も、図示することにより見出しやすくなるという利点がある。

図5は、都道府県別の「女性労働力率」と「合計特殊出生率」との散布図を描いたものである。一つ一つの点が各都道府県を示しており、x軸に「女性労働力率」、y軸に「合計特殊出生率」をとって、該当するところに点を打った。

図5 都道府県別女性労働力率と合計特殊出生率（2015年）

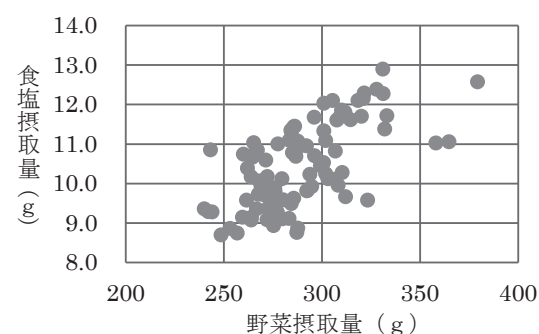


これを見ると、都道府県別の「女性労働力率」と「合計特殊出生率」との間には、ゆるやかな正の関係があることが見てとれる。もちろん、このことだけから「女性労働力率を上げれば出生率が上がる」と短絡的に結論付けることはできないが、女性の活躍が出生率向上にもつながる可能性を示唆している。

(2) 層別

散布図によって2つの変数間の関係を可視化することができるが、データに異質なものが混ざっている場合は正しい関係を読み取れない。例えば、次の散布図を見てみよう。

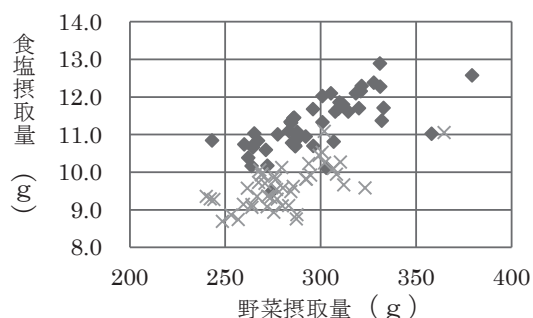
図6 都道府県別の野菜摂取量と食塩摂取量



これは、厚生労働省「平成24年国民健康・栄養調査」のデータを使って、都道府県別の野菜摂取量と食塩摂取量（ともに年齢調整済み）との散布図を描いたものである。一部に食塩摂取量が高い県であって野菜摂取量も高

い県があるものの、それらを除くと、だいたい真ん中あたりに点が集中して、野菜摂取量と食塩摂取量との間に明確な関係は見受けられないようにも見える。実はこの散布図は、男性と女性のデータが混在したものであり、男性と女性とで分けてみると図7のようになる。

図7 都道府県別の野菜摂取量と食塩摂取量（男女別）



今度は、先ほどの散布図と同じデータを使って、男性は◆、女性は×で表したものである。男女別に分けてみると、それぞれで、「野菜摂取量が多い県は食塩摂取量も多い」という関係が見てとれる。つまり、食塩摂取量と密接に関係する変数として、「野菜摂取量」と「性別」というものがあり、このうち「性別」を考慮に入れずに分析してしまうと間違った結論を導いてしまうおそれがある、ということである。これを避けるためには、上記のようにデータを性別で分けて分析する必要がある。このようにデータを分割することを「層別」という。なお、都道府県別データを用いて、平均寿命と野菜摂取量との関係进行分析すると、一般的に言われている「野菜摂取は健康によい」という関係があまり明確に出てこない。これは、上記のような「野菜摂取量が多い県は食塩摂取量も多い傾向がある」ということとも関係している。せっかく野菜摂取量が多くても食塩摂取量が多いと、血圧の上昇によって平均寿命にマイナスの影響を与え、野菜摂取によるプラスの影響を相殺してしまうからである。

(3) シンプソンのパラドックス

データの層別が重要であることのもう一つの例として、シンプソンのパラドックスとよばれるものを紹介する。これは、全体が各層

の加重平均で表されるときに、加重平均の重みが異なることにより、各層における結果と全体での結果が相反するものになることがある、というものである。

1970年代のアメリカ・カリフォルニア大学バークレー校において「大学院進学合格率が、男性の方が女性より高いのは男女差別ではないか」という問題提起があった。実際、男女別の合格率は表1のとおり、男性の方がかなり高くなっている。

表1 カリフォルニア大学バークレー校大学院の合格率（1973）

	志願者数	合格率
男性	8442	44%
女性	4321	35%

出所) ウィキペディア（英語版）“Simpson's paradox”
https://en.wikipedia.org/wiki/Simpson%27s_paradox

しかし、これを学部別に見ると、多くの学部において女性の方が男性より合格率が高い、という現象が見てとれるのである。

表2 学部別・男女別の合格率（上位6学部）

学 部	男 性		女 性	
	志願者数	合格率	志願者数	合格率
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

出所) 同上

すなわち、多くの学部において女性の合格率の方が男性の合格率より高いものの、女性は合格率の低い学部（表2だと学部EやF）への志願者が多いために全体の合格率（各学部の合格率の加重平均になる）が低めになるのに対し、男性は合格率の高い学部（学部AやB）への志願者が多いために全体の合格率が高くなり、全体で見たときと学部ごとで見たときに男女別合格率の高低が逆転してしまうのである。これを避けるためには、やはり学部ごとに層別してデータを見ていく必要がある。

(4) 相関係数

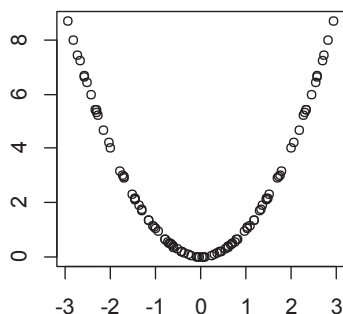
2変数の間の関係を端的に表す指標が相関

係数であり、一方の変数が増えるときにもう一方の変数も増える傾向があるときはプラス、逆に一方の変数が増えるときにもう一方の変数が減少する傾向があるときはマイナスの値をとる。

相関係数は表計算ソフトで簡単に計算することができるが、簡単であるがゆえに、使い方には十分注意する必要がある。例えば、

- ①相関係数は2変数間の直線的な関係性を表すものであり、例えば図8のような場合には2つの変数間に関係があるといえるが相関係数は0である。

図8 相関係数=0となる例



- ②相関係数の値は、特異な値をとるデータ（外れ値）に大きく左右されることがある。

①、②のようなことを避けるためには、相関係数の値だけを見るのではなく、散布図を描いて2つの変数の関係を確認することが重要である。

(5) 相関関係と因果関係

2つの変数の間に相関関係があるとは、一方の変数が増えるともう一方の変数が増える（または減る）傾向があることをいう。これは、相関係数が ± 1 に近い値をとるということに等しい。

一方、2つの変数の間に因果関係があるとは、一方の変数の増減が原因となってもう一方の変数の増加（または減少）が起きることをいう。原因となる方の変数を人為的に変化させることができればもう一方の変数の増減をもたらすことができるので、因果関係は政策分析においてとくに重要である。

因果関係があれば一般的に相関関係がある

が、逆は必ずしも成り立たない。例えば、

- ①2つの変数の原因となる第三の変数の存在

ある市におけるアイスクリームの売上高と水難事故との間に正の相関関係があったからといって、因果関係があるとは考えられないであろう。これは、両方に影響を及ぼす「暑さ」という要因があって、

- ・暑いとアイスクリームの売り上げが増える
- ・暑いと水泳に出かける人が増え、事故も増える

という関係にあると考えるべきであろう。

- ②偶然、相関関係があった

「プロ野球の阪神タイガースが優勝すると景気が良くなる」という都市伝説？があるが、これは偶然の産物であろう。

因果関係については、これら以外にも「因果関係の向き」、すなわちどちらが原因になっているかを注意する必要がある。例えば、道路ごとの「交通安全標識の数」と「交通事故の発生件数」とをとって散布図を描くと、この2つの変数の間には正の相関が見られるであろう。ただし、これは「交通安全標識があるから交通事故が起きる」ことを意味しているのではない。逆に、「交通事故が多い道路に、標識を多く配置した」と考えるべきであろう。

3 公的データの使い方

(1) e-Statの利用

データを使って政策分析を行う場合、さまざまな公的データを利用することができる。その際に有用なのが、政府統計のポータルサイトであるe-Statである。わが国では各府省がさまざまな形で統計調査を実施しているが、

図9 e-Statの画面



