@IT
atmarkIT

Excelで学ぶ、やさしいデータ分析

羽山 博［著］

この連載では、データをさまざまな角度から分析し、その背後にある有益な情報を取り出す方法を学びます。

データの収集方法、データの取り扱い、分析の手法などについての考え方を具体例で説明するとともに、身近に使える表計算ソフト（Excel や Google スプレッドシート）を利用した作成例を紹介します。

必要に応じて、Python のプログラムや統計ソフト R などでの作成例にも触れることにします。

数学などの前提知識は特に問いません。肩の力を抜いてぜひとも気楽に読み進めてください。

高校生に負けない！ 社会人が学ぶべき、やさしいデータ分析

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載のスタート。今回は、なぜデータ分析の重要性が高まっているか、ビジネスに生かすために何を学ぶべきかを概観した後、連載の全体像を紹介します。

羽山博（2023 年 04 月 20 日）

もはや中学・高校生も学んでいるデータ分析&データサイエンス

ここ数年、**データ分析**や**データサイエンス**という言葉をさまざまな場面で目にすることが多くなりました。しかし、そもそも有史以前から人類は生活を向上させるために、星の動きを見たり、川の水量の増減を見たりして、種まきの時期を知ったり、災害の予測を行ってきました。そういった活動もいわばデータ分析と考えられます。さまざまな現象を数字で表したり、数学の言葉で書くようになったのもそう新しい話ではなく、今から 5000 ～ 6000 年以前には、かなり高度な数学も成立していたようです *1。

*1 『代数的構造』（遠山啓著 , ちくま文庫 , 2019）

その後、17 世紀以降、確率や統計の研究が進み、現在使われている手法がほぼ確立しています。ただ、そういった歴史をひもといたり、専門的な手法を知らなかったりしても、私たちはデータ分析がどのようなものか直感的に理解していると思います。また、最近になってデータ分析の重要性が高まっていることを肌で感じている方も多いと思います。

ちなみに、原稿執筆時（2023 年 4 月 4 日）に、Amazon の書籍を「データ分析」というキーワードと刊行年で検索し、その件数をグラフ化してみたところ図 1 のようになりました。2023 年は 105 件でしたが、まだ 1 年の約 1/4 しか過ぎていないので、105×4=420 件としてグラフにしてあります（年々、刊行点数が増えているので、さらに増えるかもしれません。あるいは、飽和状態で頭打ちになるかもしれませんが）。

→目次に戻る

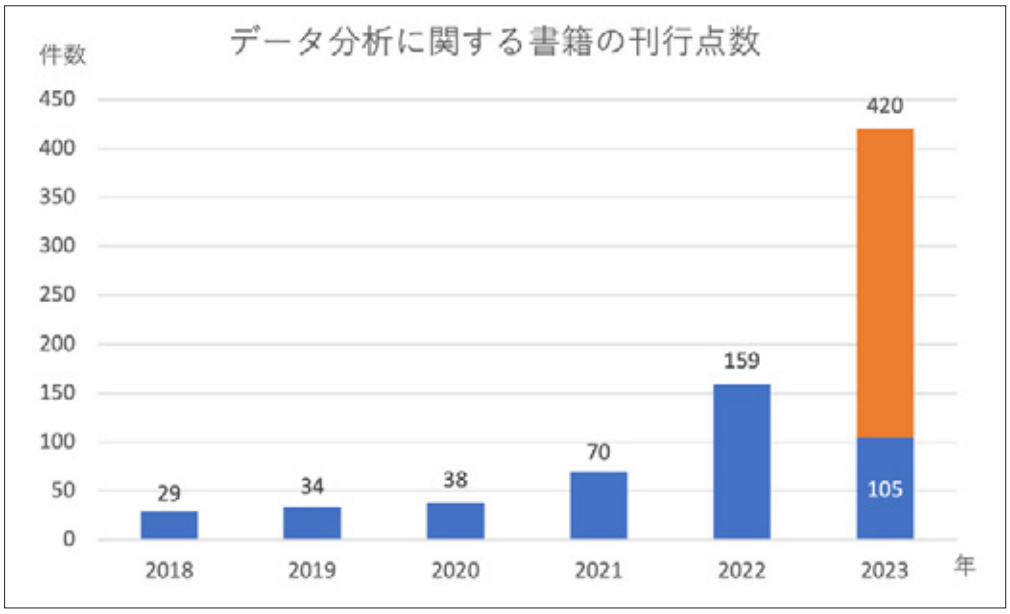


図 1　データ分析に関する書籍の刊行点数（2023 年は予測値）

データ分析関連の書籍の刊行点数が指数関数的に増えていることが分かる。なお、検索結果には高校野球のデータや入試問題の分析なども含まれているので、データ分析の手法に関係する書籍は 2020 年までは少なかった。ちなみに、このようなグラフを作成して「データ分析」の注目度を可視化することは（私たちが日常的に行っている）データ分析の一つ。

データ分析やデータサイエンスの重要性は、学校教育にも大きな影響を与えています。表 1 に示した内容は、新しい学習指導要領に取り入れられたデータ分析とデータサイエンスに関する内容です。この学習指導要領に基づく教育は、中学では 2021 年から全面実施、高校では 2022 年からすでに全面実施となっています。いかがでしょう。それ以前に高校を卒業した方には驚きを隠せないほどの大きな変化ではないでしょうか。

教科・科目	学年	項目	取り扱う主な内容（太字はこの連載でも取り扱う）
数学	中1	データの活用	**データの分布、ヒストグラム、相対度数、コンピュータを利用した表やグラフの作成**
数学	中2	データの活用	**四分位範囲、箱ひげ図**、場合の数、確率
数学	中3	データの活用	標本調査、母集団の傾向の推定
数学I	高1	データの分析	**分散、標準偏差、散布図、相関係数**
数学B	高2	統計的な推測	確率変数と確率分布（二項分布・正規分布）、区間推定、仮説検定、社会生活での数理的な考察、問題解決
情報I	高1	情報通信ネットワークとデータの活用	**オープンデータ、データの形式、量的データ、質的データ、尺度、可視化、単回帰分析**、Webスクレイピング、テキストマイニング
情報II	高2または高3	情報とデータサイエンス	**データの整形、データクリーニング、重回帰分析**、主成分分析、分類、クラスタリング、ニューラルネットワーク、画像認識

表 1　学習指導要領のうち、データ分析やデータサイエンスに関係する項目

学習指導要領、高等学校情報科「情報 I」教員研修用教材、高等学校情報科「情報 II」教員研修用教材を基に作成。各教科のうちデータ分析やデータサイエンスに関係する項目だけを抜き出して整理してみた。数学 B と情報 II は選択科目。高校の学年は目安。

→目次に戻る

表 1 に掲載した内容について全くなじみがないという方も心配には及びません。この連載で少しずつ丁寧に説明していくので、目の前に立ちはだかる高い壁と感じるよりも「読み進めれば、これだけのことが身に付けられるんだ」と期待していただくといいかと思います。なお、データ分析とは、データからその特徴を見つけ出したり、判断や予測を行ったりするための、分析の実践的な側面を表すもの、データサイエンスはさまざまな分野にまたがる理論と AI や機械学習などの応用技術までを表すものと（取りあえずは）考えておいてください。

なお、筆者が非常勤講師を務める一橋大学では 2023 年度にソーシャル･データサイエンス学部が新設され、大変な人気となっています（平均の志願倍率 3.2 倍に対し、ソーシャル・データサイエンス学部は 6.1 倍）。今後、このような知識や技能を身に付けた生徒や学生が社会の表舞台に立つことになります。それ以前に社会人となった私たちも、彼らとともに活躍するためには、新しい知識と技能を学んでおく必要があります。

この連載では、表 1 の太字で書かれた項目について少しずつ例を見ながら解説していきます（各回の内容については後でもう一度整理して掲載します）。太字で書かれた項目は、主に「記述統計」と呼ばれる分野で取り扱われる考え方や手法ですが、いずれも、データ分析を実践する上でも、AI ／機械学習の高度な手法を身に付ける上でも基礎となるものです。なお、この連載の目的や学ぶべき内容については、動画でも簡単に紹介しています。ぜひともご視聴ください（チャンネル登録・高評価もお待ちしています！）。

データ分析／データサイエンスが重要視されるのはなぜ？

では、データ分析やデータサイエンスがことさら重要視されるようになったのはなぜでしょう。これまで私たちが日常的に取り組んでいたデータ分析（例えば、図 1 のようなグラフを作成すること）とどう違うのでしょうか。学校教育のカリキュラムにまで反映せざるを得ないような、大きなブレークスルーがあったのでしょうか。

3V と呼ばれるキーワードがその大きな変化を端的に表しています。**3V** とは、いわゆるビッグデータの特徴を表す以下の 3 つの単語の頭文字です。

- **Volume**：量
- **Variety**：種類
- **Velocity**：速度

コンピュータやネットワークの性能（処理速度や記憶容量）が劇的に向上したことにより、多種多様なデータを大量に、かつリアルタイムで収集できるようになったのはもはや誰もが実感していることでしょう。しかし、その変化はおそらく想像をはるかに超えるものです。

ビッグデータ時代における情報量の計測に係る調査研究報告書（総務省）によると、とりわけ、音声や画像、センサーなどから得られたデータの量が指数関数的に増大していることが分かります。

私たちが日常的に想像するアンケートや売り上げデータの集計のような調査データ（図 2a）とは異なり、何千万件、何億件といった多種多様なデータ（図 2b）が刻々と集められ、おすすめ商品の表示や価格の予測などに活用されています。

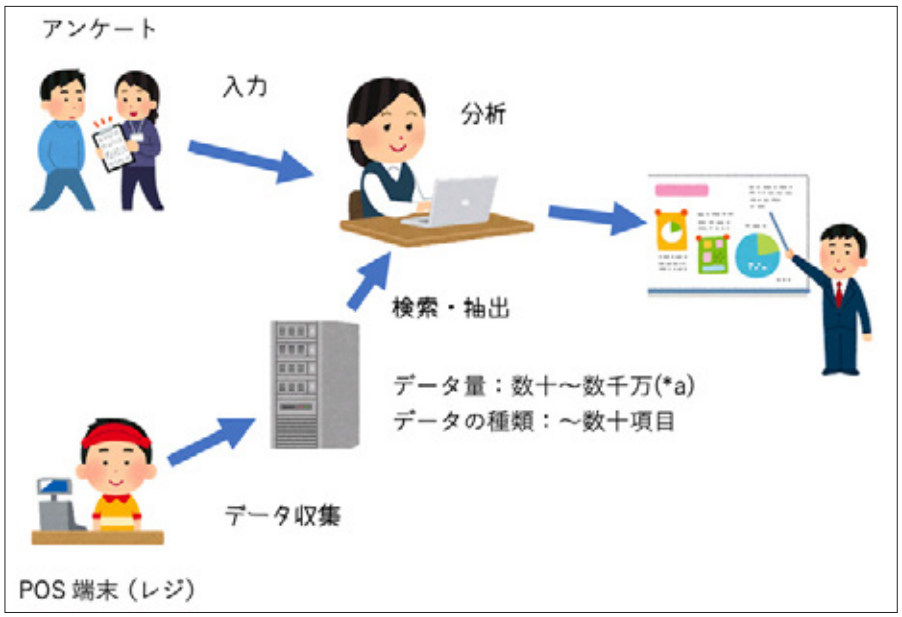


図 2a　これまでのデータ分析のイメージ

個別にデータを収集して分析が行われる。図中の（*a）は、1 万店舗のコンビニで、1 日当たり 700 人の顧客が平均 3 点の買い物をした場合を想定した値（2100 万件のデータとなる）。POS データの項目は、日時、店舗、商品名、個数、価格、ポイントカードの情報など 10 項目から多くても数十項目程度。

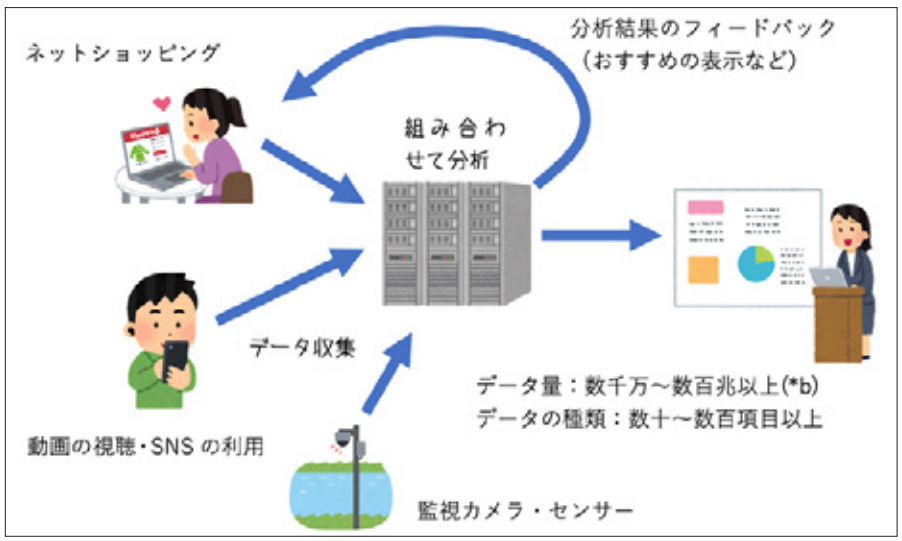


図 2b　ビッグデータ時代のデータ分析のイメージ

これまでより多種類、大量のデータが頻繁に収集される。図中の（*b）は 500 億個のセンサーからインターネット経由で収集されたデータを想定した場合の例。これらのデータを個別に分析するだけでなく、組み合わせて分析を行うこともできる。例えば、ネットショッピングである商品を購入した人に対して、その人と似たような動画を見ている人が購入した別の商品をおすすめに表示するなどのフィードバックも可能になる。

→目次に戻る

例えば、航空料金などの予測アプリを提供している Hopper という企業が公開している記事では、1 日あたり 1000 万件の検索と 10 億件の旅行データを基に、チケットの最安値を予測した結果が記されています。その記事では、一般に火曜日が最安値であると思われているにも関わらず、実際には国内線では木曜日が最安値である可能性が高いことや、ルートによる違いがあることなども紹介されています。

利用されたデータは **GDS（Global Distribution System）**と呼ばれる、世界規模の旅行関係の予約・発券システムのデータです。

このことからも分かるように、データ分析では、人が処理できる量や種類をはるかに超えたデータが対象になりつつあります。そういったデータを基に、より正確な予測が行えるというわけです。また、経験や勘を基に行われてきた予測よりも格段に精度の良い予測ができることも注目に値します。例えば、冬の降水量と育成時の平均気温、収穫期の降水量などを基にワインの品質を予測した結果が、専門家の予測よりも正確だった ***2** など、データ分析の力をセンセーショナルに伝える話は数多く知られています。

***2** 『その数学が戦略を決める』（イアン・エアーズ著、山形浩生訳、文春文庫）

人間の判断は、いかに正確・公正であろうと努めても、どうしても**認知バイアス**（歪み）の影響を受けてしまいます。例えば、私たちは自分の信念に都合のいい情報ばかりを集めてしまう傾向があります。このことは、すでに紀元前からカエサル（ジュリアス・シーザー）が “Libenter homines id, quod volunt, credunt.”（人は自らが欲するものを好んで信じる）***3** と述べている通りです。

***3** 『ガリア戦記』（カエサル著、近山金次訳、岩波文庫）

認知バイアスから逃れるのは容易ではありません。理屈では分かっていても、どうしてもそう思ってしまうのです。例えば、持ち株の株価が下がったときにロスカット（損失の拡大を防ぐために、早めに売却するなどして損失を確定すること）すべきだと分かっていても「いや、少し待てばまた上がるはずだ」という期待を持っていると、なかなかロスカットできないものです。そのうちさらに株価が下落して、売るに売れない「塩漬け」状態になってしまうことも少なくありません（筆者もいやというほど経験しています！）。そういった認知バイアスを持つのは熟練者や専門家であっても同じです。しかし、彼らは認知バイアスを回避するために、株価が何パーセント下落したら機械的にロスカットする、といったルールに従って取引を行っていたりします。それは勘や経験というよりも、むしろデータ分析によるものです。

コンピュータは認知バイアスを持たないので、冷静な判断ができます。上で述べたように、データ分析を**うまくやれば**専門家よりも正確な判断や予測ができます。といっても、もちろん経験や勘を否定しているわけではありません。何が目的なのか、何が問題なのか、分析に当たってどのような項目を盛り込めばいいのか……といったことについては、経験や勘がモノを言います。

データ分析やデータサイエンスがなぜ重視されるのかという問いに対する答えは、端的に言うと「正確な予測ができるから」ということになります。正確な予測は、例えば、疫病のまん延を防いだり、最適な治療法を探したり、より効果的な教育方法を実践したり、渋滞を減らしたり（それにより、燃料が節約できる）……と、私たちの生活や幸福度を向上させるために必要不可欠です。データ分析やデータサイエンスが、学校教育でも社会でも重視されるようになってきたのはそのためといっていいでしょう。

国際競争力を高めるため、ビジネスをより有利に進めるためといった理由ももちろん考えられます。学校教育において人材育成の必要性が叫ばれるのはそのためだとも言えるでしょう。しかし、競争に勝つことではなく、上で述べたような、よりよい生活や幸福の追求のためというのが、データ分析やデータサイエンスが重視される本質的な理由であると筆者は考えます。

一方で、与えるデータに偏りや誤りがあったり、適切な項目が選択されていなかったり、手法の適用方法を間違ったりすると、とんでもない予測が行われる危険もあります。さらには得られた結果だけが一人歩きして、誤った信念をより強固に植え付けてしまう可能性もあります。私たちが**データ分析やデータサイエンスを学ぶべき理由は、正確な予測を行うためだけでなく、取り返しのつかない誤りを防ぐためでもあります。**

この連載で取り扱う内容

ここまで、中学・高校で新たに学ぶデータ分析&データサイエンスの概要と、それらの重要性、学ぶべき理由について見てきました。そういったことを踏まえて、この連載ではまず記述統計を中心としたデータ分析の基礎を取り扱います（表 2）。これらは、表 1 で紹介した中学・高校の学習指導要領で取り上げられている内容のうち記述統計に関するものを網羅しており、さらに回帰分析やデータの取り扱いも含んだものとなっています。

回	テーマ	内容	キーワード	大分類	サンプルファイルの領域（ざっくりと）
1	**社会人が学ぶべき、やさしいデータ分析**	データ分析の必要性、身につけるべきスキル	高校情報Ⅱ、データ分析とデータサイエンス、記述統計、推測統計、AI・機械学習	**データ分析の全体像**	
2	**データ分析の進め方と、分析前に知っておきたいデータの種類**	分析の流れと分析手法、データの種類	構造化データ、非構造化データ、レコード、フィールド、定量的、定性的、尺度、データの収集方法（オープンデータ）		
3	**平均値の落とし穴 ～ 平均給与が高すぎる？！**	代表値の意味と読み解き方、尺度による代表値の使い分け	代表値（平均値、中央値、最頻値）	**記述統計（集団の特徴）**	人事・総務
4	**分散／標準偏差 ～ 給与の格差ってどれぐらい？**	散布度の意味と読み解き方、間隔尺度での散布度	散布度（分散、標準偏差）、外れ値、不偏分散と標本分散の違い、歪度、尖度		人事・総務
5	**四分位範囲と平均情報量 ～ 趣味や好みにはどれぐらいの幅があるのか？！**	順序尺度や分布に偏りのある間隔尺度の散布度、名義尺度の散布度	散布度（四分位数、四分位範囲、情報量、平均情報量）、箱ひげ図、外れ値		広告・マーケティング
6	**順位と偏差値 ～ 私の成績順位はどのあたり？**	順位や偏差値を求める	パーセント単位での順位、四分位順位、偏差値		人事
特別予告編	**グラフの種類と使い分け ～ データ可視化入門**	可視化の目的に適したグラフの種類		**記述統計（可視化）**	
7	**棒グラフで「規模や効果」を可視化 ～ どちらの広告が効果的なのか？**	規模や効果の差を可視化する	比較棒グラフ、目盛の取り方による落とし穴		広告・マーケティング
8	**折れ線グラフで「変化」を可視化 ～ 売り上げは本当に上がっているか？**	変化を可視化する	折れ線グラフ、移動平均、平均値への回帰、複合グラフ		人事・マーケティング
9	**円グラフやパレート図で「重要度」を可視化 ～ どの割合が本当に多いのか？**	重要度を可視化する／規模と割合の変化を可視化する	円グラフ・パレート図、積み上げ棒グラフ、3Dグラフに潜む落とし穴、ABC分析		総務・企画・生産
10	**ヒストグラムや箱ひげ図で「分布」を可視化 ～ 集団の特徴や外れ値を見つける**	分布を可視化する	度数分布表、ヒストグラム、箱ひげ図、バイオリン図、ピボットテーブル／ピボットグラフの活用		人事・総務
11	**クロス集計表やヒートマップで「分布」を多角的に可視化 ～ 項目同士の関連を見つける**	分布を多角的に可視化する	クロス集計表、ヒートマップ、クラスタリング		マーケティング・営業
12	**散布図を徹底活用して「関係」を可視化 ～ 関係と規模を一度に見る**	関係を可視化する	散布図、バブルチャート	**記述統計（関係）**	マーケティング・営業
13	**相関係数 ～ 気温と電気代に関係はあるのか？**	相関係数を求めて、直線的な関係の強さを知る	積率相関、スピアマンの順位相関、クラメールの連関係数、疑似相関、因果関係（とは異なる）		総務・マーケティング・営業
14	**単回帰分析による予測（線形回帰、指数回帰） ～ 排気量から中古車の価格を予測しよう**	単回帰分析による予測、指数回帰、時系列分析	単回帰分析、回帰式の可視化、説明変数、目的変数、係数、定数項、決定係数（R^2）、二乗平均平方根誤差（RMSE）、指数回帰、時系列分析	**回帰分析**	マーケティング・営業
15	**重回帰分析による予測（線形回帰、多項式回帰） ～ 年式、走行距離、排気量から中古車の価格を予測**	重回帰分析による予測、多項式回帰	重回帰分析、自由度調整済み決定係数、重回帰分析の落とし穴（多重共線性）、多項式回帰		マーケティング・営業
16	**データ分析に適したデータ形式に変換する方法と、表データを読み込む方法**	データの形式と変換方法、文字コードなどに関するトラブルシューティング	繰り返しのあるデータ、スタック形式とアンスタック形式、CSVファイル、UTF-8（BOMなし）とUTF-8（BOM付き）	**データの取り扱い**	全般

表２　連載の内容

データ分析・データサイエンスの基礎となる記述統計を中心に話を進める。さらに予測を行う方法や、データを適切に取り扱う方法についても紹介する。なお、テーマや内容、順序は連載の流れによって変更することもあるが、おおむね、この表にそって進めていくことにする。

→目次に戻る

いずれの内容についても具体例を基に、Excel などのソフトウェアを使いながらデータ分析の方法を見ていくので、考え方や手法がムリなく確実に身に付けられると思います。必要に応じて動画での解説も用意しています（今回は連載の紹介動画を用意しましたが、次回からは事例や操作も含んだ動画になります）。また、集中が途切れないように、1 回あたり 10 分程度で読めるようにします。

では、最後に少しだけ「展望」となるようなお話を付け加えておきます。具体的な内容に入るのは次回以降となるので、先の話をしても実感が湧かないかもしれませんが、ざっと読んで頭の隅にでも置いておいてください。私たちがこれから取り組んでいこうとしているデータ分析&データサイエンスは、大きく以下（図 3）の 3 つの領域に分けられます。相関や回帰分析など、さらに多くの領域に分けることもできますし、ここで言及していない分野や手法などもありますが、具体的な内容が分からないうちから細分化しても全体像がつかみにくいので、あえてざっくりと分けました。

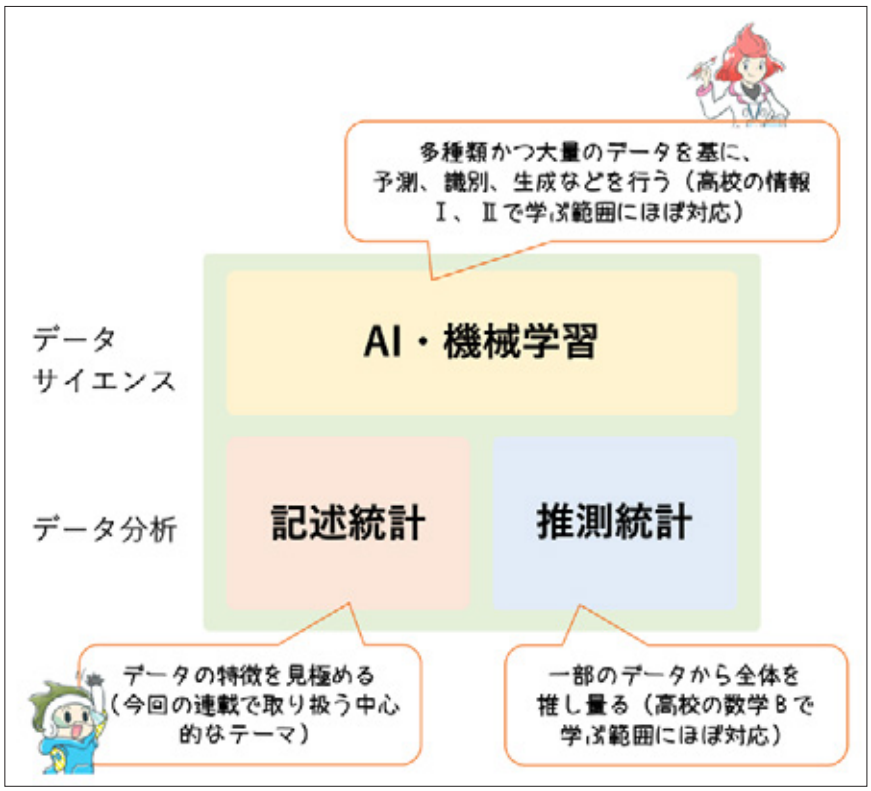


図 3　データ分析&データサイエンスの全体像

分類は大雑把なものだが、出発点と今後進むべき方向を示した。この連載では、データ分析やデータサイエンスに取り組むに当たって、まず、記述統計に当たる内容を主に取り扱う。推測統計や AI ／機械学習につながる基礎となっているので、一歩ずつ着実に理解を進めよう。

記述統計では、すでに得られたデータの**平均値**や**標準偏差**などを求めることにより、データの特徴を見極めます。単に 1 つのデータの特徴を見るだけでなく、複数のデータの関係を見たり、数字だけでは分かりにくい特徴を**可視化**したりすることによって明確にしていきます。記述統計は、次の推測統計や、さらには AI ／機械学習を理解する上での基礎にもなります。なお、図 3 には示していませんが、すでに述べたように回帰分析による予測についても、この連載で学びます。

→目次に戻る

推測統計では、一部のデータ（標本）を基に、その元となるデータ（母集団）の性質を「推し測る」方法を紹介します。そのためには、**確率分布**の知識が必要になります。つまり、「このようなデータであれば確率的にはこうなるはずだ」→「実際に得られたデータはこうだ」→「母集団はこのような分布だろう」といった推測を行うわけです。**仮説検定**を含む推測統計はデータ分析&データサイエンスのもう一つの柱となります。

AI／機械学習では、多種類かつ大量のデータを基に、今回少し触れたようなさまざまな数値予測を行ったり、識別、生成などを行ったりします。**識別**とは、例えば監視カメラに映った顔を区別したり、エックス線写真から病気を診断したりすることです。最近の話題としては、2023 年 3 月 18 日に開業した大阪駅（うめきたエリア）の顔認証改札もその一つです。**生成**とは、新しいものを作り出すことです。やはり最近話題の ChatGPT は私たちの質問に答えて、Web サイトやデータベースを検索し、適切な答えを作り出してくれます。顔写真を基に、有名な画家のタッチで油絵のような画像を作るサービスなども人気となっています。

今回の連載ではデータ分析&データサイエンスの基礎となる記述統計を主に扱いますが、その先の、推測統計や AI／機械学習についての連載も企画しています。が、あせらず、一歩ずつ基礎を固めていきましょう。というわけで、次回はデータの取り扱いについて見ていきたいと思います。データの種類、形式をきちんと見極めることはさまざまな手法を適切に利用するための大前提です。では、お楽しみに！

データ分析の進め方と、分析前に知っておきたいデータの種類

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 2 回。データ分析の流れを概観した後、取り扱うデータの種類について見ていきます。また、オープンデータを利用した簡単なデータ分析についても紹介します。

羽山博（2023 年 05 月 11 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 2 回です。前回はデータ分析とは何か、ビッグデータ時代の今、なぜデータ分析を学ぶべきなのかといったことを見た後、この連載で取り扱う内容を紹介しました。今回は、引き続き、データ分析の流れを概観した後、取り扱うデータの種類について見ていくこととします。また、オープンデータを利用した簡単なデータ分析についても紹介します。

データ分析の進め方はスパイラル！

データ分析をどのように進めるかについては、誰もがある程度は経験的に理解しているかと思います（図 1）。前回も似たような図を掲載しましたが、前回はデータ量（Volume）やデータの種類（Variety）、データが収集される速度（Velocity）に注目して、ビッグデータ時代のデータ分析の特徴を見ました。今回は、データ分析では何を行うのかという「流れ」に注目し、データ分析の進め方を見た後、取り扱うデータの種類について見ていきます。なお、以降の説明については、動画でも見られるようにしてあります。説明を丁寧に追いかけたい方は、ぜひご視聴ください。

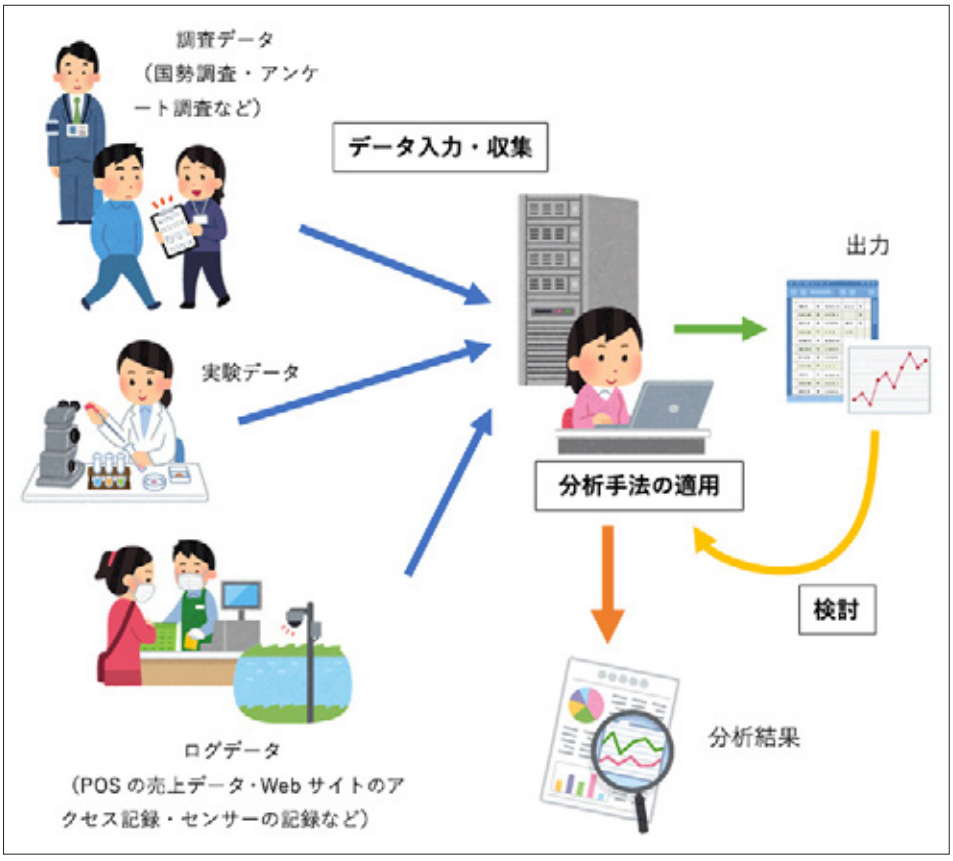


図 1　データ分析の進め方

データ分析の進め方をざっくりと図にしてみた。調査や実験などによって得られたデータをコンピュータに入力し、分析手法を適用して何らかの結果を得る。ただし、それで終わりではなく、出力された結果をさらに検討し、より詳細な分析を行う。このような作業を繰り返すことにより、最終的な分析結果を得る。なお、ログとは刻々と発生するデータを記録したもの。

図 1 で見たデータ分析の作業の流れを箇条書きにしておきます。

1. 調査や実験などから得られたデータを入力する
2. 分析の手法を適用し、データの特徴を表す値を求めたり、特徴を可視化したりする
3. 上の出力を検討し、より詳細な分析を行う必要があれば 2 に戻る
4. 最終的な分析結果としてまとめる

分析の手法については、次回以降、具体的に見ていきますが、「分析を行って結果が得られた。メデタシメデタシ」で終わるわけではありません。得られた結果から新たな疑問が浮かび上がってくることもよくあります。そのような場合には、他の項目に注目したり、別の手法を適用したりして、さらに分析を行います。必要であれば、追加の調査や実験を行うこともあります。つまり、スパイラル（らせん）のように繰り返しつつ、分析の精度を上げていくというわけです。

「らせん」とは、植物のつるのように巻き付きながら上がっていくような三次元の形のことで、英語ではヘリックス（helix）と言います。本来、蚊取り線香のような二次元の「渦巻き」がスパイラル（spiral）なのですが、繰り返しを行いながら精度を上げていくことの例えとして、日本語では「スパイラル」「スパイラル的」という表現が一般的に使われています。

コラム　データ分析へのアプローチ～仮説の探索と検証

実は、図 1 の中には記されていない重要なキーワードが 1 つあります。それは**仮説**です。

例えば、日々の活動や、店舗の責任者へのインタビューなどから「比較的、低価格の商品が売れているのではないか」とか「今後、より節約志向が高まる傾向にあるのではないか」といった仮説が得られます。データを眺めているときに何らかの気づきが得られることもあるでしょう。ちなみに、このようにして仮説を見つけ出す調査の方法を**仮説探索型**の調査と呼ぶことがあります。

仮説が立てられれば、データを集計したり、グラフによる可視化を行ったりすることによって、仮説が支持されるかどうか、詳細な分析を行います。このような調査・分析の方法は**仮説検証型**の調査と呼ばれます。

上でも述べたように、仮説が支持されると考えられる場合でも、仮説が支持されないと分かった場合でも、そこで分析が終わるわけではなく、さらなる仮説が立てられるのが普通です。例えば、分析を行っている中で、低価格の商品の売り上げの伸びは大きいが、一方で、高価格の商品の売り上げも伸びていることに気づいたとすれば「比較的購買力に余裕のある中・高齢者の利用も増えているのではないか」といった仮説も立てられます。このように、データ分析は仮説探索型と仮説検証型とに明確に分けられるのではなく、それらを繰り返しながら、より詳細な分析を行っていくことになります。

ただし、ビッグデータを取り扱う場合、日常の感覚だけでは見当が付かないことも多く、また、圧倒的なデータ量と項目数のために、分析に先立って仮説を立てることが難しい場合もあります。そのような場合には、取りあえず利用できる項目を使って、試行錯誤的にさまざまな分析を行い、何らかの特徴を発見するというアプローチも取られます。

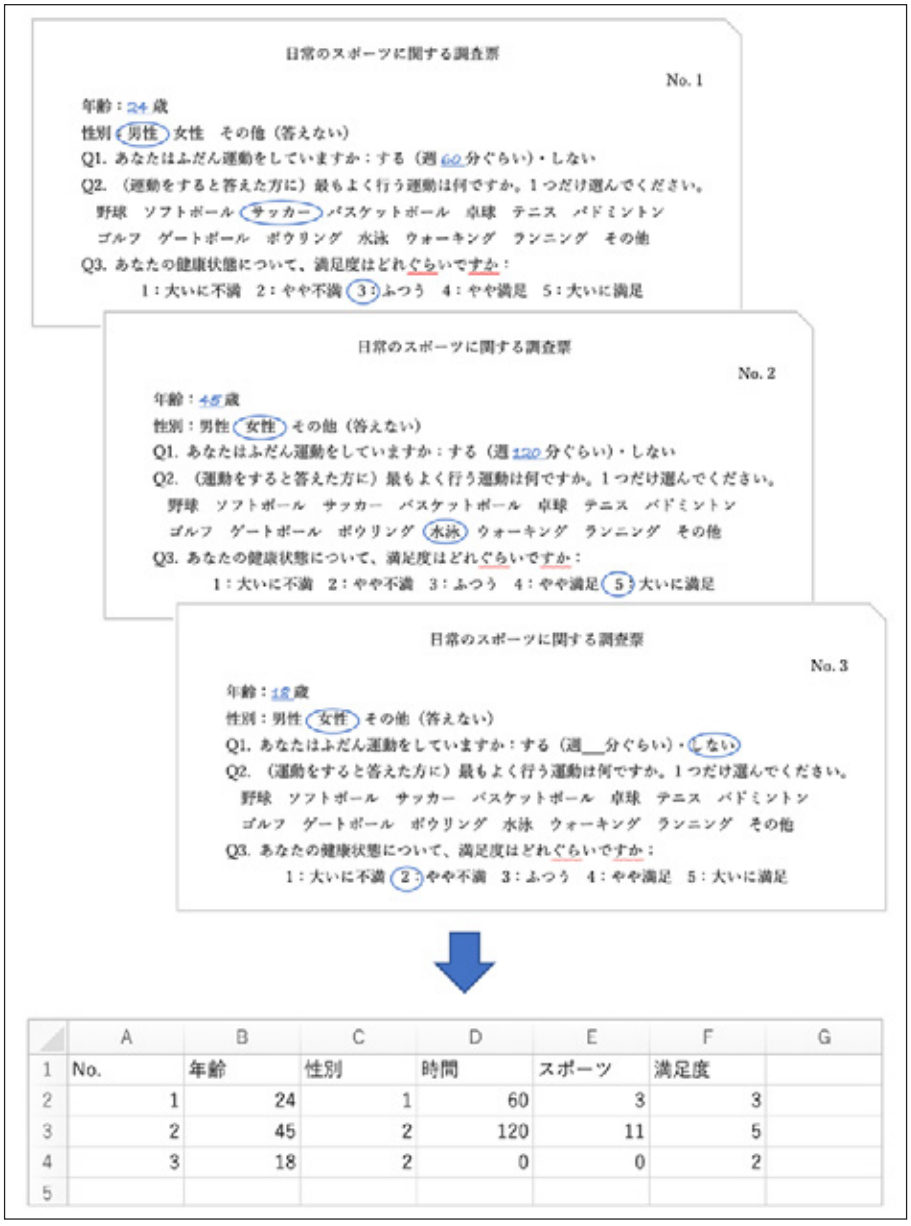
分析の前に知っておきたいデータの種類～構造化データと非構造化データ

続いて、データの入力・収集に当たって知っておくべきこととして、取り扱うデータの種類について詳しく見ていきます。

構造化データ ～ レコードとフィールド

一般に、データの種類は、構造化データと非構造化データに大きく分けることができます。

構造化データとは、文字通り構造が決まったデータのことです。端的に言うと、**レコード**と**フィールド**から成るデータです。レコードとは 1 件分のデータのことで、フィールドとはレコードに含まれる項目のことです。アンケートの調査票の例で見てみましょう（図 2）。



	A	B	C	D	E	F	G
1	No.	年齢	性別	時間	スポーツ	満足度	
2	1	24	1	60	3	3	
3	2	45	2	120	11	5	
4	3	18	2	0	0	2	
5							

図 2　構造化データの例

1 枚の調査票から得られるデータが 1 件分のデータ（＝レコード）となる。従って、このデータのレコード数は 3。それぞれの質問項目の値がフィールドに当たる。レコードには「No.」「年齢」「性別」「Q1（時間）」「Q2（スポーツ）」「Q3（満足度）」という 6 つのフィールドが含まれることも分かる。図の下は、このデータを Excel のワークシートに入力した例。Excel ではレコードやフィールドという用語は使われないが、1 レコードを 1 行に、各フィールドの値を各列に入力する。年齢や時間は数値をそのまま入力しているが、性別やスポーツなどについては、男性は 1、女性は 2 というように、区別するための値を決めて入力している。なお、Web ページに含まれる内容（項目）を表すデータも構造化データと呼ばれる（JSON-LD と呼ばれる記法で書かれる）が、この連載では、レコードとフィールドから成るデータを指すものとする。

→目次に戻る

なお、データベースでは、レコードのことを**タプル**と呼んだり、フィールドのことを**属性**と呼んだりすることもありますが、同じものと考えて差し支えありません。

定量的なデータと定性的なデータ

ところで、各項目をよく見ると、データには異なる種類のものがあることが分かります。一つは量を表す**定量的なデータ**で、もう一つは性質を表す**定性的なデータ**です。

では、どの項目が定量的なデータで、どの項目が定性的なデータであるか、ちょっと考えてみてください。なお、「No.」はレコードを識別するための番号なので、分析のためのデータ（変数）としては扱いません。

- 定量的なデータ　年齢、時間
- 定性的なデータ　性別、スポーツ、満足度

年齢や時間については、値の間隔が一定で、値の大小が意味を持っています。このようなデータは定量的なデータと考えられます。

一方、性別やスポーツについてはそうではありません。性別やスポーツも数値で表されていますが、単に区別するために与えたコードです。性別の **1**（男性）よりも **2**（女性）の方が大きいとか、スポーツの **1**（野球）よりも **2**（ソフトボール）の方が大きいということはありません。これらが定性的なデータであるということに疑問はないでしょう。

では、満足度はどうでしょう。満足度が **1** よりも **2** の方が大きいので定量的なデータと思われるかもしれません。しかし、値の間隔が一定であるとは限りません。例えば **1**（大いに不満）と **2**（やや不満）との間隔は、**2**（やや不満）と **3**（ふつう）との間隔は異なるかもしれません。**1** を付ける人は大きな不満があるので相当な低評価なのかもしれません。一方、**2** と **3** の間隔はそれほど大きくないかもしれません。というわけで、満足度は定性的なデータとなります（以下のコラムで説明する順序尺度に当たります）。

定量的なデータと定性的なデータでは、分析を行う際の取り扱いが異なります。まずは、データの種類について、違いをはっきりと認識しておくことが重要だということだけ理解しておいてください。

コラム　尺度によって分析方法が異なる

データの性質による値の表し方の違いは **尺度**（しゃくど）と呼ばれ、以下のように分けられます。

- **定量的なデータ**
  - **間隔尺度**：年齢や時間のように、値の間隔が一定で、値の大小が意味を持つもの。間隔尺度のうち、長さや重さのように原点 **0** があるもの（「ものさし」や「はかり」で測定できるようなものと考えてもよい）を、特に**比率尺度**と呼ぶこともある。
- **定性的なデータ**
  - **順序尺度**：満足度のように、大小が数値によって表されていても、数値の間隔が一定でないもの。つまり、単に順序を表しているだけのもの。
  - **名義尺度**：性別やスポーツのジャンルのように、大小が比較できないもの。

上で述べたように、尺度の違いによってデータの取り扱いや適用する分析の手法が異なります。そのことについては、それぞれの手法を見ていく中であらためて説明します。

前項の図 1 に示した調査データや実験データは、多くの場合、形式の決まった構造化データです。もちろん、次に述べる非構造化データが得られることもあります。

非構造化データ ～ 分析に適した形に変換する必要がある

非構造化データとは、文章や画像、サウンドなどのように、レコードとフィールドの形式で表されていないデータのことです。例えば、最近話題の ChatGPT では、文字で表された自然言語のテキストが入力データとなります。ログデータのうち、POS 端末から得られる売り上げデータは構造化データですが、カメラやセンサーから得られるデータは、画像やサウンドなどの非構造化データであることが少なくありません。この連載では、非構造化データはあまり扱いませんが、どのようなものか確認しておきましょう（図 3）。

図 3　非構造化データの例

上の例は、動物（ヒョウ）を撮影し、顔の部分を縦 100× 横 100（ピクセル＝点）の画像データにしたもの。各ピクセルの色は R（赤）、G（緑）、B（青）、α（不透明度）の度合いで表されている。このデータは行 × 列 × 色の 3 次元の配列の形式で表される。下の例は文章の一部。そのままでは詳しい分析ができないので、**形態素解析**と呼ばれる方法で、品詞に分解するなどの前処理を行っている。

→目次に戻る

非構造化データの場合、元のデータはレコードとフィールドで表されていません。そこで、分析に適した形式に変換して利用します。例えば、画像データであれば、各ピクセルを 3 次元の**配列**として取り扱えるようにします。ちなみに、多次元の配列のことを**テンソル**と呼ぶこともあります。

文章もやはりレコードとフィールドで表されていませんが、**形態素解析**と呼ばれる方法により、辞書に基づいて品詞に分解されます。図 3 の例であれば、1 単語のデータが 1 レコードとして 1 行で表され、それぞれの単語の品詞や活用形、読みなどがカンマやタブ文字などで区切られたフィールドとして表されます（構造化データとして取り扱えるように変換される）。

図 3 のデータは、いずれも Python のプログラムによって作成されたものです。ここではプログラムの詳細には触れませんが、興味のある方はこちらをご参照ください。形態素解析は「作って試そう！　ディープラーニング工作室」のプログラムを参考にしたもので、MeCab と呼ばれるライブラリと、UniDic と呼ばれる辞書を利用しています。

実践の第一歩として ～ ビッグデータ時代のオープンデータ利用

最後に、取り扱うデータについて、少し観点を変えて見てみましょう。

国勢調査で得られるデータと、商品の調査などのために取られるアンケートのデータには、大きな違いがあります。それは、一般に公開されるかそうでないかということです。一般に公開されているデータのことを**オープンデータ**と呼びます。なお、デジタル庁では「オープンデータ」を誰もがインターネットなどを通して容易に利用できるものと位置付けており、こちら（PDF）のように定義しています。オープンデータを利用したり、複数のオープンデータを組み合わせたりすることにより、新たな発見を得ることもデータ分析の醍醐味（だいごみ）の一つです。

調査の方法という観点からすると、全数調査であるか標本調査であるかということも違いとして挙げられます。**全数調査**では母集団全体について調査を行います。一方、**標本調査**では、母集団から標本を取り出して調査を行います。**母集団**とは調査対象全体のことで、**標本**（**サンプル**とも呼ばれます）とは母集団から取り出された一部のデータのことです。

国勢調査は、日本に住む全ての人が母集団であり、母集団全体について調査を行う全数調査です。

一方、ある店舗を利用する顧客に対して、商品の好みを調べる場合、顧客全員にアンケートを採るのは難しいので、標本調査を行うことになります。母集団はその店舗を利用している顧客全体で、店舗に訪れてアンケートに答えてくれた顧客が標本となります。

デジタル庁が運営している e-Gov データポータルから、さまざまなオープンデータにアクセスできるので、どのようなものがあるか眺めてみましょう（図 4）。

図 4　データカタログサイトのデータセット一覧

キーワードやカテゴリから利用したいデータが検索できる。画面の下の方には、データを提供している組織を選択して検索できる一覧などもある。なお、URL や画面は予告なく変更されることもあるので、これ以降の図が、実際とは異なる表示になる場合もあることに注意。

オープンデータの内容は、調査の概要やコード表などの資料的なものから、集計結果や生のデータなど多種多様です。ただし、公開されているからといって、自由に使っていいというわけではなく、利用に当たってはデータを公開している機関の利用規約とライセンスに従う必要があります。

初めてのオープンデータ活用

今回は「さわり」ということで、オープンデータを取得する方法や、ちょっとした分析例、よくある落とし穴の例を、ちらっと垣間見ることにしてみましょう。

今回のテーマはデータ分析の進め方を理解することと、データの種類や性質を知ることなので、グラフ作成の操作手順をステップごとに追いかけたり、分析手法を詳しく解説したりすることはしません。ここでは、どんなことができるかを確認していただくだけで十分です。

e-Gov データポータルでデータのダウンロード

例えば、キーワードとして「労働力調査 詳細集計」を入力して検索を行うと、キーワードに一致するデータの一覧が表示されます。その中の「労働力調査（詳細集計）_ 令和４年」をクリックすると、リンクが幾つか表示されます。［ダウンロード］をクリックして、Excel のブックをダウンロードしてみましょう（図 5）。

図５　データの詳細情報やデータを取得する

「労働力調査（詳細集計）_ 令和４年」を選んだところ。ここからデータをダウンロードしたり、データの詳細情報が取得したりできる。「2022 年 1 ～ 3 月期 ...］というリンクをクリックすると、データの詳細情報が表示され、ライセンスが「政府標準利用規約（第 2.0 版）」であることなどが分かる。利用規約の詳細については、こちら（PDF）を参照。

ダウンロードされたファイルを Excel で開くと、以下のような画面が表示されます（図 6）。

	13	14	15	16	17	18	19	20	21	22	23
4			詳細集計　第I-1表								
5			就業状態・新規就業者・転職者・現職の雇用形態についている理由・求職理由・前職の離職理由・								
6			失業期間・探している仕事の形態・就業希望の有無・非求職理由，年齢階級別15歳以上人口								
8	2022年1～3月	男女計	総数								
9	年齢階級		総数	15～24歳	15～19歳	20～24歳	25～34歳	25～29歳	30～34歳	35～44歳	35～39歳
10	就業状態，新規就業者・転職者，										
11	現職の雇用形態についている理由，求職理由，										
12	前職の離職理由，失業期間，探している仕事の形態，										
13	就業希望の有無，非求職理由，前職の離職時期		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
14	15歳以上人口	(1)	11038	1176	597	579	1285	629	656	1535	725
15	労働力人口	(2)	6855	543	116	427	1149	570	579	1352	637
16	就業者	(3)	6653	515	110	405	1104	544	560	1316	621
17	うち　新規就業者	(4)	370	143	59	84	73	43	30	48	24
18	うち　転職者	(5)	262	41	8	33	62	35	27	51	25
19	うち　非正規の職員・従業員	(6)	2073	263	84	179	231	116	115	320	143
20	現職の雇用形態についている理由（主な理由）										
21	自分の都合のよい時間に働きたいから	(7)	656	137	46	92	64	37	28	86	37
22	家計の補助・学費等を得たいから	(8)	400	53	20	32	26	9	17	62	22
23	家事・育児・介護等と両立しやすいから	(9)	207	2	0	1	27	7	20	69	35
24	通勤時間が短いから	(10)	90	12	5	7	7	4	3	13	8
25	専門的な技能等をいかせるから	(11)	174	10	3	7	21	10	11	17	6
26	正規の職員・従業員の仕事がないから	(12)	203	11	1	10	36	19	16	31	15
27	その他	(13)	273	34	8	26	40	25	15	32	15
28	失業者	(14)	202	29	7	22	45	26	19	36	16
29	（求職理由，前職の離職理由）										
30	仕事をやめたため求職	(15)	127	11	2	9	28	16	12	23	9
31	会社倒産・事業所閉鎖のため	(16)	13	0	0	－	2	1	1	3	1
32	人員整理・勧奨退職のため	(17)	12	1	－	1	2	1	1	2	0

図 6　取得したデータを表示する

詳細なデータが記録されているので、一部分のみ表示。ここでは内容の分析は行わないが、各階級の行や列の途中に合計の項目が設けられているので、利用に当たってはデータの加工が必要になる。出典：「労働力調査（詳細集計）_ 令和４年」（総務省）

統計ダッシュボードでデータのグラフ表示と分析

オープンデータを提供しているサイトがすでに分かっていれば、直接アクセスしてデータを取得するのが手っ取り早いですね。サイトによっては検索や加工のための便利なツールが用意されていることもあります。例えば、政府統計の総合窓口である e-Stat や統計ダッシュボードでは、データの検索だけでなく、グラフ化などを行うツールも用意されています。そこで、別の例でちょっとした分析も行ってみましょう。

図 7 は、統計ダッシュボードで「実質賃金」を検索して、「賃金指数」のデータを選択し、周期を「年」としたグラフと数値を示したものです。左上に表示されている［<］ボタン（データ変更ボタン）をクリックすれば、集計の周期や地域を変えることができます。

図 7　統計ダッシュボードを利用して実質賃金の推移を見る

実質賃金指数は 2020 年を 100 とした値。実質賃金が下っていることが分かる。左上の［数値］タブをクリックすると、グラフの基となる数値が一覧表示できる。また、右上にある［その他機能］をクリックするとデータを CSV ファイルとしてダウンロードすることもできる。出典：統計ダッシュボード

→目次に戻る

このように一定時間ごとに得られたデータのことを**時系列データ**と呼びます。図 7 のデータは単独で利用しても興味深いのですが、例えば正規労働者と非正規労働者の人数など、ほかのデータと組み合わせて考察を加えると、より深い分析ができます。例えば、統計ダッシュボードから「労働力調査」というキーワードで検索を行い、「正規・非正規の職員・従業員」のデータをダウンロードして Excel でグラフを作成すると図 8 のようになります。

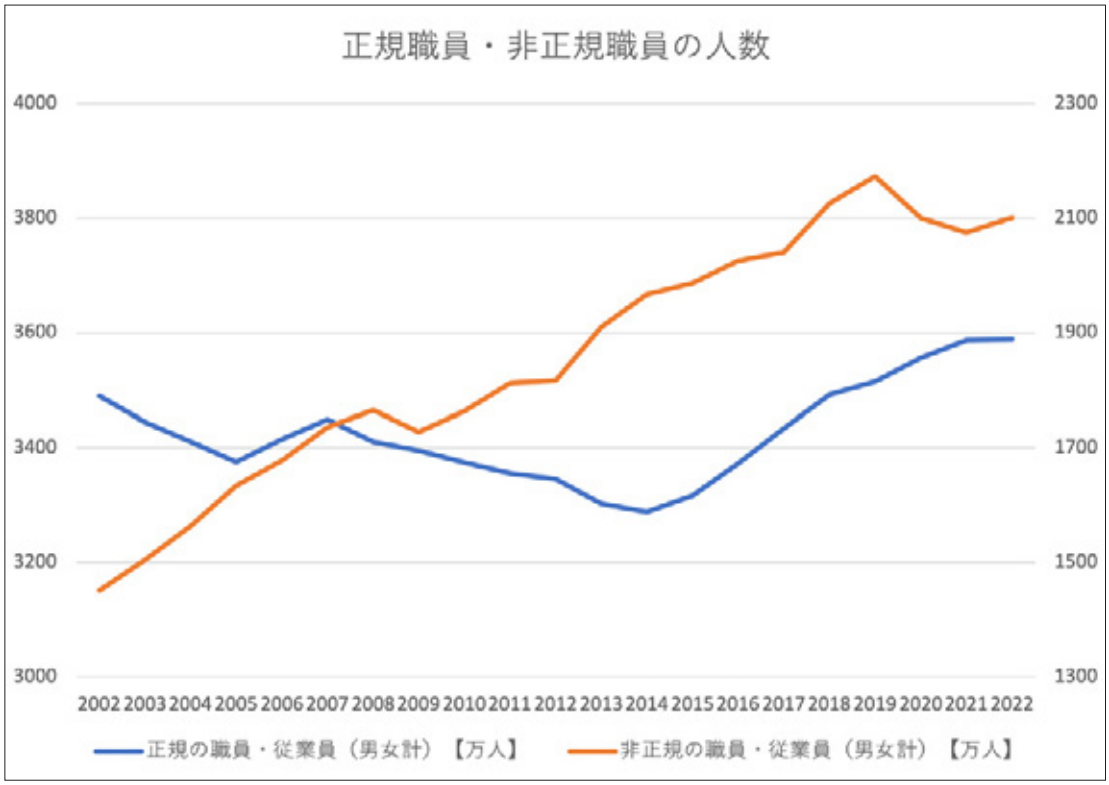


図 8　統計ダッシュボードを利用して正規労働者と非正規労働者の人数の推移を見る

左の目盛りが正規労働者の人数、右の目盛りが非正規労働者の人数。正規労働者は 2007 年～ 2014 年にかけてやや減った後、増加に転じている。非正規労働者は一貫して増加の傾向にある。統計ダッシュボードのデータを加工して作成。

実は、図 8 のグラフは非正規労働者の多さを印象付けるために、かなり恣意（しい）的に作ってあります。左の目盛りと右の目盛りとでは値の範囲が全く違いますね。一見すると、2007 年に非正規労働者の人数が正規労働者の人数を上回ったように見えます。しかし、目盛りをちゃんと見ると、（非正規労働者が増加しているとはいえ）全体としては正規労働者の方が多いことが分かります。

さらに、目盛りの間隔を変えると、人数の伸びを大きく見せたり、小さく見せたりできるので、自分の都合のいいように印象操作することもできます。もっとも、ここでは人数の変化を比較したいので、目盛りの間隔は統一しています（これに関しては良心的ですね）。確かに、データを見てみると、2002 年から 2022 年にかけて正規職員は約 100 万人増加しているのに対し、非正規職員は約 650 万人も増加していることが分かります。全体の人数については比較できないグラフですが、人数の変化については信用できるグラフだというわけです。

実質賃金のデータと正規／非正規労働者のデータを組み合わせることにより、実質賃金の低下の原因が非正規職員の急激な増加によるものではないかと推測できそうです。しかし、それが唯一の原因だと結論付けてしまうのはまだまだ早計です。実質賃金の低下や非正規職員の増加を引き起こしている真の原因がほかにあり、結果として、このようなデータが得られたのかもしれません。分析を行って何らかの特徴が発見できたとしても、いきなり結論に飛びつくのは危険です。ほかの要因やより本質的な原因などをさらに検討する必要があります。

今回はデータ分析の進め方と、データ分析を行う前に知っておきたいデータの種類についてお話した後、オープンデータを使ったデータ分析を「さわり」のレベルではありますが、ざっと紹介しました。しかし「さわり」だけでは物足りない、分析の方法を早く知りたいという方も多いでしょうから、次回以降は、統計量を求めたり、グラフによる可視化を行う方法など、データ分析の実践を少しずつ進めていきたいと思います。

分析に適した形式にデータ加工する方法についても、詳しく紹介したいところですが、一通り手法を学んだ後、連載の最後で改めて詳しく解説することにします。

というわけで、次回は、データの特徴を端的に表すために使われる代表値について、その求め方と性質、尺度による違い、落とし穴などを見ていきます。では、お楽しみに！

[データ分析]平均値の落とし穴 ～ 平均給与が高すぎる?!

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 3 回。分布の中心的な位置を表す値として代表値を取り上げ、尺度や分布によって適切な代表値を利用する必要があることを説明します。

羽山博（2023 年 06 月 01 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 3 回です。前回はデータ分析の進め方について見た後、オープンデータを利用した分析の方法を簡単に紹介しました。今回は、引き続き、分布の中心的な位置を表す値として代表値を取り上げます。代表値の求め方だけでなく、尺度や分布により、適切な代表値を利用する必要があることを説明します。なお、分布とはデータの散らばり具合、つまり、どのような値がどのような位置にどれだけあるかということです。

サンプルファイルの利用について

本稿では、表計算ソフトを使って手を動かしながら学んでいきます。そこで、表計算ソフト Microsoft Excel 用の .xlsx ファイルをダウンロードできるようにしています。デスクトップ版の Excel が手元にない場合は、Microsoft アカウントがあれば使える無料の Microsoft 365 オンライン、もしくは Google アカウントがあれば使える無料の Google スプレッドシート（Google Sheets）をお使いください。Microsoft 365 オンラインの場合は、.xlsx ファイルを OneDrive にアップロードしてから開いてください。Google スプレッドシートの場合は、.xlsx ファイルを Google ドライブにアップロードしてから開いた上で［ファイル］メニューの［Google スプレッドシートとして保存］を実行してください。

代表値として最もよく使われる平均値、でも万能ではない

私たちが小学校の算数や理科で学んだ**平均値**は、**算術平均**または**相加平均**と呼ばれるもので、全ての値を足して、個数で割れば求められます。しかし、Excel などの表計算ソフトではそういった計算を行わなくても、AVERAGE 関数を利用すれば簡単に平均値が求められます。

では、ウオーミングアップがてら、セル **B2 ～ B1001** に入力されている「勤め先収入」の平均値をセル **D2** に求めてみてください（図 1）。単位は万円です。サンプルファイル（03a.xlsx）はこちらからダウンロードできます。ここでは、Windows 11 上で、Microsoft 365（Excel 2019 以降）のデスクトップ版を使って説明を行いますが、Microsoft 365 オンライン版や Google スプレッドシートを利用する場合はこのページの上の「サンプルファイルの利用について」に示した方法でサンプルファイルを OneDrive や Google ドライブにアップロードしてからご利用ください。入力する関数はいずれも同じです。

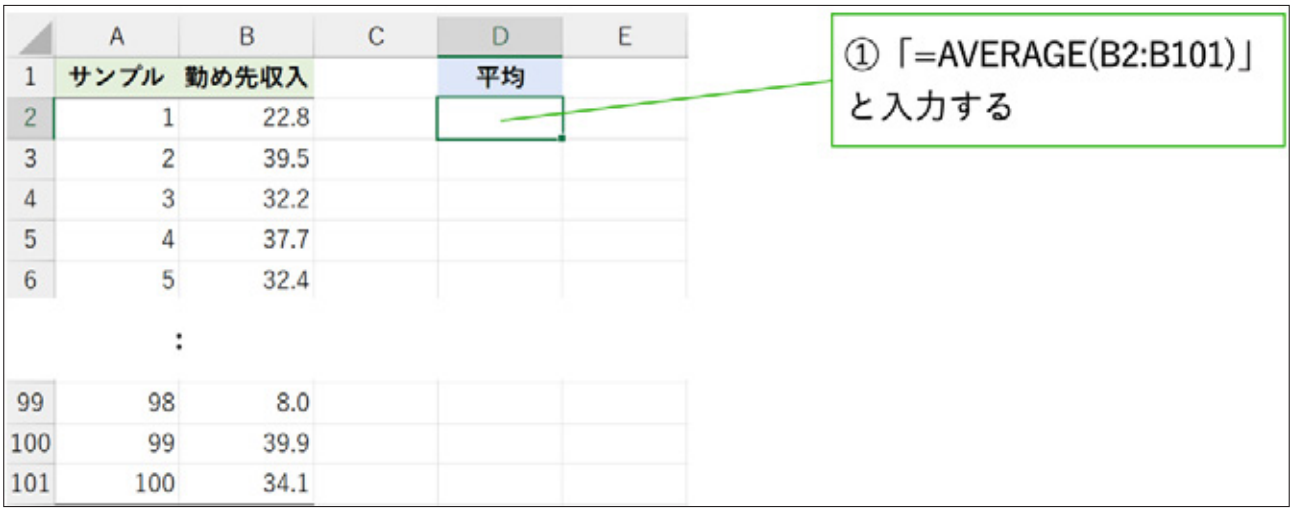


図 1　勤め先収入の平均値を求めてみる

セル **D2** に「=AVERAGE(B2:B101)」と入力して平均値を求めよう。データは架空のものだが、勤労者世帯の勤め先収入の平均値（月 49.2 万円）と一致するように作成してある。平均値の出典は、総務省統計局の家計調査の統計表に掲載された 2022 年の値（※左記のリンクをクリックすると s01.xlsx ファイルがダウンロードされます）。

答えは簡単、セル **D2** に「=AVERAGE(B2:B101)」と入力するだけです。基本の基本とも言える関数ですが、形式と説明をこの記事の最後にまとめておきます（以降、初出の関数名にリンクを設定しており、リンク先で形式と説明が見られるようにしておきます。リンク元に戻るには Web ブラウザーの［戻る］ボタンを使用してください）。

49.2 という結果が得られたでしょうか。ダウンロードしたファイルには答えのワークシートも含まれているのでそちらもご参照ください（具体的には、表計算ソフトの下部にある［平均値を求める（答え）］をクリックしてください）。

なお、具体的な操作については、これ以降の例も含めて動画で説明しています。データ量が **100 件**とかなり多いので、できるだけ効率よく操作を行いたいものですね。効率のよい方法を知りたい方は、ぜひご視聴ください。

さて、平均値を求めたところで、毎月の勤め先収入が**月 49.2 万円**もあるのは納得できないと思った方も多いのではないでしょうか（ま、そんなもんだろう、と思われる方もおられるかもしれませんが、少なくとも筆者はそんなにもらっていません……ちょっと話がそれますが、人それぞれに暗黙のうちに持っている基準と比較しているので、感じ方も人それぞれです。例えば、「月 12.3 万バーツ」だと言われると、バーツという通貨を日常的に使っていない人には高いのか安いのか判断できません。分析に当たっては、基準となる値や他の値と**比較することが重要**になってくるのですが、それについてはまた回を改めてお話します）。

話を元に戻しましょう。平均値は**代表値**としてよく使われる便利な値ですが、場合によっては実態を反映していないこともあります。実は、上で見た例では、収入が**月 100 万円**を超える人が 3 人いて、その中にはなんと**月 1,600 万円**という人もいます。このような**外れ値**と呼ばれる極端に大きな値（や小さな値）がデータに含まれていたり、**分布**（データの散らばり具合）に偏りがあると、平均値が代表値としてふさわしくないことがあります。

外れ値を見つけることはデータ分析においてとても重要なことです。上の例にある**1,600万円**という値に関しては、実際にそれだけの収入がある人がいるかもしれませんし、入力ミスなのかもしれません。分析の精度を上げるために、外れ値を除外して分析を行うこともあります。

ちなみに、TRIMMEAN 関数を使えば、上下合わせて何パーセントかの値を除外した平均値が求められます。例えば「=TRIMMEAN(B2:B101,10%)」と入力すれば、上位 **5%**と下位 **5%**を除外した平均値が求められます（**31.1 万円**になります）。

極端な値や分布の偏りに影響されにくい中央値

外れ値がある場合や、分布に偏りがある場合には、平均値の代わりに、中央値を代表値として使うこともよくあります。**中央値（メディアン）**とは、値を小さい順に並べたときに、ちょうど真ん中にある値のことです。データの件数が偶数個の場合、真ん中の値が 2 つありますが、そのときはそれらの値の平均値を中央値とします。

Excel などの表計算ソフトでは、MEDIAN 関数に値を指定すれば簡単に中央値が得られます。先ほどダウンロードしたサンプルファイル（03a.xlsx）のワークシート［中央値を求める］を開いてください。セル **D3** に「中央値」という見出しがあります。セル **D4** に MEDIAN 関数を入力してみましょう（図 2）。結果は **31.4** になるはずです。

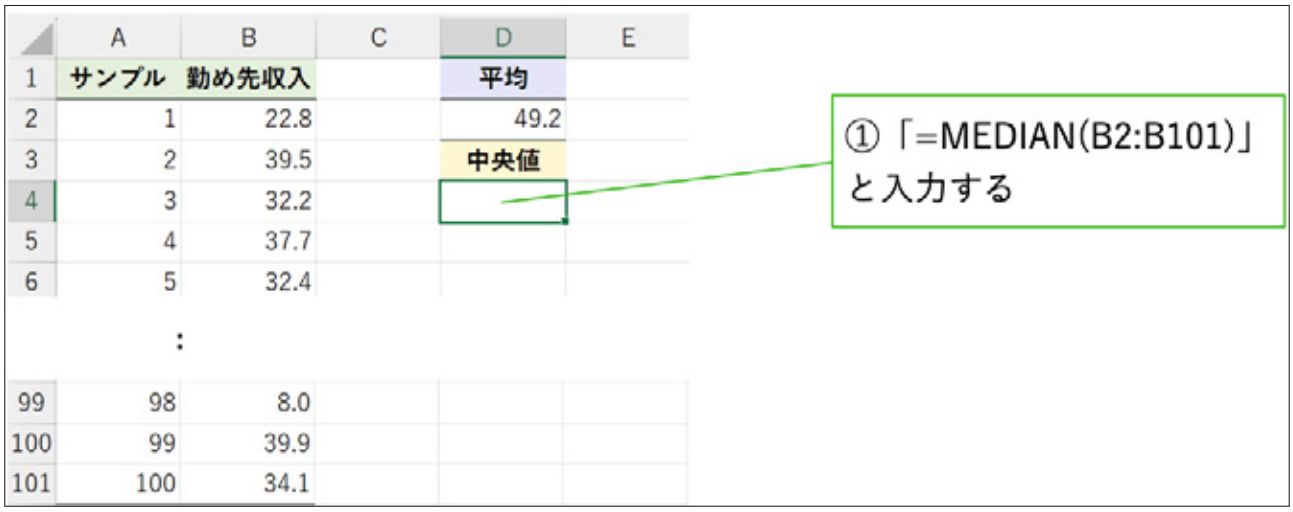


図 2　勤め先収入の中央値を求めてみる

図 1 と同じデータを使って中央値を求めてみよう。セル **D4** に「=MEDIAN(B2:B101)」と入力すればよい。この例では、中央値が平均値より小さくなる。

あくまで上のデータは架空データなので、ここで求めた値が勤労者世帯の実情を正しく反映しているわけではありませんが、一般に、給与や収入に関しては分布に偏りがあり、一部の大きな値に引きずられて平均値が大きくなっていると言われています。その場合、中央値＜平均値となります。実情はさておき、このデータに関する限り、平均値の **49.2 万円**より中央値の **31.4 万円**の方が代表値として納得できる値と言えるでしょう。

尺度によって代表値が異なる ～ 平均値／中央値／最頻値の使い分け

私たちは平均値にあまりにも慣れ親しんでいるので、何でも平均値を基準に考えてしまう「平均値信仰」とでも言ったようなものに取りつかれている傾向があります。しかし、平均値がアテにならないこともある、というのは上で見た通りです。

加えて、代表値として使える値は尺度によって異なるということも理解しておきましょう（表 1）。尺度については、前回解説しました。

尺度	利用できる代表値	データの例
間隔尺度、比率尺度	平均値	身長、体重、反応時間など
順序尺度	中央値	ランキングの順位、5段階評価など
名義尺度	最頻値	製品名、好きなスポーツの種類など

表 1　尺度と代表値

間隔尺度なら平均値、順序尺度なら中央値、名義尺度なら最頻値を代表値として使う。以下の説明にあるように、上の方に記した尺度では、下の方に示した代表値も使える。

今回の勤め先収入の例であれば、間隔尺度なので平均値を使えばいいということが分かります。基本的に表 1 の上の方に記した尺度では、下の方に記した代表値も使えます。例えば、分布に偏りのある間隔尺度のデータであれば、中央値や最頻値（後述します）が使えます。しかし、その逆はできません。例えば、名義尺度の代表値として平均値や中央値を使うことはできません。ただし、順序尺度の場合、本来は中央値または最頻値を使いますが、5 段階評価などの場合、分布に偏りがなければ、便宜的に間隔尺度と見なして平均値を使うこともあります。

最もよく現れる値も代表値として使える ～ 最頻値

では、代表値の 3 番手として登場した最頻値について見てみましょう。最頻値とは、最もよく現れる値のことです。

Excel などの表計算ソフトでは、最頻値は MODE.SNGL 関数または MODE.MULT 関数で求められます。最頻値が複数ある場合、MODE.SNGL 関数は最初に現れた最頻値を返しますが、MODE.MULT 関数は全ての最頻値を返します。

では、余暇に行うスポーツのデータを使ってセル **D4** に最頻値を求めてみてください。ここでは MODE.SNGL 関数を使うものとします（図 3）。サンプルファイル（03b.xlsx）はこちらからダウンロードできます。データはセル **B4 ～ B1003** に入力されています。

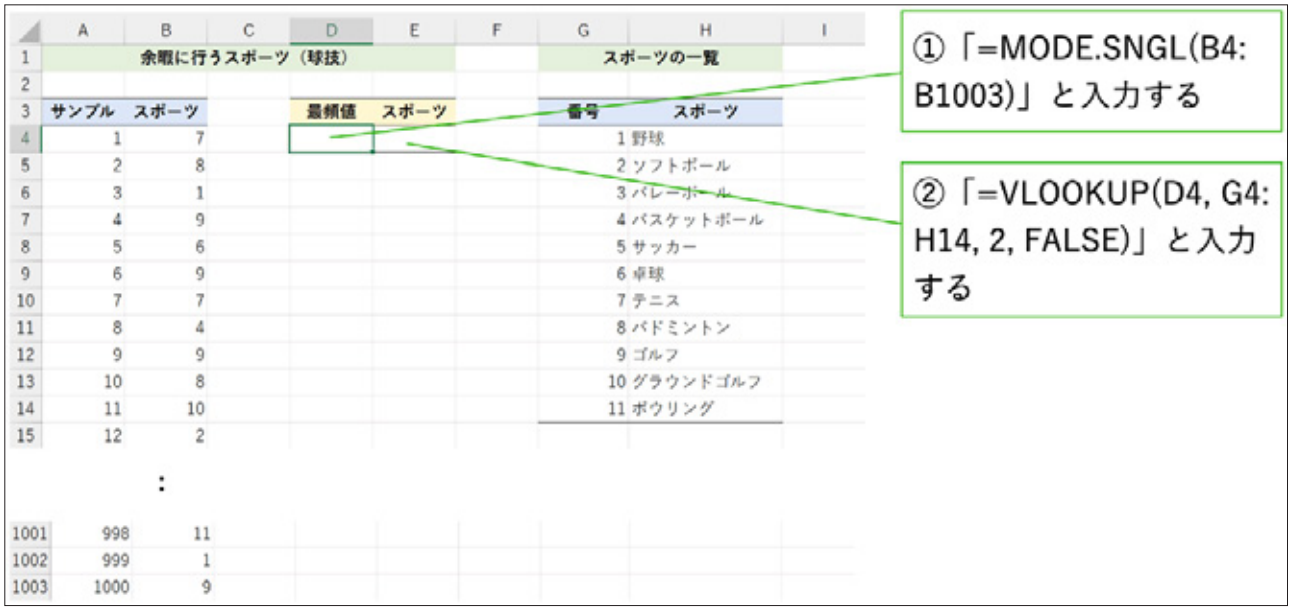


図 3　余暇に行うスポーツの最頻値を求めてみる

セル **D4** に「=MODE.SNGL(B4:B1003)」と入力して最頻値を求めてみよう。このデータは、総務省統計局の社会生活基本調査（2021 年）から球技だけを取り出して、同じ比率になるように加工したもの。スポーツの種類は［スポーツの一覧］表の［番号］列に数値で表されているが、この数値は単に種類を区別するためだけの、名義尺度のデータ。

最頻値として、**9** という値が得られれば正解です。

9 は「ゴルフ」を表しますが、結果が数字で表示されるだけだと分かりにくいので、スポーツの名前も表示できるようにしてみましょう。そのために、VLOOKUP 関数を使って、セル **E4** に「=VLOOKUP(D4,G4:H14,2,FALSE)」という式を入力します。もしくは Microsoft 365 で使える XLOOKUP 関数を使って、セル **E4** に「=XLOOKUP(D4,G4:G14,H4:H14,"",0,1)」と入力しても同じ結果が得られます。いずれの関数も、検索値（この場合は **9**）を基に表（この場合は［番号］列と［スポーツ］列を持つ［スポーツの一覧］表）を検索し、対応する値（この場合は「ゴルフ」）を取り出すためのものです。

余談ですが、上の調査は 5 年に 1 度行われており、前回の社会生活基本調査（2016 年）では、球技の最頻値はボウリングでした。ボウリングがゴルフに首位を明け渡したのは、2019 年以降のコロナ禍の影響で屋内でのスポーツができなくなったという要因があるのかもしれません（引用元のデータを見ると、全体的に屋内のスポーツが減少していることも分かります）。

度数分布表も作っておこう ～ COUNTIF 関数を使う

9 番の「ゴルフ」が最頻値であるということは分かりましたが、**1000 人**のうち、**9 番**と答えた人は何人いるのでしょうか。そこで、それぞれのスポーツについて、**度数**（データが幾つあるか）を一覧にした表を作ってみましょう。そのような表を度数分布表と呼びます。

この例では、条件付きで個数を数えるので、COUNTIF 関数を使います。引数には、データの範囲と条件を指定します。先ほどダウンロードしたサンプルファイル（03b.xlsx）のワークシート［度数分布表を作成する］を開いて、各［番号］に対する度数をセル **I4 ～ I14** に求めてみましょう（図 4）。なお、度数分布表の作成についても、動画で操作方法を説明しています。動画を見ながら 1 つ 1 つ操作を丁寧に追いかけたい方は、ぜひご視聴ください。

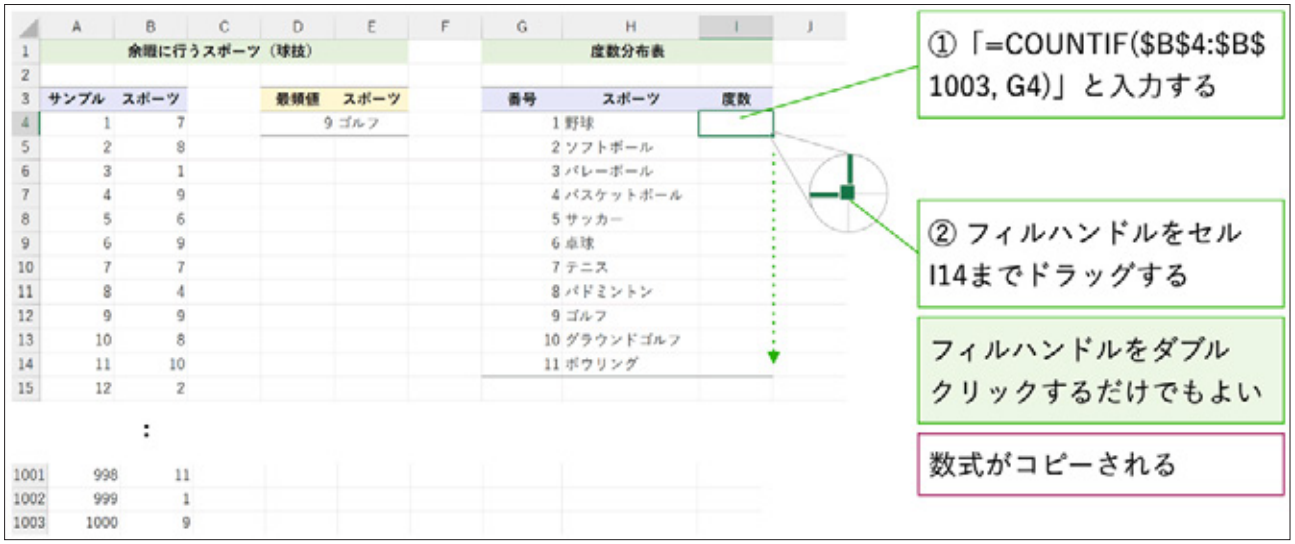


図 4　余暇に行うスポーツの度数分布表を作る

セル **I4** に「=COUNTIF($B$4:$B$1003,G4)」と入力すれば、セル **B4 ～ B1003** の範囲で、セル **G4** の値（**1**）と一致するデータの個数が求められる。セル **E4** の右下に表示されているフィルハンドル（小さな■）をセル **I14** までドラッグすれば、数式がコピーされ、全ての値が求められる。なお、フィルハンドルをダブルクリックすると、隣接するセルにデータが入力されているところまで数式がコピーされる。コピーする行数が多いときにはフィルハンドルのダブルクリックが便利。

データの範囲である B4:B1003 の列番号と行番号の頭に「$」を付け、「$B$4:$B$1003」としてあります。このように、列番号や行番号の頭に「$」を付けてセル参照を表す方法を**絶対参照**と呼びます。絶対参照の場合、数式をコピーしてもセル参照は変わりません。

一方、条件を表すために指定した「G4」のような（頭に $ を付けない）セル参照は相対参照と呼ばれます。**相対参照**の場合、数式をコピーするとコピーした方向に合わせて数式中のセル参照が変わります。この例では下方向にコピーするので、**G4** は **G5**、**G6**……と行番号が増えていきます（図 5）。

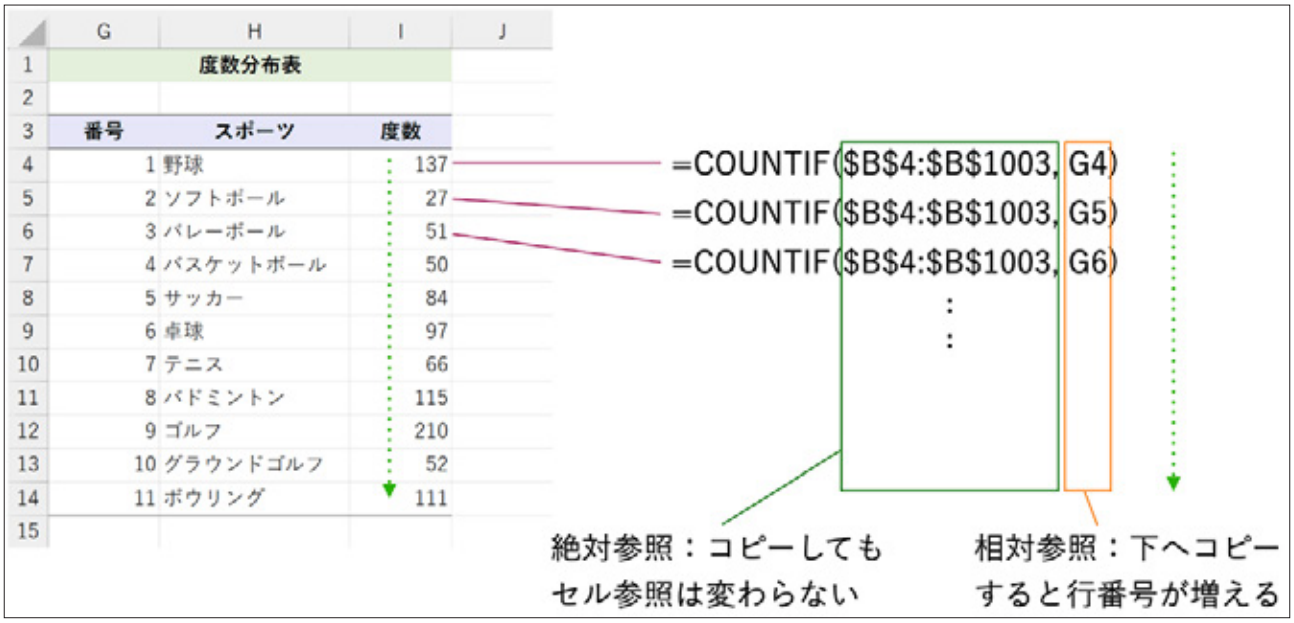


図 5　数式をコピーする（絶対参照と相対参照の違い）

絶対参照の場合、数式をコピーしても数式中のセル参照は変わらない。相対参照の場合、数式を下方向にコピーするとセル参照の行番号が増え、右方向にコピーすると列番号が増えていく。

「相対参照だとコピーしたときにセル参照が変わる」ということは、Excel の基本なので、すでにご存じの方も多いと思います。しかし、「コピーしたときにセル参照が変わるのが相対参照」と言うと、語弊があります。そもそも、相対参照とは**現在のセルから見てどの位置にあるか**というセル参照の表し方を意味します。

例えば、セル **I4** に入力した「=COUNTIF($B$4:$B$1003,G4)」に含まれる「G4」は、セル **I4** から見て「1 つ左の列で同じ行」ですね。そのセル参照を含んだ数式をセル **I5** にコピーすると「=COUNTIF($B$4:$B$1003,G5)」になりますが、この「G5」はやはりセル **I5** から見て「1 つ左の列で同じ行」です。コピーしたときに列番号や行番号が変わるのはあくまでも結果としてそうなるだけのことで、単に「1 つ左の列で同じ行」というセル参照がコピーされているだけなのです。

Excel のオプションで、数式の表示方法を R1C1 形式に変更すると、そのことがよく分かるのですが、話が本筋から外れてしまうので、これ以上は触れないことにします。

なお、Microsoft 365（Excel 2019 以降）には「スピル」と呼ばれる機能が備わっており、セル **I4** に「=COUNTIF(B4:B1003,G4:G14)」と入力して［Enter］キーを押すだけで（数式をコピーしなくても）、全ての結果が求められます。スピル機能の働きにより、1 つの数式だけで複数の結果が得られるというわけです。

→目次に戻る

Google スプレッドシートでは、セル **I4** に「=ArrayFormula(COUNTIF(B4:B1003,G4:G14))」と入力します。なお、Excel 2016 以前では、スピル機能が使えないので、図 5 のように数式をコピーするか、数式を配列数式として入力すれば、全ての結果が得られます。配列数式を入力するには、あらかじめ結果を表示したいセル範囲を選択しておき、（数式バーではなく）選択したセル範囲の最初のセルに対して「=COUNTIF(B4:B1003,G4:G14)」と数式を入力し、入力の終了時に［Enter］キーではなく、［Ctrl］＋［Shift］＋［Enter］キーを押します。ただし、この連載ではスピル機能が使えるものとして話を進めます。

間隔尺度の最頻値は度数分布表を作って求める ～ FREQUENCY 関数を使う

間隔尺度のデータでは、ほとんどの値が 1 回～数回しか現れないので、MODE.SNGL 関数や MODE.MULT 関数を使ってデータの個数を数えても最頻値は求められません。例えば、最初に見た勤め先収入の場合、**24.9** という値が最も多く現れますが、たったの 3 回だけです。

そこで、間隔尺度の場合は、データを一定の幅で区切って、その範囲に入る値の数を数えて度数分布表を作ります。その値の範囲を**階級**と呼びます。

もう少し正確に言うと、離散値（スポーツの種類を表す値や 5 段階評価のように、値が飛び飛びになっているもの）で、現れる値の種類が少ない場合には、MODE.SNGL 関数や MODE.MULT 関数を使って最頻値を求めます。一方、連続値（身長や体重など範囲内でどのような値でも取れるようなもの）で、現れる値の種類が多い場合には、度数分布表を作って最頻値を求めます。

では、勤め先収入の例で見てみましょう。度数分布表の作成には FREQUENCY 関数が便利です。サンプルファイル（03c.xlsx）はこちらからダウンロードできます。サンプルファイルを開いて、各階級の度数をセル **F6** ～ **F13** に求めてみましょう（図 6）。

	A	B	C	D	E	F	G
1	サンプル	勤め先収入		最小値	最大値	階級幅	
2	1	22.8		10	100	15	
3	2	39.5					
4	3	32.2		度数分布表			
5	4	37.7		より大	以下	度数	
6	5	32.4			10		
7	6	31.6		10	25		
8	7	30.1		25	40		
9	8	47.1		40	55		
10	9	12.8		55	70		
11	10	32.6		70	85		
12	11	42.6		85	100		
13	12	35.1		100			
14	13	13.8					

図 6　階級を設定して度数分布表を作る

セル **D6** ～セル **E13** で階級を設定している。ここでは、階級の幅を **15** とし、**10 以下**の小さな値と、**100 より大**きな値をひとまとめにした（階級数は **8** となる）。セル **F6** に「=FREQUENCY(B2:B101,E6:E12)」と入力すれば、度数分布表が作成できる。

FREQUENCY 関数には、データの範囲と階級の範囲を指定します。階級としては「以下」を表すデータの並びを指定します。ただし、最後の階級（セル **E13**）は指定しなくて構いません。スピル機能により、1 つの数式で全ての結果が求められます（図 7）。

	A	B	C	D	E	F	G
1	サンプル	勤め先収入		最小値	最大値	階級幅	
2	1	22.8		10	100	15	
3	2	39.5					
4	3	32.2		度数分布表			
5	4	37.7		より大	以下	度数	
6	5	32.4			10	3	
7	6	31.6		10	25	29	
8	7	30.1		25	40	46	
9	8	47.1		40	55	15	
10	9	12.8		55	70	2	
11	10	32.6		70	85	2	
12	11	42.6		85	100	0	
13	12	35.1		100		3	
14	13	13.8					

図 7　度数分布表から最頻値を読み取る

FREQUENCY 関数により度数分布表が作成された。**25 ～ 40** という階級の度数が **46** であることが分かる。最頻値は、階級値である（25 ＋ 40）÷2 ＝ 32.5 となる。

度数分布表から、**25 万円より大**きく、**40 万円以下**という階級に **46 人**いることが分かりました。この最も度数の大きい階級値を最頻値とします。階級値は階級の（下限＋上限）÷2 で求めます。つまり、（25 ＋ 40）÷2 ＝ 32.5 が最頻値となります。度数の **46** が最頻値ではないことに注意してください。

なお、度数分布表は COUNTIF 関数や COUNTIFS 関数を使ってデータの個数を数えることによって作成することもできます。サンプルファイル（03c.xlsx）のワークシート［度数分布表を作成する（COUNTIFS)］にはその例も含めてあります。

階級数の決め方 〜 スタージェスの公式

補足ですが、階級数を幾つにするかは、**スタージェスの公式**と呼ばれる以下の式で求めた値が目安になります。n はサンプルサイズ（得られたデータの個数）です。

$$1 + \log_2 n$$

上の例であれば、サンプルサイズが **100** なので 1 + log2100 ≈ 7.6 となります。Excel などの表計算ソフトで計算するなら LOG 関数を使って「=1+LOG(100,2)」という数式を入力すれば求められます。この値はあくまで目安ですが、図 6 や図 7 ではこの公式で求めた結果を基に、階級数を **8** としています。

コラム　どの代表値も使えない？！ ～ 複数の集団が混在している場合

代表値は分布の中心的な位置を表す便利な値ですが、半面、落とし穴もあります。すでに述べたように、外れ値がある場合や分布に偏りがある場合、平均値が代表値としてはふさわしくないといったことなどです。従って、代表値を求める前には、分布を見ることが重要です。

分布は度数分布表でも確認できますが、**ヒストグラム**（度数分布表をグラフ化したもの）を作成すると、その特徴がよく分かります。以下の図 8 の例は勤め先収入の特徴をもう少し細かく見るために、階級の幅を **5** としてヒストグラムにしたものです。グラフを使った可視化の方法については、回を改めて紹介するので、ここでは結果だけを示します。

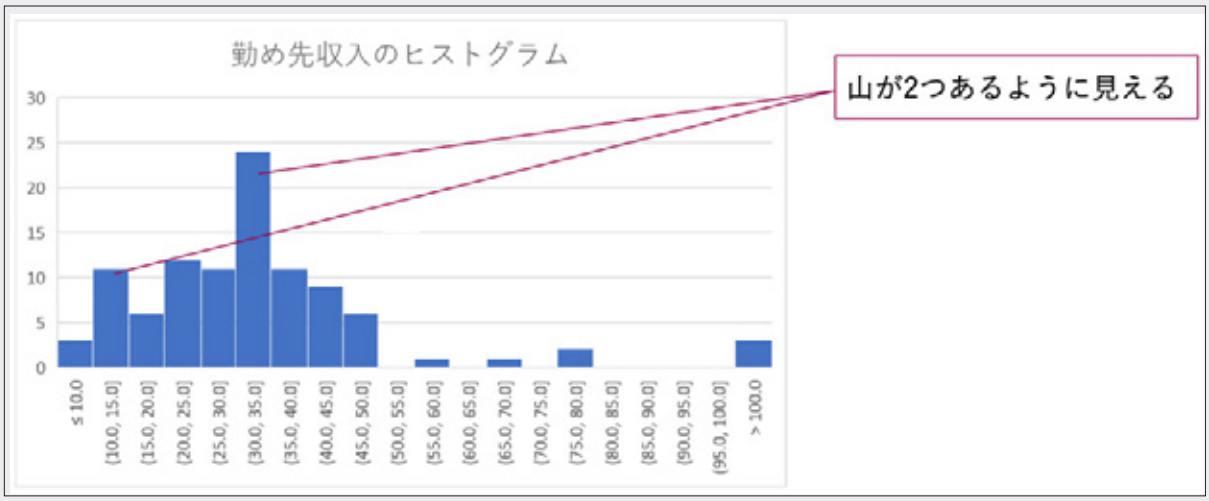


図 8　ヒストグラムにより度数分布表を可視化する

ヒストグラムは、棒グラフの棒同士の間隔を詰めたものと考えればよい。このグラフを見ると、**30 万円～ 35 万円**に大きな山があるが、**10 万円～ 15 万円**にも小さな山があるように見える。このような場合、複数の異なる集団のデータがまとめられている可能性がある。

実を言うと、このデータは、あえて **10 万円～ 15 万円**のところと、**30 万円～ 35 万円**のところに山ができるように作ったものです。つまり比較的収入の少ない集団と、そうでない集団とが混在しているというわけです。このような場合、それぞれの集団の分布に偏りがないとしても、それらの集団をまとめて求めた代表値は、全体を「代表」するのにふさわしい値ではないということになります。この例では、中央値が **31.4 万円**で、大きな山のある階級に含まれるので、問題はなさそうに思われますが、その数値だけに頼ると、収入の少ない人たちを見失いがちになります（あくまで架空のデータですが）。

極端な例も紹介しましょう。ある公園の利用者の平均年齢が **30 歳**だったとします。では、**30 歳**に近い人たちに合わせて公園を整備するのが最適な施策なのでしょうか。その公園の利用者が小学生と老人ばかりだったとすると、平均年齢は **30 歳**であっても、実際には **30 歳**に近い人はほとんどいないことになります。極端すぎて、そんな落とし穴にひっかかる人はいないだろうと思われるかもしれませんが、ターゲットを見誤ったために閑古鳥が鳴いている施設の例などを身近に見聞きしたことのある方も多いのではないでしょうか。

今回は、集団の代表値として利用される平均値、中央値、最頻値の意味や性質、分布や尺度による取り扱いの違いなどについて、Excel を使いながら具体例を見てきました。また、代表値を求めるに先立って、分布を見ることが重要であるというお話もしました。

次回は、集団の分布に関して、「散らばり具合」を表す値（分散／標準偏差、四分位範囲、平均情報量）を求める方法を尺度ごとに紹介します。では、次回もお楽しみに！

関数リファレンス：この記事で取り上げた関数の形式

関数の使いこなし方については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

代表値を求めたり、度数分布表を作ったりするために使った関数

AVERAGE 関数：平均値（算術平均）を求める

形式

AVERAGE(数値 1, 数値 2, ... , 数値 255)

引数

- **数値：**平均値を求めたい数値やセル範囲を指定する。引数は 255 個まで指定できる。

TRIMMEAN 関数：上下何パーセントかを除外した平均値（刈り込み平均）を求める

形式

TRIMMEAN(配列 , 除外する割合)

引数

- **配列：**平均値を求めたい数値の並びやセル範囲を指定する。
- **除外する割合：**上下合わせて何パーセントを除外したいかを指定する。例えば、10% を指定すると、上位 **5%**と下位 **5%**の値を除外した平均が求められる。10% の代わりに、0.1 と指定してもよい。

MEDIAN 関数：中央値を求める

形式

MEDIAN(数値 1, 数値 2, ... , 数値 255)

引数

- **数値：**中央値を求めたい数値やセル範囲を指定する。引数は 255 個まで指定できる。

MODE.SNGL 関数 , MODE.MULT 関数：最頻値を求める

形式

MODE.SNGL(数値 1, 数値 2, ... , 数値 254)

MODE.MULT(数値 1, 数値 2, ... , 数値 254)

※ MODE.SNGL 関数は最初に見つけた最頻値を返し、MODE.MULT 関数は全ての最頻値を返す。

引数

- **数値：**最頻値を求めたい数値やセル範囲を指定する。引数は 254 個まで指定できる。

COUNTIF 関数：1 つの条件に一致するデータの個数を数える

形式

COUNTIF(範囲 , 検索条件)

引数

- **範囲：**検索するデータの範囲を指定する。
- **検索条件：**検索条件を文字列として指定する。例えば、「">=10"」と指定すれば「**10 以上**」という意味になる。特定の値と一致するという条件であれば、値をそのまま書いてもよい。例えば、「"=10"」でも、単に「10」と指定しても「**10 と一致する**」という意味になる。

FREQUENCY 関数：度数分布表を作成する

形式

FREQUENCY(配列 , 区間配列)

引数

- **配列：**度数分布表の基となる数値の並びやセル範囲を指定する。
- **区間配列：** 各階級の上限を表す数値の並びやセル範囲を指定する。最後の階級の上限は指定しなくてもよい。

表を見やすくするために、または参考として取り上げた関数

VLOOKUP 関数：検索値を基に表を検索し、対応する値を取り出す

形式

VLOOKUP(検索値 , 範囲 , 列番号 , 検索方法)

引数

- **検索値：** 検索したい値を指定する。
- **範囲：** 検索値を検索する表の範囲を指定する。表の左端の列が検索される。
- **列番号：** 検索値が見つかったときに取り出したい値の列番号を指定する。
- **検索方法：** 以下の値を指定する。
  - **TRUE** または**省略** …… 近似値検索。近似値検索の場合、検索値以下の最大値が検索される。
  - **FALSE** …… 完全一致検索。

XLOOKUP 関数：検索値を基に表を検索し、対応する値を取り出す

形式

XLOOKUP(検索値 , 検索範囲 , 戻り範囲 , 見つからない場合の値 , 一致モード , 検索モード)

引数

- **検索値：** 検索したい値を指定する。
- **検索範囲：** 検索値を検索する範囲を指定する。
- **戻り範囲：** 検索値が見つかったときに、対応する値を取り出したい範囲を指定する。
- **見つからない場合の値：** 検索値が見つからなかったときに返す値を指定する。
- **一致モード：** 以下の値を指定する。
  - **0** または**省略** …… 完全一致検索。
  - **-1** …… 近似値検索。検索値以下の最大値が検索される。
  - **1** …… 近似値検索。検索値以上の最小値が検索される。
  - **2** …… ワイルドカード検索（検索文字列に含まれる * を任意の文字列、? を任意の 1 文字として検索する）。
- **検索方法：** 以下の値を指定する。
  - **1** または**省略** …… 先頭から検索する。
  - **-1** …… 末尾から検索する。
  - **2** …… 昇順に並べ替えられた範囲を二分検索する（効率のよい検索が行われる）。
  - **-2** …… 降順に並べ替えられた範囲を二分検索する（効率のよい検索が行われる）。

COUNTIFS 関数：複数の条件に一致するデータの個数を数える

形式

COUNTIF(範囲 1, 検索条件 1, 範囲 1, 検索条件 1, ..., 範囲 127, 検索条件 127)

引数

- **範囲：**検索するデータの範囲を指定する。
- **検索条件:** 検索条件を文字列として指定する。指定の方法は COUNTIF 関数と同じ。範囲と検索条件のペアは 127 個まで指定でき、全ての条件を満たしたデータの個数が返される。

LOG 関数：対数を求める

形式

LOG(数値 , 底)

引数

- **数値：**対数を求めたい値（真数）を指定する。
- **底：**対数の底を指定する。省略すると **10** が指定されたものと見なされる。

[データ分析] 分散／標準偏差～給与の格差ってどれぐらい？

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 4 回。分布のばらつきの度合いを表す値として散布度を取り上げ、尺度や分布によって適切な散布度を利用する必要があることを説明します。今回は間隔尺度・比率尺度の散布度として使われる分散／標準偏差のお話です。

羽山博（2023 年 06 月 15 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 4 回です。前回は集団の中心的な位置を表す代表値について、尺度や分布によって平均値、中央値、最頻値を使い分けることについて説明しました。今回は、集団の性質を表す値として、分布のばらつきの度合いを表す散布度を取り上げます。

やはり、尺度や分布により、分散／標準偏差、四分位範囲／四分位偏差、平均情報量／相対情報量を使い分ける必要があります。ただし、内容が少し多くなるので、今回は分散／標準偏差についてのみ見ていきます。**四分位範囲や平均情報量などについては次回取り扱います。**

サンプルファイルの利用について

本稿では、表計算ソフトを使って手を動かしながら学んでいきます。表計算ソフト Microsoft Excel 用の .xlsx ファイルをダウンロードできるようにしています。デスクトップ版の Excel が手元にない場合は、Microsoft アカウントがあれば使える無料の Microsoft 365 オンライン、もしくは Google アカウントがあれば使える無料の Google スプレッドシート（Google Sheets）をお使いください。Microsoft 365 オンラインの場合は、.xlsx ファイルを OneDrive にアップロードしてから開いてください。Google スプレッドシートの場合は、.xlsx ファイルを Google ドライブにアップロードしてから開いた上で［ファイル］メニューの［Google スプレッドシートとして保存］を実行してください。

平均値だけで集団の性質を表すのは「雑」すぎる ～ 散布度も見よう

平均値／中央値／最頻値といった代表値の意味や求め方については、前回、尺度ごとに詳しく見てきました。分布の違いによって、代表値の取り扱いには注意が必要であることも紹介しました。代表値はあくまで集団の中心的な位置を表すための値なので、集団の性質を表す値としてたった 1 つの代表値だけを使うのは、かなり「雑」だと言えます。

例えば、令和元（2019）年の国民健康・栄養調査（厚生労働省）の結果（Excel ファイルのダウンロード）では、20 歳以上の男性の平均身長は **167.7cm**、20 歳以上の女性の平均身長は **154.3cm** となっていますが、身長は人それぞれです。平均値に近い人も多いでしょうが、読売ジャイアンツの秋広選手のように身長 200cm の大柄な人もいれば、吉本新喜劇の池乃めだか師匠のように身長 149cm の小柄な人もいます。**167.7cm** という代表値だけでは、集団の姿はよく分かりませんね。

ちなみに、CDC（Centers for Disease Control and Prevention）の調査によれば、調査時期は 2015 年～ 2018 年と、日本人のデータとは異なりますが、アメリカ人の男性の平均身長は **69 インチ（≈ 175.3cm）**、女性の平均身長は **63.5 インチ（≈ 161.3cm）**でした。アメリカ人の方が日本人よりも総じて身長が高いと言えます。代表値だけで集団の姿を浮き彫りにするのは難しいですが、集団同士を比較することはできます。

代表値に頼り切るのではなく、分布を見ることによって、集団の姿がかなり分かってきます。分布を可視化するために使われるヒストグラムについても前回紹介しましたが、今回は、分布のばらつき具合を数値で表すことを考えてみたいと思います（今回も後でヒストグラムを掲載しますが、作成方法については回を改めて説明することとします。そのときのお楽しみということで……）。

分布のばらつきの度合いを表す値は**散布度**と呼ばれ、やはり、尺度によって利用できる値が異なります。そこで、尺度と、その尺度で使われる散布度をまず見ておきましょう（表 1）。前回のおさらいもかねて、代表値も合わせて示しておきました。

尺度	利用できる代表値	利用できる散布度	データの例
間隔尺度・比率尺度	平均値（算術平均）	分散／標準偏差	身長、体重、反応時間など
順序尺度	中央値	四分位範囲／四分位偏差	ランキングの順位、五段階評価など
名義尺度	最頻値	平均情報量／相対情報量	製品名、好きなスポーツの種類など

表 1　尺度と散布度

間隔尺度や比率尺度なら分散または標準偏差、順序尺度なら四分位範囲（または四分位偏差）、名義尺度なら平均情報量（または相対情報量）を散布度として使う。以下の説明にあるように、上の方に記した尺度では、下の方に記した散布度も使える。

基本的に表 1 の上の方に記した尺度では、下の方に記した散布度も使えます。例えば、間隔尺度や比率尺度では分散や標準偏差を使いますが、四分位範囲や四分位偏差を使うこともできます。特に、外れ値や分布に偏りがある場合には、そういったデータの影響を受けにくいというメリットがあります。しかし、その逆はできません。例えば、名義尺度の散布度として分散／標準偏差や四分位範囲などを使うことはできません。ただし、順序尺度の場合、本来は四分位範囲や四分位偏差を使いますが、五段階評価などで分布にあまり偏りがない場合は、便宜的に間隔尺度と見なして分散や標準偏差を使うこともあります。

間隔尺度の散布度を求める ～ まずは分散から

理屈は後回しにし、前回利用した「勤め先収入」に関連するデータを使って計算してみましょう。勤め先収入のデータは間隔尺度のデータなので、分散や標準偏差が使えます。そこで、まずは分散を Excel で求めることにします（標準偏差については後で見ます）。

サンプルファイルはこちらからダウンロードできます。今回はデータを 2 種類用意してあります。［勤め先収入 (1)］ワークシート（前回利用したもの）と、［勤め先収入 (2)］ワークシート（今回追加したもの）です。いずれも平均は **49.2** ですが、ばらつきが異なるものとなっています。

ではやってみましょう。分散は VAR.P 関数で求められます。引数にはデータの範囲を指定するだけです。セル **B2** ～セル **B101** に入力されている勤め先収入のデータを基に、分散をセル **E2** に求めてみてください（図 1）。それぞれのワークシートで分散の値を求め、比較してみましょう。具体的な操作については、標準偏差の求め方と合わせて動画で解説しているので、手順を丁寧に追いかけたい方はぜひご視聴ください（お約束のセリフですが、チャンネル登録・高評価もよろしくお願いします）。

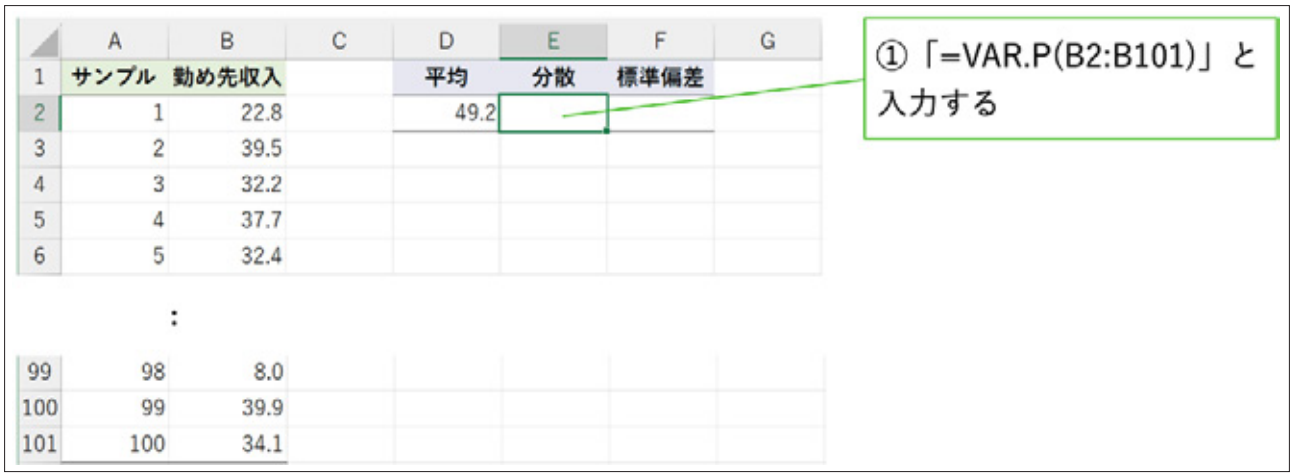


図 1　勤め先収入の分散を求めてみる

［勤め先収入 (1)］というワークシートを開き、セル **E2** に「=VAR.P(B2:B101)」と入力して分散を求めよう。［勤め先収入 (2)］についても同様。これらのデータは、総務省統計局の家計調査の統計表（Excel ファイルのダウンロード）に掲載された 2022 年の調査結果と平均値が同じになるようにしてある（ただし、分散や標準偏差は実際の値とは同じではない）。

答えは簡単、いずれのワークシートでもセル **E2** に「=VAR.P(B2:B101)」と入力するだけです。［勤め先収入 (1)］のデータでは、分散が **24824.67** となり、［勤め先収入 (2)］のデータでは **283.88** となるはずです。このことから、［勤め先収入 (1)］のばらつきの方が大きいことが分かります（が、後述するように、ちょっとした落とし穴もあります）。答えはサンプルデータの［勤め先収入 (1 の答え)］と［勤め先収入 (2 の答え)］というワークシートに含まれているので、そちらもご参照ください。

分布の違いが可視化できるように、ヒストグラムも掲載しておきます（図 2）。

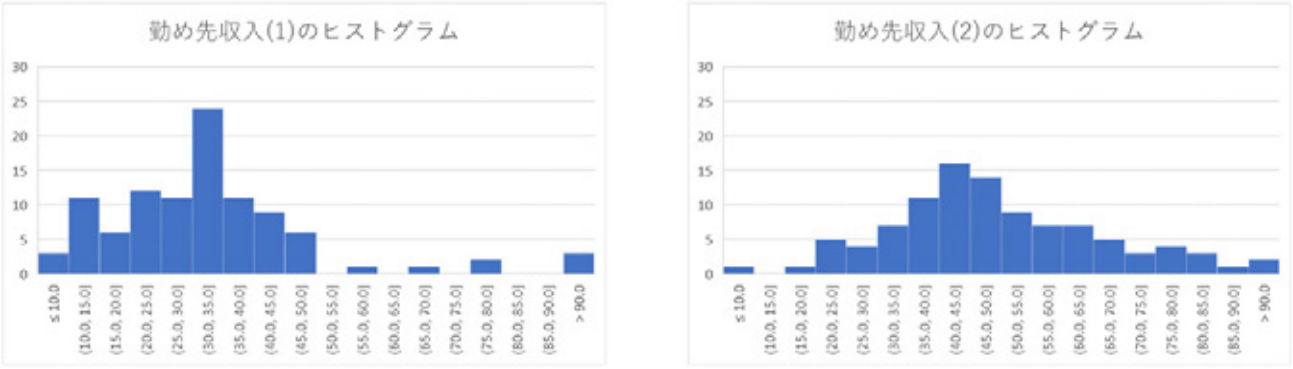


図 2　勤め先収入のヒストグラム

［勤め先収入 (1)］の分散の方が大きいのだが、ヒストグラムを見ると、平均値よりかなり小さい位置に値が集中している。分散が大きいのに、ばらつきがそれほど大きくないように見えるのは、**1600** などの極端な外れ値が含まれており、それが「90 より大」という階級にまとめられているため。［勤め先収入 (2)］は、平均値よりわずかに小さい位置に値が集中しており、なだらかに裾野が広がるような分布になっている。

［勤め先収入 (1)］の分散の方が大きいのに、値が集中して見え、［勤め先収入 (2)］の分散の方が小さいのに、広い範囲に値が散らばっているように見えます。これはいったいなぜでしょうか。すでに察しの付いた方もおられると思いますが、その犯人は「外れ値」です。前回もお話したように［勤め先収入 (1)］のデータには、外れ値と考えられる値が含まれています。

分散も極端な値があるとそれに引きずられることに注意が必要です。例えば、極端な値を上位から 3 つ（**1600、200、124**）を除外して分散を求めてみると **171.32** となり、［勤め先収入 (2)］よりもばらつきが小さくなります。つまり、比較的小さな階級に値が集中していることが分かります。分散が大きくても、ばらついていないこともあるわけです。これも外れ値による落とし穴ですね。

コラム　外れ値を検出するには

発展的なお話になりますが、外れ値の検出には、**スミルノフ・グラブス検定**などが使えます。［勤め先収入 (1)］の例では、1%水準で検定を行ったときに **1600、200、124** という 3 つの値が外れ値と見なされました。統計ソフトとしてよく使われている R（オープンソースで無料の統計解析向けプログラミング言語およびその開発実行環境）を利用すると、以下のように求められます。

```
> install.packages("outliers")  # outliers パッケージをインストール
> library(outliers)  # outliers パッケージを読み込む
> data <- c(22.8, 39.5, 32.2, ... 39.9, 34.1)  # 途中を省略してあるが全てのデータを指定する
> grubbs.test(data)  # スミルノフ・グラブス検定を行う（大きい方から 1 つ検出）

    Grubbs test for one outlier

data:  data
G = 9.793359, U = 0.021428, p-value < 2.2e-16
alternative hypothesis: highest value 1600 is an outlier

# 以下、外れ値と判断されたデータを除外して同じことを繰り返し行う
# 1%水準（p-value が 0.01 以下）の場合、1600、200、124 が外れ値と見なされる
```

R 言語のインタラクティブコンソールでスミルノフ・グラブス検定を行っている例

なお、これらの値を Excel で求めた例もサンプルファイルの［外れ値の検出］ワークシートに含めてあります。あくまで参考としてですが、サンプルファイル中には手順も記してあります。

ここで求めた **24824.67** や **283.88** などの分散の値を比較すれば、分布のばらつきの違いが分かります。しかし、値そのものが何を表しているのかが分からないですね。

分散や標準偏差の計算方法は後で説明しますが、分散の計算では途中で元の値を 2 乗するので、元の値の単位とは規模が異なる値になってしまいます。そこで、次に標準偏差の登場です。標準偏差は分散の√を取った値なので、元の値と同じ単位になります。

間隔尺度の散布度を求める ～ 標準偏差も求めよう

では、標準偏差を求めてみましょう。上で使った［勤め先収入 (1)］と［勤め先収入 (2)］のデータを使います。すでに分散を求めたのでその√を取っても構いませんが、ここでは標準偏差を求めるための STDEV.P 関数を使います。やはり、引数にはデータの範囲を指定するだけです。セル **B2** ～セル **B101** に入力されている勤め先収入のデータを基に、標準偏差をセル **F2** に求めてみてください（図 3）。それぞれのワークシートで標準偏差の値を求め、比較してみましょう。

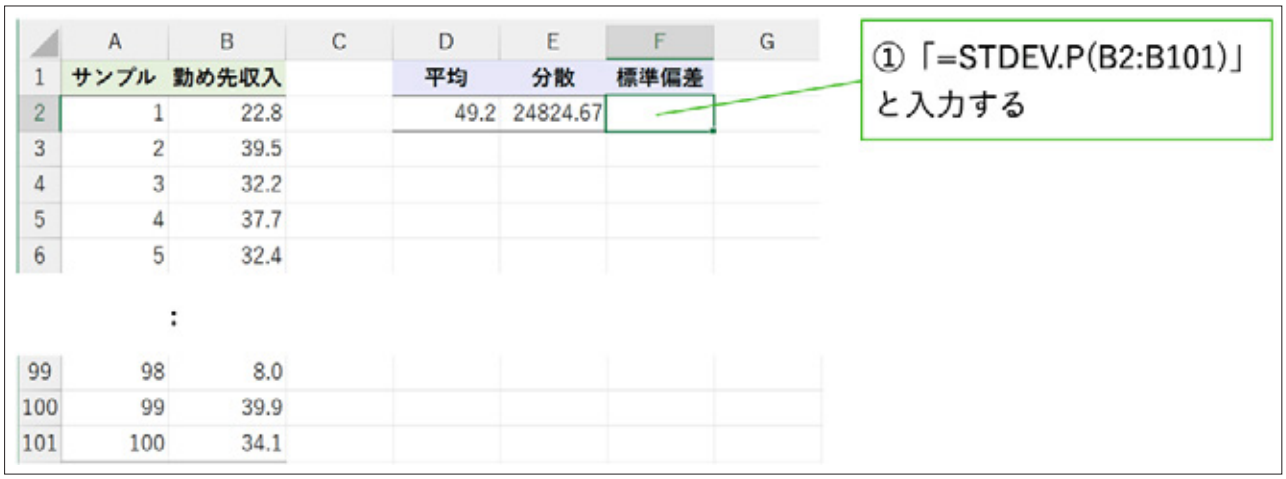


図 3　勤め先収入の標準偏差を求めてみる

［勤め先収入 (1)］というワークシートを開き、セル **F2** に「=STDEV.P(B2:B101)」と入力して標準偏差を求めよう。［勤め先収入 (2)］についても同様。

いずれのワークシートでもセル **F2** に「=STDEV.P(B2:B101)」と入力するだけです。［勤め先収入 (1)］のデータでは、標準偏差が **157.56**（万円）となり、［勤め先収入 (2)］データでは **16.85**（万円）となるはずです。当然のことながら、標準偏差からも［勤め先収入 (1)］のばらつきの方が大きいことが分かります。加えて、［勤め先収入 (1)］のばらつきが異常に大きいのではないかということも分かりますね。

分布のばらつきの度合いを表すのに、実感の湧きにくい分散なんて使わなくても、標準偏差だけで十分じゃないの、と思われる方もおられるかもしれません。しかし、統計のさまざまな計算の中で分散の値はよく使われます。分散が使われるのは、わざわざ√を求めず、分散の値のままにしておいた方が計算上便利なことも多いからです。

「ばらつきの度合い」とは平均値からどれぐらい離れているかということ

分散や標準偏差は、理屈を知らなくても、Excel の関数を使えば簡単に求めることができます。しかし、分布のばらつきの度合いを表す値として、なぜ分散や標準偏差が使われるのかを理解しておくことは重要です。そこで、VAR.P 関数や STDEV.P 関数を使わず、四則演算だけで分散や標準偏差を求めてみましょう。そうすれば、分散や標準偏差の意味がよく分かります。仕組みさえ分かればいいので、簡単なデータでやってみます。

サンプルファイルの［分散を手計算で］ワークシートを開いてください。手順は以下の通りです。ここでは分散を求める手順を見ていくことにします。

1. 各データと平均値の差を求めて 2 乗する
2. それらを全て足す
3. データの個数で割る

では、図 4 の流れに従って、数式を入力していってください。

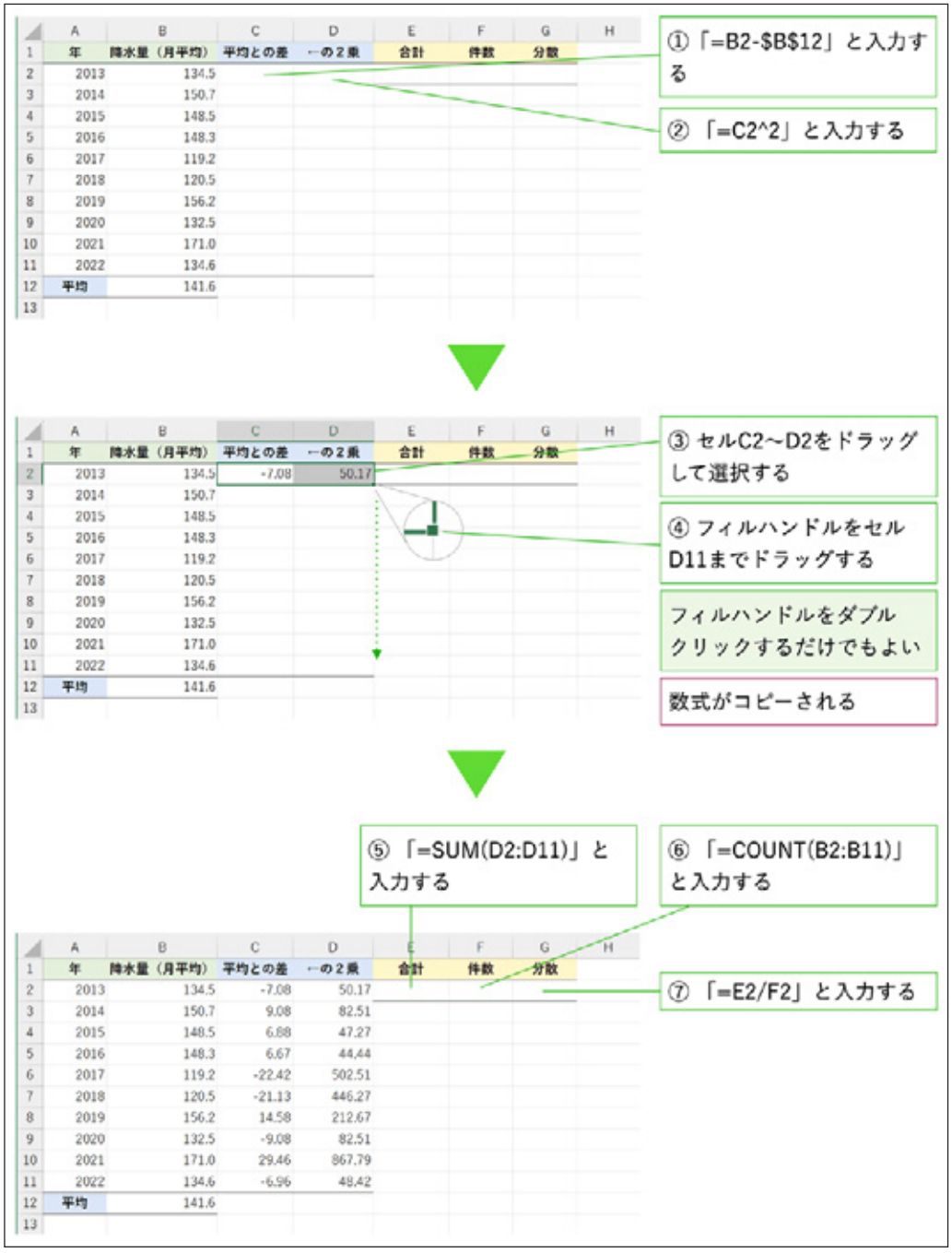


図 4　分散を手計算で求めてみる

手順通りに数式を入力すれば、分散が求められる。計算の意味については後述。なお、このデータは気象庁の「観測開始から毎月の値」から、2013 年～ 2022 年の年間降水量を取り出し、月平均の値にしたもの。単位は mm。

→目次に戻る

正しく数式が入力できたら、セル **G2** に **238.46** という値が表示されます。この値は VAR.P 関数を使って求められる値と一致します。空いているセルに「=VAR.P(B2:B11)」と入力して確認しておいてください。答えはサンプルファイルの［分数を手計算で（答え）］というワークシートに含まれているので、そちらも参照してみてください。

Microsoft 365（Excel 2019 以降）であれば、スピル機能を利用して、セル **C2** に「=B2:B11-B12」と入力し、セル **D2** に「=C2:C11^2」と入力すると、コピー操作をしなくても、セル **C2 ～ D11** の数式が全て入力できます。

Google スプレッドシートの場合は、配列数式を作成するための ARRAYFORMULA 関数を使って、セル **C2** に「=ARRAYFORMULA(B2:B11-B12)」と入力し、セル **D2** に「=ARRAYFORMULA(C2:C11^2)」と入力します。ただし、サンプルファイルを Google ドライブにアップロードすると、［分数を手計算で（答え）］ワークシートのセル **D2** には「=ARRAY_CONSTRAIN(ARRAYFORMULA(C2:C11^2), 10, 1)」という式が表示されます。この ARRAY_CONSTRAIN 関数は配列の一部だけを表示したいときに使う関数です。この例では、結果を全て求めたいので、特に ARRAY_CONSTRAIN 関数を使う必要はありません。今後、この連載では必要のない限り「=ARRAY_CONSTRAIN(ARRAYFORMULA(C2:C11^2), 10, 1)」のように入力されていても、配列数式を表す部分、つまり「=ARRAYFORMULA(C2:C11^2)」だけを掲載することにします。

では、計算の意味を考えてみましょう。「各データと平均値の差」は、それぞれのデータが平均からどれだけ離れているかということです。それらを合計すれば、各データと平均値がどれだけ離れているかの合計が求められます。しかし、「各データと平均値の差」は正になることも負になることもあるので、単純に合計すると正と負が相殺されてしまいます。そこで、合計を求めるに当たって、絶対値にしておきます。そのために、「各データと平均値の差」を 2 乗しています。そうすれば、全てが正の値になりますね（負の値の 2 乗は正の値になります。後で√を求めれば絶対値にできます）。

「各データと平均値の差」を 2 乗した値の合計を求め、データの個数で割ると、**各データと平均値がどれだけ離れているかの平均**が求められます。つまり、**分散**とは、**各データが平均値から平均的にどれだけ離れているかという値**であるということです。**標準偏差**は分散の√を取って、2 乗する前の単位に戻した値であることも分かります。

分散や標準偏差をイメージで理解しよう

うーん、まだ実感が湧かないな、という方は、以下のようなたとえ話でイメージをつかんでもらうといいのではないかと思います。一般に、公立の小学校には地域の子供たちが通っているので、小学校と各家庭との距離のばらつきは小さいですね。それらの距離（を 2 乗したもの）の平均が分散です。一方、大学には、近くの寮から通っている学生もいますが、遠距離から新幹線通学している学生もいます。つまり、大学と各家庭や寮などとの距離のばらつきの度合いは大きくなります。分散が大きくなることも納得できると思います。

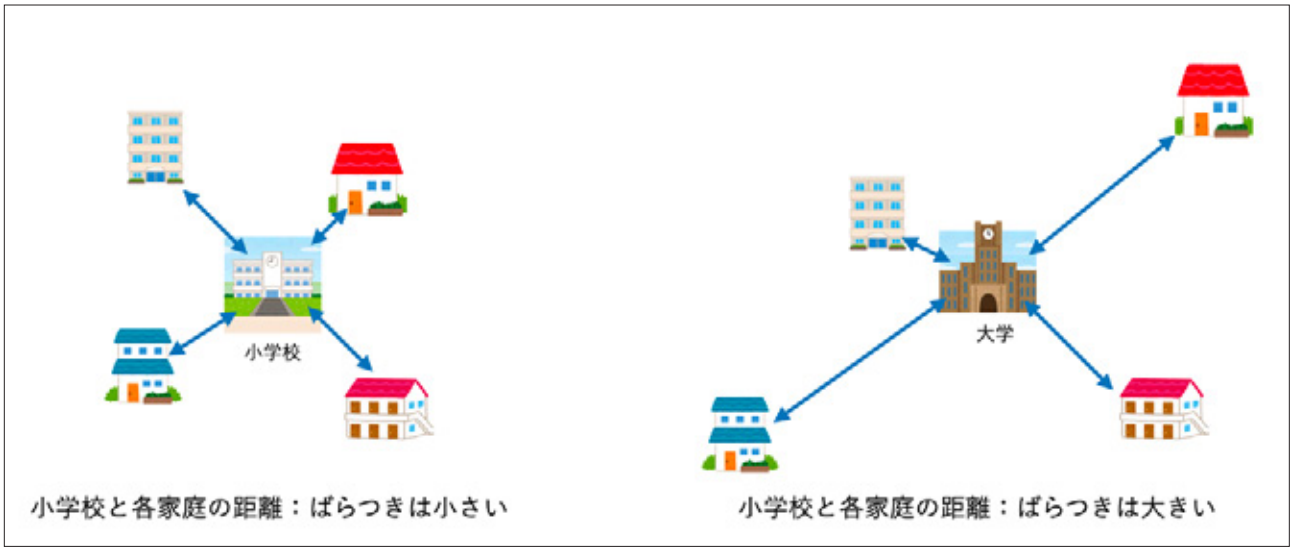


図 5　分散のイメージ

学校と各家庭との距離（の 2 乗）の平均が、ばらつきの度合いを表す。公立の小学校であれば、比較的ばらつきが小さい。大学であれば、ばらつきは大きくなる。

分散や標準偏差を数式できちんと理解しよう

さらに、分散を求めるための数式も見ておきましょう。標準偏差を s とし、分散は s^2 とすれば、図 6 のような式になります。数式が苦手な方はスルーしてもらっても構いませんが、毎度毎度「各データから平均値を引いたものを 2 乗し、それらの合計を求めて、データの個数で割る」と表現するのは面倒ですよね。数式はそういったことをできるだけ簡潔に表したものです。むしろ、面倒な計算手順を簡単に書けるだけでなく、一目見ただけでも意味が分かる便利な表し方なのだと考えてもらうと、数式がぐっと身近に感じられるのではないかと思います。

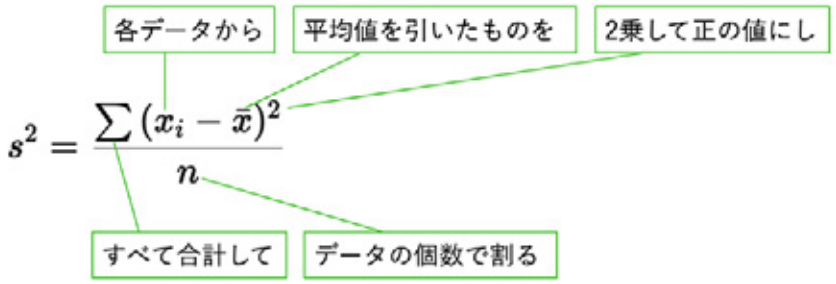


図 6　分散を求めるための式

x_i は各データ、$\bar{x}$ は平均値、n はデータの個数です。$\sum$ は合計するという意味なので、この式は、各データから平均値を引いたものを2乗し、それらの合計を求めて、データの個数で割るということになる。

→目次に戻る

コラム　その「ばらつき」って、どの「ばらつき」?! ～ 2 種類の分散と標準偏差

Excel の関数について調べたことのある方には、分散を求める関数には VAR.P 関数と VAR.S 関数があり、標準偏差を求める関数には STDEV.P 関数と STDEV.S 関数があることをご存じの方も多いと思います。これらの関数で求められる値や計算方法は表 2 のようにまとめられます。が、いったい、どう違うのでしょうか。

関数	求められる値	意味	計算方法
VAR.P	標本分散	標本そのものの分散	$\frac{\sum(x_i-\bar{x})^2}{n}$
VAR.S	不偏分散	標本から母集団の分散を推定した値	$\frac{\sum(x_i-\bar{x})^2}{n-1}$
STDEV.P	標本標準偏差	標本そのものの標準偏差	$\sqrt{\frac{\sum(x_i-\bar{x})^2}{n}}$
STDEV.S	不偏標準偏差	標本から母集団の標準偏差を推定した値	$\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$

表 2　標本分散／標本標準偏差と不偏分散／不偏標準偏差の違い

標本分散とは、取り出した値（標本やサンプルと呼ばれる）そのものの分散を表す。一方の不偏分散は取り出した標本を基に、母集団（全体）の分散を推定した値。分母が n − 1 になっているので、標本分散よりも不偏分散の方が少し値が大きくなる。標準偏差についても同じ考え方。

上の表の説明にもあるように、**標本分散**は標本そのものの分散を表す値です。一方、**不偏分散**は標本を基に母集団の分散を推定した値です（図 7）。

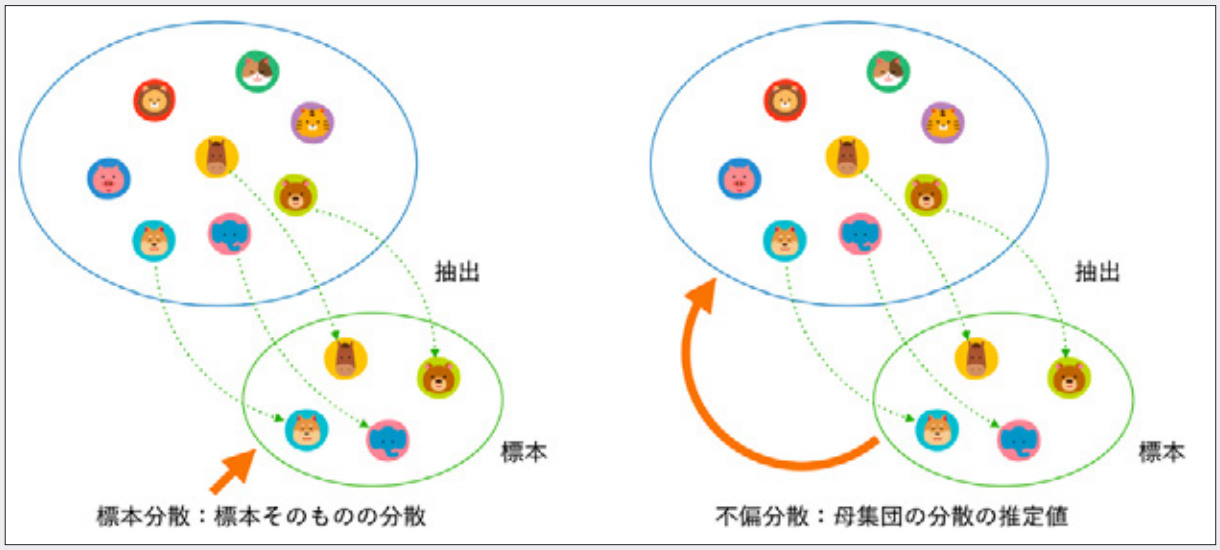


図 7　標本分差と不偏分散の違い

標本分散は標本そのものの分散。全数調査のように、標本が母集団全体である場合も標本分散を求めることになる。不偏分散は母集団から抽出した標本を基に、母集団の分散を推定した値。

勤め先収入の例には 100 件の標本があり、今回は標本そのものの分散を求めるために VAR.P 関数を使って標本分散を求めました。しかし、それらの値を基に、勤め先がある人全体の収入の分散を推定したい場合には不偏分散を使います。その場合は、VAR.S 関数を使います。引数の指定方法は全く同じです。標本標準偏差と不偏標準偏差の違いも、分散の場合と同じ考え方です。

不偏分散や不偏標準偏差の場合、分母が n ではなく n − 1 になっているのには深い理由があるのですが、取りあえず、n − 1 で割った方が適切な推定値が得られるという理解で実用上困ることはありません（が、証明について興味のある方は、「高校数学の美しい物語　不偏標本分散の意味と n-1 で割ることの証明」などをご参照ください）。

なお、文献によっては、標本分散のことを単に分散と呼び、不偏分散のことを標本分散と呼んでいるものもあります。用語が紛らわしいですが、どのようなデータを対象としてどのような分析を行っているのかが分かっていれば、どちらの意味で使っているのかは文脈から分かると思います。

Excel の中でも、関数では標本分散／不偏分散という用語が使われていますが、集計機能やピボットテーブルでは、分散／標本分散という用語が使われており、統一されていません。

コラム　歪度と尖度で分布の形を知る

ヒストグラムを見れば分布の形が可視化できますが、値が集まっているように見えるとか、なんとなくばらけているといった感覚でしか捉えられません。イメージをつかむことも大切ですが、分布の形を何らかの数値で表すことはできないでしょうか。

実は、**歪度**（わいど）と呼ばれる値により、分布の歪み（ゆがみ）を知ることができます。

また、**尖度**（せんど）と呼ばれる値により、値が集中しているかどうかを知ることができます（図 8）。

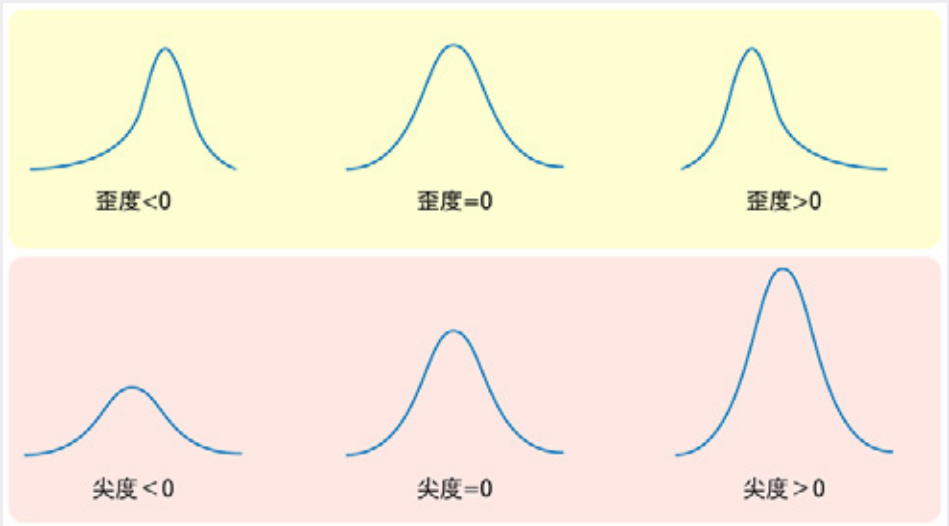


図 8　分布の形と歪度／尖度

［山が右側（大きい値の方）にあり、左側（小さい値の方）に裾野が広がった分布の場合、歪度が負になる。逆に山が左側にあり、右側に裾野が広がった分布の場合、歪度は正になる。山が尖っている（値が集中している）場合には、尖度が大きくなり、山がなだらかな場合には、尖度は小さくなる。なお、尖度の定義によっては、図の下に示した値が順に「尖度＜ 3」「尖度＝ 3」「尖度＞ 3」となっていることもある。

→目次に戻る

歪度は SKEW 関数や SKEW.P 関数で求められ、尖度は KURT 関数で求められます。いずれも引数には元のデータを指定するだけです。［勤め先収入 (1)］ワークシートと［勤め先収入 (2)］ワークシートのセル **D5** と **E5** に「=SKEW(B2:B101)」「=KURT(B2:B101)」と入力して値を比較してみてください。表 2 のような値が得られるはずです。

尺度	［勤め先収入(1)］	［勤め先収入(2)］	備考
歪度	9.69	0.42	［勤め先収入(1)］の方が左に山のある分布
尖度	95.64	-0.01	［勤め先収入(1)］の方が山が高い分布

表 3　歪度と尖度

それぞれの値と図 8 のパターンとを照らし合わせて見れば、図 2 のヒストグラムのイメージと一致していることが分かる。なお、歪度と尖度を「この記事で取り上げた関数の形式」に記した定義通りに、四則演算のみで求めた例をサンプルファイルとして用意してあるので、興味があれば参照されたい。

今回は、集団のデータのばらつきを表す散布度について、分布や尺度による取り扱いの違いを見た後、間隔尺度や比率尺度の散布度として使われる分散／標準偏差の求め方や意味を解説しました。次回は、引き続き、順序尺度で使われる四分位範囲／四分位偏差と名義尺度で使われる平均情報量／相対情報量について見ていきます。では、次回もお楽しみに！

関数リファレンス：この記事で取り上げた関数の形式

関数の使いこなし方については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます

分散や標準偏差を求めるために使った関数

VAR.P 関数：標本分散を求める

形式

VAR.P(数値 1, 数値 2, ... , 数値 255)

引数

- **数値：**標本分散を求めたい数値やセル範囲を指定する。引数は 255 個まで指定できる。

備考

以下の式に基づいて求められる。x_iは各データ、$\bar{x}$は平均値、nはデータの個数。

$$\frac{\sum(x_i - \bar{x})^2}{n}$$

VAR.S 関数：不偏分散を求める

形式

VAR.S(数値 1, 数値 2, ... , 数値 255)

引数

- **数値：**不偏分散を求めたい数値やセル範囲を指定する。引数は 255 個まで指定できる。

備考

以下の式に基づいて求められる。x_iは各データ、$\bar{x}$は平均値、nはデータの個数。

$$\frac{\sum(x_i - \bar{x})^2}{n-1}$$

STDEV.P 関数：標本標準偏差を求める

形式

STDEV.P(数値 1, 数値 2, ... , 数値 255)

引数

- **数値：**標本標準偏差を求めたい数値やセル範囲を指定する。引数は 255 個まで指定できる。

備考

以下の式に基づいて求められる。x_iは各データ、$\bar{x}$は平均値、nはデータの個数。

$$\sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$$

STDEV.S 関数：不偏標準偏差を求める

形式

STDEV.S(数値 1, 数値 2, ... , 数値 255)

引数

- **数値：**不偏標準偏差を求めたい数値やセル範囲を指定する。引数は 255 個まで指定できる。

備考

以下の式に基づいて求められる。x_iは各データ、$\bar{x}$は平均値、nはデータの個数。

$$\sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

分布の形を知るために使った関数

SKEW 関数：歪度を求める（SPSS 方式）

形式

SKEW(数値 1, 数値 2, ... , 数値 255)

引数

- **数値**：歪度を求めたい数値やセル範囲を指定する。引数は 255 個まで指定できる。

備考

統計ソフトとしてよく使われているSPSSでの計算方法（以下の式）に基づいて求められる。x_iは各データ、$\bar{x}$は平均値、nはデータの個数、sは不偏標準偏差。

$$\frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i-\bar{x}}{s}\right)^3$$

SKEW.P 関数：歪度を求める

形式

SKEW.P(数値 1, 数値 2, ... , 数値 255)

引数

- **数値**：歪度を求めたい数値やセル範囲を指定する。引数は 255 個まで指定できる。

備考

一般的に使われる以下の式に基づいて求められる。x_iは各データ、$\bar{x}$は平均値、nはデータの個数、sは不偏標準偏差。

$$\frac{1}{n}\sum\left(\frac{x_i-\bar{x}}{s}\right)^3$$

KURT 関数：尖度を求める（SPSS 方式）

形式

KURT(数値 1, 数値 2, ... , 数値 255)

引数

- **数値**：歪度を求めたい数値やセル範囲を指定する。引数は 255 個まで指定できる。

備考

統計ソフトとしてよく使われているSPSSでの計算方法（以下の式）に基づいて求められる。x_iは各データ、$\bar{x}$は平均値、nはデータの個数、sは不偏標準偏差。

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{x_i-\bar{x}}{s}\right)^4-\frac{3(n-1)^2}{(n-2)(n-3)}$$

Google スプレッドシートで配列数式を利用するための関数

ARRAYFORMULA 関数：数式を配列数式にする

形式

ARRAYFORMULA(数式)

引数

- **数式**：配列数式にしたい式を指定する。

ARRAY_CONSTRAIN 関数：配列の一部を返す

形式

ARRAY_CONSTRAIN(配列 , 行数 , 列数)

引数

- **配列**：値の並びやセル範囲、配列数式などを指定する。
- **行数**：配列から取り出す行数を指定する。
- **列数**：配列から取り出す列数を指定する。

［データ分析］四分位範囲と平均情報量 ～ 趣味や好みにはどれぐらいの幅があるのか？！

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 5 回。分布のばらつきの度合いを表す値として散布度を取り上げ、尺度や分布によって適切な散布度を利用する必要があることを説明します。順序尺度の散布度として使われる四分位範囲と、名義尺度の散布度として使われる平均情報量のお話です。

羽山博（2023 年 06 月 29 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 5 回です。前回は分布のばらつきの度合いを表す値として、間隔尺度や比率尺度のデータで使われる分散／標準偏差について、Excel の関数を使った値の求め方や、その意味について説明しました。今回は、順序尺度のデータで使われる四分位範囲（または四分位偏差）と、名義尺度のデータで使われる平均情報量（または相対情報量）を取り上げます。

サンプルファイルの利用について

本稿では、表計算ソフトを使って手を動かしながら学んでいきます。表計算ソフト Microsoft Excel 用の .xlsx ファイルをダウンロードできるようにしています。デスクトップ版の Excel が手元にない場合は、Microsoft アカウントがあれば使える無料の Microsoft 365 オンライン、もしくは Google アカウントがあれば使える無料の Google スプレッドシート（Google Sheets）をお使いください。Microsoft 365 オンラインの場合は、.xlsx ファイルを OneDrive にアップロードしてから開いてください。Google スプレッドシートの場合は、.xlsx ファイルを Google ドライブにアップロードしてから開いた上で［ファイル］メニューの［Google スプレッドシートとして保存］を実行してください。

順序尺度や分布に偏りのある間隔尺度の散布度 ～ 四分位範囲／四分位偏差

分布のばらつきの度合いを表す値として**散布度**が利用されることと、尺度によって利用できる散布度が異なることについては前回もお話ししました。おさらいもかねて、代表値と散布度の一覧（表 1）を再掲しておきます。

尺度	利用できる代表値	利用できる散布度	データの例
間隔尺度・比率尺度	平均値（算術平均）	分散／標準偏差	身長、体重、反応時間など
順序尺度	中央値	四分位範囲／四分位偏差	ランキングの順位、五段階評価など
名義尺度	最頻値	平均情報量／相対情報量	製品名、好きなスポーツの種類など

表 1　尺度と散布度

間隔尺度や比率尺度なら分散または標準偏差、順序尺度なら四分位範囲（または四分位偏差）、名義尺度なら平均情報量（または相対情報量）を散布度として使う。上の方に記した尺度では、下の方に記した散布度も使える。

今回は順序尺度や分布に偏りのある間隔尺度で使われる四分位範囲／四分位偏差から見ていきます。まず、四分位範囲のイメージをつかみましょう。ざっくり言うと、データを小さい方から順に並べたときに、全体の 1/4 から 3/4 にあたる範囲が四分位範囲です。図 1 をご覧ください。これは、平成 29（2017）年告示の中学校学習指導要領（PDF）で取り上げられている四分位範囲の定義です。

図 1　四分位範囲とは（中央値を除外する方法）

全体を半分に分け、中央値を求める。次に、中央値より小さな値のグループと、中央値より大きな値のグループを作る。つまり、中央値を除外してグループを分ける。データが偶数個の場合はちょうど真ん中の値がないが、その場合は中央にある 2 つの値を除外する。次に、小さいグループの中央値を求める。これが第 1 四分位数（Q1）となる。また、大きいグループの中央値が第 3 四分位数（Q3）となる。データが偶数個の場合は、中央にある 2 つの値の平均を Q1 や Q3 とする（全体の中央値である Q2 についても同様）。Q3-Q1 が四分位範囲の値となる。

データを小さい方から順に並べて 4 つに区切ったとき、最初の区切り位置にある値を**第 1 四分位数**と呼びます。また、3 番目の区切り位置にある値を**第 3 四分位数**と呼びます。具体的な値の求め方は図 1 に示した通りです。ただ、いちいち第何四分位数と言うのは面倒なので、第 1 四分位数を「Q1」、第 2 四分位数を「Q2」というように簡単に表すこともよくあります。ちなみに、第 2 四分位数（Q2）は中央値です。

四分位範囲とは、**第 1 四分位数（Q1）から第 3 四分位数（Q3）までの範囲のこと**で、その値は、第 3 四分位数（Q3）－ 第 1 四分位数（Q1）で求められます。つまり、四分位範囲は、中央値を中心として、全体の半分程度の個数のデータがどの範囲にあるかということを表します。四分位範囲は **IQR**（「Inter Quartile Range」の略）とも呼ばれます。

四分位偏差は四分位範囲の値を 2 で割った値です。四分位偏差は中央値からの平均的なばらつきの大きさを表しています。

実は、四分位数にはさまざまな定義があります。図 2 のように、中央値を除外せずに分けていく方法もその一つです。

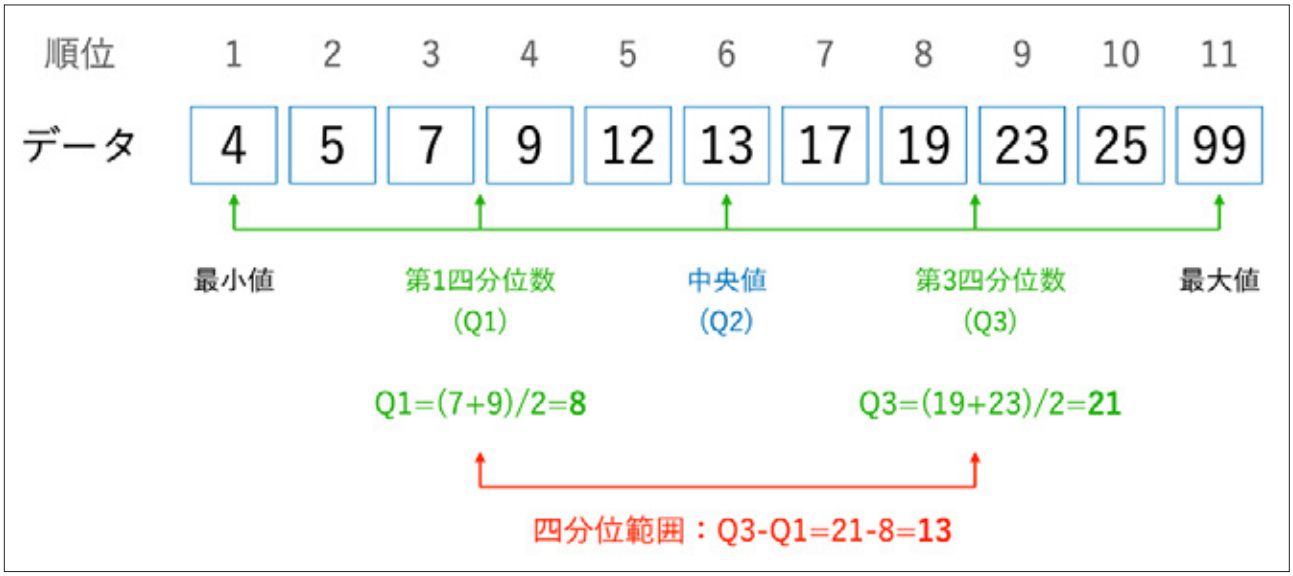


図 2　四分位範囲とは（中央値を含める方法）

全体を半分に分け、中央値を求める。次に、中央値以下の小さな値のグループと中央値以上の大きな値のグループを作る。つまり、中央値を含めてグループを分ける。ただし、データが偶数個の場合、中央値がデータの中にないので、中央値を含めずにグループを分ける。

実は、……ちゃぶ台をひっくり返すようですが、Excel ではこれらの方法とは異なる計算方法が使われます。なにっ、今までの話はなんだったんだ、と思われるかもしれませんが、実際のところ、データ数が多くなればどの方法を使ってもそれほど差が出ません。また、ばらつきの程度を把握するという意味では実用上の問題はありません。これまでのお話は四分位数や四分位範囲、四分位偏差の考え方を知るためのお話ということで、次に、Excel を使って実際に値を求めてみましょう。

四分位範囲／四分位偏差を求めてみよう（1）～ QUARTILE.EXC 関数を使う

Excel では、四分位数を求める関数として QUARTILE.EXC 関数と QUARTILE.INC 関数が利用できます。それぞれ「クォータイル・エクスクルーシブ」「クォータイル・インクルーシブ」と読みます。先ほども述べたように、これらの関数では、図 1 や図 2 の方法とは異なる方法で四分位数が求められます。が、その方法を知らなくても、関数を入力するだけで簡単に答えが得られます。そこで、細かな理屈は後のコラムにまとめることとして、まず、QUARTILE.EXC 関数を使って四分位数と四分位範囲／四分位偏差を求めてみましょう。

サンプルファイルはこちらからダウンロードできます。［四分位範囲（EXC）］ワークシートを開いてみてください。データはあまり意味のなさそうな架空の値ですが、実は、筆者の高校時代の数学の実力テストを思い出す値です。私の点数は 100 点満点中たったの 17 点でした。が、しかし、周囲を見渡しても同じような点数の人たちばかり。あまりにも難しすぎたのです。しかし、その中でも満点近い点数を取る天才的な同級生もいました。このデータは、そういった外れ値や分布に偏りのある間隔尺度のデータ、あるいは、趣味や嗜好（しこう）、スポーツなどの成績をランキング（順位）で表した順序尺度のデータだと想像してください。

では、QUARTILE.EXC 関数を入力して、第 1 四分位数から第 3 四分位数までを求めましょう（図 3）。引数には以下の 2 つの値を指定します。

- データの範囲：セル **A2 ～ A12** に入力されている
- 四分位数の位置（何番目の四分位数を求めるか）：セル **C3 ～ C5** に入力されている値を使う

QUARTILE.EXC 関数では、四分位数の位置として **1 未満**の値や **4 以上**の値は指定できません。というわけで、QUARTILE.EXC 関数はセル **D3 ～ D5** に入力してください。四分位範囲の値と四分位偏差の計算は単なる四則演算なので簡単に求められます。それぞれ、セル **D8** と **D9** に入力することにしましょう。

なお、具体的な操作については動画で解説しているので、手順を丁寧に追いかけたい方はぜひご視聴ください（お約束のセリフですが、チャンネル登録・高評価もよろしくお願いします）。

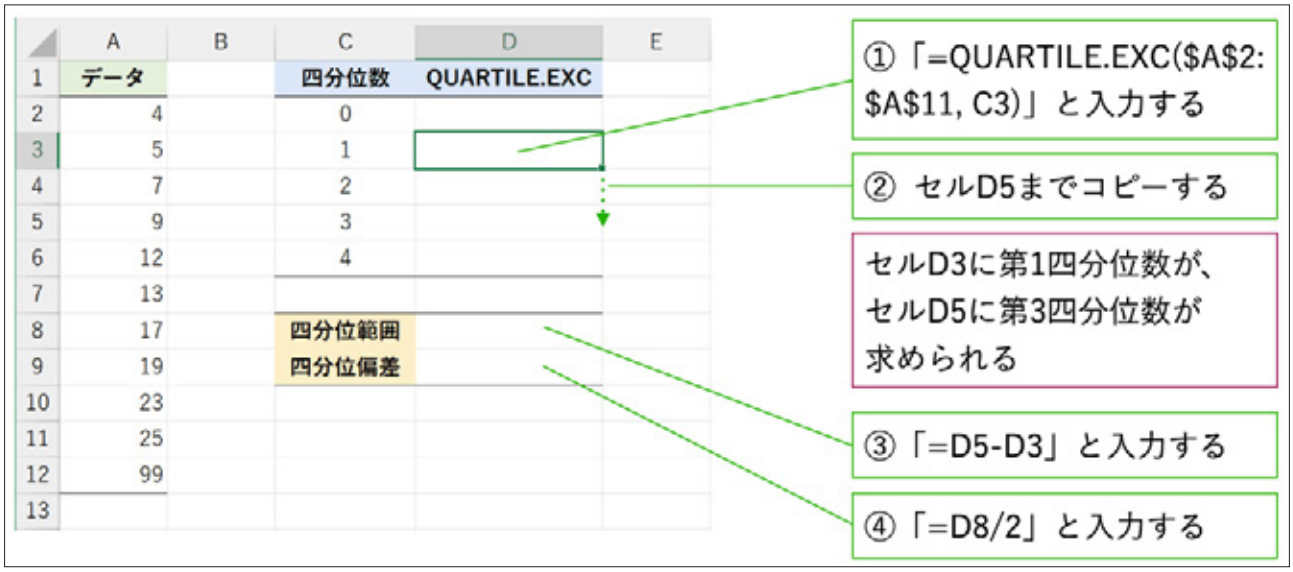


図 3　QUARTILE.EXC 関数を使って四分位範囲／四分位偏差を求める

セル **D3** に「=QUARTILE.EXC($A$2:$A$12,C3)」と入力し、セル **D5** までコピーする。スピル機能を利用するなら、セル **D3** に「=QUARTILE.EXC(A2:A12,C3:C5)」と入力するだけでよい（Google スプレッドシートの場合は「=ARRAYFORMULA(QUARTILE.EXC(A2:A12,C3:C5))」となる）。セル **D8** には「**=D5-D3**」と入力し、セル **D9** には「=D8/2」と入力すればよい。

数式を正しく入力できれば、四分位数として順に **7**、**13**、**23** という値が得られるはずです。セル **D8** の四分位範囲の値は **16** となり、セル **D9** の四分位偏差は **8** となります。答えはサンプルデータの［四分位範囲（EXC の答え）］というワークシートに含まれているので、そちらもご参照ください。

ところで、すでにお気づきかと思いますが、データの中には **99** という大きな値があります。しかし、四分位範囲や四分位偏差は、順位に基づく計算なので、極端な値の影響を受けにくくなっています。例えば、**99** という値が **99999** のようなきわめて大きな値であっても四分位範囲や四分位偏差は変わりません。

一方、前回取り上げた分散や標準偏差は極端な値の影響を強く受けます。試しに、空いているセルに「=STDEV.P(A2:A12)」と入力して標本標準偏差を求めてみてください。**25.5** という大きな値になることが分かります（平均値は **21.2** ですが、平均値を中心とした標準偏差の範囲内に **99** 以外の値が全て収まってしまいます）。

四分位範囲／四分位偏差を求めてみよう（2）～ QUARTILE.INC 関数を使う

次に、QUARTILE.INC 関数を使って四分位数、四分位範囲の値、四分位偏差を求めてみましょう。［四分位範囲（INC）］ワークシートを開いてください。関数の形式は同じです。以下のような引数を指定します。

- データの範囲：セル **A2** ～ **A12** に入力されている
- 四分位数の位置（何番目の四分位数を求めるか）：セル **C2** ～ **C6** に入力されている値を使う

QUARTILE.INC 関数では、引数に **0** を指定すると最小値が求められ、**4** を指定すると最大値が求められます。では、図 4 の流れに従って数式を入力していってください。

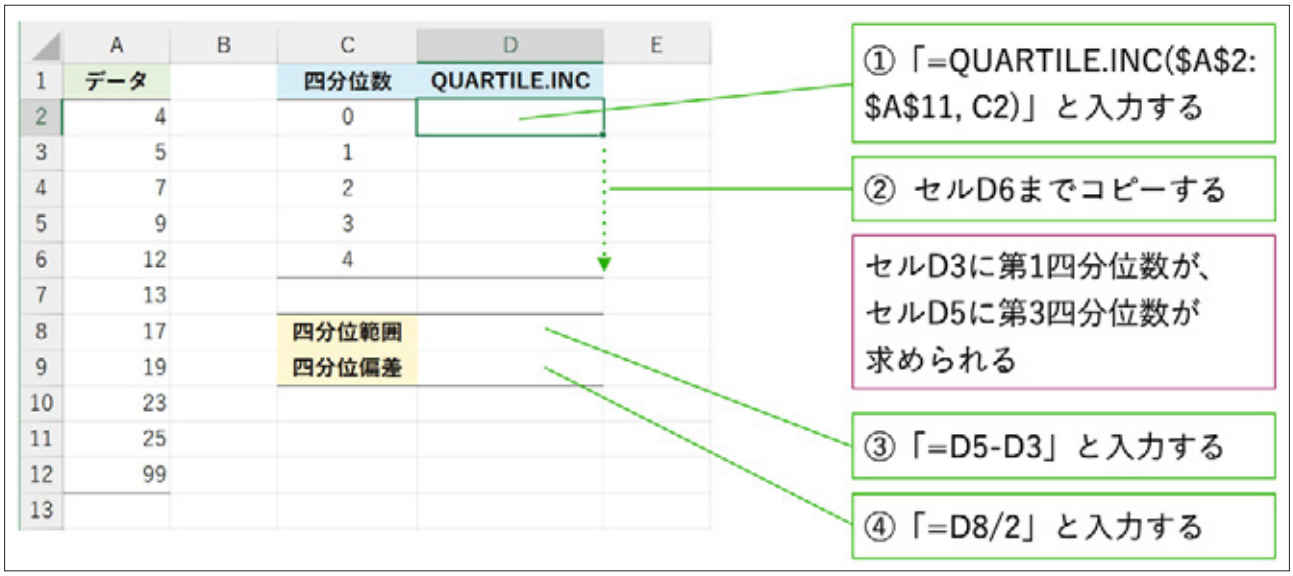


図 4　QUARTILE.INC 関数を使って四分位範囲／四分位偏差を求める

セル **D2** に「=QUARTILE.INC($A$2:$A$12,C2)」と入力し、セル **D6** までコピーする。スピル機能を利用するなら、セル **D2** に「=QUARTILE.INC(A2:A12,C2:C6)」と入力するだけでよい（Google スプレッドシートの場合は「=ARRAYFORMULA(QUARTILE.INC(A2:A12,C2:C6))」となる）。セル **D8** には「=D5-D3」と入力し、セル **D9** には「=D8/2」と入力すればよい。

数式を正しく入力できれば、四分位数が順に **4**、**8**、**13**、**21**、**99** となり、セル **D8** の四分位範囲の値が **13**、セル **D9** の四分位偏差が **6.5** となります。答えはサンプルデータの［四分位範囲（INC の答え）］というワークシートに含まれているので、そちらもご参照ください。

以上のように、四分位数は関数を入力するだけで求められます。といっても、実際にどういう計算が行われているのか分からないとちょっとモヤモヤした気になりますね。そこで、以下のコラムに詳細な計算方法を記しておきました。話がかなり細かくなるので、次に進みたい方はコラムを飛ばして、次の「四分位範囲を可視化するのに使われる箱ひげ図」や、その次の「名義尺度の散布度を求める ～ 平均情報量っていったい何？」に進んでいただいてもけっこうです。

コラム　QUARTILE.EXC 関数と QUARTILE.INC 関数の計算方法

第 1 四分位数は小さい方から 25%の位置にある値と考えられ、第 3 四分位数は小さい方から 75%の位置にある値と考えられます。一般に、QUARTILE.EXC 関数は 0%と 100%を含まない計算で、QUARTILE.INC 関数は 0%と 100%を含む計算だと言われますが、意味がちょっと分かりにくいですね。そこで、それぞれの計算の考え方と四則演算のみでやってみた場合の手順を記します。以下の例は、サンプルファイルの［四分位範囲を手計算で］というワークシートに含まれています。データの件数を n とします。

QUARTILE.EXC 関数の場合（第 1 四分位数の求め方）

QUARTILE.EXC 関数では、k 番目の値の位置を k/(n + 1) と考えます。全体を n + 1 個と数えるので、25%の位置は (n + 1)×0.25 番目となります。その位置にある値を第 1 四分位数とします。例えば、今回のサンプルデータでは n = 11 なので、(11 + 1)×0.25 = 3 となり、3 番目にある 7 という値が第 1 四分位数になります（図 3 に示した手順で実行した場合と同じ結果）。第 3 四分位数の場合は **0.25** の代わりに **0.75** を掛けて、位置を求めます。

しかし、25%の位置を表す値や 75%の位置を表す値が整数にならないこともあります。例えば、最後の **99** という値を削除すると n = 10 になるので、(10 + 1)×0.25 = 2.75 となります。その場合は **2** 番目にある **5** という値と次の **3** 番目のある **7** という値の間隔、つまり 7 − 5 = 2 を小数部分の **0.75** で比例配分します。したがって、5 + (7 − 5)×0.75 = 6.5 が第 1 四分位数となります。

これらの計算例は［四分位範囲を手計算で］というワークシートに含まれており、位置が整数になる場合も、そうならない場合にも対応した計算をしているので、ぜひ参照してみてください。以下の手順は、**99** という値を削除した場合の例です。

計算方法	サンプルデータから99という値を取り除いた場合の例	
(n+1)×25%の値を求める	(10+1)×0.25 = 2.75	…… a とする
aの整数部を求める	2	…… b とする
aの小数部を求める	0.75	…… c とする
bの位置にある値を求める	2番目にあるのは5	…… d とする
bの次の位置にある値を求める	3番目にあるのは7	…… e とする
d+(e−d)×cを求める	5+(7−5)×0.75 = 6.5	…… Q1

表 2　QUARTILE.EXC 関数の場合（第 1 四分位数の求め方）

図 1 の中央値を除外して四分位数を求める方法と似ているが、25%の位置が整数にならない場合には (c) で求めた小数点以下の値を使って補間を行う。第 3 四分位数を求める場合は、(a) の値を求めるときに 25%の代わりに 75%を指定し、同様の方法で計算する。統計ソフト SPSS で標準的に使われている方法。

QUARTILE.INC 関数の場合（第 1 四分位数の求め方）

QUARTILE.INC 関数では、先頭を **0** として位置を表します。そのため、k 番目の値の位置を (k − 1)/(n − 1) とします。全体を n − 1 個と数えるので、25%の位置は (n − 1)×0.25 + 1 となります。**1** を足しているのは、先頭を元の並びに合わせて 1 番とするためです。求められた値が整数であれば、その位置にある値を第 1 四分位数とします。第 3 四分位数の場合は **0.25** の代わりに **0.75** を掛けて、位置を求めます。

25%や 75%の位置が整数にならない場合はやはり小数部分で補間を行います。今回のサンプルデータでは n = 11 なので、(11 − 1)×0.25 + 1 = 3.5 となり、3 番目にある **7** という値と次の 4 番目にある **9** という値の間隔、つまり 9 − 7 = 2 を **0.5** で比例配分し、7 + (9 − 7)×0.5 = 8 を第 1 四分位数とします（図 4 に示した手順で実行した場合と同じ結果）。

計算方法	サンプルデータでの例	
(n−1)×25%+1の値を求める	(11−1)×0.25+1 =3.5	…… aとする
aの整数部を求める	3	…… bとする
aの小数部を求める	0.5	…… cとする
bの位置にある値を求める	3番目にあるのは7	…… dとする
bの次の位置にある値を求める	4番目にあるのは9	…… eとする
d+(e−d)×cを求める	7+(9−7)×0.5 = 8	…… Q1

表 3　QUARTILE.INC 関数の場合（第 1 四分位数の求め方）

図 2 の中央値を含めて四分位数を求める方法と似ているが、25%の位置が整数にならない場合には c で求めた小数点以下の値を使って補間を行う。第 3 四分位数を求める場合は、a の値を求めるときに 25%の代わりに 75%を指定し、同様の方法で計算する。統計ソフト R で標準的に使われている方法。

→目次に戻る

統計ソフトの R（オープンソースで無料の統計解析向けプログラミング言語およびその開発実行環境）では、標準では QUARTILE.INC 関数と同じ計算が行われますが、全部で 9 種類の計算方法が選べるようになっています。R では quartile ではなく、quantile という関数を使います。quantile 関数に type=6 を指定すると QUARTILE.EXC 関数と同様の計算が行われます（ただし、最小値と最大値も返されます）。

```
> data <- c(4, 5, 7, 9, 12, 13, 17, 19, 23, 25, 99)
> quantile(data)  # 何も指定しないと QUARTILE.INC 関数と同様の結果
  0%  25%  50%  75% 100%
   4    8   13   21   99
> quantile(data, type=6)  # type=6 を指定すると QUARTILE.EXC 関数と同様の結果
  0%  25%  50%  75% 100%
   4    7   13   23   99
```

R 言語のインタラクティブコンソールで四分位数を計算した例

四分位範囲を可視化するのに使われる箱ひげ図

四分位範囲を可視化し、分布を把握したり、外れ値を見つけたりするには**箱ひげ図**などが便利です。詳細な手順については、回を改めて、グラフを利用した可視化の方法を解説する予定なので、ここでは、箱ひげ図がどのようなものであるかということと、箱ひげ図の見方だけを簡単に紹介することにします。以下の図は、サンプルファイルの［箱ひげ図］ワークシートに含まれています（ただし、2023 年 6 月の段階では、Google スプレッドシートには箱ひげ図の機能がないので、グラフは表示されません）。

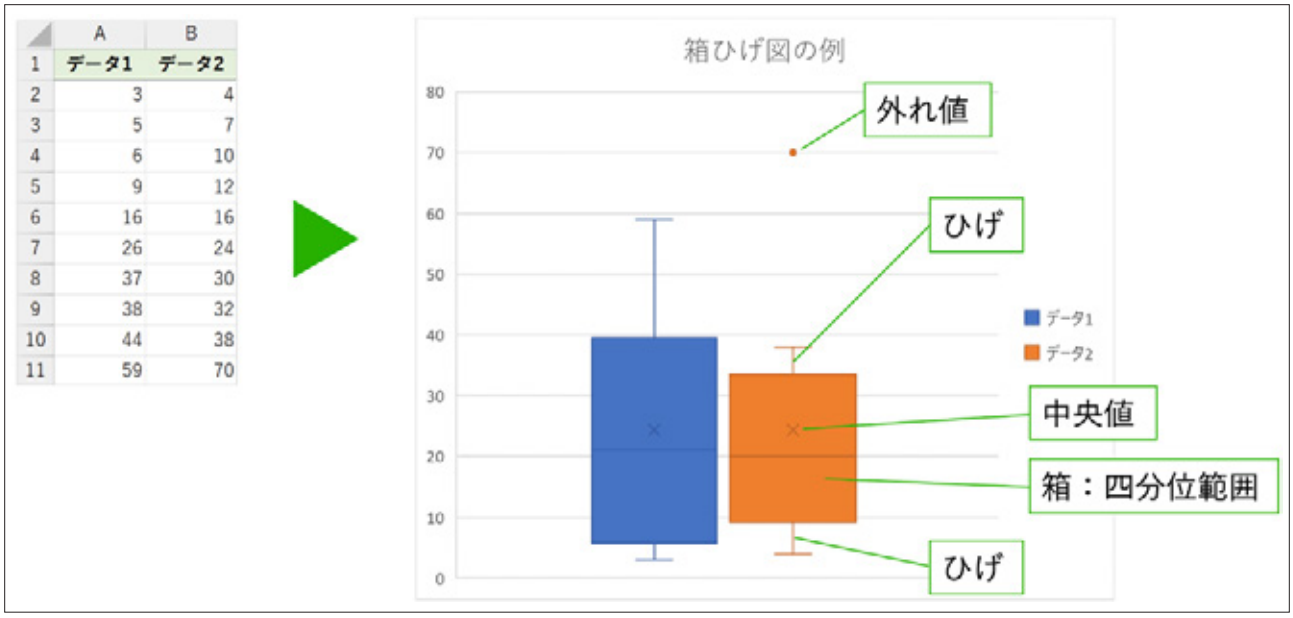


図 5　箱ひげ図とその見方

特徴が分かりやすいようなデータを 2 つ用意して、箱ひげ図を作成した例。四分位範囲（IQR）が「箱」で描かれ、中央値の値がその中に「×」で示される。「箱」から上下に伸びている線が「ひげ」で、Excel の場合、「ひげ」の上限は Q3 + 1.5× 四分位範囲の値以下の最も近い値の位置となり、下限は Q1 − 1.5×IQR 以上の最も近い値の位置となる。「ひげ」の上限と下限にある短い横線は**キャップスライン**とも呼ばれる。上限や下限からはみ出した値（外れ値）は○で示される。

→目次に戻る

左（データ 1）の例は「箱」が全体の下の方にあるので、小さい値が比較的多くなっていることが分かります。右（データ 2）の例は、極端に離れたところに大きな値がありますが、それ以外の部分のほぼ中央に「箱」があるので、外れ値以外の分布にはあまり偏りがないことが分かります。

図 5 は、Excel で箱ひげ図を作成した例ですが、特に何も指定しないと、四分位範囲を求めるために QUARTILE.EXC 関数で求めた四分位数が使われます。なお、平成 29（2017）年告示の中学校学習指導要領（PDF）では、上限と下限の定義が上とは異なっており、最大値を「ひげ」の上限とし、最小値を「ひげ」の下限とした図になっていることに注意が必要です（外れ値があると「ひげ」が大きく伸びます）。

名義尺度の散布度を求める ～ 平均情報量っていったい何

ここからは、名義尺度データのばらつきの度合いを表すために使われる**平均情報量**について見ていきます。まず、サンプルファイルをダウンロードして［平均情報量］ワークシートを開いてください。G 列から I 列に表示されているのは、第 3 回の記事で紹介した「余暇に行うスポーツ（球技）」の度数分布表です（図 6）。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	余暇に行うスポーツ（球技）						度数分布表			平均情報量を求める			
2													
3	サンプル	スポーツ		最頻値	スポーツ		番号	スポーツ	度数	確率P_i	$-\log P_i$	$P_i(-\log P_i)$	
4	1	7		9	ゴルフ		1	野球	137				
5	2	8					2	ソフトボール	27				
6	3	1					3	バレーボール	51				
7	4	9					4	バスケットボール	50				
8	5	6					5	サッカー	84				
9	6	9					6	卓球	97				
10	7	7					7	テニス	66				
11	8	4					8	バドミントン	115				
12	9	9					9	ゴルフ	210				
13	10	8					10	グラウンドゴルフ	52				
14	11	10					11	ボウリング	111				
15	12	2						合計	1000		平均情報量		
16	13	10									相対情報量		
17	14	9											

図 6 「余暇に行うスポーツ（球技）」の度数分布表

B 列の回答を集計したものが G 列～ I 列の度数分布表。例えば、**野球（1）**と答えた人が **137 人**いることが分かる。全体の人数は **1000 人**。

このようなデータで、ばらつきが大きいとか小さいということはどういうことでしょうか。例えば、1000 人中 1000 人が「ゴルフ」と答えたとすると、答えが集中しているのでばらつきがないですね。一方、全てのスポーツの人数がほとんど同じであるとすると、答えはほぼバラバラです（この例では 11 種類のスポーツがあるので、それぞれの度数が 1000/11 = 91 程度の場合）。つまりばらつきが大きくなります。そのようなばらつきの度合いを表すのが平均情報量です。名義尺度の散布度については、このようなイメージを持っていただくといいと思います。

といっても、これだけでは、なんとなく分かったような分からないような、という状態かもしれません。しかし、ここでも理屈は後回しにします。とにもかくにも、Excel を使って平均情報量を求めてみましょう。図 7 の数式に従って、手順通りに進めていけば計算ができます。

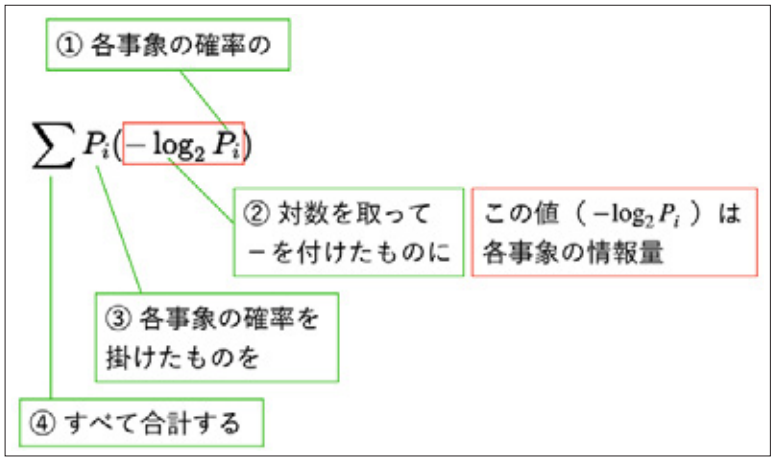


図 7　平均情報量を表す式

「事象」とは、それぞれの「できごと」や「事柄」ということ。上の例では、「野球」や「ソフトボール」などの回答が何件あったかということ。数式の中の$-\log_2 P_i$は各事象の情報量を表す。それに確率を掛けたもの、つまり$P_i(-\log_2 P_i)$は、各事象の情報量を確率で重み付けした値。それらを合計すると平均情報量が求められる（後のコラムで解説するように、これは各事象の情報量を度数で重み付けして、全体の個数で割った値、つまり各事象の情報量の重み付け平均と同じ）。

まだ、情報量や平均情報量の意味がよく分からないかもしれませんが、手を動かして操作を追いかけてみると、少しずつその意味に気づくこともあるはずです。……が、やはり理屈が分からないとモヤモヤするという方は後掲の「コラム　情報量と平均情報量の定義」を先にお読みください。以降の操作については、動画で解説しているので、手順を丁寧に追いかけたい方はぜひご視聴ください。

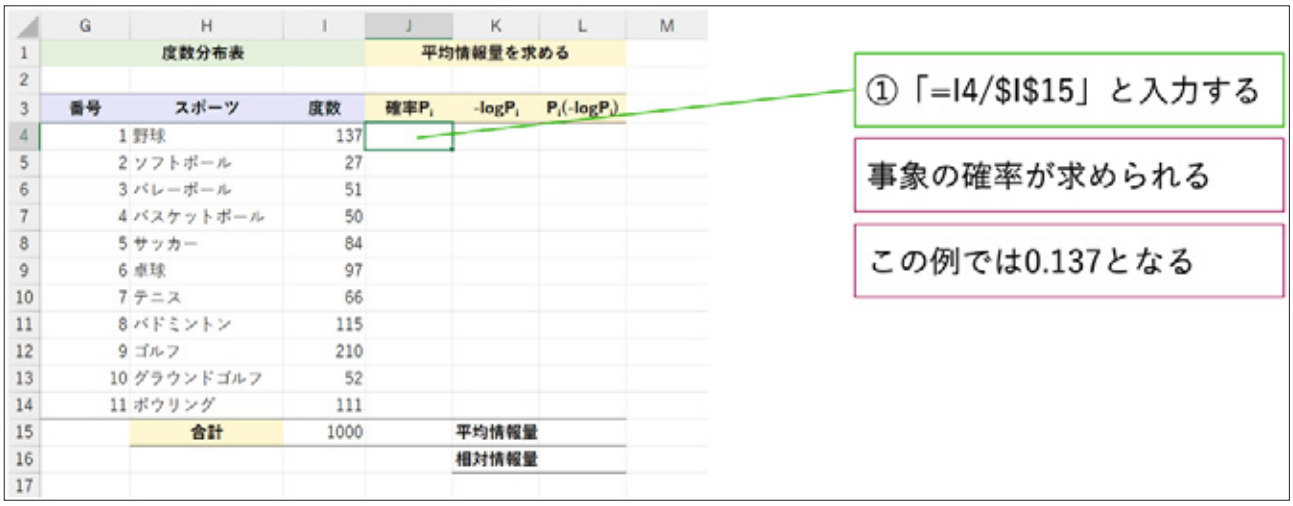
平均情報量の定義としては、マイナスをΣの外側に出した以下のような式が使われるのが普通です。

$$-\sum P_i \log_2 P_i$$

これは、図 7 の式を変形したもので、答えは同じです。後のコラムで説明する「理屈」と照らし合わせると、理解しやすいので、図 7 の式を使っています。

→目次に戻る

図 7 の式の中にある P_i はそれぞれの事象（できごと）が起こる確率です。例えば、このアンケートでは **137 人**が「野球」と答えましたが、全体の人数が **1000 人**なので、確率は 137/1000 = 0.137 です。まず、この P_i の値を求めるところからスタートしましょう（図 8）。

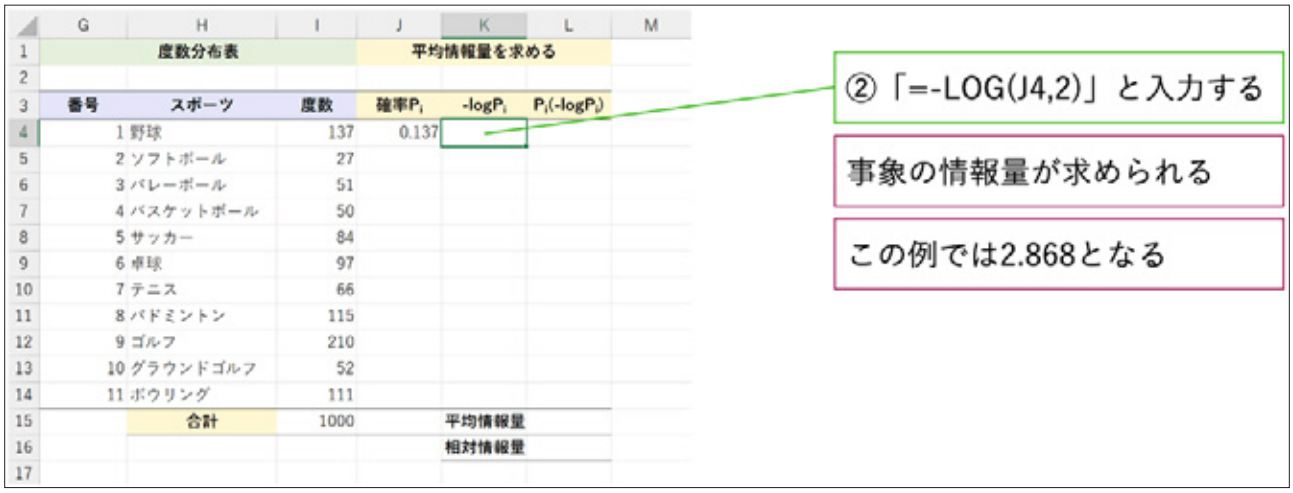


	G	H	I	J	K	L	M
1		度数分布表		平均情報量を求める			
2							
3		番号	スポーツ	度数	確率P_i	$-\log P_i$	$P_i(-\log P_i)$
4		1	野球	137			
5		2	ソフトボール	27			
6		3	バレーボール	51			
7		4	バスケットボール	50			
8		5	サッカー	84			
9		6	卓球	97			
10		7	テニス	66			
11		8	バドミントン	115			
12		9	ゴルフ	210			
13		10	グラウンドゴルフ	52			
14		11	ボウリング	111			
15			合計	1000		平均情報量	
16						相対情報量	
17							

図 8　事象の確率を求める

セル **J4** に「=I4/$I$15」と入力する。これで、「野球」の確率が求められる。数式を正しく入力すると、セル **J4** に **0.137** と表示される。「I15」を絶対参照の「$I$15」にしているのは、後で数式をコピーするため。

次に $-\log_2 P_i$ を求めます（図 9）。これは、各事象の情報量です。log は「対数」でしたね。log の後に小さく書かれた値は「底（てい）」と呼ばれます。「確かに、高校でそんなことを学んだような気がするけど、すっかり忘れてしまったよ」という方もご心配なく。対数の意味や計算方法を忘れていても、LOG 関数を使えば値が求められます。



	G	H	I	J	K	L	M
1		度数分布表		平均情報量を求める			
2							
3		番号	スポーツ	度数	確率P_i	$-\log P_i$	$P_i(-\log P_i)$
4		1	野球	137	0.137		
5		2	ソフトボール	27			
6		3	バレーボール	51			
7		4	バスケットボール	50			
8		5	サッカー	84			
9		6	卓球	97			
10		7	テニス	66			
11		8	バドミントン	115			
12		9	ゴルフ	210			
13		10	グラウンドゴルフ	52			
14		11	ボウリング	111			
15			合計	1000		平均情報量	
16						相対情報量	
17							

図 9　事象の情報量を求める

セル **K4** に「=-LOG(J4,2)」と入力する。これで、「野球」が **137 人**であったときの情報量が求められる。LOG 関数の 2 番目の引数には「底」の **2** を指定する。数式を正しく入力すると、セル **K4** に **2.868** と表示される。

続いて、$P_i(-\log_2 P_i)$ を求めます（図 10）。これは、情報量を確率で重み付けした値でしたね。図 8 で求めた値と図 9 で求めた値の積です。

→目次に戻る

図 10　事象の情報量を確率で重み付けする

セル **L4** に「=J4*K4」と入力する。情報量に確率を掛けて重み付けした値が求められる。数式を正しく入力すると、セル **K4** に **0.393** と表示される。

ここまでで、「野球」について、情報量を確率で重み付けした値が求められました。他のスポーツについても同様の計算を行います。図 8 ～図 10 で入力した式（セル **J4 ～ L4**）を全てのスポーツについて（14 行目まで）コピーしましょう（図 11）。

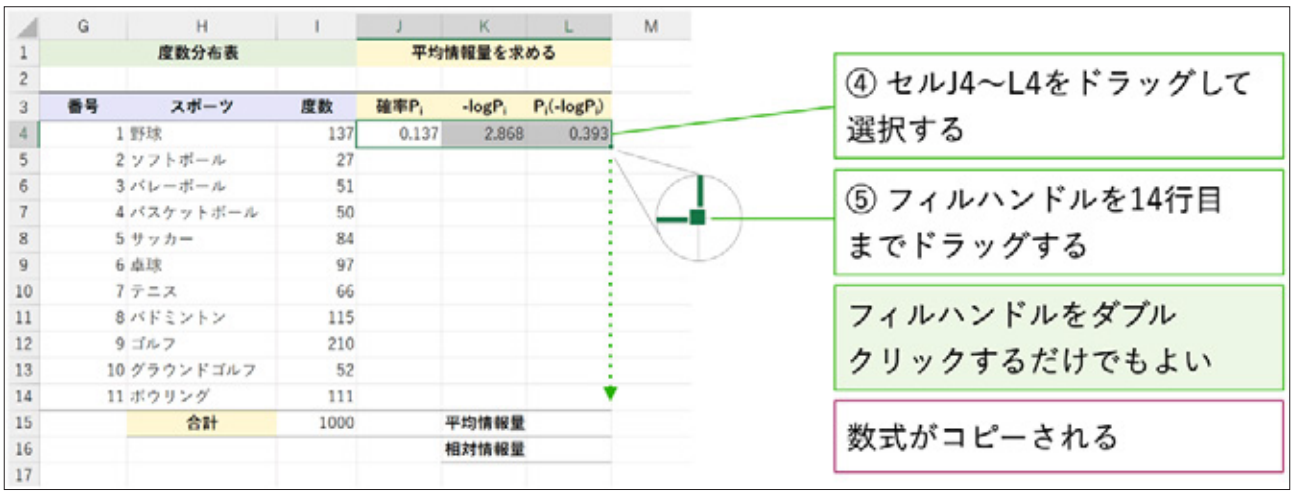
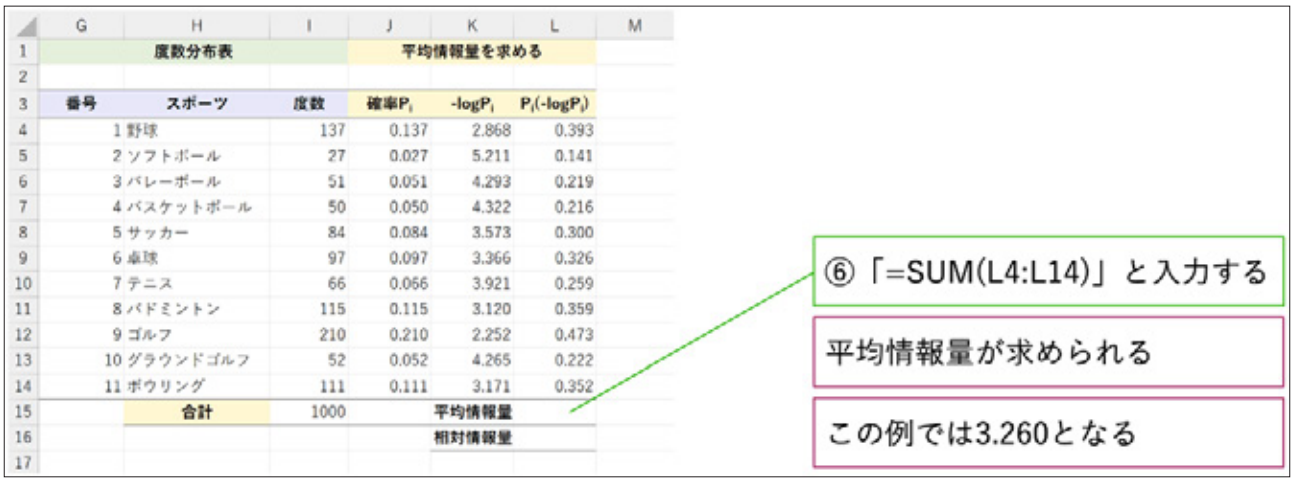


図 11　全ての事象の情報量を確率で重み付けした値を求める

セル **J4 ～ L4** を選択し、14 行目までコピーする。これで、全ての事象の情報量を確率で重み付けした値が一気に求められる。

最後に、L 列の値を合計すれば、平均情報量が求められます（図 12）。



	G	H	I	J	K	L	M
1		度数分布表			平均情報量を求める		
2							
3		番号	スポーツ	度数	確率P_i	$-\log P_i$	$P_i(-\log P_i)$
4		1	野球	137	0.137	2.868	0.393
5		2	ソフトボール	27	0.027	5.211	0.141
6		3	バレーボール	51	0.051	4.293	0.219
7		4	バスケットボール	50	0.050	4.322	0.216
8		5	サッカー	84	0.084	3.573	0.300
9		6	卓球	97	0.097	3.366	0.326
10		7	テニス	66	0.066	3.921	0.259
11		8	バドミントン	115	0.115	3.120	0.359
12		9	ゴルフ	210	0.210	2.252	0.473
13		10	グラウンドゴルフ	52	0.052	4.265	0.222
14		11	ボウリング	111	0.111	3.171	0.352
15			合計	1000		平均情報量	
16						相対情報量	
17							

図 12　全ての事象の情報量を確率で重み付けした値を求める

セル **L15** に「=SUM(L4:L14)」と入力する。これで平均情報量が求められる。結果は **3.260** となる。

いかがでしょう。セル **L15** に **3.260** という値が表示されたら正解です。セル **L16** の相対情報量のお話はちょっと後回しにして、［平均情報量（答え）］というワークシートを開いて作成例を確認しておいてください。

スピル機能を使うなら、セル **J4** に「=I4:I14/I15」、セル **K4** に「=-LOG(J4#,2)」、セル **L4** に「=J4#*K4#」と入力するだけで（コピーの操作は行わなくても）、セル **J4** ～ **L14** までの値が全て求められます。あとはセル **L15** で合計を求めるだけです。「J4#」などはスピル機能によって入力された範囲を表します（つまり「**J4**:**J14**」と同じ意味です）。

なお、Google スプレッドシートの場合は、セル **J4** に「=ARRAYFORMULA(I4:I14/I15)」、セル **K4** に「=ARRAYFORMULA(-LOG(J4:J14,2))」、セル **L4** に「=ARRAYFORMULA(J4:J14*K4:K14)」と入力します。

ちなみに、セル **L15** に、図 8 ～ 12 の計算を全て組み合わせた「=SUM(-LOG(I4:I14/I15,2)*(I4:I14/I15))」という数式を入力するだけで、**3.260** という答えを得ることもできます（Google スプレッドシートなら、「=ARRAYFORMULA(SUM(-LOG(I4:I14/I15,2)*(I4:I14/I15)))」）。

答えのワークシートには、全ての度数が同じ場合（平均情報量が最大になる場合）の例も含めてあります。平均情報量の最大値は、項目（カテゴリ）数を k とすると、それぞれの確率 P_i が全て 1/k なので、

$$
\begin{aligned}
-\sum P_i \log_2 P_i &= -\sum \frac{1}{k} \log_2 \frac{1}{k} \\
&= -k \frac{1}{k} \log_2 \frac{1}{k} \quad \text{.....同じ値を}k\text{個足すので}k\text{倍した} \\
&= -\log_2 \frac{1}{k} \\
&= \log_2 k
\end{aligned}
$$

となります。この場合であれば、11 個の項目があるので $\log_2 11 = 3.459$ となります。このように、平均情報量の最大値はカテゴリ数によって変わりますが、平均情報量を $\log_2 k$ で割れば、最大値を **1** にそろえることができます。それが**相対情報量**です。［平均情報量］ワークシートに戻って試してみましょう。セル **L16** に「=L15/LOG(COUNT(I4:I14),2)」と入力すれば相対情報量が求められます。平均情報量が **3.260** だったので、相対情報量は 3.260/3.459 = 0.942 となるはずです。相対情報量が **1** に近いことから、回答が比較的ばらけている（特定のスポーツのみに集中していない）ということが分かります。

なお、全ての人が「野球」と答えた場合は、他の項目がそもそも存在しないので、P = 1 となり、平均情報量は **0** となります。

コラム　情報量と平均情報量の定義

平均情報量の定義を理解するに当たっては、まず、情報量がどのような値であるかを理解しておく必要があります。しかし、そもそも情報量が大きいとか小さいというのはどういうことでしょうか。

例えば「飛行機が空港に着陸した」というありふれた情報よりも、少し古いニュースですが「飛行機がハドソン川に不時着水した」という情報の方が、情報量が大きいような気がします。なんとなく感覚としては分かるのですが、数式として定義したいものです。

そこで、珍しいこと、つまり、起こる確率が小さいことが起こった（ということが分かった）場合に、情報量が大きくなるものとしましょう。そのためには、確率の逆数を求めるといいですね。確率 P が小さくなるほど、

$$\text{その逆数} \frac{1}{P}$$

は大きくなります。

ただし、

$$\text{確率の逆数}\frac{1}{P}$$

をそのまま情報量として使うのではなく、対数を取ったものを情報量とします。その理由は後述しますが、対数を取ると、

$$\log\frac{1}{P} = -\log P$$

となります（底は 2 です。以降、底の表記は省略します）。これが情報量の定義で、単位は**ビット**（bit）です。上の式の右辺は、以下に示す対数の公式を適用して求めたものです。

$$\log\frac{1}{x} = -\log x$$

対数を取る理由は、掛け算を足し算として表したいからです。簡単な例で見てみましょう。正確な（1 ～ 6 のどの目も同じ確率で出る）サイコロを振ったとき、偶数の目が出たこと（A とします）が分かった後に、それが 6 でなかった（つまり 2 か 4 であった）ことが分かった場合（B とします）は、最初に考えた情報

- 偶数の目が出た → **6個のうちの3個** …… $A = \frac{6}{3}$
- さらに、6ではないことが分かった → **3個のうちの2個** …… $B = \frac{3}{2}$

量の大きさ（確率の逆数）は以下のようになります。

- 最初から2か4であることが分かった → **6個のうちの2個**……$C = \frac{6}{2}$

このように情報を小出しにしても、最初から、2 か 4 であることが分かったとしても、得られた情報量は同じです。

$$A \times B = C$$

これらの値は、以下のような掛け算の関係になっています。

$$\frac{6}{\cancel{3}} \times \frac{\cancel{3}}{2} = \frac{6}{2}$$

値を当てはめて確認してみましょう。

$$A \times B = C$$

です。確かに A×B = C になっています。このように掛け算で表される式を、足し算で表すことを考えてみます。そのためには対数を取ればいいですね。

のとき、

$$\log A + \log B = \log C$$

です。これも対数に関係する公式そのままです。A や B の値をそのまま使うのではなく、対数を取った値を使うと、新たな情報が加えられたときに情報量を足し算として表すことができます。というわけで、対数を取った値を情報量として使うことにしたわけです。

話を元に戻して、次に平均情報量の定義です。平均情報量は情報量の期待値（度数で重みを付けた平均）です。つまり、それぞれの情報量に度数を掛けて、個数で割ったものです。それぞれの度数を F_i とすると、

$$\frac{\sum F_i(-\log P_i)}{n}$$

と表せます。確率は、度数を全体の個数で割ったものなので、

$$P_i = \frac{F_i}{n}$$

です。つまり、$F_i = nP_i$ となります。これを上の式に代入しましょう。

$$\begin{aligned}\frac{\sum \cancel{n} P_i(-\log P_i)}{\cancel{n}} &= \sum P_i(-\log P_i) \\ &= -\sum P_i \log P_i\end{aligned}$$

はい、これで平均情報量を求める式が得られました。なお、対数の意味や公式、情報量の定義、平均情報量については、「[AI・機械学習の数学］指数と対数（対数編）」でも解説しています。ぜひご参照ください。

今回は、集団のデータのばらつきの度合いを表す散布度について、順序尺度で使われる四分位範囲／四分位偏差と、名義尺度で使われる平均情報量／相対情報量の求め方や意味をそれぞれ解説しました。次回は、個々のサンプルの値が集団の中でどの位置にあるかということを考えます。単なる順位だけでなく、パーセント単位での順位を求めたり、偏差値を求めたりする方法を見ていきます。では、次回もお楽しみに！

関数リファレンス：この記事で取り上げた関数の形式

関数の使いこなし方については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

四分位数を求めるために使った関数

QUARTILE.EXC 関数：四分位数を求める（SPSS の標準的な方法）

形式

QUARTILE.EXC(値の並び , 四分位数)

引数

- **値の並び**：四分位数を求めたい元のデータを指定する。
- **四分位数**：**1** ～ **3** を指定する。**1** なら第 1 四分位数、**2** なら第 2 四分位数（中央値）、**3** なら第 3 四分位数が求められる。

QUARTILE.INC 関数：四分位数を求める（S や R の標準的な方法）

形式

QUARTILE.INC(値の並び , 四分位数)

引数

- **値の並び**：四分位数を求めたい元のデータを指定する。
- **四分位数**：**0** ～ **4** を指定する。**0** なら最小値、**1** なら第 1 四分位数、**2** なら第 2 四分位数（中央値）、**3** なら第 3 四分位数、**4** なら最大値が求められる。

対数を求めるために使った関数

LOG 関数：対数を求める

形式

LOG(数値 , 底)

引数

- **数値**：対数を求めたい値（真数）を指定する。
- **底**：対数の底を指定する。省略すると **10** が指定されたものとみなされる。

[データ分析] 順位と偏差値～私の成績順位はどのあたり？

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 6 回。集団の中での位置をパーセント単位で求めたり、偏差値を求めたりする方法と、その考え方を説明します。偏差値は大学や高校のランク付けによく使われていますが、序列を付けるためのものではなく、異なる分布の集団の間でも位置が比較できるとても便利な値です。

羽山博（2023 年 07 月 20 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 6 回です。これまでは、集団のばらつきの度合いを表す散布度について見てきました。具体的には、前々回は間隔尺度や比率尺度で使われる分散／標準偏差を、前回は、順序尺度で使われる四分位範囲／四分位偏差と、名義尺度で使われる平均情報量／相対情報量を紹介しました。今回は、集団そのものの特徴から少し視点を変えて、集団の中にある個々の値に注目していきます。個々の値が集団の中でどの位置にあるのかをパーセント単位で求める方法や、異なる分布の集団の間で位置を比較できる偏差値について見ていきます。

サンプルファイルの利用について

本稿では、表計算ソフトを使って手を動かしながら学んでいきます。表計算ソフト Microsoft Excel 用の .xlsx ファイルをダウンロードできるようにしています。デスクトップ版の Excel が手元にない場合は、Microsoft アカウントがあれば使える無料の Microsoft 365 オンライン、もしくは Google アカウントがあれば使える無料の Google スプレッドシート（Google Sheets）をお使いください。Microsoft 365 オンラインの場合は、.xlsx ファイルを OneDrive にアップロードしてから開いてください。Google スプレッドシートの場合は、.xlsx ファイルを Google ドライブにアップロードしてから開いた上で［ファイル］メニューの［Google スプレッドシートとして保存］を実行してください。

まずは単純に順位を求めてみよう

集団の中での順位を求める方法については、誰もが直感的に理解していることと思います。そこで、以下（表 1）のような成績のデータに順位を付けてみてください。まずは手作業でやってみましょう。作業しやすいように、データは降順（大きい値から小さい値への順）に並べ替えてあります。

成績	95	90	82	81	78	69	67	54	54	35	24	12
順位												

表 1　成績のデータに順位を付けてみましょう（［順位］欄に番号を振る）

皆さんはどのように順位を付けたでしょうか。以下（表 2）のような付け方をした人が多いと思います。

成績	95	90	82	81	78	69	67	54	54	35	24	12
順位	1	2	3	4	5	6	7	8	8	10	11	12

表 2　成績のデータに順位を付けた例（1）

→目次に戻る

54 点の人が 2 人いるので、同じ **8 位**という順位を与えました。その次は **10 位**としたわけです。つまり、同じ点数の人が複数いる場合には、上位の順位を与えて、後ろは欠番とするわけですね。Excel の RANK.EQ 関数で順位を求めるとこれと同じ結果になります。

別の考え方もあります。**8 位**と **9 位**が同じ得点だったので、上位と下位の平均の順位、つまり (8 + 9)÷2 = 8.5 を 2 人に与えるという考え方です。その場合は、以下（表 3）のような順位になります。

成績	95	90	82	81	78	69	67	54	54	35	24	12
順位	1	2	3	4	5	6	7	8.5	8.5	10	11	12

表 3　成績のデータに順位を付けた例（3）

Excel の RANK.AVG 関数では、このような順位の求め方になります。

では、さっそく Excel を使って……と行きたいところですが、それぞれの関数を使ったときにどのような結果になるかを予想してからやることにしましょう。上とはちょっと違った例です（表 4）。同じ順位が 3 人いることにも注意してください。

成績	95	90	82	78	78	69	67	54	54	54	24	12
RANK.EQでの順位	1	2	3	4	4	6	7	8	8	8	11	12
RANK.AVGでの順位	1	2	3	4.5	4.5	6	7	9	9	9	11	12

表 4　成績のデータに順位を付けてください（表 1 ～ 3 と成績の内容が異なることに注意）

予想は当たっていたでしょうか。**54 点**が 3 人いますが、RANK.EQ 関数の場合は上位の順位を与えるので、いずれも **8 位**となります。その次が **11 位**ですね。RANK.AVG 関数の場合は、**8 位**、**9 位**、**10 位**の上位と下位の平均なので、全て **9 位**を与えます。その次が **11 位**になるのは同じです。

では、Excel で試してみましょう。サンプルファイル（06a.xlsx）をこちらからダウンロードし、[順位を求める] ワークシートを開いて試してみてください。以下で説明する操作については動画でも解説しているので、手順を丁寧に追いかけたい方はぜひご視聴ください。

RANK.EQ 関数でも、RANK.AVG 関数でも、第 1 引数には対象となる値を指定し、第 2 引数には、全ての値を指定します。第 3 引数には降順での順位か昇順での順位かを指定します。降順の場合は **0** を指定するか省略し、昇順の場合は **0 以外**の値を指定します。降順の場合は値が大きいほど上位になり、昇順の場合は値が小さいほど上位になります。なお、値は並べ替えられていなくても構いません。

図 1 に RANK.EQ 関数を入力する手順を示します。手順に従って操作してみてください。

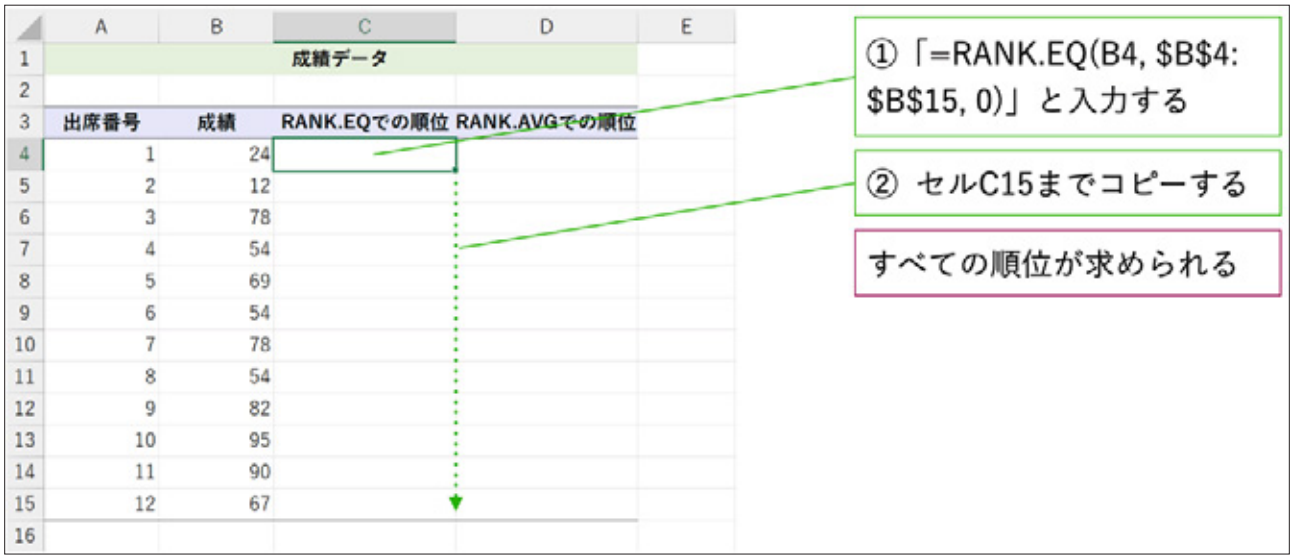


図 1　RANK.EQ 関数を使って順位を求める

［順位を求める］ワークシートを開き、セル **C4** に「=RANK.EQ(B4,$B$4:$B$15,0)」と入力して順位を求めよう。

セル **C4** に「=RANK.EQ(B4,$B$4:$B$15,0)」と入力し、セル **C15** までコピーしてください。スピル機能を使うならセル **C4** に「=RANK.EQ(B4:B15,B4:B15,0)」と入力するだけで全ての順位が求められます。Google スプレッドシートで配列数式を使うなら、セル **C4** に「=ARRAYFORMULA(RANK.EQ(B4:B15,B4:B15,0))」と入力すれば同じ結果が得られます。

RANK.AVG 関数については、関数名が異なるだけで、引数の指定は全く同じです。セル **D4** に「=RANK.AVG(B4,$B$4:$B$15,0)」と入力して、セル **D15** までコピーすると、図 2 のような結果が得られます。スピル機能を使うなら「=RANK.AVG(B4:B15,B4:B15,0)」となり、Google スプレッドシートで配列数式を使うなら、「=ARRAYFORMULA(RANK.EQ(B4:B15,B4:B15,0))」となります。

	A	B	C	D	E
1			成績データ		
2					
3	出席番号	成績	RANK.EQでの順位	RANK.AVGでの順位	
4	1	24	11	11	
5	2	12	12	12	
6	3	78	4	4.5	
7	4	54	8	9	
8	5	69	6	6	
9	6	54	8	9	
10	7	78	4	4.5	
11	8	54	8	9	
12	9	82	3	3	
13	10	95	1	1	
14	11	90	2	2	
15	12	67	7	7	
16					

図 2　RANK.EQ 関数と RANK.AVG 関数で求めた順位

成績順に並べ替えられていないので少し分かりづらいが、予想した順位と同じ結果が得られた。RANK.EQ 関数では、**4 位**が 2 名、**8 位**が 3 名となり、RANK.AVG 関数では、**4.5 位**が 2 名、**9 位**が 3 名となる。作成例は［順位を求める（答え）］ワークシートに含まれている。

図 2 の結果を見ると、実際に試験を受けてある点数を取った人が何位であるかは分かりますが、例えば、その試験で **80 点**を取ったとすればどの辺りの位置にいるかは分かりません。実は、RANK.EQ 関数や RANK.AVG 関数では、順位を求める対象となる値（この場合なら **80 点**）がデータの中に存在しないとエラーになってしまいます。そこで、次に、データの中に存在しない値を指定しても順位が求められる方法を見ていくことにします。そのためには、値が全体の何パーセントの位置にあるかを求めます。

図 1 や図 2 のデータは出席番号順に並んでいるので、自分の成績や順位が一目で見つけられます。しかし、データを分析する場合には、成績の降順または昇順に並んでいた方が便利です。並べ替えもデータ分析のために使われる基本的な手法なので、番外編として並べ替えの方法も解説する予定です。

80 点を取ったら第何位？

順位を基にして、ある値が全体の何パーセントの位置にあるかを求めるには、PERCENTRANK.EXC 関数または PERCENTRANK.INC 関数が使えます。

これまでに見てきたデータを例に、**80 点**を取ると全体の何パーセントの位置にいるかを求めてみましょう。サンプルファイル（06b.xlsx）をこちらからダウンロードし、［順位（PERCENTRANK.EXC）］ワークシートを開いて以下の操作を行いましょう。これについても、操作は動画でも解説しています。手順を丁寧に追いかけたい方はぜひご視聴ください。

図 3 に PERCENTRANK.EXC 関数を入力する手順を示します。

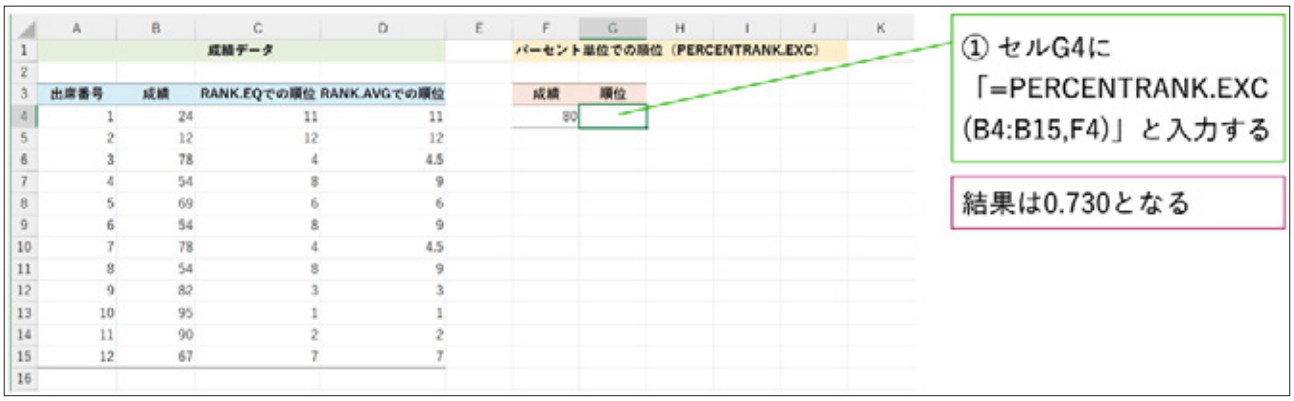


図 3　パーセント単位での順位を求める（PERCENTRANK.EXC 関数）

［順位（PERCENTRANK.EXC）］ワークシートを開き、セル **G4** に「=PERCENTRANK.EXC(B4:B15,F4)」と入力して順位を求めよう。作成例は［順位（PERCENTRANK.EXC 答え）］ワークシートを参照。データの最小値より小さい値や、データの最大値より大きい値を指定するとエラーになることに注意。

PERCENTRANK.EXC 関数の引数にはデータの範囲と順位を求めたい値を指定します。関数の実行結果は **0.730**、つまり **73%**となるはずです。ただし、Google スプレッドシートでは、有効桁数の取り扱いが異なるので **0.731** と表示されます（Excel では切り捨て、Google スプレッドシートでは四捨五入が行われているようです）。

PERCENTRANK.INC 関数の場合も、関数名が違うだけで、引数の指定方法は同じです。［順位（PERCENTRANK.INC）］ワークシートを開いて、セル **G4** に「=PERCENTRANK.INC(B4:B15,F4)」と入力してみてください。結果は **0.772**、つまり **77.2%**となります。作成例は［順位（PERCENTRANK.INC 答え）］ワークシートに含まれています。上でも述べた通り、Google スプレッドシートでは、有効桁数の取り扱いが異なるので **0.773** と表示されます。

これらの関数の違いは、PERCENTRANK.EXC 関数では **0％**と **100％**を除いた範囲で順位を求め、PERCENTRANK.INC 関数では **0%**と **100%**を含めた範囲で順位を求めるということです。……が、それだけだと意味が分からないですね。**80 点**など、元のデータに存在しない値を指定した場合の順位の求め方についても謎です。そこで、詳細な計算方法を後のコラムにまとめておきました。先を急ぐ方はコラムを飛ばしてもらっても構いません。

実際のところ、データ数が多ければ、どちらの関数を使ってもほぼ同じ結果になるので、順位を求めるという目的において、実用上での違いはほとんどありません。

コラム　PERCENTRANK.EXC 関数と PERCENTRANK.INC 関数の計算方法

PERCENTRANK.EXC 関数では、RANK.EQ 関数と同じ方法で求めた**昇順**の順位を k とし、全体の個数を n とすると、k/(n + 1) でパーセント単位の順位が求められます。ただし、指定した値が元のデータに含まれない場合には、1 つ下の値と 1 つ上の値の間で補間が行われます。

例えば、図 3 のデータでは、**80** という値は元のデータに含まれていません。そこで、1 つ下の値である **78**（昇順なら **8 位**）と、1 つ上の値である **82**（昇順なら **10 位**）の間で補間を行います（図 4）。

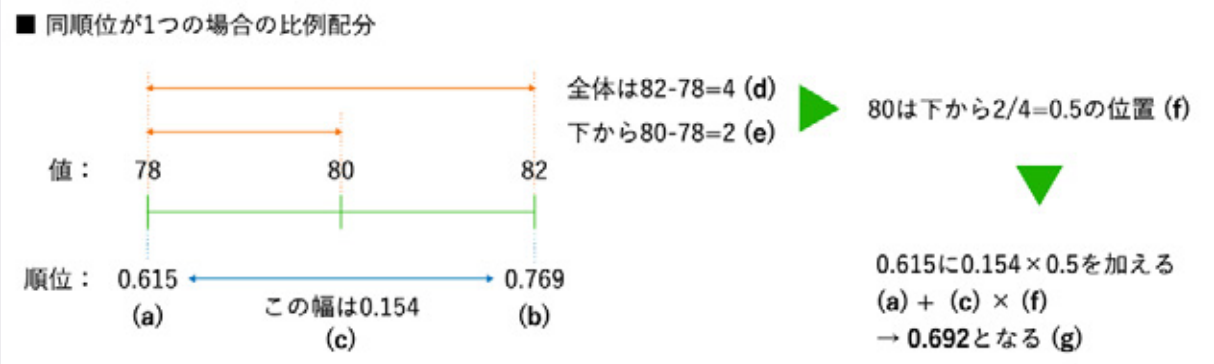


図 4　補間の方法（同順位が 1 つの場合）

下位の順位からどれだけ離れているかを基に、パーセント単位での順位の幅を配分する。この例では、**80** という値は **78** から **82** のちょうど半分の位置にあるので、**0.615** から **0.769** までの幅を半分進んだ位置（＝ **0.692**）が答えとなる。

→目次に戻る

計算の手順を追いかけると以下のようになります。

- 計算の基となる下位と上位のパーセント単位での順位と幅を求める
  - **78 点**の順位は昇順で **8 位**：k = 8, n = 12 なので、k/(n + 1) = 8/13 = 0.615 …… (a)
  - **82 点**の順位は昇順で **10 位**：k = 10, n = 12 なので、k/(n + 1) = 10/13 = 0.769 …… (b)
  - (b) と (a) の幅：0.769 − 0.615 = 0.154 …… (c)
- 比例配分する
  - 上位と下位の値の差：82 − 78 = 4 …… (d)
  - 80 点と下位の値の差：80 − 78 = 2 …… (e)
  - (d) と (e) の比：2/4 = 0.5 …… (f)
  - (a) + (c)×(f) を求める：0.615 + 0.154×0.5 = 0.615 + 0.77 = 0.692 ……(g)

このように、同順位が 1 つだけであれば単純に補間すればいいのですが、同順位が複数個ある場合には、補間した値をさらに補間する必要があります。この例では、**78 点**が 2 つあるので、図 4 の補間を行った後、さらに図 5 のような補間を行います。

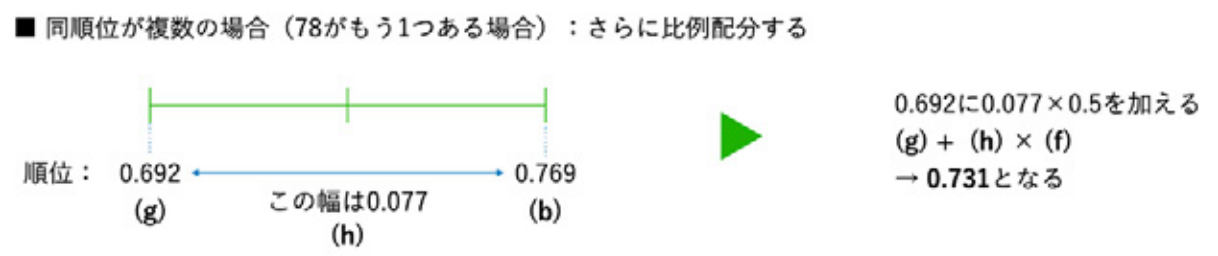


図 5　補間の方法（同順位が複数ある場合）

図 4 のように比例配分して求めた順位を下位の順位とし、さらに比例配分する。この例では、上で求めた **0.692** から **0.769** までの幅を半分進んだ位置（**= 0.731**）が答えとなる。

計算の手順は以下の通りです。

- **78 点**がもう 1 つあるので、さらに比例配分する
  - 上位と (g) のパーセント単位での順位の差：0.769 − 0.692 = 0.077 …… (h)
  - (g) + (h)×(f) を求める：0.692 + 0.077×0.5 = 0.692 + 0.0385 = 0.7305 …… 小数点以下 3 桁までで四捨五入すると **0.731**（これが答え）

PERCENTRANK.EXC 関数や PERCENTRANK.INC 関数では、有効桁数の既定値が小数点以下 3 桁となっているので、Excel の場合、特に何も指定しないと小数点以下 3 桁目より下位の桁は切り捨てられます。そのため、上で見た計算とは小数点以下 3 桁目の値が異なることがあります（PERCENTRANK.EXC 関数の場合、**0.7305** の最後の桁が切り捨てられるので、結果は **0.730** となります）。

一方の、PERCENTRANK.INC 関数では、RANK.EQ 関数と同じ方法で求めた昇順の**順位**を k とし、全体の個数を n とすると、(k − 1)/(n − 1) でパーセント単位の順位が求められます。やはり、指定した値が元のデータに含まれない場合には、1 つ下の値と 1 つ上の値の間で補間が行われます。補間の方法は PERCENTRANK.EXC 関数と同じです。

サンプルデータの［計算方法］ワークシートには上記の手順で計算した例と、PERCENTRANK.EXC 関数や PERCENTRANK.INC 関数での検算結果が含まれています。手順はかなり長いですが、詳しく知りたい方はぜひご参照ください。

上位 10%に入るには何点取ればいい?

PERCENTRANK.EXC 関数や PERCENTRANK.INC 関数では、ある点数が全体の中の何パーセントの位置にあるかを求めましたが、逆に、全体の中の何パーセントの位置にいるには何点取ればいいかを知りたいこともあるでしょう。例えば、上位 **10%**に入るには何点取ればいいか、といった場合です。

上位 **10%**の位置を知るには、下位から **90%**の位置を求めます。利用する関数は、PERCENTILE.EXC 関数と PRECENTILE.INC 関数です。サンプルファイル（06c.xlsx）をこちらからダウンロードし、［上位 10%（PERCENTILE.EXC)］ワークシートを開いて以下の操作を行いましょう。操作は動画でも解説しています。手順を丁寧に追いかけたい方はぜひご視聴ください。

図 6 に PERCENTILE.EXC 関数を入力する手順を示します。

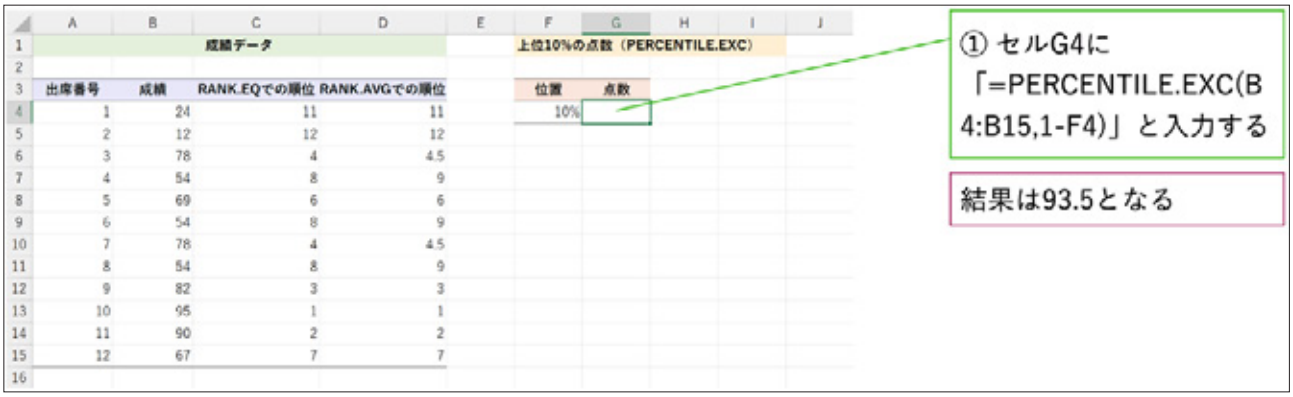


図 6　上位 10 パーセントの位置の点数を求める（PERCENTILE.EXC 関数）

［上位 10%（PERCENTILE.EXC)］ワークシートを開き、セル **G4** に「=PERCENTILE.EXC(B4:B15,1-F4)」と入力して順位に対応する点数を求めよう。
作成例は［上位 10%（PERCENTILE.EXC 答え)］ワークシートを参照。PRECENTRANK.EXC 関数で求められる位置の最小値より小さい値や、最大値より大きい値を指定するとエラーになることに注意。

PERCENTILE.EXC 関数の第 1 引数にはデータの範囲を、第 2 引数には下位からの割合（パーセント）を指定します。この場合は上位 **10%**に当たる値を求めたいので、第 2 引数の指定は、**1** から **10%**を引いて、下位から **90%**としています。関数の実行結果は **93.5** となります。つまり **93.5** 点を取れば上位 **10%**に入れるということです。

PERCENTILE.INC 関数の場合も、関数名が違うだけで、引数の指定方法は同じです。［上位 10％（PERCENTILE.INC）］ワークシートを開いて、セル **G4** に「=PERCENTILE.INC(B4:B15,1-F4)」と入力してみてください。結果は **89.2** となります。

これらの関数の違いも、PERCENTILE.EXC 関数では **0％**と **100％**を除いた範囲で値を求め、PERCENTILE.INC 関数では **0%**と **100%**を含めた範囲で値を求めるということです。これについては、前回説明した四分位数を求めるための QUARTILE.EXC 関数や QUARTILE.INC 関数の計算方法と同様です。四分位数の場合は **25%**や **75%**を指定して計算しましたが、PERCENTILE.EXC 関数や PERCENTILE.INC 関数ではそれ以外の値も指定できるというわけです。詳細については、前回のコラムで解説しているので、そちらをご参照ください。

このようにして得られた値を**パーセンタイル値**と呼びます。図 6 の例であれば、**93.5** が **90 パーセンタイル値**ということになります。PERCENTILE.EXC 関数や PRECENTILE.INC 関数の第 2 引数に **25%**を指定した場合は第 1 四分位数が、**75%**を指定した場合は第 3 四分位数が求められます。それぞれ QUARTILE.EXC 関数や QUARTILE.INC 関数の第 2 引数に**1**や**3**を指定した場合と同じ結果です。なお、PERCENTILE.EXC 関数と PRECENTILE.INC 関数についても、データ数が多ければ、ほぼ同じ値が返されるので、位置を求めるという目的において、実用上での違いはほとんどありません。

前回紹介した統計ソフトの R（オープンソースで無料の統計解析向けプログラミング言語およびその開発実行環境）の quantile 関数は、実は四分位数だけを求める関数ではなく、パーセンタイル値を求めるのにも使えます。その場合、第 2 引数に下位からの割合（パーセント）を指定します。引数に type=6 を指定すると PERCENTILE.EXC 関数と同じ計算が行われ、type 引数を指定しないと PRECENTILE.INC 関数と同じ計算が行われます。

これまでに見てきた関数は、全て順位を基にしたものです。しかし、試験の成績などの間隔尺度のデータは平均値の近くに値が集まっているのが一般的なので、単純に順位や比率で表すよりも、分布を反映したような値を求めたいこともあると思います。例えば、平均点が **60 点**の試験であれば、**60 点**から **70 点**あたりには多くの人がいますが、同じ **10 点**の幅でも **90 点**から **100 点**の人はそれほど多くないと考えられます。そこで、分布を想定し、どの位置までにどれぐらいの人がいるかを求めていくことにします。

平均値と標準偏差の間にどれぐらいの人がいるの？

身長や知能テストの成績、各種測定値の誤差などは、一般に**正規分布**と呼ばれる分布に従う（当てはまる）ものと考えられます。正規分布とは、以下の式で表される分布です。

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad (1)$$

πは円周率、μが平均値（母集団の平均）、σは標準偏差（母集団の標準偏差）です。exp(X) は、自然対数の底 e ≈ 2.71821 の X 乗を表します。

この式の意味については、ここでは詳しく触れません。まずは、後で見るグラフをきちんと読み取れるようにしておきましょう。なお、式の意味は「AI・機械学習の数学入門」の第 14 回で詳しく解説しています。興味のある方はぜひご参照ください（動画での説明もこちらにあります）。

数式だけではイメージが湧かないかもしれないので、(1) 式を基に作成したグラフを眺めてみましょう。ここでは、令和元（2019）年の国民健康・栄養調査（厚生労働省）の結果（Excel ファイルのダウンロード）から引用した 20 歳代男性の身長の平均値と標準偏差を利用します。平均値は **171.5** で、標準偏差は **6.6** となっています（以下、身長のデータは全てこの資料から引用しています）。

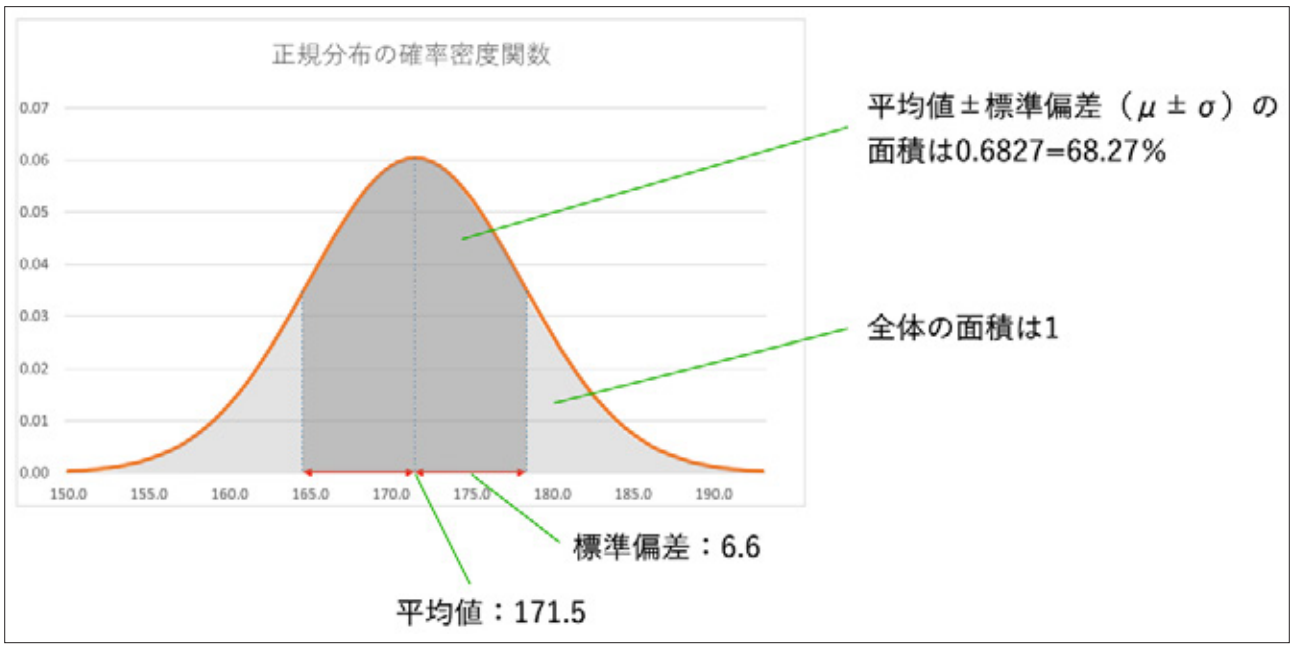


図 7　正規分布のグラフ（確率密度関数）

このグラフ（オレンジの線）は正規分布の**確率密度関数**と呼ばれる。確率密度関数は、平均値の位置で山が最も高く、左右に裾野が広がる形になっている。横軸の値は、理論的には−∞〜∞までだが、取り得る値の範囲に限定して考えるのが一般的（身長なら広く見積もって **0cm 〜 300cm**、試験の成績なら **0 〜 100** など）。縦軸の値については後述。

図 7 にも示しましたが、正規分布では、グラフと X 軸で囲まれた範囲（アミカケ全体）の面積は **1**、つまり **100%**となります。また、$\mu \pm \sigma$の範囲（アミカケが濃くなっている部分）に全体の **68.27%**が含まれ、$\mu \pm 2\sigma$の範囲に全体の **95.45%**が含まれることが分かっています。重要なことなので、箇条書きでもう一度整理しておきます。

- $\mu \pm \sigma$の範囲：全体の **68.27%**
- $\mu \pm 2\sigma$の範囲：全体の **95.45%**

この例であれば、全体の **68.27%**が身長 **164.9cm** から **178.1cm** であるということが分かります。また、全体の **95.45%**が **158.3cm ～ 184.7cm** となります。

確率密度関数を見る上での留意点が 1 つあります。それは、**確率密度関数の縦軸の値は確率ではない**ということです。例えば、x=170.0 のとき、f(x) の値は **0.0589** ですが、これは身長が **170cm** である確率が **0.0589** だということではありません（この値は次に示す累積分布関数の微分係数に当たる値です）。

x の値に対する f(x) **までの面積**を関数 F(x) として表すと、図 8 の右のグラフができます。この関数は**累積分布関数**と呼ばれます。全体の面積が **1** なので、F(x) の値は、x までの値を取る確率になることが分かります。正規分布などの連続分布では、値 x に対する確率は求められませんが、ある値 x **までの**確率は累積分布関数によって求めることができます。

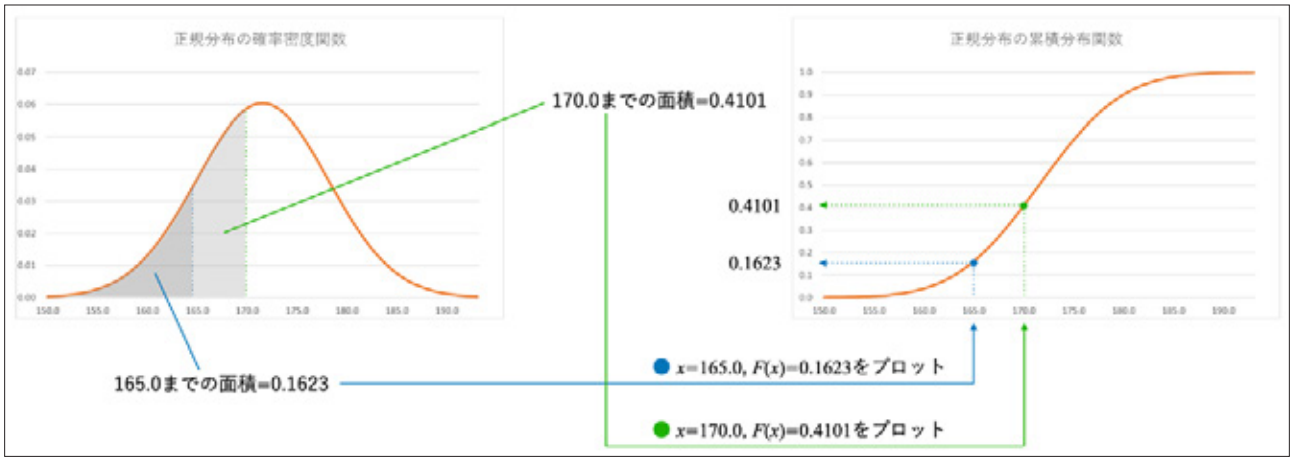


図 8　正規分布のグラフ（確率密度関数と累積分布関数）

例えば、x = 165.0 までの面積（左のグラフの濃いアミカケの面積）は **0.1623** となっており、x = 170.0 までの面積（左のグラフのアミカケ全体の面積）は **0.4101** となる。これらの値をプロットしていったものが、右側の累積分布関数のグラフとなる。これは確率密度関数を積分した値をプロットしたものと考えられる。

累積分布関数では、横軸が x を表し、縦軸が累積確率 F(x) を表します。従って、この例では、横軸が身長、縦軸がその身長**まで**の人の割合となります。身長 **165.0cm** までの人は全体の **16.23%**、170.0cm までの人は全体の **41.01%**であることが累積分布関数のグラフから読み取れます。

20 歳代の男性で身長 175cm なら全体のどのあたり？

ここまでで、グラフの見方は分かったと思いますが、身長が 170.0cm のときの確率密度関数の 0.0589 という値や、そのときの累積分布関数の 0.4101 という値はどうやって求められるのでしょうか。(1) 式を思い出してください。(1) 式の x に 170.0 を代入すれば、確率密度関数 f(x) の値 0.0589 が求められます。また、その式を−∞～170.0 まで積分すれば、累積分布関数 F(x) の値 0.4101 が求められます。とはいえ、この計算はかなり難しいですね。例えば、次に身長が 175.0cm のときの累積分布関数の値を知りたい（身長が 175.0cm なら全体のどのあたりの位置にいるかを知りたい）と思っても、簡単には計算できそうにありません。

そこで、Excel の NORM.DIST 関数の出番です。この関数を使えば、x の値と平均値μ、標準偏差σを指定するだけで、確率密度関数や累積分布関数の値が求められます。では、サンプルファイル（06d.xlsx）をこちらからダウンロードし、［累積分布の値］ワークシートを開いて以下の操作を行いましょう。具体的な操作については動画でも解説しています。

図 9 に NORM.DIST 関数を入力する手順を示します。

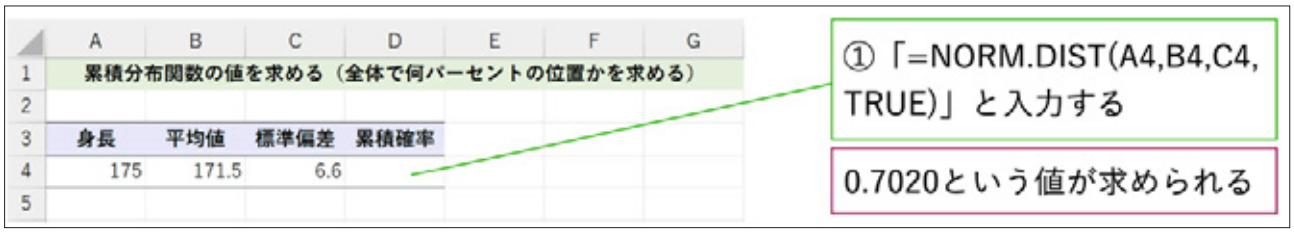


図 9　身長 175cm の人がどのあたりにいるかを NORM.DIST 関数で求める

［累積分布の値］ワークシートを開き、セル **D4** に「=NORM.DIST(A4,B4,C4,TRUE)」と入力しよう。結果は **0.7020** となる。
作成例は［累積分布の値（答え）］ワークシートを参照。

NORM.DIST 関数の引数には、x の値、平均値、標準偏差、関数の形式（**FALSE** なら確率密度関数、**TRUE** なら累積分布関数）を指定します。この例では、セル **D4** に「=NORM.DIST(A4,B4,C4,TRUE)」と入力すれば、**0.7020** という結果が得られます。つまり、身長 **175cm** の人は身長の小さい方から数えて全体の **70.2%**の位置にいることが分かります。ここでは、確率密度関数の値を求めても「どの位置にいるか」という目的には合わないので、最後の引数に **TRUE** を指定して累積分布関数の値を求めました。

実は図 7 や図 8 のグラフは、NORM.DIST 関数に **150 ～ 193** までの値を **1** 刻みで順に指定していき、確率密度関数の値や累積分布関数の値を求めて、折れ線グラフにしたものです（散布図を使っても作成できます）。グラフの作成例もサンプルファイル（06d.xlsx）の末尾にある［正規分布の確率密度関数］ワークシートと［正規分布の累積分布関数］ワークシートに含まれているので、興味のある方はご参照ください。

なお、ここで求めた正規分布の累積確率は、**左側確率**または**下側確率**と呼ばれることもあります。また、1 − 累積確率は、**右側確率**または**上側確率**とも呼ばれます（正規分布以外の分布でも同じように呼ばれます）。

正規分布で全体の上位 10%に入るための身長は？

これまでに見てきた「身長が何 cm ならどのあたりの位置にいるのか」とは逆に「どのあたりの位置にいる人は身長何 cm なのか」を知りたいこともあります。つまり、x の値から F(x) の値を求めるのとは逆に F(x) の値から x の値を求めたい、ということです。Excel にはそのような場合に使える NORM.INV 関数がちゃんと用意されています。

では、先ほどダウンロードしたサンプルファイル（06d.xlsx）の［累積分布の逆関数］ワークシートを開いて以下の操作を行いましょう。この操作についても動画で解説しています。

図 10 に NORM.INV 関数を入力する手順を示します。

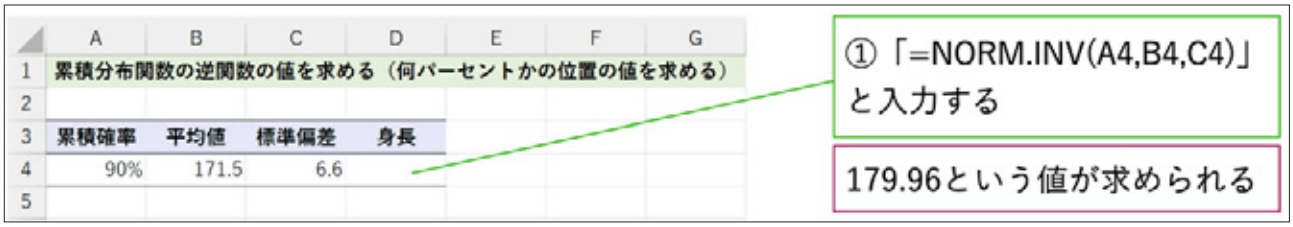


図 10　全体の 90%の位置に当たる身長はいくからを求める

［累積分布の逆関数］ワークシートを開き、セル **D4** に「=NORM.INV(A4,B4,C4)」と入力しよう。上位 **10%**ということは、累積確率が **90%**の位置となる。関数の実行結果は **179.96** となる。
作成例は［累積分布の逆関数（答え）］ワークシートを参照。

NORM.INV 関数の引数には、累積確率、平均値、標準偏差を指定します。この例では、セル **D4** に「=NORM.INV(A4,B4,C4)」と入力すると、**179.96** という結果が得られます。つまり、上位 **10%**に入るための身長は **179.96cm** であることが分かります。累積確率は、下位から何パーセントであるかを指定することに注意してください。

言うまでもないことかもしれませんが、身長が上位 **10%**に入っているかどうかというのは、あくまでも身長という物理的な値が大きいか小さいということであって、いいか悪いかという価値判断とは全く関係のないことです。

20 歳代の 170cm と 60 歳代の 170cm ではどちらが高身長？ ～ 偏差値

ところで、20 歳代で **170cm** の男性と 60 歳代の **170cm** の男性がいたとき、どちらの方が身長が高いと言えるでしょう。どちらも同じ値だから同じ、というのも一つの答えですが、20 歳代男性の平均値が **171.5cm**、60 歳代男性の平均値が **167.4cm** であることから、集団の中での位置を考えると、60 歳代男性の **170cm** の方が高いと言えそうです。

このことは、平均値の違いだけでなく、標準偏差の違いによっても変わってきます。実は、20 歳代男性の平均値も 30 歳代男性の平均値も **171.5** ですが、20 歳代男性の標準偏差は **6.6**、30 歳代男性の標準偏差は **5.5** となっています。30 歳代男性の方が、わずかですが平均値の近くに集まっているということですね。ということは、平均値から離れた値は 30 歳代男性の方が少ないことになります。例えば、**175cm** という身長であれば、30 歳代男性の方が高いことになります。

平均値や標準偏差が異なる集団同士でも、値がどのあたりの位置にあるかを比較したい場合に便利なのが、標準化変量や偏差値です。

標準化変量は、データの値 x_i から平均値μを引き、標準偏差σで割った値です。数式で表すと以下のようになります。

$$\frac{x_i - \mu}{\sigma}$$

このような計算を行うと、平均値が **0**、標準偏差が **1** の分布になるように値が調整されます。そうすれば、元の分布の平均値と標準偏差が異なっていても比較ができるというわけです。

重回帰分析やロジスティック回帰などの機械学習でも、学習を効率よく行ったり、回帰式の係数を比較しやすくするために標準化を行うことがあります（こちらに詳しい説明があります）。

ただ、標準化変量の値は日常の感覚ではちょっとイメージしづらい値になります。そこで、標準化変量に **10** を掛けて **50** を足した値もよく使われます。それが**偏差値**です。数式で表すと、以下のようになります。

$$\frac{x_i - \mu}{\sigma} \times 10 + 50$$

偏差値は平均値が **50**、標準偏差が **10** になります。試験の成績に比較的近いイメージなので、感覚的にも分かりやすいですね。

では、20 歳代男性で **170cm** の場合と 60 歳代男性で **170cm** の偏差値を求めてみましょう。先ほどダウンロードしたサンプルファイル（06d.xlsx）の［偏差値］ワークシートを開き、上の式に従って図 11 のように計算を行いましょう。平均値と標準偏差は表に入力されています。具体的な操作については動画でも解説しているので、ぜひご参照ください。

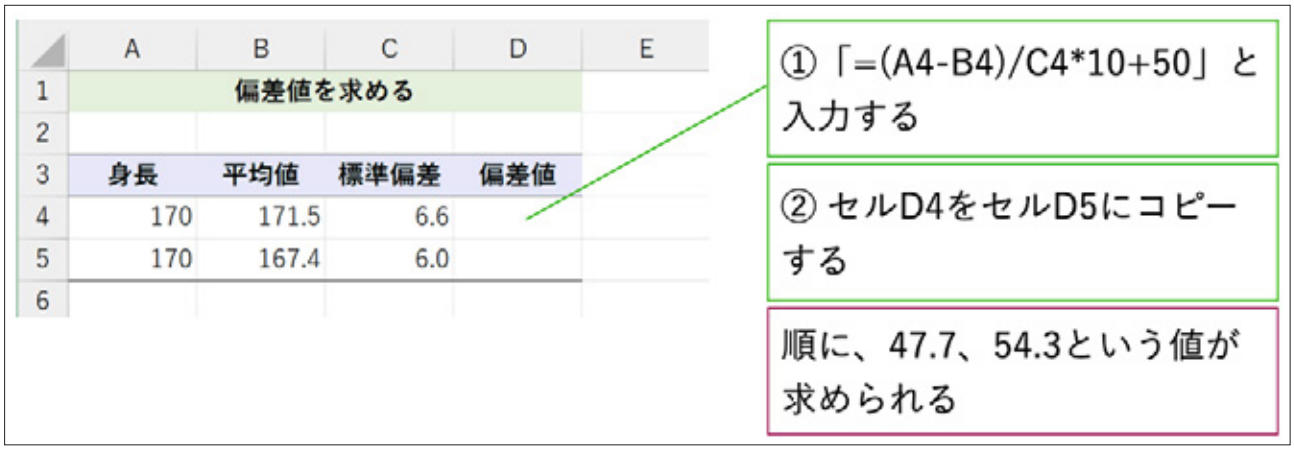


図 11　偏差値を求める

［偏差値］ワークシートを開き、セル **D4** に「=(A4-B4)/C4*10+50」という式を入力し、セル **D5** にコピーする。答えはそれぞれ **47.7**、**54.3** となる。
作成例は［偏差値（答え）］ワークシートを参照。

セル **D4** に「=(A4-B4)/C4*10+50」という式を入力し、セル **D5** にコピーすると、それぞれの偏差値が求められます。結果は、20 歳代男性で **170cm** の人の偏差値が **47.7**、60 歳代男性で **170cm** の人の偏差値が **54.3** となるので、60 歳代男性で **170cm** の人の方が（集団の中では）身長が高いと言えます。なお、図 11 では、数式通りに計算を行いましたが、Excel には標準化変量を求めるための STANDARDIZE 関数も用意されています。従って「=STANDARDIZE(A4,B4,C4)*10+50」と入力しても同じ結果が得られます。関数を使うとかえって式が長くなりますが、標準化を行っているということはよく分かります。［偏差値（答え）］ワークシートには STANDARDIZE 関数を使った例も含めてあります。

偏差値は入試の難易度など、高校や大学のランク付けによく使われるので、あまりいい印象が持たれていないかもしれませんが、このように、異なる集団の間でも位置を比較するのに使えるとても便利な値なのです。

……というわけで、今回は、各データの位置を知ることに焦点を当て、単純な順位だけでなく、パーセント単位での順位や偏差値を求めました。逆に、何パーセントの範囲に入るには何点を取る必要があるかといった計算の方法についても見てきました。次回以降、何回かに分けて、グラフを利用した可視化によってデータを分析する方法を見ていきます。

次回は、規模や効果の大きさやその差を可視化するために使われる棒グラフについて、基本から応用までを見ることにします。棒グラフは誰もがすでに知っている基本中の基本ですが、意外な落とし穴もあります。そういったことについても触れることにします。では、次回もお楽しみに！

この記事で取り上げた関数の形式

関数の使いこなし方については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

順位を求めるために使った関数

RANK.EQ 関数：順位を求める（同じ値は同じ順位とする）

形式

RANK.EQ(数値 , セル範囲 , 順序)

引数

- **数値：**順位を求めたい値を指定する。
- **セル範囲：**データが入力されているセル範囲を指定する。
- **順序：0** を指定するか省略すると降順に、**0 以外の数値**を指定すると昇順に並べたときの順位を返す（Excel や Google スプレッドシートでは、**0** は **FALSE**、**0 以外**は **TRUE** と見なされるので、降順の場合には **FALSE** または省略、昇順の場合は **TRUE** と考えてもよい）。

RANK.AVG 関数：順位を求める（同じ値は上位と下位の平均とする）

形式

RANK.AVG(数値 , セル範囲 , 順序)

引数

- **数値：**順位を求めたい値を指定する。
- **セル範囲：**データが入力されているセル範囲を指定する。
- **順序：0** を指定するか省略すると降順に、**0 以外の値**を指定すると昇順に並べたときの順位を返す（Excel や Google スプレッドシートでは、**0** は FALSE、**0 以外**は **TRUE** と見なされるので、降順の場合には **FALSE** または省略、昇順の場合は **TRUE** と考えてもよい）。

パーセント単位での順位やパーセンタイル値を求めるために使った関数

PERCENTRANK.EXC 関数：パーセント単位での順位を求める（0%と 100%は含まない）

形式

PERCENTRANK.EXC(データの並び , 値 , 有効桁数)

引数

- **データの並び**：データ全体を指定する。
- **値**：パーセント単位での順位を求めたい値を指定する。
- **有効桁数**：結果を小数点以下何桁まで求めるかを指定する。指定した桁以降は切り捨てられる（Google スプレッドシートでは四捨五入される）。省略すると **3** が指定されたものと見なされる。

PERCENTRANK.INC 関数：パーセント単位での順位を求める（0%と 100%を含む）

形式

PERCENTRANK.INC(データの並び , 値 , 有効桁数)

引数

- **データの並び**：データ全体を指定する。
- **値**：パーセント単位での順位を求めたい値を指定する。
- **有効桁数**：結果を小数点以下何桁まで求めるかを指定する。指定した桁以降は切り捨てられる（Google スプレッドシートでは四捨五入される）。省略すると **3** が指定されたものと見なされる。

PERCENTILE.EXC 関数：パーセンタイル値を求める（0%と 100%は含まない）

形式

PERCENTILE.EXC(データの並び , 率)

引数

- **データの並び**：データ全体を指定する。
- **率**：求めたいパーセンタイル値の位置を **0** ～ **1** の範囲で指定する。ただし、最小値の位置より小さい値や最大値の位置より大きな値を指定するとエラーになる。

PERCENTILE.INC 関数：パーセンタイル値を求める（0%と 100%を含む）

形式

PERCENTILE.INC(データの並び , 率)

引数

- **データの並び**：データ全体を指定する。
- **率**：求めたいパーセンタイル値の位置を **0** ～ **1** の範囲で指定する。

正規分布の確率密度関数や累積分布関数を求めるために使った関数

NORM.DIST 関数：正規分布の確率密度関数や累積分布関数の値を求める

形式

NORM.DIST(値 , 平均 , 標準偏差 , 関数の形式)

引数

- **値：**確率密度関数や累積分布関数の `x` に当たる値を指定する。
- **平均：**分布の平均値を指定する。
- **標準偏差：**分布の標準偏差を指定する。
- **関数の形式：FALSE** を指定すると確率密度関数の値を求める。**TRUE** を指定すると累積分布関数の値を求める。

NORM.INV 関数：正規分布の累積分布関数の逆関数の値を求める

形式

NORM.INV(確率 , 平均 , 標準偏差)

引数

- **確率：**累積分布関数の累積確率（左側確率）を指定する。
- **平均：**分布の平均値を指定する。
- **標準偏差：**分布の標準偏差を指定する。

標準化のために使った関数

STANDARDIZE 関数：標準化変量を求める

形式

STANDARDIZE(値 , 平均 , 標準偏差)

引数

- **値：**標準化したい値を指定する。
- **平均：**分布の平均値を指定する。
- **標準偏差：**分布の標準偏差を指定する。

［データ分析］グラフの種類と使い分け ～データ可視化入門【特別予告編】

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の特別予告編。次回から数回に分けてグラフを利用した可視化の方法を見ていきます。それに先だって、今回は可視化の目的と手法を概観します。「何を見たい」→「どのグラフを使うのか」→「何がうれしいのか」という流れをひととおり確認し、次回以降のお話にスムーズに入れるようにします。

羽山博（2023 年 08 月 17 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の特別予告編です。第 6 回（前回）までは、平均値や標準偏差などの値を求めることによって、データを分析する方法を見てきました。特に、前回は、集団の中での位置を知るために、パーセント単位での順位や偏差値などを求めました。

次回からは「可視化による分析」をテーマとして、何回かに分けて幾つかの事例を見ていきます。話の内容や展開がこれまでと少し変わるので、今回は特別予告編として、可視化の目的とそのために利用するグラフの種類などを概観します。次回以降の具体的な内容にスムーズに入れるよう、可視化の方法を大局的に捉えておきましょう。

問題意識と分析の目的、可視化の方法について ～ ケーススタディーを中心に

可視化の方法と分析の流れを追いかけていくにあたって、まずはその全体像を眺めておきましょう。具体的なお話は次回からですが、簡単な事例を取り上げ、可視化と分析を進めていくことになります。内容としては、グラフ作成に関するソフトウェアの機能を網羅するというよりは、可視化の目的やデータをどう捉えるか、グラフをどう読み解くかといった「考え方」が中心となります。目的やデータの種類、形式によってアプローチの方向や可視化の方法は千差万別ですが、1 つのケースを追いかけることによって、考え方を知る糸口をつかもうというわけです。

私たちは、データ分析の目的や手法うんぬんの前に、何らかの「問題意識」を持っているはずです。

- どちらの Web サイトに広告を出せば効果的なのか
- ここ数年、営業成績が下がってきているのではないか
- 不良品が発生する原因は何か
- どうすれば不良品を減らせるのか
- 商品の価格に最も影響があるのはどのような要因か
- 商品の評価に極端な賛否両論があるのではないか
- すぐにお客様に提供できる商品はどれか

……などなど、枚挙にいとまがありません（ちなみに、みなさんはご自分の業務でどのような問題意識をお持ちでしょうか）。

「問題意識」は、分析の目的に直結します。そこで、目的に対してどのグラフを使えば、適切な可視化ができるのかを次にまとめておきます。それぞれの目的や可視化の方法ごとに分析の流れを確認しておこうというわけです。次回以降、ケーススタディーを通して、データの扱い方やグラフ作成の手順、分析の観点などを具体的に追いかけます。

規模や効果の差を比較するには棒グラフ

Web サイトに広告を出すという業務の例であれば、どの Web サイトに広告を出せば効果的なのかを知りたいといった問題意識が湧き起こってきます（図 1 の左側）。そのような例では、それぞれの Web サイトでの売上を比較する**棒グラフ**を作成するのが可視化による分析の典型的な方法です。それにより、規模や効果の差を見ることができます（図 1 の右側に示したデータ分析の効用）。

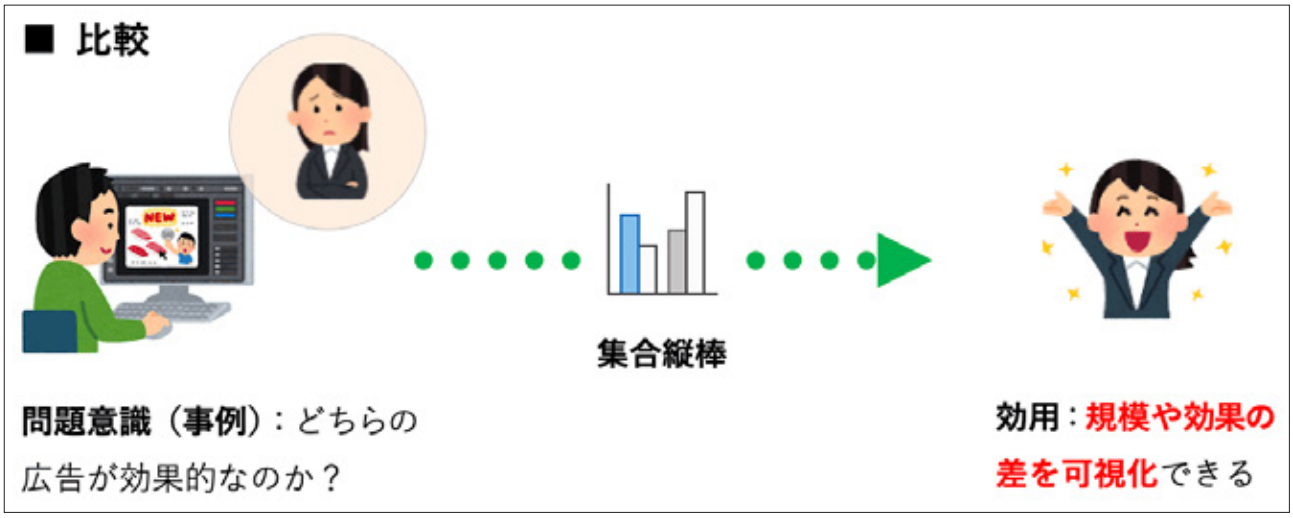


図 1　規模や効果の差を可視化したいときには棒グラフが適している

図の左上にある「比較」というのは、分析の目的を端的に表したキーワード。広告の効果を比較したい場合は、それぞれの Web サイトでの売上を比較すればよい。そのためには棒グラフを使うのが定石。なお、中央のグラフのアイコンとして、Excel の［挿入］タブの［グラフ］グループの中で選択すべきボタンを示してある（以降も同様）。

次回のテーマは棒グラフですが（慣れ親しんだ棒グラフとあなどるなかれ、です。グラフ化に当たっての前処理が必要になったり、意外な落とし穴などもあります……乞うご期待です）、連載では、さらに以下のような事例を何回かに分けて取り扱います。図の見方は同じなので、以降は、簡単に目的や事例など列挙するにとどめます。

時系列での変化を見るには折れ線グラフ

横軸を時間、縦軸を売上などの数値として、**折れ線グラフ**を作成すれば、時系列での変化を可視化できます。その際、1 つの系列をグラフにするだけでなく、複数の系列を比較することも重要です。例えば、日本の GDP と諸外国の GDP の変化を見れば、諸外国と比べて日本の景気が良くなっているのか悪くなっているのかが分かります。また、賃金と物価の変化を見れば、賃金は上がっていないのに物価だけが上がっているなどの特徴も見られます（もちろん、別のパターンになるかもしれません）。

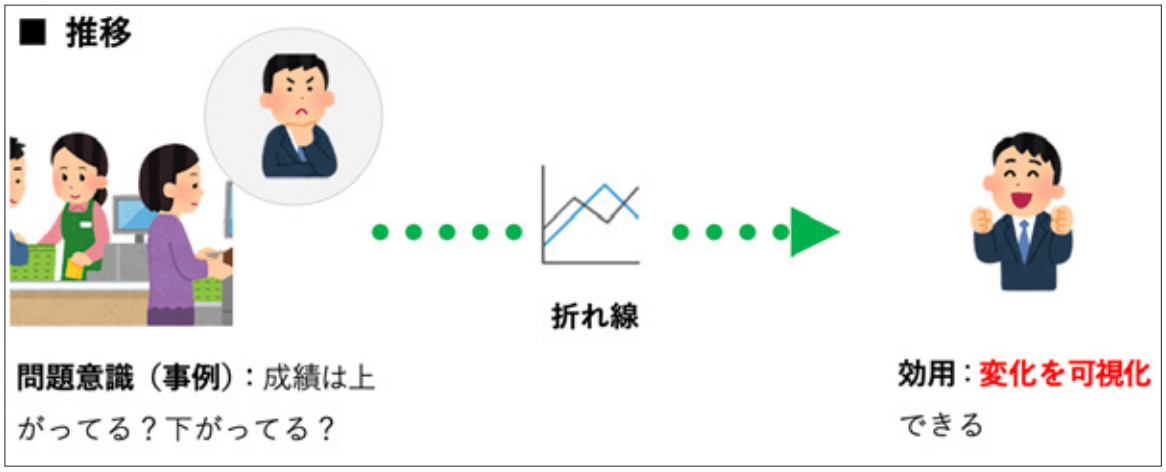


図 2　時系列での変化を可視化したいときには折れ線グラフが適している

折れ線グラフでは変化を見ることができる。売上や収入、株価の分析などによく使われる。単純にデータを折れ線グラフにするだけでなく、比較することにより目的のデータの特徴を明確にしたり、移動平均を求めることによりトレンドを見ることもできる（これも次回以降のお楽しみです）。

全体の中での割合 ＝ 重要度を見るには円グラフ／パレート図

円グラフも棒グラフや折れ線グラフと同様、小学校の算数などから長年慣れ親しんでいる基本的なグラフです。円グラフは、割合（比率）を可視化するために利用します。ある項目が全体の中でどれぐらいの割合を占めるのかが分かれば、その項目の重要度が分かります。売上に貢献している商品はどれか、故障の原因の大半は何か、といったことが分かり、以降の方針を策定するのに役立てることができます（グラフを作成しただけで安心してしまって、重要度という観点を欠いてしまうこともありがちです。要注意ですね）。

パレート図は、そういった方針の策定に役立てることを強く意識したグラフです。パレート図は、重要度をランク付けする ABC 分析に使われます。

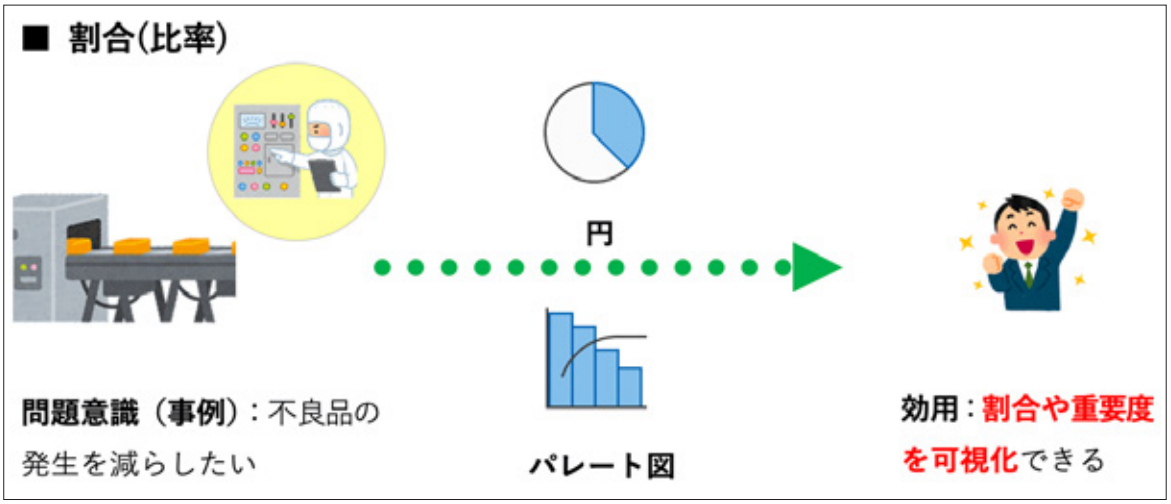


図 3　割合（比率）を基に重要度を可視化したいときには円グラフやパレート図が適している

円グラフでは割合（比率）を見ることができる。政党の支持率、商品のシェア、不良品の原因の割合などを分析するのに使われる。

パレート図では、上位 70%を占める項目や上位 90%を占める項目などを可視化できる。それにより重要度を A、B、C の 3 つのランクに分け、方針を策定するのに役立てることができる（**ABC 分析**と呼ばれる）。

→目次に戻る

中心の位置や広がり（分布）を見るにはヒストグラム

分布を可視化するために使われる**ヒストグラム**については、連載の第 3 回で取り上げました。

また、**箱ひげ図**については第 5 回で取り上げました。

いずれも詳細な作成方法は割愛しましたが、分布を可視化することにより、中心の位置を知ったり、データのばらつきを見たり、さらには、外れ値を発見するために使われることについて詳しく説明しました。そのため、次回以降では、そういった話のおさらいもしつつ、ヒストグラムや箱ひげ図の作成手順、グラフの見せ方を変える方法などに重点を置くことにします。

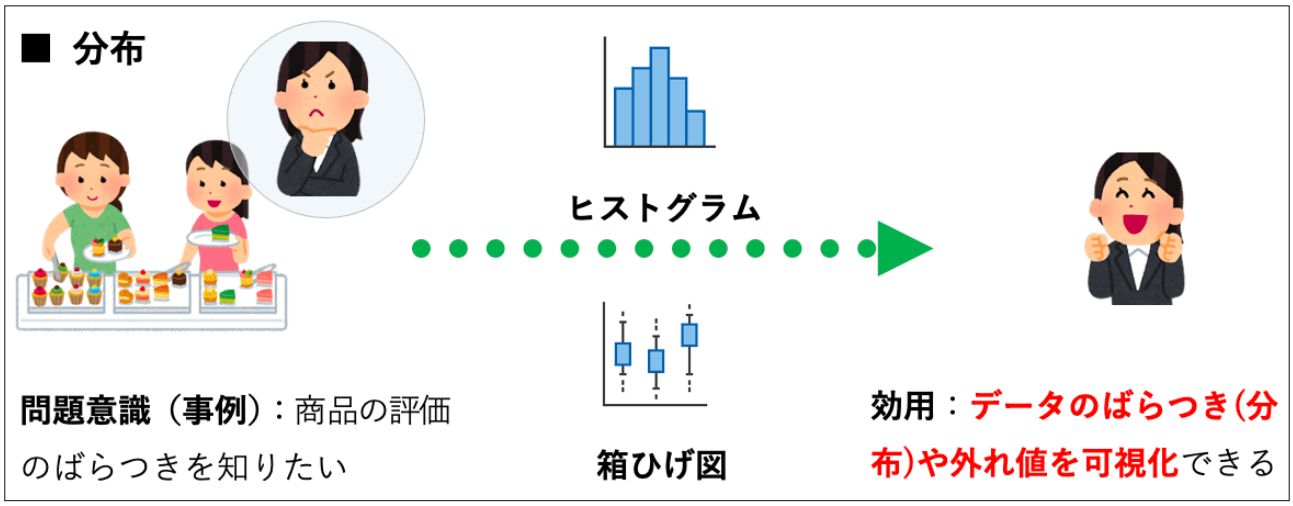


図 4　分布を可視化したいときにはヒストグラムや箱ひげ図が適している

ヒストグラムでは中心の位置や集中の度合い、データの広がりが分かる。
箱ひげ図では大半のデータがどの範囲にあるかが分かる。いずれも、外れ値を見つけるのに役立つ。

どの位置の値が大きいかを見るにはヒートマップ

ヒートマップは、値の大きさにより色を変えたグラフです。Excel ではグラフ機能にヒートマップは含まれていませんが、条件付き書式を使うと簡単に作成できます。例えば、中古車の在庫状況を可視化し、どの年式のどの価格帯の在庫が豊富なのかを一目で分かるようにしたり、重回帰分析を行うに当たって、相関行列の値が大きい部分を可視化したりするのに使われます（似た項目を重複して使っていないかを調べることができます。重回帰分析はこの連載の終盤で取り扱うので、そちらもお楽しみに）。また、グループ分けしたデータの特徴をつかむのにも便利です（『数学 ×Python プログラミング入門』の第 5 回で、クラスター分析の結果をヒートマップで可視化する例を紹介しています）。

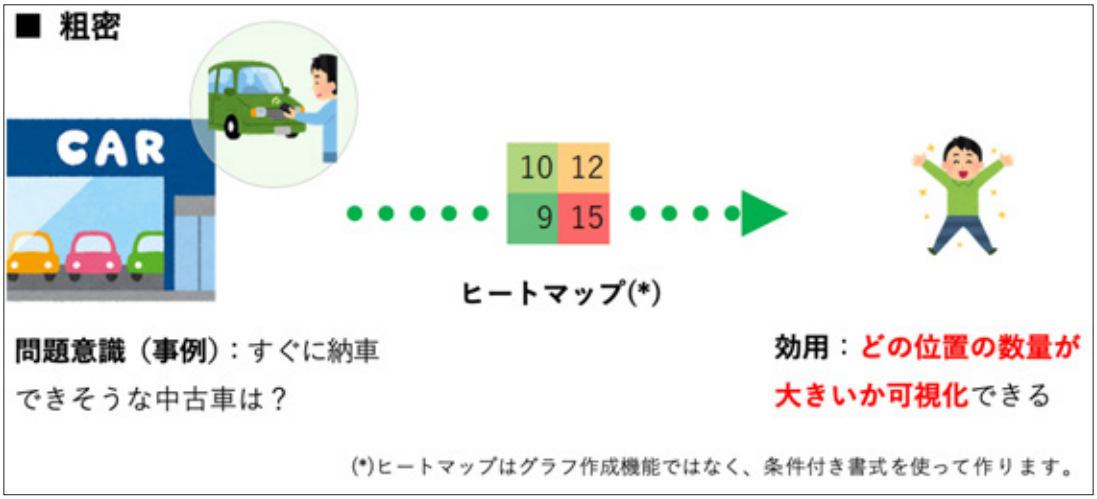


図 5　どの位置の値が大きいかを可視化するにはヒートマップが適している

条件付き書式を利用すれば、値の大きさによってセルを色分けできる。値の大きな位置（や小さな位置）が可視化できるので、どこにデータが集中しているのかを調べたり、項目同士の関係の強さを見たりすることができる。中央のアイコンは Excel のボタンではなく、ヒートマップのイメージを示したもの。

項目同士の関係を見るには散布図

散布図は項目同士の関係を可視化するのに使われます。例えば、中古車の年式と価格の関係、気温とビールの売上の関係、年齢と給与の関係など、さまざまな関係が可視化できます。（やはりこの連載の後半で取り扱う）相関係数を求めると、関係の強さが数値として表されますが、その場合でも散布図を作っておくと、直線的な関係なのか、指数関数的な関係なのかといったことが直感的に読み取れます（数値だけでは分かりにくいです）。また、散布図も外れ値の発見に使えます。

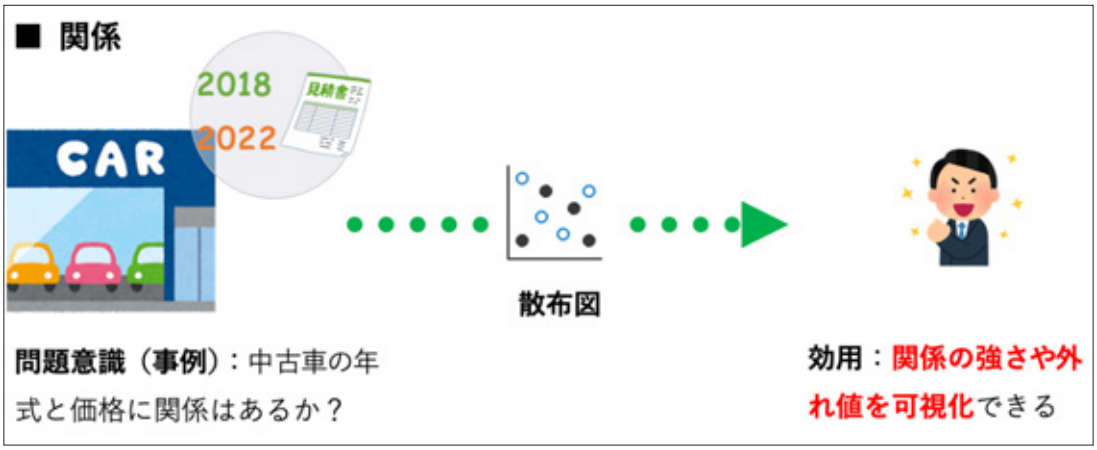


図 6　項目同士の関係を可視化するには散布図が適している

散布図は、横軸（X）に当たる項目の値と縦軸（Y）に当たる項目の値の交わった位置に「点」をプロットしたグラフ。例えば、X が **2018**（年）で、Y が **100**（万円）なら **(2018, 100)** の位置に点を表示する。多くの点がプロットされると、その形から関係が分かる（例えば、右上がりになっていれば、一方が増えれば他方も増える）。極端に離れた位置に点があれば、外れ値と考えられる。

次回からのお話で、分析の目的に合ったグラフはどれなのか、前処理としてデータをどのように加工すればいいのか、作成されたグラフをどう読み解くのか……といった可視化によるデータ分析の流れや考え方がケーススタディーを通してひととおり学べます。というわけで、次回からの新展開をお楽しみに！

［データ分析］棒グラフで「規模や効果」を可視化 ～どちらの広告が効果的なのか？

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 7 回。グラフを利用して規模や効果の差、つまり大きさの差を可視化する方法や、考え方などについて説明します。具体的には棒グラフを使いますが、慣れ親しんだ棒グラフでも、作成時の準備や意外な落とし穴など、改めて考慮すべき点がたくさんあります。

羽山博（2023 年 08 月 31 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 7 回です。前回の特別予告編では、可視化の目的と利用するグラフ、その効用などについて整理しました。幾つかの事例を取り上げ、ケーススタディーを通して、分析の流れと考え方を追いかけるということでしたが、今回はその第 1 弾です。

まずは、規模や効果を可視化するために、棒グラフの活用について考えていくことにします。これまで何気なく使っていたグラフかもしれませんが、作成する際の前処理や誤った使い方など、考慮すべき点が幾つもあります。なお、Excel のあまり知られていないテクニックや関数についても併せて紹介します。

サンプルファイルの利用について

本稿では、表計算ソフトを使って手を動かしながら学んでいきます。表計算ソフト Microsoft Excel 用の .xlsx ファイルをダウンロードできるようにしています。デスクトップ版の Excel が手元にない場合は、Microsoft アカウントがあれば使える無料の Microsoft 365 オンライン、もしくは Google アカウントがあれば使える無料の Google スプレッドシート（Google Sheets）をお使いください。Microsoft 365 オンラインの場合は、.xlsx ファイルを OneDrive にアップロードしてから開いてください。Google スプレッドシートの場合は、.xlsx ファイルを Google ドライブにアップロードしてから開いた上で［ファイル］メニューの［Google スプレッドシートとして保存］を実行してください（Google スプレッドシート独自の機能を使っている場合は、ファイルを共有して参照できるようにします。その場合は、該当する箇所で使い方を記します）。

たかが棒グラフとあなどるなかれ ～ 売り上げの差を可視化してみよう

データの特徴を可視化するための方法として、おそらく誰もが最も慣れ親しんでいるのが棒グラフでしょう。数値の大きさに合わせて棒の長さを変えるだけなので、考え方に難しいところはなさそうです。実際、棒グラフなら小学生の頃から何度も作成したことがあると思います。2017 年告示、2020 年全面実施の学習指導要領（PDF ファイル）でも小学 3 年生で学ぶことになっています。しかし、収集されたデータを棒グラフとして表すには、手描きにしても、ソフトウェアを使うにしても、幾つかのステップを踏む必要があります。また、分析を行う上でも、意外に奥の深いところがあります。

では、始めましょう。図 1 をご覧ください。これは、ユーザーが Web サイトの広告を最初にクリックした日のクリック数とその広告から得られた売り上げの一覧です（架空のデータです）。同じユーザーが複数回同じ広告をクリックすることもあるかもしれませんが、ここでは最初にクリックした日を基準に回数をカウントしています（2 回目以降は数えないものとします）。このデータを基に、サイト A に広告を出した場合とサイト B に広告を出した場合の売り上げを比較できるグラフを作成し、**どちらのサイトの広告が有効なのか**分析してみましょう。

Web サイトの利用とその効果を分析するためには、Google Analytics などのツールが利用できます。さまざまなデータの収集や可視化、予測などを行うことができるので、実務ではそういったツールを活用するのが一般的です。ここでは、ツールの利用方法ではなく、データ分析の「ツボ」に触れることが目的なので、Excel などの身近なツールを利用し、簡単な事例を分析するプロセスを追いかけます（事例はきわめて単純化したものなので、データの値は現実離れしたものになっているかもしれませんが、ご容赦のほど）。

なお、Google Analytics について知りたい方は公式ページや公式の無料オンライン講座、『Google アナリティクス 4 のやさしい教科書。』（山野勉著、MdN）、『Google アナリティクス 4 設定・分析のすべてがわかる本』（小川卓著、ソーテック社）などの書籍をご参照ください。

サンプルファイルをこちらからダウンロードし、［売上一覧］ワークシートを開いて取り組んでみてください。Google スプレッドシートの場合はこちらのサンプルファイルを開いてメニューバーの［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください（曜日の表示形式を Google スプレッドシートに合わせた形式にしてあるだけで、内容は同じです）。

	A	B	C	D	E	F	G
1	媒体別広告売上一覧						
2							
3			クリック数（回）		売上（千円）		
4	日付	曜日	サイトA	サイトB	サイトA	サイトB	
5	7月24日	月	94	50	94	284	
6	7月25日	火	130	87	180	84	
7	7月26日	水	137	135	181	107	
8	7月27日	木	87	60	170	225	
9	7月28日	金	246	171	360	77	
10	7月29日	土	175	132	346	281	
11	7月30日	日	217	179	360	378	
12							

図 1　広告の回数と売上金額のデータ

売上金額は E 列と F 列に入力されている。このデータを基に、売り上げの差を比較するための棒グラフを作成してみよう。本来なら長期にわたるデータが必要だが、簡単にするため 1 週間分のデータとしてある。なお、必要に応じて計算を行うなど、データを加工しても構わない。

さて、皆さんはどのようなグラフを作成したでしょうか。最初に考えるべきことは、どのデータをグラフにするかということですね。取りあえず売上金額をそのまま棒グラフにした方は、図 2 のようなグラフになったのではないかと思います。グラフを見ると曜日ごとの売り上げの傾向が分かっていいかも……と思われるかもしれませんが、ここでの目的はあくまでもどちらのサイトの広告が有効なのかを知りたいということです。もちろん、曜日ごとの傾向を見ることも重要ですし、これはこれで興味深いのですが、図 2 のグラフだとサイトの優劣が分かりにくいですね。

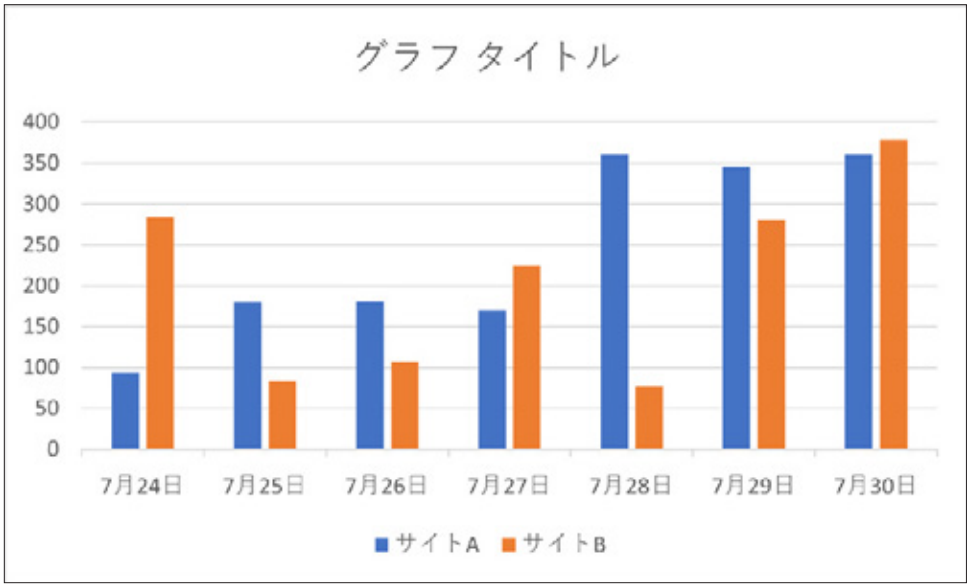


図 2　売り上げをそのまま棒グラフにしたもの（売り上げ全体は比較できない）

系列の見出しとデータ、横軸の項目を選択して単純に棒グラフを作った例。グラフタイトルなどの設定は一切変更していない。サイト A の売り上げが週の後半に上がっており、サイト B の売り上げが週の最初と最後に上がっているような印象は得られるが、サイト A とサイト B の売り上げ全体を比較するのにはあまり適してはいない。作成例は［売上一覧（日付ごとのグラフ）］ワークシートに含まれている。

目的に合ったグラフを作る前に、図 2 のグラフをどのようにして作成したかを説明しておきます（ここでの目的からは外れますが、作成手順を知ることはグラフ作成のスキルを上げる上で役に立つので）。なお、以下の手順と、以降の図 7 までのグラフ作成方法や考え方については、動画でも解説しています。Excel での手順を丁寧に追いかけたい方はぜひご視聴ください。

- セル **A4 ～ A11** をドラッグして選択する（これが横（項目）軸のラベルになる）
- ［Ctrl］キーを押しながら、セル **E4 ～ F11** をドラッグして選択する（これらが系列の見出しとデータになる）
- ［挿入］タブを開き、［縦棒／横棒グラフの挿入］－［集合縦棒］を選択する
    - Google スプレッドシートの場合は、メニューから［挿入］－［グラフ］を選択し、［グラフエディタ］の［グラフの種類］のリストから［縦棒グラフ］を選択する

棒グラフを作成するときには、横（項目）軸のラベルと、系列の見出しとデータを範囲指定する必要があります。この例では横（項目）軸のラベル（セル **A4 ～ A11**）と系列の見出しとデータ（セル **E4 ～ F11**）が**離れた位置にあることに注意が必要です。離れた複数の範囲を選択するには［Ctrl］キーを押しながら、それぞれの範囲をドラッグ**します（意外に知られていない操作ですね）。

Excel では、セル **E4 ～ F11** をドラッグして棒グラフを作成するだけでも同様のグラフが表示できます。しかし、横（項目）軸に対する値が設定されないので、横軸に日付ではなく **1**、**2**、……といった連番が表示されます。

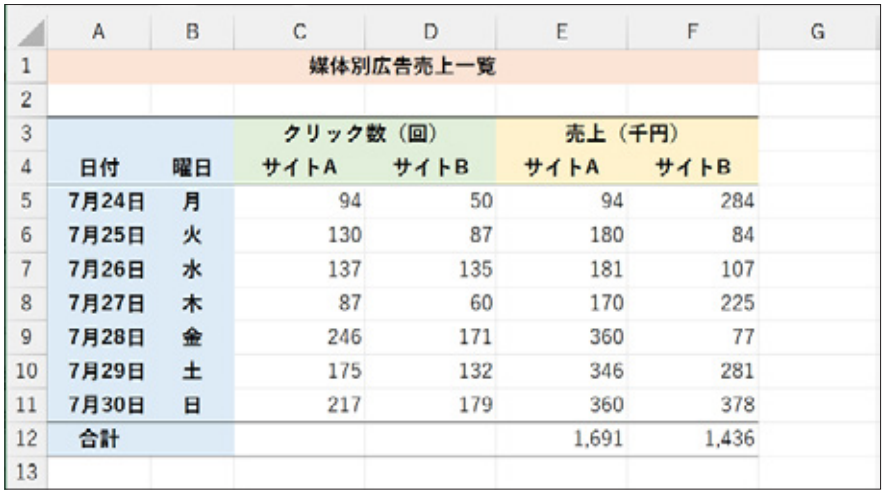
Google スプレッドシートでは、セル **E4 ～ F11** をドラッグして棒グラフを作成すると、E 列が X 軸の値と見なされてしまい、F 列だけのグラフになってしまいます（Google スプレッドシートでは、Excel の「横（項目）軸」は「X 軸」と呼ばれます）。その場合は［列 E をラベルとして使用］チェックボックスをオフにすれば、E 列もグラフ化する系列として扱われ、E 列と F 列のグラフが表示されます。

試行錯誤的にいろいろなグラフを作っているうちに、思いも寄らなかった発見があることもまれではありません。しかし、目的に合った可視化の方法を選べば、効率よく分析が進められるのも事実です。目的を意識していれば、見通しがつくようになります。ムダに試行錯誤しなくても、さまざまな角度から分析ができるようになるはずです。……というわけで、回りくどくなってしまいましたが、次に、売り上げ全体を比較できるようなグラフを作ってみることにします。

グラフの種類を選択するにあたっては、横軸がどのような項目であるかを意識しておく必要があります。横軸が日付を表す値であり、時系列での**変化**を見たいのであれば、折れ線グラフの方が適しています。また、横軸も縦軸も変数であり、それらの**関係**を見たい場合には、散布図が適しています（横軸が X、縦軸が Y になります）。サイト A とサイト B の値を比較したい場合（後述）のように、横軸が**カテゴリ**を表す場合は棒グラフが適しています。

グラフ作成のための前処理を行う（1）～ 売り上げを集計してからグラフ化する

サイト A とサイト B の売り上げを比較するなら、あらかじめ売上金額を集計しておく必要があります。**分析の目的に合った可視化を行うためには、集計が必要になる場合がある**ということです。集計の方法には、SUM 関数を使う方法、集計機能を使う方法、ピボットテーブルを使う方法がありますが、ここでは SUM 関数を使って合計を求めておきましょう。SUM 関数については説明するまでもないと思いますが、この連載では一応初出なので、最後に関数の形式をまとめてあります。



	A	B	C	D	E	F	G
1	媒体別広告売上一覧						
2							
3			クリック数（回）		売上（千円）		
4	日付	曜日	サイトA	サイトB	サイトA	サイトB	
5	7月24日	月	94	50	94	284	
6	7月25日	火	130	87	180	84	
7	7月26日	水	137	135	181	107	
8	7月27日	木	87	60	170	225	
9	7月28日	金	246	171	360	77	
10	7月29日	土	175	132	346	281	
11	7月30日	日	217	179	360	378	
12	合計				1,691	1,436	
13							

図 3　売上金額を集計した表

セル **E12** に「=SUM(E4:E11)」と入力し、セル **F12** にコピーする。これらの合計の値をグラフ化するとよい。これまで何気なくグラフを作成してきた人も多いかもしれないが、グラフを作成する前には、集計などの前処理が必要になることを意識しておくことも重要。

図 3 を作成する手順は以下の通りです。

- セル **E12** に「=SUM(E4:E11)」と入力する
- セル **F12** にコピーする

この表は、上でダウンロードしたファイルの［売上集計］ワークシートに含まれているので、それを基に売り上げが比較できるグラフを作成してみましょう（図 4）。手順は図 4 の後に箇条書きで示しています。

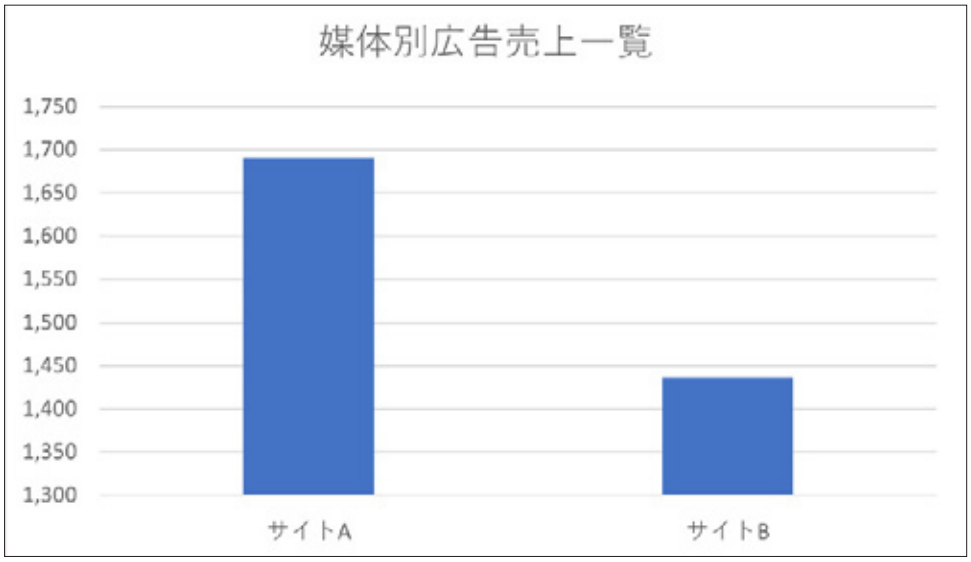


図 4　売上金額の比較ができるグラフを作成する

合計だけをグラフ化する。セル **E4** ～ **F4** とセル **E12** ～ **F12** を選択して棒グラフにすればよい。売り上げを集計した時点でも想像は付くが、グラフを見れば、サイト A の方が売り上げの合計が大きいことがよく分かる。……が、実はまだ落とし穴がある（後述）。作成例は［売上集計（合計のグラフ）］ワークシートに含まれている。

- セル **E4** ～ **F4** をドラッグして選択する（これが横（項目）軸のラベルになる）
- ［Ctrl］キーを押しながら、セル **E12** ～ **F12** をドラッグして選択する（これらが系列のデータになる）
- ［挿入］タブを開き、［縦棒／横棒グラフの挿入］－［集合縦棒］を選択する

Google スプレッドシートの場合は、以下のような操作で同様のグラフが作成できます。

- セル **E12** ～ **F12** をドラッグして選択する（これらが系列のデータになる）
- メニューから［挿入］－［グラフ］を選択し、［グラフエディタ］の［グラフの種類］のリストから［縦棒グラフ］を選択する（この段階では各日付のグラフも表示される）
- ［グラフエディタ］の［行と列を切り替える］チェックボックスをオンにする
- ［グラフエディタ］の［X 軸を追加］をクリックし、右端に表示される［データ範囲の選択］ボタンをクリックする
- セル **E4** ～ **F4** をドラッグして選択し、［OK］ボタンをクリックする（これが X 軸のラベルになる）

セル **E4** ～ **F12** を選択して棒グラフを作成し、5 行目～ 11 行目を非表示にするという方法でも同様のグラフが作成できます（ただし、Google スプレッドシートでは、［行と列を切り替える］チェックボックスをオンにし、［列 E を見出しとして使用］チェックボックスをオフにしておく必要があります）。

→目次に戻る

実際のところ、金額を集計した段階でサイト A に広告を出した方が、売り上げが大きいことは分かりますが、可視化するとそのことが顕著に分かります。が、しかし、です。この可視化には落とし穴があります。皆さんはどう思われるでしょうか。すでに答えが分かっている方もおられるかとは思いますが、サイト A の広告クリック数の方が多いので、サイト A からの売り上げが大きいのは当然といえば当然です。確かに集客に関してはサイト A が勝っていますが、**広告の効果を見るなら、1 クリック当たりの売上金額を見ておく必要があります。**

同様の例は枚挙にいとまがありません。例えば、GDP（国内総生産）ではなく 1 人当たり GDP を見る、人口ではなく人口密度（人口 /km^2）を見る、都道府県別感染者数ではなく、都道府県の人口 1000 人当たりの感染者数を見る……などです。他にもどんな例があるか考えてみるといいですね。

グラフ作成のための前処理を行う（2）～ 広告 1 クリック当たりの売り上げをグラフ化する

では、1 クリック当たりの売上金額を求めて、グラフを作成してみましょう。グラフ化に必要な値は図 5 のセル **G12** とセル **H12** の値だけですが、セル **G5** ～ **H11** に毎日の値も併せて求めてあります。［売上集計（1 回あたり）］ワークシートを基にグラフを作成してみてください。念のため、図 5 の作成方法も図の後に箇条書きで示しておきます。

	A	B	C	D	E	F	G	H	I
1	媒体別広告売上一覧								
2									
3			クリック数（回）		売上（千円）		1クリックあたりの売上		
4	日付	曜日	サイトA	サイトB	サイトA	サイトB	サイトA	サイトB	
5	7月24日	月	94	50	94	284	1.00	5.68	
6	7月25日	火	130	87	180	84	1.38	0.97	
7	7月26日	水	137	135	181	107	1.32	0.79	
8	7月27日	木	87	60	170	225	1.95	3.75	
9	7月28日	金	246	171	360	77	1.46	0.45	
10	7月29日	土	175	132	346	281	1.98	2.13	
11	7月30日	日	217	179	360	378	1.66	2.11	
12	合計		1,086	814	1,691	1,436	1.56	1.76	
13									

図 5　クリック 1 回当たりの売上金額を求める

G 列と H 列に、1 クリック当たりの売り上げを求める。「売上／クリック数」で計算すればよい。グラフ作成のためにはセル **G12** の値（**1.56**）とセル **H12**（**1.76**）の値を使えばよい。

- セル **C12** に「=SUM(C5:C11)」と入力する
- セル **C12** をセル **D12** にコピーする
- セル **G5** に「=E5/C5」と入力する
- セル **G5** をセル **H5** にコピーする
- セル **G5** ～ **H5** を選択し、セル **G12** ～ **H12** までコピーする

上記のコピー操作を行うと、表の下の方にあらかじめ設定されている書式が崩れてしまいます。そのような場合には、コピーした後に表示される［貼り付けオプション］ボタンをクリックして［書式なしコピー］を選択すると、元の書式に戻ります。

図 5 のように集計した段階で、1 クリック当たりの売り上げではサイト B が勝っているように思われます。続いてグラフを作成してみましょう（図 6）。手順は、図 4 で売上金額をグラフ化した例とほぼ同じです。

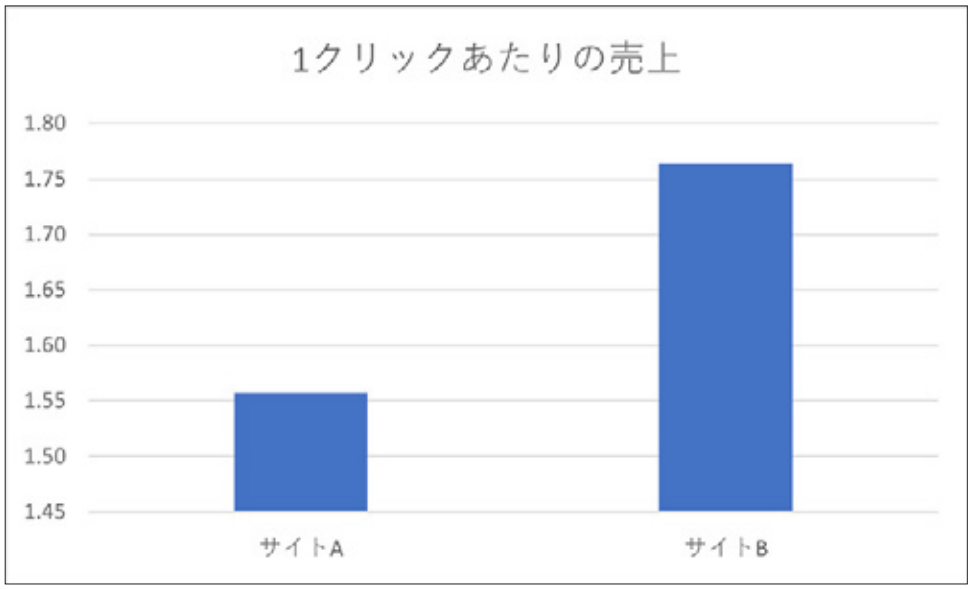


図 6　クリック 1 回当たりの売上金額をグラフ化する（差が強調されたグラフ）

1 クリック当たりの売上金額はサイト A が **1.56**、サイト B が **1.76** なので、サイト B の方が優秀であることが読み取れそうである。とはいえ、1 週間のデータだけでは、1.76 − 1.56 = 0.2 （**200 円**）という差は「たまたま」だったかもしれないので、これだけで差が大きいのか小さいのかは明確には分からない。作成例は［売上集計（1 回あたりグラフ）］ワークシートに含まれている。

- セル G4 ～ H4 をドラッグして選択する（これが横（項目）軸のラベルになる）
- ［Ctrl］キーを押しながら、セル **G12** ～ **H12** をドラッグして選択する（これらが系列のデータになる）
- ［挿入］タブを開き、［縦棒／横棒グラフの挿入］－［集合縦棒］を選択する

Google スプレッドシートの場合は、以下のような操作になります。

- セル **G12** ～ **H12** をドラッグして選択する（これらが系列のデータになる）
- メニューから［挿入］－［グラフ］を選択し、［グラフエディタ］の［グラフの種類］のリストから［縦棒グラフ］を選択する（この段階では各日付のグラフも表示される）
- ［グラフエディタ］の［行と列を切り替える］チェックボックスをオンにする
- ［グラフエディタ］の［X 軸を追加］をクリックし、右端に表示される［データ範囲の選択］ボタンをクリックする
- セル **G4** ～ **H4** をドラッグして選択し、［OK］ボタンをクリックする（これが X 軸のラベルになる）

売上金額はサイト A の方が大きいですが、図 6 のグラフを見ると、1 クリック当たりの売り上げはサイト B の方が大きいので、「サイト B の広告の方がユーザーに届いている」と確信を持って言えそうです。しかし、ここにも大きな落とし穴があります。次に、そのことについて見ていきましょう。

棒グラフの大きな大きな落とし穴 ～ 目盛りの取り方に注意

Excel で特に何も指定せずにグラフを作成すると、縦（数値）軸の目盛りが自動的に設定されます。そのため、ごく小さな差が強調されすぎて、あたかも大きな差があるように見えることがあります。逆に、目盛りの設定を変えれば、大きな差があっても、あまり差がないように見せることもできてしまいます（いずれにしても悪用禁止ですね）。これがグラフによる可視化の大きな落とし穴の一つです。

この例では、1 週間しかデータを取っていないので、「差がある」と言うにはかなり無理があります。しかし、図 6 のグラフだけを見せられると「サイト B の方が優れている」という印象が与えられてしまいます。では、誤ったイメージを伝えないためにはどうすればいいでしょうか。縦（数値）軸の目盛りを適切に設定すればいいですね。そこで、縦（数値）軸の目盛りの最小値を **0**、最大値を **2** に設定してグラフを描いてみましょう。手順は図 7 の後に箇条書きで示してあります。ちなみに、Google スプレッドシートでグラフを作成した場合には、特に何も指定しなくても、最小値が **0**、最大値が **2** のグラフ（図 7 と同様のグラフ）になります。利用するソフトウェアによって、可視化した際の印象が異なることもあるので、目盛りなどの設定はソフトウェア任せにせず、自分で変更できるようにしておきたいものです。

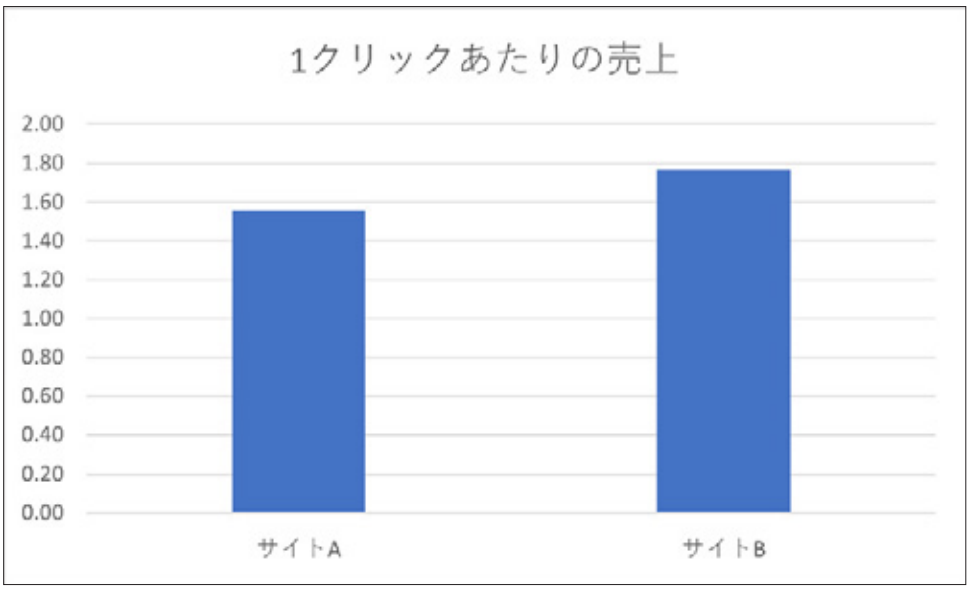


図 7　クリック 1 回当たりの売上金額をグラフ化する（目盛りを修正する）

縦（数値）軸の目盛りの最小値を **0**、最大値を **2** に設定したグラフ。差がありそうな傾向はうかがえるが、それほど大きな差とは感じられない。こちらのグラフの方が「誠実」な可視化と言える。作成例は［売上集計（1 回あたりグラフ目盛修正）］ワークシートに含まれている。

- 縦（数値）軸を右クリックし、［軸の書式設定］を選択する
- ［最小値］に **0** を入力し、［最大値］に **2** を入力する

→目次に戻る

人は得てして自分の信念に合うようにモノの見方を変えてしまうものです。差があることを主張したいからといって、意識するとしないとに関わらず、図 6 のようなグラフを作成するのはご法度です。この例はあまりにもあからさまなので、まんまとだまされてしまう人はいないかもしれません。しかし、世の中には印象操作のために巧妙に加工したようなグラフも数多くはびこっています。

T.TEST 関数を使って、いずれかのセルに「=T.TEST(G5:G11,H5:H11,1,3)」と入力すると、**0.17** という t 検定の結果が得られます。この値が **0.05 以下**あるいは **0.01 以下**であれば「サイト B の方が大きい」と言えますが、そうでないので「サイト B の方が大きいとは言えない」ことになります（ただし、差がないとも言い切れません）。なお、一般に、データ数が多くなれば T.TEST 関数の値は小さくなるので、わずかな差であっても「差がある」という結果が得られてしまいます。また、その差は対象となる問題によっても重要度が変わってきます。生命に関わるような領域であれば、ごくわずかな差でも無視できないかもしれません。適切に判断するためには、効果量、検出力、サンプル数の検討が必要になりますが、統計的検定のお話は、今回のテーマから大きく外れるので、ここでは割愛します。興味のある方は『事例で学ぶ Excel 統計』（羽山博著、日経 BP）の第 5 章などをご参照ください。

可視化のメリットは数値だけでは分からない特徴が明確になることですが、逆に、それほど大きくない差異を実態よりも大きく見せてしまうことがあるというデメリットも理解しておきましょう。

コラム　費用対効果も考慮する必要がある

仮にサイト B の方が広告の効果が高いということが分かったとしても、サイト B にのみ広告を出せばいいかというとそういうわけでもありません。現実問題として、費用対効果（コストパフォーマンス）を考える必要があります。例えば、サイト B の広告料が高いのであれば、費用対効果は小さくなってしまいます。

例えば、広告の予算が **10,000 円**で、ユーザーが広告にアクセスしたときの広告料が表 1 のようになっている場合、いずれかのサイトに全て広告予算をつぎ込めば売り上げが最大化されます（ここでは、同じユーザーが何回広告をクリックしても広告料は変わらないものとします）。

ケース	広告料（A）	広告料（B）	クリック数（A）	クリック数（B）	売上（A）	売上（B）
ケース1	1.0	1.1	10,000	9,091	15,600	**16,000**
ケース2	1.0	1.2	10,000	8,333	**15,600**	14,667

表 1　広告料と売り上げの関係（[売上（A）] と [売上（B）] の単位は千円）

ケース 1 では、サイト B に広告料を全てつぎ込んだ方が有利（**16,000 千円**の売り上げ）。ケース 2 では、サイト B の広告料が少し高くなるので、サイト A に広告料を全てつぎ込んだ方が有利（**15,600 千円**の売り上げ）。

この例はあまりにも単純で、制約となるのが広告予算だけなので、どちらか一方に広告を集中させればいいというつまらない結果になってしまいますが、複数の制約がある場合には線形計画法により最適な配分を求めることができます。ちなみに、上の表の値は以下の Python プログラムで求められます（**シンプレックス法**と呼ばれる方法が使われます）。広告料が表 1 のケース 1 のようになっている場合の例です。

```
from scipy.optimize import linprog

c = [-1.56, -1.76]  # 目的関数:-1.56x0-1.76x1（最小値が求められるので、符号を変えておく）
A = [[1.0, 1.1]]  # 制約関数（広告料）:1.0x0+1.1x1
b = [10000] # 制約量:1.0x0+1.1x1 ≦ 10000

x0_bounds = (0, None)  # x0 の下限と上限
x1_bounds = (0, None)  # x1 の下限と上限

res = linprog(c, A_ub=A, b_ub=b, bounds=(x0_bounds, x1_bounds))
print(res.fun)  # 最小値のみを出力

# 出力例（符号を変えれば最大値となる）
-15999.999999999998  # 小数点以下は誤差。最大値は 16000 となる
```

リスト 1　広告料によって売り上げがどう変わるかを Python で求める

scipy.optimize は、最適化問題を解くための関数を提供しているモジュールで、linprog は線形計画法を解くための関数。linprog では目的関数が最小となる変数 x_0, x_1 ... の値や、その場合の最小値などが求められる（上の例では目的関数の値を最大化したいので、係数の符号を変えてある）。A = [[1.0, 1.1]] を A = [[1.0, 1.2]] に書き換えると、ケース 2 の結果が求められる。ここでは、制約関数が 1 つだけなので極端な結果しか得られないが、一般には複数の制約関数を設定し、最適な配分を計算する。

コードの意味について興味のある方は、科学技術計算のためのパッケージである SciPy のドキュメントをご参照ください。

実際には、上のコラムでも触れた費用対効果はもちろんのこと、それぞれのサイトの性質や利用者の違いなどについても考慮しながら、どのように広告を出せばよいかを検討することになると思います。最後に、蛇足ですが、やや先走った話もしておきます。

データをもう少し詳しく見てみよう ～ 時系列データはタイムラグに注意

今回の目的はサイト A とサイト B の売り上げの比較なので、時系列による分析は行っていません。しかし、時間的な要因を分析すれば、より豊かな情報が得られる可能性があります。考えられることの一つとして**タイムラグ**があります。例えば、収入が増えたからといって、すぐに住宅や自動車などの大きな支出が増えるわけではありません。ある程度の期間が経過してから支出が増えることが考えられますね。

今回のデータにも実はタイムラグがあります。クリック数については、サイト A もサイト B も同じようなパターンですが、売り上げについてはサイト B が 1 日ずれたパターンになっています。そのことを確かめてみましょう。まず、列を作るための TOCOL 関数を使って、月曜日のクリック数が入力されている行（5 行目）に対して、火曜日の売り上げを表示する、といったように、クリック数に対して翌日の売り上げが表示されるようにしてみましょう（図 8）。便宜的に、日曜日のクリック数に対して、前の週の月曜日の売り上げが対応するようにしてあります。以降の操作については、動画でも解説しているので、手順を追いかけたい方はぜひご参照ください。

図 8　サイト B の売り上げを 1 日ずらしたグラフ

セル **G5** に「=TOCOL((F6:F11,F5))」と入力してサイト B の売り上げを 1 日ずらしたデータを作成する。例えば 7 月 24 日（月）のクリック数に対しては、7 月 25 日（火）の売り上げの値が表示されるようにする。ただし、7 月 30 日（日）のサイト B の売り上げは、便宜的にその前の週の 7 月 24 日（月）のものを使う。スピル機能により、セル **G5** ～ **G11** の全ての結果が表示される。

TOCOL 関数のかっこが 2 重になっているのは、F6:F11,F5 をひとまとめにして第 1 引数に指定するためです。内側のかっこがないと、第 1 引数として F6:F11 が指定され、第 2 引数として F5 が指定されたものと見なされるので、正しい結果が得られません（エラーになってしまいます）。なお、Google スプレッドシートでは {} で囲み、セミコロンで区切ることにより列（縦）方向の配列を作ることができます（カンマで区切ると行（横）方向の配列になります）。従って「={F6:F11;F5}」または「=TOCOL({F6:F11;F5})」と入力すれば同じ結果になります。

続いて、サイト A の売り上げ（E 列）とサイト B の 1 日後の売り上げ（G 列）をグラフ化しましょう（図 9）。作成手順は図 8 の後に箇条書きで示しておきます。

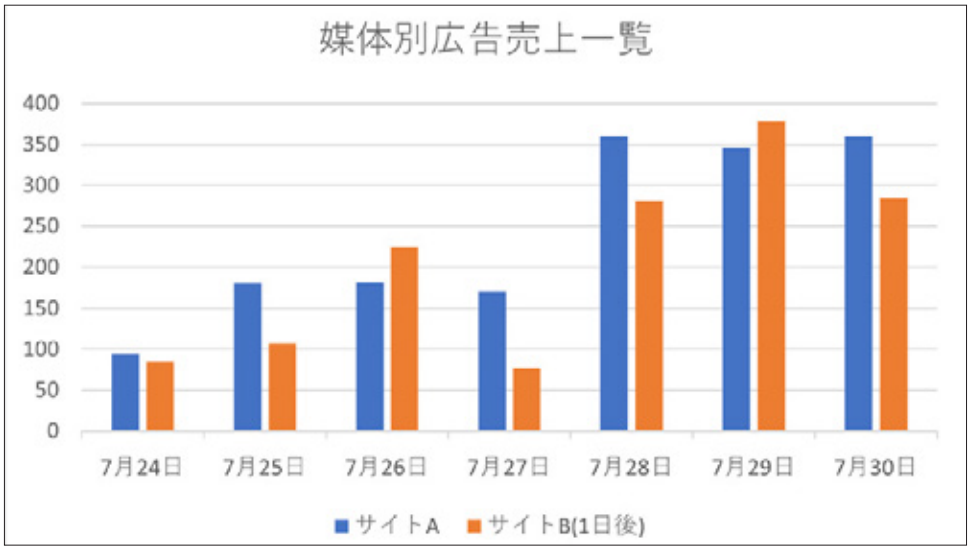


図 9　サイト B の売り上げを 1 日ずらしたグラフ

TOCOL 関数を使ってサイト B の売り上げを 1 日ずらしたデータを基にグラフを作成。例えば 7 月 24 日（月）のグラフ（オレンジの棒）はサイト B の 7 月 25 日（火）の売り上げの値を基にしたもの。ただし、7 月 30 日（日）のサイト B の売り上げはその前の週の 7 月 24 日（月）のもの。サイト A の当日売り上げとサイト B の翌日売り上げが似たようなパターンで推移していることが分かる。作成例は［売上一覧（日付をずらしたグラフ）］ワークシートに含まれている。

グラフ作成の方法は図 2 と同様です。ここでは棒グラフをそのまま使っていますが、変化のパターンを見るなら折れ線グラフの方が見やすいと思います（が、今回は棒グラフだけを取り扱うことにします）。

- セル **A4** ～ **A11** をドラッグして選択する（これが横（項目）軸のラベルになる）
- ［Ctrl］キーを押しながらセル **E4** ～ **E11** をドラッグして選択する（これが最初の系列のデータになる）
- ［Ctrl］キーを押しながらセル **G4** ～ **G11** をドラッグして選択する（これが次の系列のデータになる）
- ［挿入］タブを開き、［縦棒／横棒グラフの挿入］－［集合縦棒］を選択する
    - ・Google スプレッドシートの場合は、メニューから［挿入］－［グラフ］を選択し、［グラフエディタ］の［グラフの種類］のリストから［縦棒グラフ］を選択する

図 9 を見ると、サイト A では広告が最初にクリックされたその日に購買行動が起こり、サイト B では広告が最初にクリックされた翌日に購買行動が起こっているのではないかという仮説が立てられそうです。そこで、クリック数とその日の売り上げとの相関係数、クリック数と翌日の売り上げの相関係数をそれぞれ求めてみましょう。

では、[売上一覧（日付をずらした相関）] ワークシートを開いてみてください。相関係数については回を改めて詳しく解説するので、ここでは、セル **E14** ～ **F15** に入力されている関数とその結果を確認していただくだけでけっこうです（図 10）。相関係数を求めるには CORREL 関数を使います。

	A	B	C	D	E	F	G	H	I
1	媒体別広告売上一覧								
2									
3			クリック数（回）		売上（千円）				
4	日付	曜日	サイトA	サイトB	サイトA	サイトB	サイトA(1日後)	サイトB(1日後)	
5	7月24日	月	94	50	94	284	180.0	84.0	
6	7月25日	火	130	87	180	84	181.0	107.0	
7	7月26日	水	137	135	181	107	170.0	225.0	
8	7月27日	木	87	60	170	225	360.0	77.0	
9	7月28日	金	246	171	360	77	346.0	281.0	
10	7月29日	土	175	132	346	281	360.0	378.0	
11	7月30日	日	217	179	360	378	94.0	284.0	
12									
13				相関係数	サイトA	サイトB			
14				当日	**0.9178**	0.0107			
15				翌日	0.0307	**0.8331**			
16									

図 10　クリック数と当日の売り上げとの相関、翌日の売り上げとの相関を求める

相関係数を求めるには CORREL 関数を使う。引数にはクリック数と対応する売り上げを指定すればよい。サイト A では当日の売り上げとの相関が **0.9178** と高く、サイト B では翌日の売り上げとの相関が **0.8331** と高くなっている。

各セルに入力されている関数と作成手順は以下の通りです。

- セル **G5** に「=TOCOL((E6,E11,E5))」と入力する（サイト A の 1 日後の売り上げを求める）
- セル **G5** をセル **H5** にコピーする（サイト B の 1 日後の売り上げを求める）
- セル **E14** に「=CORREL(C5:C11,E5:E11)」と入力する（サイト A のクリック数と当日の売り上げとの相関係数を求める）
- セル **E15** に「=CORREL(C5:C11,G5:G11)」と入力する（サイト A のクリック数と翌日の売り上げとの相関係数を求める）
- セル **E14** ～ **E15** をセル **F14** ～ **F15** にコピーする（サイト B についても同様に相関係数を求める）

一方の変数の値が増えると他方の変数の値も増える場合、相関係数の値は **1** に近くなります（正の相関）。逆に、一方の変数の値が増えると他方の変数の値が減る場合、相関係数の値は **-1** に近くなります（負の相関）。変数同士に関係がないと考えられる場合には、相関係数は **0** に近くなります（無相関）。

結果を見ると、サイト A ではクリック数と当日の売り上げに強い正の相関があり、サイト B ではクリック数と翌日の売り上げに強い正の相関があることが分かります。やはり、サイト B では最初に広告がクリックされた翌日に購買活動が起こっているように思われます。つまり、サイトの性質に何らかの違いがあり、サイトを利用するユーザー層やユーザーに与える広告の影響が異なっているのではないか、ということが示唆されます。

実際のところ、最初に広告をクリックした日と商品を購入した日の記録や、ユーザーのプロフィールを詳細に追いかければ、このような傾向は読み取れます。しかし、洞察力を働かせれば、限られたデータからでもさらなる分析につながるヒントが得られることが分かると思います。逆に、高度なツールを利用していても、単に結果の数値を鵜呑み（うのみ）にしているだけでは本質に迫ることができないということも言えそうですね。

時系列データでは、日数や月数など、期間をずらしながら、自分自身との相関係数（自己相関）を求めていくと、周期的なパターン（季節性変動）が見いだされることがあります。そのような季節性変動による分析や予測も行われますが、それについてはいずれ機会を改めて、ということにしましょう。

今回は、規模や効果の可視化を行うために棒グラフを作成しました。単に操作を行うだけでなく、目的に合った分析のために、前処理を行ったり、作成されたグラフをどのように読み解くかについて考えたりしました。次回は、時間的な変化の可視化をテーマとしたケーススタディーを通して、折れ線グラフの作成や利用、読み解き方の留意点などを見ていきます。次回も、落とし穴や意外に知られていない機能なども紹介します。どうぞお楽しみに！

この記事で取り上げた関数の形式

関数の使いこなし方については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

合計を求めるために使った関数

SUM 関数：合計を求める

形式

SUM(数値 1, 数値 2, ... , 数値 255)

引数

数値： 合計を求めたい値を指定する。数値は 255 個まで指定できる。数値にセル範囲を指定すると、そのセル範囲の値が合計される。「=SUM(A1:A10)」とした場合、引数の個数はセルの個数（10 個）ではなく、1 個となる。

平均値の差の検定のために使った関数

T.TEST 関数：t 検定（平均値の差の検定）を行う

形式

T.TEST(配列 1, 配列 2, 検定の指定 , 検定の種類)

引数

- **配列 1 ／配列 2：** 平均値の差を検定したい値の並びを指定する。
- **検定の指定：1** なら片側検定、**2** なら両側検定。
- **検定の種類：1** なら対応のある **2** 群、**2** なら等分散の **2** 群、**3** なら非等分散の **2** 群。

指定した範囲のデータを 1 列に並べるために使った関数

TOCOL 関数：列を作る

形式

TOCOL(配列 , 無視する値 , 方向)

引数

- **配列:** 列を作りたいデータの並び。複数の範囲を指定するときには (A1:A3, C1:C3) のようにかっこで囲む必要がある。
- **無視する値：** 以下の値を指定する。
  - **0** または省略：全ての値を利用する。
  - **1**：空白を無視する。
  - **2**：エラーを無視する。
  - **3**：空白とエラーを無視する。
- **方向: FALSE** を指定するか省略すると行（横）方向に結合する。**TRUE** を指定すると列（縦）方向に結合する。[配列]が A1:B3 の場合、**FALSE** または**省略**であれば A1 → B1 → A2 → B2 → A3 → B3 という並びの列が作られる。**TRUE** の場合は、A1 → A2 → A3 → B1 → B2 → B3 という並びの列が作られる。

相関係数を求めるために使った関数

CORREL 関数：相関係数を求める

形式

CORREL(配列 1, 配列 2)

引数

- **配列：** 相関係数を求めたいそれぞれの値の並びを指定する。

［データ分析］折れ線グラフで「変化」を可視化 ～売り上げは本当に上がっているか？

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 8 回。グラフを利用して時間的な変化を可視化する方法と、それに関連するさまざまな考え方を追いかけます。具体的には折れ線グラフを使いますが、データの取り扱い、結果の見方などに関して、考慮すべき点や見落としがちな点について見ていきます。

羽山博（2023 年 09 月 14 日）

読者の皆さんは、データを見るだけで、売り上げや成績が「上がっている」あるいは「下がっている」と即座に判断できるでしょうか。データ分析に慣れた人であれば、まずはデータをグラフ化して目視で確認します。図 1 は、ある生徒の成績をグラフ化した図です。

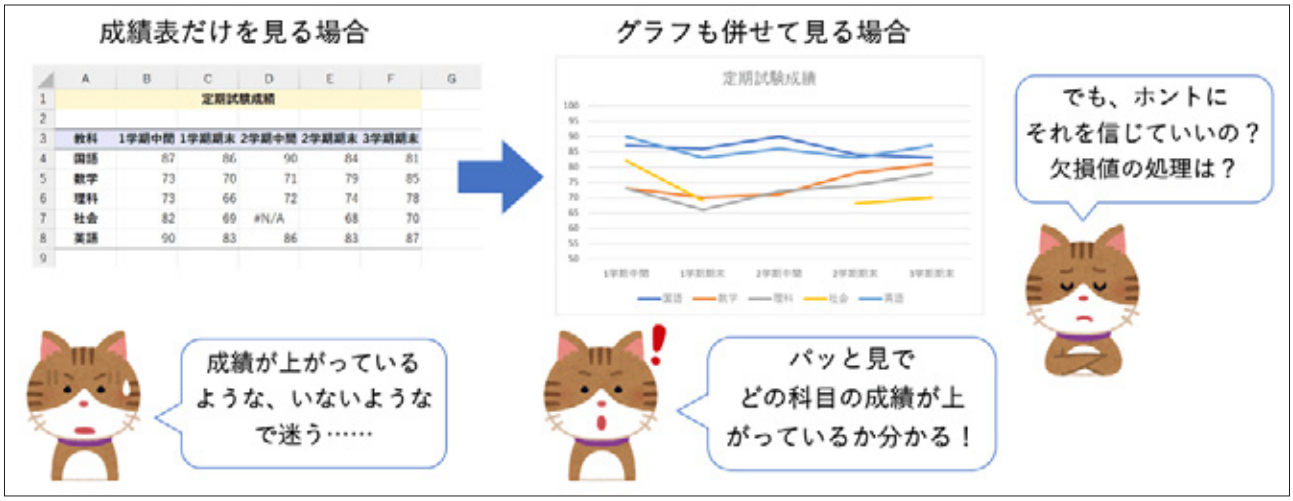


図 1　成績データをグラフ化して目視で確認

何の変哲もない、皆さんがよく使う「折れ線グラフ」ですね。意外に思われるかもしれませんが、折れ線グラフはデータ分析に必須で大活躍します。折れ線グラフを使えば、時間による売り上げや成績の変化が一目で分かるようになります。この記事では、折れ線グラフをどうやって役立てればよいかを詳しく説明します。

また、折れ線グラフの作り方がよくないと誤解を招くケースがあります。記事の後半では、誤解を避けるコツやさらなる応用の方法などを説明します。ぜひ無料会員登録して全文をお読みください。

この記事は、データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 8 回です。第 7 回の棒グラフから、今回の折れ線グラフ、円グラフ／パレート図、ヒストグラム、ヒストグラム、第 12 回の散布図まで、1 つずつ可視化の基礎を学んでいきます。これらグラフの目的と効用などについて、特別予告編で簡単に整理していますので、事前に確認しておくとより理解が深まるでしょう。

この記事で学べること

今回は以下のようなポイントについて、分析の方法や落とし穴を見ていきます。

- 折れ線グラフに潜む落とし穴 …… 簡単そうに見えて、見当外れなグラフに？
- 折れ線グラフでデータを比較 …… 世界の中で日本の地位はどう変化したか？
- 折れ線グラフにだまされるな …… 可処分所得は増えているのか？
- 折れ線グラフでトレンドをつかむ …… 移動平均で株価の動きを予測！
- 棒グラフと折れ線グラフの組み合わせ …… 規模の異なるデータの変化を可視化／比較！

折れ線グラフは小学 4 年生で学びますが、意外に奥が深いです。データ分析の基本なのでしっかりと身に付けておきましょう！　では、サンプルファイルの利用についての説明の後、本編に進みましょう。

サンプルファイルの利用について

本稿では、表計算ソフトを使って手を動かしながら学んでいきます。表計算ソフト Microsoft Excel 用の .xlsx ファイルをダウンロードできるようにしています。デスクトップ版の Excel が手元にない場合は、Microsoft アカウントがあれば使える無料の Microsoft 365 オンライン、もしくは Google アカウントがあれば使える無料の Google スプレッドシート（Google Sheets）をお使いください。Microsoft 365 オンラインの場合は、.xlsx ファイルを OneDrive にアップロードしてから開いてください。Google スプレッドシートの場合は、.xlsx ファイルを Google ドライブにアップロードしてから開いた上で［ファイル］メニューの［Google スプレッドシートとして保存］を実行してください（Google スプレッドシート独自の機能を使っている場合は、ファイルを共有して参照できるようにします。その場合は、該当する箇所で使い方を記します）。

折れ線グラフに潜む落とし穴 ～ 成績の変化を可視化、でもそれでいいの？

「この記事で学べること」で触れた最初のテーマからです。売り上げや成績、事故の件数など、時系列で並んださまざまなデータは折れ線グラフを使って可視化することにより、変化を捉えることができます。そこで、できるだけシンプルな例として、図 2 の成績データ（架空データ）を基にケーススタディーを行ってみましょう。折れ線グラフの作成なんて簡単と思われるかもしれませんが（確かに操作そのものは簡単ですが）、分析は簡単にいかないこともあります。折れ線グラフを作成し、グラフからどんなことが言えそうか考えてみましょう。サンプルファイルをこちらからダウンロードし、［成績］ワークシートを開いて取り組んでみてください。

	A	B	C	D	E	F	G
1	定期試験成績						
2							
3	教科	1学期中間	1学期期末	2学期中間	2学期期末	3学期期末	
4	国語	87	86	90	84	81	
5	数学	73	70	71	79	85	
6	理科	73	66	72	74	78	
7	社会	82	69	#N/A	68	70	
8	英語	90	83	86	83	87	
9							

図 2　定期試験のデータ

このデータを基に折れ線グラフを作成し、成績についての分析を行ってみよう。気付いたことをできるだけたくさん列挙してほしい。なお、社会の 2 学期中間試験は、その時間に体調を崩して試験を受けられなかったので、利用できる値がないことを表すエラー値 #N/A を入力してある。

折れ線グラフの作成に関しては、難しいところはありません。図 3 のようなグラフが作成できると思います。手順は、図 3 の後に箇条書きで示してありますが、動画でも詳しく解説しています。手順を一つ一つ追いかけたい方はぜひご視聴ください。

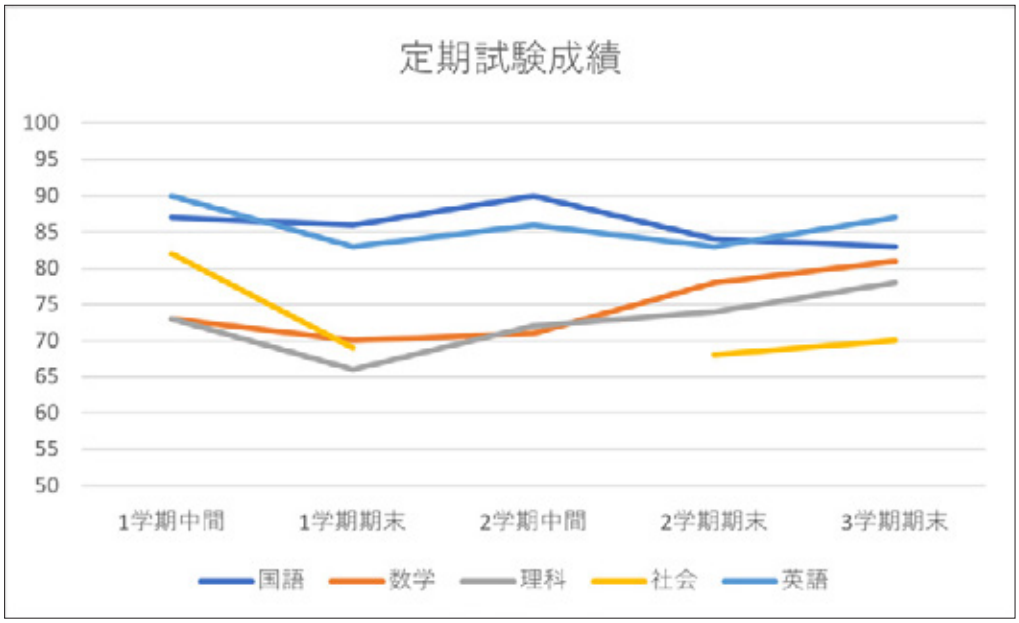


図 3　定期試験のデータを折れ線グラフにする

素直に折れ線グラフを作成した結果。ただし、縦（数値）軸の最小値と最大値を変更し、グラフのタイトルとしてセル **A1** の値が表示されるようにしてある。このグラフを基に分析を行ってみよう。社会の 2 学期中間試験のデータがないので線がつながらないが、今のところは線をつながなくてもよい（線をつなぐ方法については後述）。

グラフ作成の手順は以下の通りです。

- セル **A3** ～ **F8** のいずれかのセルをクリックする（アクティブセル領域がグラフデータの範囲となる）
- ［挿入］タブを開き、［折れ線／面グラフの挿入］－［折れ線］を選択する
  - ・Google スプレッドシートの場合は、メニューから［挿入］－［グラフ］を選択し、［グラフエディタ］の［グラフの種類］のリストから［折れ線グラフ］を選択する

最初の手順でグラフ化する範囲を指定していないことが気になる方もおられるかもしれません。Excel では、グラフ作成や並べ替え、集計などの際に特に範囲を指定しないと、アクティブセル領域が指定されたものと見なされます。アクティブセル領域とは、アクティブセルを含み、空白のセルで囲まれた範囲です。図 1 の例であれば、セル **A3** ～ **F8** を選択しなくても、その範囲のいずれかのセルをクリックしておけばいいというわけです。

続けて、縦（数値）軸の最小値と最大値を指定し、グラフのタイトルを変更します。

- 縦（数値）軸を右クリックし、［軸の書式設定］を選択する
- ［最小値］に **50** を入力し、［最大値］に **100** を入力する（縦（数値）軸の目盛りが変わる）
  - ・Google スプレッドシートの場合は［グラフエディタ］の［カスタマイズ］をクリックし、［縦軸］をクリックして［最小値］に **50** を入力し、［最大値］に **100** を入力する
- グラフ中の「グラフ タイトル」をクリックして選択する
- 「グラフ タイトル」をもう 1 回クリックして、「定期試験成績」という文字列を入力する
  - ・Google スプレッドシートの場合はグラフのタイトル部分をダブルクリックし、「定期試験成績」という文字列を入力する

グラフのタイトルは、上の手順のように直接入力しても構いませんが、デスクトップ版の Excel では、以下の操作でセルの内容をグラフのタイトルに表示できます。

- 「グラフ タイトル」をクリックして選択する
- 数式バーをクリックして入力できるようにする
- 「=」を入力する
- セル **A1** をクリックする

このように操作すれば、セル **A1** に入力されている「定期試験成績」という文字列がグラフのタイトルとして表示されるようになります。また、セル **A1** の内容を変更すれば、グラフのタイトルもそれに合わせて変更されます。

図 2 を見ると「定期試験成績」という文字列はセル **C1** ～ **D1** 辺りに表示されているように見えますが、データは **A1** に入力されています。これはセル **A1** ～ **F1** の範囲内で中央にそろえるように設定してあるからです。デスクトップ版の Excel の場合、選択範囲内で中央にそろえるには、セル **A1** ～ **F1** を選択し、［ホーム］タブにある［配置］グループの［配置の設定］ボタン（右下にある小さな矢印が表示されたボタン）をクリックします。［セルの書式設定］ダイアログの［配置］タブが表示されるので、［横位置］のリストから［選択範囲内で中央］を選択します。

なお、セル結合を行っても同様の表示にできますが、セル結合を行うと列の移動／挿入や並べ替えなどがうまくいかなくなることがあります。

さて、図 3 のグラフからどのようなことに気づいたでしょうか。

- 1 学期の期末試験では成績が下がっている → 中間試験の成績がよかったので、油断した？
- 2 学期から 3 学期にかけて成績が上がっている → 成績が下がったので奮起した？
- 理科系よりも文科系の方が好成績。ただし、社会は 1 学期中間を除いてあまりよくない → いわゆる暗記科目は苦手？
- 数学と理科は成績が上がってきている → 何かコツをつかんだのか？　奮起のたまものか？

ここまではまだまだウオーミングアップです。上に示した箇条書きの項目は高校生でも（というか小学生でも）容易に思いつくことと思います。しかし、この連載の第 1 回は「高校生に負けない」というううたい文句から始まりました。ここで、上の分析を全否定して、高校生をぎゃふんと言わせてみましょう。

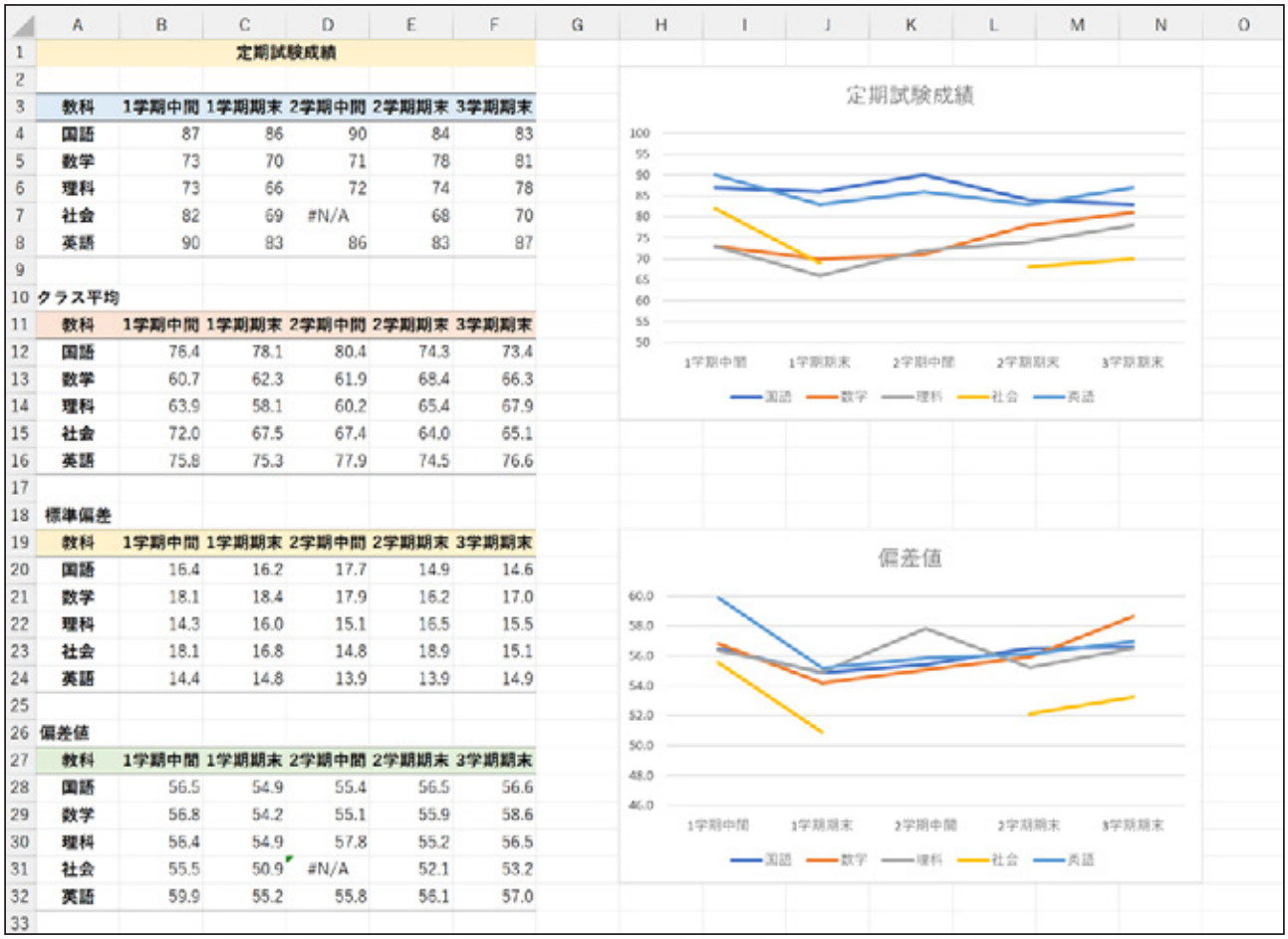
- 試験の難易度が教科によって異なるかもしれない → 理科系よりも文科系が好成績とは一概には言えない
- 同じ教科でも中間試験や期末試験など、期によっても難易度が異なるかもしれない → 成績が上がっているとか下がっているとは一概には言えない

例えば、1 学期には数学と理科の成績があまり良くなく、2 学期以降は成績が上がっているようです。しかし、1 学期の試験は難しい試験だったので、全体的に成績がよくなかったからかもしれません。そのため、2 学期からは易しめの問題になったのかもしれません。

では、どうすればより適切な分析ができるでしょうか。それには、この連載の第 6 回で解説した偏差値を使うといいでしょう。クラスあるいは学年の平均値と標準偏差を基に偏差値を求めれば、教科ごと、期ごとの比較ができます。やはり架空のデータですが、クラスの平均値と標準偏差を含む［偏差値グラフ］ワークシートを用意してあるので、参考にしてください。

素点や偏差値を使わずに、クラスや学年での順位を使って折れ線グラフを作るという手もあります。

偏差値の求め方はすでに第 6 回でやった通りで、グラフの作成方法は上で見たのと同じなので、結果だけ示しておきます（図 4）。この例では、社会を除き、全般的に文科系、理科系の差はなさそうです。



	A	B	C	D	E	F
1	定期試験成績					
2						
3	教科	1学期中間	1学期期末	2学期中間	2学期期末	3学期期末
4	国語	87	86	90	84	83
5	数学	73	70	71	78	81
6	理科	73	66	72	74	78
7	社会	82	69	#N/A	68	70
8	英語	90	83	86	83	87
9						
10	クラス平均					
11	教科	1学期中間	1学期期末	2学期中間	2学期期末	3学期期末
12	国語	76.4	78.1	80.4	74.3	73.4
13	数学	60.7	62.3	61.9	68.4	66.3
14	理科	63.9	58.1	60.2	65.4	67.9
15	社会	72.0	67.5	67.4	64.0	65.1
16	英語	75.8	75.3	77.9	74.5	76.6
17						
18	標準偏差					
19	教科	1学期中間	1学期期末	2学期中間	2学期期末	3学期期末
20	国語	16.4	16.2	17.7	14.9	14.6
21	数学	18.1	18.4	17.9	16.2	17.0
22	理科	14.3	16.0	15.1	16.5	15.5
23	社会	18.1	16.8	14.8	18.9	15.1
24	英語	14.4	14.8	13.9	13.9	14.9
25						
26	偏差値					
27	教科	1学期中間	1学期期末	2学期中間	2学期期末	3学期期末
28	国語	56.5	54.9	55.4	56.5	56.6
29	数学	56.8	54.2	55.1	55.9	58.6
30	理科	56.4	54.9	57.8	55.2	56.5
31	社会	55.5	50.9	#N/A	52.1	53.2
32	英語	59.9	55.2	55.8	56.1	57.0
33						

図 4　定期試験の偏差値を折れ線グラフにする

セル **B28** に「=(B4-B12)/B20*10+50」と入力し、セル **F32** までコピーすると全ての偏差値が求められる。グラフにすると、1 学期中間は英語をはじめとして全般的に好成績だが、1 学期の期末に下降。2 学期以降は徐々に回復。特に数学は素点でも偏差値でも上昇傾向。社会は 1 学期期末以降あまり良くないように思われる。しかし、社会以外は素点ほどには文科系、理科系の差はない。理科は素点では上昇傾向にあるが、偏差値を見ると浮き沈みがある。2 学期中間の好成績はたまたまヤマが当たったのか、興味のある単元が試験範囲だったのかもしれない。

さて、懸案事項となっていた、欠損値を線でつなぐ方法を紹介しておきましょう。手順を箇条書きで示しておきます。意外に知られていない操作です。結果の図は単に社会の線がつながるだけなので省略します。

- グラフを選択し、［グラフのデザイン］タブをクリックする
- ［データの選択］ボタンをクリックする
- ［データソースの選択］ダイアログが表示されるので、［非表示および空白のセル］ボタンをクリックする
- ［非表示および空白のセル］ダイアログが表示されるので、［空白セルの表示方法］の［データ要素を線で結ぶ］を選択し、［#N/A を空のセルとして表示］のチェックをオンにする
- ［OK］ボタンをクリックして［非表示および空白のセル］ダイアログを閉じる
- ［OK］ボタンをクリックして［データソースの選択］ダイアログを閉じる

Google スプレッドシートでは以下の操作になります。

- ［グラフエディタ］の［カスタマイズ］をクリックし、［グラフの種類］をクリックする
- ［null 値を表示］のチェックマークをオンにする

コラム　連勝の後は連敗が来る？ 〜 平均値への回帰

プロ野球などのスポーツで、推しのチームが何連勝もして喜んでいると、その後なぜか負けが込んでくることがあるように思われます。極端な場合、10 連勝の後に 10 連敗などということもありますね。実は、この現象は**平均値への回帰**と呼ばれます。たまたまいい成績を取ったとしても、長い目で見ると平均的な（実力を反映した）成績に落ち着くというわけです。

上の例では、1 学期中間でいい成績を取ったことも、1 学期期末で振るわなかったのも、たまたまかもしれません。「中間ではいい成績だったのに、期末はどうしたんだ。たるんでるんじゃないか」などと叱ったとしても、その後、成績が持ち直したのは叱った効果ではなく、単なる平均値への回帰かもしれません。にもかかわらず、親や指導者が「叱ると成績が上がる」と思い込んでしまうのは短絡的です。それでも成績が上がらなかったときに「成績が上がらなかったのは叱り方が足りないからだ」と、さらにエスカレートするのは、生徒にとっては不幸でしかありません。

世界の中で日本の地位はどう変化したか 〜 データは比較してこそ違いが見えてくる

ここまでは、分析の観点などを分かりやすく説明するために架空のデータを使ってきました。しかし、ケーススタディーとしてはリアリティーが足りないかもしれませんね。そこで、ここからは実際のデータを使って、可視化と分析に取り組むことにしましょう。試験の成績の例でも、**他者との比較によって本質が見えてくる**ことに気づかれたかと思います。そういった例を見ていきます。

皆さんは昨年（2022 年）の日本の 1 人当たり GDP（国内総生産）は G7 先進国首脳会議の中で第何位かご存じでしょうか。……と、急に聞かれても答えられる人はあまりいないと思います。IMF（国際通貨基金）の調査によると 3 万 3822 ドルで、実は第 7 位、つまり最下位です。ちなみに 1988 年（2 万 5575 ドル）〜 1996 年（3 万 9164 ドル）は第 1 位でした。

その地位の変化を可視化してみましょう。日本のデータだけを折れ線グラフにすると、上昇している／下降している／浮き沈みがあるといったことは分かります。しかし、他の国々と比較しないと、地位の変化は見えてきません。図 5 は G7 に加えて、同じアジアの国ということで、中国と台湾、韓国を含め、1990 年〜 2022 年までの 1 人当たり GDP を折れ線グラフにしたものです。

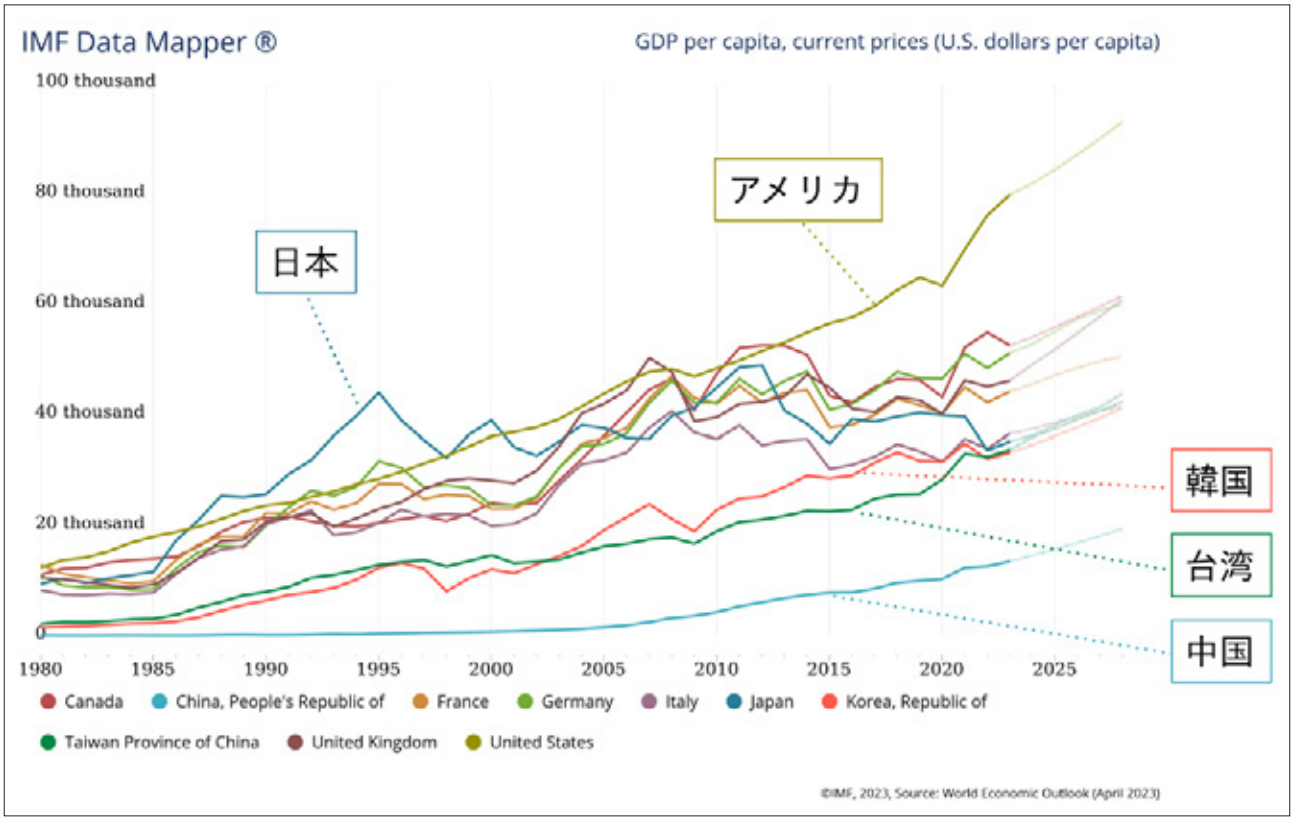


図 5　G7 ＋中国、台湾、韓国の 1 人当たり GDP の変化を折れ線グラフにする

このグラフは上で述べた IMF の Web サイトで作成したグラフに、幾つかの国名をラベルとして追加したもの。国や地域、年を選ぶだけで、ヒートマップや折れ線グラフが自動的に作成される。また、アニメーションも実行できる。2023 年以降の点線は予測値。Excel のファイルをダウンロードして自分でグラフを作成することもできるが、こちらの方が手軽で便利。

グラフの色分けが微妙でかなり込み入っているので分かりにくいかもしれませんが、傾向としては、アメリカが 1 人勝ちの様相を示しています。ヨーロッパは波があるものの、イタリアを除き、2010 年以降おおむね安定しています。残念ながら、日本は負け組のようです。中国、台湾、韓国もほぼ右肩上がりで、特に台湾と韓国は日本と同じ水準に達しています。中国は GDP 全体ではアメリカに次いで世界第 2 位なのですが、人口が多いので 1 人当たりの GDP はかなり低くなっています。また、貧富の差も大きいようです。ちなみに、日本は GDP 全体では世界第 3 位です。

日本が負け組に転落した理由としては、1900 年代末期のバブル崩壊後の需要の低下や生産年齢人口の減少、生産性の伸び悩み（内閣府、平成 27 年度年次経済財政報告）、2000 年代のアメリカでの IT バブル崩壊の影響を大きく受けたこと（内閣府、世界経済の潮流／世界経済白書）などが挙げられているようです。筆者は経済に関しては門外漢なので、これ以上の言及は避けますが、比較することによって、時系列による変化だけでなく、日本の地位や立場のようなものが見えてくることが分かります。

可処分所得は増えているのか ～ 折れ線グラフは「切り取り」にご注意

ここからは、折れ線グラフに関する注意点や便利な使い方などをオムニバス的に幾つか紹介します。まず、切り取りによる印象操作についてです。

図 6 をご覧ください。これは、内閣府の「家計可処分所得・家計貯蓄率四半期別速報（参考系列）」（Excel ファイル）を基に家計可処分所得のデータをグラフ化したものです。2021 年は新型コロナ禍の影響が拡大してきたせいでしょうか、可処分所得が減少していますが、2012 年から一貫して増加しています。なお、こちらの Excel ファイルに、グラフ化に使ったデータだけをまとめてあります。Google スプレッドシートの場合はこちらのサンプルファイルを開いて、メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください（縦軸の目盛りの設定を合わせてあるだけで、内容は同じです）。

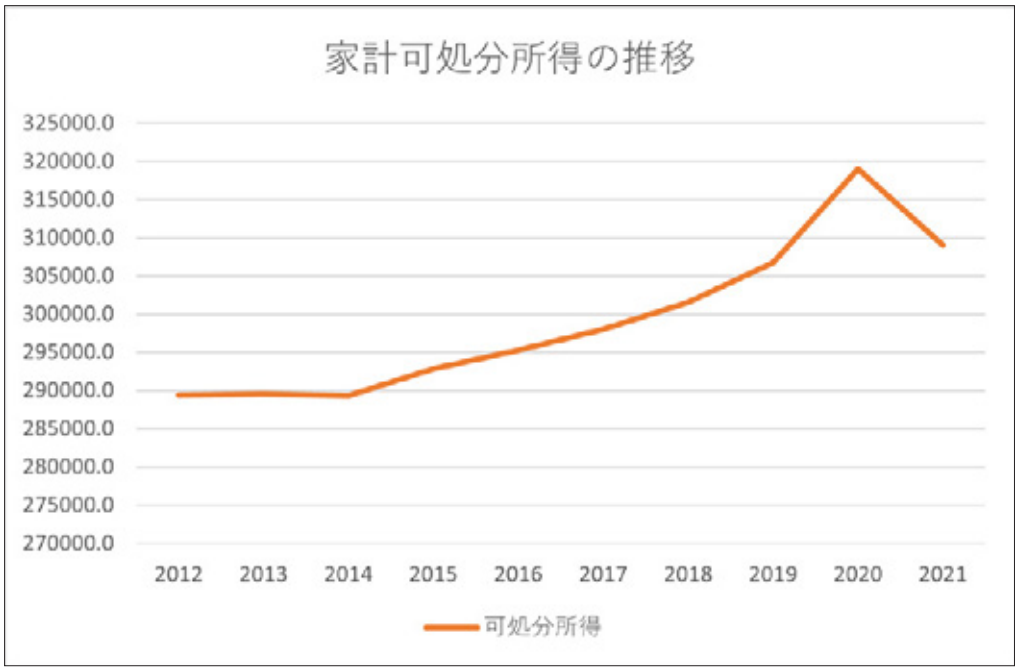


図 6　2012 年から 10 年間の家計可処分所得の変化を折れ線グラフにする（単位：10 億円）

2012 年から 2021 年にかけて、家計可処分所得（総額なので単位は 10 億円）はかなり増加しているように思われる。データには個人商店の値も含まれる。

個人的には、生活が楽になったとは実感できないのですが、データを見る限り、可処分所得は増加しているように思われます。しかし、これは**切り取りによる印象操作**です。実は、基となるデータは 1994 年から 2022 年までの値が記録されています。その時期も含めてグラフを作成すると、図 7 のようになります。

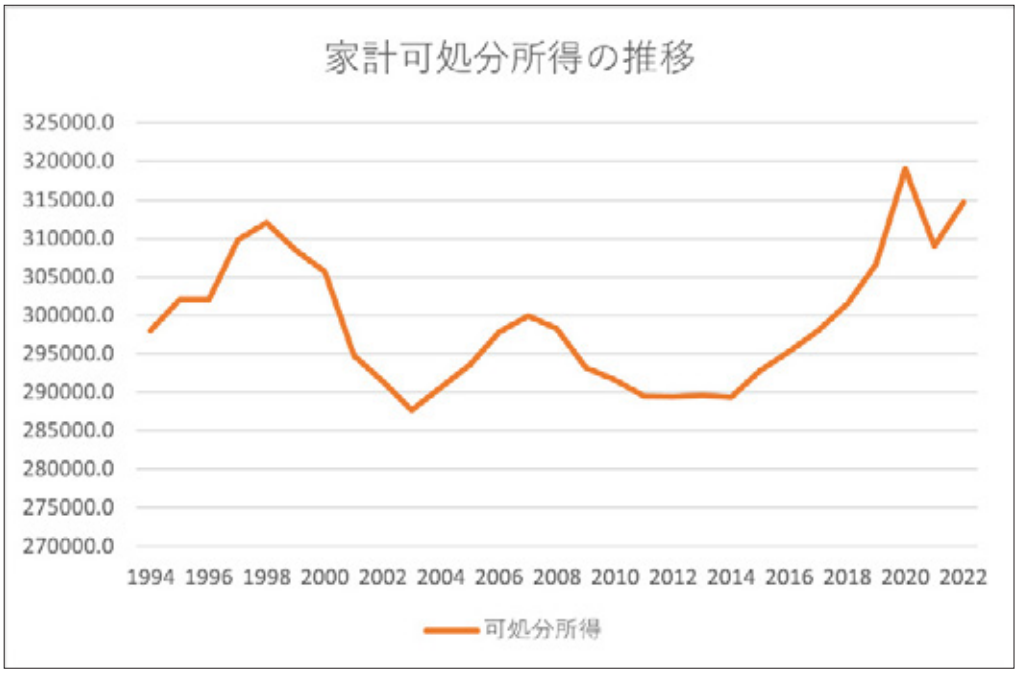


図 7　1994 年から 29 年間の家計可処分所得の変化を折れ線グラフにする（単位：10 億円）

1998 年にピークがあり、その後低下。2019 年にようやく 1998 年の水準に戻ったことが分かる。

図 6 は、家計可処分所得が増加していることを印象づけるため、一部のデータを切り取ってグラフ化したものになっていたということです。もちろん、もっと長い目で見れば、家計可処分所得は増加の傾向にあるのかも知れませんが、自分の主張に都合のいい部分だけを切り取るのはよくありませんね。

なお、上のグラフは名目値をプロットしたもので、実質の家計可処分所得は 1994 年以降ほぼ一貫して増加しています。ただし、実質値は、名目家計可処分所得を最終消費支出の名目値／実質値で割って求められた参考値なので、実質最終消費支出が大きくなると、実質家計可処分所得の値は大きくなってしまいます。また、上記のデータには貯蓄額などの値も含まれており、何かと興味深いのですが（貯蓄額はほぼ一貫して減っています）、ここは切り取りによって印象が変わるというお話なので、これぐらいにとどめておきます。

コラム　直感だけに頼らず、データに頼ろう

個人的なお話になってしまうのですが、筆者は最近のバイクブームに触発され、40 年ぶりにバイクにまたがるようになりました。そういったリターンライダーが増加したせいか、毎日のようにバイクによる死亡事故がニュースになっています。筆者の感覚としては、バイクの事故が増えているような気がしています。

実際のところ、バイクの死亡事故は増えているのでしょうか。ちょっと調べてみました。警視庁による「30日以内交通事故死者の状況について」の PDF ファイルを基に、二輪車に乗車していた人の死亡者数を折れ線グラフにしてみると、図 8 のようになります。こちらの Excel ファイルに、グラフ化に使ったデータだけをまとめてあります。

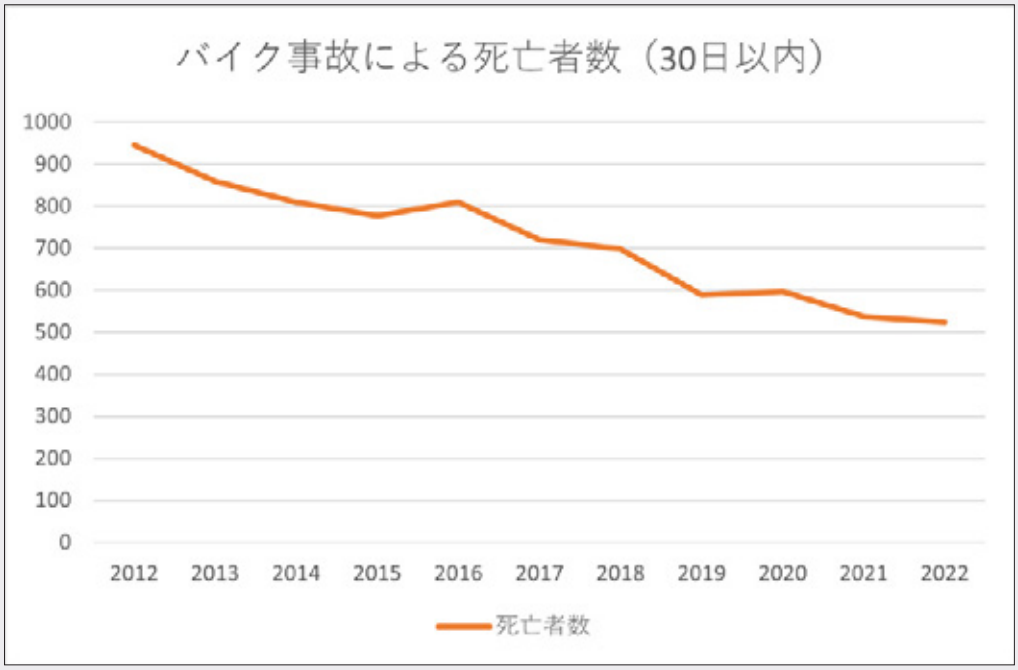


図 8　バイク事故による死亡者数の推移

直感とは違って、バイク事故による死亡者数は減少の傾向にある。直感も重要だが、データによる理解も重要。かといって、油断は大敵。安全運転を心がけよう（ここには四輪車のデータは掲載していないが、四輪車とバイクの普及台数を考え合わせれば、当然のことながらバイク事故による死亡者の割合は大きい）。

データによると、バイクの死亡事故は増えているわけではなく、むしろ減っているようです。しかし、死亡事故が増えているように思うのはなぜでしょうか。理由の 1 つは**認知バイアス**です。バイクに再び乗るようになる前も同じようにニュースが報じられていたのかもしれませんが、その時には興味がなかったので、全く気にとめなかったのだと思われます。しかし、バイクに興味を持ち出すとバイク関連のニュースが強く印象に残るというわけです。理由のもう 1 つは、Web サイトや SNS のターゲティングにより、興味のあるニュースや情報がよく表示されるようになったからだと思われます。事故に関するニュースは筆者自身には警鐘を鳴らすものとして有益ですが、陰謀論や極端な考え方にハマってしまうのもこのためだと考えられています。**直感も重要ですが、データをきちんと見ることも重要**ですね。

ただし、警視庁のページによると、東京都ではバイクの事故は増えているようです。また、通勤時、50歳代、単独事故と右折時、頭部と胸部の損傷による死者が多いようです。頭部はヘルメットが外れた場合の死者が多く、胸部プロテクターを装着している人はわずかなようです（筆者はもちろん装着しています）。全体的に事故が減っているとはいえ、油断大敵です。バイクに乗る乗らないに関わらず、皆さんも交通にはどうぞお気を付けください。

株価の動きを予測する ～ トレンドを見るには移動平均が便利

移動平均とは、何日分かの平均値を 1 日ずつずらしながら順に求めたものです。短期移動平均と長期移動平均を求めて折れ線グラフにすると、株価などの上昇／下落のトレンド（傾向）の変化が分かります。

図 9 は、Apple の株価（終値）の 5 日移動平均と 15 日移動平均を求め、折れ線グラフにしたものです。データは Stooq という Web サイトで［Download data in csv file...］をクリックして、2023 年 6 月 1 日～ 8 月 31 日までの 3 カ月間のデータを取得したものです。それを見やすくしたもの（Excel ファイル）をこちらに置いておきました。グラフの作り方は図 9 の後に箇条書きで示してあります。動画でも解説しているので、手順を一つ一つ追いかけたい方はぜひご視聴ください。

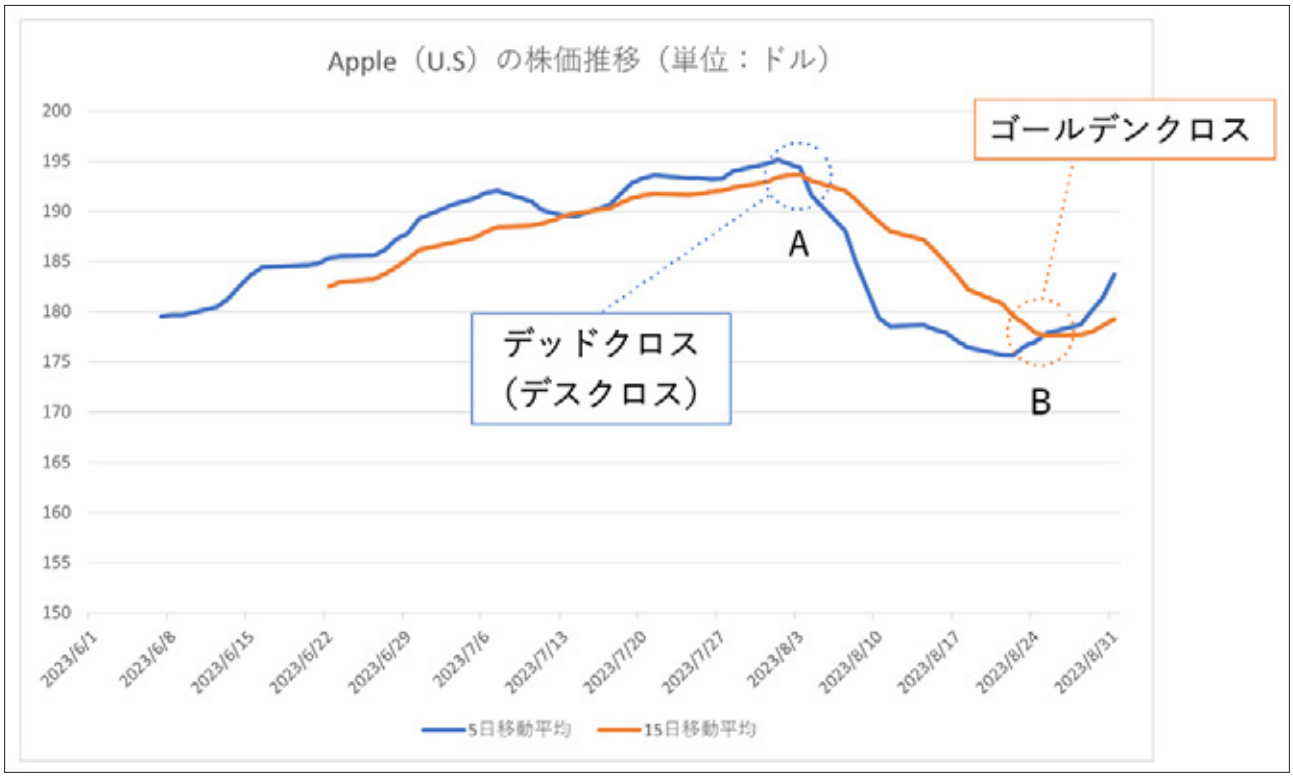


図 9　短期移動平均と長期移動平均（2023 年 6 月 1 日～ 8 月 31 日の Apple の株価）

短期（5 日）移動平均と長期（15 日）移動平均をプロットしたグラフ。短期移動平均が長期移動平均を上から下にクロスする箇所（A の部分）を「デッドクロス」または「デスクロス」と呼び、株価が下がるシグナルとなる。逆に、短期移動平均が長期移動平均を下から上にクロスする箇所（B の部分）を「ゴールデンクロス」と呼び、株価が上がるシグナルとなる。

図 9 のグラフの作り方は以下の通りです。

- 移動平均を求める手順：
  - ・セル **G8** に「=AVERAGE(E4:E8)」と入力する
  - ・セル **G8** をセル **G67** までコピーする
  - ・セル **H18** に「=AVERAGE(E4:E18)」と入力する
  - ・セル **H18** をセル **H67** までコピーする
- グラフを作成する手順：
  - ・セル **A3** ～ **A67** をドラッグして選択する
  - ・セル **G3** ～ **H67** を［Ctrl］キーを押しながらドラッグして選択する
  - ・［挿入］タブを開き、［折れ線／面グラフの挿入］－［折れ線］を選択する

Google スプレッドシートの場合、移動平均の求め方は同じですが、グラフの作成時に縦軸の最小値と最大値を指定する必要があります。

- グラフを作成する手順：
  - ・セル **A3** ～ **A67** をドラッグして選択する
  - ・セル **G3** ～ **H67** を［Ctrl］キーを押しながらドラッグして選択する
  - ・メニューから［挿入］－［グラフ］を選択し、［グラフエディタ］の［グラフの種類］のリストから［折れ線グラフ］を選択する。
  - ・［グラフエディタ］の［カスタマイズ］をクリックし、［縦軸］をクリックして［最小値］に **150** を入力し、［最大値］に **200** を入力する

図 9 を見ると、8 月 3 日頃に短期移動平均が長期移動平均を上から下にクロスしています。長期移動平均のなだらかな変化に対して、短期移動平均が急に下がってきたということなので、その後株価が下落することが予測されます（この部分をデッドクロスまたはデスクロスと呼びます）。その後、8 月 24 日頃には、短期移動平均が長期移動平均を下から上にクロスしています。こちらは、短期移動平均が急に上がってきたということなので、株価が上昇することが予測されます（この部分をゴールデンクロスと呼びます）。このように、**短期移動平均と長期移動平均はトレンド（傾向）の変化を簡単に見るのに便利**です。もちろん、デッドクロスやゴールデンクロスの兆候が現れたからといって、必ずしもその後株価がそれぞれ下落したり上昇したりするとは限りません（7 月 13 日頃にデッドクロスの兆候がありますが、すぐに持ち直しています）。

規模の異なるデータの変化を可視化／比較する ～ 棒グラフと折れ線グラフの複合グラフを作る

前回、ちょっと先取りした話として時系列データのタイムラグについて説明しました。前回は、利用するグラフを棒グラフのみとし、縦（数値軸）の目盛りを 1 つだけに限定していたので、図 10 に示したサイト A の売り上げ（E 列）とサイト B の 1 日後の売り上げ（G 列）を棒グラフで簡易的に比較しました。

実のところ、サイト B のクリック数（D 列）とサイト B の 1 日後の売り上げ（G 列）をグラフにすれば、Web サイトにアクセスしてから購買行動に移るタイムラグが直接的に可視化できます。しかし、クリック数と売り上げはデータの規模（値の範囲）が異なるので、棒グラフだけではパターンが比較しづらくなります。そのような場合、**第 2 軸を指定し、異なる目盛りを指定すれば、異なる規模のデータでも、パターンの変化が比較できる**ようになります。また、グラフを見やすくするためには、両方の系列を棒グラフにするのではなく、一方の系列を折れ線グラフにした「複合グラフ」の方が、より比較がしやすくなります。というわけで、サイト B のクリック数を折れ線グラフ（第 2 軸）として、1 日後の売り上げを棒グラフとして表した複合グラフを作成してみたいと思います。

	A	B	C	D	E	F	G	H
1	媒体別広告売上一覧							
2								
3			クリック数（回）		売上（千円）			
4	日付	曜日	サイトA	サイトB	サイトA	サイトB	サイトB(1日後)	
5	7月24日	月	94	50	94	284	84.0	
6	7月25日	火	130	87	180	84	107.0	
7	7月26日	水	137	135	181	107	225.0	
8	7月27日	木	87	60	170	225	77.0	
9	7月28日	金	246	171	360	77	281.0	
10	7月29日	土	175	132	346	281	378.0	
11	7月30日	日	217	179	360	378	284.0	
12								

図 10　Web サイトの広告クリック数と売り上げ

サイト B の広告クリックが 1 日後の購買活動に結びついていることを可視化するには、サイト B のクリック数（D 列）と、サイト B の 1 日後の売り上げ（G 列）の変化のパターンを見るとよい。ただし、値の範囲が異なるので、集合縦棒グラフだけではパターンが比較しづらい。

図 11 が複合グラフの作成例です。操作の手順は図 11 の後に箇条書きで示してあります。こちらから Excel のブックをダウンロードし、［売上一覧］ワークシートを開いて操作してみてください。これについても、動画で操作の手順を解説しているので、一つ一つ追いかけたい方はぜひご視聴ください。なお、Google スプレッドシートの場合はこちらのサンプルファイルを開いて、メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください（曜日の表示形式を Google スプレッドシートに合わせた形式にしてあるだけで、内容は同じです）。

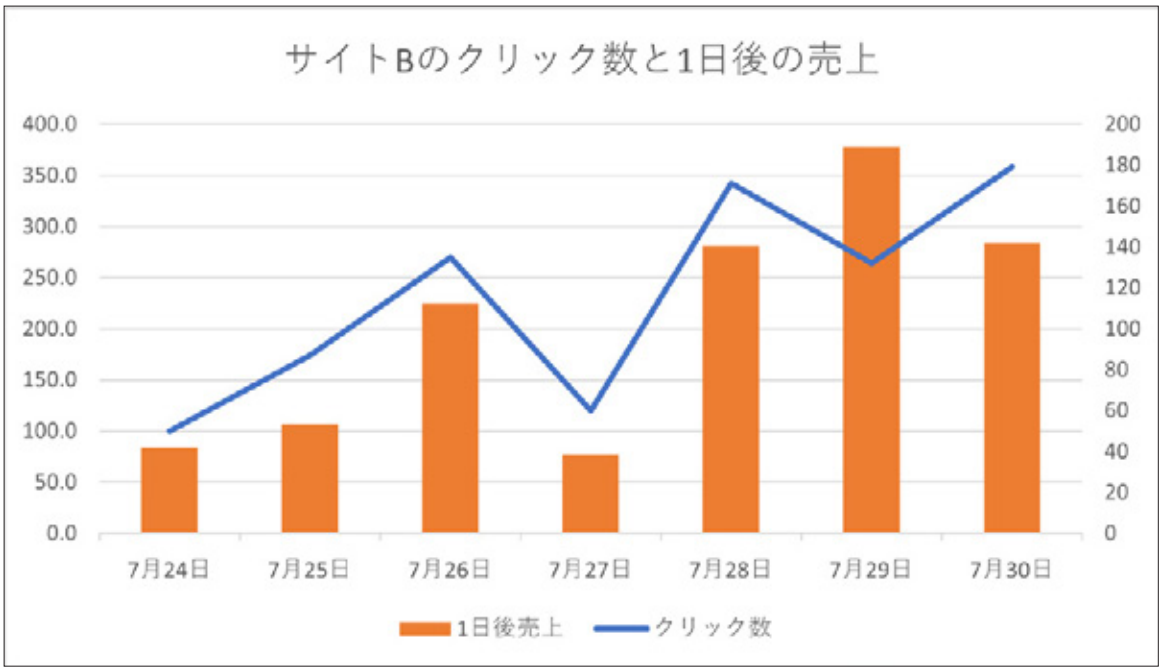


図 11　広告クリック数と 1 日後の売り上げを複合グラフにする

複合グラフにすると、広告のクリック数（折れ線グラフ）と 1 日後の売り上げ（棒グラフ）が同じようなパターンになっていることがより明確に分かる。クリック数の目盛りは右側の第 2 軸となっていることに注目。第 2 軸を指定すれば、規模の異なるデータでもパターンを比較することができる。作成例は上でダウンロードしたファイルの［売上一覧（日付をずらしたグラフ）］ワークシートに含まれている。

グラフ作成の手順は以下の通りです。なお、この記事の執筆時点では、オンライン版の Excel の場合、複合グラフや第 2 軸を設定したブックをアップロードして表示／編集することはできますが、作成することはできないようです。

- セル **A4** ～ **A11** をドラッグして選択する
- セル **D4** ～ **D11** を［Ctrl］キーを押しながらドラッグして選択する
- セル **G4** ～ **G11** を［Ctrl］キーを押しながらドラッグして選択する
- ［挿入］タブを開き、［複合グラフ］－［集合縦棒 － 第 2 軸の折れ線］を選択する
    - ・Google スプレッドシートでは、メニューから［挿入］－［グラフ］を選択し、［グラフエディタ］の［グラフの種類］のリストから［折れ線］の一覧にある［複合グラフ］を選択する

ここまでの操作で、クリック数が棒グラフ、1 日後の売り上げが折れ線グラフの複合グラフが作られます。しかし、ここでは、逆にクリック数を折れ線グラフに、1 日後の売り上げを棒グラフにしたいので、以下の操作でグラフの種類を変更する必要があります。

- 作成されたグラフをクリックして選択する
- ［グラフのデザイン］タブをクリックする
- ［グラフの種類の変更］ボタンをクリックする
- ［サイト B］の［グラフの種類］から［折れ線］を選択する
- ［サイト B］の［第 2 軸］のチェックボックスをクリックしてオンにする
- ［サイト B（1 日後）］の［グラフの種類］から［集合縦棒］を選択する

Google スプレッドシートでは、以下のような操作になります。

- ［グラフエディタ］の［系列］をクリックし、［すべての系列に適用］というリストから［サイト B］を選択する
- ［形式］の下の［種類］リストから［折れ線］を選択する
- ［軸］リストから［右軸］を選択する
- ［系列］の［サイト B］と表示されているリストをクリックし［サイト B(1 日後)］を選択する
- ［形式］の下の［種類］リストから［縦棒］を選択する

凡例には、クリック数が［サイト B］、1 日後の売り上げが［サイト B（1 日後）］と表示されているので、ちょっと分かりにくいですね。そこで系列名も変えておきましょう。

- ［グラフのデザイン］タブをクリックする
- ［データの選択］ボタンをクリックする
- ［凡例項目（系列）］リストの［サイト B］をクリックし、［編集］ボタンをクリックする
- ［系列名］に「クリック数」と入力し、［OK］ボタンをクリックする
- ［凡例項目（系列）］リストの［サイト B（1 日後）］をクリックし、［編集］ボタンをクリックする
- ［系列名］に「1 日後売上」と入力し、［OK］ボタンをクリックする
- ［OK］ボタンをクリックして［データの選択］ダイアログを閉じる

Google スプレッドシートでは、以下のような操作になります。

- グラフの上に表示されている凡例をクリックして選択する
- ［サイト B（1 日後）］という凡例をダブルクリックして「1 日後売上」と入力する
- ［サイト B］という凡例をダブルクリックして「クリック数」と入力する

作成されたグラフから、サイト B のクリック数と 1 日後の売り上げが同じパターンで増減していることが明確に読み取れますね。上でも述べましたが、複合グラフを利用して、**左右にそれぞれの軸を設ければ、異なる規模のデータでもパターンの変化が比較できる**ようになります。

今回は、時系列での変化を見るために折れ線グラフを作成しました。その中で、自分の位置を確かめるためには比較することが重要であることも見ました。また、切り取りによる印象操作という落とし穴や、移動平均によりトレンドの変化をつかむ方法、複合グラフの第 2 軸を利用して規模の異なるデータの変化を比較する方法についても紹介しました。

次回は、重要度の可視化をテーマとしたケーススタディーを通して、円グラフやパレート図の作成や利用、読み解き方の留意点などを見ていきます。次回も、落とし穴や意外に知られていない機能などを紹介します。どうぞお楽しみに！

［データ分析］円グラフやパレート図で「重要度」を可視化 ～ どの割合が本当に多いのか

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 9 回。グラフを利用して「重要度」を可視化する方法と、それに関連するさまざまな考え方を追いかけます。具体的には円グラフやパレート図、積み上げ棒グラフなどを使いますが、データの取り扱い、結果の見方などに関して、考慮すべき点や見落としがちな点について、ケーススタディーを通して見ていきます。

羽山博（2023 年 10 月 12 日）

読者の皆さんは、割合（比率）を可視化するのに円グラフを使うことは百も承知だと思います。しかし、割合が何らかの目的に対する「重要度」を反映した値であることについては、あまり意識されないことも多いようです。今回は、その「重要度」に焦点を当て、可視化による比較の方法や分析の例、落とし穴などについて見ていきます。

「比率」と「割合」は、同じような意味ですが、「比率」は項目同士の値の比較という意味合いの言葉です。例えば、A と B の比率は **1:1.5** といった感じの使い方です。一方の「割合」は、全体に対してその項目の占める大きさといった意味合いです。A の割合は全体の **30%**、B の割合は全体の **45%**といった感じです。

図 1 の上側は、2020 年～ 2022 年の電動キックボードの事故件数を相手別にグラフ化したものです。このグラフをパッと見て、どのような印象を受けるでしょうか。また、下側の図は、2022 年の不正アクセスによる被害の認知件数をパレート図と呼ばれるグラフにしたものです。被害を減らすためにはどこから手を付けていけばいいでしょうか。

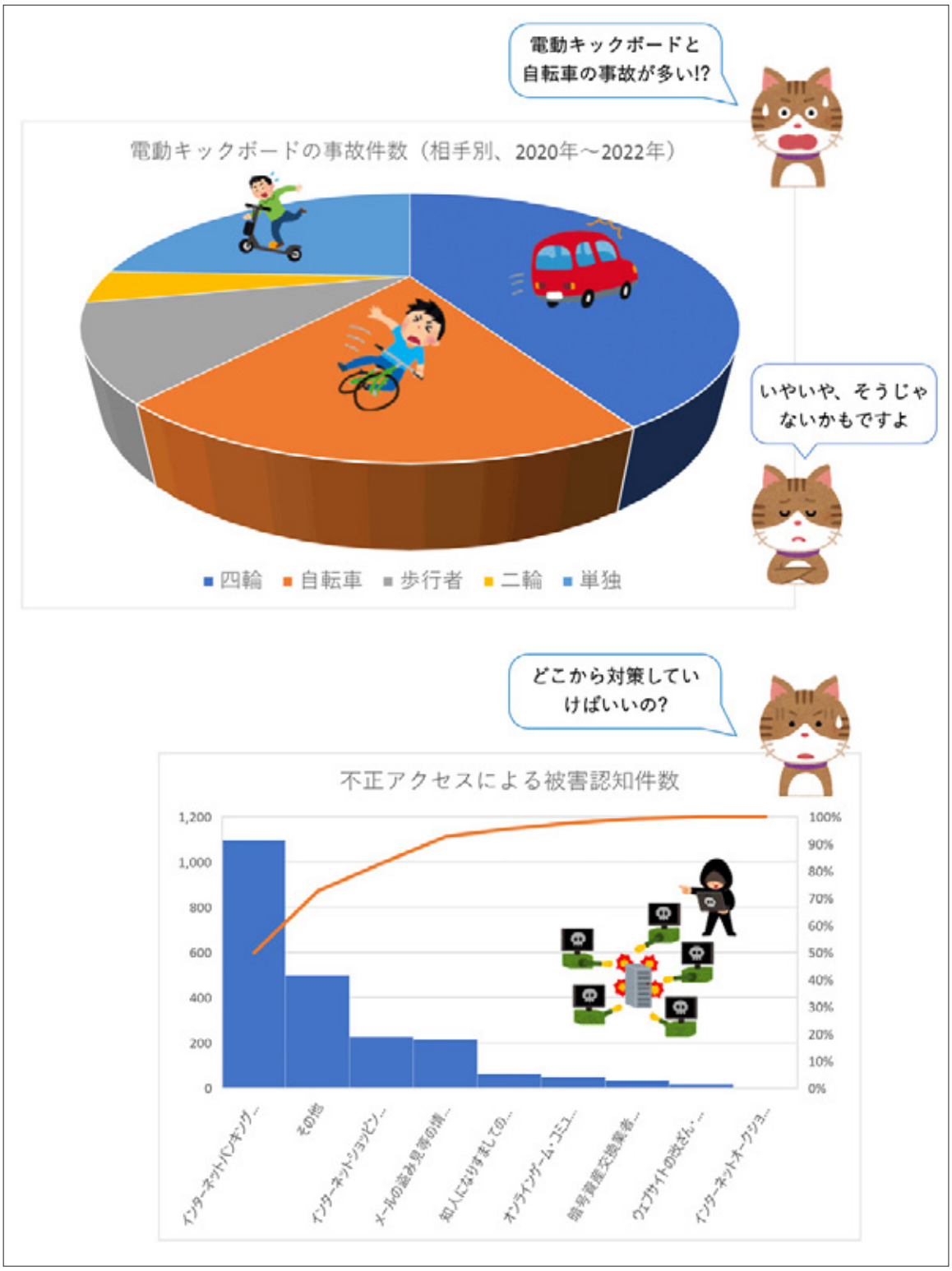


図 1　重要度を可視化する（落とし穴あり!）

上側の図は、一見すると、対自転車の事故件数が多く、四輪との事故件数がそれに続くように思われる。また、単独事故はそれほど多くないように見える。はたして、電動キックボードと自転車の事故は本当に多いのだろうか。また、下側の図を基に、不正アクセスの被害を防ぐための対策に優先順位を付けるにはどうすればいいだろうか。

→目次に戻る

出典となるデータは、上側が警察庁の交通事故分析資料から閲覧できる「令和 4 年における交通事故の発生状況について」の PDF ファイルです。下側は、総務省のページに掲載されている［別紙］の PDF ファイルです。

今回は重要度を可視化するというテーマで、幾つかの例を見ていきます。図 1 に関する問いの答えを探りながら、事故や故障、不良品などを激減させたり、売り上げを伸ばしたりするためには何に注目すればいいかを見てきます。また、重要度が社会情勢や個人の嗜好（しこう）によってどのように変化しているのかを考えます。ぜひ無料会員登録して全文をお読みください。

この記事は、データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 9 回です。第 7 回の棒グラフから、前回の折れ線グラフ、今回の円グラフ／パレート図、ヒストグラム、ヒートマップ、第 12 回の散布図まで、1 つずつ可視化の基礎を学んでいきます。これらグラフの目的と効用などについて、特別予告編で簡単に整理していますので、事前に確認しておくとより理解が深まるでしょう。

この記事で学べること

今回は以下のようなポイントについて、分析の方法や落とし穴を見ていきます。

- 円グラフに潜む落とし穴 …… 3D グラフにすると割合が違って見える？
- 重要な項目を見つけるには …… トラブルの原因は 9 割が ××× ？
- 規模と割合の変化を可視化する …… 好まれるスポーツの変遷を見てみよう！

まずは、小学生の頃から（幼稚園や保育園の頃から？）使い慣れている円グラフからスタートします。棒グラフや折れ線グラフと同様、円グラフもよく使われるにもかかわらず、グラフを作っただけで安心してしまい、どう読み解き、どう活用するかを考えられることが少ないように思われます。戦略や方針を立てる上でとても役に立つツールなので、いま一度その扱い方をしっかりと身に付けておきましょう！　では、サンプルファイルの利用についての説明の後、本編に進みましょう。

サンプルファイルの利用について

本稿では、表計算ソフトを使って手を動かしながら学んでいきます。表計算ソフト Microsoft Excel 用の .xlsx ファイルをダウンロードできるようにしています。デスクトップ版の Excel が手元にない場合は、Microsoft アカウントがあれば使える無料の Microsoft 365 オンライン、もしくは Google アカウントがあれば使える無料の Google スプレッドシート（Google Sheets）をお使いください。Microsoft 365 オンラインの場合は、.xlsx ファイルを OneDrive にアップロードしてから開いてください。Google スプレッドシートの場合は、.xlsx ファイルを Google ドライブにアップロードしてから開いた上で［ファイル］メニューの［Google スプレッドシートとして保存］を実行してください（Google スプレッドシート独自の機能を使っている場合は、ファイルを共有して参照できるようにします。その場合は、該当する箇所で使い方を記します）。

3D グラフに潜む落とし穴 ～ 円グラフで割合を可視化してみると？

答えから先に言うと、図 1 の上側に示した 3D 円グラフは、割合を見誤ってしまう危険性のある不適切なグラフです。元データは以下の通りです（図 2）。このデータを基に、ウオーミングアップがてら、まずは 2D の円グラフを作成してみましょう。後で 2D グラフと 3D グラフを見比べてみることにします。

	A	B	C	D	E
1	電動キックボードの事故件数（相手別、2020年～2022年）				
2					
3	相手	件数			
4	四輪	31			
5	自転車	14			
6	歩行者	8			
7	二輪	3			
8	単独	18			
9	合計	74			
10					

図 2　電動キックボードの事故件数（相手別）のデータ

このデータを基に 2D 円グラフを作成してみよう。円グラフでは割合が可視化されるが、利用するデータは件数だけでよい（割合をあらかじめ求めておく必要はない）。

サンプルファイルをこちらからダウンロードし、［相手別事故件数］ワークシートを開いて取り組んでみてください。Google スプレッドシートの場合はこちらのサンプルファイルを開いて、メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。

作成されたグラフは以下の通りです。操作の手順については、図 3 の後に箇条書きで示しておきます。なお、動画の解説も用意してあるので、操作を一つ一つ追いかけたい方はぜひご視聴ください。

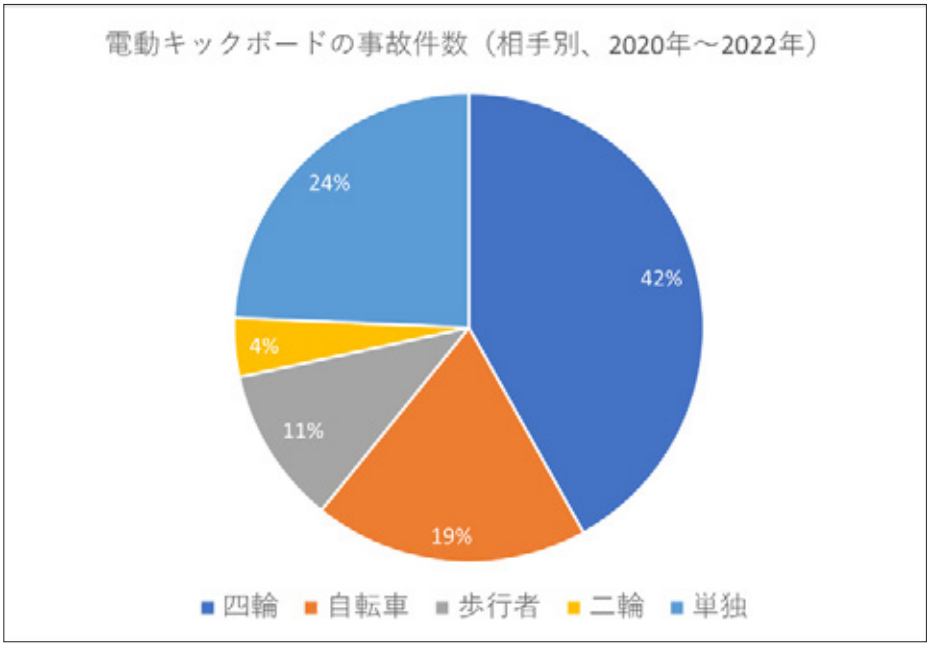


図 3　電動キックボードの事故件数を円グラフにして割合を見る

実際の割合も分かるように、グラフ内にデータラベルを表示した。当然のことながら、円の面積がそれぞれの割合を反映している。一目で、最も多い事故の相手が分かる。

手順は以下の通りです。

- セル **A3** ～セル **B8** をドラッグして選択する
- ［挿入］タブを開き、［円またはドーナツグラフの挿入］－［円］を選択する
  - ・Google スプレッドシートの場合はメニューバーから［挿入］－［グラフ］を選択し、［グラフの種類］のリストから［円グラフ］を選択する

これだけで円グラフが作成されます。Google スプレッドシートでは「四輪」や「自転車」などの分類名と割合が自動的に表示されますが、Excel の場合は単に円と凡例が表示されるだけです。そこで、割合を円グラフの中に表示するようにしましょう。

- データ系列を右クリックし、［データラベルの追加］を選択する
- 表示されたデータラベルを右クリックし、［データラベルの書式設定］を選択する
- ［データラベルの書式設定］ウィンドウで［パーセンテージ］のチェックマークをオンにし、［値］のチェックマークをオフにする
- ［ラベルの位置］の［内部外側］ボタンをクリックしてオンにしておく

［値］のチェックマークを先にオフにしてしまうと、何も選択されていない状態になるので、せっかく表示されたデータラベルが消えてしまいます。先に［パーセンテージ］のチェックマークをオンにしましょう。また、［ラベルの位置］は、標準では［自動調整］になっていますが、この例では、円グラフの表示が少し小さくなります。［内部外側］にした方が円グラフが大きく表示され、見やすくなります。タイトルを変更したり、データラベルのフォントサイズやフォントの色を変更したりして表示を整えれば完成です。

では、2D 円グラフと 3D 円グラフを見比べてみましょう（図 4）。3D 円グラフだと、割合が大きく違って見えますね。

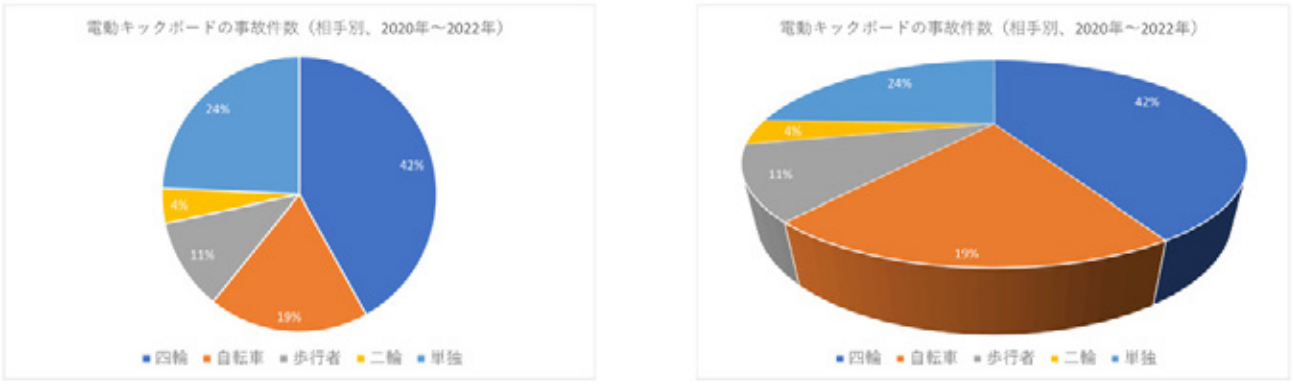


図 4　2D 円グラフと 3D 円グラフを見比べてみる

対四輪は **42%**、対自転車は **19%**となっており、対自転車の事故が対四輪の半分以下であるにもかかわらず、3D 円グラフにすると、手前にある対自転車がかなり大きな割合になっているように見える。また、単独事故は **24%**と、対自転車よりも多いのに、3D 円グラフではかなり小さく見える。なお、違いが顕著に分かるようにするため、右側の 3D 円グラフでは奥行きを深くして遠近感を強調してある（もちろん、標準の設定でも手前の項目が大きく見えてしまう）。

「この記事で学べること」で最初に触れたように、3D グラフにすると割合が違って見えてきます。シンプルな円グラフだとパッとしないので、見栄えをよくしたいというのは人情ですが、3D グラフにすると、手前にある項目が大きく見えてしまうのです。これは、円グラフに限らず、3D グラフ全般に見られる罠（わな）なので注意が必要です。そもそも、必然性のない 3D グラフの利用は避けた方がいいでしょう。

もっとも、件数そのものが少ない（全体で **74 件**）ので、図 3 の割合が電動キックボードの事故に関する特質を表していると断言することはできません。なお、2023 年 7 月に道路交通法が改正され、16 歳以上であれば免許なしで電動キックボードを公道で運転できるようになり、時速 6km/h 以下であれば歩道の走行も可能になりました。自転車や歩行者を巻き込んだ事故の増加が懸念されています（最近、歩道上での事故も大きく報道されていました）。なお、図 2 のデータは警察庁に報告された件数なので、対四輪の事故以外については報告そのものが上がっていない可能性も大きいと思われます。

コラム　グラフデータの範囲を間違って変なグラフができてしまったら

円グラフに限らず、どの種類グラフでも、グラフデータの範囲を間違えると、明らかにおかしいグラフ（図5）ができてしまうことがあります。

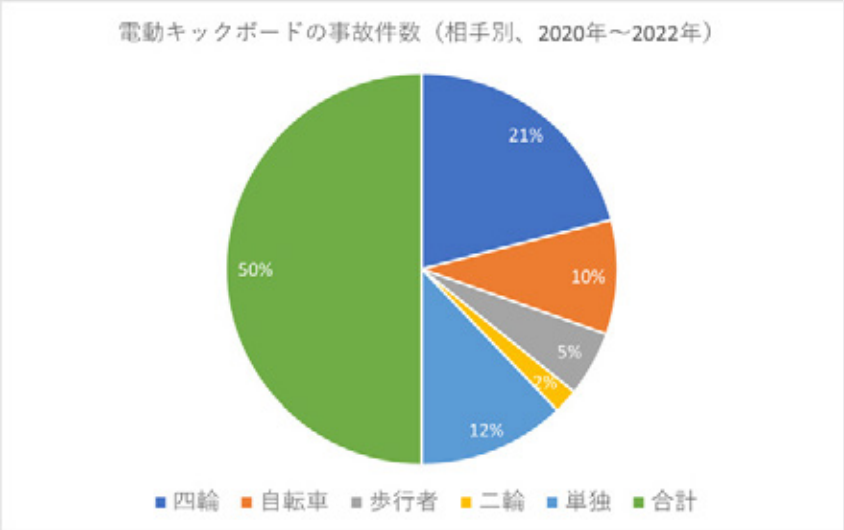


図5　グラフデータの範囲を間違ったグラフ

合計がグラフ化するための系列に含まれてしまっている。何も考えずに円グラフを作成するとこうなってしまう。思ったようなグラフにならなかった場合に、適切に修正する方法も身に付けておこう。

間違ったグラフができる原因の多くは、グラフデータの範囲を指定していないことがほとんどです。Excelでは、特に範囲を指定しないとアクティブセル領域（アクティブセルを含み、空白のセルで囲まれた範囲）がグラフ化や並べ替えなどの範囲と見なされます。そのため、本来はグラフデータに含めるべきでなかった合計の値などがグラフに含まれてしまうことがあるというわけです。

たいていの人は、グラフが表示された時点で間違いに気づくのですが、重要なのはそこからのリカバリーです。そこで、わざと間違ったグラフを作成し、それを正しいグラフにするための手順を見ていくことにしましょう。以下に箇条書きで操作を示します。

- セル **A3** ～セル **B8** のいずれかのセルをクリックする（範囲を選択しない）
- ［挿入］タブを開き、［円またはドーナツグラフの挿入］－［円］を選択する
  - Google スプレッドシートの場合は、メニューバーから［挿入］－［グラフ］を選択し、［グラフの種類］のリストから［円グラフ］を選択する

セル **A3** ～ **B8** のいずれかのセルをクリックすると、セル **A3** ～ **B9** がアクティブセル領域になり、グラフデータに合計が含まれてしまいます。図5のようなグラフになったのはそのためです。

グラフデータの範囲は以下の操作で修正できます。

- グラフをクリックして選択する
- セル **B9** の右下に表示されているハンドル（小さな■）をセル **B8** までドラッグする

Google スプレッドシートの場合は以下のように操作します。

- グラフを右クリックして［データ範囲］を選択する
- ［グラフエディタ］の画面で［データ範囲］を「A3:B8」に修正する
  - 右端の［データ範囲を選択］ボタン（田の形のボタン）をクリックして、セル **A3** ～ **B8** をドラッグしてもよい

逆に、アクティブセル領域がどの範囲であるかを理解していると、（元のデータに合計行が含まれていない場合など）いちいち範囲指定をしなくても適切なグラフを作成したり、並べ替えを行ったりすることができます。特に、対象となる範囲が大きい場合には効率のよい操作ができます。「アクティブセル領域」は、Excel を使いこなすのに必要不可欠なキーワードです。

コラム　グラフを構成する要素

グラフにはさまざまな要素が含まれているので、書式などの設定項目が多岐にわたります。何回かの試行錯誤の後、ようやく目的の設定項目にたどりついたという経験をお持ちの方も多いと思います。確実に設定項目にたどりつくには、グラフの各要素の名前を知っておくのが近道です。図 6 でそれぞれの要素と名前を確認しておきましょう。

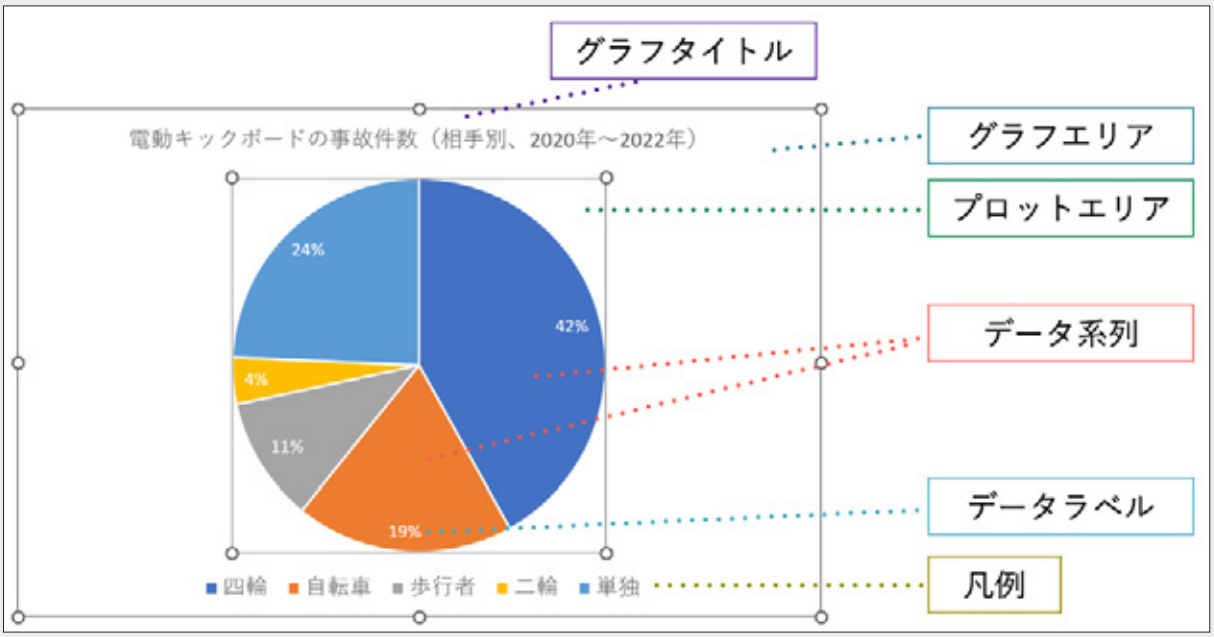


図 6　グラフに含まれる要素

上から順に、**グラフタイトル**は文字通りグラフ全体のタイトル、**グラフエリア**はグラフの全ての要素を含む領域、**プロットエリア**はグラフが描かれる領域、**データ系列**はデータの並び（引き出し線が多いと見づらくなるので、2 つしか描いていないが、一連のデータの並びのこと）、**データラベル**は各データの名前や値、**凡例**はデータ系列の名前。

上に示した要素以外にも、棒グラフや折れ線グラフでは**縦（数値）軸、横（項目）軸**などがあります。マウスポインタを位置付けると、ポップヒントに名前が表示されるので、一度確認しておくといいでしょう。また、それぞれの要素を右クリックすると［グラフタイトルの書式設定］や［グラフエリアの書式設定］などのように、ショートカットメニューにそれぞれの要素の書式設定を行うための項目が表示されます。

なお、データ系列やデータラベルなど、複数の要素から成り立っているものは、クリックして選択すると、それらの要素全てが選択でき、もう一度クリックすると、クリックした要素だけが選択できます。円グラフの 1 つの扇型の部分や、棒グラフの 1 つの棒の色を変えたいときにこの操作を使います。

重要度をランク付けする ～ パレート図を使って ABC 分析を行う

全体に対する個々の項目の割合や比率を表現するためのグラフとしては、円グラフだけでなくパレート図も使えます。パレート図は割合の大きい項目から順に棒グラフを作成し、それらの値の累計を折れ線グラフにしたものです。図 7 のデータは、不正アクセスが行われた後にどのような被害があったか、届け出などにより 2022 年に認知された件数をまとめたものです。出典は総務省のページに掲載されている［別紙］の PDF ファイルです。ただし、これ以外に表面化していない例もあるかもしれません。

	A	B	C
1	**不正アクセスによる被害認知件数（2022年）**		
2			
3	**区分**	**件数**	
4	インターネットバンキングでの不正送金等	1,096	
5	インターネットショッピングでの不正購入	227	
6	メールの盗み見等の情報の不正入手	215	
7	知人になりすましての情報発信	63	
8	オンラインゲーム・コミュニティサイトの不正操作	50	
9	暗号資産交換業者等での不正送信	32	
10	ウェブサイトの改ざん・消去	17	
11	インターネットオークションの不正操作	0	
12	その他	500	
13	**合計**	2,200	
14			

図 7　不正アクセスによる被害の認知件数

不正アクセスによって、どのような被害があったかをまとめた表。一見して不正送金等が多いのは分かるが、パレート図を使って、件数をランク付けしてみよう。

このデータを基に図 7 のようなパレート図を作ってみましょう。サンプルファイルをこちらからダウンロードし、［不正アクセス］ワークシートを開いて取り組んでみてください。手順は、図 7 の後に箇条書きで記しておきます。また、これについても動画で解説しているので、操作を一つ一つ追いかけたい方はぜひご視聴ください。

なお、Google スプレッドシートの場合はこちらのサンプルファイルを開いて、メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。

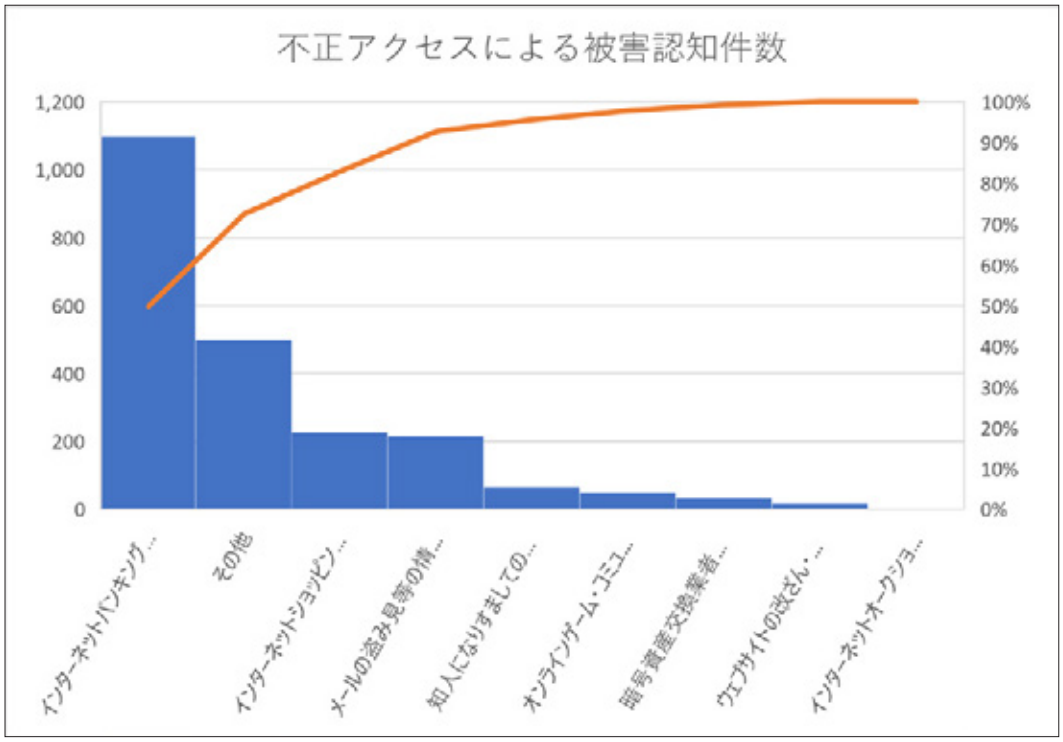


図 7　不正アクセスによる被害の認知件数をパレート図にする

棒グラフはそれぞれの件数を降順に並べたもの。自動的に値の降順に並べ替えられたグラフになるので、元のデータを降順に並べ替えておく必要はない。折れ線グラフは件数の累計。右側の第 2 軸が全体に占める割合を表す。

手順は以下の通りです。

- セル **A3** ～ **B12** を選択する
- ［挿入］タブを開き、［統計グラフの挿入］－［パレート図］を選択する

これだけでパレート図が作成できます。あとはタイトルを変更するだけです。ただし、パレート図では、セルの内容をグラフタイトルに表示することはできないので、タイトルは自分で入力する必要があります。

一方、Google スプレッドシートにはパレート図を作成する機能がありません。そこで、件数の降順に並べ替えを行った後、件数の累計を基に割合を累計し、棒グラフと折れ線グラフの複合グラフを作成する必要があります。

- 並べ替える
  - ・セル **A3** ～ **B12** を選択する
  - ・［Tab］キーを押し、アクティブセルをセル **B3** に位置付けておく
  - ・メニューバーから［データ］－［範囲を並べ替え］－［列 B を基準に降順で範囲を並べ替え］を選択する

- 累計を求めてパーセント表示にする
  - ・セル **A3** に「累計」と入力する
  - ・セル **C4** に「=B4/B13」と入力する
  - ・セル **C5** に「=C4+B5/$B$13」と入力する
  - ・セル **C5** をセル **C12** までコピーする
  - ・セル **C4** ～ **C12** を選択し、ツールバーの［表示形式をパーセントに設定］ボタンをクリックする
- 複合グラフを利用してパレート図を作る
  - ・セル **A3** ～ **C12** を選択する
  - ・メニューバーから［挿入］－［グラフ］を選択し、［グラフの種類］のリストから［複合グラフ］を選択する
  - ・系列（棒グラフの部分でよい）を右クリックし、［系列］－［累計］を選択
  - ・［グラフエディタ］の［カスタマイズ］画面で、［系列］の［軸］のリストから［右軸］を選択する

Google スプレッドシートでは、棒グラフの間隔を調整できないので、棒と棒の間にスペースが空いてしまいますが、図 7 と同様のグラフが作成できます。

さて、作成されたパレート図をどのように分析していけばいいでしょうか。図 9 のように、第 2 軸の **70%**の位置から左に向かって線を引き、折れ線とぶつかったところで線を下に引きます。さらに、第 2 軸の **90%**の位置から左に向かって線を引き、折れ線とぶつかったところで線を下に引きます。すると、横軸が 3 つの部分に分けられます。

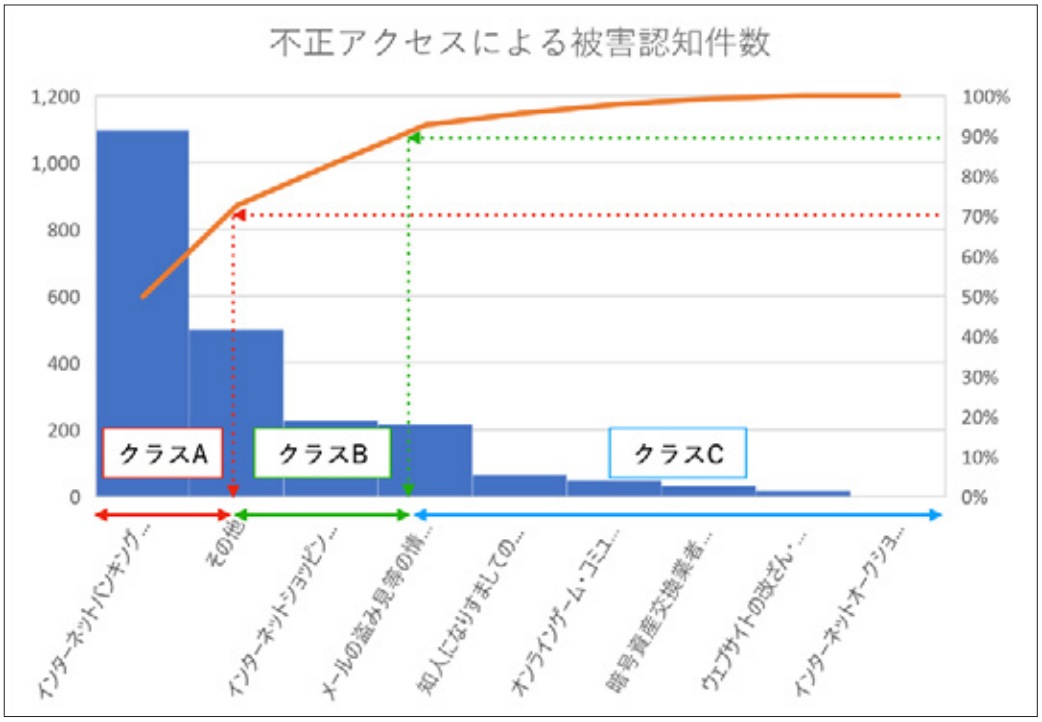


図 9　不正アクセスの被害に関する ABC 分析

ここでは全体の 70%までをクラス A とし、90%までをクラス B とした。クラス A は「インターネットバンキングでの不正送金」と「その他」、クラス B は「インターネットショッピングでの不正購入」と「メールの盗み見等の情報の不正入手」となっている。クラス C は右側の残りの部分。

3 つの部分の左側をクラス A とし、中央をクラス B とします。右側はクラス C です。

クラス A は全体の **70%**を占める重要な項目と考えられます。つまり、この部分にある項目に対策を施せば、不正アクセスによる被害の 70%は（理屈としては）防げるというわけです。

次にクラス B です。さらに、クラス B への対策を施せば 90%の問題が解決することになります。

クラス C は残りの 10%です。全体に占める割合が小さいので放置しておいていいかというと、そういうわけでもありません。例えば、件数は少なくても、行政や医療などに関連する Web サイトの改ざんやサーバーデータの消去などがあれば、国民生活への影響は甚大です。**件数の多さは必ずしも重要度の大きさであるとは限りません**が、重要度を測る一つの指標にはなります。

この例では「その他」の件数がかなり多く、その部分の実態がつかめないので、ひとまず「その他」は除外して ABC 分析を行ってもいいかと思われます（もちろん、「その他」にどのような事例があるのかを把握しておく必要はありますが）。ただし、実際に「その他」を除外して、分析を行ってもほぼ同じ結果になります。

このように全体を A、B、C という 3 つのクラスに分けて分析していくことを **ABC 分析**と呼び、さまざまな分野で活用されています。例えば、機械などの故障の原因についての ABC 分析を行えば、対応すべき問題に優先順位を付けるのに役立ちます。また、商品の売れ行きについての ABC 分析を行えば、主力商品として推していくべきものはどれか、あるいは、テコ入れすべき商品はどれかといった戦略の策定に役立ちます。

一般に、Amazon などのネットショッピングでは、取り扱う商品の種類が多いので、クラス C の項目が極めて多くなります。そのように右側に「尾」を引いている部分を「**ロングテール**」と呼びます。ロングテールの商品は重要度が低いというわけではなく、少量でも幅広く売れ続けるので、安定した売り上げに貢献します（クラス A の主力商品のみに依存していると、その商品の人気が落ちたときの影響が大きくなります）。ただし、取り扱う商品が多い分、いかに効率よく在庫管理を行うかがカギとなります。

規模と割合の変化を可視化する ～ 積み上げ縦棒グラフの利用

上の例では、2022 年のデータを基にパレート図を作成し、ABC 分析を行いました。しかし、時系列での変化も気になりますね。そこで、不正アクセスによる被害がどのように変化しているかを見てみましょう。実は、出典の PDF ファイルには過去 5 年間のデータが掲載されています（図 10）。

	A	B	C	D	E	F	G
1	不正アクセスによる被害認知件数（2018年～2022年）						
2							
3	区分＼年	2018	2019	2020	2021	2022	
4	インターネットバンキングでの不正送金等	330	1,808	1,847	693	1,096	
5	インターネットショッピングでの不正購入	149	376	172	349	227	
6	メールの盗み見等の情報の不正入手	385	329	234	175	215	
7	知人になりすましての情報発信	199	60	81	65	63	
8	オンラインゲーム・コミュニティサイトの不正操作	24	30	26	71	50	
9	暗号資産交換業者等での不正送信	169	22	18	20	32	
10	ウェブサイトの改ざん・消去	13	19	10	8	17	
11	インターネットオークションの不正操作	29	47	6	4	0	
12	その他	188	269	412	131	500	
13	合計	1,486	2,960	2,806	1,516	2,200	
14							

図 10　不正アクセスによる被害の認知件数（2018 年～ 2022 年）

このデータを基に、積み上げ縦棒グラフを作り、件数の変化を可視化してみよう。グラフを作成したら、どのようなことが言えそうか、分析してみよう。

このデータを基に、不正アクセスによる被害の認知件数がどのように変化しているかを可視化してみましょう。円グラフでは時系列での変化が可視化できないので、積み上げ縦棒グラフを使います。図 11 のようなパレート図を作ってみましょう。上で使ったサンプルファイルの［不正アクセス（5 年間）］ワークシートを開いて取り組んでみてください。手順は、図 11 の後に箇条書きで記しておきます。これについても動画で解説しているので、操作の手順を一つ一つ追いかけたい方はぜひご視聴ください。

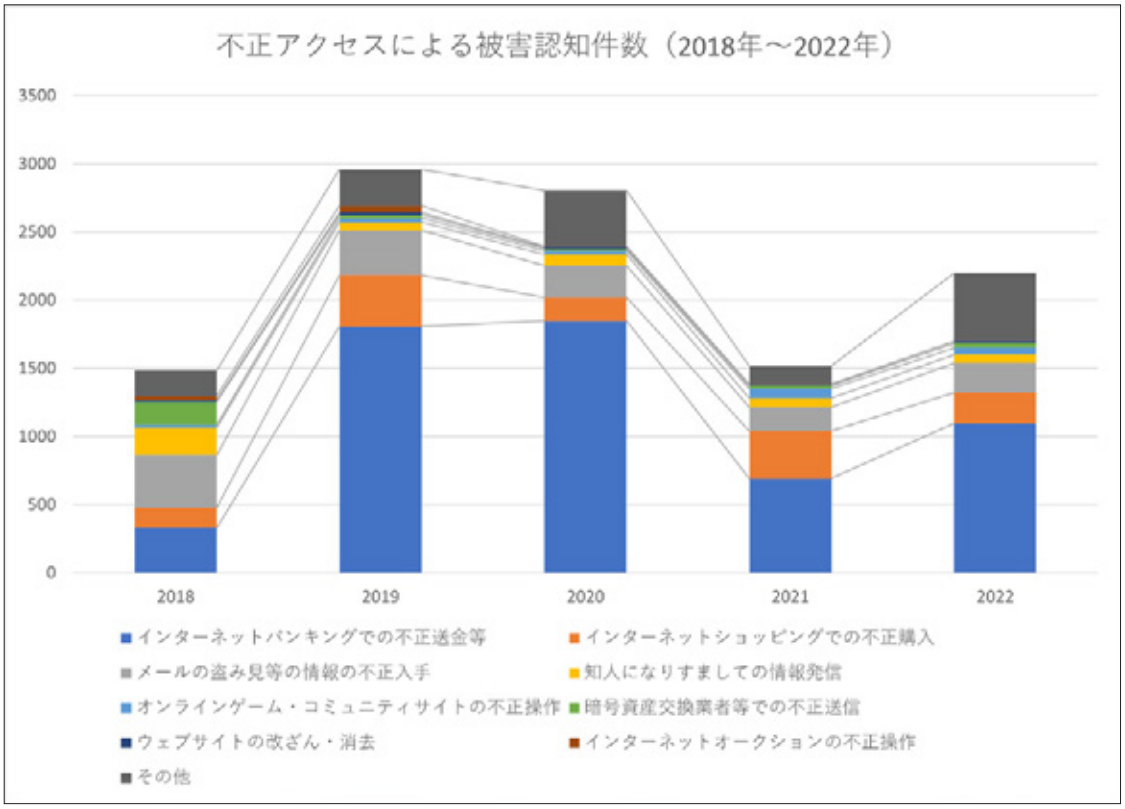


図 11　不正アクセスによる被害の認知件数を積み上げ縦棒グラフにする（2018 年～ 2022 年）

件数全体としては、年による波があるが、その要因の大部分は「インターネットバンキングでの不正送金等」のように見える。2018 年に多かった「暗号資産交換業者等での不正送信」は、2019 年以降少なくなっている。「インターネットオークションの不正操作」も減少している。

積み上げ縦棒グラフの作成手順は以下の通りです。

- セル **A4** ～ **F12** を選択する
  - ・3 行目を選択しないのは「年」の値（2018 や 2019 など）がグラフに含まれないようにするため
- ［挿入］タブを開き、［縦棒グラフ／横棒グラフの挿入］－［積み上げ縦棒］を選択する
- ［グラフのデザイン］タブを開き、［行／列の切り替え］ボタンをクリックする
- ［グラフのデザイン］タブを開き、［データの選択］ボタンをクリックする
- ［横（項目）軸ラベル］の［編集］ボタンをクリックする
- ［軸ラベルの範囲］ボックスをクリックし、セル **B3** ～ **F3** を選択する

［OK］ボタンをクリックしてダイアログボックスを閉じておいてください。グラフのサイズが小さいと棒の部分が密集して見づらいかもしれないので、グラフをドラッグして見やすいサイズに変更しておきましょう。図 11 のように各項目を結ぶ細い線（斜めの線）を表示するには以下の操作を行います。

- ［グラフのデザイン］タブを開き、［グラフ要素を追加］－［線］－［区分線］を選択する

あとはタイトルを指定するだけです。

Google スプレッドシートでは、以下のように操作します。

- セル **A3** ～ **F12** を選択する
- メニューバーから［挿入］－［グラフ］を選択し、［グラフの種類］のリストから［積み上げ縦棒グラフ］を選択する
- ［グラフエディタ］の［設定］画面で以下に示すチェックボックスを操作する（すでに設定されていれば操作は不要）
  - ・［行と列を切り替える］をオンにする
  - ・［列 A を見出しとして使用］をオンにする
  - ・［行 3 をラベルとして使用］をオンにする

Google スプレッドシートでは、Excel の区分線に対応する機能がないので、タイトルを設定すれば完成です。

作成されたグラフからどのようなことが読み取れそうでしょうか。年によって件数に波がありますが、その大部分は「インターネットバンキングでの不正送金等」によるものと思われます。詳細については背景となるできごとを精査しないと分かりませんが、不正送金に対応してセキュリティを強化しても、また新たな手口が登場し、またそれに対応し……という「いたちごっこ」になっているのかも知れません。いずれにしても、2019 年以降は「インターネットショッピングでの不正購入」と合わせて、お金にかかわる不正行為が大半を占めているようです。

少し細かくなりますが、「知人になりすましての情報発信」「暗号資産交換業者等での不正送信」「インターネットオークションの不正操作」については、2019 年以降、実数も割合も減っているようです。これらについてはグラフよりも数値を見た方が分かりやすいかもしれません。背景としては、不正アクセスへの対策強化が考えられます。特に、暗号資産については、2018 年 1 月のコインチェック事件以来、取引所のセキュリティ対策だけでなく、ユーザー側でも二要素認証を徹底するなど、意識の向上があったのではないかと思われます。ちなみに、暗号資産の口座数は 2018 年から 2021 年にかけて倍以上に増えています（日本暗号資産取引業協会の統計情報（PDF ファイル）による）。とはいえ「対策が強化されたのではないか」というのはあくまで仮説です。さらなる分析を行い、対策などに役立てていくには、実際にどのような出来事があり、どのような対策が取られたのかを詳しく調べる必要があります。逆に、実際に何らかの対策を行った後、トラブルが減少したというグラフが提示できれば、対策の有効性に対する説得力が高まります。

割合の変化だけを可視化するには ～ 100%積み上げ縦棒グラフの利用

今回は、事故や不正アクセスなどちょっと負の側面のデータばかりだったので、多少は楽しいデータも取り扱ってみましょう（といっても雰囲気を明るくすることが目的ではないのですが）。図 12 のデータは、総務省の社会生活基本調査のデータから、普段行うスポーツの人数のうち、球技のみを取り出して作成した表です。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	スポーツの種類別行動者数（2011年~2021年）単位：千人														
2															
3	年	野球	ソフトボール	バレーボール	バスケットボール	サッカー	卓球	テニス	バドミントン	ゴルフ	ボウリング	球技合計	その他合計（参考）	総合計	
4	2011	9,728	4,565	6,043	4,486	6,832	7,272	6,318	7,363	10,139	21,137	83,883	128,015	211,898	
5	2016	8,143	3,017	5,136	4,864	6,770	7,664	5,625	7,559	8,900	14,334	72,012	132,949	204,961	
6	2021	7,051	1,667	3,906	4,051	5,339	5,465	3,814	6,842	7,738	5,703	51,576	118,639	170,215	
7															

図 12　普段行うスポーツの種類と人数（2011 年～ 2021 年）

調査は 5 年ごとに行われる。このデータは最近の 3 回分。スポーツの種類は他にもあるが、ここでは 3 回の調査に共通して現れる球技だけを取り出した。人数は 10 歳以上の 19 万人のサンプルから計算された推定値で、そのスポーツを年に 1 日以上行った人の数を合計したもの。このデータを基に人気のスポーツの変化を可視化してみよう。なお、表には参考として球技以外の「その他合計」と「総合計」も含めてある。

この表を基に、それぞれの球技をたしなむ人の割合がどのように変化しているかを可視化してみましょう。ここでは、上で見た積み上げ縦棒グラフではなく、割合の変化だけを見るために 100%積み上げ縦棒グラフにします。積み上げ縦棒グラフであれば、球技人口の減少も可視化できますが、全体の人数（N 列）も減少しているので、球技人口の減少が強調されすぎるからです（全体に対する球技人口の割合を求めてグラフ化すれば意味のあるものになります）。

サンプルファイルをこちらからダウンロードし、［普段行う球技］ワークシートを開いて取り組んでみてください。手順は図 13 の後に箇条書きで記しておきます。ただし、積み上げ縦棒グラフとほとんど同じなので、動画での解説は省略します。Google スプレッドシートの場合はこちらのサンプルファイルを開いて、メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。

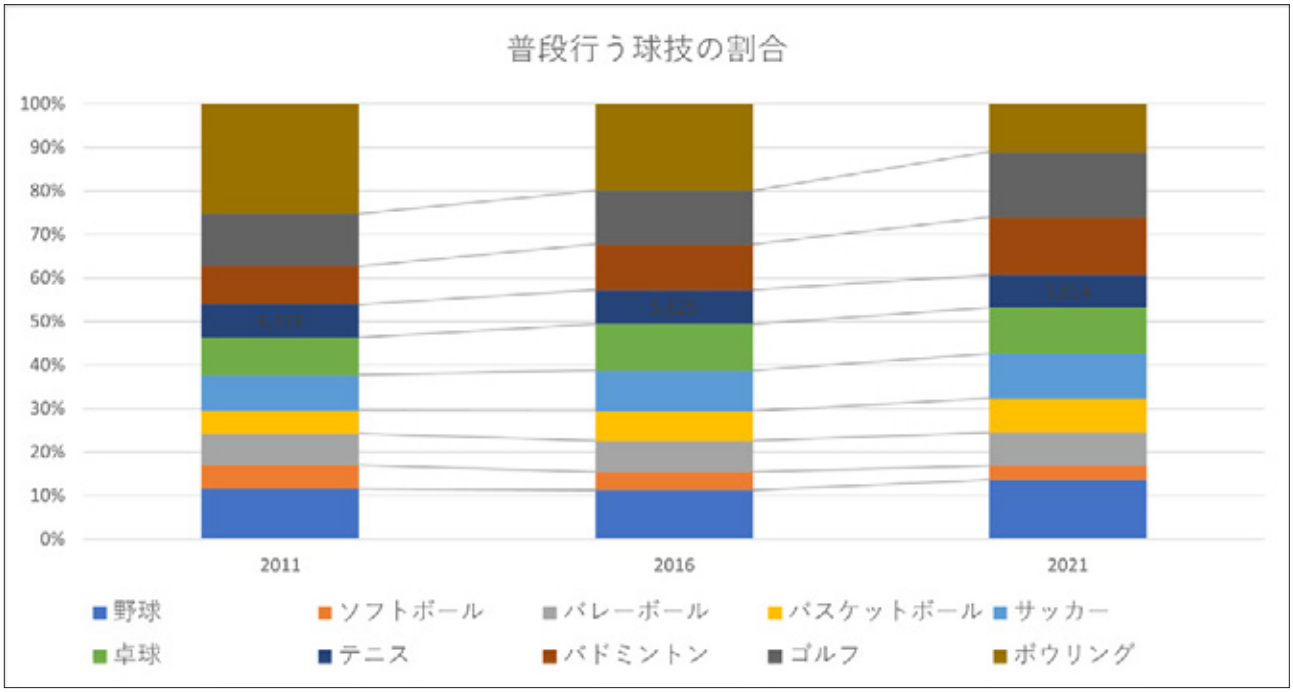


図 13　普段行うスポーツの割合の変化（2011 年～ 2021 年）

ボウリングの割合が減少し続けている。特に 2016 年から 2021 年にかけての減少が顕著。一方、野球、バドミントン、ゴルフなどの割合がわずかに増加しているように見える。

→目次に戻る

100%積み上げ縦棒グラフの作成手順は以下の通りです。

- セル **B3** ～ **K6** を選択する
  - ・A 行目を選択しないのは「年」の値（2011 や 2016 など）がグラフに含まれないようにするため
- ［挿入］タブを開き、［縦棒グラフ／横棒グラフの挿入］－［100%積み上げ縦棒］を選択する
- ［グラフのデザイン］タブを開き、［行／列の切り替え］ボタンをクリックする
- ［グラフのデザイン］タブを開き、［データの選択］ボタンをクリックする
- ［横（項目）軸ラベル］の［編集］ボタンをクリックする
- ［軸ラベルの範囲］ボックスをクリックし、セル **A4** ～ **A6** を選択する
- ［OK］ボタンをクリックして、ダイアログボックスを閉じる
- ［グラフのデザイン］タブを開き、［グラフ要素を追加］－［線］－［区分線］を選択する

Google スプレッドシートでは、以下のように操作します。

- セル A3 ～ K6 を選択する
  - ・メニューバーから［挿入］－［グラフ］を選択し、［グラフの種類］のリストから［100%積み上げ縦棒グラフ］を選択する
- ［グラフエディタ］の［設定］画面で以下に示すチェックボックスを操作する（すでに設定されていれば操作は不要）
  - ・［行と列を切り替える］をオフにする
  - ・［行 3 を見出しとして使用］をオンにする
  - ・［列 A をラベルとして使用］をオンにする

ボウリングが減少しているのは、ボウリングの斜陽化が原因なのかもしれませんが、2021 年の減少については新型コロナ禍の影響も大きいのでしょう。野球、バドミントン、ゴルフなど、屋外でできる球技に関しては、2016 年から 2021 年かけては増加の傾向にあります。もちろん、スポーツそのものの人気の変化もあると思われます。バスケットボールやサッカーなどは着実に増加しています。

残念ながら、1970 年代の第一次ボウリングブームの終焉（しゅうえん）の後、1990 年代にやや盛り上がりを見せたものの、ボウリング場の数は年々減少しています（日本ボウリング協会の報道資料（PDF ファイル）による）。にもかかわらず、ボウリングの割合がそれほど小さくない（サッカーより多い！）のを意外に思われる方もおられるかもしれません。実は、社会生活基本調査のデータを見ると、ボウリングの平均行動日数はかなり少なくなっています。つまり、年に数回しかしない人が大多数だというわけです。とすると、「普段行うスポーツ」というのはかなり語弊がありますね。そこで、2021 年のデータについて平均行動日数が週 1 日未満の場合と週 1 日以上の場合に分けて人数を集計し、どのスポーツが「普段行う」ものなのか「たまに行う」ものなのか、違いを可視化してみたいと思います。データは図 14 の通りです。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	スポーツの種類別頻度別行動者数（2021年）単位：千人												
2													
3	頻度	野球	ソフトボール	バレーボール	バスケットボール	サッカー	卓球	テニス	バドミントン	ゴルフ	グラウンドゴルフ	ボウリング	
4	週1回未満	5,689	1,368	2,876	2,984	4,049	4,575	2,383	5,972	6,042	1,375	5,378	
5	週1回以上	1,098	224	852	839	1,024	723	1,274	651	1,542	509	215	
6	合計	6,787	1,592	3,728	3,823	5,073	5,298	3,657	6,623	7,584	1,884	5,593	
7													

図 14　普段行うスポーツの種類と人数（2021 年頻度別）

2021 年の調査項目には、高齢者向けに考案された「グラウンドゴルフ」も含まれている。合計の人数が図 12 と一致しない理由は不明だが、恐らく頻度について未回答のデータがあったものと思われる。このデータを基に割合を比較するグラフを作成してみよう。

この表を基に、それぞれの球技を行う頻度を比較するグラフを作成してみましょう。図 13 のような 100%積み上げ縦棒グラフを使っても構いませんが、時系列での比較ではないので、100%積み上げ横棒グラフの方がよさそうです。先ほどと同じファイルの［ふだん行う球技（頻度）］ワークシートを開いて取り組んでみてください。手順は図 15 の後に箇条書きで記しておきます。こちらも手順は積み上げ縦棒グラフや 100%積み上げ縦棒グラフとほとんど同じなので、動画での解説は省略します。

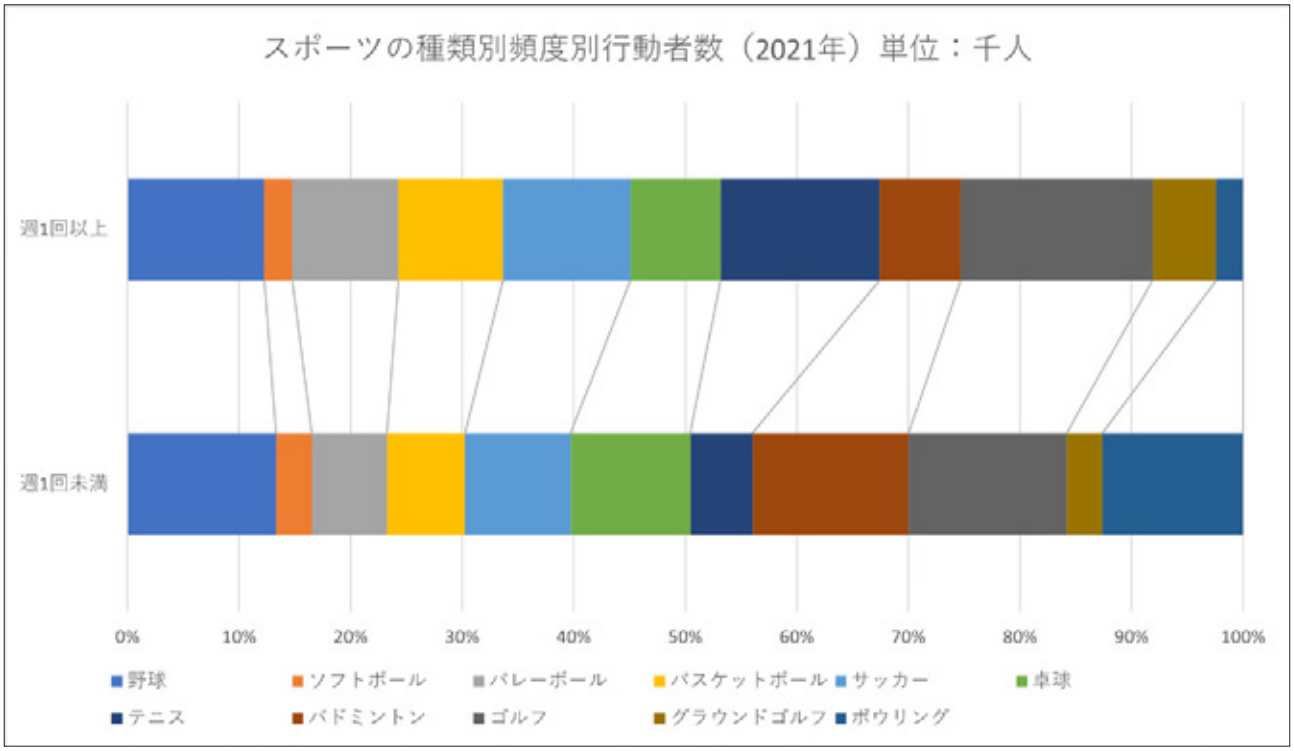


図 15　普段行うスポーツの割合の頻度別比較（2021 年）

グラフの右端を見ればボウリングを行う人のほとんどが週 1 回未満であることが可視化できる（週 1 回以上の人はボウリングを行ったと回答した人のわずか **3.8%**）。バドミントンも休日に公園で遊んだり、気分転換に行う程度かもしれない。バレーボール、バスケットボール、サッカー、テニス、ゴルフなどでは、週 1 回以上の人がかなり多い。

100%積み上げ横棒グラフの作成手順は以下の通りです。

- セル **A3** ～ **L5** を選択する
- ［挿入］タブを開き、［縦棒グラフ／横棒グラフの挿入］－［100%積み上げ横棒］を選択する
- ［グラフのデザイン］タブを開き、［行／列の切り替え］ボタンをクリックする
- ［グラフのデザイン］タブで［グラフ要素を追加］－［線］－［区分線］を選択する

Google スプレッドシートでは、以下のように操作します。

- セル **A3** ～ **L5** を選択する
  - ・メニューバーから［挿入］－［グラフ］を選択し、［グラフの種類］のリストから［100%積み上げ横棒グラフ］を選択する
- ［グラフエディタ］の［設定］画面で以下に示すチェックボックスを操作する（すでに設定されていれば操作は不要）
  - ・［行と列を切り替える］をオフにする
  - ・［行 3 を見出しとして使用］をオンにする
  - ・［列 A をラベルとして使用］をオンにする

出典のデータには年齢層別の人数も含まれています。バレーボール、バスケットボール、サッカー、テニスは、若年層の人数が多くなっており、週 1 回以上の割合が大きいのは学校での部活に参加しているためと考えられます。ゴルフに関しては、いわゆるゴルフ中毒に陥り、練習場に行かずにはいられない人が多いのかもしれません。ゴルフの年齢層は全体的にかなり高めです。

なお、図 15 のような割合（比率）の比較には、ドーナツグラフも使えます。サンプルファイルにはドーナツグラフの作成例も含めてあるので、ぜひご参照ください。ただし、Google スプレッドシートのドーナツグラフでは複数の輪を同時に表示することができないので、サンプルファイルには含めてありません。

今回は、円グラフを使った割合の可視化と落とし穴の確認から始め、パレート図による ABC 分析、積み上げ縦棒グラフによる割合の変化などについて見てきました。割合の大きさが必ずしも重要度の高さであるとは限りませんが、方針の立案などに役立つ、一つの手がかりになることは確かです。

次回は、集団の全体像を見るためのケーススタディーを通して、ヒストグラムや箱ひげ図について、作成方法や設定の変更方法を詳しく見ていきます。次回も、落とし穴や意外に知られていない機能なども紹介します。どうぞお楽しみに！

［データ分析］ヒストグラムや箱ひげ図で「分布」を可視化 ～ 集団の特徴や外れ値を見つける

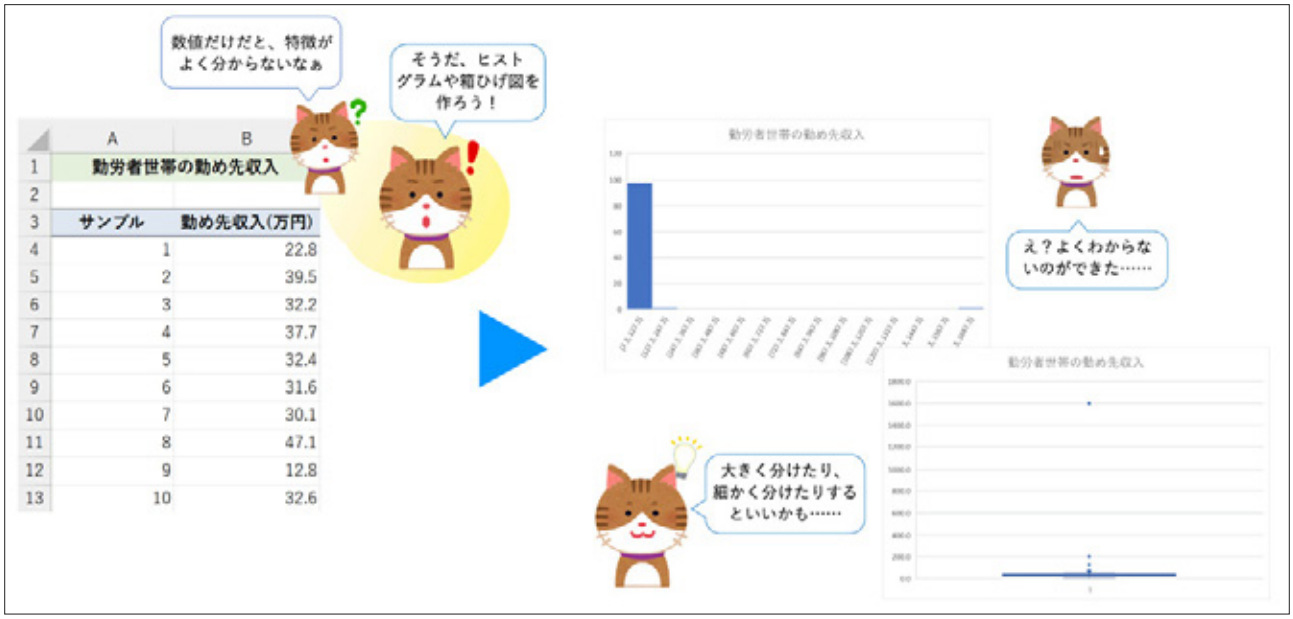
データ分析を初歩から学ぶ連載の第 10 回。グラフを使って集団の特徴や外れ値を可視化します。ヒストグラムや箱ひげ図の作成方法と、ピボットテーブル／ピボットグラフによる視覚的な分析のコツを、ケーススタディを通して学びましょう。

羽山博（2023 年 10 月 26 日）

いきなりですが、読者のみなさんは「分析」という言葉の意味をじっくりと考えてみたことはあるでしょうか。「分」は、分けるということですね（刀で左右に切り離す）。「析」はちょっと難しいですが、細かく分けるという意味です（斤＝「おの」で木を切る）。つまり、分析とは大きく分けたり、細かく分けたりすること……なのですが、そのためには**全体像を見て、どのように分けるかを決める**必要があります。というわけで、今回は可視化により全体像を見ることと、全体が何らかの特徴によってどう切り分けられるかを見ることに焦点を当てます。

具体的には、ヒストグラムと箱ひげ図を利用します。ヒストグラムは、集団の代表値を求めるお話（この連載の第 3 回）の中で紹介しました。また、箱ひげ図については散布度を求めるお話（この連載の第 5 回）の中で紹介しました。いずれも、どのようなグラフなのかをお見せしただけで、作成手順については触れていませんでした。そこで、今回はそれらのグラフの具体的な作成手順や書式の設定方法、より詳細な分析方法を見ていくことにします。

ところで、ヒストグラムや箱ひげ図を作成すると、図 1 のような、いびつなグラフが作られることもよくあります。データは架空のものですが、勤労者世帯の勤め先収入の平均値 (月 49.2 万円) と一致するように作成してあります。平均値の出典は、総務省統計局の家計調査の統計表（Excel ファイル）に掲載された 2022 年の値です。



	A	B
1	勤労者世帯の勤め先収入	
2		
3	サンプル	勤め先収入(万円)
4	1	22.8
5	2	39.5
6	3	32.2
7	4	37.7
8	5	32.4
9	6	31.6
10	7	30.1
11	8	47.1
12	9	12.8
13	10	32.6

図 1　全体像を可視化して「切り分ける」ポイントを知る

数値だけでは分からない特徴を可視化するために、ヒストグラムや箱ひげ図を作成し、まずは分布を見てみよう。この例のように整ったグラフにならないことも多いが、そういう場合こそ、切り分けるポイントが分かりやすい。値が集中している箇所と離れている箇所を大きく分けたり、値が集中している箇所を細かく分けて調べていくとよい。

統計学の教科書で紹介されているヒストグラムは、真ん中あたりに山があって、左右に裾が広がっている「整った」形のグラフが多いようです。ヒストグラムがどのようなものかを理解するにはいいのですが、実際には図 1 に示したような、いびつな形のヒストグラムになることがよくあります。しかし、整った形のグラフにはそれ以上の特徴があまりありません。むしろ、いびつな形のグラフからの方が、興味深い特徴を見つけやすいものです。

今回は集団の特徴を可視化するというテーマについて、幾つかの例を見ていきます。つまり、図 1 のグラフをどう切り分けて特徴を見つけていくかということです。また、データを多角的に分析するには、ピボットテーブル／ピボットグラフも便利です。ピボットテーブル／ピボットグラフに苦手意識を持つ人も多いようですが、分かりやすく説明していきます。

この記事は、データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 10 回です。第 7 回の棒グラフ、第 8 回の折れ線グラフ、第 9 回の円グラフ／パレート図から、今回のヒストグラム、箱ひげ図、第 11 回のクロス集計表、ヒートマップ、第 12 回の散布図まで、1 つずつ可視化の基礎を学んでいきます。これらのグラフの目的と効用などについて、特別予告編で簡単に整理していますので、事前に確認しておくとより理解が深まるでしょう。

この記事で学べること

今回は以下のようなポイントについて、分析の方法や目の付け所を見ていきます。

- ヒストグラムの作成と適切な設定 …… 全体的な特徴と部分の特徴を見る
- 箱ひげ図の作成と適切な設定 …… 外れ値を見つける
- ピボットテーブル／ピボットグラフを利用した多角的な分析 …… 属性別にヒストグラムを作成、比較する

では、ヒストグラムの作成から見ていきましょう。まず、特に何も指定せずに図 1 のようなヒストグラムを作り、書式の設定を変えながら何が読み取れるかを見ていきます。では、サンプルファイルの利用についての説明の後、本編に進みましょう。

サンプルファイルの利用について

本稿では、表計算ソフトを使って手を動かしながら学んでいきます。表計算ソフト Microsoft Excel 用の .xlsx ファイルをダウンロードできるようにしています。デスクトップ版の Excel が手元にない場合は、Microsoft アカウントがあれば使える無料の Microsoft 365 オンライン、もしくは Google アカウントがあれば使える無料の Google スプレッドシート（Google Sheets）をお使いください。Microsoft 365 オンラインの場合は、.xlsx ファイルを OneDrive にアップロードしてから開いてください。Google スプレッドシートの場合は、.xlsx ファイルを Google ドライブにアップロードしてから開いた上で［ファイル］メニューの［Google スプレッドシートとして保存］を実行してください（Google スプレッドシート独自の機能を使っている場合は、ファイルを共有して参照できるようにします。その場合は、該当する箇所で使い方を記します）。

ヒストグラムから特徴を読み取る ～ いびつなグラフこそ情報の宝庫

最初に見た勤め先収入のデータはサンプルファイル（後述）のセル **B4** ～ **B103** までに入力されています。セル **B3** の項目見出しを含めて、ヒストグラムを作成し、設定を変更してみましょう。図 2 の左側が特に何も指定せずにヒストグラムを作った例で、右側が書式を変更して適切な形式にしたものです。サンプルファイルをこちらからダウンロードし、［勤め先収入 1］ワークシートを開いて取り組んでみてください。Google スプレッドシートの場合はこちらのサンプルファイルを開いて、メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。

手順は図の後に箇条書きで示しておきます。ただし、タイトルなど、データ分析そのものにあまり関係のない設定については省略してあります。なお、動画でも手順を解説しているので、操作を一つ一つ追いかけたい方はぜひご視聴ください。

図 2　勤め先収入のヒストグラム（取りあえず作成したものと書式を設定したもの）

上側のグラフは特に何も指定せずに作成したヒストグラム。値が集中している左端の部分を詳しく表示し、それ以外の値が離れている部分をまとめたものが下側のグラフ。

→目次に戻る

手順は以下の通りです。

- セル **B3** ～セル **B103** をドラッグして選択する
- ［挿入］タブを開き、［統計グラフの挿入］－［ヒストグラム］を選択する
  - Google スプレッドシートの場合は、メニューバーから［挿入］－［グラフ］を選択し、［グラフの種類］のリストから［ヒストグラム グラフ］を選択する

これで上側のグラフが作成できます。左端の棒は、収入が **7.3 万円**より大きく、**127.3 万円**以下である人数を表しますが、かなり幅が広いですね。これではほとんどの人がこの範囲に入ってしまいます。そこで、階級の幅を **5 万円**にして、その部分を詳しく見られるようにしましょう。また、収入が **127.3 万円**より大きい人の数は少ないので、その部分をまとめてしまいましょう。操作を続けます。

- 横軸を右クリックし、［軸の書式設定］を選択する
- ［ビンの幅］に 5 と入力する（階級の幅が 5 万円になる）
- ［ビンのオーバーフロー］のチェックマークをオンにし、「100」と入力する（**100 万円**より大きな階級がまとめられる）
- ［ビンのアンダーフロー］のチェックマークをオンにし、「10」と入力する（**10 万円**以下の階級がまとめられ、キリのいい目盛りになる）

Google スプレッドシートの場合は、以下のように操作します

- グラフを右クリックし［軸］-［横軸］を選択し、［グラフエディタ］を表示する
- ［カスタマイズ］タブをクリックする
- ［最小値］に「10」を入力する
- ［最大値］に「100」を入力する

いかがでしょう。下側のグラフを見ると、全体的に山が左に寄っていますね。つまり、大半の人がこの山の中にいて、収入のかなり大きな人が少数いるということが分かります。特に、収入が **30 万円**より大きく **35 万円**以下の階級の人数が多いようですが、**10 万円**より大きく **15 万円**以下の階級にも小さな山があるようです。このことから、勤労者世帯が、少数の高収入の層、多数の人が属する **30 万円**程度の層、一定数の低収入の層に分かれることが示唆されます。

なお、Google スプレッドシートでは、最小値以下の階級がまとめられたり、最大値より大きな階級がまとめられるのではなく、単に最小値以下や最大値より大きな階級が表示されなくなるだけです。また、階級の幅も自動的に決められるので、このままでは、上で説明したような 3 つの階層に分かれるという示唆は得られません（ただし、最大値を「55」に設定して、より詳細に表示すると、図 2 の下側と同じようなグラフになります）。

図 2 から、勤め先収入の最頻値は、最も度数の大きな階級の下限と上限の平均、つまり、(35 + 30)÷2 = **32.5（万円）**であることが分かります。また、空いているセルに「=AVERAGE(B4:B103)」と入力すると、平均値が **49.2（万円）**であることが分かり、「=MEDIAN(B4:B103)」と入力すると、中央値が **31.4（万円）**であることも分かります。平均値が大きくなっているのは、分布に偏りがあり、少数の大きな値に引きずられているからですね。

箱ひげ図により四分位範囲を可視化する ～ 大半のデータがどの範囲にあるかを知る

勤め先収入が 3 つの階層に分かれるのではないかということについて、さらに掘り下げていきたいところですが（後で見るのでお楽しみに！）、その前に、箱ひげ図の作成に取り組んでおきましょう。箱ひげ図を作成すると四分位範囲が可視化できます。つまり、順位を基にして、全体の **25%～ 75%**（中央部分に位置する半数）が属する範囲が分かります。また、大きく離れた値（外れ値）も可視化できます。

では、上で見たファイルの［勤め先収入 2］ワークシートを開いて、図 3 のように箱ひげ図を作成し、設定を変更してみましょう。データは［勤め先収入 1］ワークシートと全く同じです。グラフを作成しやすいように別のワークシートにしてあるだけです。残念ながら、Google スプレッドシートには今のところ箱ひげ図の機能がないので、サンプルファイルには含めていません。

手順は図の後に箇条書きで示しておきます。なお、動画でも手順を解説しているので、操作を一つ一つ追いかけたい方はぜひご視聴ください。

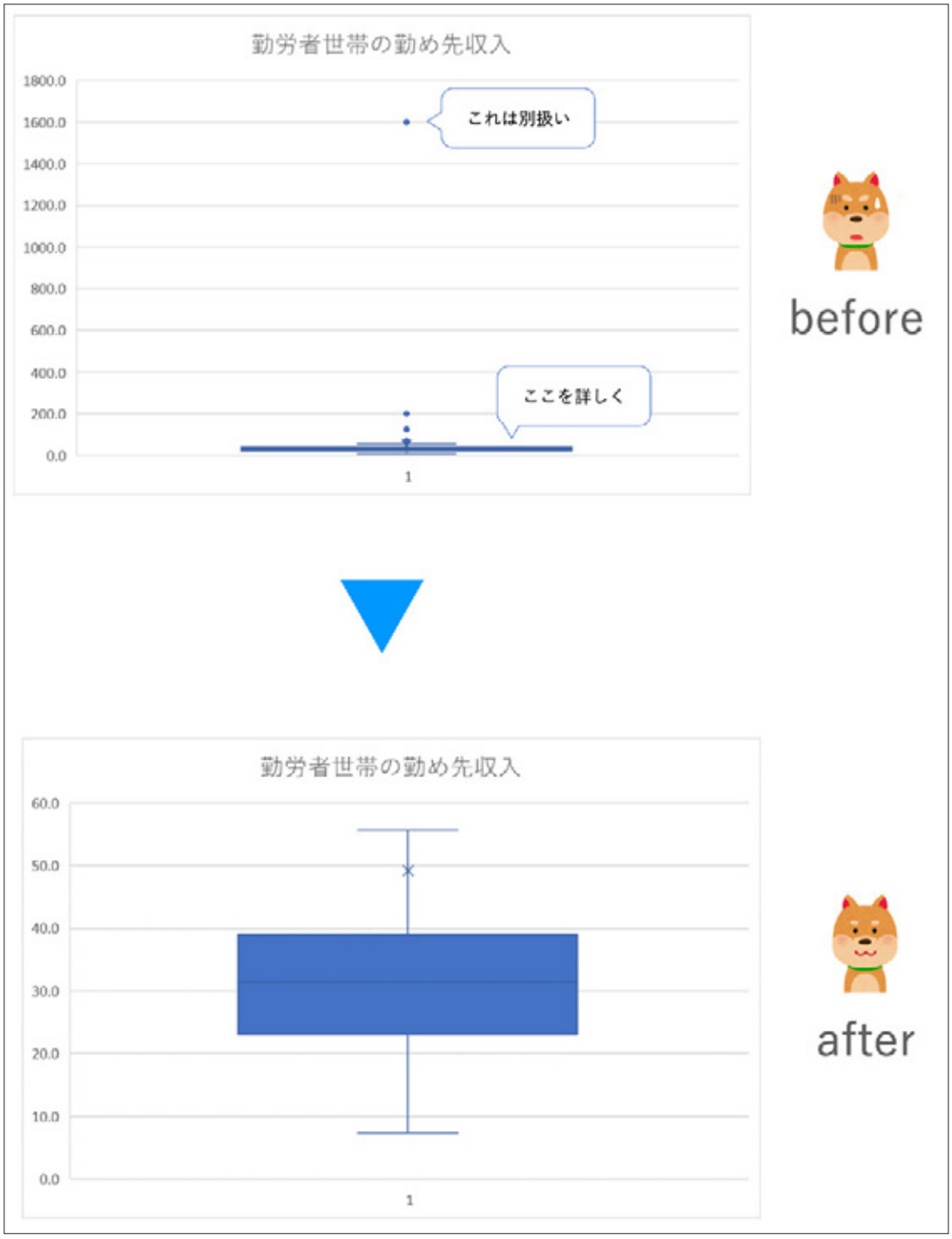


図 3　勤め先収入の箱ひげ図（取りあえず作成したものと書式を設定したもの）

上側のグラフは特に何も指定せずに作成した箱ひげ図。値が集中している下の部分を詳しく表示するために、小さな●で示された外れ値を除外したものが下側のグラフ。下側のグラフの中央の四角い部分が四分位範囲。四分位範囲の中央にある横線が中央値。上の × が平均値。「ひげ」の上限（上の横線の位置）は第 3 四分位数＋ 1.5× 四分位範囲の値以下の最も近い値となり、下限（下の横線の位置）は第 1 四分位数－ 1.5× 四分位範囲以上の最も近い値となる。ひげの外側が外れ値となる。

手順は以下の通りです。

- セル **B3** ～セル **B103** をドラッグして選択する
- ［挿入］タブを開き、［統計グラフの挿入］－［箱ひげ図］を選択する

これで上側のグラフが作成できます。外れ値が幾つかあることはよく分かりますが、値が集中している部分は下の方にわずかに表示されているだけです。外れ値（Excel では**特殊ポイント**と呼ばれます）を表示しないようにしてみましょう。

- データ系列（グラフの箱やひげの部分）を右クリックし、［データ系列の書式設定］を選択する
- ［データ系列の書式設定］作業ウィンドウの［系列のオプション］ボタンをクリックする
- ［特異ポイント］をクリックしてチェックマークをオフにする

図 3 の下側のグラフになり、四分位範囲がよく分かるようになりました。**20 万円〜 40 万円**あたりですね。この四分位範囲は QUARTILE.EXC 関数で求められる値を基に描かれています。QUARTILE.INC 関数で求められる値を基にしたい場合には、上で見た［データ系列の書式設定］作業ウィンドウで［四分位数計算］の［包括的な中央値］をクリックして、設定をオンにします。

空いているセルに「=QUARTILE.EXC(B4:B103,1)」と入力すれば、第 1 四分位数が **23.05** であることが分かります。また、「=QUARTILE.EXC(B4:B103,3)」と入力すれば、第 3 四分位数が **39.08** であることが分かります。

なお、外れ値として表示されている小さな●にマウスポインタを位置付けると、その値がポップアップ表示されます。図 3 の左の例では、上から順に **1600**、**200**、**124**、**68.7** の 4 つが外れ値と見なされています。外れ値の検出には、この連載の第 4 回で紹介したスミルノフ・グラブス検定なども使えます。

コラム　バイオリン図では値が集中している箇所も分かる

箱ひげ図の四分位範囲は四角で表されているので、ヒストグラムのような「山」が表せません。そこで、度数を反映したような表示にできる**バイオリン図（バイオリンプロット）**が使われることもあります。残念ながら Excel や Google スプレッドシートにはバイオリン図の機能がないので、Python や R などを使う必要があります。

以下の例は、Python から Excel のデータを読み込み、バイオリン図を作成したものです。このリンクをクリックすれば、ブラウザが起動し、Google Colaboratory で以下のコードが表示されます（Google アカウントでのログインが必要です）。［ドライブにコピー］ボタンをクリックすれば、自分の Google ドライブにコピーできます。コード（リスト 1）の部分をクリックして［Shift］+［Enter］キーを押せばグラフ（図 4）が描画されます。ぜひ試してみてください。

```
import pandas as pd
import matplotlib.pyplot as plt

# データの読み込み
df = pd.read_excel("https://github.com/Gessys/data_analysis/raw/main/10a.xlsx", sheet_name="勤め先収入1", usecols="A:B", skiprows=2, index_col="サンプル")

# バイオリン図の作成
graph = plt.violinplot(df.loc[:, "勤め先収入(万円)"],
                showmedians=True, showmeans=True, quantiles=[0.25, 0.75])
graph['cquantiles'].set(color="C1") # 四分位範囲をオレンジ色で表示する
graph['cmedians'].set(color="C2", linewidth=0.5, linestyle="dotted") #中央値を緑色で表示する
graph['cmeans'].set(color="C3", linewidth=0.5, linestyle="dotted") # 平均値を赤色で表示する
plt.ylim([0, 200])
plt.xticks(ticks=[1],labels=["income"])
plt.show()
```

リスト1　バイオリン図を表示するためのコード

詳細については割愛するが、pandas モジュールの read_excel 関数で Excel のデータを読み込み、matplotlib.pyplot モジュールの violinplot 関数にデータや中央値の表示（showmedians）、平均値の表示（shoemeans）、四分位範囲の表示（quantiles）などの引数を指定して描画を行う。

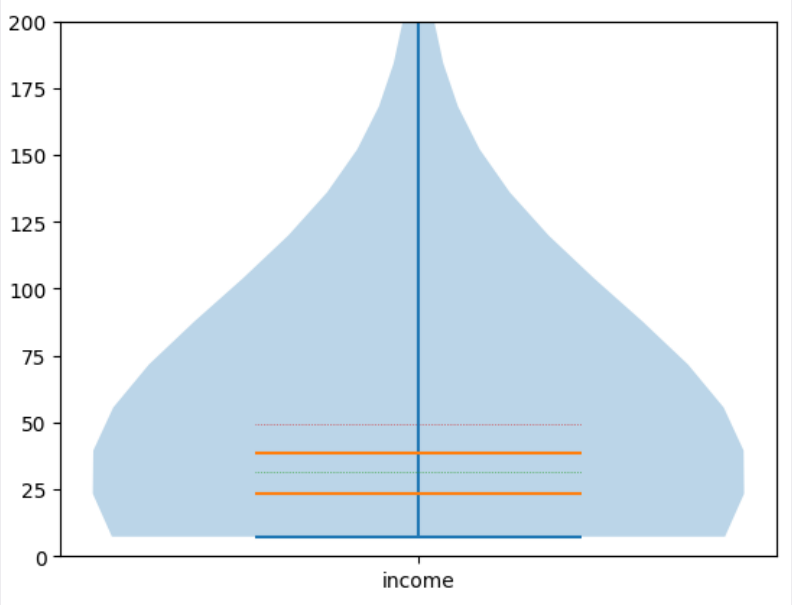


図 4　バイオリン図の例

四分位範囲がオレンジの実線で、中央値が緑の点線で表示されている。赤の点線で表示されている平均値が四分位範囲の外にあることも分かる。最小値は青の実線で表示されているが、y 軸の範囲を限定したので最大値は表示されていない。どのあたりの度数が大きいかも分かるのがバイオリン図のメリット。

ピボットテーブル／ピボットグラフを活用する ～ 属性別にヒストグラムを作る

実は、勤め先収入のデータは、性別や雇用形態が全て「込み」になっています。答えから先に言うと、中央値よりも左の階級に小さな山があるのは、性別や雇用形態による収入の格差が原因です。性別や雇用形態を含めたデータは図 5 のようなものになっています。

	A	B	C	D	E
1	勤労者世帯の勤め先収入				
2					
3	サンプル	勤め先収入(万円)	性別	正規/非正規	
4	1	22.8	F	N	
5	2	39.5	F	N	
6	3	32.2	F	N	
7	4	37.7	F	N	
8	5	32.4	F	R	
9	6	31.6	F	R	
10	7	30.1	F	N	
11	8	47.1	M	R	
12	9	12.8	F	N	
13	10	32.6	F	R	

図 5　勤め先収入のデータ（性別・雇用形態別）

データはセル **B3** ～ **D103** に入力されており、3 行目は項目の見出しになっている。性別は **F** が女性、**M** が男性を表す。雇用形態は **R** が正規職員、**N** が非正規職員を表す。性別と雇用形態で勤め先収入の分布がどう違うかを知りたいのだが、単純にヒストグラムを作るのはちょっと面倒。そこで、ピボットテーブル／ピボットグラフを使う。

この例では、明らかに性別や雇用形態といった属性が分かるようになっていますが、現実には、隠された要因によるものであることもよくあります。ともあれ、性別と雇用形態別にヒストグラムを作成し、それらを比較してみましょう。

これまでの方法で、男性だけのヒストグラム、女性だけのヒストグラム……のようにグラフを作成してもいいのですが、かなり面倒ですね。そこで、ピボットテーブル／ピボットグラフを使いましょう。ピボットグラフを使えば直接グラフが作成できます。

手順は以下の通りです。ステップごとに図を示しておくので、操作が細かくなるので、丁寧に進めてください（といっても、いくらでもやり直しはできます）。動画でも手順を解説しているので、手順を確認しながら進めるのが苦手な方や、ステップごとに操作を追いかけたい方は、ぜひ動画をご視聴ください。

では、手順です。サンプルファイルをこちらからダウンロードし、［勤め先収入 3］ワークシートを開いて取り組んでみてください。Google スプレッドシートではピボットテーブルの作成とグラフ作成の機能が別になっているので、Excel での操作の後にまとめて手順を掲載します（Microsoft 365 オンラインでは、後述の［グループ化］ができないため、説明を割愛します）。

Excel での操作手順

- セル **A3** ～ **D103** のいずれかのセルをクリックしておく
- ［挿入］タブを開き、［ピボットグラフ］－［ピボットグラフとピボットテーブル］を選択する
- ［ピボットテーブルの作成］ダイアログボックスの［テーブルまたは範囲を選択］がオンになっていることを確認する
- ［テーブル／範囲］が「勤め先収入 3!A3:D103」になっていることを確認する
- ［ピボットテーブルレポートを作成する場所を選択してください］の下の［既存のワークシート］をクリックしてオンにする
- ［場所］ボックスをクリックし、セル **F3** をクリックする

これで、空のピボットテーブルと空のピボットグラフが作成されます。画面は図 6 のようになります。

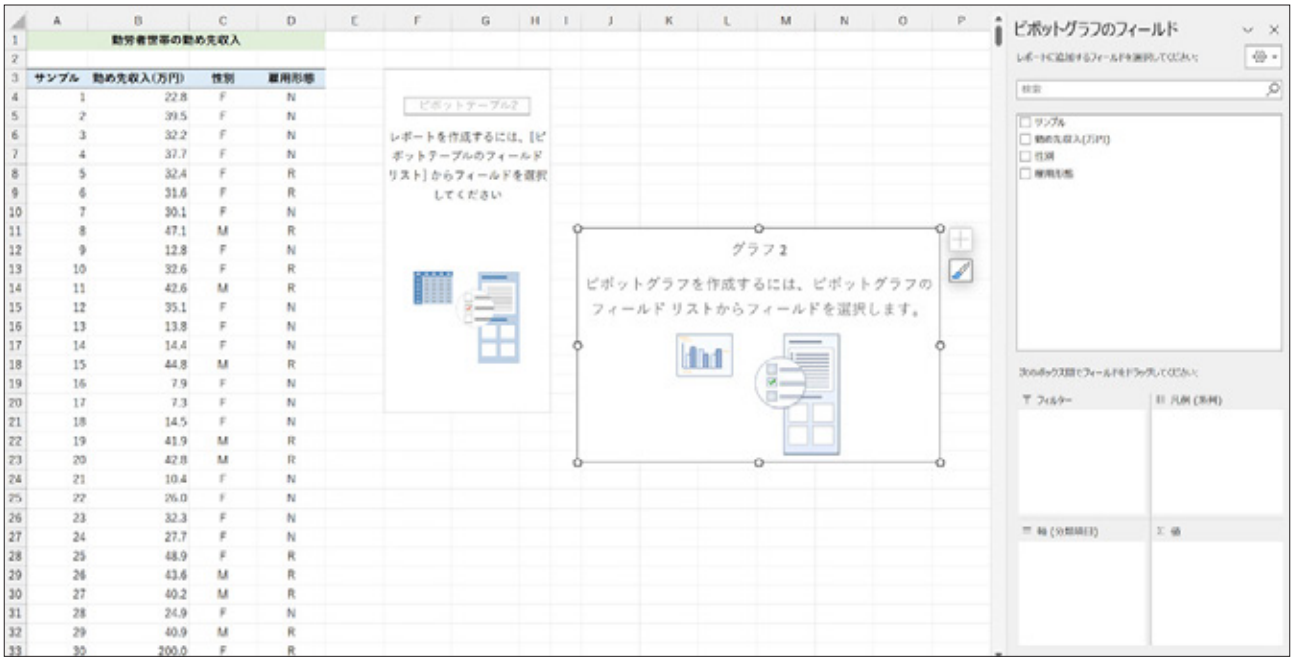


図 6　空のピボットテーブルと空のピボットグラフが作成された

この段階では、どの項目のどの値を集計し、グラフ化するかが指定されていないので、ピボットテーブルもピボットグラフも空の状態になっている。ここから、集計項目や集計の方法などを指定していく。

まず、勤め先収入ごと、性別ごとに人数を集計しましょう。

- ［ピボットテーブルのフィールド］のリストにある［勤め先収入（万円）］を［軸（分類項目）］の欄にドラッグする
- ［ピボットテーブルのフィールド］のリストにある［性別］を［凡例（系列）］の欄にドラッグする
- ［ピボットテーブルのフィールド］のリストにある［サンプル］を［Σ値］の欄にドラッグする
- ［Σ値］の欄の「合計／サンプル」をクリックし、［値フィールドの設定］を選択する
- ［選択したフィールドのデータ］リストから［個数］を選択し、［OK］をクリックする

この段階では、勤め先収入の値（F 列）に対する人数が、1 つずつ表示されています。たとえば、**7.3 万円**の女性が **1 人**、**7.9 万円**の女性が **1 人**といったぐあいです（図 7）。

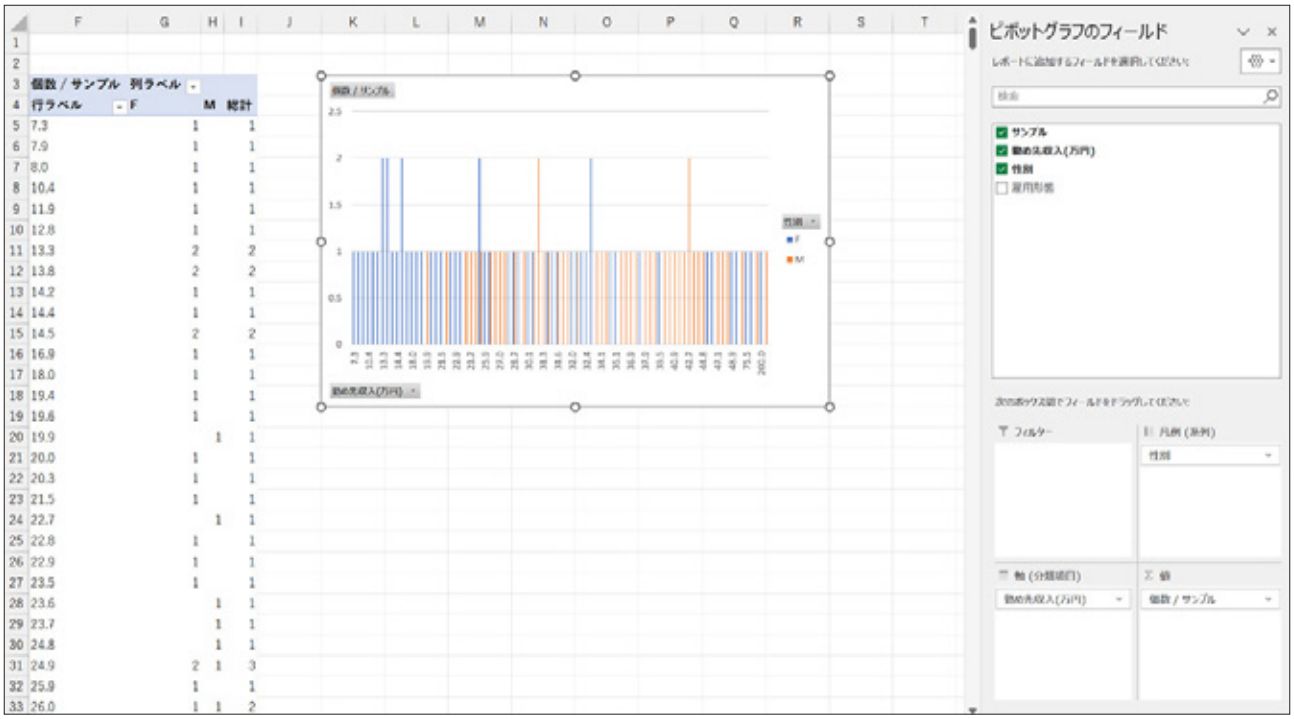


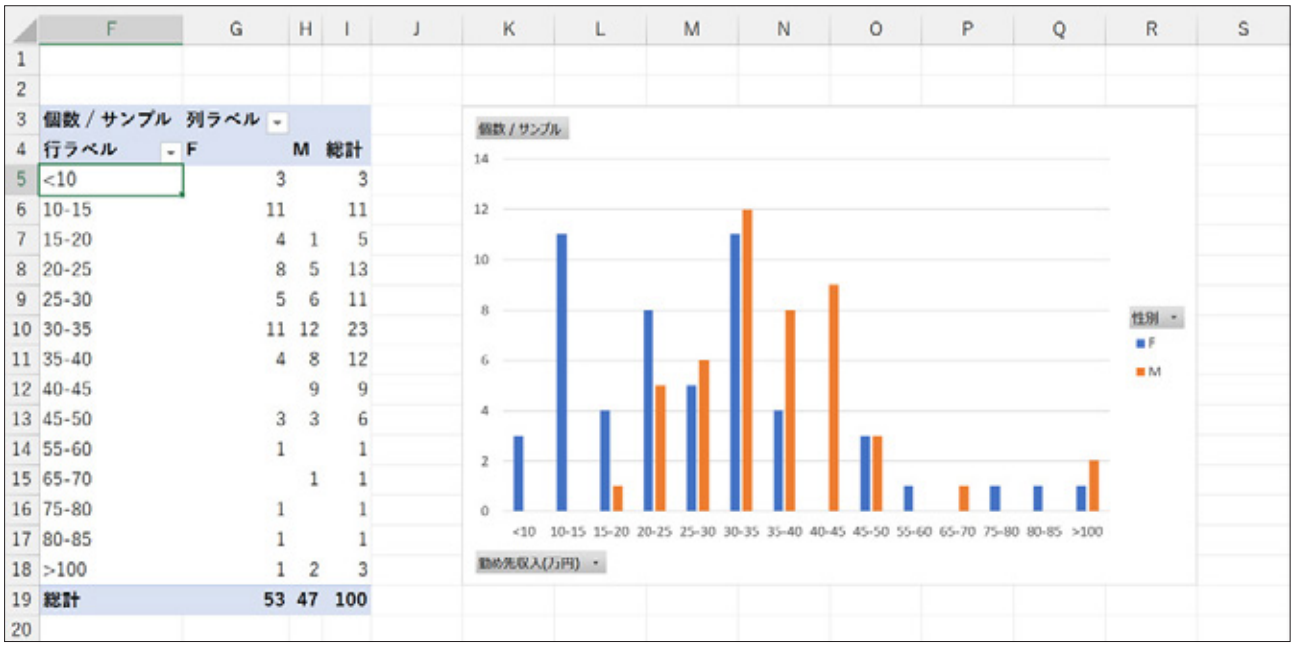
図 7　性別ごとに勤め先収入の人数を集計する

勤め先収入が縦方向（F 列）に、性別が横方向（4 行目）に並ぶようにして、それぞれの人数を集計する。［Σ値］に指定した項目は、そのままだと合計が求められる。人数、つまりデータの個数を求めたいので、個数を求めるように集計の方法を変えておく。

これではヒストグラムにならないので、勤め先収入を階級に区切って人数が集計されるようにしましょう。

- F 列の値（どれでもいい）を右クリックして［グループ化］を選択する
- ［グループ化］ダイアログボックスで［先頭の値］に「10」を入力、［末尾の値］に「100」を入力、［単位］に「5」を入力する

これで、勤め先収入が階級に区切られ、性別ごとのヒストグラムが作成できました（図 8）。



個数 / サンプル	列ラベル		
行ラベル	F	M	総計
<10	3		3
10-15	11		11
15-20	4	1	5
20-25	8	5	13
25-30	5	6	11
30-35	11	12	23
35-40	4	8	12
40-45		9	9
45-50	3	3	6
55-60	1		1
65-70		1	1
75-80	1		1
80-85	1		1
>100	1	2	3
総計	53	47	100

図 8　勤め先収入をグループ化し、階級を作る

10 万円未満の人数と **100 万円**より大きい人数がまとめられた。各階級は「以上〜未満」となる。例えば **10** 〜 **15** という階級は **10 万円**以上 **15 万円**未満の人数。ただし、［末尾の値］の直前の階級のみ「以上〜以下」となる。この例では、**95** 〜 **100** という階級が **0 人**なのでグラフには表示されていないが、**95** 〜 **100** の階級だけ、**95 万円**以上 **100 万円**未満の人数となる。

このままでも分布の特徴は分かりますが、ヒストグラムらしくするために、棒の間隔を「0」にしておきましょう。

- 系列（グラフの棒の部分）を右クリックして［データ系列の書式設定］を選択する
- ［系列のオプション］ボタンをクリックする
- ［系列の重なり］に「0」（%）を入力する
- ［要素の間隔］に「0」（%）を入力する

グラフは以下のようになります（図 9）。

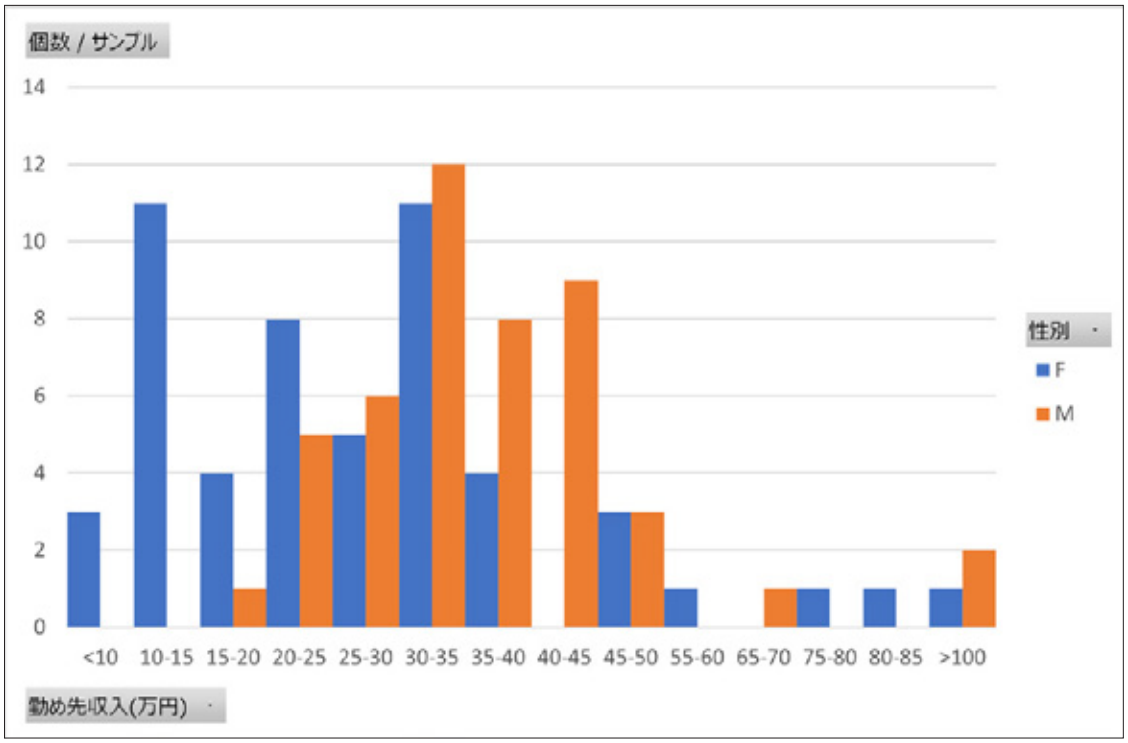


図 9　性別ごとに勤め先収入のヒストグラムを作成した例（完成例）

全体的に女性の収入が低いことが分かる。男性のヒストグラム（オレンジ色）は **30 万円**～ **35 万円**の山が最も高くなっている。女性のヒストグラム（青色）は **30 万円**～ **35 万円**の山と **10 万円**～ **15 万円**の山が最も高い。

性別によるヒストグラムを見ると、明らかに男女間で収入の格差があることが分かります。特に、女性で **10 万円**～ **15 万円**のところに山ができていることも分かりますね。もちろん、このデータは架空のデータなので、実態を表しているわけではありませんが、実態としては、2021 年の男性の収入を **100** とすると、女性の収入は **75.2** となっています（内閣府男女共同参画局の「男女間賃金格差（我が国の現状）」による。一般労働者の給与水準）。ちなみに、サンプルデータの中央値は男性が **34.9 万円**、女性が **25.9 万円**で、ほぼその割合（34.9/25.9 = **0.74**）になるようにしてあります。

ピボットテーブル／ピボットグラフでは、項目（フィールド）や集計項目をドラッグ操作で自由に入れ替えられるので、さまざまな方向からの分析ができます。以下に、幾つかの例を示しておきます（図 10、図 11）。［データ系列の書式設定］ウィンドウの右上に表示されている［×］ボタンをクリックすれば、［データ系列の書式設定］ウィンドウが閉じられ、［ピボットグラフのフィールド］ウィンドウが表示されます。ぜひ、試してみてください。

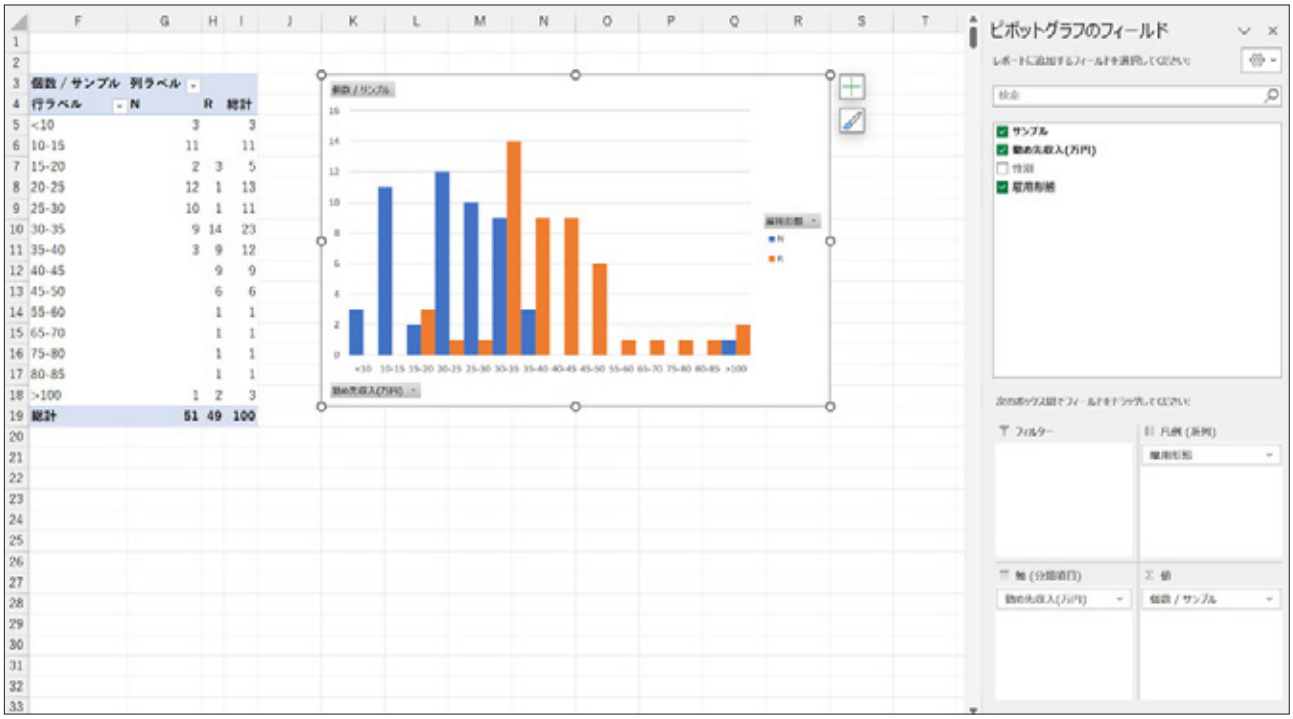


図 10　雇用形態ごとに勤め先収入のヒストグラムを作成した例

［ピボットテーブルのフィールド］に表示されている［性別］のチェックマークをオフにし、［雇用形態］を下の［凡例（系列）］の欄にドラッグすると、雇用形態ごとのヒストグラムが作成できる。全体的に非正規職員の収入が低いことが見て取れる。

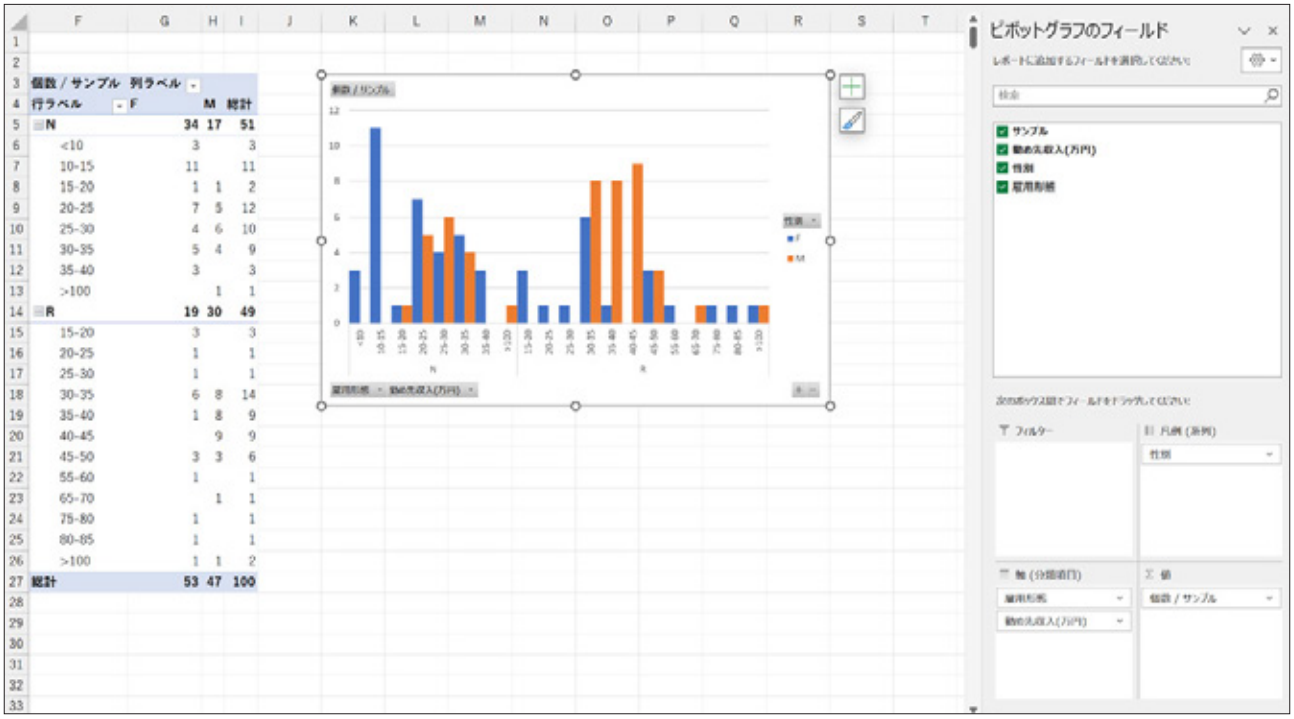


図 11　男女ごと、雇用形態ごとに勤め先収入のヒストグラムを作成した例

［ピボットテーブルのフィールド］に表示されている［性別］を下の［凡例（系列）］の欄にドラッグし、［雇用形態］を［軸（分類項目）］の欄の［勤め先収入（万円）］の上にドラッグすると、男女ごと、雇用形態ごとのヒストグラムができる。女性の非正規職員に **10 万円～ 15 万円**の大きな山があることが分かる。

→目次に戻る

Google スプレッドシートでの操作手順

Google スプレッドシートでの操作は以下の通りです。こちらのサンプルファイルを開いて、メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。［勤め先収入 3］ワークシートにデータが入力されています。

- セル **A3** ～ **D103** のいずれかのセルをクリックしておく
- メニューバーから［挿入］－［ピボットテーブル］を選択する
- ［ピボットテーブルの作成］ダイアログボックスの［データ範囲］が「勤め先収入 3!A3:D103」になっていることを確認する
- ［挿入先］の下の［既存のワークシート］をクリックしてオンにする
- ［データ範囲を選択］ボタン（田のマークのボタン）をクリックして、セル **F3** をクリックし、［OK］をクリックする
- ［作成］ボタンをクリックする

これで、空のピボットテーブルが作成されます。画面の右側にピボットテーブルエディタが表示されるので、以下のように操作しましょう。

- 右側に表示されている項目一覧の［勤め先収入（万円）］を、左側の［行］の下にドラッグする
- ［総計を表示］のチェックマークをクリックしてオフにする
- 右側に表示されている項目一覧の［性別］を、左側の［列］の下にドラッグする
- ［総計を表示］のチェックマークをクリックしてオフにする
- 右側に表示されている項目一覧の［サンプル］を、左側の［値］の下にドラッグする
- ［集計］ボックスからの［COUNT］を選択する

集計が行われますが、図 7 で見たような勤め先収入に対する人数（F 列）が、1 つずつ表示された表になります。そこで、勤め先収入の階級を設定します。

- 作成されたピボットテーブルの［勤め先収入（万円）］の値（F 列の値ならどれでもいい）を右クリックして［ピボットグループのルールを作成］を選択する
- ［グループ化のルール］ダイアログボックスで、［最小値］に「10」、［最大値］に「100」、［間隔のサイズ］に「5」を入力し、［OK］ボタンをクリックする

これで、図 8 で見たようなピボットテーブルが作成されます。後は縦棒グラフを作成するだけです。

- 作成されたピボットテーブル（セル **F3** ～ **H18**）のいずれかのセルをクリックする
- メニューバーから［挿入］－［グラフ］を選択する
- グラフエディタで［グラフの種類］のリストから［縦棒グラフ］を選択する

ピボットテーブルエディタで項目を入れ替えると、グラフも自動的に表示し直されるので、雇用形態別のヒストグラムなども簡単に作成できます。ただし、図 11 と同様の、男女ごと、雇用形態ごとのヒストグラムにするには、以下のような操作を行います。なお、ピボットテーブルエディタが表示されていない場合には、セル **F3** ～ **H18** に作成されているピボットテーブルの左下にマウスポインタを位置付け、［編集］ボタン（鉛筆のアイコンのボタン）をクリックしてください。

- ピボットテーブルエディタで、［雇用形態］を、左側の［行］の下、［勤め先収入（万円）］の上にドラッグする

ただし、この段階では、女性だけのヒストグラムしか表示されないので、グラフエディタで系列を追加する必要があります。

- グラフを右クリックし［データ範囲］を選択する
- グラフエディタで［系列を追加］ボタンをクリックし、［データ範囲の選択］ボタン（田のマークのボタン）をクリックして、セル **I4** ～ **I24** をドラッグする
- ［OK］ボタンをクリックする

ピボットテーブルエディタでの設定とグラフエディタでの設定が正しくできれば、図 12 のようなグラフになります。

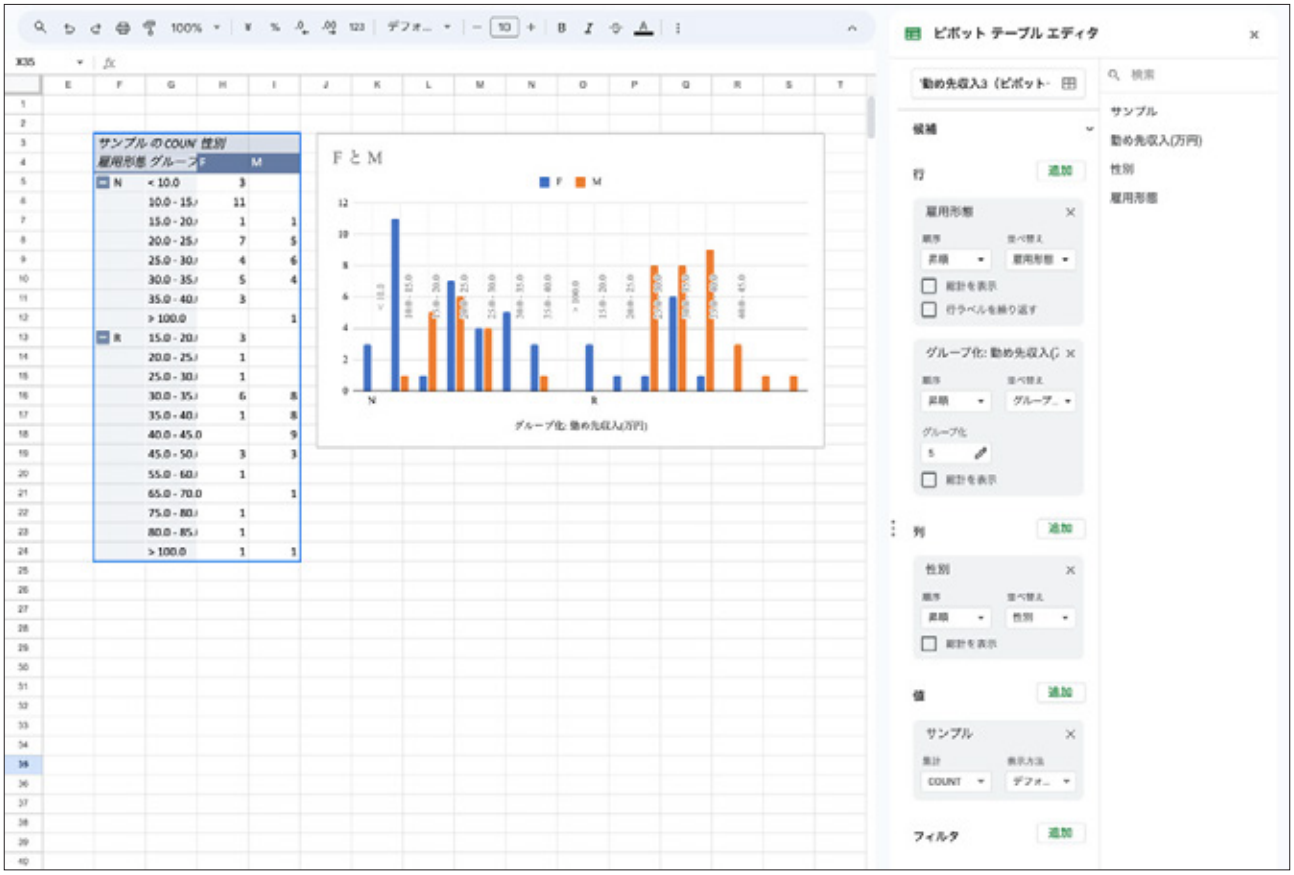


図 12　男女ごと、雇用形態ごとに勤め先収入のヒストグラム（Google スプレッドシート）

Excel で作成した図 11 の例と同じグラフ。Google スプレッドシートでは、棒の間隔を変えることができないので、やや見た目は異なるが、分析を行う上での支障はない。

今回は、ヒストグラムや箱ひげ図により、集団の特徴を可視化する方法を見てきました。まず、全体像を見た上で、データを切り分けていくことにより、外れ値を発見したり、集団の中に含まれる幾つかの特徴や「層」を見いだしたりする流れを紹介しました。

次回は、クロス集計表を作成したり、ヒートマップを作成することにより、複数の項目同士がどう関係しているか、その関係の中からどのような特徴が見いだせるかを、ケーススタディーを通して追いかけていきます。次回もどうぞお楽しみに！

［データ分析］クロス集計表やヒートマップで「分布」を多角的に可視化～項目同士の関連を見つける

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 11 回。グラフを利用して分布や項目同士の関係を多角的に可視化します。ピボットテーブルの詳細な取り扱いとヒートマップによる視覚的な分析について、ケーススタディを通して学びましょう。

羽山博（2023 年 11 月 16 日）

前回はヒストグラムと箱ひげ図を利用して分布を可視化し、集団の特徴を見極めるというお話をしました。最後に、男女別や雇用形態別に勤め先収入のヒストグラムを作成する例を紹介しましたが、今回はいわばその続きです。具体的にはピボットテーブルを利用してクロス集計表を作成し、さらにヒートマップにより値の大きな箇所を可視化します。

前回見た性別や雇用形態の例では、各項目の内容はそれほど多くない（男性／女性、正規／非正規など）ので、1 つのグラフの中に複数のヒストグラムを同時に描いてもそれほど見づらくはありませんでした。しかし、図 1 のような場合だとどうでしょう。このデータは 1000 種類のワインについて、価格と口コミの評価を一覧にしたものです（架空のデータです）。現実のデータ処理ではこの程度のデータ量は「ざら」にありますし、価格や評価など、項目の値が多くの階級に分けられることもあります。そのままヒストグラムを作っても図 1 に示したグラフのように、訳の分からないものになってしまいます。

今回のテーマは、クロス集計表を利用した関係の可視化です。また、後半ではグループの可視化についても紹介します。……といっても、どのような分析を行うのかまだイメージが湧きませんよね。図 1 のグラフは見づらいので、以降は使いませんが、一部分だけ取り出してみると、以下のような値がグラフ化されていることが分かります。

- **2000** 円以上 **3000** 円未満のワインでは、**3.5** 以上 **4.0** 未満の点数が付いた商品が一番多く、104 種類ある
- **10000** 円以上 **11000** 円未満のワインでは **4.5** 以上 **5.0** 以下の点数が付いた商品が一番多いが、8 種類だけ（最大値が **5.0** なので最後の階級だけは「以下」となる）

今回は、図 1 に関連して述べた問題（見づらいグラフ）を解決し、箇条書きで示したような分析をもっと簡単にできるようにすることを目指します。そのための一つの方法として、ピボットテーブルを利用してクロス集計表を作成し、ヒートマップとして色分けして表す方法を見ていこうというわけです。それにより、分布や項目同士の関係を見やすくします。見やすい可視化ができれば、何か面白い発見があるかもしれません。

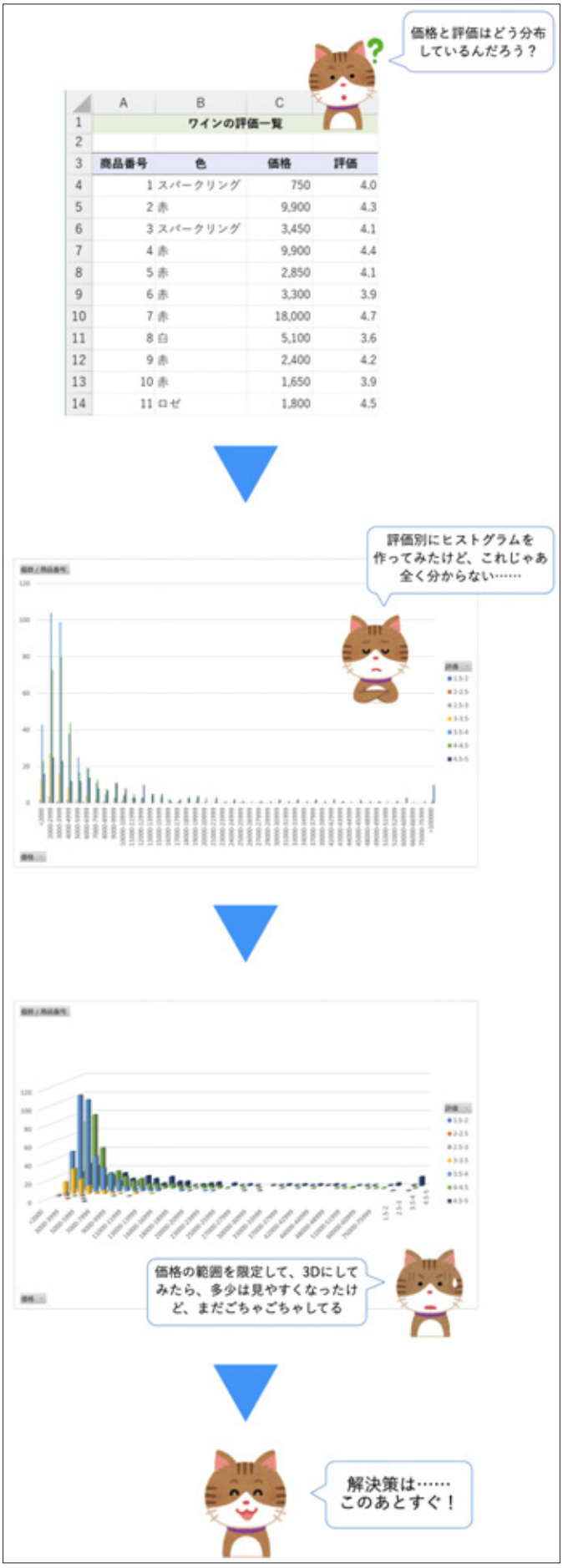


図 1　項目の内容が細かく分かれる場合でも、関係や分布を可視化したい

ワインの価格と評価の関係を可視化したいのだが、グラフ化する系列が多くなると、それらを同時に可視化することにはかなり無理がある。なお、データはセル **A4** ～ **A1003** に 1000 件入力されている。評価の値はその商品の評価の平均値。

→目次に戻る

この記事は、データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 11 回です。第 7 回の棒グラフ、第 8 回の折れ線グラフ、第 9 回の円グラフ／パレート図、第 10 回のヒストグラム／箱ひげ図、今回のクロス集計表、ヒートマップ、第 12 回の散布図まで、1 つずつ可視化の基礎を学んでいきます。これらのグラフの目的と効用などについて、特別予告編で簡単に整理していますので、事前に確認しておくとより理解が深まるでしょう。

この記事で学べること

今回は以下のようなポイントについて、分析の方法や目の付けどころを見ていきます。

- ピボットテーブルを利用した多角的な分析 …… グループ化や計算の方法を指定し、クロス集計表を作成する
- ヒートマップを利用した可視化 …… 値の集中している箇所や異なるグループの可視化

では、ピボットテーブルを利用したクロス集計表の作成から見ていきましょう。続いて、条件付き書式を使ってヒートマップを作成します。サンプルファイルの利用についての説明の後、本編に進みましょう。

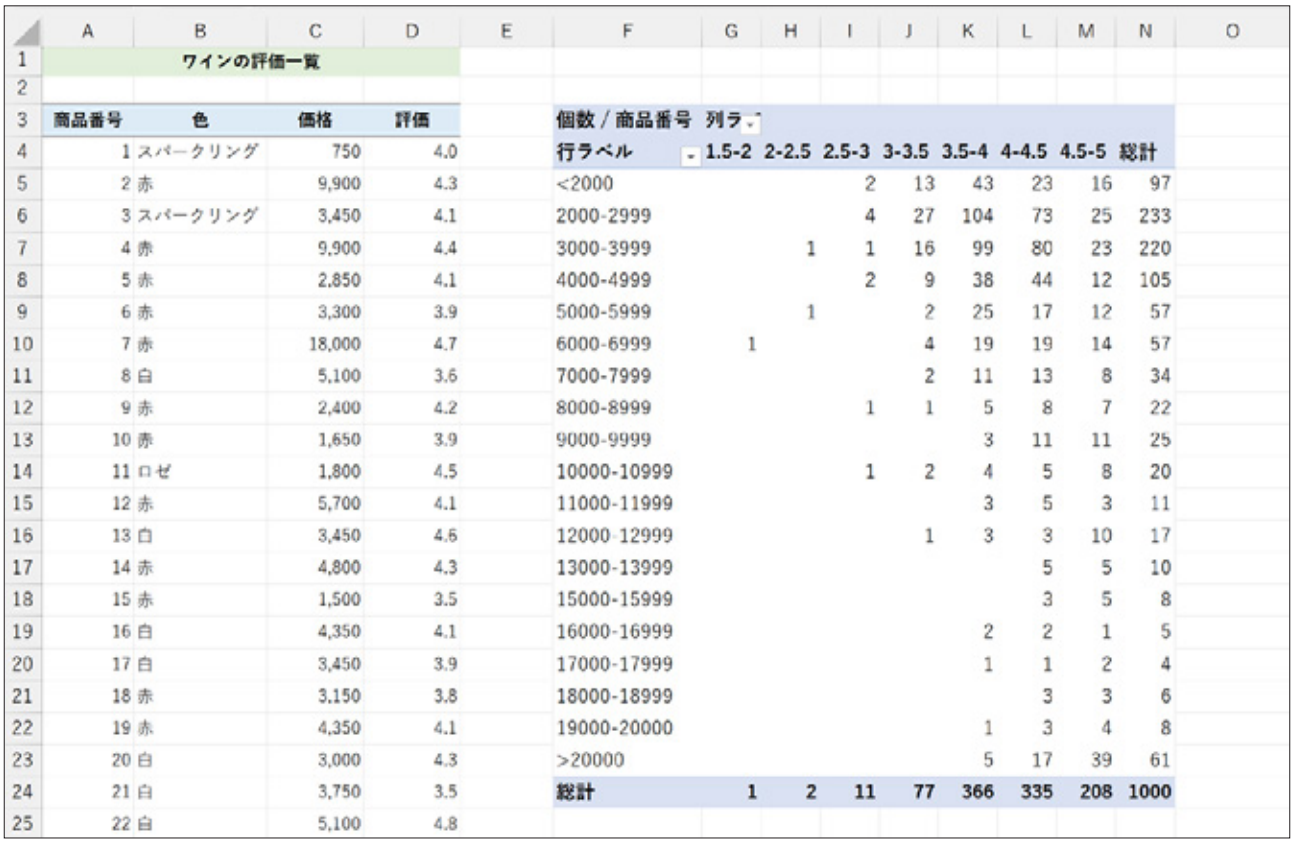
サンプルファイルの利用について

本稿では、表計算ソフトを使って手を動かしながら学んでいきます。表計算ソフト Microsoft Excel 用の .xlsx ファイルをダウンロードできるようにしています。デスクトップ版の Excel が手元にない場合は、Microsoft アカウントがあれば使える無料の Microsoft 365 オンライン、もしくは Google アカウントがあれば使える無料の Google スプレッドシート（Google Sheets）をお使いください。Microsoft 365 オンラインの場合は、.xlsx ファイルを OneDrive にアップロードしてから開いてください。Google スプレッドシートの場合は、.xlsx ファイルを Google ドライブにアップロードしてから開いた上で［ファイル］メニューの［Google スプレッドシートとして保存］を実行してください（Google スプレッドシート独自の機能を使っている場合は、ファイルを共有して参照できるようにします。その場合は、該当する箇所で使い方を記します）。

ピボットテーブルを利用してクロス集計表を作る 〜 関係を可視化するためのデータを作成

クロス集計表とは、行と列に項目が並んでいて、その交わった位置に値がある表のことです。実は、前回のお話の中でも既にクロス集計表を作成しています。従って、図 2 のようなクロス集計表も簡単に作成できると思います。サンプルファイルをこちらからダウンロードし、［ワインの価格と評価］ワークシートを開いて取り組んでみてください。Google スプレッドシートの場合はこちらのサンプルファイルを開いて、メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。

手順は図の後に箇条書きで示しておきます。ただし、タイトルや列幅など、データ分析そのものにあまり関係のない設定については省略してあります。なお、動画でも手順を解説しているので、操作を一つ一つ追いかけたい方はぜひご視聴下さい。



ワインの評価一覧

商品番号	色	価格	評価
1	スパークリング	750	4.0
2	赤	9,900	4.3
3	スパークリング	3,450	4.1
4	赤	9,900	4.4
5	赤	2,850	4.1
6	赤	3,300	3.9
7	赤	18,000	4.7
8	白	5,100	3.6
9	赤	2,400	4.2
10	赤	1,650	3.9
11	ロゼ	1,800	4.5
12	赤	5,700	4.1
13	白	3,450	4.6
14	赤	4,800	4.3
15	赤	1,500	3.5
16	白	4,350	4.1
17	白	3,450	3.9
18	赤	3,150	3.8
19	赤	4,350	4.1
20	白	3,000	4.3
21	白	3,750	3.5
22	白	5,100	4.8

個数 / 商品番号	列ラ							
行ラベル	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5	総計
<2000			2	13	43	23	16	97
2000-2999			4	27	104	73	25	233
3000-3999		1	1	16	99	80	23	220
4000-4999			2	9	38	44	12	105
5000-5999		1		2	25	17	12	57
6000-6999	1			4	19	19	14	57
7000-7999				2	11	13	8	34
8000-8999			1	1	5	8	7	22
9000-9999					3	11	11	25
10000-10999			1	2	4	5	8	20
11000-11999					3	5	3	11
12000-12999				1	3	3	10	17
13000-13999						5	5	10
15000-15999						3	5	8
16000-16999					2	2	1	5
17000-17999					1	1	2	4
18000-18999						3	3	6
19000-20000					1	3	4	8
>20000					5	17	39	61
総計	1	2	11	77	366	335	208	1000

図 2　ワインの価格と評価のクロス集計表

前回の知識だけでこの表は作成できる。価格だけでなく評価もグループ化して階級に分けてあることに注目。例えば、**2000 円**未満のワインでは、評価が **2.5** 以上 **3** 未満の商品が **2** つあり、**3** 以上 **3.5** 未満の商品が **13** 種類ある……といったように、表から値が読み取れる。図 1 の集合縦棒グラフや 3D 縦棒グラフはこのクロス集計表を基に作成したもの。

以下の手順で進めていきましょう。Google スプレッドシートでの操作手順は後でまとめておきます（Microsoft 365 オンラインでは、ピボットテーブルの作成はできますが、後述の［グループ化］ができないため、説明を割愛します）。

Excel での操作手順

- セル **A3** ～ **D1003** のいずれかのセルをクリックしておく
- ［挿入］タブを開き、［ピボットテーブル］ボタンをクリックする（［ピボットテーブルの作成］ダイアログボックスが表示される。ただし、Excel のバージョンによってはダイアログボックスの名前が［テーブルまたは範囲からのピボットテーブル］となっている）
- ［テーブル／範囲］が「ワインの価格と評価 !$A$3:$D$1003」になっていることを確認する
- ［ピボットテーブルレポートを作成する場所を選択してください］の下の［既存のワークシート］をクリックしてオンにする
- ［場所］ボックスをクリックし、セル **F3** をクリックして、［OK] ボタンをクリックする

これで、空のピボットテーブルと空のピボットグラフが作成されます。画面は図 3 のようになります。

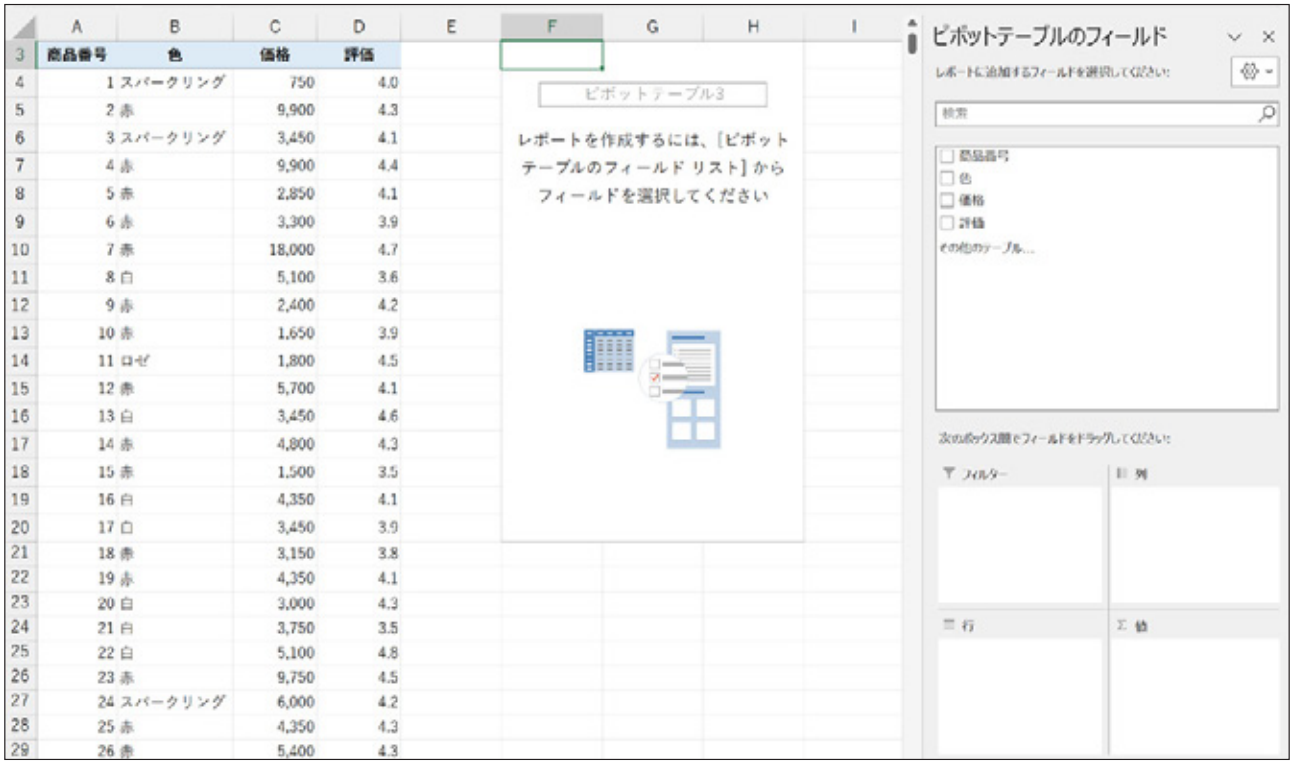


図 3　空のピボットテーブルと空のピボットグラフが作成された

この段階では、どの項目のどの値を集計し、グラフ化するかが指定されていないので、ピボットテーブルもピボットグラフも空の状態になっている。ここから、集計項目や集計の方法などを指定していく。

価格ごと、評価ごとにデータの件数を集計しましょう。

- ［ピボットテーブルのフィールド］のリストにある［価格］を［行］の欄にドラッグする
- ［ピボットテーブルのフィールド］のリストにある［評価］を［列］の欄にドラッグする
- ［ピボットテーブルのフィールド］のリストにある［商品番号］を［Σ値］の欄にドラッグする
- ［Σ値］の欄の「合計／商品番号」（Mac 版の場合は［i］のアイコン）をクリックし、［値フィールドの設定］を選択する
- ［選択したフィールドのデータ］リストから［個数］を選択し、［OK］をクリックする

この段階では、価格が細かく表示されており（F 列）、評価も細かく表示されています（4 行目）。表の中にはそれぞれの個数が表示されています。例えば、**1200 円**で平均評価が 3.0 の商品が 1 つ、**1500 円**で平均評価が 3.4 の商品は 2 つあるというわけです（図 4）。

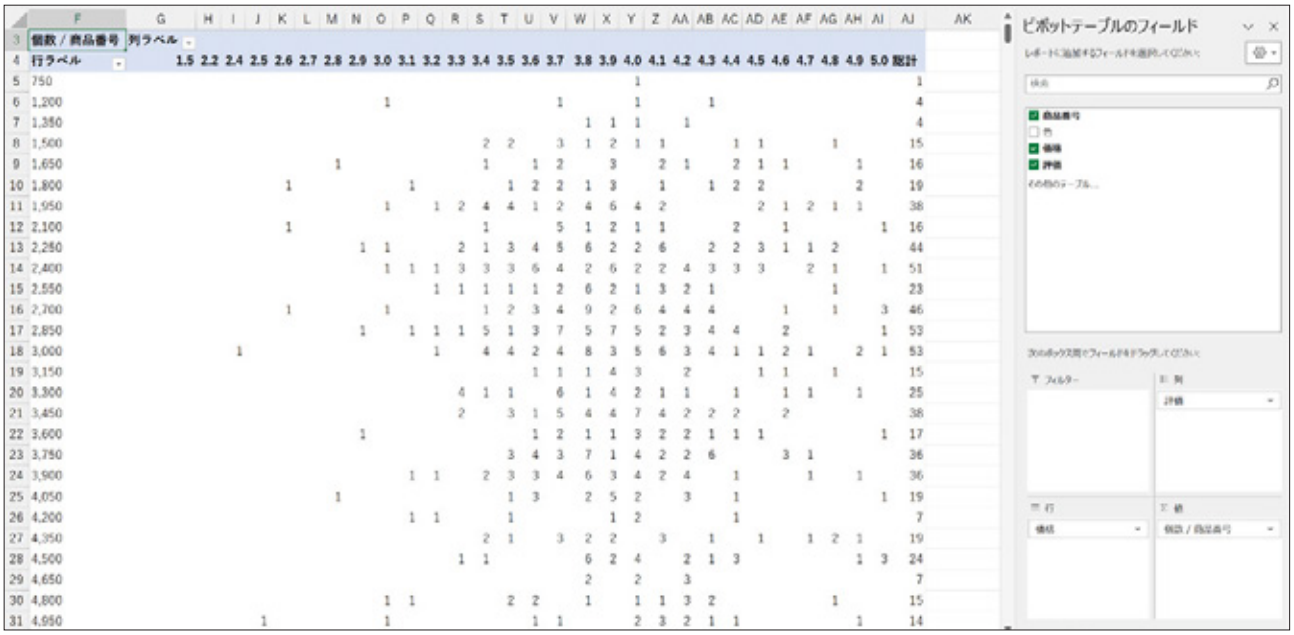


図 4　価格と評価のクロス集計表（作成途中）

価格が縦方向（F 列）に、評価が横方向（4 行目）に並ぶようにして、それぞれの件数を集計する。［Σ値］に指定した項目は、そのままだと合計が求められる。商品の個数を求めたいので、個数を求めるように集計の方法を変えておく。

図 4 の行と列はあまりにも細かく分かれているので、価格と評価の関係がよく分かりません。そこで、価格は **1000 円**ごとに、評価は **0.5 点**ごとに区切ることにしましょう。かなり高額なワインも幾つかあるので、**20000 円**より高い商品はひとまとめにします。

- セル **F5** ～ **F147** のいずれかを右クリックして［グループ化］を選択する
- ［グループ化］ダイアログボックスで［先頭の値］に「2000」を入力、［末尾の値］に「20000」を入力、［単位］に「1000」を入力する
- セル **G4** ～ **AI4** のいずれかを右クリックして［グループ化］を選択する
- ［グループ化］ダイアログボックスで［先頭の値］に「1」を入力、［末尾の値］に「5」を入力、［単位］に「0.5」を入力する

これで、価格と評価が階級に区切られ、その価格帯で、ある評価を得た商品の件数が求められます（図 2 で見たものです）。この段階まで操作を進めた結果は［ワインの価格と評価（ピボットテーブル）］ワークシートに作成してあります。

Google スプレッドシートでの操作手順

- セル **A3** ～ **D1003** のいずれかのセルをクリックしておく
- メニューバーから［挿入］－［ピボットテーブル］を選択する
- ［ピボットテーブルの作成］ダイアログボックスの［データ範囲］が「ワインの価格と評価 !**A3**:**D1003**」になっていることを確認する
- ［挿入先］の下の［既存のワークシート］をクリックしてオンにする
- ［データ範囲を選択］ボタン（田のマークのボタン）をクリックして、セル **F3** をクリックし、［OK］をクリックする
- ［作成］ボタンをクリックする

これで、空のピボットテーブルが作成されます。画面の右側にピボットテーブルエディタが表示されるので、以下のように操作しましょう。

- 右側に表示されている項目一覧の［価格］を、左側の［行］の下にドラッグする
- 右側に表示されている項目一覧の［評価］を、左側の［列］の下にドラッグする
- 右側に表示されている項目一覧の［商品番号］を、左側の［値］の下にドラッグする
- ［集計］ボックスから［COUNT］を選択する

価格と評価の両方に階級を設定しましょう。

- 作成されたピボットテーブルの［価格］の値（F 列の値ならどれでもいい）を右クリックして［ピボットグループのルールを作成］を選択する
  - ・［グループ化のルール］ダイアログボックスで、［最小値］に「2000」、［最大値］に「20000」、［間隔のサイズ］に「1000」を入力し、［OK］ボタンをクリックする
- 作成されたピボットテーブルの［評価］の値（4 行目の値ならどれでもいい）を右クリックして［ピボットグループのルールを作成］を選択する
  - ・［グループ化のルール］ダイアログボックスで、［最小値］に「1」、［最大値］に「5」、［間隔のサイズ］に「0.5」を入力し、［OK］ボタンをクリックする

これでクロス集計表が作成できました。

ヒートマップにより頻度を色分けする ～ 値の大小や関係を可視化

最初に示した図 1 の見づらいグラフは、前項で作成したピボットテーブルの値を棒グラフにしたものです。価格 × 評価という 2 次元の軸があり、頻度を高さで表しているので、かなり込み入った図になってしまったというわけです。そこで、ヒートマップを使い、頻度を高さではなく色で表すことにします。ここでは、頻度の多いセルを赤で、頻度の少ないセルを白で表示することにしましょう。ヒートマップはグラフの機能ではなく、条件付き書式で指定します。

- セル **G5** ～ **M23** をドラッグして選択する（右端の［総計］列を含めないように注意しましょう）
- ［ホーム］タブを開く
- ［条件付き書式］－［カラー スケール］－［赤、白のカラー スケール］を選択する

Google スプレッドシートでは以下のように操作します。

- セル **G5** ～ **M23** をドラッグして選択する（右端の［総計］列を含めないように注意しましょう）
- メニューバーから［表示形式］－［条件付き書式］を選択し、右側に表示される［条件付き書式設定ルール］作業ウインドウで［カラースケール］をクリックする
- ［プレビュー］の色が表示されている部分をクリックして、［白→赤］を選択する
- ［完了］ボタンをクリックする
- ［条件付き書式設定ルール］作業ウインドウの右上の［×］をクリックして、作業ウィンドウを閉じておく

条件付き書式を指定するとピボットテーブルは以下のようになります（図 5）。

	F	G	H	I	J	K	L	M	N	O
1										
2										
3	個数 / 商品番号	列ラベル								
4	行ラベル	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5	総計	
5	<2000			2	13	43	23	16	97	
6	2000-2999			4	27	104	73	25	233	
7	3000-3999		1	1	16	99	80	23	220	
8	4000-4999			2	9	38	44	12	105	
9	5000-5999		1		2	25	17	12	57	
10	6000-6999	1			4	19	19	14	57	
11	7000-7999				2	11	13	8	34	
12	8000-8999			1	1	5	8	7	22	
13	9000-9999					3	11	11	25	
14	10000-10999			1	2	4	5	8	20	
15	11000-11999					3	5	3	11	
16	12000-12999				1	3	3	10	17	
17	13000-13999						5	5	10	
18	15000-15999						3	5	8	
19	16000-16999					2	2	1	5	
20	17000-17999					1	1	2	4	
21	18000-18999						3	3	6	
22	19000-20000					1	3	4	8	
23	>20000					5	17	39	61	
24	総計	1	2	11	77	366	335	208	1000	
25										

図 5　ヒートマップの作成例

どのセルの値が大きいかが一目で分かる。濃い赤のセルが頻度の大きなセル。ヒートマップは、いわば図 1 の 3D グラフを上から覗き込んで、高い棒と低い棒を色分けしたようなものとも考えられる。この例は［ワインの価格と評価（ヒートマップ）］ワークシートに作成してある。

図 9 のヒートマップを見ると、以下のようなことが一目で見て取れます。

- 全体的に価格の高いワインは、評価の高い商品が多い
- **2000 円**以上 **4000 円**未満のワインで、**3.5** 以上 **4.5** 未満の点数が付いた商品が多い

つまり、満足度の高いワインが欲しければある程度の予算が必要だということと、（口コミの数は売り上げを反映していると考えられるので）どの価格帯の商品が入手しやすく、その評価はどれぐらいかということが分かります。また、以下のようなことも言えそうです。

- 安いワインでも低評価はそれほど多くない
- **5000 円**以上 **10000 円**未満のワインには評価の低いものが若干ある
- **20000 円**以上のワインは評価が高い

最初のは「安いワインだからまあこんなもんだろう（値段の割にはいい）」、次は「そこそこの値段だったのに、期待したほどではなかった」という評価の商品があったのかもしれません。最後は、実際に出来の良いワインで、それだけの商品を購入する人はやはり舌が肥えているということでしょうか。意地悪な見方をすれば「高いワインだから、きっといいものだろう」と無意識のうちに考える心理的な要因も考えられなくもないですが。

カラースケールには赤、黄、緑など、もっと派手なものもありますが、1 型 2 色覚（P）の人や 2 型 2 色覚（D）の人にとっては赤と緑の区別が付きづらいので、図 5 のような配色が望ましいでしょう。どうしても見づらい配色にせざるを得ない場合には、値やラベルなどを表示して区別できるようにしておきましょう（この例では、値が表示されてはいますが、色分けに重きを置いているので赤と白の配色にしてあります）。筆者は「色のシミュレータ」（iPad/iPhone 用はこちら、Android 用はこちら）というアプリを使って、見づらい配色がないかチェックしています。

ピボットテーブルの計算方法を変える ～ 値ではなく割合を可視化する

ところで、図 5 のピボットテーブル／ヒートマップを見て、疑問に思う点はないでしょうか。例えば、**2000 円**以上 **3000 円**未満のワインで、**3.5** 以上 **4.0** 未満の点数が付いた商品は 104 種類あります（図 **6** の上、○で囲んだ箇所）。一方、**20000 円**より高いワインでは、**4.5** 以上 **5.0** 以下の点数が付いた商品が 39 種類あります（図 6 の右下、○で囲んだ箇所）。

	F	G	H	I	J	K	L	M	N	O
1										
2										
3	個数 / 商品番号	列ラベル								
4	行ラベル	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5	総計	
5	<2000			2	13	43	23	16	97	
6	2000-2999			4	27	104	73	25	233	
7	3000-3999		1	1	16	99	80	23	220	
8	4000-4999			2	9	38	44	12	105	
9	5000-5999		1		2	25	17	12	57	
10	6000-6999	1			4	19	19	14	57	
11	7000-7999				2	11	13	8	34	
12	8000-8999			1	1	5	8	7	22	
13	9000-9999					3	11	11	25	
14	10000-10999			1	2	4	5	8	20	
15	11000-11999					3	5	3	11	
16	12000-12999				1	3	3	10	17	
17	13000-13999						5	5	10	
18	15000-15999						3	5	8	
19	16000-16999					2	2	1	5	
20	17000-17999					1	1	2	4	
21	18000-18999						3	3	6	
22	19000-20000					1	3	4	8	
23	>20000					5	17	39	61	
24	総計	1	2	11	77	366	335	208	1000	
25										

図 6　単純に値を色分けしただけだと割合の大きさが分からない

104 の方が 39 よりも濃く表示されているので、その部分のワインが多いことが分かるが、評価を問題にするのであれば、高額なワインが高評価であることが分かりづらい。

確かに数の上では 104 の方が多いので、**2000 円**以上 **3000 円**未満で **3.5** 以上 **4.0** 未満程度のワインが多く、入手しやすいということは分かります。しかし、評価の良さはこのままでは比較できません。**2000 円**以上 **3000 円**未満のワインは全部で 233 種類あるので、104 というのは全体の **44.6%**です。**20000 円**より高いワインは 61 種類あるので、そのうちの 39 種類は全体の **63.9%**です。高価なワインの方が断然高評価なのに、ヒートマップによる色分けではそれほど目立っていませんね。

そこで、ピボットテーブルの集計方法として、単にデータの件数を求めるのではなく、同じ価格帯の商品全体に占める件数の割合を求めるように変更してみましょう。そのためには、行の総計に対する各セルの割合を求めた表を別に作って……と、面倒な操作が必要になりそうだと思われるかもしれません。しかし、ピボットテーブルではそのような計算も簡単にできます。

- ［Σ値］の欄の「合計／商品番号」（Mac 版の場合は［i］のアイコン）をクリックし、［値フィールドの設定］を選択する
- ［値フィールドの設定］ダイアログボックスで［計算の種類］タブをクリックする
- ［計算の種類］リストから［行集計に対する割合］を選択し、［OK］をクリックする
- Google スプレッドシートでは、［ピボットテーブルエディタ］作業ウィンドウの左側、［値］の下に表示されている［商品番号］の［表示方法］リストから［行集計に対する割合］を選択する

設定ができると、ピボットテーブル／ヒートマップは図 7 のようになります。

	F	G	H	I	J	K	L	M	N
1									
2									
3	個数 / 商品番号	列ラベル							
4	行ラベル	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5	総計
5	<2000	0.0%	0.0%	2.1%	13.4%	44.3%	23.7%	16.5%	100.0%
6	2000-2999	0.0%	0.0%	1.7%	11.6%	44.6%	31.3%	10.7%	100.0%
7	3000-3999	0.0%	0.5%	0.5%	7.3%	45.0%	36.4%	10.5%	100.0%
8	4000-4999	0.0%	0.0%	1.9%	8.6%	36.2%	41.9%	11.4%	100.0%
9	5000-5999	0.0%	1.8%	0.0%	3.5%	43.9%	29.8%	21.1%	100.0%
10	6000-6999	1.8%	0.0%	0.0%	7.0%	33.3%	33.3%	24.6%	100.0%
11	7000-7999	0.0%	0.0%	0.0%	5.9%	32.4%	38.2%	23.5%	100.0%
12	8000-8999	0.0%	0.0%	4.5%	4.5%	22.7%	36.4%	31.8%	100.0%
13	9000-9999	0.0%	0.0%	0.0%	0.0%	12.0%	44.0%	44.0%	100.0%
14	10000-10999	0.0%	0.0%	5.0%	10.0%	20.0%	25.0%	40.0%	100.0%
15	11000-11999	0.0%	0.0%	0.0%	0.0%	27.3%	45.5%	27.3%	100.0%
16	12000-12999	0.0%	0.0%	0.0%	5.9%	17.6%	17.6%	58.8%	100.0%
17	13000-13999	0.0%	0.0%	0.0%	0.0%	0.0%	50.0%	50.0%	100.0%
18	15000-15999	0.0%	0.0%	0.0%	0.0%	0.0%	37.5%	62.5%	100.0%
19	16000-16999	0.0%	0.0%	0.0%	0.0%	40.0%	40.0%	20.0%	100.0%
20	17000-17999	0.0%	0.0%	0.0%	0.0%	25.0%	25.0%	50.0%	100.0%
21	18000-18999	0.0%	0.0%	0.0%	0.0%	0.0%	50.0%	50.0%	100.0%
22	19000-20000	0.0%	0.0%	0.0%	0.0%	12.5%	37.5%	50.0%	100.0%
23	>20000	0.0%	0.0%	0.0%	0.0%	8.2%	27.9%	63.9%	100.0%
24	総計	0.1%	0.2%	1.1%	7.7%	36.6%	33.5%	20.8%	100.0%
25									

図 7　割合に注目したピボットテーブル／ヒートマップを作成する

高額なワインはデータの件数（口コミの数）が少ないので、正確さにはやや欠けるが、値段が高くなるほど評価の高い商品が多くなっていることがより鮮明に分かる。この例は［ワインの価格と評価（比率のヒートマップ）］ワークシートに作成してある。

ヒートマップを作成したときに、大きな値（あるいは小さな値）の集まっている箇所に偏りがあることが分かれば、項目同士に何らかの関係があることが分かります。図 7 の例であれば、右下の値がやや大きいので、価格が高くなるほど、評価も高くなることが分かります。逆に、右上に大きな値が集まっているのなら、価格が安いほど評価が高いことになります。右側のどの行にも同じぐらい大きな値が集まっていれば、価格に関係なく比較的高評価だということになりますね。このように、大きな値がどのあたりに集まっているかによって項目同士の関係も見えてくるというわけです。

ヒートマップでクラスターを可視化する ～ グループ分けを見やすくする

ヒートマップは項目同士の関係を可視化するだけでなく、集団が幾つかのクラスター（グループ）から成り立っていることを可視化するのにも使えます。データをその値によって「似たもの同士」を集め、幾つかのグループに分けるには**クラスタリング**と呼ばれる機械学習の手法が使われます。ここでは、クラスタリングによりグループ分けされた結果をヒートマップとして表示する例を紹介しておきます。

とはいうものの、残念ながら Excel にはクラスタリングの機能がありません。そのため、クラスタリングは既に行われたものとして、その結果のデータを使うことにします。図 8 は、**k-means** 法と呼ばれる方法で年齢とインターネットの利用時間を元にクラスタリングを行った結果です（データは架空のものです。また、プログラムは後でまとめて掲載します）。

k-means 法については、「[AI・機械学習の数学] 総和を表すΣは機械学習に必須の記号」の 5 ページ目で基本的な考え方を紹介しています。なお、実際に k-means 法でクラスタリングを行う Python のプログラムをこの記事の最後のコラムで参考として掲載しておきます。

インターネット利用時間のデータ

	A	B	C	D	E
1	No.	Sex	Age	Minutes	
2	1	F	55	55	
3	2	M	34	24	
4	3	F	22	240	
5	4	F	53	99	
6	5	M	53	154	
7	6	M	28	98	
8	7	M	47	174	
9	8	M	46	203	
10	9	F	71	49	
11	10	F	40	78	
12	11	F	83	24	
13	12	M	31	184	
14	13	M	26	207	

年齢と利用時間を基に、k-means法で
6つのグループにクラスタリング

	A	B	C	D	E	F
1	No.	Sex	Age	Minutes	Group	
2	1	F	55	55	5	
3	2	M	34	24	1	
4	3	F	22	240	3	
5	4	F	53	99	2	
6	5	M	53	154	4	
7	6	M	28	98	1	
8	7	M	47	174	4	
9	8	M	46	203	3	
10	9	F	71	49	5	
11	10	F	40	78	1	
12	11	F	83	24	0	
13	12	M	31	184	3	
14	13	M	26	207	3	

グループの番号が
追加された

図 8　年齢とインターネットの利用時間を基にクラスタリングを行った結果

この例では、6 つのグループに分けている。例えば年齢（Age）が 55 歳、インターネットの利用時間（Minutes）が 55 分の人は **5 番**のグループ、年齢が 34 歳、インターネットの利用時間が 24 分の人は **1 番**のグループに振り分けられた。なお、データは 160 件あり、セル **A1** ～ **E161** に入力されている（1 行目は項目見出し、E 列がクラスタリングの結果）。

このデータを基にピボットテーブル／ヒートマップを作成してみましょう。ここでは、グループと年齢ごとにインターネット利用時間の**平均**を求めてください。サンプルファイルはこちらからダウンロードできます。Google スプレッドシートの場合はこちらのサンプルファイルを開いて、メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。なお、Microsoft 365 オンラインでは、ピボットテーブルの作成はできますが、［グループ化］ができないため、説明を割愛します。

では、[クラスタリング］ワークシートを開いて取り組んでみてください。手順は図 9 の後に箇条書きで示しておきますが、既に見た手順とほぼ同じなので、まずは独力でチャレンジしてみましょう。

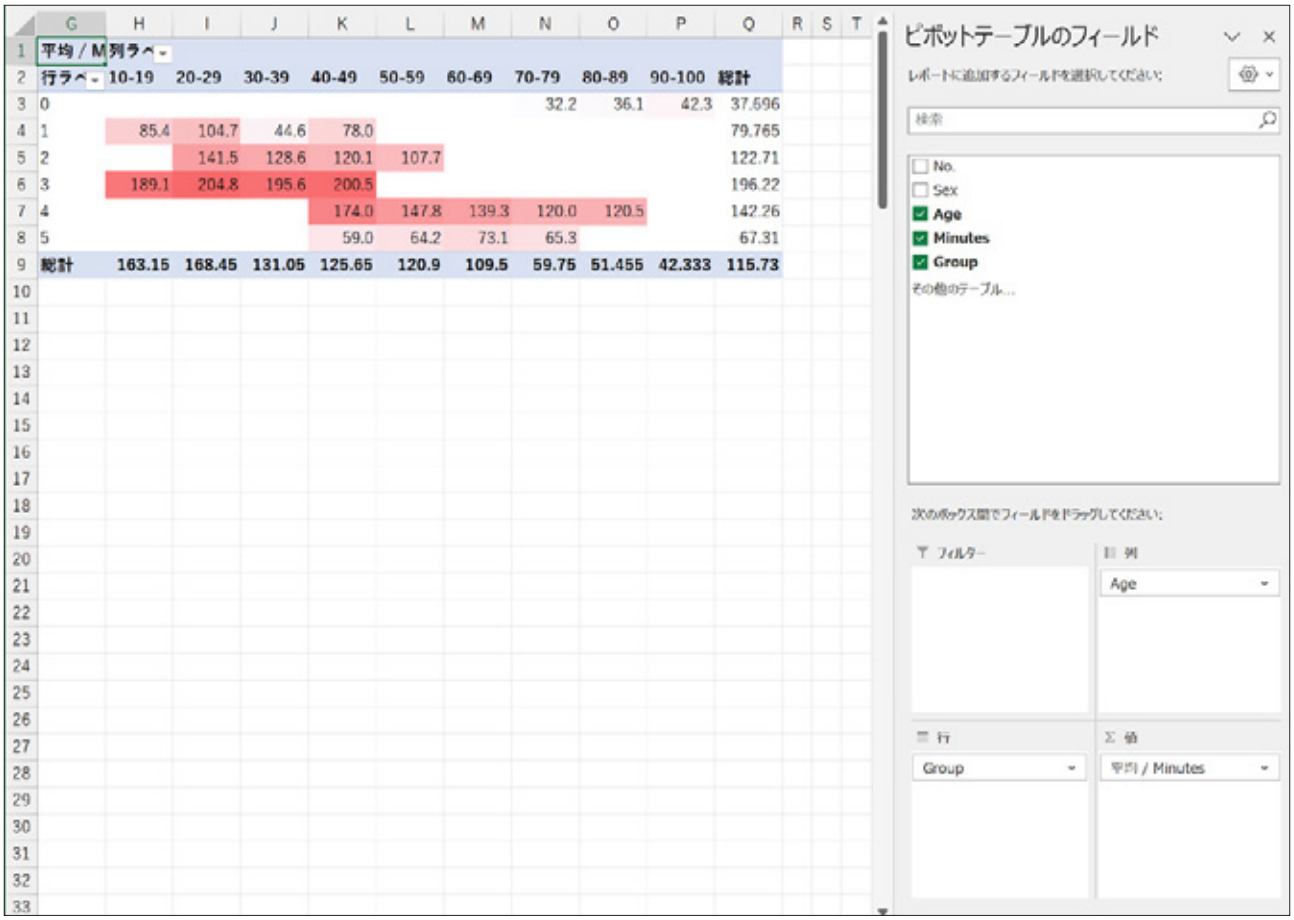


図 9　クラスタリングの結果を基にピボットテーブル／ヒートマップを作成する

ピボットテーブルを作成し、条件付き書式を使って色分けしてみた。6 つのグループごと、年齢ごとにインターネットの利用時間が可視化できる。それぞれのグループがどのような特徴を持つのかを見つけるのに役立つ。完成例は［時間を集計］ワークシートに作成されている。

手順は以下の通りです。これについても、Google スプレッドシートでの操作は後でまとめておきます。

Excel での操作手順

- セル **A1** ～ **E161** のいずれかのセルをクリックしておく
- ［挿入］タブを開き、［ピボットテーブル］ボタンをクリックする（［ピボットテーブルの作成］ダイアログボックスが表示される。ただし、Excel のバージョンによってはダイアログボックスの名前が［テーブルまたは範囲からのピボットテーブル］となっている）
- ［テーブル／範囲］が「クラスタリング !$A$1:$E$161」になっていることを確認する
- ［ピボットテーブルレポートを作成する場所を選択してください］の下の［既存のワークシート］をクリックしてオンにする
- ［場所］ボックスをクリックし、セル **G1** をクリックする
- ［OK］ボタンをクリックする

これで空のピボットテーブルができるので、以下のように進めます。

- ［ピボットテーブルのフィールド］のリストにある［Group］を［行］の欄にドラッグする
- ［ピボットテーブルのフィールド］のリストにある［Age］を［列］の欄にドラッグする
- ［ピボットテーブルのフィールド］のリストにある［Minutes］を［Σ値］の欄にドラッグする
- ［Σ値］の欄の「合計／ Minutes」（Mac 版の場合は［i］のアイコン）をクリックし、［値フィールドの設定］を選択する
- ［選択したフィールドのデータ］リストから［平均］を選択し、［OK］をクリックする

続いて、年齢（Age）を 10 歳ごとの階級に分けます。

- セル **H2** ～ **BS2** のいずれかを右クリックして［グループ化］を選択する
- ［グループ化］ダイアログボックスで［先頭の値］に「10」を入力、［末尾の値］に「100」を入力、［単位］に「10」を入力する

これで、ピボットテーブルの完成です。求められた平均値は小数点以下の桁数が多くなるので、［ホーム］タブの［数値］グループにある［小数点以下の表示桁数を減らす］ボタンなどを利用して、見やすくしておくといいでしょう。

ヒートマップの作成は簡単ですね。

- セル **H3** ～ **P8** をドラッグして選択する
- ［ホーム］タブを開く
- ［条件付き書式］－［カラー スケール］－［赤、白のカラー スケール］を選択する

Google スプレッドシートでの操作手順

- セル **A1** ～ **E161** のいずれかのセルをクリックしておく
- メニューバーから［挿入］－［ピボットテーブル］を選択する
- ［ピボットテーブルの作成］ダイアログボックスの［データ範囲］が「クラスタリング !A1:E161」になっていることを確認する
- ［挿入先］の下の［既存のワークシート］をクリックしてオンにする
- ［データ範囲を選択］ボタン（田のマークのボタン）をクリックして、セル **G1** をクリックし、［OK］をクリックする
- ［作成］ボタンをクリックする

これで、空のピボットテーブルが作成されます。続けましょう。

- 右側に表示されている項目一覧の［Group］を、左側の［行］の下にドラッグする
- 右側に表示されている項目一覧の［Age］を、左側の［列］の下にドラッグする
- 右側に表示されている項目一覧の［Minutes］を、左側の［値］の下にドラッグする
- ［集計］ボックスから［AVERAGE］を選択する

年齢（Age）に 10 歳ごとの階級を設定しましょう。

- 作成されたピボットテーブルの［Age］の値（2 行目の値ならどれでもいい）を右クリックして［ピボットグループのルールを作成］を選択する
- ［グループ化のルール］ダイアログボックスで、［最小値］に「10」、［最大値］に「100」、［間隔のサイズ］に「10」を入力し、［OK］ボタンをクリックする

これでクロス集計表が作成できました。ヒートマップの作成は簡単ですね。

- セル **H3** ～ **P8** をドラッグして選択する
- メニューバーから［表示形式］－［条件付き書式］を選択し、右側に表示される［条件付き書式設定ルール］作業ウインドウで［カラースケール］をクリックする
- ［プレビュー］の色が表示されている部分をクリックして、［白→赤］を選択する
- ［完了］ボタンをクリックする
- ［条件付き書式設定ルール］作業ウインドウの右上の［×］をクリックして、作業ウィンドウを閉じておく

条件付き書式を設定すると、上で見た図 9 のような表になります。ただし、クラスタリングによって分けられたグループがどのような特徴を持つものなのかを吟味するのは人間の役割です。**0 番**のクラスターは高齢者のグループで、インターネットの利用時間が短めです。**1 番**のクラスターには若年層の人が集中しています。**2 番**はもう少し年齢が高めです。働き盛りのグループでしょうか。**3 番**のクラスターは 1 番や 2 番と似ていますが、インターネットの利用時間がかなり長いアクティブユーザーのようです。**4 番**と **5 番**は年齢的には中高年といったところですが、**5 番**の方がインターネットの利用時間が短めですね。というわけで、グループに名前を付けるのは人間がやるべきことですが、ヒートマップを利用すると、グループのまとまりや、グループ間の違いが可視化できるというわけです。

1 ～ **3 番**のクラスターは年齢構成が似ているように思われますが、どの年代の人が多いかは図 9 からは分かりません。それを知るためには、各グループを構成する人の人数を集計する必要があります。サンプルファイルには人数を集計した結果（[人数を集計] ワークシート）も含めてあります。それを見ると、**2 番**は 40 歳代の人が多く、**3 番**は 10 歳代の人が多いことが分かります。なお、年齢とインターネットの利用時間を基に、クラスターごと色分けした散布図を作成すれば、もう少し詳しいことが分かりそうです……が、それについては、次回のお話とします。

コラム　k-means 法によるクラスタリングのプログラム

図 8 のようにクラスタリングを行い、グループ分けを行うプログラムは以下に掲載しておきます。このリンクをクリックすれば、ブラウザが起動し、Google Colaboratory で以下のコードが表示されます（Google アカウントでのログインが必要です）。［ドライブにコピー］ボタンをクリックすれば、自分の Google ドライブにコピーできます。コードの部分をクリックして［Shift］+［Enter］キーを押せば、プログラムが実行され、クラスタリングを行った結果が Excel のファイルとして作成されます。コードの詳細についてはこの記事の範囲を大きく逸脱するので割愛しますが、コード中のコメントと説明を参照していただければだいたいの意味は分かると思います。

```
import pandas as pd
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

# データの読み込み
df = pd.read_excel("https://github.com/Gessys/data_analysis/raw/main/11b.xlsx", sheet_name=" 年齢と利用時間 ")
data = df.loc[:, "Age":"Minutes"]  # 年齢と時間だけを取り出す

# データを標準化する（年齢と時間では値の大きさが異なるので）
sc = StandardScaler()
data_sc = sc.fit_transform(data)

# k-means 法によるクラスタリング
model = KMeans(n_clusters=6, random_state=0, n_init="auto") # 6グループに分ける
model.fit(data_sc)

# 元のデータにクラスター番号を追加する
cluster_no = pd.DataFrame(model.labels_, columns=["Group"])
dfresult = pd.concat([df, cluster_no], axis=1)

# Excel ファイルに書き出す
dfresult.to_excel("11b_add.xlsx", index=False)
```

リスト 1　k-means 法によるクラスタリングを行うプログラム

sklearn.cluster モジュールの KMeans 関数により k-means 法のモデルを用意し、標準化したデータ data_sc に当てはめる。クラスタリングされた結果は model.labels_ で取得できるので、元のデータ df と取得したクラスター番号をつないで Excel のファイルに書き出す。

→目次に戻る

作成された Excel ファイルは以下の手順でダウンロードできます。

- Google Colaboratory の左側の領域に表示されているファイル一覧の中にある「11b_add.xlsx」の右側にマウスポインタを合わせる
  - ・ファイル一覧が表示されていない場合は、画面の左端にあるフォルダの形のアイコンをクリックする
- 3 つの点が縦に表示されたアイコンをクリックし、［ダウンロード］を選択する

図 8 のサンプルファイル（11b.xlsx）は、元のファイルと 11b_add.xlsx ファイルをひとまとめにして、書式などを設定したものです。もちろん、データを Excel に移さず、ピボットテーブルの作成からヒートマップの作成までを Python で実行することもできます。コードは以下の通りです。リスト 1 のコードに続けて実行すれば、ピボットテーブル（クロス集計表）とヒートマップが作成されます。

```
# 年齢の階級を作る
bins = range(10, 100, 10)
dfresult["AgeClass"] = pd.cut(dfresult["Age"], bins, right=False)

# クロス集計表を作る
from pandas.core.groupby import grouper
cross_table = pd.pivot_table(dfresult, values="Minutes", index="Group",
columns="AgeClass", aggfunc="mean", fill_value=0)
cross_table
```

リスト 2　ピボットテーブルによりクロス集計表を作る

pandas モジュールの cut 関数により、クラスタリングされた結果（dfresult）の年齢を **10** 以上 **101** 未満まで **10** 刻みの階級に分けた列を作る。引数 right=False を指定すると階級が「～以上～未満」となり、Excel の結果と同じになる。引数 right を指定しないと階級は「～より大～以下」となる。クロス集計表は pandas モジュールの pivot_table 関数や crosstab 関数で作成できる。ここでは pivot_table 関数を使った。引数 values に集計する項目を指定し、引数 index には行の見出しを、引数 columns には列の見出しを指定する。引数 aggfunc には集計の方法を指定。ここでは平均を表す **"mean"** を指定した。引数 fill_value は、値が存在しないとき（NaN のとき）に代わりに出力する値を指定する。

```
# ヒートマップを作成する
import seaborn as sns
sns.heatmap(cross_table, annot=True, fmt=".1f", cmap="Reds")
```

リスト 3　ヒートマップを表示するプログラム

ヒートマップは seaborn モジュールの heatmap 関数にクロス集計表を指定するだけでできる。引数 annot は値を表示するかどうかの指定、fmt は値の書式。ここでは、小数点以下 1 桁まで表示するようにした。引数 cmap は配色の指定。

実行結果は以下のようになります（図 10）。

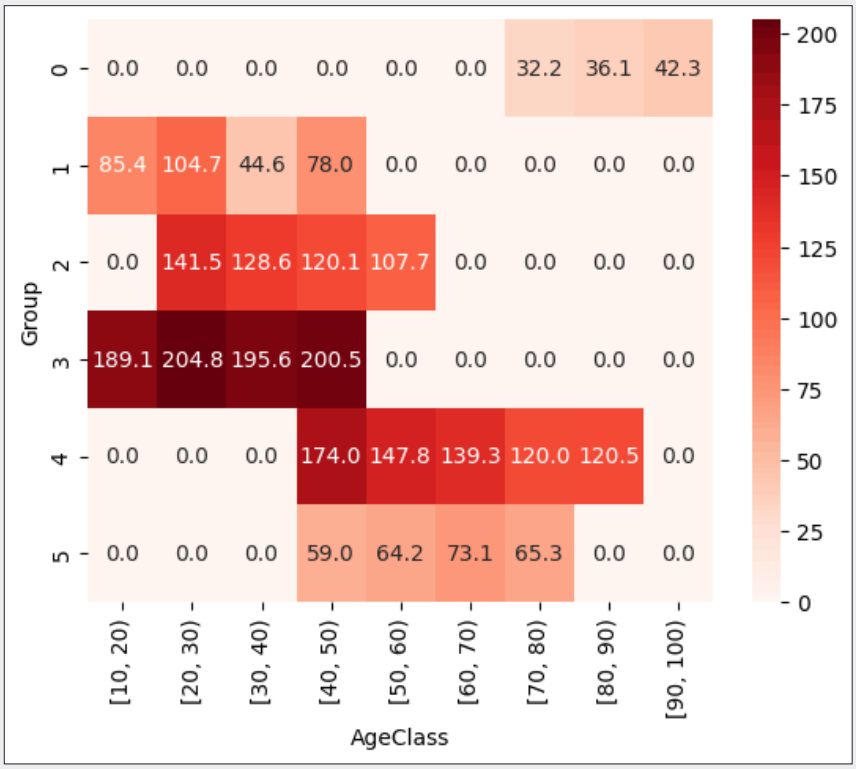


図 10　クラスタリングの結果を基にピボットテーブル／ヒートマップを作成する

Excel で作成したヒートマップと同様の結果になる。画像を右クリックして［名前を付けて保存］を選択すれば、ヒートマップを PNG ファイルとして保存できる。

今回は、ピボットテーブルを利用してクロス集計表を作成し、ヒートマップを作成しました。ヒートマップでは、値の大小を色分けすることにより、項目同士の関係を可視化したり、集団を幾つかのグループに分けるヒントを得ることができます。

次回も「関係」に注目し、散布図を作成することとします。関係の可視化だけでなく、外れ値の可視化や、値の飽和（サチュレーション＝測定限界を超えた値が数多く存在すること）などの検出もできます。また、多くの項目同士の関係を見るため、複数の散布図をまとめて表示したり、グループごとに色分けして散布図を表示したりする方法も併せて紹介します。次回もどうぞお楽しみに！

［データ分析］散布図を徹底活用して「関係」を可視化 ～ 関係と規模を一度に見る

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 12 回。グラフを利用して項目同士の関係や、その中での値の大きさを可視化します。散布図やバブルチャートの詳細な取り扱いと視覚的な分析について、ケーススタディを通して学びましょう。

羽山博（2023 年 12 月 07 日）

前回はピボットテーブルによるクロス集計表の作成とヒートマップを利用して分布を可視化しました。今回は、散布図とバブルチャートによって分布を可視化します。

いずれも項目間の関係や分布を可視化するのに便利ですが、ヒートマップは変数が名義尺度や順序尺度である場合、または、階級に分けられている間隔尺度の場合に便利です。前回見たワインの価格は間隔尺度ですが、それを階級に分けて利用しました。評価はそもそも順序尺度ですが、便宜的に間隔尺度として取り扱い、得られた平均値によって階級を分けていました。ちなみに尺度については第 2 回で説明しています。

今回取り扱う散布図は、主として変数が間隔尺度の場合に使われます。そこで、前回のデータをそのまま（階級に分けずに）散布図を作成してみます。今回は、売上金額を含めた表を基に、価格と評価、そして売上金額の関係を散布図やバブルチャートで分析していくことにします（図 1）。前回同様、データは架空のものです。

図 1　散布図による分布の可視化+規模の可視化=バブルチャート!

ワインの価格と評価の関係に対する、売上金額の大きさを可視化したい。散布図だと、2 つの変数（価格と評価）の関係しか分からない。しかしバブルチャートなら、2 つの変数の関係だけでなく、もう 1 つの変数（売上金額）の大きさも可視化できる。

→目次に戻る

今回のテーマは、散布図を利用した 2 つの変数の関係の可視化することです。さらにバブルチャートを利用して 2 つの変数の関係だけでなくもう 1 つの変数の規模（大きさ）も可視化します。後半のコラムでは、散布図を色分けしてグループを可視化する方法と、数多くの項目について散布図をまとめて作る方法についても紹介します。

この記事は、データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 12 回です。特に、第 7 回から今回の散布図／バブルチャートまでは「可視化シリーズ」として、グラフの使い方と分析の観点について解説しています。第 7 回の棒グラフ、第 8 回の折れ線グラフ、第 9 回の円グラフ／パレート図、第 10 回のヒストグラム／箱ひげ図、第 11 回のクロス集計表／ヒートマップも併せてご参照ください。これらのグラフの目的と効用などについて、特別予告編で簡単に整理しています。事前に確認しておくとより理解が深まるでしょう。

この記事で学べること

今回は以下のようなポイントについて、分析の方法や目の付けどころを見ていきます。

- 散布図による分布の可視化 …… 間隔尺度のデータの関係を可視化する
- バブルチャートによる分布と規模の可視化 …… 散布図の中で値の大きさを可視化する
- 散布図やバブルチャートの活用 …… 複数の散布図を一度に作ったり、グループにより色分けしたりする

では、基本の基本である散布図の作成から見ていきます。サンプルファイルの利用についての説明の後、本編に進みましょう。

サンプルファイルの利用について

本稿では、表計算ソフトを使って手を動かしながら学んでいきます。表計算ソフト Microsoft Excel 用の .xlsx ファイルをダウンロードできるようにしています。デスクトップ版の Excel が手元にない場合は、Microsoft アカウントがあれば使える無料の Microsoft 365 オンライン、もしくは Google アカウントがあれば使える無料の Google スプレッドシート（Google Sheets）をお使いください。Microsoft 365 オンラインの場合は、.xlsx ファイルを OneDrive にアップロードしてから開いてください。Google スプレッドシートの場合は、.xlsx ファイルを Google ドライブにアップロードしてから開いた上で［ファイル］メニューの［Google スプレッドシートとして保存］を実行してください（Google スプレッドシート独自の機能を使っている場合は、ファイルを共有して参照できるようにします。その場合は、該当する箇所で使い方を記します）。

散布図を作成して関係を可視化する ～ 外れ値や項目の特徴も見つけよう

では、さっそく散布図を作成してみましょう。サンプルファイルをこちらからダウンロードし、[ワインの売り上げ(1)]ワークシートを開いて取り組んでみてください。Google スプレッドシートの場合はこちらのサンプルファイルも利用できます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。

手順は図の後に箇条書きで示しておきます。ただし、タイトルやグラフの表示位置、サイズなど、データ分析そのものにあまり関係のない設定については省略してあります。なお、動画でも手順を解説しているので、操作を一つ一つ追いかけたい方はぜひご視聴ください。

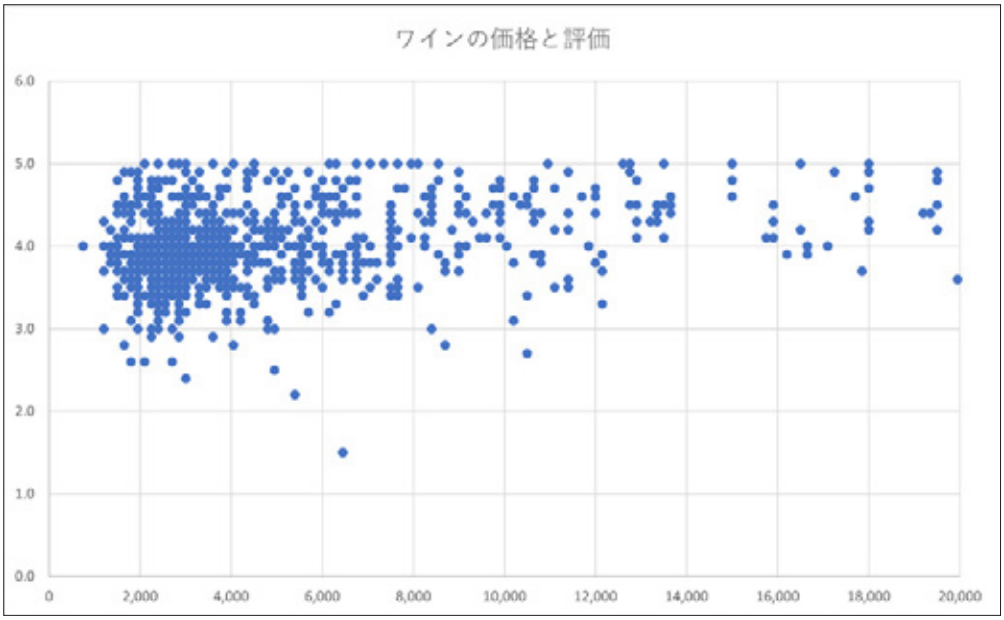


図 2　ワインの価格と評価の散布図（完成イメージ）

前回は、価格と評価をグループ化し、階級に分けてヒートマップを作成したが、価格は間隔尺度であり、評価も便宜的に間隔尺度として扱うので、素直に散布図として表してみよう。

以下の手順で進めていきましょう。

Excel での操作手順

- セル **C3** ～ **D1003** を選択する
- ［挿入］タブを開き、［散布図（X, Y）またはバブルチャートの挿入］ボタンをクリックする
- ［散布図］を選択する

Google スプレッドシートでの操作手順

- セル **C3** ～ **D1003** を選択する
- メニューバーから［挿入］－［グラフ］を選択する
- ［グラフエディタ］作業ウインドウで［グラフの種類］リストから［散布図］を選択する

セル **C3** ～ **D1003** を選択するにはドラッグ操作よりも［名前］ボックスに「C3:D1003」と入力する方が簡単です。これで図 3 のような散布図が作成されます。

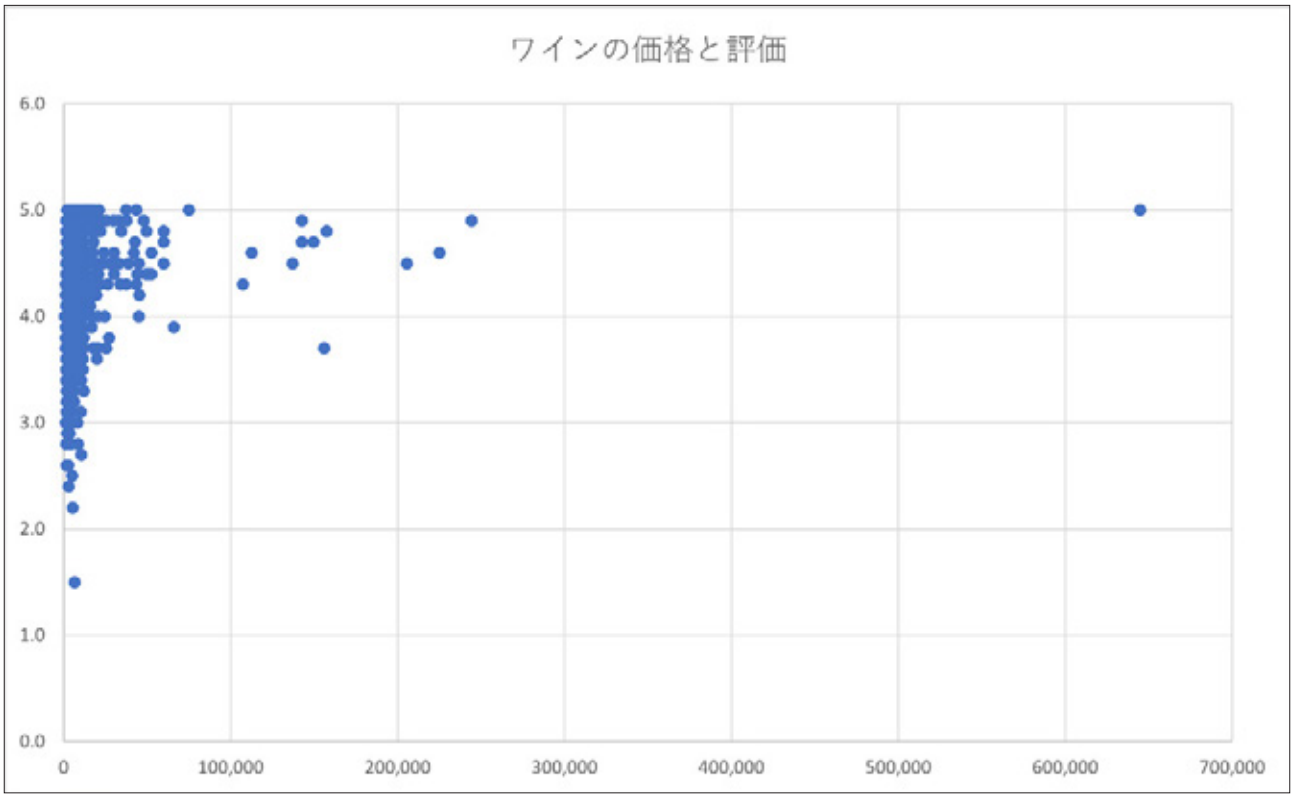


図 3　ワインの価格と評価の散布図（外れ値の発見）

価格を横軸（X 軸）、評価を縦軸（Y 軸）とした散布図が作成された。一見して **65 万円**近くに外れ値（高価なワイン）があることが分かる。価格と評価はおおむね比例しているようだが、低価格のワインほど評価のばらつきが大きいようにも見える。

図 3 から、**65 万円**（実際の値は **64 万 5000 円**）のワインが外れ値として存在することが分かります。そこで、高価なワインは除外し、値が密集している部分を詳細に見るために、横軸の最大値を変更してみましょう。ここでは、最大値を **20000** とします。

Excel での操作手順

- 横軸の目盛を右クリックし［軸の書式設定］を選択する
- ［軸の書式設定］作業ウインドウで［最大値］に「20000」と入力する
  - ［最大値］という項目が表示されていない場合は、［軸のオプション］ボタンをクリックし、［>軸のオプション］という項目をクリックすれば、［最大値］が入力できるようになる

Google スプレッドシートでの操作手順

- グラフを右クリックし［軸］－［横軸］を選択する
- ［グラフエディタ］作業ウインドウの［カスタマイズ］タブで［最大値］に「20000」と入力する

これで、図 2（完成イメージ）のような散布図が作成されます。前回見た通り、**2,000 円**から **3,000 円**で、評価が **3.5** から **4.5** あたりの商品が多いことが分かりますね。さらに、グラフからもう一つ重要なことが読み取れます。それは、評価が **5.0** で頭打ちになっているということです。そもそも、**2,000 円**から **3,000 円**あたりの **5.0** という評価と、**20,000 円**あるいはそれ以上の価格のワインの **5.0** は同じ価値を持つ評価でしょうか。品質の良い高価なワインであれば「☆ 5 つではなく☆ 100 個を付けたい」というように、もっと高い評価を付けたい人も多いのではないでしょうか。ネットショッピングの口コミは 5 段階評価が多いので、図 2 のようなグラフになってしまいますが、より繊細かつ厳密に評価するのであれば、100 点満点にした方がよさそうだということも分かりますね。

ワインの評価としては、**パーカーポイント**と呼ばれる評価が有名です。パーカーポイントでは、色や風味、質などを基に、100 点満点で値が与えられます。詳細については、サッポロビールのワインに関するページなどをご参照ください。

バブルチャートを作成して規模も可視化する ～ 売り上げに貢献しているクラスを見つける

図 2 や図 3 を見ると、ある価格のワインがどのような評価を得ているかが分かりますが、そのワインの売り上げがどの程度であるかは分かりません。手頃に購入できるワインは、単価は安くても売上数量が多いので、合計の売上金額も大きいと考えられそうです。一方、高価なワインは、単価が高いので、売上数量が少なくても合計の売上金額は大きいかもしれません。そこで、散布図に売上金額のような「規模」（大きさ）を表す値を反映させるためにバブルチャートを作ってみましょう。

作成の方法は簡単です。横軸に対応する項目と縦軸に対応する項目、規模に対応する項目を指定してグラフを作成するだけです。

では、散布図の作成に利用したファイルの［ワインの売り上げ（2）］ワークシートを開いて取り組んでみてください。［ワインの売り上げ（1）］と同じデータですが、作業しやすいように別のワークシートにしておきました。箇条書きにした以下の手順でバブルチャートを作成してみましょう。これについても、動画での解説も用意してあるので、操作を一つ一つ追いかけたい方はぜひご視聴ください。なお、現在のところ、Microsotf 365 オンラインにはバブルチャートの機能がありません。

Excel での操作手順

- セル **C3** ～ **D1003** とセル **F3** ～ **F1003** を選択する
- ［挿入］タブを開き、［散布図（X, Y）またはバブルチャートの挿入］ボタンをクリックする
- ［バブル］を選択する

Google スプレッドシートでの操作手順

- セル **C3** ～ **D1003** とセル **F3** ～ **F1003** を選択する
- メニューバーから［挿入］－［グラフ］を選択する
- ［グラフエディタ］作業ウインドウで［グラフの種類］リストから［バブルチャート］を選択する
- ［サイズ］ボックスで「売上金額」を選択する

グラフ化するセルの範囲を選択するには、ドラッグ操作と［Ctrl］＋ドラッグ操作で離れた範囲を選択するよりも［名前］ボックスに「C3:D1003,F3:F1003」と入力する方が簡単です。ただし、Google スプレッドシートでは、この指定ができないので、［名前］ボックスに「C3:D1003」と入力した後、［Ctrl］キーを押しながらセル **F3** ～ **F1003** をドラッグする必要があります。別の方法としては、E 列の見出しをクリックし［列を非表示］を選択した後、［名前］ボックスに「C3:F1003」と入力するのが簡単です。

この段階では、図 4 のようなグラフになっています。

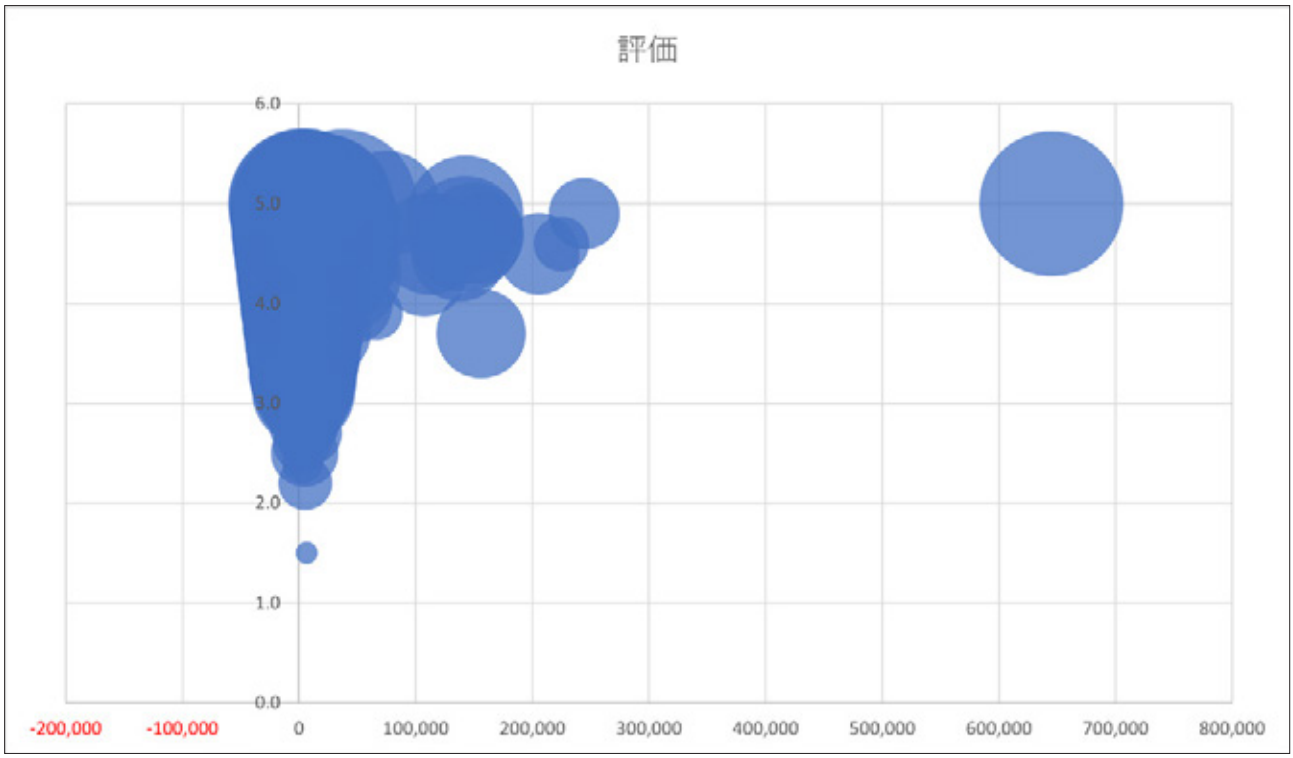


図 4　ワインの価格、評価、売り上げのバブルチャート（作成途中）

取りあえず作成されたバブルチャート。**約 65 万円**の高価なワインはかなり売り上げが大きいことが分かる。手頃な価格のワインに関してはバブル（円）が重なっているのでよく分からない。横軸の最小値と最大値、バブルのサイズを調整しよう。

→目次に戻る

続けて、以下の手順で最小値と最大値の設定やバブルサイズの変更を行っておきましょう。ただし、Google スプレッドシートではバブルサイズが自動的に決められ、変更ができないようなので、その操作は省略します。

Excel での操作手順

- 横軸の目盛を右クリックし［軸の書式設定］を選択する
- ［軸の書式設定］作業ウインドウで［最大値］に「0」と入力する
- ［軸の書式設定］作業ウインドウで［最小値］に「20000」と入力する
  - ・［最大値］という項目が表示されていない場合は、［軸のオプション］ボタンをクリックし、［>軸のオプション］という項目をクリックすれば、［最大値］が入力できるようになる
- データ系列（バブルの部分ならどこでもよい）をクリックして作業ウインドウの表示を［系列の書式設定］作業ウインドウに切り替える
  - ・作業ウインドウが表示されていない場合は、データ系列を右クリックし［データ系列の書式設定］を選択するとよい
- ［バブルサイズの調整］ボックスに「15」と入力する
-

Google スプレッドシートでの操作手順

- グラフを右クリックし［軸］－［横軸］を選択する
- ［グラフエディタ］作業ウインドウの［カスタマイズ］タブで［最大値］に「20000」と入力する

これで、図 5 のようなバブルチャートになります。

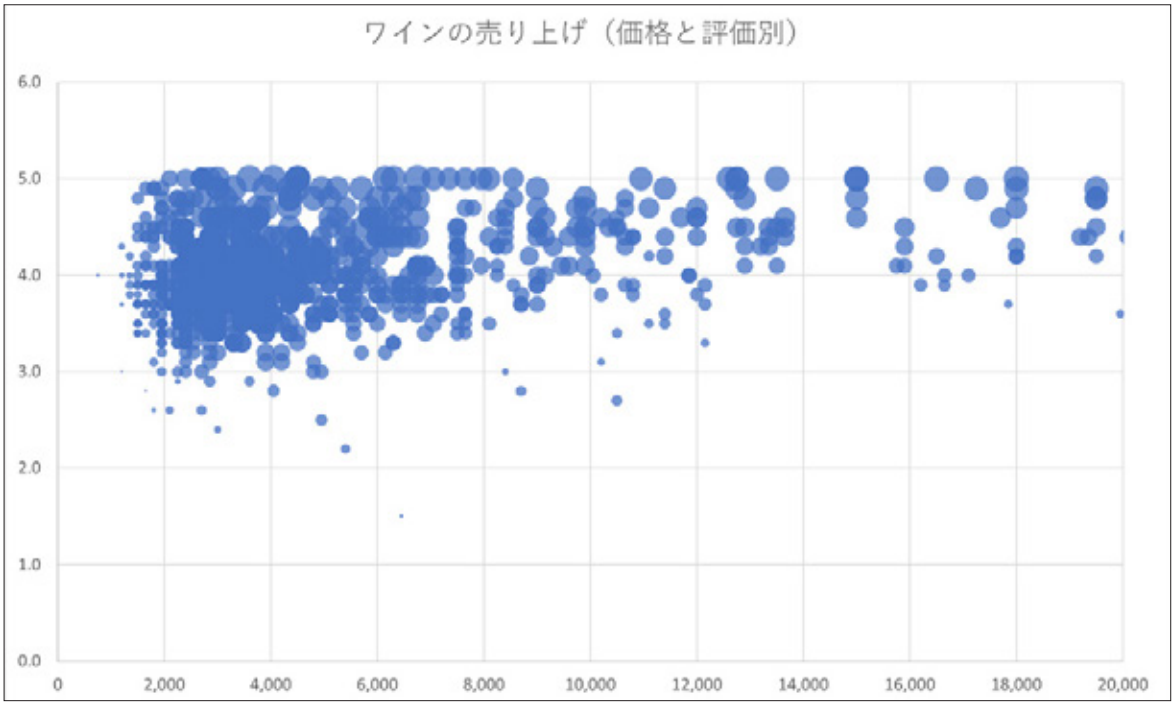


図 5　ワインの価格、評価、売り上げのバブルチャート（完成例）

2,000 円以下の安価なワインや評価の低いワインはバブルのサイズが小さいことからから概して売上金額も小さいことが分かる。
その一方で、価格が高くなっても、合計の売上金額は手頃に入手できるワインとそれほど変わらないことも分かる。

→目次に戻る

価格の高いワインはそれほど売上数量が大きいわけではありませんが、売上金額はお手頃なワインとそれほど変わりません。数量が少なくても売り上げが上がるということは、在庫のためのスペースが少なくて済むということですね。これは大きなメリットです（同時に、保管方法に特別な配慮が必要になったり、盗難に遭った際のリスクが大きくなったりしますが）。なお、バブルチャートでは系列のデータ数が多いとバブルが重なり過ぎて、見づらくなることがあります。その場合はバブルのサイズを小さくするとある程度見やすくなります。例えば、バブルサイズを「5」に変更すると、図 5 の値が集中している箇所が見やすくなります。

コラム　バブルチャートをグループごとに色分けする

前回のコラムで k-means 法を利用したクラスタリングの例と、それをヒートマップとして表示する方法を紹介しました。散布図やバブルチャートでも同様にグループによる色分けができれば、より多角的な可視化ができますね。残念ながら、Excel ではかなり難しいので、Python でのプログラムで作成した例を紹介します（図 6）。

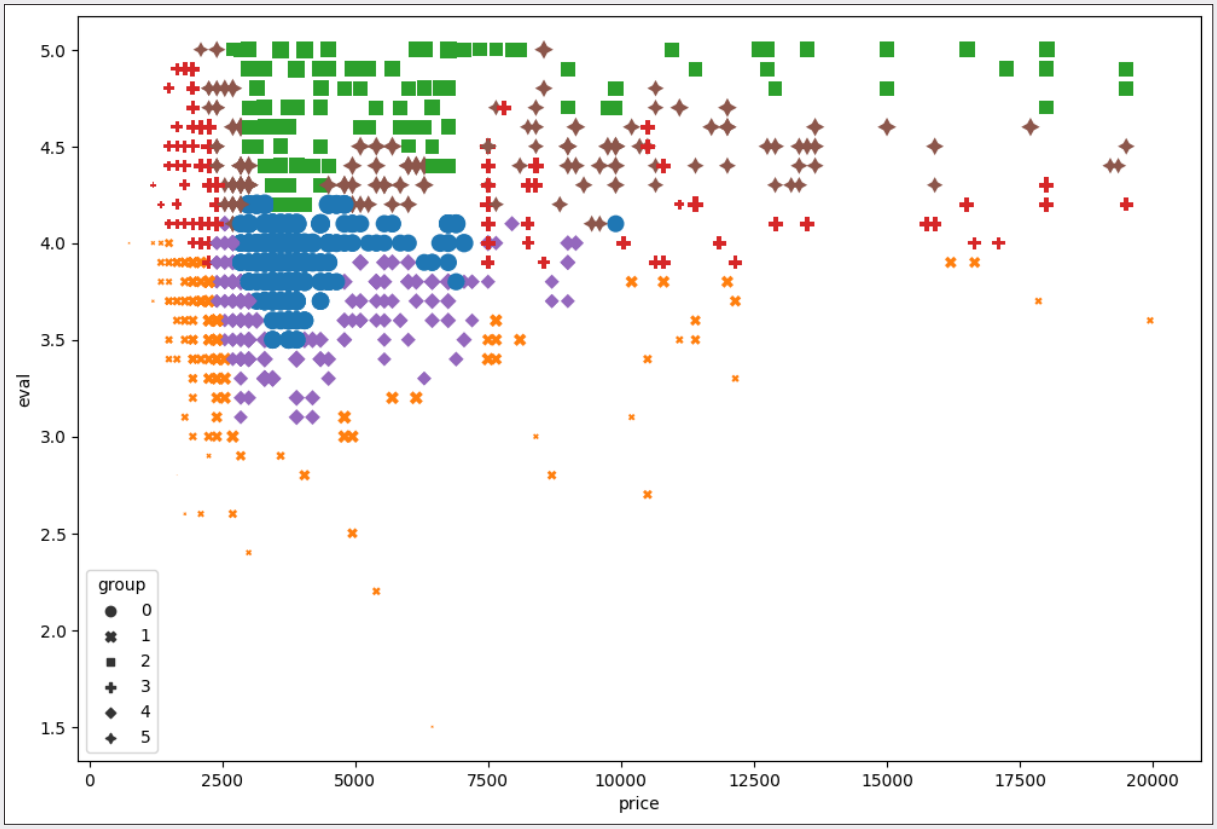


図 6　ワインの価格、評価、売り上げのバブルチャート（グループごとの色分け）

赤と緑、青と紫など、色分けが見づらい部分があるので、マーカーの形でも区別できるようにした。マーカーのサイズが比較しづらくはなるが、グループ分けされていることはよく分かる。

図 6 から、●で表されている **0 番**のグループは価格がお手頃で評価は中ぐらいといったこと、× で表されている **1 番**のグループが低価格または低評価のグループであり、売り上げがそう大きくないことが分かります。また、**2 番**のグループが価格を問わず高評価であり、売り上げも比較的大きいことも分かります。ただ、それ以上のことは少し分かりにくいですね。

ワインのデータを使って、k-means 法により 6 つのグループを作るコードは以下の通りです。これは前回紹介したコードとほとんど同じです。このリンクをクリックすれば、ブラウザが起動し、Google Colaboratory で以下のコードが表示されます（Google アカウントでのログインが必要です）。［ドライブにコピー］ボタンをクリックすれば、自分の Google ドライブにコピーできます。コードの部分をクリックして［Shift］+［Enter］キーを押せば、プログラムが実行され、クラスタリングを行った結果が dfresult という名前のデータフレームとして作成されます。

コードの詳細についてはこの記事の範囲を大きく逸脱するので割愛しますが、コード中のコメントと説明を参照していただければだいたいの意味は分かると思います。なお、上記のリンクにはグループごとに平均値を求めるためのコードや、その値を基にヒートマップを作成するためのコードも含めてあります。さらなる分析に活用できるので、ぜひご参照ください。

```
import pandas as pd
from sklearn.cluster import KMeans
from sklearn.preprocessing import MinMaxScaler

# データの読み込み
df = pd.read_excel("https://github.com/Gessys/data_analysis/raw/main/12a.xlsx", sheet_name="ワインの売り上げ（1）", skiprows=2, usecols="A:F")
data = df.loc[:, ["価格", "評価", "売上金額"]]

# データをスケーリングする（値の大きさが異なるので、0～1になるように調整する）
sc = MinMaxScaler()
data_sc = sc.fit_transform(data)

# k-means 法によるクラスタリング
model = KMeans(n_clusters=6, random_state=0, n_init="auto")
model.fit(data_sc)

# 元のデータにクラスタ番号を追加する
cluster_no = pd.DataFrame(model.labels_, columns=["グループ"])
dfresult = pd.concat([df, cluster_no], axis=1)
```

リスト 1　k-means 法によるクラスタリングを行うコード

sklearn.cluster モジュールの KMeans 関数により k-means 法のモデルを用意し、MinMaxScaler により最小値を **0**、最大値を **1** に調整（スケーリング）したデータ data_sc に当てはめる。前回のコードでは StandardScaler を使って標準化を行ったが、今回はバブルチャートのマーカーのサイズを指定する必要があるので、売上金額をスケーリングした結果が負の値にならないように、MinMaxScaler を使っている。クラスタリングされた結果は model.labels_ で取得できるので、元のデータ df と、取得したクラスタ番号をつないで dfresult という名前のデータフレームにする。

リスト 1 の結果を基に、価格を横軸に、評価を縦軸に、売上金額をバブルのサイズにし、さらにグループによる色分けを行うコードは以下の通りです。散布図やバブルチャートは matplotlib.pyplot モジュールの scatter 関数やデータを簡単に可視化するのに便利な seaborn モジュールの scatterplot 関数で作成できます。ここでは scatterplot 関数を使っています。

```
!pip install japanize_matplotlib  # これは最初に 1 回実行しておけばよい

import matplotlib.pyplot as plt
import matplotlib.cm as cm
import seaborn as sns
import japanize_matplotlib

# 20000 未満のデータだけを取り出す
dfbubble = dfresult.loc[dfresult.価格 < 20000]
dfsize = data_sc[dfresult.価格 < 20000, 2] * 200  # マーカーのサイズを計算

# バブルチャートを表示する
plt.figure(figsize=(12, 8))
kwargs = {"linewidth": 0}  # 枠線を表示しない
sns.scatterplot(x=dfbubble.価格 , y=dfbubble.評価 ,
                c=cm.tab10(dfbubble.グループ ),  # マーカーの色の指定
                s=dfsize,  # マーカーのサイズの指定
                style=dfbubble.グループ ,  # マーカーの形の指定
                **kwargs)
plt.show()
```

リスト 2　グループによって色分けしたバブルチャートを作成するコード

項目名には「価格」「評価」などの日本語文字が含まれているが、Google Colaboratory では、そのままだとグラフのタイトルなどに日本語文字が表示できない。そのため、!pip コマンドにより japanize_matplotlib モジュールをインストールし、それを import しておく。scatterplot 関数の引数 c に指定した cm.tab10 は、10 色のカラーテーブルを表す。グループは 6 つあるので、10 色のうちの 6 色が割り当てられる。引数 style にはマーカーの形を指定する。これも 6 つのグループに対応する形を割り当てる（図 6 の左下に表示されている凡例を参照）。

コラム　多数の項目同士の散布図を一度に作成する

今回利用したワインのデータでは、項目（列）の数はそれほど多くありませんでした。しかし、項目が多くなると、項目同士の組み合わせも多くなるので、手作業で散布図を作るのはかなり面倒です。そのような場合はプログラムによって散布図をまとめて描画するのが得策です。

図 10 は、seaborn モジュールの pairplot 関数を使って、数多くの項目同士の散布図を作成した例です。利用するデータは scikit-learn のデータセットとして用意されている糖尿病関係のサンプルデータで、age（年齢）、sex（性別）、bmi（体格指数）、bp（血圧）、s1（総コレステロール値）、s2（悪玉コレステロール値）、s3（善玉コレステロール値）、s4（= s1/s3）、s5（中性脂肪の対数）、s6（血糖値）、target（1 年後の糖尿病の進行度）という項目が含まれています。

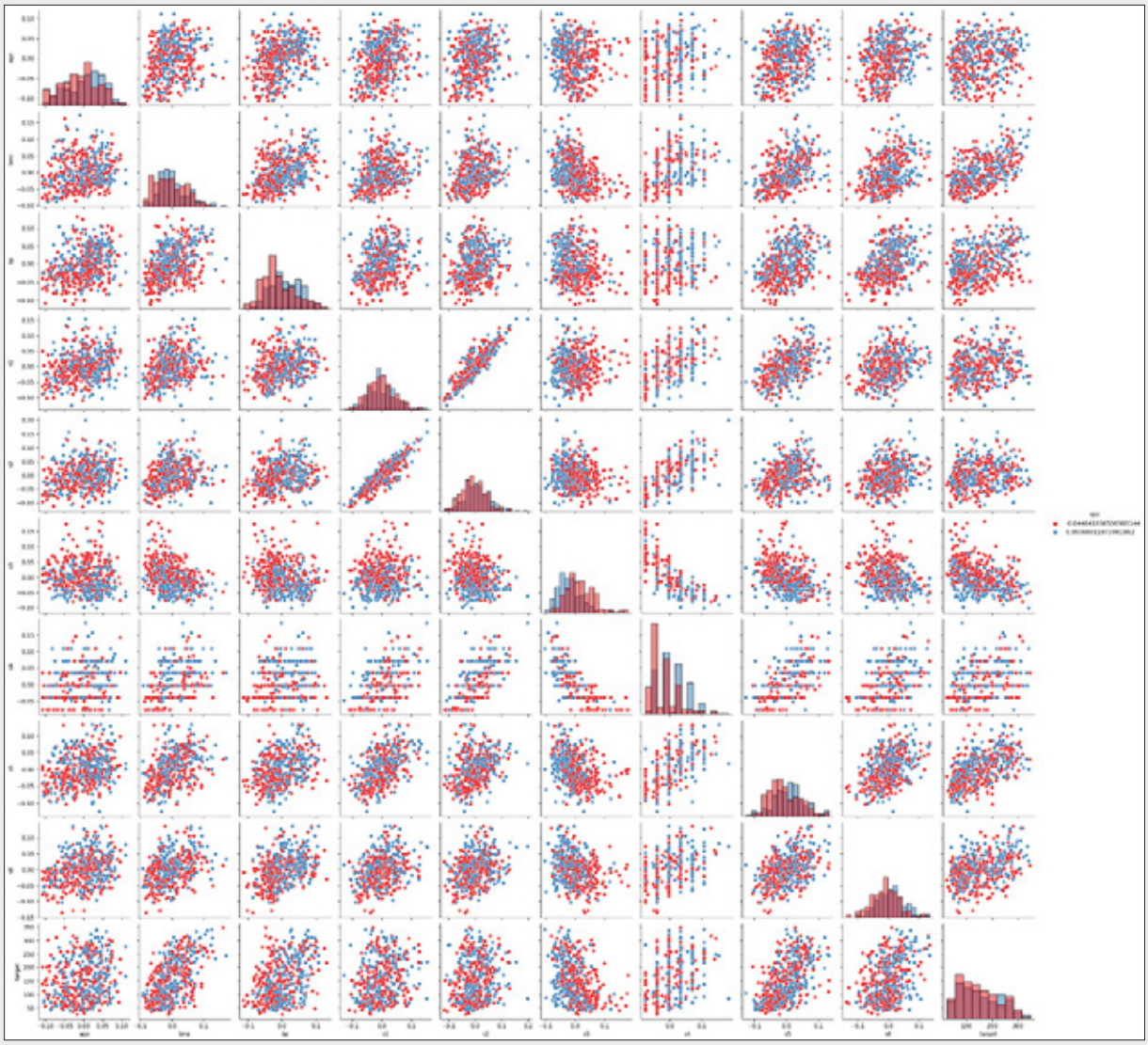


図 7　多数の項目同士の散布図を一気に作成した例

age ～ s6 までの項目同士の散布図。対角線上にある自分自身（age と age など）のグラフはヒストグラムとなっている。また、性別により色分けしてある。糖尿病の進行度と正の相関がありそうな項目は bmi、bp、s4、s5 あたり。糖尿病の進行度と s3（善玉コレステロール値）には負の相関があるように見える。

コードはこのリンクから参照できます。リンクをクリックすれば、ブラウザが起動し、Google Colaboratory で以下のコードが表示されます（Google アカウントでのログインが必要です）。このリンクには、相関行列を求めてヒートマップを作成するためのコードも含めてあります。併せてご参照ください。

```
from sklearn.datasets import load_diabetes  # 糖尿病に関するデータ
import seaborn as sns

df_diabetes = load_diabetes(as_frame=True).frame  # サンプルデータを読み込む
sns.pairplot(df_diabetes,  # 利用するデータ
             hue="sex",  # 性別による色分けを行う
             diag_kind="hist",  # 対角線上はヒストグラムとする
             palette = "Set1")  # 配色は Set1 を使う
```

リスト 3　多くの項目同士の散布図を一気に作成するためのコード

複数の項目同士の散布図は seaborn モジュールの piarplot 関数にデータを指定するだけでできる。ただし、データ数と組み合わせが多いので、実行にはかなりの時間がかかることに注意。

今回は、散布図を利用して間隔尺度の項目間の関係を可視化しました。バブルチャートを利用すれば、それらの関係に加えて、規模を可視化し、さらなる分析に役立てることができます。バブルチャートの色分けや多数の散布図の作成などについては、コラムで Python のプログラムを紹介しました。今回で「可視化シリーズ」はひと区切りです。

次回は関係の強さを数値で表すために、相関係数を計算するとともに、その仕組みについても見ていきます。それ以降の回帰分析による予測へとつながっていくお話です。次回もどうぞお楽しみに！

[データ分析] 相関係数～気温と電気代に関係はあるのか？

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 13 回。変数同士の関係の強さを表す相関係数の計算内容を仕組みから理解します。Excel を使って手を動かしながら、相関係数の意味や求め方、落とし穴などについて学んでいきましょう。

羽山博（2024 年 01 月 11 日）

前回は散布図やバブルチャートを作成して、「変数同士の関係」を**可視化**しました。

今回は、「変数同士の関係」を**数値で表す**ことを考えてきます。**相関係数**は変数同士の関係の強さを表す値としてよく知られていますが、**相関**という言葉だけが一人歩きして、不適切な解釈や使われ方をすることもよくあります。

そこで、相関係数の求め方と意味を確認した後、ケーススタディを通して分析していきます。。その中で、相関係数に関する落とし穴についても解説します。相関係数（ピアソンの積率相関）は間隔尺度の変数同士で使われますが、順序尺度や名義尺度の場合に使われる、関係の強さを表す値（順位相関やクラメールの連関係数）についても最後に紹介します。

まず、問題提起です。図 1 をご覧ください。このデータは気温と家庭での CO_2 排出量の値です（2021 年度）。気温の出典は気象庁のページから札幌、東京、那覇の値を取り出したものです。CO_2 排出量の出典は、環境省による、北海道、関東甲信、沖縄のデータです。

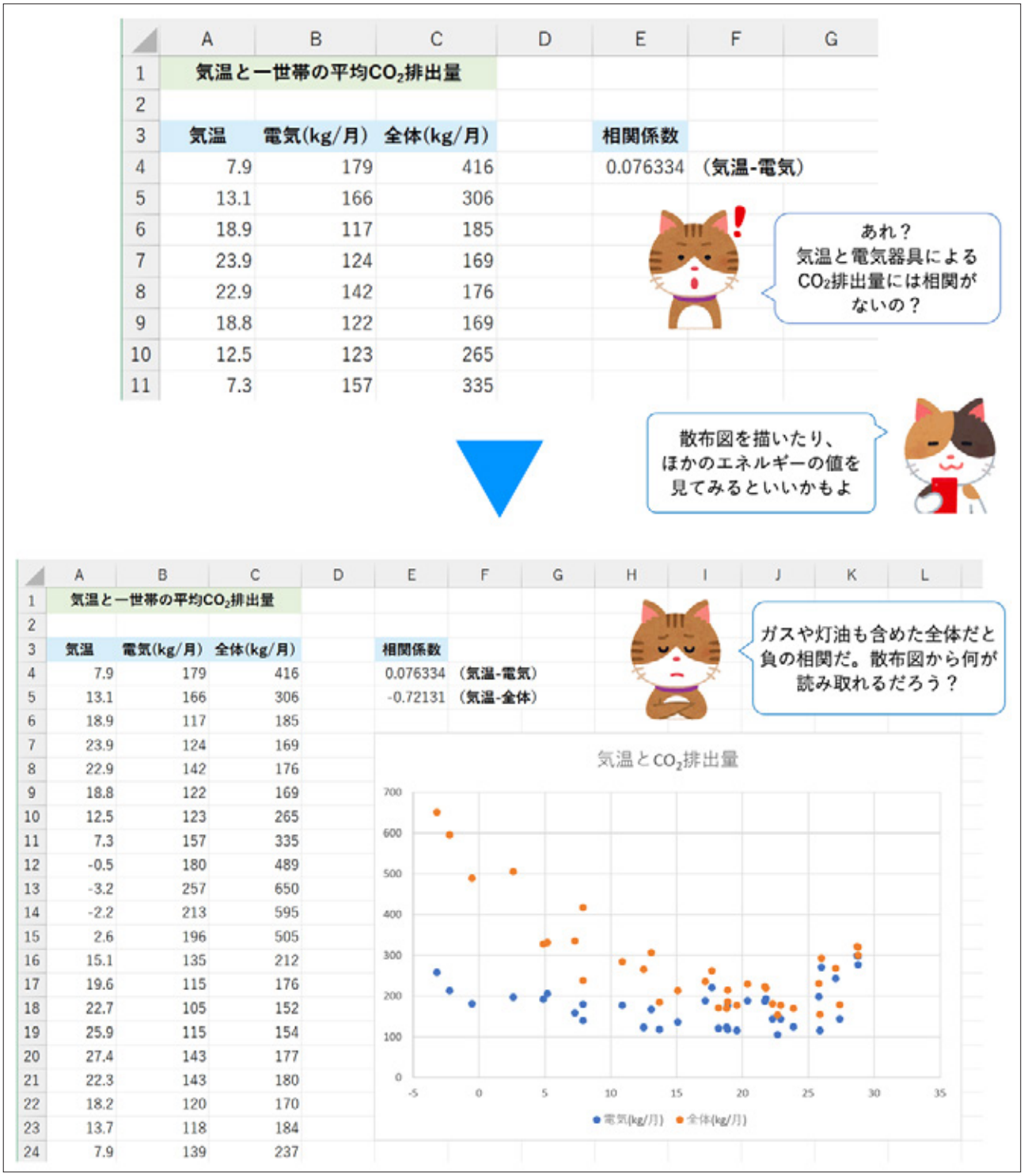


図 1　気温と CO_2 排出量の関係を調べる

相関係数が **1** に近ければ正の相関（一方が増えれば他方も増える）、**-1** に近ければ負の相関（一方が増えれば他方は減る）となるが、気温と電気器具による CO_2 排出量の相関係数は **0** に近い（無相関）。ということは、気温が上がっても下がっても、電気器具による CO_2 排出量は変わらないということなのだろうか。個人的には夏の電気代が高いので直感に反するのだが……。

北海道、関東甲信といってもかなり広いので、気温を札幌と東京で代表させるのはやや無理がありますが、図 1 を見ると気温と電気による CO_2 の排出量の相関係数は **0** に近いので、相関はないようです。暑い時期にはエアコンがフル稼働し、夏の電力不足が毎年のように話題になるので、気温が上がれば CO_2 の排出量も増えるような気がしますが、どうもそうではないように見えます。

一方、ガスや灯油なども含めると、相関係数が **-1** に近い値になるので、負の相関が見られます。つまり、気温が下がるほど CO_2 の排出量が増えるというわけです。

なぜ、このような結果になるのかを考えていきたいと思います。……が、その前に、そもそも相関係数とはどのような計算で求められる値なのか、どういう意味を持つのかをきちんと確認しておきましょう。今回はそこからスタートです。

この記事は、データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 13 回です。第 3 回〜第 6 回は代表値や散布度について、データ可視化入門【特別予告編】の後、第 7 回〜第 12 回は可視化をテーマに解説してきました。今回から第 15 回までは「関係」に注目し、相関係数や回帰分析などについて見ていきます。トップページから全体の目次が参照できます。

この記事で学べること

今回は以下のようなポイントについて、分析の方法や目の付け所を見ていきます。

- 相関係数の求め方と意味 …… 相関係数の計算方法を図形的に見ながら意味を理解する
- 相関係数の落とし穴（1）…… 相関係数が小さくても関係が強いこともある
- 相関係数の落とし穴（2）…… 疑似相関（見た目の相関）にご注意
- 相関係数の落とし穴（3）…… 相関関係は因果関係ではない

では、基本の基本である相関係数の計算方法から見ていきます。サンプルファイルの利用についての説明の後、本編に進みましょう。

サンプルファイルの利用について

本稿では、表計算ソフトを使って手を動かしながら学んでいきます。表計算ソフト Microsoft Excel 用の .xlsx ファイルをダウンロードできるようにしています。デスクトップ版の Excel が手元にない場合は、Microsoft アカウントで使える無料の Microsoft 365 オンライン、もしくは Google アカウントで使える無料の Google スプレッドシート（Google Sheets）をお使いください。Microsoft 365 オンラインの場合は、.xlsx ファイルを OneDrive にアップロードしてから開いてください。Google スプレッドシートの場合は、ファイルを共有して参照できるようにします。リンクを開き、メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。

相関係数の基本をざっとおさらい

相関係数についてはこの連載の中で既に少し触れていますし、初歩の初歩なのでご存じの方も多いかと思いますが、理解を確実にするために、おさらいから始めましょう。相関係数は 2 つの変数にどの程度**直線的な**関係があるかを表す値です（図 2）。

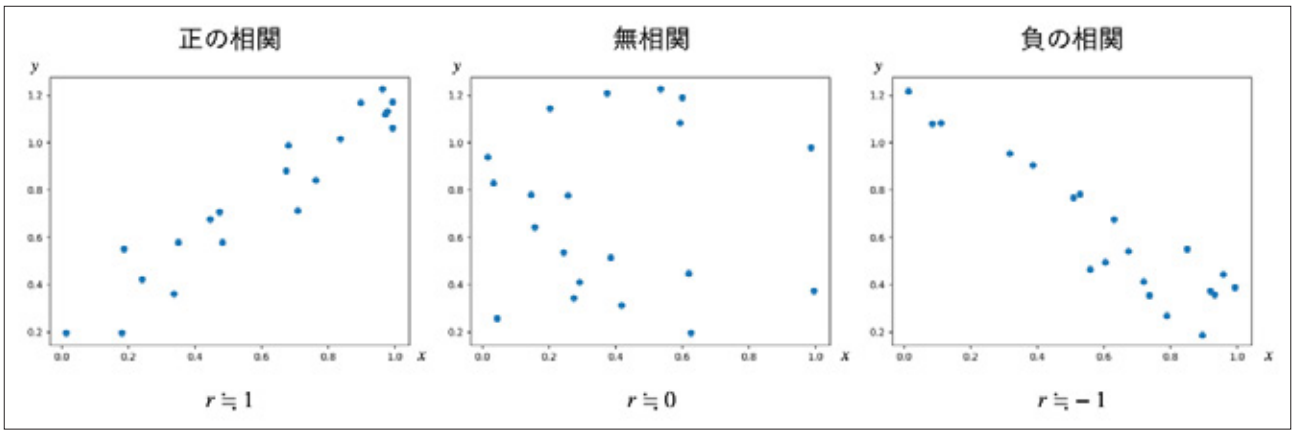


図 2　相関係数と変数同士の関係の強さ

相関係数は r という文字で表されることも多い。r が **1** に近い場合は、図の左側のように、変数 x の値が増えればそれに対する変数 y の値も増える（正の相関）、逆に、r が **-1** に近い場合は、図の右側のように、変数 x の値が増えればそれに対する変数 y の値は減る。r が **0** に近い場合は、図の中央のように、変数 x の値の大小と変数 y には直線的な関係はない（無相関）。

明確な基準はありませんが、一般に、相関係数 r の値は以下のように評価されます。

- $0 \leq |r| \leq 0.2$ …… ほぼ相関はない
- $0.2 < |r| \leq 0.4$ …… 弱い相関がある
- $0.4 < |r| \leq 0.7$ …… 相関がある
- $0.7 < |r| \leq 1$ …… 強い相関がある

相関係数はあくまでも変数同士の**直線的な**関係の強さを表すので、相関係数が 0 に近くても（直線的ではないが）何らかの関係があることもあります。また、複数のグループが混在している場合、それぞれのグループに分けると相関関係が見られることもあります。図 2 の中央の図も、y の値が **1.0** 以上のグループと **1.0** 未満のグループのデータが混在しているのかもしれません。y の値が小さいグループには負の相関がありそうに見えますね。このことについては、また後述することにします。

相関係数の値が **1** あるいは **-1** に近い場合は、「直線的な関係がある」と言えますが、直線の傾きを表すものではありません。相関係数はあくまで関係の強弱を表すもので、変数同士がどれだけ「同じ方向に動く」かということを表します（次の項で説明する相関係数の図形的な意味を見ればよく分かります）。

では、相関係数を求めてみましょう。相関係数を求めるには、Excel の CORREL 関数 *1 が使えます。

*1　この連載では、通常、記事の末尾にある「関数リファレンス」セクションで Excel 関数の書式を詳しく紹介しています。しかし、今回の記事で扱う Excel 関数は、全て過去の記事で解説しているため、「関数リファレンス」セクションは掲載しません。今回取り上げる各 Excel 関数の書式については、初出時の関数に設定しているリンク先（ジャンプ先は過去記事の「関数リファレンス」）を参照してください。

CORREL 関数についても、既にご存じの方も多いと思います。簡単な例（架空の事例）を使って関数の使い方を確認しておきましょう（図 3）。サンプルファイル（13a.xlsx）をこちらからダウンロードし、[相関係数を求める]ワークシートを開いて取り組んでみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。

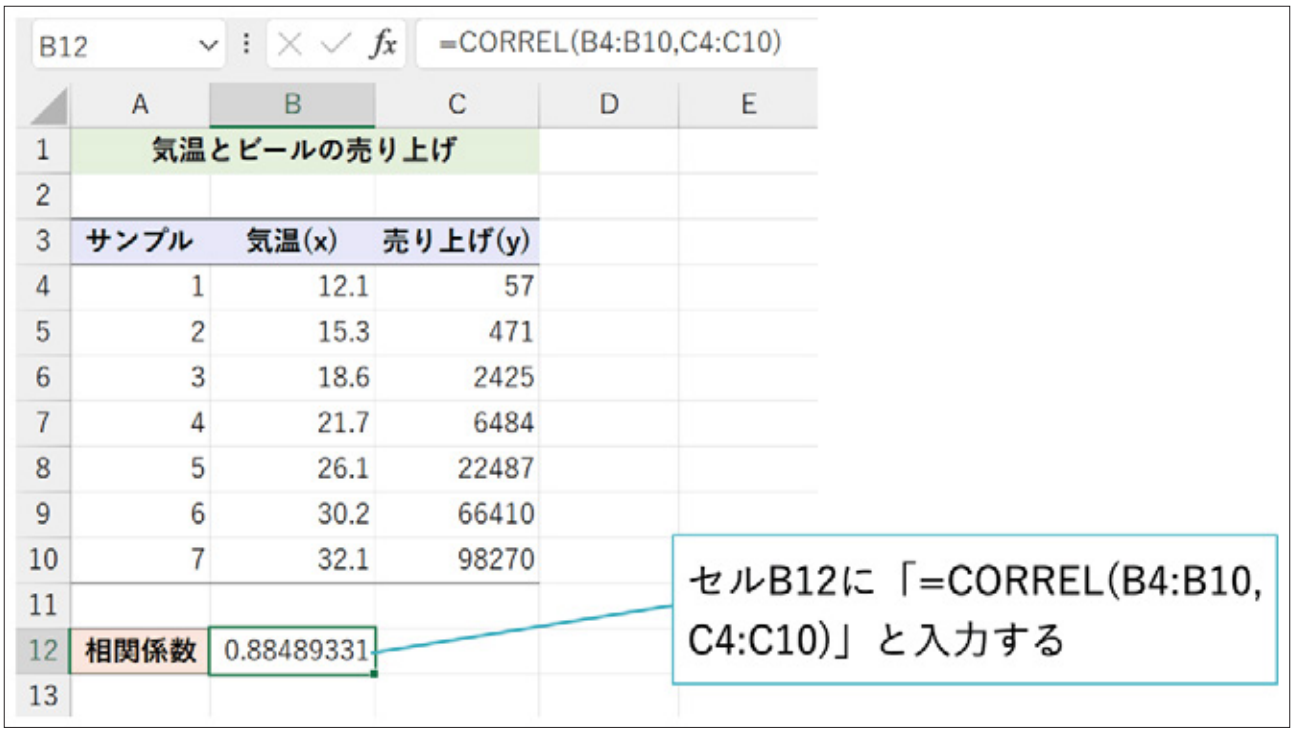


図 3　CORREL 関数を使って相関係数を求める

CORREL 関数には変数 x の範囲と、それに対する変数 y の範囲を指定すればよい。具体的には＝CORREL(B4:B10,C4:C10) と入力すればよい。この例では、0.88... という値が求められ、強い正の相関があることが分かる。つまり、気温が上がればビールの売り上げも上がるということになる。

実は、図 3 の例はさらに分析の余地があるのですが、それについては後述することにします。

実際には、ビールの売り上げはこのデータのように極端ではありません。夏場がピークであることは確かですが、春先から徐々に増えていきます。また、気温の上がる時期だけでなく、12 月にも売り上げが多いようです（アサヒグループホールディングスのファクトブックなどによる）。

相関係数の計算方法を図形的に見てみよう ～ 相関係数の意味がよく分かる！

CORREL 関数が入力できれば、相関係数は簡単に求められます。しかし、完全に「ブラックボックス」ですね。いったいどういう根拠で、どういう計算が行われているのかが全く分かりません。実用上、苦労することはないかもしれませんが、やはり理屈を理解した上で使いたいものです。

そこで、相関係数の意味を図形的に理解できるようにしてみましょう。まず、相関係数の定義を掲載します。数式が登場するので、数式の苦手な方はちょっと腰が引けてしまうかもしれませんが、ご心配なく。単純な四則演算レベルの計算しか出てきませんし、重要なのは分子だけです。なお、以下の解説と定義通りに計算する方法については動画も用意してあります。相関係数の図形的な意味と計算方法を一つ一つ丁寧に追いかけたい方はぜひご視聴ください。

$$r = \frac{\sum\left((x_i - \bar{x})(y_i - \bar{y})\right)}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \qquad (1)$$

（1）式の意味と計算の手順は図 4 のようにまとめられます。分子と分母に分けて説明していますが、図形的に理解する上では、分母は完全に無視してもらって構いません（分母は相関係数の値が **-1** ～ **1** の範囲になるように調整しているだけです）。

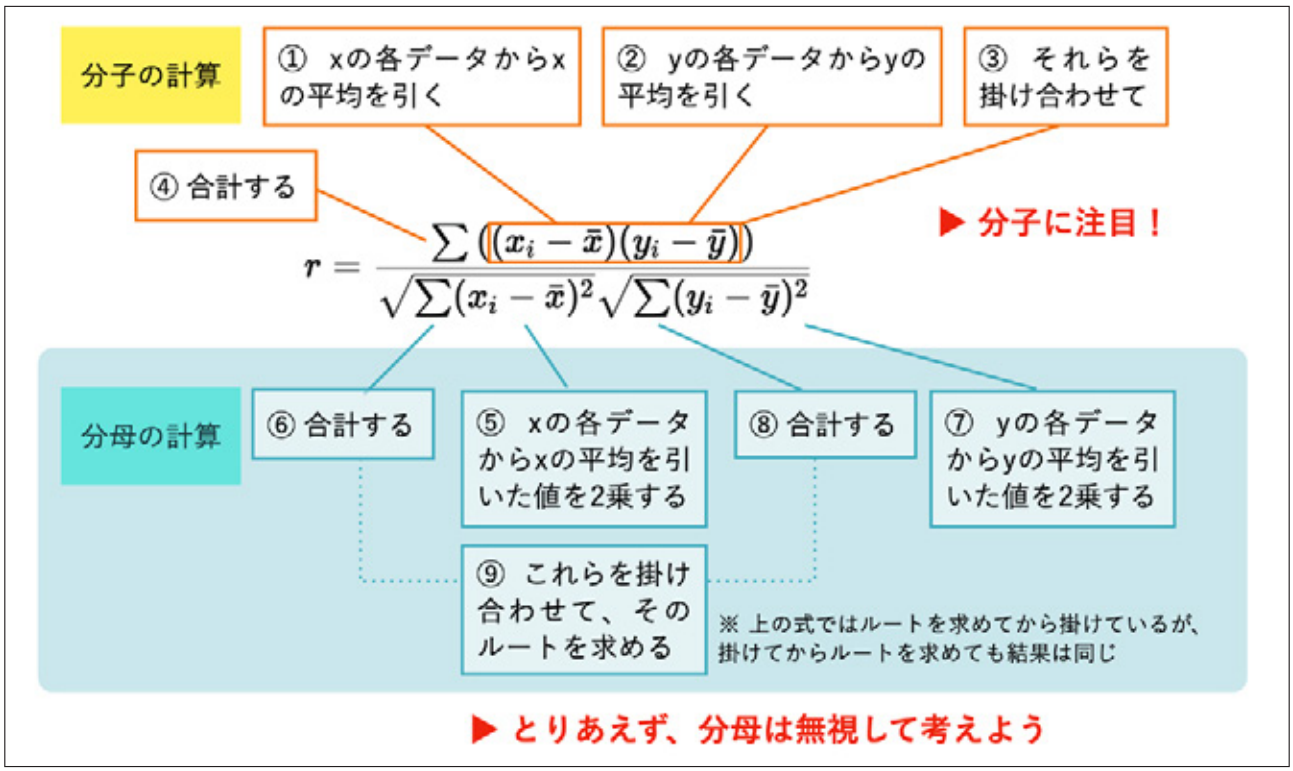


図 4　相関係数を求めるための数式とその計算方法

x_iは一方の変数の各データで、y_iはもう一方の変数の各データ。
$\bar{x}$と$\bar{y}$はそれぞれの平均値。$\sum$は合計するという記号。
分子が$(x_i - \bar{x})$と$(y_i - \bar{y})$の積

になっていることに注目して、以下の解説を読み進めよう。

分子の計算では各データの値から平均値を引いています。これは、原点（0, 0）を平均値の位置に移動しているのと同じことです。つまり、平均値が原点（0, 0）になります。そして、それらの積 $(x_i - \bar{x})(y_i - \bar{y})$ を求めています。その計算を図として表し、意味を考えてみましょう（図 5）。

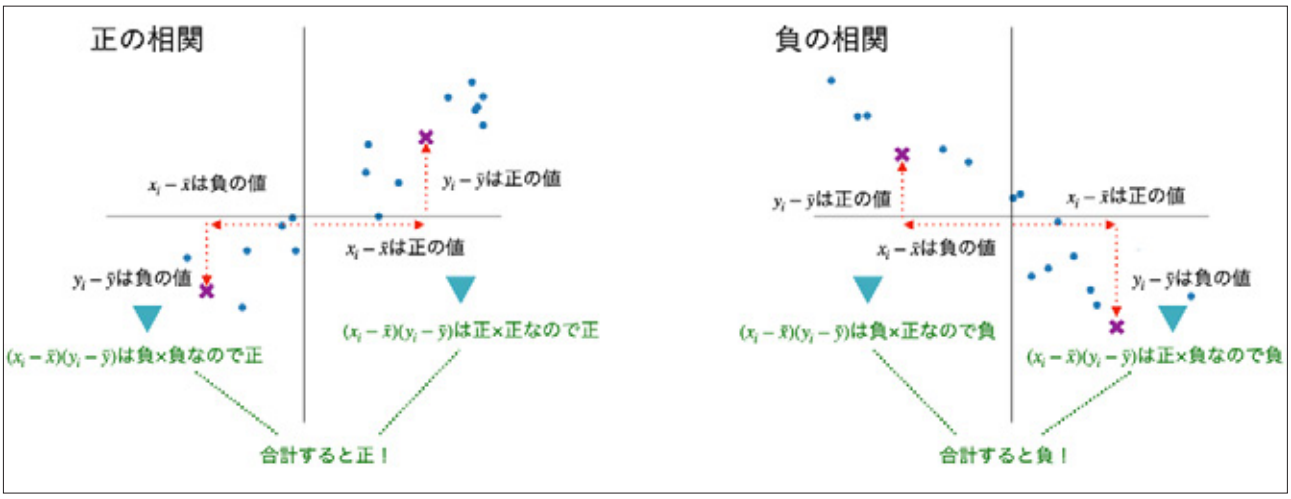


図 5　相関係数の図形的な意味

正の相関の例では平均値が原点になるので、原点を通るような右肩上がりのグラフになる。このとき、右上の × 点の X 軸の値と Y 軸の値を乗算すると、正 × 正なので正の値になる。左下の × 点は、負 × 負なのでやはり正の値になる。それらを全て合計すれば正の値になる。
負の相関の例では同様に原点を通るような右肩下がりのグラフになる。このとき、右下の × 点の X 軸の値と Y 軸の値を乗算すると、正 × 負なので負の値になる。左上の × 点は負 × 正なのでやはり負の値になる。それらを全て合計すれば負の値になる。
なお、無相関の場合は、正と負が相殺されて **0** に近い値になる。

図 5 からも分かるように、

$$\sum (x_i - \bar{x})(y_i - \bar{y})$$

の値が正になる場合が正の相関、負になる場合が負の相関です。

いかがでしょう。そういうことだったのか、と納得していただけたでしょうか。

→目次に戻る

ちなみに、（1）式の分子は、

$$\text{ベクトル}(x_i - \bar{x})\text{とベクトル}(y_i - \bar{y})\text{の内積}$$

です。分母は、

$$\text{それぞれのベクトルの大きさ}|x_i - \bar{x}|\text{と}|y_i - \bar{y}|\text{の積}$$

です。三角関数の以下の公式と全く同じですね。

$$\cos\theta = \frac{X \cdot Y}{|X||Y|} = \frac{\sum(X_i Y_i)}{\sqrt{\sum X_i^2}\sqrt{\sum Y_i^2}}$$

X_i の部分に、

$$(x_i - \bar{x})$$

を代入し、Y_i の部分に、

$$(y_i - \bar{y})$$

を代入すれば、（1）式と同じになります。つまり、相関係数とは、

$$\text{ベクトル}(x_i - \bar{x})\text{とベクトル}(y_i - \bar{y})\text{のなす角度の}\cos\theta\text{の値}$$

なのです。ベクトルの向きが同じであれば、$\theta = 0$ なので、$\cos\theta = 1$ （正の相関）です。ベクトルの向きが逆であれば、$\theta = \pi$ なので、$\cos\theta = -1$ （負の相関）です。ベクトルが直交するなら $\theta = \pi/2$ なので、$\cos\theta = 0$ （無相関）です。

相関係数を定義通りに計算してみよう

図 4 で見た（1）式の意味と計算の手順に合わせて数式を入力すれば、相関係数の計算ができそうですね。実際にやってみましょう。……といっても、相関係数の意味が十分に理解できていれば、CORREL 関数を使えばいいので、ここからの計算は絶対に必要というわけではありません。先を急ぐ方は、次の項「相関係数の落とし穴（1）」に進んでいただいてけっこうです。

図 6 は完成例です。上で見たサンプルファイルの［相関係数の計算］ワークシートを開き、表 1（図 6 の後に掲載しています）に従って数式を入力していきましょう。Google スプレッドシートの例については後述します。

	A	B	C	D	E	F	G	H	I
1	相関係数の考え方を理解するためのサンプル（気温とビールの売り上げ）								
2									
3	サンプル	気温(x)	売り上げ(y)	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	
4	1	12.1	57	-10.2	-28029.3	285898.7143	104.04	7.86E+08	
5	2	15.3	471	-7	-27615.3	193307	49	7.63E+08	
6	3	18.6	2425	-3.7	-25661.3	94946.75714	13.69	6.59E+08	
7	4	21.7	6484	-0.6	-21602.3	12961.37143	0.36	4.67E+08	
8	5	26.1	22487	3.8	-5599.29	-21277.28571	14.44	31352001	
9	6	30.2	66410	7.9	38323.71	302757.3429	62.41	1.47E+09	
10	7	32.1	98270	9.8	70183.71	687800.4	96.04	4.93E+09	
11	平均	22.3	28086.2857		合計Σ	1556394.3	339.98	9.1E+09	
12							相関係数	0.884893	
13							検算	0.884893	
14									

図 6　相関係数を定義通りに求めてみる

（1）式の定義に従って計算してみた例。入力すべき数式や関数は以下の箇条書きと表 1 にまとめてあるので、サンプルファイルに入力して計算の手順を確認しておくとよい。「ブラックボックス」の中身（相関係数の意味と計算の方法）が理解できていれば、いちいち手順通りに計算しなくても、CORREL 関数を使えばよい。

→目次に戻る

各セルに入力されている数式や関数をコピーしやすいように箇条書きで示しておきます。なお、従来の Excel（2019 以前のバージョン）で =B4:B10-B11 のような配列数式を入力するには、入力したい範囲のセルを選択した状態で先頭のセルに配列数式を入力し、入力終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す必要があります。

- セル **B11**：=AVERAGE(B4:B10)
- セル **C11**：=AVERAGE(C4:C10)
- セル **D4**：=B4:B10-B11
- セル **E4**：=C4:C10-C11
- セル **F4**：=D4:D10*E4:E10
- セル **F11**：=SUM(F4:F10)
- セル **G4**：=D4:D10^2
- セル **G11**：=SUM(G4:G10)
- セル **H4**：=E4:E10^2
- セル **H11**：=SUM(H4:H10)
- セル **H12**：=F11/SQRT(G11*H11)
- セル **H13**：=CORREL(B4:B10,C4:C10)

表 1 では上記の内容を一覧としてまとめ、数式の意味や計算結果を「備考」に記しています。

（1）式の分子

セル	入力されている数式や関数	備考
B11	=AVERAGE(B4:B10)	$\bar{x}$の値
C11	=AVERAGE(C4:C10)	$\bar{y}$の値
D4	=B4:B10-B11	$x_i - \bar{x}$の全ての値がセルD4～D10に求められる*
E4	=C4:C10-C11	$y_i - \bar{y}$の全ての値がセルE4～E10に求められる*
F4	=D4:D10*E4:E10	$(x_i - \bar{x})(y_i - \bar{y})$の全ての値がセルF4～F10に求められる*
F11	=SUM(F4:F10)	$\sum((x_i - \bar{x})(y_i - \bar{y}))$の値 …… **分子の値**

（1）式の分母と相関係数

セル	入力されている数式や関数	備考
G4	=D4:D10^2	$(x_i - \bar{x})^2$の全ての値がセルG4～G10に求められる*
G11	=SUM(G4:G10)	$\sum(x_i - \bar{x})^2$の値
H4	=E4:E10^2	$(y_i - \bar{y})^2$の全ての値がセルH4～H10に求められる*
H11	=SUM(H4:H10)	$\sum(y_i - \bar{y})^2$の値
H12	=F11/SQRT(G11*H11)	相関係数の値（F11は分子の値、SQRT(G11*H11)は分母の値）
H13	=CORREL(B4:B10,C4:C10)	CORREL関数を使った検算結果

表 1　図 6 で入力した数式や関数の一覧

この表の通りに入力していけば、（1）式での計算方法が理解できる。Google スプレッドシートでは、* で示した箇所の数式は、ARRAYFORMULA 関数の引数に指定する必要がある（後述）。

上の例ではスピル機能を使って複数の計算を一度に行っていますが、Excel のスピル機能を使わずに計算した例もサンプルファイルの［相関係数の計算（スピルを使わない）］ワークシートに含めてあります。その場合、例えば

$$\text{セル}\mathbf{D4}\sim\mathbf{D10}\text{の}x_i - \bar{x}\text{の全ての値}$$

を求めるのであれば、セル **D4** に =B4-$B$11 と入力し、セル **D4** をセル **D10** までコピーします。サンプルファイルで確認してみてください。

Google スプレッドシートでは、Excel のスピル機能や配列数式に相当する機能が ARRAYFORMULA という関数で提供されています。例えば、セル **D4** に入力する数式であれば =B4:B10-B11 の代わりに =ARRAYFORMULA(B4:B10-B11) と入力します。詳細については、こちらのサンプルファイルを開き、メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてご参照ください。

相関係数の落とし穴（1）～ 相関係数が 0 に近くても関係が強いこともある

冒頭で紹介した気温と CO_2 排出量の例を思い出してください。図 1 の下の部分を再掲します。気温と電気による CO_2 の排出量については、相関はないようです。一方、ガスや灯油なども含めると、負の相関が見られます。つまり、気温に関わらず電気による CO_2 排出量は変わらず、全体で見ると気温が下がるほど CO_2 の排出量が増えるというわけです。夏や冬には電気の消費量が増えそうなので、この結果は直感に反しますね。

サンプルファイル（13b.xlsx）はこちらからダウンロードできます。［気温と CO_2 排出量 (散布図付き)］ワークシートを開くと図 7 の内容が確認できます。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。なお、散布図の作成方法についてはここでは触れません。前回の「散布図を徹底活用して「関係」を可視化 ～ 関係と規模を一度に見る」をご参照ください。

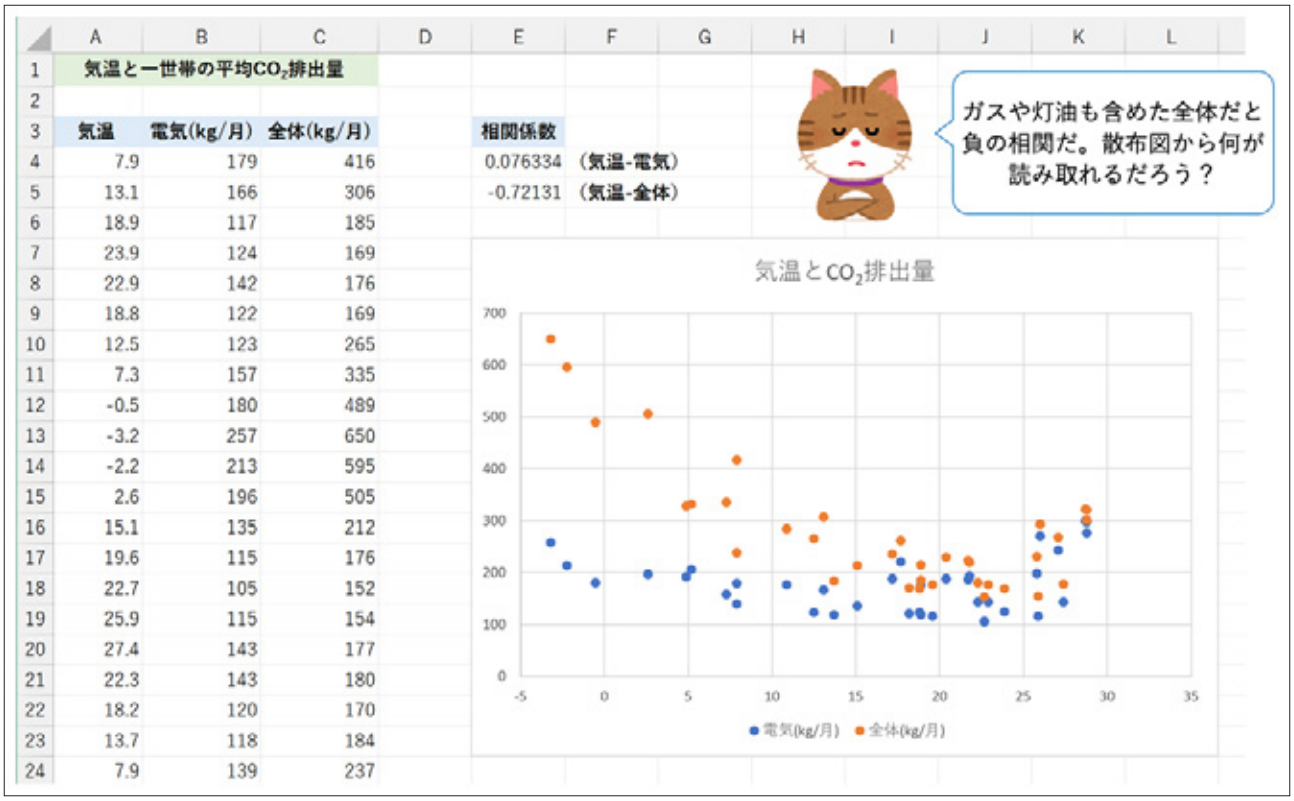


図 7　気温と CO2 排出量の関係

セル **E4** には =CORREL(A4:A39,B4:B39) が入力されており、セル **E5** には =CORREL(A4:A39,C4:C39) が入力されている。気温と電気による CO_2 排出量の相関係数は **0** に近い。つまり直線的な関係はない（無相関）。しかし、気温と（灯油やガスも含めた）全体のエネルギーによる CO_2 排出量の相関係数は **-1** に近い（負の相関）。これはどういうことを表しているのだろうか。

まず、全体の CO_2 排出量について考えてみましょう。これは簡単です。気温があまりに低いと電気での暖房が追いつかないからです。その場合は灯油などの消費が多くなるので、全体の CO_2 排出量が多いというわけです。気温が高いとき（図の右側）は気温と CO_2 排出量はやや比例しているようですが、エアコンがフル稼働し、灯油などは使われないので、電気による CO_2 排出量と全体の CO_2 排出量にはさほど差がありません。

次に、気温と電気による CO_2 排出量の関係です。相関係数を見ると「関係がない」と言えそうですが、実際のところ、夏と冬にはエアコンも使われるので、春や秋よりも CO_2 排出量が多くなりそうです。確かに、直線的な関係ではありませんが、実はU字形のグラフが描けるような関係になっています。そこで、サンプルファイルの［気温と CO_2 排出量 (電気のみ)］ワークシートを開き、気温と電気だけの散布図を見てみましょう（図 8）。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	気温と一世帯の平均CO_2排出量				平均気温	16.63056							
2													
3	気温	電気(kg/月)	平均で折り返し		相関係数								
4	7.9	179	-179		0.076334	（気温-電気）							
5	13.1	166	-166		0.910794	（気温-平均で折り返し）							
6	18.9	117	117										
7	23.9	124	124										
8	22.9	142	142										
9	18.8	122	122										
10	12.5	123	-123										
11	7.3	157	-157										
12	-0.5	180	-180										
13	-3.2	257	-257										
14	-2.2	213	-213										
15	2.6	196	-196										
16	15.1	135	-135										
17	19.6	115	115										
18	22.7	105	105										
19	25.9	115	115										
20	27.4	143	143										
21	22.3	143	143										
22	18.2	120	120										
23	13.7	118	-118										
24	7.9	139	-139										

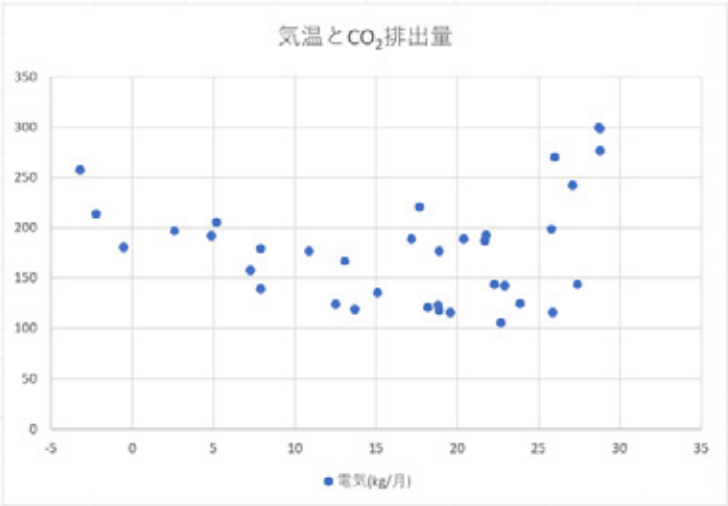


図 8　気温と電気による CO_2 排出量の関係

図 7 では、全体の CO_2 排出量が多かったので、電気の CO_2 排出量があまり目立たなかったが、U字形の分布になっているように見える。「平均で折り返し」という項目については以下で解説する。

どうやら、気温と電気による CO_2 排出量は直線的な関係ではなくU字形の関係があるようです。そこで、ちょっと乱暴ですが、気温が平均よりも低い場合には、CO_2 排出量にマイナスを付けてみましょう。そうすれば、気温の平均の位置で値が折り返され、U字形が直線に近くなるはずです。

セル **C4** には、`=IF(A4:A39<F1,-B4:B39,B4:B39)` という式が入力されています。スピル機能を使わないのであれば、セル **C4** に `=IF(A4<$F$1, -B4, B4)` と入力し、セル **C39** までコピーすれば同じ結果になります。Google スプレッドシートの場合は、`=ARRAYFORMULA(IF(A4:A39<F1,-B4:B39,B4:B39))` と入力します。

相関係数は、図 7 で見たのと同じ式で求められます。セル **E5** の値は **0.91**... とかなり **1** に近い値です。つまり、正の相関になります。気温が低いときと高いときに電気による CO_2 排出量が大きくなるという直感は間違っていなかったようです。

なお、図 3 で見た気温とビールの売り上げについても「**1** に近い値が求められたので正の相関がある」と安心してしまわずに散布図を描くようにしましょう。直線的ではない関係があるかもしれないからです。散布図は図 9 のようになります。こちらは、最初に利用したサンプルファイル（13a.xlsx）の［気温とビールの売り上げ］ワークシートに含まれています。

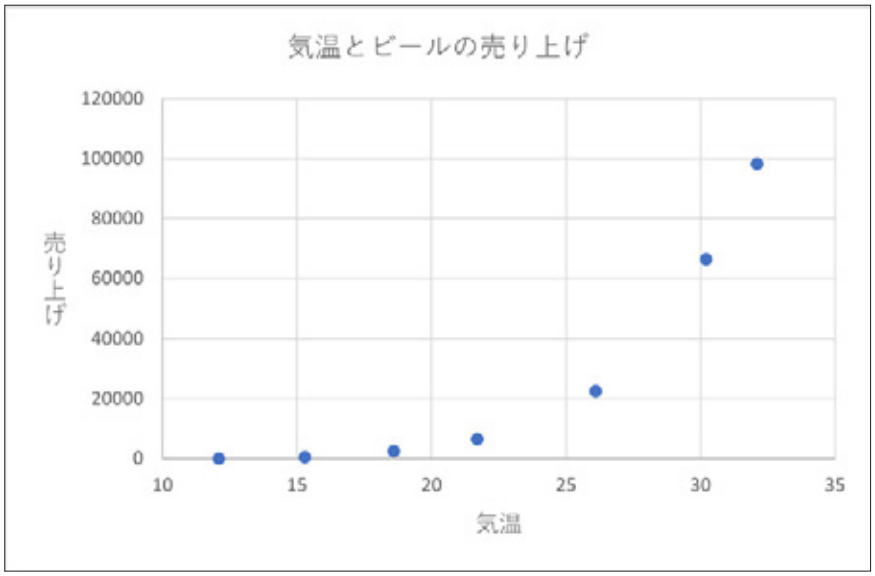
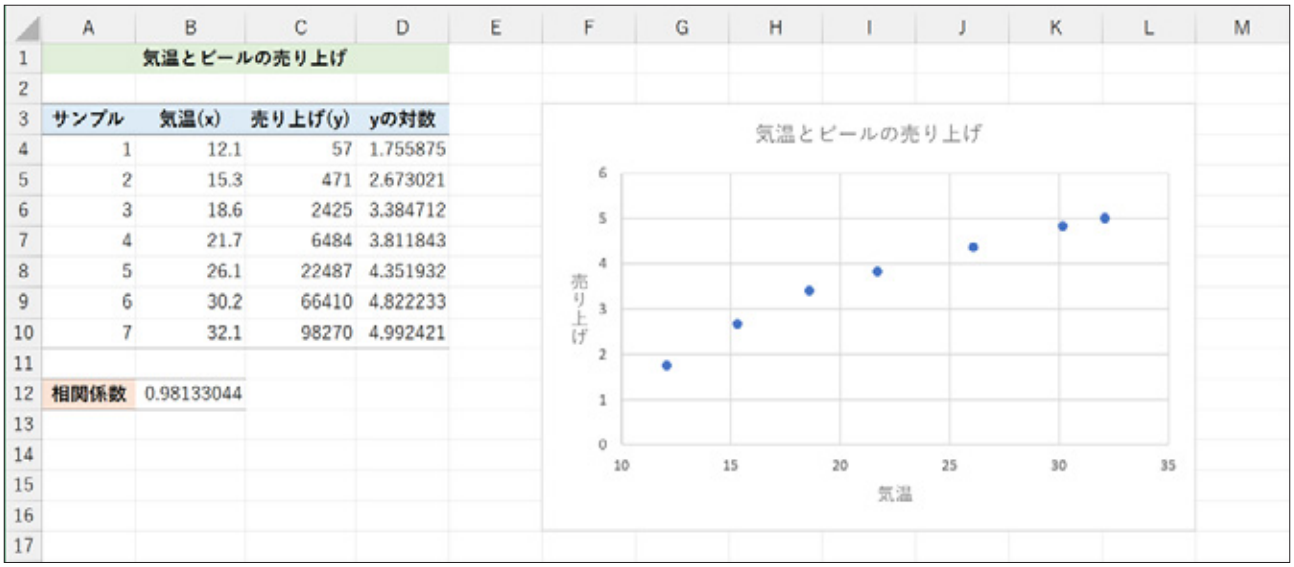


図 9　気温とビールの売り上げを散布図にしてみる

図 3 の気温とビールの売り上げの散布図を作成してみると、直線的な関係ではなく指数関数的な関係であることが分かる。

図 9 から、気温とビールの売り上げには指数関数的な関係がありそうだということが分かります。ということは、対数を取れば直線的な関係になるはずです。そこで、売り上げの対数を取った値を基に散布図を作成し、さらに相関係数も求めてみます。サンプルファイル（13a.xlsx）の［気温とビールの売り上げ（対数）］を開くと、図 10 のような画面が表示されます。



	A	B	C	D
1	気温とビールの売り上げ			
2				
3	サンプル	気温(x)	売り上げ(y)	yの対数
4	1	12.1	57	1.755875
5	2	15.3	471	2.673021
6	3	18.6	2425	3.384712
7	4	21.7	6484	3.811843
8	5	26.1	22487	4.351932
9	6	30.2	66410	4.822233
10	7	32.1	98270	4.992421
11				
12	相関係数	0.98133044		

図 10　気温とビールの売り上げの対数

散布図を描くと、やや弓なりになっているが、図 9 に比べると直線的であることが分かる。セル **D4** には =LOG(C4:C10) が入力されており、売り上げの対数が求められている（底は **10**）。セル **B12** には =CORREL(B4:B10,D4:D10) と入力されており、気温と売り上げの対数の相関係数が求められている。元の値で求めたときよりも、さらに **1** に近い値になっていることが分かる。

気温とビールの売り上げのデータは、あえてこのような結果になるように作成した架空のデータですが、表面的な数値だけに頼らず、データをさまざまな角度から見ることの大切さはご理解いただけたと思います。

相関係数の落とし穴（2）～ 疑似相関にご注意

相関係数に関連する落とし穴の一つとして、疑似相関がよく知られています。架空の話ですが、ビールの売り上げと水難事故に正の相関が見られたとします。ビールが売れると水難事故が起こるのでしょうか。これは**見た目の相関（疑似相関）**です。

ビールが売れるのは一般に夏場の暑い時期です。その時期には海や川で水遊びする人が多いので水難事故が多くなるのは当然と言えば当然です。真相は、気温とビールの売り上げ、気温と水難事故に相関があるという話だったわけです（図 11）。

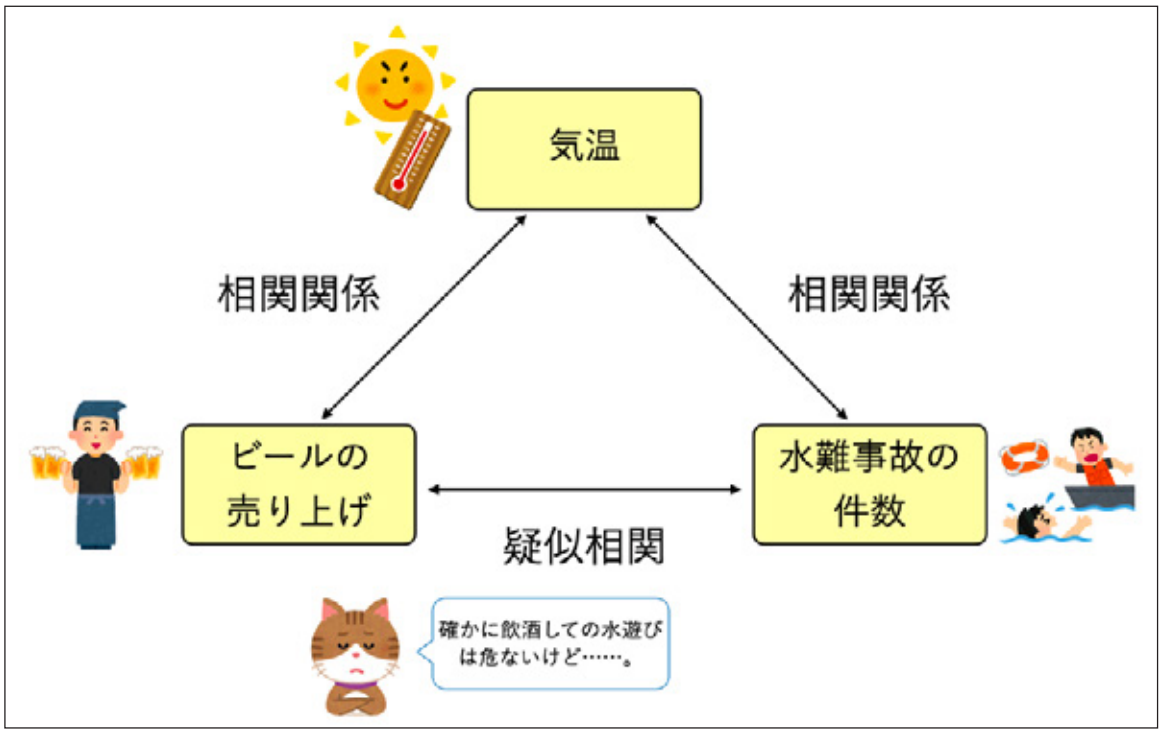


図 11　疑似相関の例

この例はいかにも疑似相関だと分かるが、実際には疑似相関だと気付かない場合もある。データの背景を考えることも重要。

また、単なる偶然やデータ数の少なさが原因ということもあります。ニコラス・ケイジの映画出演数とプールで溺死する人の数に正の相関があるといった例などが有名です。

相関係数の落とし穴（3）～ 相関関係は必ずしも因果関係ではない

相関係数のもう一つの落とし穴は、相関関係を**因果関係**（＝原因と結果の関係）と誤解してしまうことです。

気温とビールの売り上げについては、因果関係と考えてもよさそうですが、そうではない（むしろ因果関係ではない）場合もよくあります。気温が上がる（原因）とビールが売れる（結果）ことはあるでしょうが、ビールが売れる（原因）と気温が上がる（結果）というのは現実的ではないので、気温とビールの売り上げについては、因果関係の可能性があると解釈してもよさそうです。

上でも触れましたが、実際にはビールの売り上げは 12 月にも上がります（おそらく忘年会シーズンのため）。気温はビールの売り上げが上がる原因の一つですが、唯一無二の原因でもありません。他の要因があることも考慮する必要があります。

しかし、相関係数が高くても「因果関係の可能性を示唆する」とは解釈できない事例も数多くあります。例えば、ゲームに費やす時間と成績に負の相関がある場合、ゲームにのめり込んでしまうから勉強をしなくなるのか、勉強するのが嫌だからゲームに逃避しているのかは分かりません（図 12）。安易に因果関係と見なすのは危険です。やはり背景を調べることが重要です。

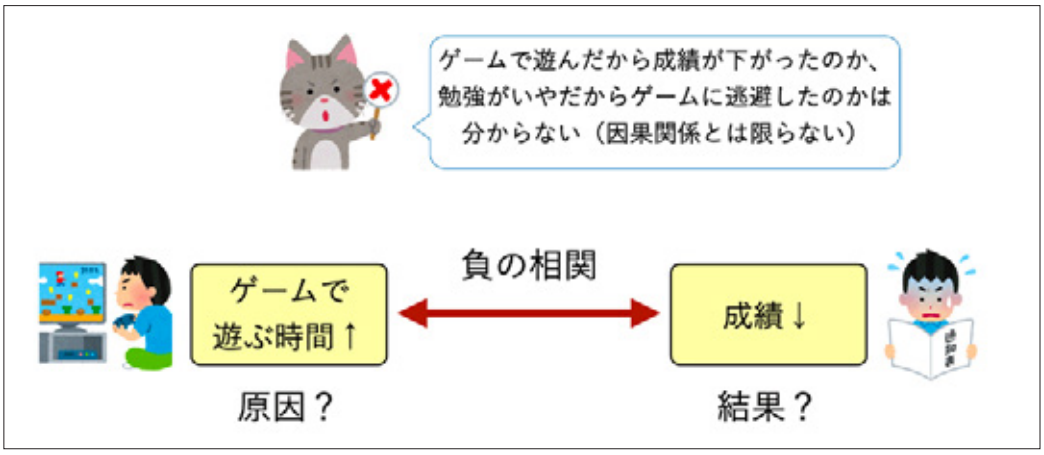


図 12　相関関係は必ずしも因果関係ではない

人は 2 つの出来事が同時に起こると因果関係だと思う傾向（認知バイアス）があるので、相関関係は因果関係と混同されやすい。しかし、相関関係は必ずしも因果関係ではないことに注意。

発展：順序尺度や名義尺度で、関係を表す数値を求めるには

分布に偏りがある場合には、間隔尺度のデータでも、元の値ではなく順位を使って代表値（中央値）や散布度（四分位範囲）を求めることがあります。相関係数についても同様に、元の値ではなく順位を使って求めた「**順位相関**」の方が適切です。また、元のデータが順位を表す値である場合（順序尺度の場合）にも順位相関を使います。

一方、名義尺度の場合は**クラメールの連関係数**などを使います。

以降の例については、作成例をこちら（13c.xlsx）に用意してあるので、解説と併せてご参照ください。Google スプレッドシートのサンプルはこちらです。

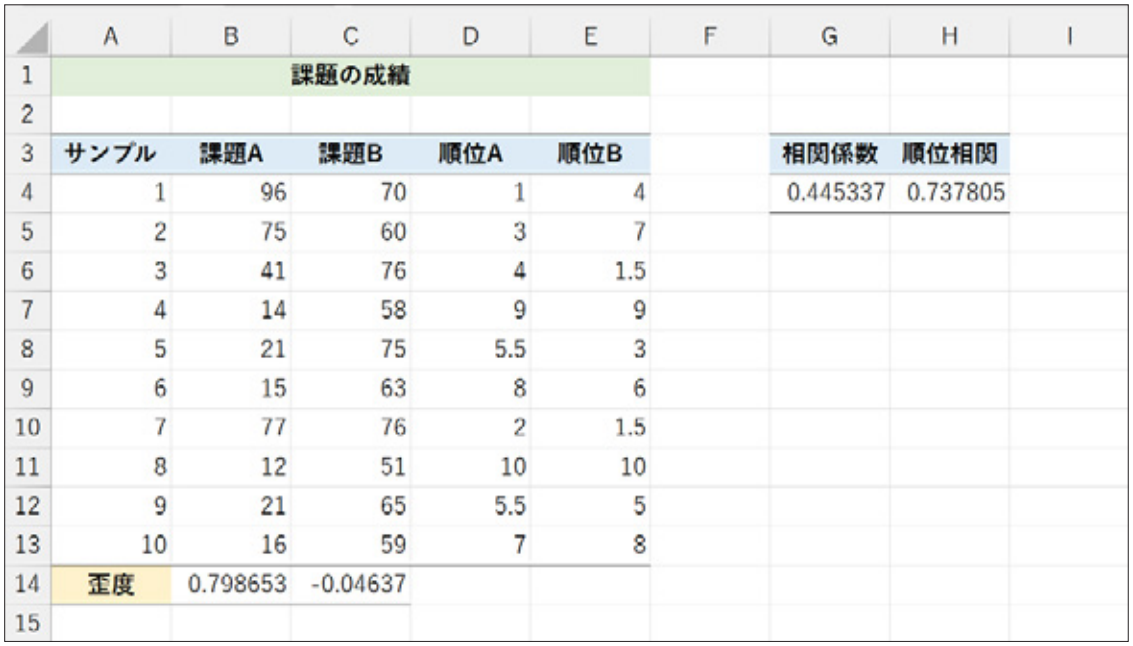
これまでに見てきた相関係数は**ピアソンの積率相関**と呼ばれるもので、間隔尺度のデータに対して使われるものです。

スピアマンの順位相関

順位相関には**スピアマンの順位相関**や**ケンドールの順位相関**がありますが、ここではスピアマンの順位相関を紹介します。スピアマンの順位相関 r_s は以下の式で表されます。

$$r_s = 1 - \frac{(6 \times \sum (x_i - y_i)^2)}{n(n^2 - 1)}$$

といっても、実はこの式で計算しなくても、順位を表す値を CORREL 関数に指定すれば、同じ値が求められます。なお、同じ順位がある場合には、順位を RANK.AVG 関数で求めて CORREL 関数で相関係数を求めます。［順位相関を求める（同順位がある場合）］ワークシートを開くと図 13 の表が参照できます。なお、以下の操作と、次項のクラメールの連関係数を求める操作については動画も用意してあります。数式を入力する手順を一つ一つ丁寧に追いかけたい方はぜひご視聴ください。



	A	B	C	D	E	F	G	H	I
1	課題の成績								
2									
3	サンプル	課題A	課題B	順位A	順位B		相関係数	順位相関	
4	1	96	70	1	4		0.445337	0.737805	
5	2	75	60	3	7				
6	3	41	76	4	1.5				
7	4	14	58	9	9				
8	5	21	75	5.5	3				
9	6	15	63	8	6				
10	7	77	76	2	1.5				
11	8	12	51	10	10				
12	9	21	65	5.5	5				
13	10	16	59	7	8				
14	歪度	0.798653	-0.04637						
15									

図 13　順位相関を求める

10 人に課題 A と課題 B を与えたときの成績（架空データ）。歪度（わいど）を見ると課題 A の成績は小さい値が多いことが分かる。分布に歪みがあるので、順位相関を使うとよい。元のデータで求めた相関係数の値は比較的小さいが、順位相関の値はかなり大きい。歪度については本連載の第 4 回「分散／標準偏差 ～ 給与の格差ってどれぐらい?」のコラム「歪度と尖度で分布の形を知る」も参照のこと。

各セルに入力されている関数は以下の通りです。

セル	入力されている数式	備考
B14	=SKEW(B4:B13)	課題Aの歪度
C14	=SKEW(C4:C13)	課題Bの歪度
D4	=RANK.AVG(B4:B13,B4:B13,0)	課題Aの順位（同順位は順位の平均）
E4	=RANK.AVG(C4:C13,C4:C13,0)	課題Bの順位（同順位は順位の平均）
G4	=CORREL(B4:B13,C4:C13)	相関係数の値
H4	=CORREL(D4:D13,E4:E13)	順位相関の値

表 2　図 13 で入力した数式の一覧

スピル機能を使わない場合は、セル **D4** に =RANK.AVG(B4,$B$4:$B$13,0) を、セル **E4** に =RANK.AVG(C4,$C$4:$C$13,0) を入力し、セル **D4** と **E4** をセル **13** 行目までコピーする。Google スプレッドシートでは、セル **D4** に =ARRAYFORMULA(RANK.AVG(B4:B13,B4:B13,0)) を、セル **E4** に =ARRAYFORMULA(RANK.AVG(C4:C13,C4:C13,0)) と入力する。

クラメールの連関係数

名義尺度の変数同士の場合、**クラメールの連関係数**と呼ばれる値で関係の強さを求めることができます。連関係数は元のデータと、期待値（偏りがないと考えた場合の値）とのズレを表すような値です。ズレが大きいということは偏りがあるということです。連関係数はそれほど大きな値にならないので、一般に **0.1** 以上であれば何らかの関係があると言われています。図 14 の例は［連関係数］ワークシートに含まれています。

	A	B	C	D	E	F	G	H	I	J	K	L
1		出身地とお雑煮の種類						期待値				
2												
3		出身地	味噌	すまし汁	小豆汁	合計		出身地	味噌	すまし汁	小豆汁	合計
4		北海道	32	42	6	80		北海道	31.64706	41.76471	6.588235	80
5		東北	14	65	3	82		東北	32.43824	42.80882	6.752941	82
6		関東	26	84	7	117		関東	46.28382	61.08088	9.635294	117
7		関西	94	21	4	119		関西	47.075	62.125	9.8	119
8		中国	28	44	17	89		中国	35.20735	46.46324	7.329412	89
9		四国	45	48	8	101		四国	39.95441	52.72794	8.317647	101
10		九州	30	51	11	92		九州	36.39412	48.02941	7.576471	92
11		合計	269	355	56	680		合計	269	355	56	680
12												
13		カイ二乗値を求めるための表										
14												
15		出身地	味噌	すまし汁	小豆汁	合計						
16		北海道	0.003936	0.001326	0.052521	0.05778						
17		東北	10.48049	11.50343	2.085694	24.0696						
18		関東	8.889358	8.599842	0.720764	18.21						
19		関西	46.77548	27.22359	3.432653	77.4317						
20		中国	1.475429	0.130588	12.75959	14.3656						
21		四国	0.637175	0.423939	0.012131	1.07324						
22		九州	1.123389	0.183729	1.546967	2.85409		カテゴリ数	3	（小さい方）		
23		合計	69.3853	48.0664	20.6103	138.062		連関係数	0.318616			
24												

図 14　クラメールの連関係数を求める

出身地と正月のお雑煮の種類についてアンケートを採ったデータ。農林水産省の「全国のいろいろな雑煮」を参考にして作成したものだが、値は架空のもの。沖縄ではお雑煮を食べる習慣があまりないとのことなので、データには含めていない。クラメールの連関係数は **0.32** 程度なので、出身地と正月のお雑煮の種類には偏りがある（出身地によってお雑煮の種類は異なる）と言えそうである。

各セルに入力されている関数は以下の通りです。合計は SUM 関数で求めていますが、簡単なので以下の表では省略しています。

セル	入力されている数式	備考
H4	=E4:E10*B11:D11/E11	期待値の全ての値がセル**H4**～**J10**に求められる
B16	=(B4:D10-H4:J10)^2/H4:J10	「（実測値と期待値のズレ）の2乗÷期待値」の全ての値がセル**B15**～**D20**に求められる
H22	=MIN(COUNTA(B3:D3),COUNTA(A4:A10))	カテゴリ数の小さい方（お雑煮の種類＝**3**）
H23	=SQRT(E23/(E11*(H22-1)))	連関係数の値

表 3　図 14 で入力した数式の一覧

スピル機能を使わない場合は、セル **H4** に =$E4*B$11/$E$11 を入力し、セル **J10** までコピーする。セル **B16** には =(B4-H4)^2/H4 を入力し、セル **D22** までコピーする。Google スプレッドシートでは、セル **H4** に =ARRAYFORMULA(E4:E10*B11:D11/E11) を、セル **B16** に =ARRAYFORMULA((B4:D10-H4:J10)^2/H4:J10) と入力する。

期待値とは、2 つの変数に関係がないと考えられるときの値です。つまり、出身地とは関係なく、お雑煮の種類は一定の割合で、偏りがないと考えられるときの値です。セル **B11** ～ **E11** を見れば、全体では、味噌、すまし汁、小豆汁が、それぞれ **269/680**、**355/680**、**56/680** の割合になっていることが分かります。北海道出身者の数はセル **E4** の 80 人なので、一定の割合であるとすれば、北海道の味噌は 80×269÷680 になるはずです。セルアドレスを使った数式で表すなら =E4*B11/E11 となります。同様に北海道のすまし汁は =E4*C11/E11、小豆汁は =E4*D11/E11 で期待値が求められます。変化する部分を範囲として表せば、=E4:E10*B11:D11/E11 という数式で全ての期待値が求められるわけです。

セル **B16** ～ **D22** に入力されている数式 =(B4:D10-H4:J10)^2/H4:J10 は、実測値（セル **B4** ～ **D10**）と期待値（セル **H4** ～ **J10**）の差を二乗し、それが期待値に対してどれぐらいの割合であるかを求めたものです。つまり、期待値に対するズレの程度を表します。セル **E23** はそれらの合計です。これが**カイ二乗値**と呼ばれる値です。つまり、カイ二乗値は以下のように表されます。

$$\chi^2 = \sum \frac{(\text{実測値} - \text{期待値})^2}{\text{期待値}}$$

連関係数 V は以下の式で求められます。n はデータの個数、k はカテゴリ数の小さい方です。

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

カイ二乗値は**独立性の検定**と呼ばれる統計的検定を行うための基本です。独立性の検定では、カイ二乗値を基に、カイ二乗分布の右側確率を求めます……が、それについては、この連載の続編である推測統計編で取り扱うことになります（ただ、かなり先になりそうですが）。

今回は、相関係数を求めて変数同士の直線的な関係の強さを評価する方法を紹介しました。直線的な関係がなく、相関係数が **0** に近くても、何らかの関係がある場合も考えられます。また、疑似相関に注意することや、相関関係が必ずしも因果関係ではないことについても触れました。さらに、発展的な内容として順序尺度や名義尺度の場合に関係の強さを求める方法についても紹介しました。

次回は、単回帰分析による予測の方法を見ていきます。どうぞお楽しみに！

［データ分析］単回帰分析による予測（線形回帰、指数回帰）～排気量から中古車の価格を予測しよう

データ分析の初歩からステップアップしながら学んでいく連載の第 14 回。既知のデータから未知の値を「予測」する回帰分析の式の可視化や、求め方、実際の予測を、Excel を使って手を動かしながら学んでいきましょう。直線の式だけでなく指数関数の式での予測や時系列分析についても触れます。

羽山博（2024 年 02 月 01 日）

これまでは、集団の性質や変数同士の関係など、何らかの「特徴」を見極める方法を紹介してきました。例えば前回は、相関係数を求めて変数同士の関係を数値で表す方法を紹介しました。

今回からは、既に得られたデータから、未知の値を「予測」することに焦点を当て、回帰分析に取り組みます。回帰分析の方法を紹介するだけでなく、予測の精度を上げるための工夫についても見ていきます。今回取り上げるのは回帰分析による予測の第一歩、単回帰分析です。Excel を使って手を動かしながら単回帰分析にチャレンジしてみましょう。

単回帰分析とは、ある変数 x の値からもう一方の変数 y の値を予測するための式（回帰式）を求めたり、**回帰式**を基に予測したりすることです。例えば、図 1 のように排気量から中古車の本体価格を予測するといった例がそれに当たります。なお、図 1 のデータでは実際のメーカーや車種の名称が使われていますが、本体価格などの値は架空のものです。

図 1　中古車のデータ（ここでは排気量と本体価格に注目する）

中古車の価格はさまざまな要因で決められるが、ここでは排気量と本体価格のみに注目する。取りあえず散布図を描いて回帰式を可視化したものがこの例。実際には、予測の精度を上げるために、相関係数や散布図から得た手掛かりを基に準備（前処理）を行う必要がある。が、最初の一歩として、このまま回帰分析を行ってみよう。

図 1 に表示されている散布図には、各データの近くを通る点線の直線が引かれていますね。その直線を表す式が回帰式です。回帰式を利用すれば排気量から本体価格が予測できるというわけです。

回帰分析では、x に当たる変数を**説明変数**と呼び、y に当たる変数を**目的変数**と呼びます。中古車データの例であれば、排気量が説明変数 x、本体価格が目的変数 y となります。

本来は、図 1 のように相関係数を求めたり、散布図を描いたりしてデータの特徴を見極め、より良い予測のための前処理を行ってから回帰分析に取り組むのが筋です。しかし、そもそも回帰分析の方法を知らないと実感が湧かないでしょうから、まずは、前処理なしで回帰分析を行ってみます。Excel ではグラフを作成したり、幾つかの関数を使ったりするだけで簡単に回帰分析や予測ができてしまうので、先に結果を出してみようというわけです。その後、図 1 の下に記されている問題意識に対処して、前処理を行い、予測の精度を向上させる方法を見ていくことにします。

ここでは、話を単純化するために回帰式を「直線を表す式」としていますが、実は直線でなくても構いません。例えば、x の値が大きくなると y の値が急激に大きくなるような場合、指数関数を回帰式とした方が当てはまりがよくなります（指数回帰）。なお、最後のコラムでは、値の増減に波があるデータについて予測するための時系列分析についても簡単に触れます。

回帰式が直線などの一次関数の場合を**線形回帰**と呼び、回帰式が指数関数の場合を**指数回帰**と呼びます。また、説明変数が 1 つだけ（x のみ）の回帰分析を**単回帰**と呼び、説明変数が複数（x_1, x_2, ...）の場合は**重回帰分析**と呼びます。今回は特に必要がなければ、線形回帰の単回帰分析を「**回帰分析**」と呼ぶことにします。

この記事は、データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 14 回です。第 13 回から第 15 回までは「関係」に注目し、相関係数や回帰分析などについて見ていきます。前回（第 13 回）は、相関係数の意味や求め方に加え、順序尺度や名義尺度での変数同士の関係を見極める方法を紹介しました。今回は、説明変数から目的変数の値を予測するために、単回帰分析の考え方や取り扱いを見ていきます。次回（第 15 回）は、複数の説明変数を利用する重回帰分析や、回帰式が二次以上の式になる多項式回帰に取り組む予定です。連載のトップページから全体の目次が参照できます。

この記事で学べること

今回は以下のようなポイントについて、分析の方法や目の付けどころを見ていきます。

- 回帰式の可視化 …… 散布図に回帰式を表す直線（または曲線）を表示する
- 回帰式の求め方 …… 単回帰分析の回帰式（係数と定数項）を求める
- 回帰式による予測 …… 回帰式に値を代入して予測する
- 回帰分析のための準備 …… 予測の精度を上げるための前処理を行う
- 指数回帰についても理解する …… 直線的でない関係でも回帰分析を行う

では、散布図を作成して回帰式を可視化した後、基本の基本である回帰式の係数と定数項の求め方に進みます。サンプルファイルの利用についての説明の後、本編に進みましょう。

サンプルファイルの利用について

本稿では、表計算ソフトを使って手を動かしながら学んでいきます。表計算ソフト Microsoft Excel 用の .xlsx ファイルをダウンロードできるようにしています。デスクトップ版の Excel が手元にない場合は、Microsoft アカウントで使える無料の Microsoft 365 オンライン、もしくは Google アカウントで使える無料の Google スプレッドシート（Google Sheets）をお使いください。Microsoft 365 オンラインの場合は、.xlsx ファイルを OneDrive にアップロードしてから開いてください。Google スプレッドシートの場合は、ファイルを共有して参照できるようにします。リンクを開き、メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。

視覚的に回帰分析を理解しよう ～ 回帰式の可視化

既に述べたように、回帰分析とは、各データの最も近くを通る回帰式を求めたり、その式を基に予測したりすることです。ここでは、線形回帰に取り組むので、回帰式は直線を表す一次式になります。単回帰なので説明変数 x は 1 つです。つまり、中学校で学んだ、

$$y = ax + b$$

が回帰式となります。式の中のそれぞれの文字の意味は以下の通りです。

- **変数：**
  - y …… 目的変数（ここでは、中古車の本体価格）：この値を予測したい
  - x …… 説明変数（ここでは、排気量）：予測にはこの値を使う
- **定数：**
  - a …… 係数：直線の傾き
  - b …… 定数項：x = 0 のときの y の値。切片とも呼ばれる

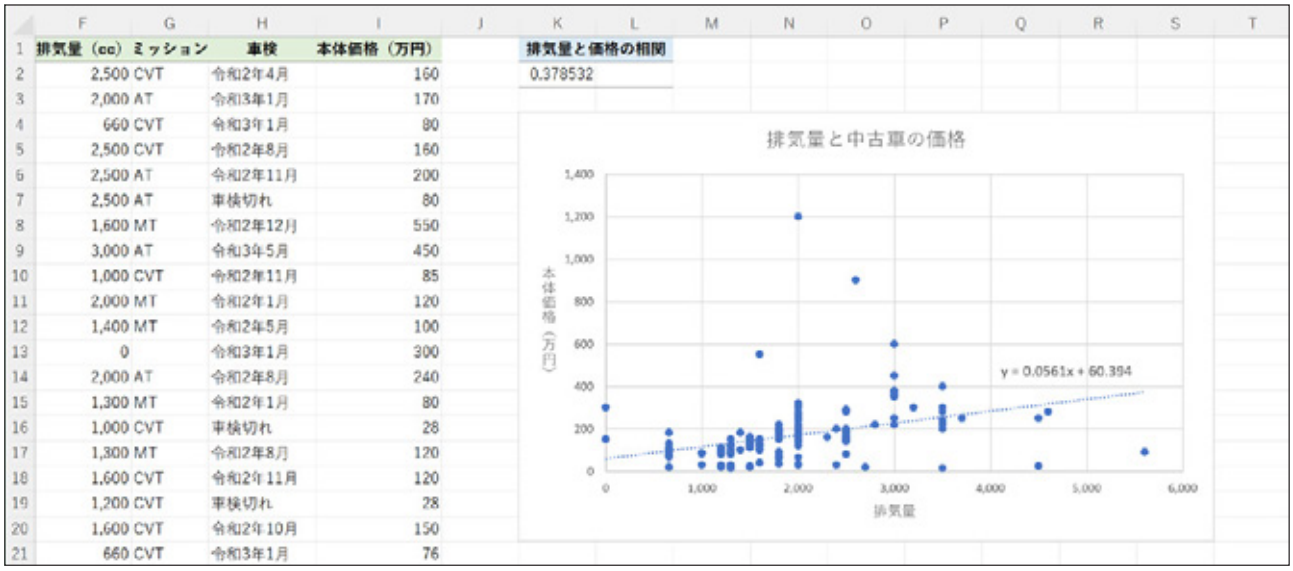
回帰式 y = ax + b を求めるということは、既に分かっている幾つかの y と x の値を基に、それらの値の近くを通る直線（つまり係数 a と定数項 b の値）を求めるということです。「近くを通る」というのはずいぶんとあいまいな言い方ですが、正確な意味については後述するので、今のところはあまり気にせずに進めましょう。

まず、図 1 に示したような散布図を描いてデータの全体像を見ておきます。散布図の描き方については、この連載の第 12 回で既に見ましたが、おさらいをかねて手順を示しておきます。その後、回帰式で表される直線（近似曲線）を描画し、さらに、回帰式を直線の近くに表示してみましょう。

相関係数は、セル **K2** にあらかじめ「=CORREL(I2:I171,F2:F171)」と入力して求められており、**0.378...** という値になっています（弱い正の相関）。ある程度直線的な関係があるものと考えられます。

では、サンプルファイルをこちらからダウンロードして［中古車価格］ワークシートを開き、図 2 の後に記した箇条書きの手順に従って取り組んでみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。

この手順と次の「関数を使って回帰式の係数と定数項を求める」の手順については動画も用意してあります。操作方法を一つ一つ丁寧に追いかけたい方はぜひご視聴ください。



	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	排気量（cc）	ミッション	車検	本体価格（万円）		排気量と価格の相関									
2	2,500	CVT	令和2年4月	160		0.378532									
3	2,000	AT	令和3年1月	170											
4	660	CVT	令和3年1月	80											
5	2,500	CVT	令和2年8月	160											
6	2,500	AT	令和2年11月	200											
7	2,500	AT	車検切れ	80											
8	1,600	MT	令和2年12月	550											
9	3,000	AT	令和3年5月	450											
10	1,000	CVT	令和2年11月	85											
11	2,000	MT	令和2年1月	120											
12	1,400	MT	令和2年5月	100											
13	0		令和3年1月	300											
14	2,000	AT	令和2年8月	240											
15	1,300	MT	令和2年1月	80											
16	1,000	CVT	車検切れ	28											
17	1,300	MT	令和2年8月	120											
18	1,600	CVT	令和2年11月	120											
19	1,200	CVT	車検切れ	28											
20	1,600	CVT	令和2年10月	150											
21	660	CVT	令和3年1月	76											

図 2　中古車データの回帰式を可視化する

散布図には回帰式で表される近似曲線を表示することができる。また、回帰式も表示できる。外れ値が幾つかあるように見えるので、それらのデータを除外すると回帰式の当てはまりがよくなりそうである（が、そのための前処理については後述する）。回帰式は y = 0.0561x + 60.394 であることが分かる。つまり係数 a が **0.0561**、定数項 b が **60.394** となる。

手順は以下の通りです。排気量のデータはセル **F1** ～ **F171** に、本体価格のデータはセル **I1** ～ **I171** に入力されています。セル **F1** とセル **I1** には見出しが入力されているので、値は 2 ～ 171 行目の 170 件となります。

Excel での操作手順

- セル **F1** ～ **F171** とセル **I1** ～ **I171** を選択する
- ［挿入］タブを開き、［散布図（X, Y）またはバブルチャートの挿入］ボタンをクリックする
- ［散布図］を選択する（これで散布図が作成される）
- 系列（作成された散布図のいずれかの点）を右クリックし、［近似曲線の追加］を選択する
- ［近似曲線の書式設定］作業ウィンドウで［グラフに数式を表示する］チェックボックスをオンにする

Google スプレッドシートでの操作手順

- セル **F1** ～ **F171** とセル **I1** ～ **I171** を選択する
- メニューバーから［挿入］－［グラフ］を選択する
- ［グラフエディタ］作業ウインドウで［グラフの種類］リストから［散布図］を選択する（これで散布図が作成される）
- ［グラフエディタ］作業ウインドウで［カスタマイズ］をクリックする
- ［系列］をクリックして下位の設定項目を表示し、［トレンドライン］チェックボックスをオンにする
- ［ラベル］のリストから［方程式を使用］を選択する

グラフ化するセルの範囲を選択するには、ドラッグ操作と［Ctrl］＋ドラッグ操作で離れた範囲を選択するよりも［名前］ボックスに「F1:F171,I1:I171」と入力する方が簡単です。ただし、Google スプレッドシートでは、この指定ができないので、［名前］ボックスに「F1:F171」と入力した後、［Ctrl］キーを押しながらセル **I1** ～ **I171** をドラッグする必要があります。

回帰式の係数 a は、直線の傾きを表すので、排気量が **1cc** 大きくなると本体価格は **0.0561 万円**（＝ **561 円**）だけ上がることが分かります。定数項は x の値、つまり排気量が **0**（エンジンがない）のときの本体価格ということになります。この例では年式や車検などについては考慮しておらず、車体の価値が排気量だけで決まるという前提なので、定数項の値の **60.394**（万円）は車体に価値がないときにも必要な固定費と考えることができます。

回帰式の係数と定数項を求めよう

散布図には回帰式も表示できるので、回帰式の係数 a や定数項 b がいくらであるかも分かります。y = ax + b の x に、排気量の値を代入すれば、本体価格 y が予測できますね。しかし、グラフから a と b の値を読み取り、セルに入力して計算するのはあまり賢明な方法とはいえません。出力結果を手作業で入力し直すのは手間がかかるだけでなく、ミスの原因ともなるからです。そこで、関数を使って回帰式の係数 a と定数項 b を求めることにしましょう。

回帰式 y = ax + b の係数 a を求めるには SLOPE 関数を使い、定数項 b を求めるには INTERCEPT 関数を使います。引数には、既に分かっている目的変数 y の値（既知の y）と既に分かっている説明変数 x の値（既知の x）を指定します。同じサンプルファイルの［単回帰分析］ワークシートを開き、図 3 のように SLOPE 関数と INTERCEPT 関数を入力してみましょう。排気量の値（既知の x）はセル **F2** ～ **F171** に、それぞれの本体価格（既知の y）はセル **I2** ～ **I171** に入力されています。手順は図の中にも記してありますが、図の後にも箇条書きで記しておきます。

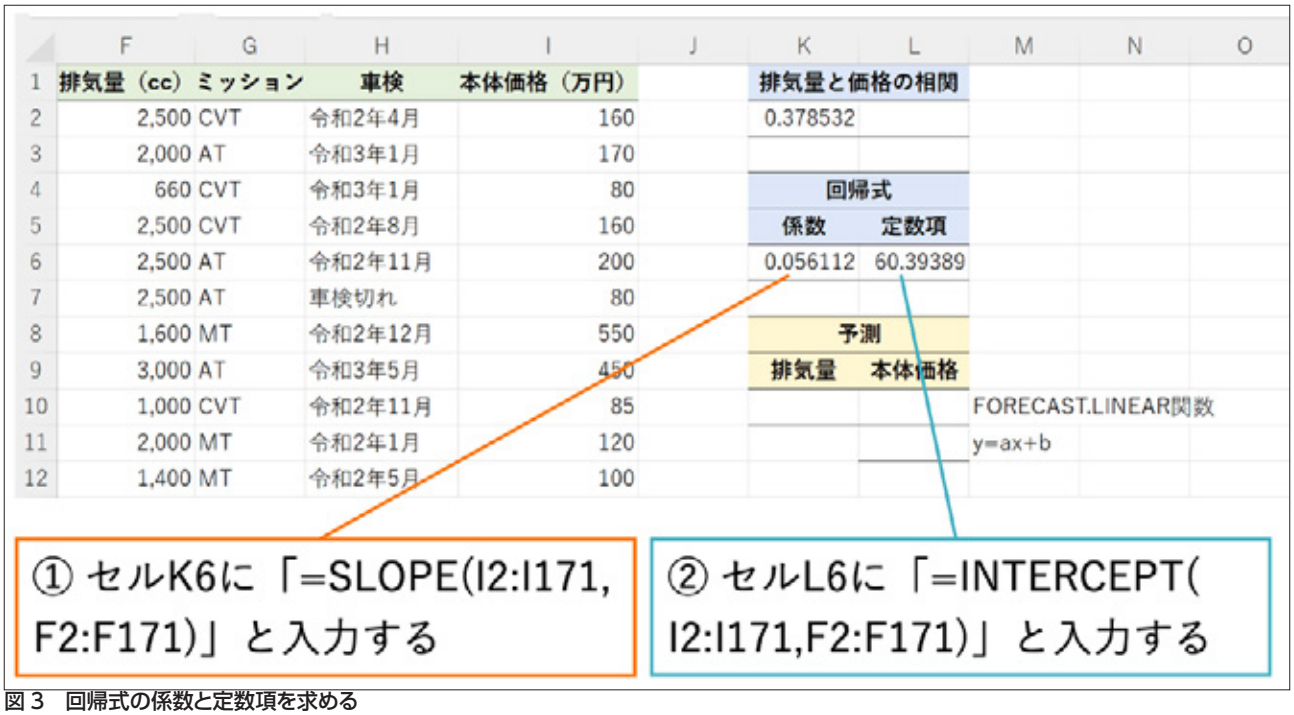


図 3　回帰式の係数と定数項を求める

SLOPE 関数を使って回帰式の係数を求め、INTERCEPT 関数を使って定数項を求める。いずれも、引数には既に分かっている目的変数 y の値（既知の y）と、既に分かっている説明変数 x の値（既知の x）を指定する。

手順は以下の通りです。

- セル **K6** に「=SLOPE(I2:I171,F2:F171)」と入力する
- セル **L6** に「=INTERCEPT(I2:I171,F2:F171)」と入力する

図 3 から、回帰式は以下のようになることが分かります。散布図に表示された値よりも詳細な値が求められていますね（小数点以下の表示桁数を増やせば、有効数字 15 桁まで表示できます）。

$$y = 0.056112x + 60.39389$$

回帰式を利用した予測を行ってみよう

続いて、回帰式を利用した予測を行います。例えば、排気量が **1500cc** の場合、x = 1500 なので、本体価格は以下のようになります。

$$
\begin{aligned}
y &= 0.056112 \times 1500 + 60.39389 \\
&= 144.56189
\end{aligned}
$$

しかし、このような数式を使って自分で計算しなくても、FORECAST.LINEAR 関数を使えば、予測に使いたい x の値と、既知の y の値、既知の x の値を指定するだけで、予測値が求められます（図 4）

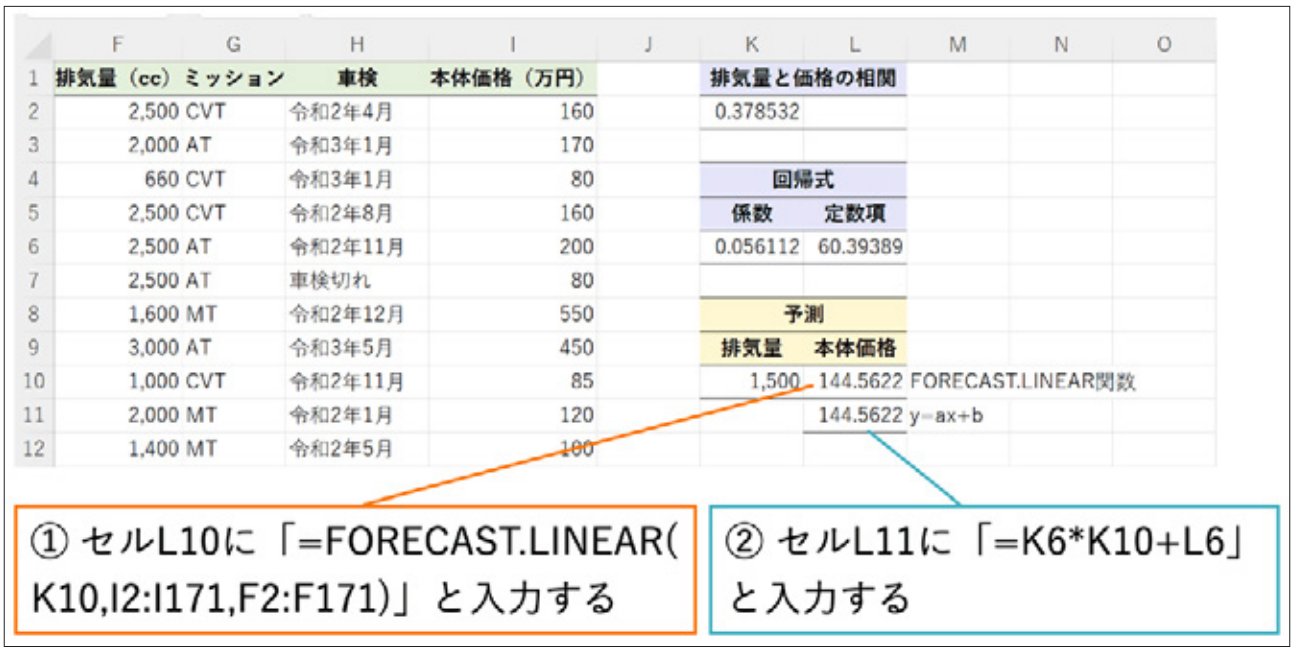


図 4　排気量から中古車の価格を予測する

セル **L10** では FORECAST.LINEAR 関数を使って予測値を求めている。一方、セル **L11** では回帰式に値を代入して予測値を求めている。当然のことながらこれらの値は一致する。上で見た計算とは答えが異なるのは係数や定数項の小数点以下の桁数の違いによるもの。

手順は以下の通りです。

- セル **K10** に「**1500**」と入力する
- セル **L10** に「=FORECAST.LINEAR(K10,I2:I171,F2:F171)」と入力する
- セル **L11** に「=K6*K10+L6」と入力する

ここまでは「習うより慣れろ」主義で、取りあえず回帰分析を行ってみました。しかし、回帰分析の仕組みについてはあまり詳しくはお話ししていませんでした。最初に、回帰式は各データの近くを通る直線であるとお話ししましたが、「近くを通る」の意味をきちんと見ておきましょう。

回帰分析の考え方（単回帰分析の場合）

回帰式の意味を図 5 で考えてみます（話を簡単にするためにこれまでのデータとは少し違う値を使っています）。回帰式は目的変数のそれぞれの値（実測値）と直線（予測値）との誤差ができるだけ小さくなるように係数 a と定数項 b を決めたものです。ここでは考え方が分かればいいので、a と b の値を求めるための道筋を示すにとどめます。

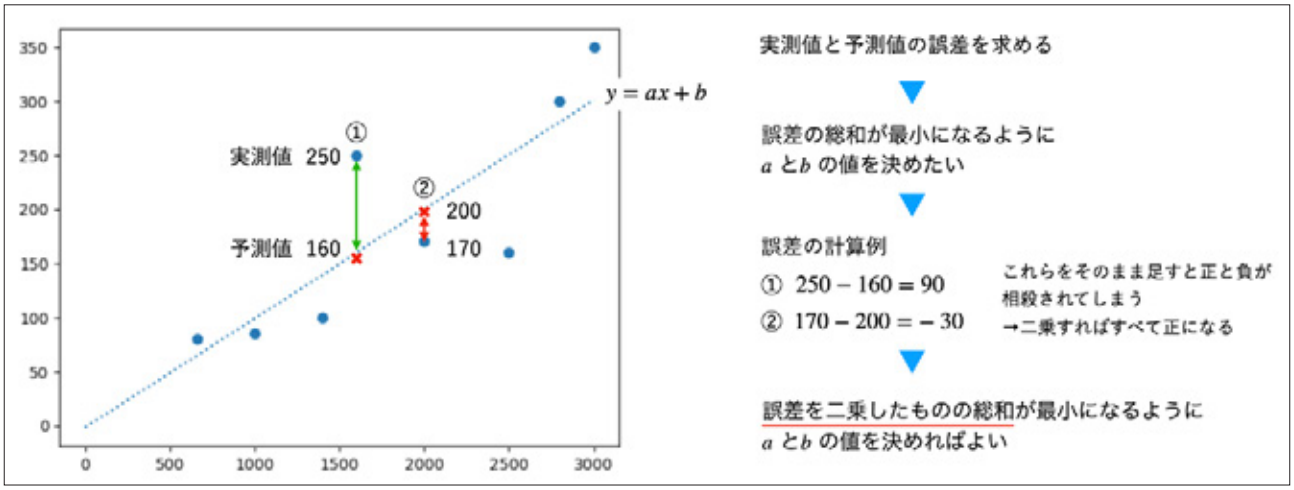


図 5　回帰式の意味（係数と定数項の求め方）

実測値は●で表された点、予測値は回帰式で表される直線上の × 点。それぞれの実測値と予測値の差（誤差）の合計が最小になるように a と b の値を決めたい。ただし、誤差は正の値になることも負の値になることもあるので、単純に合計を求めるとそれらの値が相殺されてしまう。そこで、誤差の絶対値を求めるために誤差を二乗しておく。その合計が最小になるように a と b の値を決めればよい。

図 5 の数値と照らし合わせながら、以下の文章を読み進めてください。

例えば、左から 4 番目の点を見てみましょう。実際には **1600cc** の中古車が **250 万円**であったとします。単位の万円を省略して表すと、実測値は **250** です。ここで、回帰式 y = ax + b に x = 1600 を代入してみたところ、予測値が y = 160 になったとします。すると実測値と予測値の差は 250 − 160 = 90 です。これは正の値ですね。

左から 5 番目の点についても見てみましょう。**2000cc** の中古車の本体価格、つまり実測値は **170** です。回帰式 y = ax + b に x = 2000 を代入したところ、予測値が y = 200 になったとします。実測値と予測値の差は、170 − 200 = -30 です。こちらは負の値になります。

実測値と予測値の誤差を最小にするには、それぞれの誤差を全て足した合計が最小になるようにすればいいですね。しかし、上で求めたように誤差は正になることもあれば負になることもあるので、そのまま足してしまうと正の値と負の値が相殺されてしまいます。そこで、誤差を二乗して全て正の値にします。そして、その総和が最小になるような a と b の値を求めます。つまり、**誤差の二乗和を最小にするように a と b の値を決める**というわけです。このような方法を**最小二乗法**と呼びます。

ここでは、実測値と予測値の誤差の二乗和を最小にする a と b の値を求めればいい、というところまでたどり着ければ十分です。それぞれの値を求めるための計算は SLOPE 関数と INTERCEPT 関数に任せることにしましょう。

SLOPE 関数と INTERCEPT 関数を使えば回帰式の係数 a と定数項 b の値が簡単に求められますが、上の説明だけでは物足りない（SLOPE 関数や INTERCEPT 関数で実際にどのように計算されているのかが気になる）という方もいると思います。結果だけ示しておくと以下のようになります。

$$a = \frac{\sum ((x_i - \bar{x})(y_i - \bar{y}))}{\sum (x_i - \bar{x})^2}$$

$$b = \bar{y} - a\bar{x}$$

なぜこのような式で求められるのか興味のある方は、こちらのリンク先に詳しい話を掲載しているので、ぜひご参照ください。微分（偏微分）の知識が必要になりますが、基本的に四則演算のみで理解できます。

回帰分析の精度を上げるには

最初に述べたように、回帰分析には、適切に分析するための準備が必要です。その際に留意すべきポイントを 2 点挙げておきます。他にも考慮すべきことはあるのですが、今回はこの 2 つに絞って見ておきましょう。

- 変数同士にどのような関係があるかを見極める（**モデルの選択**）
- 外れ値や欠損値などがないかどうか調べる（**データクリーニング**）

これまでは、説明変数と目的変数が直線的な関係であるという前提で、線形回帰モデル（$y = ax + b$ などの一次式で表される回帰式）を利用した例を見てきましたが、相関係数を求めたり、散布図を描いてみたりすると「関係はあるが直線的な関係ではない」と考えられる場合もあります。そのような場合には、最初に少し触れた指数回帰（$y = b + a^x$）や、次回お話しする予定の多項式回帰（$y = ax^2 + bx + c$ など二次以上の項があるもの）など、より適切と考えられるモデルを利用します。ただ、今回の中古車データの例に関しては、線形回帰モデルで進めることにします。

外れ値の発見については、散布図が役に立ちます。この連載の第 10 回で紹介した箱ひげ図や、第 4 回で紹介したスミルノフ・グラブス検定なども使えます。

散布図の中で特に目立つ **1200 万円**と **900 万円**の中古車は、プレミアの付いた希少価値のある旧車です（元のデータでは 81 行目と 105 行目にあります）。また、排気量が 0 の中古車は EV（電気自動車）です（13 行目と 68 行目にあります）。これらの値を除外すると、より直線的な関係になりそうですね。

この例には欠損値は存在しませんが、CORREL 関数、SLOPE 関数、INTERCEPT 関数、FORECAST.LINEAR 関数では、説明変数または目的変数の値が存在しないペアや、文字列が入力されているセルは無視されるので、欠損値があっても、存在するデータだけで計算が行われます。ただし、欠損値があることを確認しておくことは重要です。

では、外れ値を除外したデータで回帰分析を行ってみましょう。［外れ値の除外］ワークシートを開いて取り組んでみてください。データを 4 行削除したので、目的変数はセル **I2** ～ **I167**、説明変数はセル **F2** ～ **F167** となります。手順はこれまでと全く同じです（図 6）。

	F	G	H	I	J	K	L	M
1	排気量（cc）	ミッション	車検	本体価格（万円）		排気量と価格の相関		
2	2,500	CVT	令和2年4月	160		0.516905		
3	2,000	AT	令和3年1月	170				
4	660	CVT	令和3年1月	80		回帰式		
5	2,500	CVT	令和2年8月	160		係数	定数項	
6	2,500	AT	令和2年11月	200		0.055183	49.73046	
7	2,500	AT	車検切れ	80				
8	1,600	MT	令和2年12月	550		予測		
9	3,000	AT	令和3年5月	450		排気量	本体価格	
10	1,000	CVT	令和2年11月	85		1,500	132.5056	
11	2,000	MT	令和2年1月	120				

図 6　排気量から中古車の価格を予測する（外れ値を除去した例）

入力すべき関数はこれまでに見たものばかり。引数として指定するセル範囲が異なるだけ。相関係数が **0.516...** となっており、最初に見た例よりも直線的な関係になっていることが分かる。極端に大きな本体価格の影響を受けなくなったので、予測値は少し小さくなっている。

手順は以下の通りです。セル **K2** にはあらかじめ「=CORREL(I2:I167,F2:F167)」と入力されています。

- セル **K6** に「=SLOPE(I2:I167,F2:F167)」と入力する
- セル **L6** に「=INTERCEPT(I2:I167,F2:F167)」と入力する
- セル **K10** に「**1500**」と入力する
- セル **L10** に「=FORECAST.LINEAR(K10,I2:I167,F2:F167)」と入力する

最初に見た例（図 5）では、相関係数は **0.378...** と弱い正の相関になっていますが、外れ値を除外すると **0.516...** となり、正の相関が少し強くなります。

回帰分析の精度を評価するには

回帰分析の精度を評価するためには R^2（決定係数）や RMSE（二乗平均平方根誤差）などが使われます。単回帰分析の場合、決定係数は相関係数の二乗で求められます。決定係数は値が大きいほど精度が良く、RMSE は値が小さいほど精度が良いと考えられます。RMSE を求めるための式は以下の通りです。

$$\sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$$

差の二乗和を求めるには、SUMXMY2 関数が使えるので、どのセルでもいいので「=SQRT(SUMXMY2(I2:I167, K6*F2:F167+L6)/COUNT(I2:I167))」と入力してみてください。Google スプレッドシートでは、「=ARRAYFORMULA(SQRT(SUMXMY2(I2:I167, K6*F2:F167+L6)/COUNT(I2:I167)))」と入力してください。RMSE の値は **84.657...** となります。

ちなみに、最初の例であれば、［単回帰分析］ワークシートで「=SQRT(SUMXMY2(I2:I171, K6*F2:F171+L6)/COUNT(I2:I171))」と入力します。Google スプレッドシートでは、「=ARRAYFORMULA(SQRT(SUMXMY2(I2:I171, K6*F2:F171+L6)/COUNT(I2:I171)))」と入力してください。RMSE の値は **128.734...** となります。

RMSE 値を比べると、外れ値を除外したことにより、精度が上がっていることが分かりますね。

Excel の散布図では［近似曲線の書式設定］作業ウィンドウで［グラフに R-2 値を表示する］チェックボックスをオンにすると、近似曲線の近くに決定係数が表示されます。Google スプレッドシートでは、［グラフエディタ］作業ウインドウの［カスタマイズ］画面で［系列］の下の［決定係数を表示する］チェックボックスをオンにします。

説明変数と目的変数の関係が直線的でない場合 ～ 指数回帰の例

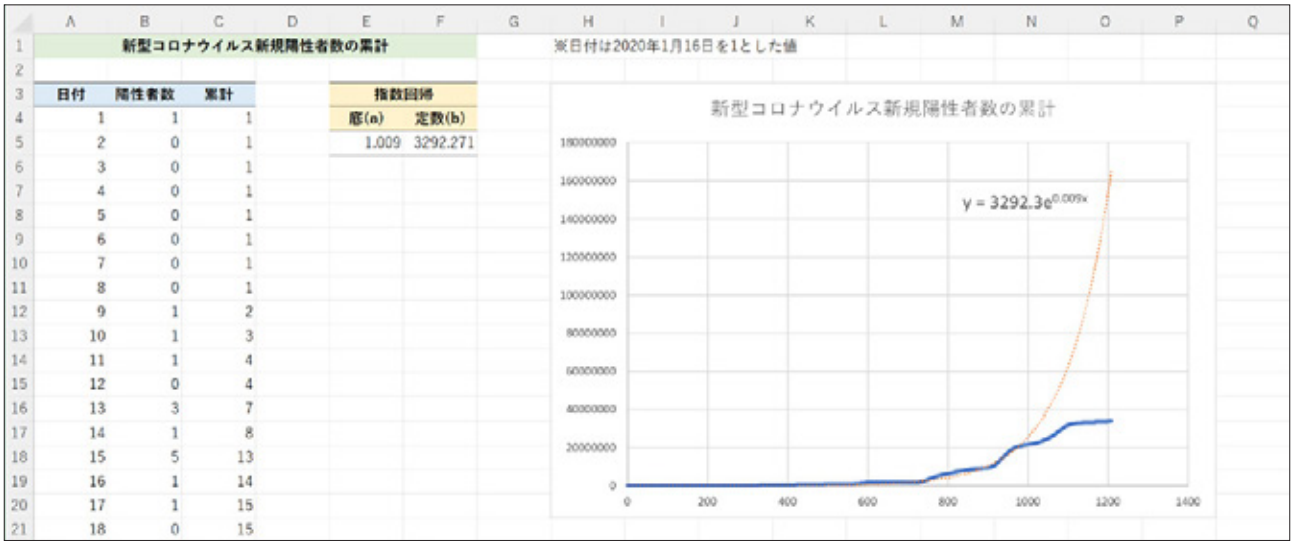
最初に触れたように、回帰式は直線を表すような式でなくても構いません。回帰分析に先立って散布図を作成したところ、x の値が大きくなると y の値が急激に大きくなるように見えたとしましょう。そのような場合には、以下のような指数関数を回帰式とした方が適切な場合があります。例えば、細菌の増殖や感染者数の増加などの例がそれに当たります。

$$y = b * a^x$$

ただし、細菌や感染者数はある程度増えると、その先は増加率が低下します。そのようなモデルを表すにはロジスティック方程式がよく使われます。ただ、今回の話題からは外れてしまうので、ここでは指数回帰のみを取り扱います。後のコラムで説明するように、変曲点（増加率が減少に転じる点）より前の部分では、ロジスティック方程式（の解であるロジスティック関数）を指数関数で近似できます。

指数回帰の場合、LOGEST 関数を使えば定数 b や底 a を一度に求めることができます（y の対数を求めて、これまでに見たような回帰分析を行っても構いませんが計算は少し面倒です）。図 7 の例は 2020 年 1 月 16 日～2023 年 5 月 8 日（5 類感染症に移行した日）までの新型コロナウイルス感染症新規陽性者数の累計データです。このデータは厚生労働省のオープンデータの「新規陽性者数の推移（日別）」を基に作成したものです。

サンプルファイルをこちらからダウンロードして［陽性者数の累計］ワークシートを開き、回帰式の底 a と定数 b を求め、散布図を描いて近似曲線を追加してみましょう。図 7 の後に示した手順に従って進めてみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。なお、この手順については動画も用意してあります。操作方法を一つ一つ丁寧に追いかけたい方はぜひご視聴ください。



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	新型コロナウイルス新規陽性者数の累計							※日付は2020年1月16日を1とした値									
2																	
3	日付	陽性者数	累計		指数回帰												
4	1	1	1		底(a)	定数(b)											
5	2	0	1		1.009	3292.271											
6	3	0	1														
7	4	0	1														
8	5	0	1														
9	6	0	1														
10	7	0	1														
11	8	0	1														
12	9	1	2														
13	10	1	3														
14	11	1	4														
15	12	0	4														
16	13	3	7														
17	14	1	8														
18	15	5	13														
19	16	1	14														
20	17	1	15														
21	18	0	15														

図 7　新型コロナウイルス感染症新規陽性者数の累計を予測する

回帰式は $y = 3292.271 \times 1.009^x$ となっている。散布図の青い●は実測値（間隔が狭いので線に見えている）、赤の点線は回帰式を元に描いた近似曲線。ただし、散布図には底を $e = 2.71828...$ とした $y = 3292.3 \times e^{0.009x}$ が回帰式として表示されている（$a^x = e^{cx}$ とすると $c = \log_e a$ で求められる）。

図 7 は以下の手順で作成されています。A 列は日付のシリアル値をそのまま使ってもいいのですが、実際に計算すると、定数項があまりにも小さくなって見づらいので、2020 年 1 月 16 日を **1** とした値にしてあります。これが説明変数 x に当たり、C 列の累計が目的変数 y に当たります。データは 1212 行目まで入力されています。

回帰式の係数と底の計算

- セル **E5** に「=LOGEST(C4:C1212,A4:A1212,TRUE,FALSE)」と入力する
  - スピル機能が使えない場合は、あらかじめセル **E5** ～ **F5** を選択しておき、「=LOGEST(C4:C1212,A4:A1212,TRUE,FALSE)」と入力して、入力終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す

近似曲線の描画

Excel での操作手順

- セル **A3** ～ **A1212** とセル **C3** ～ **C1212** を選択して散布図を描画する
- 系列（作成された散布図のいずれかの点）を右クリックし、［近似曲線の追加］を選択する
- ［近似曲線の書式設定］作業ウィンドウで［指数回帰］オプション（または［指数近似］オプション）をオンにする
- ［グラフに数式を表示する (E)］チェックボックスをオンにする

Google スプレッドシートでの操作手順

- セル **A3** ～ **A1212** とセル **C3** ～ **C1212** を選択して散布図を描画する
  - ・ただし、離れた範囲が選択しづらいので、セル **A3** ～ **C1212** を選択して散布図を作成し、後で B 列の系列を削除した方が簡単
- ［グラフエディタ］作業ウインドウで［カスタマイズ］をクリックする
- ［系列］をクリックして下位の設定項目を表示し、［トレンドライン］チェックボックスをオンにする
- ［種類］のリストから［指数関数］を選択する
- ［ラベル］のリストから［方程式を使用］を選択する

なお、Google スプレッドシートでは、LOGEST 関数で求められる回帰式の定数と底は Excel で求めた値と同じですが、近似曲線は実測値に近くなるように表示されるようです（この例では $y = 81796 \times e^{0.00533x}$ となります）。Google スプレッドシートのサンプルファイルには、LOGEST 関数で求められた回帰式による予測値もプロットしてあります（Excel の近似曲線と同様のものになります）。

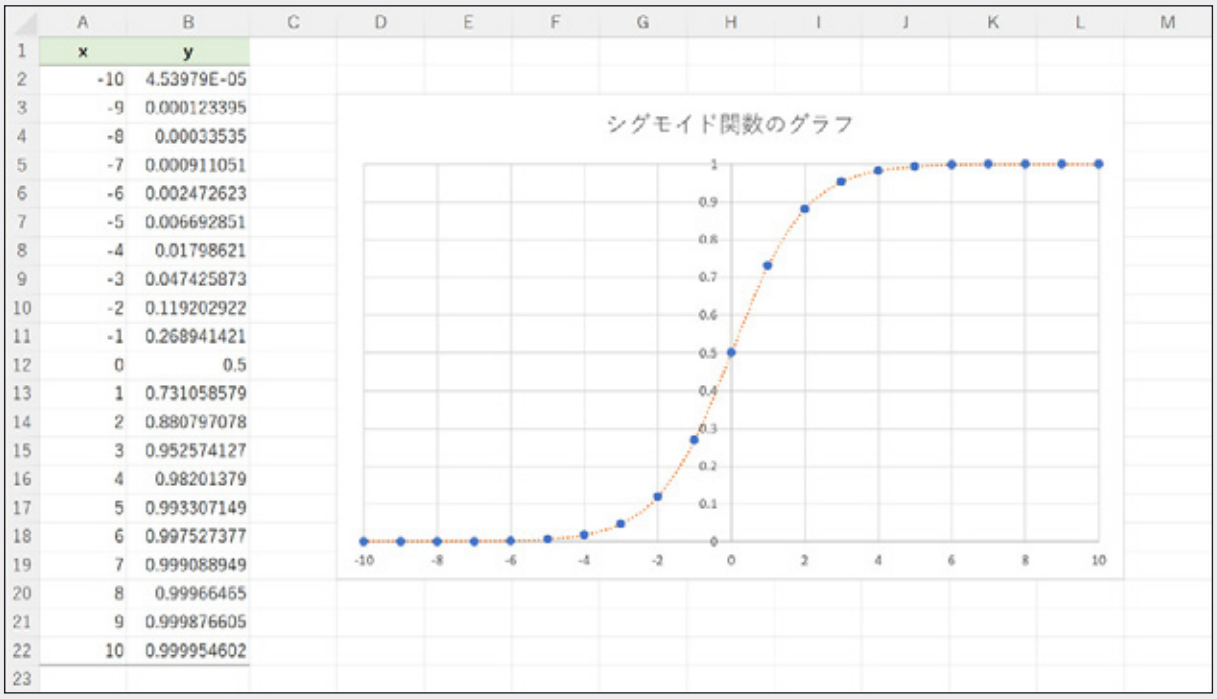
最後に、話がそれてしまいますが、今回使ったデータの性質に関連のあるちょっと発展的なお話をコラムとして掲載しておきます。

コラム　細菌の数や感染者数などはロジスティック方程式でモデル化される

細菌の数や感染者数などはある程度増加すると増加率が低下していきます。そのような例を表すモデルとしてはロジスティック方程式がよく知られています。例えば、ロジスティック関数の特殊な形であるシグモイド関数の式は以下の通りです。目的変数 y は **0 ～ 1** の値を取るので（確率なので）、以下の式では p と表しています。

$$p = \frac{1}{1 + e^{-x}}$$

図 8 はシグモイド関数のグラフです。増加率が低下していき、やがて一定のレベルに収束していく様子が分かりますね。上で見たサンプルファイルの [シグモイド関数] ワークシートにこの例が含まれています。シグモイド関数の値が B 列に求められており、散布図（平滑線とマーカー）を作成してあります。



	A	B
1	x	y
2	-10	4.53979E-05
3	-9	0.000123395
4	-8	0.00033535
5	-7	0.000911051
6	-6	0.002472623
7	-5	0.006692851
8	-4	0.01798621
9	-3	0.047425873
10	-2	0.119202922
11	-1	0.268941421
12	0	0.5
13	1	0.731058579
14	2	0.880797078
15	3	0.952574127
16	4	0.98201379
17	5	0.993307149
18	6	0.997527377
19	7	0.999088949
20	8	0.99966465
21	9	0.999876605
22	10	0.999954602
23		

図 8　シグモイド関数の例

x の値が増えていくと y の増加率が増えていき、急激に y の値が大きくなるが、しばらくすると増加率が減少し、y の値が一定のレベルに収束する。シグモイド関数では y の値は **0 ～ 1** の範囲となる。なお、シグモイド関数の増加率が増えている部分（この例であれば **-5 ～ 0** あたり）は指数関数で近似できる。

入力されている数式は以下の通りです。EXP 関数は e のべき乗を求める関数です。

- セル **B2** に「=1/(1+EXP(-A2:A22))」と入力する（シグモイド関数の値が求められる）
  - Google スプレッドシートの場合は「=ARRAYFORMULA(1/(1+EXP(-A2:A22)))」と入力する

図 8 を見ると、シグモイド関数の逆関数を使えば（y を x に変換すれば）直線的な値になることも分かりますね。シグモイド関数の逆関数は、以下の**ロジット関数**と呼ばれるものです。対数は底が e = 2.7182... の自然対数です。

$$x = \log \frac{p}{1-p}$$

例えば、セル **C1** に自然対数を求める LN 関数を使って「=LN(B2:B22/(1-B2:B22))」と入力すれば、元の **-10 ～ 10** という値が得られます。

既に述べたように、ロジスティック関数（シグモイド関数）の変曲点よりも前の部分は、指数関数で近似できます。サンプルファイルには、それらを比較したグラフも含めてあります。

蛇足ながら、上で紹介したロジット関数とシグモイド関数を使えば、分類を行うためのロジスティック回帰の計算が簡単にできます。それについては、また機会があればお話ししたいと思います。

コラム　周期的に変化するデータには時系列分析が使える

皆さんもご存じのように、陽性者数の増加には波があります。2022 年の冬に第 8 波が来た後、しばらくは落ち着きを見せ、5 類感染症になった 2023 年 5 月以降には第 9 波が来たと言われています。このように波があるデータを基に、その周期などを検出して予測するには、FORECAST.ETS 関数による時系列分析が役に立ちます。なお、この関数を含めて以下の説明で用いる時系列分析の関数は Excel デスクトップ版でのみ利用でき、Microsoft 365 オンラインや Google スプレッドシートでは利用できません。

ここでは、波が比較的顕著に現れている 2022 年 1 月 1 日以降のデータのみを使い、使い方のみ紹介することにします。図 9 は、2023 年 5 月 9 日から 12 月 31 日までの感染者数を予測した例です。Excel 用の［時系列分析］ワークシートを開いて、数式やグラフを確認してみてください。

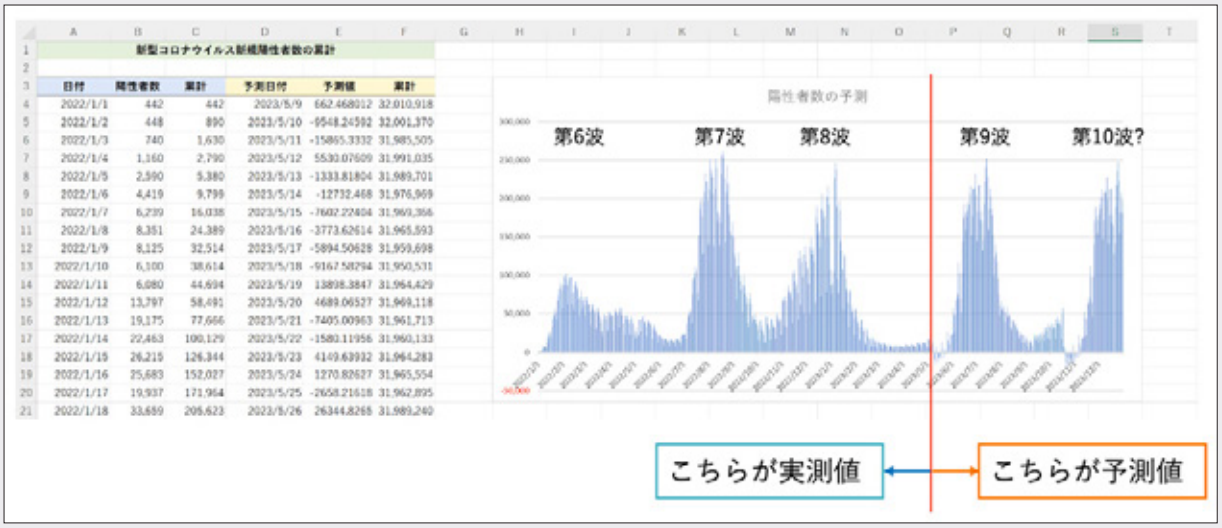


図 9　時系列分析による陽性者の累計の予測例

2023 年 1 月 1 日～ 5 月 8 日の陽性者数を基に、2023 年 5 月 9 日～ 12 月 31 日について予測してみた。2023 年 5 月以降の第 9 波が予測できている。同じ周期で陽性者数が増えるのであれば、2023 年 12 月以降にも陽性者数が増える予測になっている（実際、2024 年に入って変異株 JN.1 への置き換わりが進んでおり、第 10 波が到来しているという声もある）。

入力すべき関数は以下のたった 1 つだけです。

- セル **E4** に「=FORECAST.ETS(D4:D240,B4:B496,A4:A496)」と入力する

予測を行いたい日付はセル **D4** ～ **D240** に入力されています。予測の基となるデータ（陽性者数）はセル **B4** ～ **B496** に入力されており、それに対するタイムライン（日付）はセル **A4** ～ **A496** に入力されています。それらの値を指定して、E 列で予測される陽性者数を求めているというわけです。

右側のグラフは C 列と E 列を棒グラフにしたものです。離れた範囲を 1 つの系列とするのはちょっと難しいですが、以下のように、後から系列の範囲を追加する方法が簡単です。

- セル **A3** ～ **B496** の範囲を選択して、取りあえず元のデータで棒グラフを作っておく
- 系列をクリックして選択し、数式バーに入力されている系列の指定を「=SERIES(時系列分析 !$B$3,(時系列分析 !$A$4:$A$733, 時系列分析 !$D$4:$D$240),(時系列分析 !$B$4:$B$733, 時系列分析 !$E$4:$E$240),1)」に修正する

図 9 を見ると、かなり正確に波（季節性）が予測できていることが分かります。なお、波の周期は、FORECAST.ETS.SEASONALITY 関数で求められます。どのセルでも構わないので、「=FORECAST.ETS.SEASONALITY(B4:B496,A4:A496)」と入力すると、**164** という値が得られます。164 日つまり 5 カ月半の周期で波が来ていることが分かります。

今回は、単回帰分析により回帰式の係数と定数項を求めたり、回帰式を利用して予測したりする方法を紹介しました。また、回帰式を可視化するために、散布図に近似曲線を追加する方法についても触れました。しかし、事例として取り上げた中古車の価格は、排気量だけで決まるものではありません。

そこで、次回は年式や走行距離なども含めた複数の説明変数を利用した回帰分析（重回帰分析）について見ていきます。数値として表されている値だけでなく、メーカーなどの文字列データ（カテゴリを表すデータ）も説明変数として取り扱ったり、新たな特徴量を作り出して説明変数に加えたりする方法についても触れます。さらに、前回見た気温と電気器具による CO_2 排出量のように、U 字形の関係になっている場合にも回帰分析を行うために、多項式回帰を利用する方法も紹介します。次回も、どうぞお楽しみに！

関数リファレンス：この記事で取り上げた関数の形式

関数の利用例については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

単回帰分析のために使った関数

SLOPE 関数：単回帰分析による回帰式の係数（直線の傾き）を求める

形式

SLOPE(既知の y, 既知の x)

引数

- **既知の y**：既に得られている目的変数の値（セル範囲）を指定する。
- **既知の x**：既に得られている説明変数の値（セル範囲）を指定する。

INTERCEPT 関数：単回帰分析による回帰式の定数項（切片）を求める

形式

INTERCEPT(既知の y, 既知の x)

引数

- **既知の y**：既に得られている目的変数の値（セル範囲）を指定する。
- **既知の x**：既に得られている説明変数の値（セル範囲）を指定する。

FORECAST.LINEAR 関数：単回帰分析による予測値を求める

形式

FORECAST.LINEAR(未知の x, 既知の y, 既知の x)

引数

- **未知の x**：予測に使いたい説明変数の値を指定する。この x に対する y の値が求められる。
- **既知の y**：既に得られている目的変数の値（セル範囲）を指定する。
- **既知の x**：既に得られている説明変数の値（セル範囲）を指定する。

指数回帰分析のために使った関数

LOGEST 関数：指数回帰による回帰式 $y = b \times a^x$ の底 a や定数項 b を求める

形式

LOGEST(既知の y, 既知の x, [定数], [補正項])

引数

- **既知の y**：既に得られている目的変数の値（セル範囲）を指定する。
- **既知の x**：既に得られている説明変数の値（セル範囲）を指定する。
- **定数**：定数項 b の値を求めるかどうかを以下の値で指定する。
  - **TRUE** または**省略** …… b の値を求める
  - **FALSE** …… b の値を **1** とする
- **補正項**：検定などで使われる詳細な情報を求めるかどうかを以下の値で指定する。
  - **TRUE** …… 補正項を求める
  - **FALSE** または**省略** …… 補正項は求めない（底と定数項のみを求める）

備考

この関数は底と定数項などの値を一度に返すので、スピル機能が使えない場合は配列数式として入力する必要がある。また、説明変数 x は複数あっても構わない。つまり $y = b \times a_1^{x1} \times a_2 x^{2\times} \cdots \times anxn$ のような回帰式の底と定数項が求められる。返される値は $an, \ldots, a_2, a_1, b$ の順であることに注意。

差の二乗和を求めるために使った関数

SUMXMY2 関数：$\Sigma (x_i - y_i)^2$ の値を求める

形式

SUMXMY2(x, y)

引数

- **x**：x の値が入力されているセル範囲を指定する。
- **y**：y の値が入力されているセル範囲を指定する。

対数や指数を求めるために使った関数

LN 関数：自然対数を求める

形式

LN(数値)

引数

- **数値**：自然対数を求めたい値（真数）を指定する。自然対数の底は $e = 2.71828...$ となる。

EXP 関数：e^n を求める

形式

EXP(数値)

引数

- **数値：** e の指数 n に当たる値を指定する。自然対数 e ＝ 2.71828... の n 乗が求められる。

時系列分析のために使った関数

FORECAST.ETS 関数：時系列分析を行う

形式

FORECAST.ETS(期日 , 値 , タイムライン , ［季節性］ , ［補間］ , ［集計］)

引数

- **期日：** 予測したい日付や時間などを指定する。
- **値：** 既に分かっている値を指定する。
- **タイムライン：** 値に対する日付や時間などを指定する。
- **季節性：** 周期を計算するかどうかを指定する。
  - **0**：…… 季節性がないものと見なす。
  - **1** または**省略**：…… 周期を自動的に計算する。
  - **その他の値**：…… 指定した値を周期とする。
- **補間：** 欠損値の取り扱いを以下の値で指定する。
  - **0** …… 欠損値を **0** とする。
  - **1** または**省略**：…… 隣接するセルの平均で補間する。
- **集計：** タイムラインに同じ日付や時刻の値がある場合にどの手法で集計するかを以下の値で指定する。
  - **1** または**省略** …… 平均値（AVERAGE）
  - **2** …… 数値（＝数値を含むセル）の個数（COUNT）
  - **3** …… データ（＝空白ではない全てのセル）の個数（COUNTA）
  - **4** …… 最大値（MAX）
  - **5** …… 中央値（MEDIAN）
  - **6** …… 最小値（MIN）
  - **7** …… 合計（SUM）

備考

この関数は Excel デスクトップ版（Excel 2016 以降）でのみ利用でき、Microsoft 365 オンラインや Google スプレッドシートでは利用できない。

FORECAST.ETS.SEASONALITY 関数：季節性の周期を求める

形式

FORECAST.ET.SEASONALITY(値 , タイムライン , ［補間］ , ［集計］)

引数

- **値**：既に分かっている値を指定する。
- **タイムライン**：値に対する日付や時間などを指定する。
- **補間**：欠損値の取り扱いを以下の値で指定する。
  - **0** …… 欠損値を **0** とする。
  - **1** または**省略** …… 隣接するセルの平均で補間する。
- **集計**：タイムラインに同じ日付や時刻の値がある場合にどの手法で集計するかを以下の値で指定する。
  - **1** または**省略** …… 平均値（AVERAGE）
  - **2** …… 数値（＝数値を含むセル）の個数（COUNT）
  - **3** …… データ（＝空白ではない全てのセル）の個数（COUNTA）
  - **4** …… 最大値（MAX）
  - **5** …… 中央値（MEDIAN）
  - **6** …… 最小値（MIN）
  - **7** …… 合計（SUM）

備考

この関数は Excel デスクトップ版（Excel 2016 以降）でのみ利用でき、Microsoft 365 オンラインや Google スプレッドシートでは利用できない。

[データ分析] 重回帰分析による予測（線形回帰、多項式回帰）～年式、走行距離、排気量から中古車の価格を予測

データ分析の初歩からステップアップしながら学んでいく連載の第 15 回。複数の説明変数を基に目的変数の値を予測する重回帰分析について、Excel を使って手を動かしながら学んでいきましょう。カテゴリーなどの数値ではないデータを説明変数として利用する方法や、二次関数などの多項式を基に回帰分析する方法も紹介します。

羽山博（2024 年 02 月 22 日）

前回は単回帰分析により、説明変数 x の値から目的変数 y の値を予測するための回帰式を求めたり、回帰式を基に予測を行ったりしました。

今回は、説明変数が複数ある場合の**重回帰分析**に取り組みます。図 1 の例であれば、年式が説明変数 x_1、走行距離が説明変数 x_2、排気量が説明変数 x_3 となり、本体価格が目的変数 y となります。図 1 のデータでは実際のメーカーや車種の名称が使われていますが、本体価格などの値は架空のものです。

	A	B	C	D	E	F	G	H	I
1	メーカー	名称	タイプ	年式	走行（km）	排気量（cc）	ミッション	車検	本体価格（万円）
2	スバル	レガシィ	セダン	2016	55,000	2,500	CVT	令和2年4月	160
3	マツダ	CX-5	SUV	2018	40,000	2,000	AT	令和3年1月	170
4	スズキ	アルト	コンパクト	2019	20,000	660	CVT	令和3年1月	80
5	日産	ティアナ	セダン	2016	55,000	2,500	CVT	令和2年8月	160
6	マツダ	CX-9	SUV	2017	35,000	2,500	AT	令和2年11月	200
7	トヨタ	ハイエース	バン	2019	40,000	2,500	AT	車検切れ	80
8	アルファロメオ	ジュリアスプリ	クーペ	1962	70,000	1,600	MT	令和2年12月	550
9	ポルシェ	911	スポーツカー	2020	15,000	3,000	AT	令和3年5月	450
10	トヨタ	パッソ	コンパクト	2017	30,000	1,000	CVT	令和2年11月	85
11	マツダ	ロードスター	スポーツカー	2017	40,000	2,000	MT	令和2年1月	120
12	スズキ	バレーノ	コンパクト	2016	65,000	1,400	MT	令和2年5月	100

	K	L	M	N	O
1	相関行列				
2		年式	走行	排気量	本体価格
3	年式	1.000	-0.371	-0.112	-0.279
4	走行	-0.371	1.000	0.123	-0.388
5	排気量	-0.112	0.123	1.000	0.517
6	本体価格	-0.279	-0.388	0.517	1.000
7					

	K	L	M	N	O	P
8	重回帰分析					
9		排気量	走行	年式	定数項	
10	係数	0.057959	-0.00174	-5.98853	12204.43	
11	係数の標準誤差	0.005042	0.000142	0.67449	1362.932	
12	R^2, 標準誤差	0.646035	59.55898	#N/A	#N/A	
13	F値, 自由度	98.55734	162	#N/A	#N/A	
14	二乗和（回帰, 誤差）	1048829	574658	#N/A	#N/A	
15						
16		年式	走行	排気量	本体価格	
17	予測値	2018	20000	1500	171.7315	
18						

図 1　重回帰分析を利用して中古車の価格を予測する

中古車のデータを基に各項目同士の相関係数を求めてみると、年式と本体価格、走行距離と本体価格では負の相関があり、排気量と本体価格には正の相関があることが分かる。つまり、年式、走行距離、排気量という複数の説明変数を利用すれば、単回帰分析よりも本体価格がよく予測できそうである。そのための重回帰分析について考え方と操作の手順を追いかけよう。タイプの違いや、輸入車かどうかといった名義尺度のデータを重回帰分析で利用するための方法についても見ていく。

図 1 のデータは、前回のデータと同じものです。このデータを使って重回帰分析を行います。簡単な関数を利用するだけでできますが、説明変数の範囲がやや複雑になるので、「名前」機能などを使って、引数を簡単に指定できるようにします。

年式と本体価格に負の相関がある（年式が大きくなれば価格が下がる＝新しい車ほど安い）という結果には違和感がありますね。これには理由があります。前回は排気量に注目して電気自動車（EV）を除外し、本体価格に注目して極端に高価な中古車を外れ値として除外しましたが、年式にも外れ値があるということです（例えば、8 行目のアルファロメオは 1962 年式ですね）。このことについては、後ほど触れるので、取りあえずこのまま重回帰分析を進めましょう。

最後に、二次式以上の回帰式を利用した多項式回帰の例についても紹介します。具体的には、第 13 回で取り上げた、気温と電気器具による CO_2 排出量のデータのようなU字形の分布について回帰分析を行います。重回帰分析のちょっとした応用で簡単にできます。

この記事は、データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 15 回です。第 13 回から第 15 回までは「関係」に注目し、相関や回帰分析などについて見ていきます。前回は、1 つの説明変数から目的変数の値を予測する単回帰分析を紹介しました。今回は、複数の説明変数から目的変数の値を予測する重回帰分析の考え方や取り扱いを見ていきます。また、回帰式が二次以上の式になる多項式回帰にも取り組みます。トップページから全体の目次が参照できます。

この記事で学べること

今回は以下のようなポイントについて、分析の方法や目の付けどころを見ていきます。

- 重回帰分析による回帰式の求め方 …… 重回帰式の係数と定数項を求める
- 重回帰分析による予測 …… 回帰式に値を代入して予測を行う
- 名義尺度のデータを説明変数にする方法 …… カテゴリー変数を数値化する
- 重回帰分析での留意点を知る …… 多重共線性の意味と確認の方法
- 多項式回帰についても理解する …… 二次関数のような直線的でない関係でも回帰分析を行う

今回は、変数同士の相関係数を全て求めるところからスタートします。サンプルファイルの利用についての説明の後、本編に進みましょう。

サンプルファイルの利用について

本稿では、表計算ソフトを使って手を動かしながら学んでいきます。表計算ソフト Microsoft Excel 用の .xlsx ファイルをダウンロードできるようにしています。デスクトップ版の Excel が手元にない場合は、Microsoft アカウントで使える無料の Microsoft 365 オンライン、もしくは Google アカウントで使える無料の Google スプレッドシート（Google Sheets）をお使いください。Microsoft 365 オンラインの場合は、.xlsx ファイルを OneDrive にアップロードしてから開いてください。Google スプレッドシートの場合は、ファイルを共有して参照できるようにします。リンクを開き、メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。

相関行列を作成しよう ～ 説明変数を選定する

今回の例では、年式、走行距離、排気量を説明変数とすることにしていますが、本来はどの項目（変数）を説明変数とするかを選定する必要があります。そのためには、目的変数と何らかの関係がある項目を選ぶ必要があります。変数同士の関係を知るには相関係数を求めたり、散布図を描くというのが定石ですね。そこで、それぞれの相関係数を求めてみましょう。

といっても、幾つかの CORREL 関数にいちいち異なる範囲を指定するのはとても面倒です。そこで、図 2 では、範囲に名前を付けて関数を入力しやすくしています。ただし、ブック全体で同じ名前を使っているので、以下の操作については既に設定済みとなっています（つまり、後述するサンプルファイルでは図 2 の操作は必要ないので、操作方法の確認だけしておいてください）。

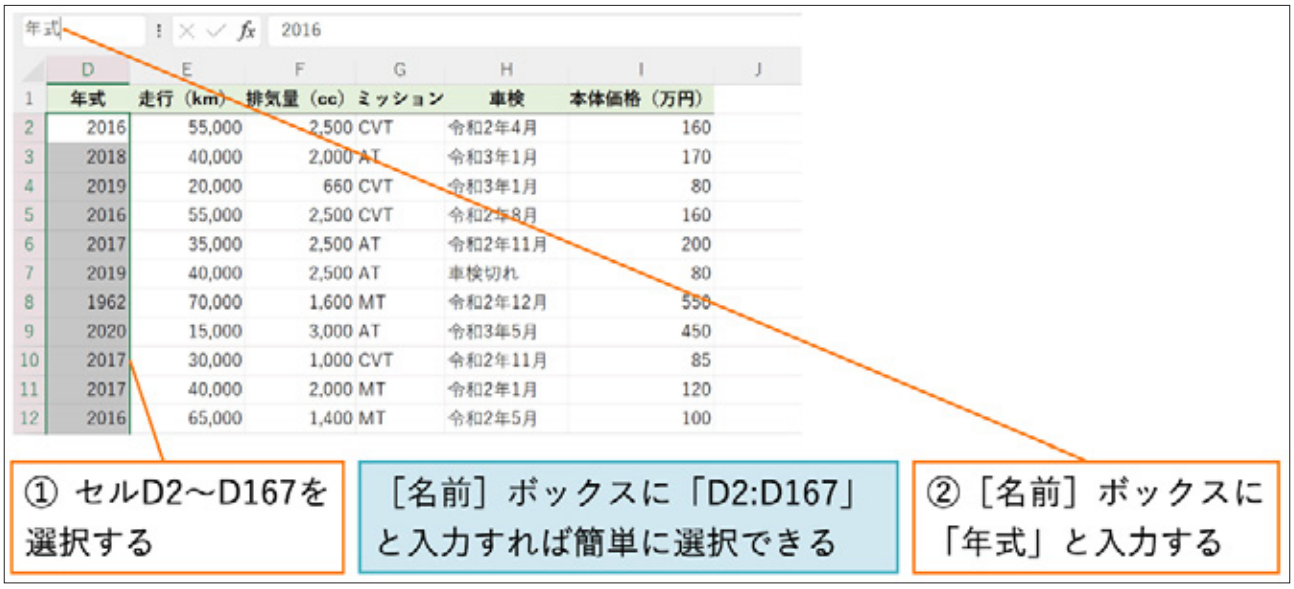


図 2　数式を簡単にするためにあらかじめ範囲に名前を付けておく

［数式］タブで［名前の定義］を選択すれば、範囲に名前を付けておくことができる。数式の中では、範囲指定の代わりにここで付けた名前が使える。ここでは、年式という名前をブック全体で使える名前として、セル **D2** ～ **D167** に付けている。走行、排気量、本体価格についても同様に名前を付けておくとよい。

サンプルファイルをこちらからダウンロードして[中古車価格]ワークシートを開いてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。

実際の作業は不要ですが、事前に設定した際の手順は以下の通りです。データは 167 行目まで入力されています。

- ［名前］ボックスに「D2:D167」と入力してセル **D2** ～ **D167** を選択する
  - ・セル **D2** をクリックしてから［Ctrl］＋［Shift］＋［↓］キーを押してもセル **D2** ～ **D167** が選択できる
- ［名前］ボックスに「年式」と入力する

同じ手順で走行という名前をセル **E2** ～ **E167** に、排気量という名前をセル **F2** ～ **F167** に、本体価格という名前をセル **I6** ～ **I167** に付けてあります。このようにして、範囲に名前を付けておくと、数式や関数の引数としてセルアドレスを指定する代わりに名前が使えるようになります。

Excel では、年式については、セル **D1** ～ **D167** を選択し、［数式］タブの［定義された名前］グループにある［選択範囲から作成］ボタンをクリックし、[選択範囲から名前を作成］ダイアログボックスで［以下に含まれる範囲から名前を作成］の［上端行］をチェックしても作成できます。上端行が名前となり、その下の範囲が名前で表される範囲となります。

なお、ブック全体ではなく、特定のワークシートで使われる名前にするには［数式］タブの［定義された名前］グループにある［名前の定義］ボタンをクリックし、［新しい名前］ダイアログボックスで、［範囲］のリストからワークシート名を選択して、名前と範囲を指定します。

続いて、図 3 で CORREL 関数を入力します。ここから実際の操作に入るので、サンプルファイルの［中古車価格］ワークシートを開き、図の後に記した箇条書きの手順に従って取り組んでみてください。また、ここから重回帰分析の結果までの作成例は［中古車価格（作成例）］ワークシートに含まれています。

図 2 と図 3 の手順については動画も用意してあります。操作方法を一つ一つ丁寧に追いかけたい方はぜひご視聴ください。

では、CORREL 関数を入力しましょう。手順は図 3 に示した通りです。名前機能を使えば、数式が簡潔に入力できることが分かりますね。

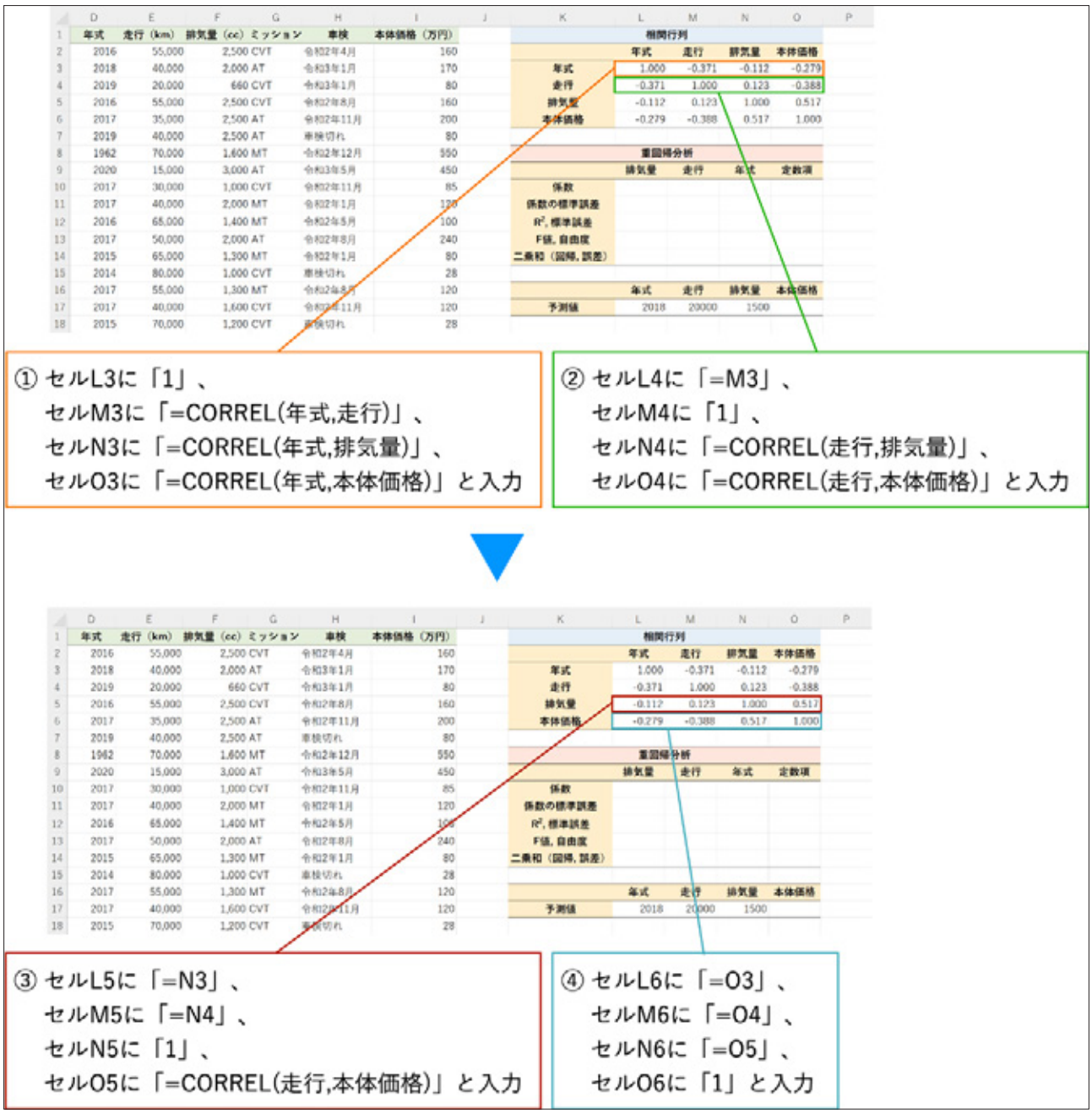


図 3　CORREL 関数を利用して相関行列を作成する

自分自身との相関係数は **1** なので、表の対角線の位置（セル **L3**、**M4**、**N5**、**O6**）には単に「1」と入力するだけで構わない。セル **M4** には =CORREL(年式 , 走行) と入力する。対角線を挟んで対称の位置にある値は同じ。例えば、セル **L4** には =M3 と入力するだけでよい。

手順は図中に示した通りですが、以下に箇条書きにしておきます。

- 3 行目
  - セル **L3**：1
  - セル **M3**：=CORREL(年式 , 走行)
  - セル **N3**：=CORREL(年式 , 排気量)
  - セル **O3**：=CORREL(年式 , 本体価格)
- 4 行目
  - セル **L4**：=M3
  - セル **M4**：1
  - セル **N4**：=CORREL(走行 , 排気量)
  - セル **O4**：=CORREL(走行 , 本体価格)
- 5 行目
  - セル **L5**：=N3
  - セル **M5**：=N4
  - セル **N5**：1
  - セル **O5**：=CORREL(排気量 , 本体価格)
- 6 行目
  - セル **L6**：=O3
  - セル **M6**：=O4
  - セル **N6**：=O5
  - セル **O6**：1

実は、文字列をセル参照に変換するための INDIRECT 関数を使うと、もっと簡単に相関行列が作成できます。手順はセル **L3** に =CORREL(INDIRECT($K3), INDIRECT(L$2)) と入力し、セル **O6** までコピーするだけです。

INDIRECT 関数を使いこなすには、文字列とセル参照の違いを明確に理解しておく必要があります。例えば、「"A4"」という**文字列**は、単に「"A"」という文字と「"4"」という文字が並んだだけのデータですが、A4 というセル参照はセル A4 を指し示すものです。例えば、INDIRECT("A4") とすると、「"A4"」という文字列がセル参照 A4 に変換されます。

=CORREL(INDIRECT($K3), INDIRECT(L$2)) であれば、セル **K3** に入力されている「" 年式 "」という文字列が INDIRECT 関数によって年式という名前（セル参照）に変換されます。同様に、セル **L2** に入力されている「" 年式 "」という文字列もやはり年式という名前に変換されます。従って、この式は =CORREL(年式 , 年式) と入力したのと同じことになります。その式を右のセル **M3** にコピーすると、引数の **L$2** の列番号だけが変わり **M$2** になるので、関数は =CORREL(INDIRECT($K3), INDIRECT(M$2)) となります。セル **M2** に入力されている「" 走行 "」という文字列が INDIRECT 関数によって走行という名前に変換されるので、=CORREL(年式 , 走行) と入力したのと同じことになります。このようにして、セル **O6** までの全ての値が求められます。

INDIRECT 関数はちょっと難しいですが、サンプルファイルの［INDIRECT 関数の利用例］ワークシートに幾つか例を掲載しているので、参考にしてください。ただし、INDIRECT 関数が分かりにくいという方や、INDIRECT 関数を使うとそのワークシートを使う他の人が理解できないかもしれないという懸念のある方は、図 3 の手順で示したように入力した方がいいかと思われます。

相関行列のお話が長くなりましたが、これで目的変数と関係のありそうな説明変数の候補が選定できます。

コラム　全ての変数の組み合わせについて散布図を描くには

説明変数を選択したり、外れ値の除外を行ったりするために、それぞれの散布図を描いて変数同士の関係を可視化したいところですが、Excel ではかなり面倒です（操作そのものは簡単ですが、組み合わせが多いので 1 つずつ散布図を描いていくのは手間が掛かります）。そこで、Python のプログラムを書いて、全ての組み合わせについて散布図を描いてみます。

連載の第 12 回で紹介した seaborn モジュールの pairplot 関数を使えば簡単です。サンプルプログラムはこちらから参照できます。リンクをクリックすれば、ブラウザが起動し、Google Colaboratory で以下（リスト 1）のコードが表示されます（Google アカウントでのログインが必要です）。

```
!pip install japanize-matplotlib  # グラフに日本語を表示するために必要。最初に１回だけ実行すればよい

import pandas as pd
import japanize_matplotlib
import seaborn as sns
df = pd.read_excel("https://github.com/Gessys/data_analysis/raw/main/15a.xlsx",
                   sheet_name="中古車価格", usecols="A:I")
sns.pairplot(df)
```

リスト 1　全てのペアについての散布図を描くためのコード

seaborn モジュールの pairplot 関数を使えば、全ての変数同士の散布図が描かれる。同じ変数同士（対角線上の部分）についてはヒストグラムが描かれる。

実行結果は図 4 の通りです。外れ値や分布の歪み、変数同士の関係が可視化できるので、前処理やモデルの選択などに役立てることができます。

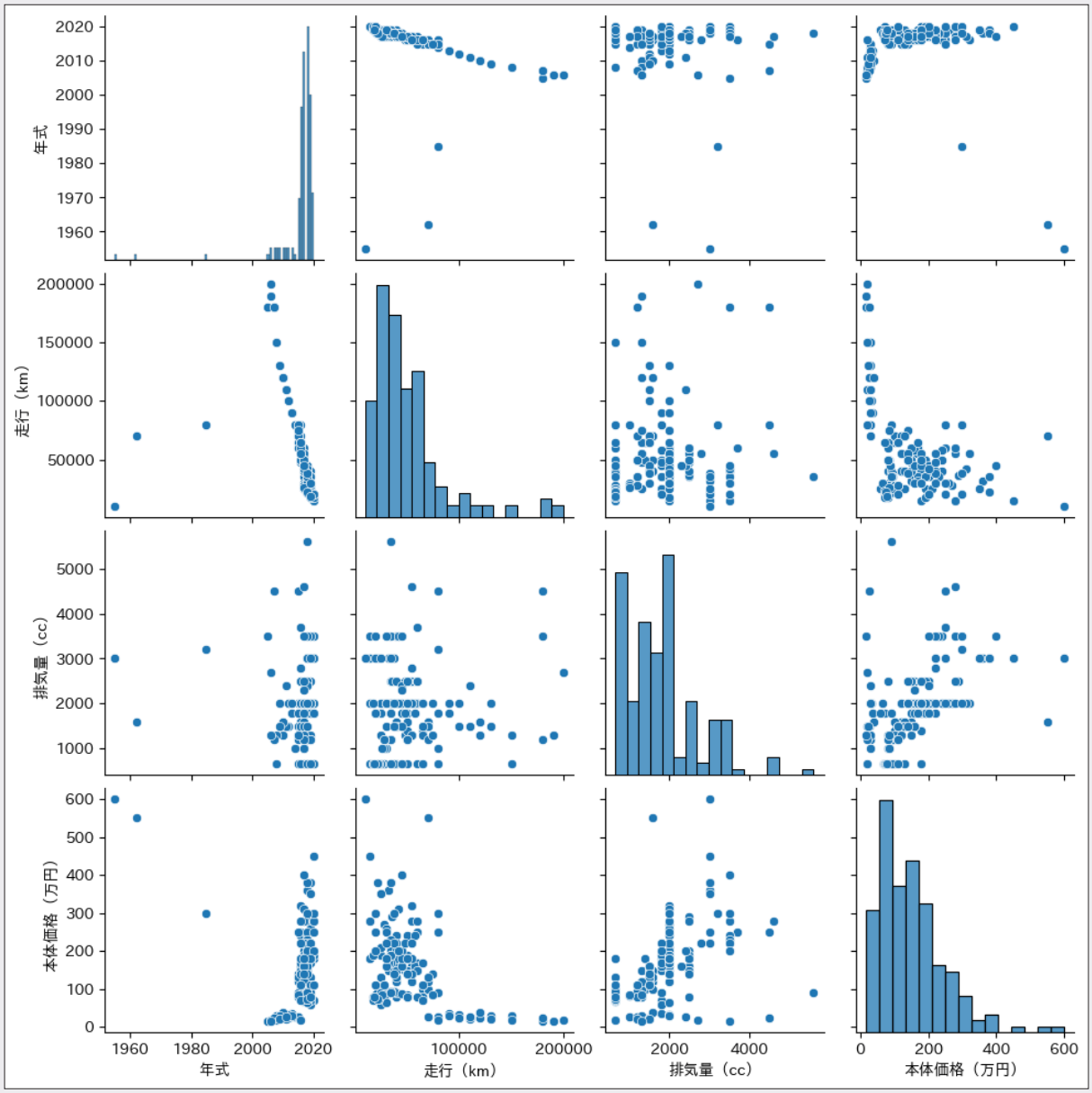


図 4　変数同士の組み合わせ全てについて散布図を描く

散布図を見ると、年式が極端に古い中古車が 3 つあることや、排気量が **5000cc** を超える特殊な車両があることも分かる。それらを除外すれば、年式と本体価格が直線的な関係になりそうである（このことについては、後のコラムで少し触れる）。走行距離については、**10 万 km** を超えると本体価格がほとんど下限に近くなることが分かる。さらに、ヒストグラムを見れば、排気量が小さい中古車（軽自動車）が多いことも分かる。軽自動車を別扱いにすることも考えられるが、今回はこのまま進めることとする。

重回帰分析を行う ～ 重回帰式の係数と定数項を求める

では、重回帰分析に取り組みましょう。重回帰分析では、説明変数が複数あるので、回帰式は以下のように表されます。

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n \qquad (1)$$

前回は、中学で学んだ一次関数 $y = ax + b$ と同じように単回帰式の定数項を最後に書きましたが、回帰分析では定数項を前に書くのが一般的です。(1) 式では a_0 が定数項です。係数や変数が複数個あり、a, b, c, ... のようなアルファベットを使っているとかえって煩雑になるので、a_0, a_1, ...、x_1, x_2 ... のように添え字を付けて区別します。

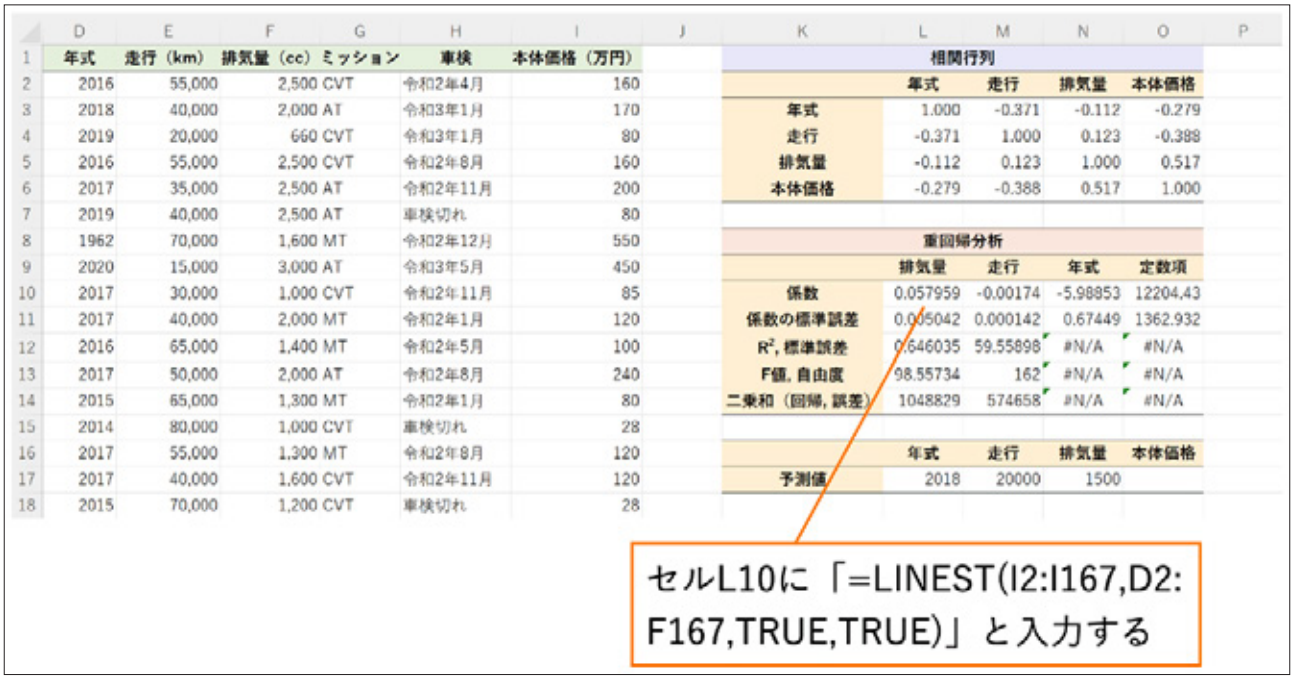
単回帰分析は、データの近くを通る（実測値と予測値の差の二乗和が最小になる）直線を描くイメージでしたが、説明変数が 2 つの重回帰分析は、データの近くを通る平面を描くイメージです。説明変数が 3 つ以上になると、さらに次元が増えるので図として表すのが難しくなりますが、同様にデータの近くを通る「超平面」を描くようなイメージです。(1) 式には x_0 の項が書かれていませんが、最初の項が a_0x_0 であり、$x_0 = 1$ であるものと考えられます。

中古車のデータでは、本体価格が目的変数 y に当たり、年式が説明変数 x_1 に、走行距離が説明変数 x_2 に、排気量が説明変数 x_3 に当たります。具体的に書けば、以下のようになりますね。

$$\text{本体価格} = a_0 + a_1 \times \text{年式} + a_2 \times \text{走行} + a_3 \times \text{排気量} \qquad (2)$$

というわけで、説明変数と目的変数を基に、それぞれの係数と定数項を求めてみましょう。そのためには LINEST 関数を使います。以下の例では補正項と呼ばれる値も求めてみます。**補正項**は、重回帰式の当てはまりの良さや係数の有効性についての検定を行うのに使われます。

手順は図 5 の後に記した通りです。セル **L10** に =LINEST(I2:I167,D2:F167,TRUE,TRUE) と入力しましょう。ただし、スピル機能が使えない場合は、あらかじめ結果が求められるセル範囲（セル **L10** ～ **O14**）を選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押してください。以降、この操作を「**配列数式を入力する**」と呼ぶことにします。



	D	E	F	G	H	I
1	年式	走行（km）	排気量（cc）	ミッション	車検	本体価格（万円）
2	2016	55,000	2,500	CVT	令和2年4月	160
3	2018	40,000	2,000	AT	令和3年1月	170
4	2019	20,000	660	CVT	令和3年1月	80
5	2016	55,000	2,500	CVT	令和2年8月	160
6	2017	35,000	2,500	AT	令和2年11月	200
7	2019	40,000	2,500	AT	車検切れ	80
8	1962	70,000	1,600	MT	令和2年12月	550
9	2020	15,000	3,000	AT	令和3年5月	450
10	2017	30,000	1,000	CVT	令和2年11月	85
11	2017	40,000	2,000	MT	令和2年1月	120
12	2016	65,000	1,400	MT	令和2年5月	100
13	2017	50,000	2,000	AT	令和2年8月	240
14	2015	65,000	1,300	MT	令和2年1月	80
15	2014	80,000	1,000	CVT	車検切れ	28
16	2017	55,000	1,300	MT	令和2年8月	120
17	2017	40,000	1,600	CVT	令和2年11月	120
18	2015	70,000	1,200	CVT	車検切れ	28

相関行列				
	年式	走行	排気量	本体価格
年式	1.000	-0.371	-0.112	-0.279
走行	-0.371	1.000	0.123	-0.388
排気量	-0.112	0.123	1.000	0.517
本体価格	-0.279	-0.388	0.517	1.000

重回帰分析				
	排気量	走行	年式	定数項
係数	0.057959	-0.00174	-5.98853	12204.43
係数の標準誤差	0.005042	0.000142	0.67449	1362.932
R^2, 標準誤差	0.646035	59.55898	#N/A	#N/A
F値, 自由度	98.55734	162	#N/A	#N/A
二乗和（回帰, 誤差）	1048829	574658	#N/A	#N/A

	年式	走行	排気量	本体価格
予測値	2018	20000	1500	

図 5　重回帰分析により係数や定数項を求める

目的変数の値はセル **I2:I167** に入力されている、説明変数の値はセル **D2**:**F167** に入力されている。
第 3 引数には定数項を求めるか、定数項を **0** とするかを指定する。ここでは **TRUE** を指定して定数項を求める。
最後の引数には補正項を求めるかどうかを指定する。ここでは **TRUE** を指定して補正項も求める。係数と定数項だけを求めるのであれば、**FALSE** でも構わない。
LINEST 関数だけで全ての値が求められるが、係数が元の項目と逆の順序（排気量、走行、年式）になることに注意。

図 5 の解説にも記してありますが、**求められた係数は元の項目と逆の順序になっている**ことに注意してください。得られた値を（2）の回帰式に当てはめると、以下のようになります。

$$\text{本体価格} = 12204.43 - 5.98853 \times \text{年式} - 0.00174 \times \text{走行} + 0.57959 \times \text{排気量} \qquad (3)$$

セル **L11** ～ **O14** の値が補正項です。これらの補正項を使った検定については、この連載の次に企画している推測統計編で取り扱うことになるので、これ以上は触れませんが、セル **L12** の値（R^2：決定係数）とセル **M12** の値（SE：標準誤差）は、重回帰式の当てはまりの良さを評価するのに使われるので、少しばかり気にしておいてください（決定係数は後でまた登場します）。

SE（標準誤差）は RMSE（二乗平均平方根誤差）と似ていますが、SE は、実測値と予測値の差の二乗和（セル **M14** の値）を自由度（セル **M13** の値）で割ってその√を求めたもので、RMSE は実測値と予測値の差の二乗和をデータの個数（この例では 166）で割ってその√を求めたものです。なお、セル **M13** の自由度は回帰分析での検定に使われる標準誤差の自由度で、データの個数－説明変数の個数－ 1 で求められます（この例なら 166 － 3 － 1 ＝ 162）。

説明変数として利用している年式や走行距離、排気量は値の大きさが異なるので、元の値を使って重回帰分析を行っても、係数の大きさをそのままでは比較できません。係数の大きさを比較できるようにするには、元の値を標準化したものを説明変数とするとよいでしょう。

標準化の方法はこの連載の第 6 回で紹介した通り、元の値から平均値を引き、標準偏差で割るだけです（STANDARDIZE 関数を使ってもできます）。標準化を行うと、全体の平均が **0**、標準偏差が **1** となるように値が調整されます。

予測に使う値についても、元の値ではなく、標準化した値を指定します。サンプルファイルの［(参考）標準化］ワークシートにその例を含めてあるのでご参照ください。

重回帰分析による予測を行う

前項で見た（2）式に年式と走行距離、排気量を代入すれば本体価格が予測できます。（2）式は以下の通りでしたね（再掲）。係数と定数項はセル **L10** ～ **N10** で求められています。

$$\text{本体価格} = a_0 + a_1 \times \text{年式} + a_2 \times \text{走行} + a_3 \times \text{排気量} \qquad (2)$$

例えば、セル **L17** ～ **N17** に入力されている年式、走行距離、排気量を基に本体価格を予測するなら、=O10+N10*L17+M10*M17+L10*N17 と入力すれば、171.7315 という結果が得られます（セル **O18** などの空いているセルに入力してみてください）。しかし、係数や定数項を求めずに、予測値だけを求めたいのであれば、TREND 関数を使うのが簡単です。セル **O17** に =TREND(I2:I167,D2:F167,L17:N17,TRUE) と入力しましょう（図 6）。

	D	E	F	G	H	I	J	K	L	M	N	O	P
1	年式	走行（km）	排気量（cc）	ミッション	車検	本体価格（万円）			相関行列				
2	2016	55,000	2,500	CVT	令和2年4月	160			年式	走行	排気量	本体価格	
3	2018	40,000	2,000	AT	令和3年1月	170		年式	1.000	-0.371	-0.112	-0.279	
4	2019	20,000	660	CVT	令和3年1月	80		走行	-0.371	1.000	0.123	-0.388	
5	2016	55,000	2,500	CVT	令和2年8月	160		排気量	-0.112	0.123	1.000	0.517	
6	2017	35,000	2,500	AT	令和2年11月	200		本体価格	-0.279	-0.388	0.517	1.000	
7	2019	40,000	2,500	AT	車検切れ	80							
8	1962	70,000	1,600	MT	令和2年12月	550			重回帰分析				
9	2020	15,000	3,000	AT	令和3年5月	450			排気量	走行	年式	定数項	
10	2017	30,000	1,000	CVT	令和2年11月	85		係数	0.057959	-0.00174	-5.98853	12204.43	
11	2017	40,000	2,000	MT	令和2年1月	120		係数の標準誤差	0.005042	0.000142	0.67449	1362.932	
12	2016	65,000	1,400	MT	令和2年5月	100		R^2, 標準誤差	0.646035	59.55898	#N/A	#N/A	
13	2017	50,000	2,000	AT	令和2年8月	240		F値, 自由度	98.55734	162	#N/A	#N/A	
14	2015	65,000	1,300	MT	令和2年1月	80		二乗和（回帰, 誤差）	1048829	574658	#N/A	#N/A	
15	2014	80,000	1,000	CVT	車検切れ	28							
16	2017	55,000	1,300	MT	令和2年8月	120			年式	走行	排気量	本体価格	
17	2017	40,000	1,600	CVT	令和2年11月	120		予測値	2018	20000	1500	171.7315	
18	2015	70,000	1,200	CVT	車検切れ	28							

セルO17に「=TREND(I2:I167,D2:F167,L17:N17,TRUE)」と入力する

図 6　重回帰分析による予測を行う

TREND 関数には、既知の目的変数 y と既知の説明変数 x、予測に使う新しい x の値を指定する。
最後の引数は定数項を求めるか、**0** と見なすかの指定。**TRUE** を指定すれば定数項を求めることになる。

2018 年式で走行距離 **20000km**、排気量が **1500cc** の中古車の価格は **171.7315 万円**と予測できました。取りあえず、これで重回帰分析による係数と定数項の計算、予測の方法については一通り見たことになります。

ここからは、重回帰分析の精度を評価し、より良い予測ができるようにする方法について見ていきましょう。精度を評価するための決定係数や RMSE については前回の単回帰分析でも簡単に紹介しましたが、今回はそれらに加えて自由度調整済み決定係数の求め方も説明します。

重回帰分析の精度を可視化／評価する

重回帰分析によって求めた回帰式の当てはまりの良さを可視化するには、

$$\text{目的変数}y\text{とその予測値}\hat{y}\text{を散布図にする}$$

のが最も簡単です。サンプルファイルの［中古車価格（散布図）］ワークシートをご覧ください。I 列の本体価格と、J 列で求めた予測値を散布図にし、決定係数（R^2 値）も併せて表示します。手順は図 7 の後に箇条書きで記してあります。

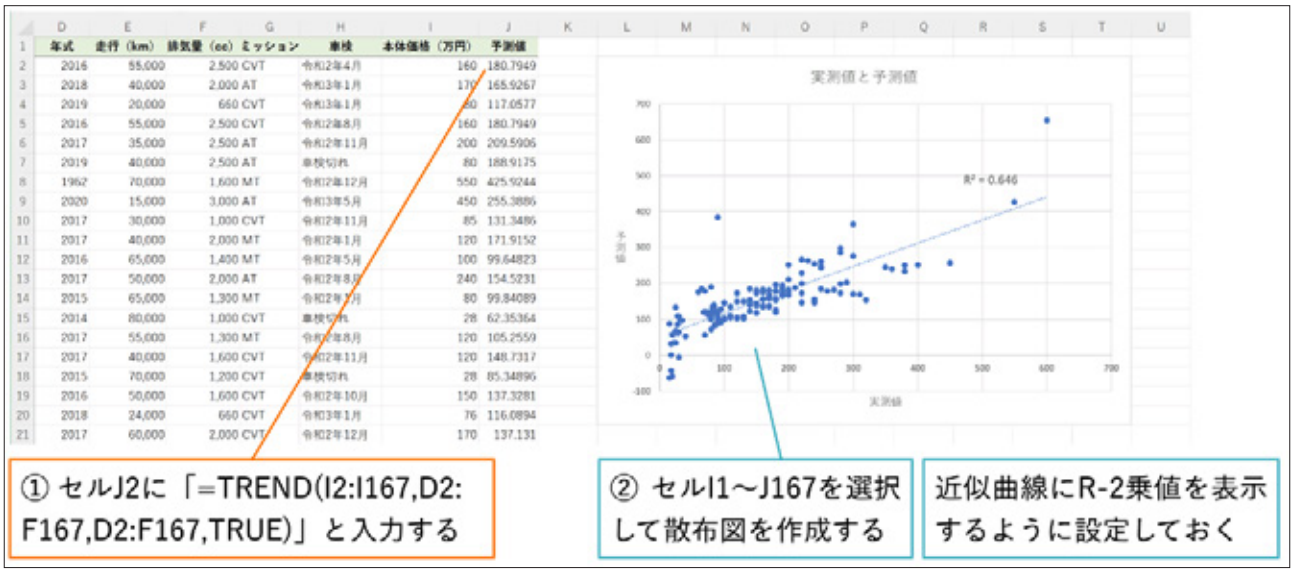


図 7　回帰式に当てはまりの良さを可視化する

実測値と予測値の散布図が直線的になれば、回帰式の当てはまりが良いと考えられる。決定係数（R^2）は LINEST 関数の補正項でも求められているが、散布図の近似曲線に表示することもできる。

手順は以下の通りです。散布図の位置やサイズ、書式の設定などについては省略しています。

散布図の作成（Excel での操作手順）

- セル **J2** に =TREND(I2:I167,D2:F167,D2:F167,TRUE) と入力する（予測値が全て求められる）
- スピル機能が使えない場合は配列数式として入力してもいいが、セル **J2** に =TREND(I2:I167,D2:F167,D2:F2,TRUE) と入力してセル **J167** までコピーする方が簡単
- セル **H1** ～ **I167** を選択する
- ［挿入］タブを開き、［散布図（X, Y）またはバブルチャートの挿入］ボタンをクリックする
- ［散布図］を選択する（これで散布図が作成される）
- 系列（作成された散布図のいずれかの点）を右クリックし、［近似曲線の追加］を選択する
- ［近似曲線の書式設定］作業ウィンドウで［グラフに R-2 乗値を表示する］チェックボックスをオンにする

散布図の作成（Google スプレッドシートでの操作手順）

- セル **J2** に =ARRAYFORMULA(TREND(I2:I167,D2:F167,D2:F167,TRUE)) と入力する（予測値が全て求められる）
- セル **H1** ～ **I167** を選択する
- メニューバーから［挿入］－［グラフ］を選択する
- ［グラフエディタ］作業ウインドウで［グラフの種類］リストから［散布図］を選択する（これで散布図が作成される）
- ［グラフエディタ］作業ウインドウで［カスタマイズ］をクリックする
- ［系列］をクリックして下位の設定項目を表示し、［トレンドライン］チェックボックスをオンにする
- ［系列］の下の［決定係数を表示する］チェックボックスをオンにする

図 7 を見ると、回帰式の当てはまりは比較的良いように思われます。決定係数は **0.646** となっていますが、一般に **0.5** 以上あれば当てはまりが良いと考えられます。

ここでは予測値を全て TREND 関数で求めているので、毎回、係数と定数項の計算が行われます。従って、係数と定数項が既に求められているのであれば、回帰式に従って予測値を求めた方が速く計算できます。といっても、この程度の規模であれば計算の速さはそれほど変わりません。

決定係数に加えて、RMSE（二乗平均平方根誤差）も求めておきましょう。RMSE を求めるための式は以下の通りでしたね。

$$\sqrt{\frac{1}{n}\sum(y_i-\hat{y}_i)^2} \qquad (4)$$

y_i は I 列の値です。

$$\hat{y}_i$$

は y の予測値なので J 列の値に当たりますね。n はサンプルサイズ（データ数）です。従って、図 7 の表を基にこの式に従って RMSE が計算できます……が、実は（4）式の一部の、

$$\sum (y_i - \hat{y}_i)^2 \qquad (5)$$

は LINEST 関数の補正項で既に求められています（**誤差の二乗和**あるいは**残差平方和**などと呼ばれます）。というわけで、［中古車価格］ワークシートに戻って計算してみましょう（図 8）。

図 8　RMSE の値を求める

LINEST 関数で求めた補正項のうち、セル **M14** が誤差の二乗和となっている。この値をデータの件数で割り、その√を求めれば RMSE の値となる。

手順は図 8 に示した通りです。以下のように入力しましょう。

- セル **K20**：COUNT 関数を使ってデータの個数を求めるため、=COUNT(I2:I167) と入力
- セル **K22**：誤差の二乗和をデータの個数で割って、その√を求めるため、=SQRT(M14/K20) と入力
- セル **L20**：COUNTA 関数を使って説明変数の個数を求めるため、=COUNTA(D1:F1) と入力

セル **L20** で求めている説明変数の個数はここでは使いません。次のコラムで使うので、ついでに求めておいただけです。

コラム　決定係数に関する留意点 ～ 自由度調整済み決定係数などのお話

一般に、決定係数は説明変数の数が多くなると値が大きくなります。そのため、説明変数の数が多い場合には、説明変数の数を考慮した自由度調整済み決定係数が使われます。自由度調整済み決定係数は以下の式で求められます。詳細については、こちらなどをご参照ください。

決定係数 R^2

$$1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2} \qquad (6)$$

自由度調整済み決定係数 $adj\ R^2$

$$1-\frac{\frac{\sum(y_i-\hat{y}_i)^2}{n-k-1}}{\frac{\sum(y_i-\bar{y})^2}{n-1}} = 1-(1-R^2)\frac{n-1}{n-k-1} \qquad (7)$$

$\hat{y}_i$はy_iの予測値、$\bar{y}$はyの平均値です。

n はサンプルサイズ（データの件数）、k は説明変数の個数を表します。例えば、セル **L12** に決定係数が求められており、セル **K20** にデータの件数が、セル **L20** に説明変数の個数が求められているのであれば、セル **K25** に =1-(1-L12)*(K20-1)/(K20-L20-1) と入力すれば、自由度調整済み決定係数が求められます（結果は **0.6395** となります）。［中古車価格（重回帰）］サンプルファイルには（7）式によって計算した結果も含めてあるので、参照してみてください。

ちなみに、（6）式の分母は**全変動の二乗和**と呼ばれます。この値は回帰変動の二乗和と誤差の二乗和（図 8 のセル **L14** とセル **M14** の和）となっています。分子は**誤差の二乗和**ですね、従って、空いているセルに「=1-M14/(L14+M14)」と入力しても決定係数が求められます。同様に、自由度調整済み決定係数は「=1-(M14/(K20-L20-1))/((L14+M14)/(K20-1))」でも求められます。

注意点：決定係数が大きいから当てはまりがよいとは限らない？！

もう一点、補足です。元のデータでは RMSE が **58.8370** となっており、決定係数は **0.646035** となっています。さらに分析の精度を上げるために、年式が極端に古い中古車と排気量が **5000cc** を超える特殊な車両を除外すると、RMSE は **52.0031** と小さくなるのですが、決定係数も **0.643666** とわずかに小さくなってしまいます（サンプルファイルの［参考（外れ値の除外）］ワークシートに計算例があります）。

誤差が小さくなったのに、決定係数で表される回帰式の当てはまりは悪くなったということでしょうか。実は、これは決定係数を求めるための（6）式で分母（全変動の二乗和）が分子（回帰の変動の二乗和）に比べて小さくなったためです。従って、決定係数が大きいからといって当てはまりがよいとか、決定係数が小さいから当てはまりが悪いと直ちには言えない場合もあることに注意が必要です。このことについては、こちらに詳しい解説があります。

特徴量を作成する

ここまでは、中古車の価格が年式、走行距離、排気量といった値で決まるものとして分析を進めてきました。しかし、輸入車であるか国産車であるかによっても中古車の価格が異なることは十分考えられます（一般に輸入車の方が高価だと考えられますね）。このように、元々のデータには存在しなかった項目（特徴量）を新たに追加することによって、分析の精度を向上させることができる可能性もあります。

では、輸入車であるか、国産車であるかをどのようにして数値で表せばいいでしょうか。輸入車か国産車かというように、2 つのカテゴリーに分けられる場合は、輸入車を **1** とし、国産車を **0** とする方法が使えます。回帰式の中での、その項を $a_i x_i$ とすると、輸入車の場合、その項の値は $a_i \times 1 = ai$ となり、国産車の場合は $a_i \times 0 = 0$ となりますね。輸入車には ai だけ価格が上乗せされるというわけです。

サンプルファイルの［中古車価格（特徴量）］ワークシートを使って手順を追いかけてみましょう。ここからの手順については動画も用意してあります。操作方法を一つ一つ丁寧に追いかけたい方はぜひご視聴ください。

まず、輸入車であるかどうかを表す項目を作成しましょう（図 9）。UNIQUE 関数を使ってメーカーの一覧を作成し、輸入車であれば **1** を、国産車であれば **0** を入力した表を作ります。次に、XLOOKUP 関数を使ってそれぞれの中古車が輸入車であるか国産車であるかを B 列に表示されるようにします。

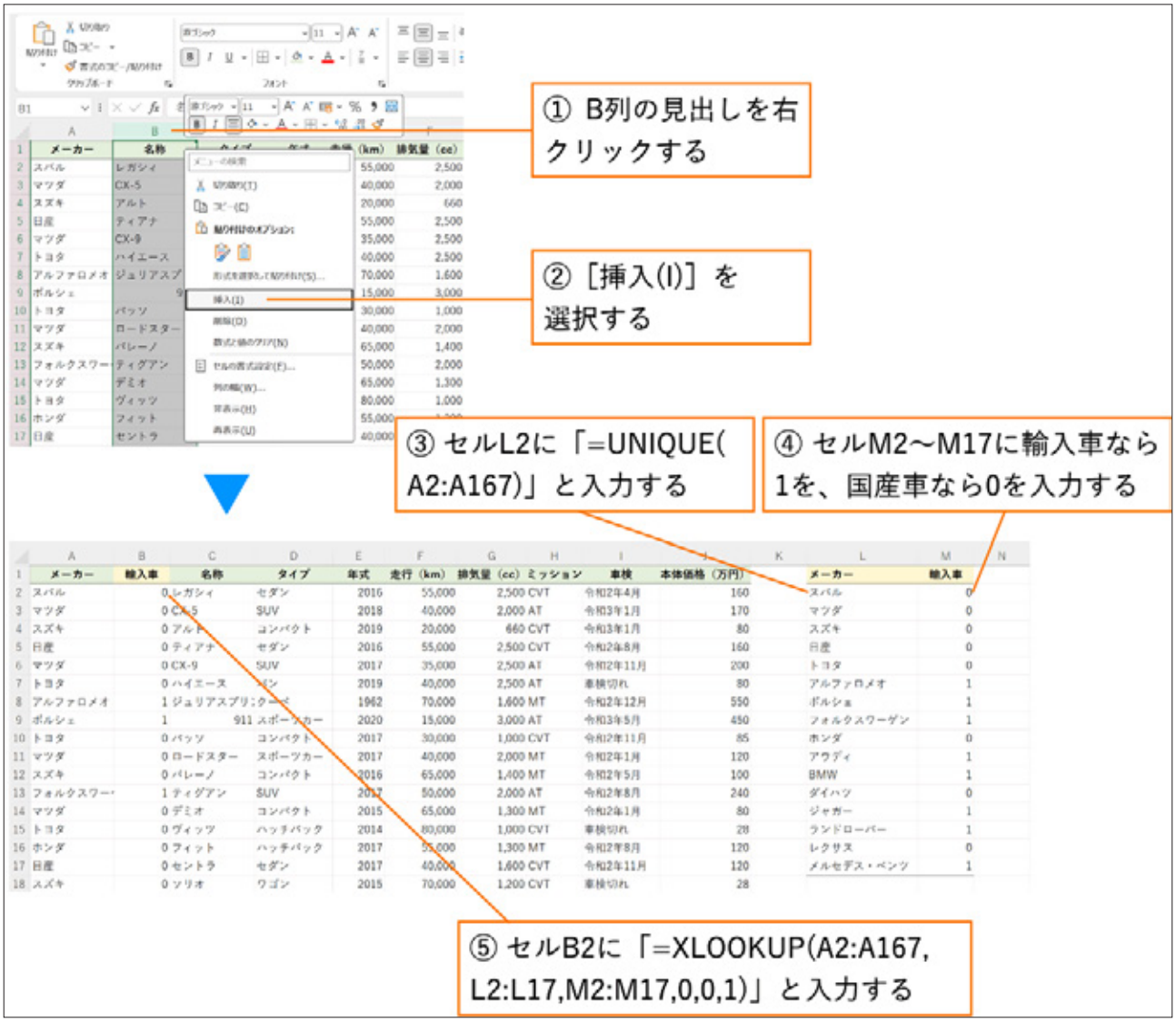


図 9　輸入車かどうかを表す特徴量を作成する

Excel 2019 以前のバージョンでは、以下の手順に登場する UNIQUE 関数が使用できません。それらのバージョンをお使いの場合は、Microsoft 365 オンラインなどで代用するか、［中古車価格（特徴量作成例）］ワークシートを参照するだけにしてください。

- B 列の見出しを右クリックし、［挿入］を選択する（B 列の手前に新しい列が挿入される。セル **B1** に「輸入車」という見出しを入力しておくとよい）
- セル **L2** に `=UNIQUE(A2:A167)` と入力する（メーカーの一覧表が作成される）
- セル **M2** ～ **M17** に、輸入車なら **1** を、国産車なら **0** を読者自身が手動で 1 つずつ入力する
- セル **B2** に `=XLOOKUP(A2:A167,L2:L17,M2:M17,0,0,1)` と入力する
  - スピル機能が使えない場合は、配列数式として入力する
  - Google スプレッドシートでは `=ARRAYFORMULA(XLOOKUP(A2:A167,L2:L17,M2:M17,0,0,1))` と入力する

レクサスはトヨタ自動車の高級車ブランドなので、ここでは国産車としておきます。UNIQUE 関数を使えば、複数回登場する値を 1 つにまとめた重複のないリストが作成できます。XLOOKUP 関数は、セル **A2** ～ **A167** のメーカー名をセル **N2** ～ **N17** から検索し、それに対応するセル **O2** ～ **N17** の値（**0** か **1** の値）を返します。XLOOKUP 関数はこの連載の第 3 回で取り上げましたが、あまりなじみのない方は、こちらで形式を確認しておいてください。

これで新たな特徴量（説明変数）が作成できました。あとはこれまでと同様に重回帰分析を行うだけです。セルアドレスをそのまま使うと見づらいので、あらかじめ以下のような名前を付けてあります（B 列の手前に列を挿入すると列番号が変わってしまうので、作成例である［中古車価格（特徴量作成例）］ワークシートの範囲に名前を付けてあります）。

- セル **E2** ～ **G167** に説明変数 1 という名前を付けておく（年式、走行、排気量に当たる）
- セル **B2** ～ **B167** に説明変数 2 という名前を付けておく（輸入車に当たる）

LINEST 関数を入力すれば係数や定数項が求められます。ただし、説明変数の範囲が離れた列にあるので、HSTACK 関数を使って列をつなぐことにします。セル **M2**1 に =LINEST(J2:J167,HSTACK(説明変数 1, 説明変数 2),TRUE,TRUE) と入力してください（図 10）。セル **J2** ～ **J167** の値は［中古車価格］ワークシートのセル **I2** ～ **I167** と同じ値なので、本体価格という名前を使って、=LINEST(本体価格 ,HSTACK(説明変数 1, 説明変数 2),TRUE,TRUE) と入力しても同じ結果になります。

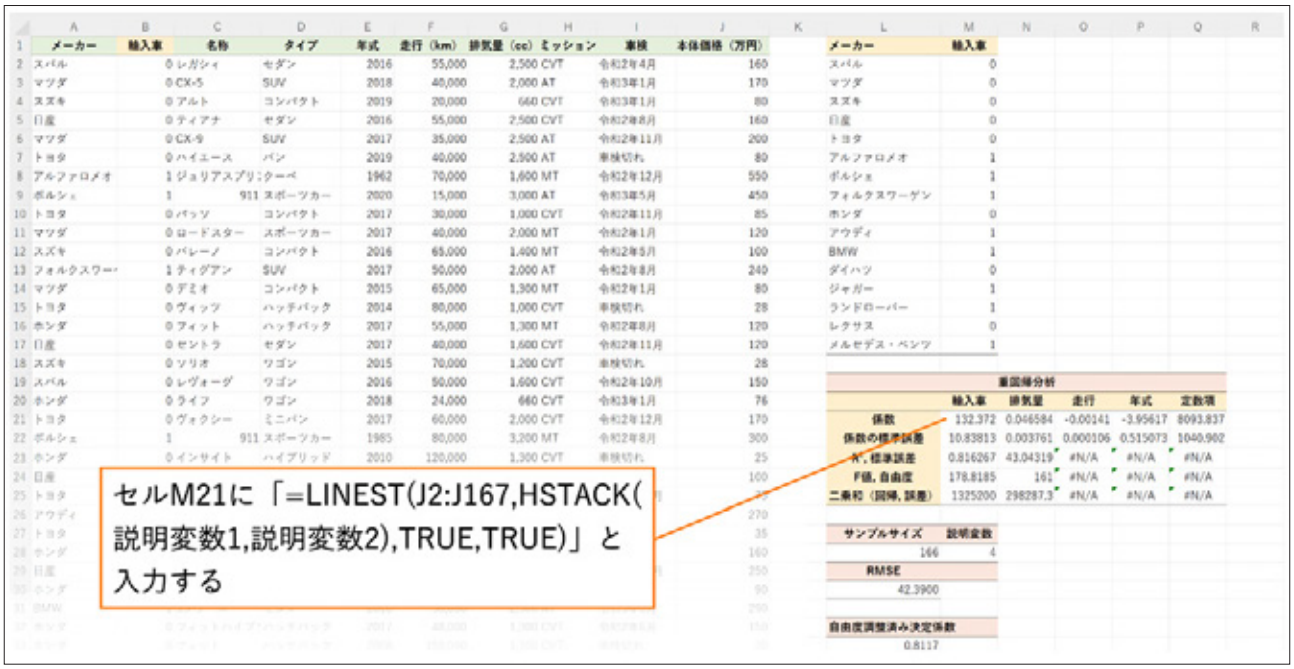


図 10　輸入車かどうかを表す特徴量を追加して重回帰分析を行う

LINEST 関数に指定する説明変数の範囲が離れた位置にあるので、HSTACK 関数で列をつないでひとまとめにする。それ以外は既に見た通り。

決定係数や RMSE の求め方は既に見た通りです。実際に入力されている式については［中古車価格（特徴量作成例）］をご参照ください。これまで（［中古車価格（作成例）］ワークシート）は決定係数が **0.6460**、RMSE が **58.8370** でしたが、「輸入車」という特徴量を追加すると、決定係数が **0.8163**、RMSE が **42.3900** となり、精度がかなり上がることが分かります。

名義尺度のデータを数値化する

中古車の本体価格に関係があり、数値で表せない値は他にどのようなものがあるでしょうか。

候補として考えられるのは車両のタイプとミッション（変速機の種類）です。スポーツカーや SUV は高価で、コンパクトカーやワゴンは比較的安価であると思われるので、車両のタイプは説明変数の候補として有力です。しかし、車両のタイプは種類が多く、表が大きくなり過ぎてしまうので、ここでは、分かりやすさを優先させることとし、ミッション（CVT、AT、MT の 3 種類）を例に取り上げます。一概には言えませんが、CVT は比較的小さな車種で、AT は比較的大きな車種で、MT は比較的古い車種やスポーツカーで採用されています。

車検の期限も関係ありそうですが、ここでは名義尺度の変数の取り扱いを見るので、取り上げていません（車検費用は本体価格には含まれないという理由もありますが）。ただし、車検切れかそうでないかといった分け方もアリかと思われます。車検切れの車両はメンテナンスされずに放置されていた、程度のよくないものであると考えられるからです。車検の期限は数値で表せますが、車検の残りがあるということと、車検切れであるということとは質的に異なるというわけです。実際、車検切れの車両は全て 100 万円未満です。

前項で見たように、カテゴリーが 2 つに分けられる場合は一方を **0**、もう一方を **1** とすればうまくいきました。カテゴリーが 3 つ以上の場合には、それぞれのカテゴリーを項目にし、一致する場合には **1** を、一致しない場合には **0** を入れます。このようにして作られた変数のことを**ダミー変数**と呼びます（図 11）。

図 11　名義尺度のデータをダミー変数で数値化する

カテゴリーを表す名称を項目名とし、一致する場合は 1 を、一致しない場合は 0 を入れる。回帰式では、それぞれの変数に係数が掛けられるが、例えば、CVT の中古車は AT と MT が 0 なので、CVT の係数だけが掛けられることになる。

→目次に戻る

実は、**カテゴリーが n 個の場合、ダミー変数は n − 1 個だけで十分です**。図 11 では全てのカテゴリーを項目にしていますが、例えば MT という項目がなくても構いません。CVT と AT の両方が **0** であれば MT であることが分かるからです。

では、ダミー変数を用意して重回帰分析を行ってみましょう。やるべきことは単純ですが、手作業だと面倒なのでできるだけ自動化しましょう。［中古車価格（ダミー変数）］ワークシートを開いて、どのような数式が入力されているか確認していただくだけで構いません（図 12）。

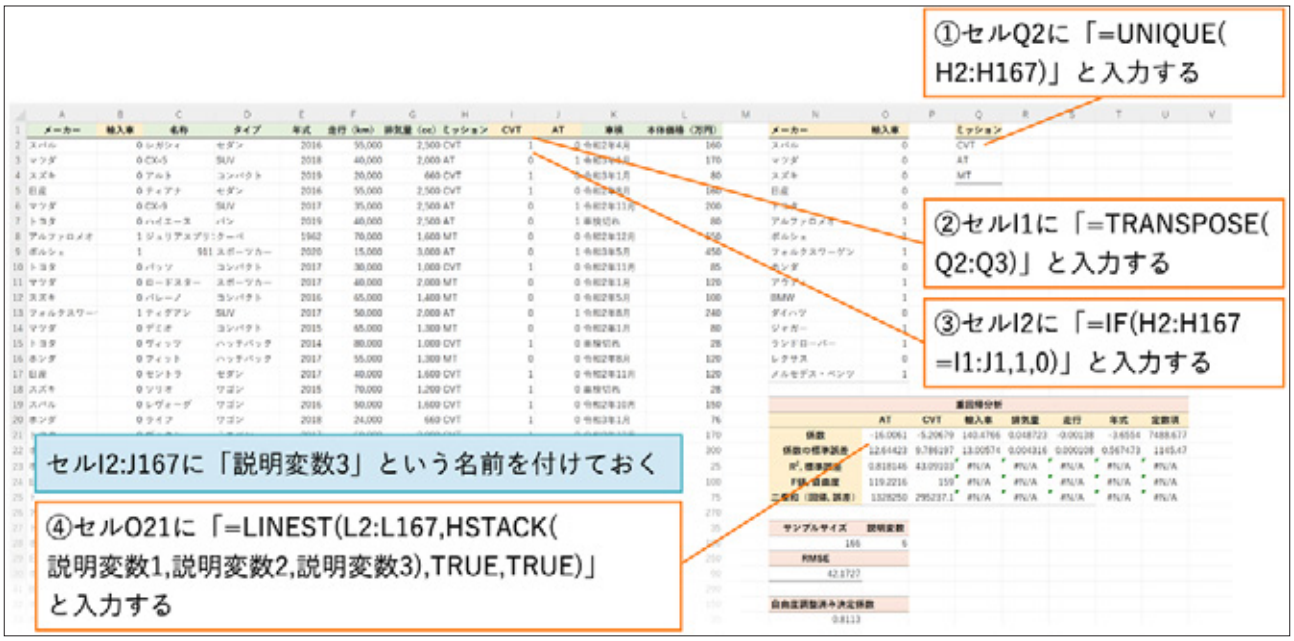


図 12　ミッションの違いをダミー変数として表して重回帰分析を行う

UNIQUE 関数を使ってミッションのリストを作成し、TRANSPOSE 関数を使って行と列を入れ替えれば、見出し（セル **I1** ～ **J1**）が作成できる。IF 関数を使い、H 列の値とセル **I1** ～ **J1** の見出しが等しければ **1**、等しくなければ **0** とすればよい。セル **I2** ～ **J167** に説明変数 3 という名前を付けておけば、LINEST 関数の引数が簡潔になる。

手順は以下の通りです。TRANSPOSE 関数は行と列を入れ替えるための関数です。以下の手順でも Excel 2019 以前のバージョンでは UNIQUE 関数が使用できないので、Microsoft 365 オンラインなどで代用するか、［中古車価格（ダミー変数）］ワークシートを参照するだけにしてください。

- セル **Q2** に「=UNIQUE(H2:H167)」と入力する（ミッションの一覧を作成する）
- セル **I1** に「=TRANSPOSE(Q2:Q3)」と入力する（セル **Q2** ～ **Q3** を横に並べる）
- セル **I2** に「=IF(H2:H167=I1:J1,1,0)」と入力する
  - ・スピル機能が使えない場合は、配列数式として入力する
  - ・Google スプレッドシートでは「=ARRAYFORMULA(IF(H2:H167=I1:J1,1,0))」と入力する
- セル **I2** ～ **J167** に説明変数 3 という名前を付けておく
- セル **O21** に「=LINEST(L2:L167,HSTACK(説明変数 1, 説明変数 2, 説明変数 3),TRUE,TRUE)」と入力する

決定係数、RMSE の求め方は既に見た通りです。いずれもごくわずかですが改善されたようです。

コラム　重回帰分析の落とし穴 ～ 多重共線性にご注意

説明変数を適切に選択すれば、重回帰分析の精度は良くなります。しかし、説明変数を増やせばいいというものでもありません。似たような説明変数を複数使うと、かえって回帰式の精度が悪くなることがあります。このことを**多重共線性**と呼びます。

多重共線性を検出するには、説明変数同士に強い相関のある項目がないかを確認します。

また、**VIF** と呼ばれる値も多重共線性の目安となります。VIF は相関行列の逆行列の対角要素の値となるので、MINVERSE 関数を使うと簡単に求められます（図 13）。一般に、VIF の値が **10** 以上になると多重共線性が疑われるので、説明変数をどれか 1 つに減らすことを検討する必要があります。［中古車価格（VIF）］ワークシートに例があるので、参照してみてください。

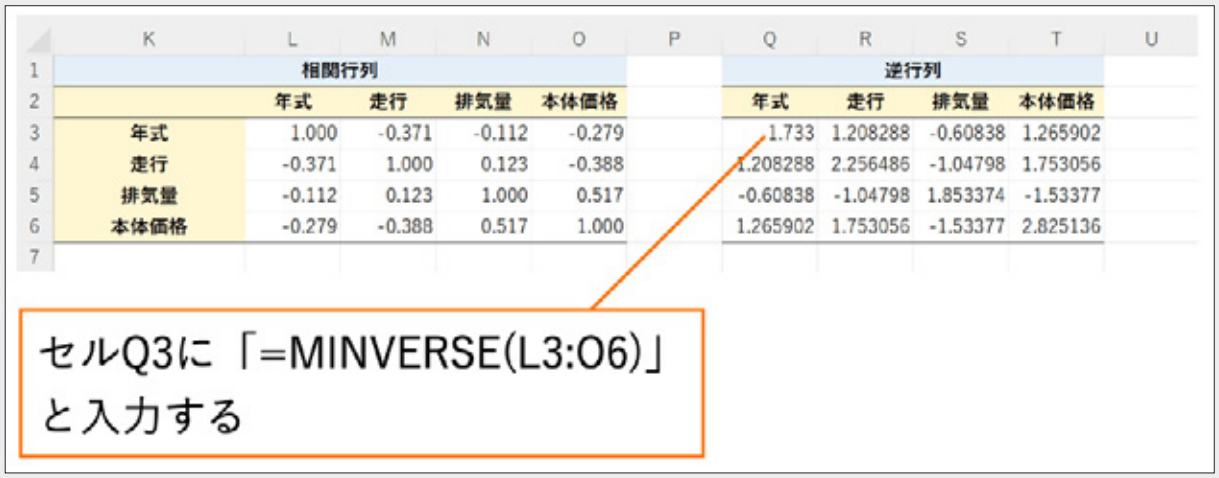


図 13　VIF の値を求める

相関行列の逆行列の値を MINVERSE 関数で求める。対角要素（セル **Q3**、**R4**、**S5**、**T6**）の値が 10 以上になると多重共線性が疑われる。この例では多重共線性は見られない（サンプルファイルには多重共線性が見られる極端な例も含めてある）。

多項式回帰による回帰分析を行う ～ 気温と電気器具による CO_2 排出量を例に

ここからは多項式回帰のお話に移ります。この連載の第 12 回で見たように、気温と電気器具による CO_2 排出量はU字形の分布になっていました。ということは下に凸な二次関数で近似できそうです。二次関数は以下のような式で表されますね。

$$y = ax^2 + bx + c$$

実は、LINEST 関数を使うと、このような多項式を利用した回帰分析もできます。単に x_2 の項を用意するだけです。それ以外の方法はこれまでと全く同じなので、図 14 でサクッと確認しておきましょう。三次以上の関数でも同様です。

サンプルファイルはこちらからダウンロードできます。［気温と CO_2 排出量］ワークシートを開いて、入力されている数式を確認してください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。

以下の手順については動画も用意してあります。操作方法を一つ一つ丁寧に追いかけたい方はぜひご視聴ください。

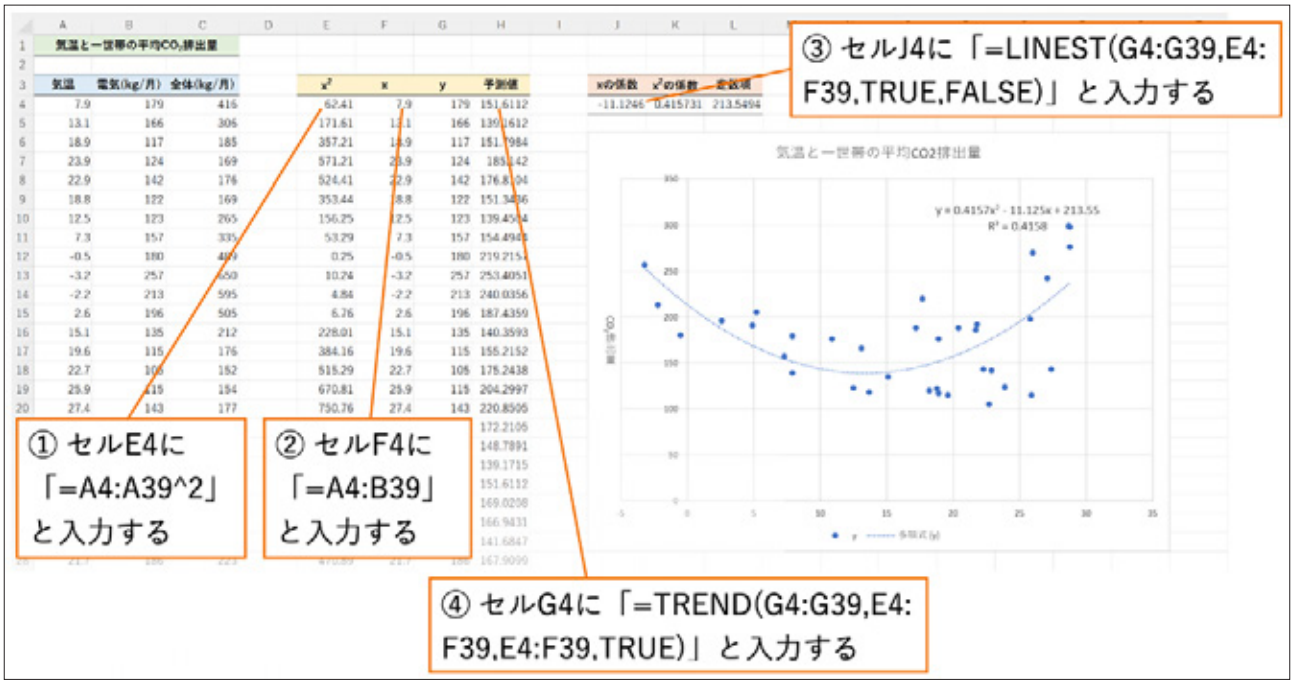


図 14　多項式回帰による回帰分析を行う

E 列が x^2 の値、F 列と G 列は x と y なので、A 列と B 列の値をそのまま表示してある。目的変数はセル **G4** ～ **G39** の範囲、説明変数はセル **E4** ～ **F39** となる。ここでは、LINEST 関数の最後の引数に **FALSE** を指定して、補正項を求めず、係数と定数項だけを求めている。右側の散布図はセル **A3** ～ **B39** を基に作成したもの。

手順は以下の通りです。散布図は図 7 のような実測値と予測値の当てはまりの良さを見るためのものではなく、元のデータの散布図とそれに対する近似曲線を描くためのものです。

多項式回帰による回帰分析を行う

- セル **E4** に「=A4:A39^2」と入力する
  - ・スピル機能が使えない場合は配列数式として入力する
  - ・Google スプレッドシートでは「=ARRAYFORMULA(A4:A39^2)」と入力する
- セル **F4** に「=A4:B39」と入力する
  - ・Google スプレッドシートでは「=ARRAYFORMULA(A4:B39)」と入力する
- セル **J4** に「=LINEST(G4:G39,E4:F39,TRUE,FALSE)」と入力する（係数と定数項が求められる）
- セル **J4** に「=TREND(G4:G39,E4:F39,E4:F39,TRUE)」と入力する

散布図の作成（Excel での操作手順）

- セル **A3** ～ **B39** を選択する
- ［挿入］タブを開き、［散布図（X, Y）またはバブルチャートの挿入］ボタンをクリックする
- ［散布図］を選択する（これで散布図が作成される）
- 系列（作成された散布図のいずれかの点）を右クリックし、［近似曲線の追加］を選択する
- ［近似曲線の書式設定］作業ウィンドウで［多項式近似］オプションをオンにする
- ［次数］ボックスに「2」を入力する
- ［グラフに数式を表示する］チェックボックスをオンにする
- ［グラフに R-2 乗値を表示する］チェックボックスをオンにする

散布図の作成（Google スプレッドシートでの操作手順）

- セル **A3** ～ **B39** を選択する
- メニューバーから［挿入］－［グラフ］を選択する
- ［グラフエディタ］作業ウインドウで［グラフの種類］リストから［散布図］を選択する（これで散布図が作成される）
- ［グラフエディタ］作業ウインドウで［カスタマイズ］をクリックする
- ［系列］をクリックして下位の設定項目を表示し、［トレンドライン］チェックボックスをオンにする
- ［種類］のリストから［多項式］を選択する
- ［多項式次数］のリストから「2」を選択する
- ［ラベル］のリストから［方程式］を選択する
- ［決定係数を表示する］チェックボックスをオンにする

多項式近似の近似曲線の次数は、Excel では **6** まで、Google スプレッドシートでは **10** まで指定できます。次数を上げると、近似曲線がよりデータの近くを通るようになりますが、あまりにも元のデータへの当てはまりが良くなり過ぎて、未知のデータにうまく当てはまらないことがあります（**過剰適合**や**過学習**とも呼ばれます）。過学習が起こっているかどうかを調べるには、元のデータを学習用と検証用にランダムに分け（例えば全体の 75%を学習用に、25%を検証用にするなど）、学習用のデータで係数と定数項を求め、検証用のデータでの当てはまりを見るという方法があります（**ホールドアウト法**と呼ばれます）。

今回は、重回帰分析により回帰式の係数と定数項を求めたり、回帰式を利用して予測を行ったりする方法を紹介しました。また、より良い予測ができるように特徴量を作成したり、名義尺度のデータを数値化したりしました。さらに、多重共線性など、重回帰分析における留意点などについても触れました。最後に、気温と電気器具による CO_2 排出量のように、U 字形の関係になっている場合にも回帰分析を行うために、多項式回帰を利用する方法も紹介しました。

今回のケーススタディでは、データの形式を整えることの重要性にも気付いたと思います。そこで、次回から数回に分けてデータの形式や取り扱いについて整理したいと思います（本来は分析に先だって知っておくべきことなのですが、実際に分析に取り組んでみないとその必要性を実感しにくいので、あえて後回しにしていました）。というわけで、次回もお楽しみに！

関数リファレンス：この記事で取り上げた関数の形式

関数の利用例については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

相関行列を作成するために使った関数

INDIRECT 関数：参照文字列を基にセルを間接参照する

形式

INDIRECT(参照文字列 , 参照形式)

引数

参照文字列：セル参照を表す文字列を A1 形式または R1C1 形式で指定します。

参照形式：参照文字列の形式を指定する。

TRUE または**省略** …… 参照文字列は A1 形式で表す。

FALSE ……参照文字列は R1C1 形式で表す。

重回帰分析のために使った関数

LINEST 関数：重回帰分析による回帰式 $y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n$ の係数 $a_1 \sim a_n$ や定数項 a_0 を求める

形式

LINEST(既知の y, 既知の x, [定数], [補正項])

引数

- **既知の y**：既に得られている目的変数の値（セル範囲）を指定する。
- **既知の x**：既に得られている説明変数の値（セル範囲）を指定する。
- **定数**：定数項 a_0 の値を求めるかどうかを指定する。
  - **TRUE** または**省略** …… 定数項 a_0 の値を求める
  - **FALSE** …… 定数項 a_0 の値を 0 とする
- **補正項**：検定などで使われる詳細な情報を求めるどうかを指定する。
  - **TRUE** …… 補正項を求める
  - **FALSE** または**省略** …… 補正項は求めない（係数と定数項のみを求める）

備考

この関数は係数や定数項などの値を一度に返すので、スピル機能が使えない場合は配列数式として入力する必要がある。係数と定数項の値は a_n, ... , a_2, a_2, a_0 の順に返されることに注意。

TREND 関数：重回帰分析による予測値を求める

形式

TREND(既知の y, 既知の x, 未知の x, [定数])

引数

- **既知の y**：既に得られている目的変数の値（セル範囲）を指定する。
- **既知の x**：既に得られている説明変数の値（セル範囲）を指定する。
- **未知の x**：予測に使いたい説明変数の値（セル範囲）を指定する。この x に対する y の値が求められる。
- **定数**：重回帰式の定数項 a_0 の値を求めるかどうかを指定する。
  - **TRUE** または**省略** …… 定数項 a_0 の値を求める
  - **FALSE** …… 定数項 a_0 の値を **0** とする

データの個数を数えるために使った関数

COUNT 関数：数値の個数を求める

形式

COUNT(数値 1, 数値 2, ... , 数値 255)

引数

- **値：**数値の個数を求めたい値やセル範囲を指定する。引数は 255 個まで指定できる。

COUNTA 関数：データの個数を求める

形式

COUNTA(値 1, 値 2, ... , 値 255)

引数

- **値：**データの個数を求めたい値やセル範囲を指定する。引数は 255 個まで指定できる。

備考

COUNT 関数は数値の個数を数え、COUNTA 関数は数値や文字列、エラー値など全ての種類のデータの個数を数えるのに使う。

データの形式を整える数えるために使った関数

UNIQUE 関数：重複のないデータのリストを作る

形式

UNIQUE(配列 , [列の比較], [単一の値])

引数

- **配列：**基となるデータの配列やセル範囲を指定する。
- **列の比較：**検索の方向を指定する。
  - **TRUE** …… 列方向（右に向かって）検索する
  - **FALSE** または**省略** …… 行方向（下に向かって）検索する

単一の値：1 回だけ登場する値を返すかどうかを指定する。

- **TRUE** …… 1 回だけ登場した値のみを返す
- **FALSE** または**省略** …… 全ての値について複数回登場した値は 1 つにまとめて返す

備考

Excel デスクトップ版では Excel 2021 以降のバージョンで使用できる。Microsoft 365 オンラインや Google スプレッドシートでは使用できる。

TRANSPOSE 関数：行と列を入れ替えた配列を返す

形式

TRANSPOSE(配列)

引数

- **配列**：基となるデータの配列やセル範囲を指定する。

HSTACK 関数：複数の配列を横方向にまとめる

形式

HSTACK(配列 1, 配列 2, ... , 配列 254)

引数

- **配列**：基となるデータの配列やセル範囲を指定する。引数は 254 個まで指定できる。

備考

HSTACK 関数は複数の離れた列を 1 つの範囲にまとめるのに便利です。複数の配列を縦方向にまとめる（複数の離れた行を 1 つの範囲にまとめる）には VSTACK 関数が使われます。

逆行列を求めるために使った関数

MINVERSE 関数：逆行列を求める

形式

MINVERSE(配列)

引数

- **配列**：基となるデータの配列やセル範囲を指定する。

備考

引数に指定する配列やセル範囲の行数と列数は同じ（正方行列）である必要があります。

データ分析に適したデータ形式に変換する方法と表データを読み込む方法

データ分析の初歩から学んでいく連載の第 16 回（最終回）。分析に適した形にデータを入力／変換する方法を、Excel を使って手を動かしながら学んでいきましょう。スタック形式のレコードをアンスタック形式に変換する方法、CVS ファイルや Web ページからデータを読み込む方法などについて解説します。

羽山博（2024 年 03 月 28 日）

前回は重回帰分析により、複数の説明変数を基に目的変数の値を予測する方法を学びました。その中で、データの形式を整えることの重要性について気付いたかと思います。**データの取り扱いについては、本来であれば分析に先立って考えておくべき**ことです。しかし、さまざまな手法を体験することを優先したので、あえて連載の最後にまとめることとしました。

今回は、実際にレコードをどのように構成するか、データをどのように並べるか（スタック形式／アンスタック形式）といったデータの表現に関する問題について見た後、ファイルに保存されているデータの形式（CSV）と文字コード（UTF-8、BOM）にまつわる問題について、幾つかの「困った」事例を取り上げ、トラブルシューティング的に見ていくこととします（図 1 ～ 3）。なお、当初の予定ではデータの取り扱いを数回に分けて解説する予定でしたが、第 2 回で、必要最低限のお話（構造化データと非構造化データ、レコードとフィールドなど）については済ませてあるので、これで一通りのお話が終わりとなり、全 16 回にわたる連載は今回が最終回となります。

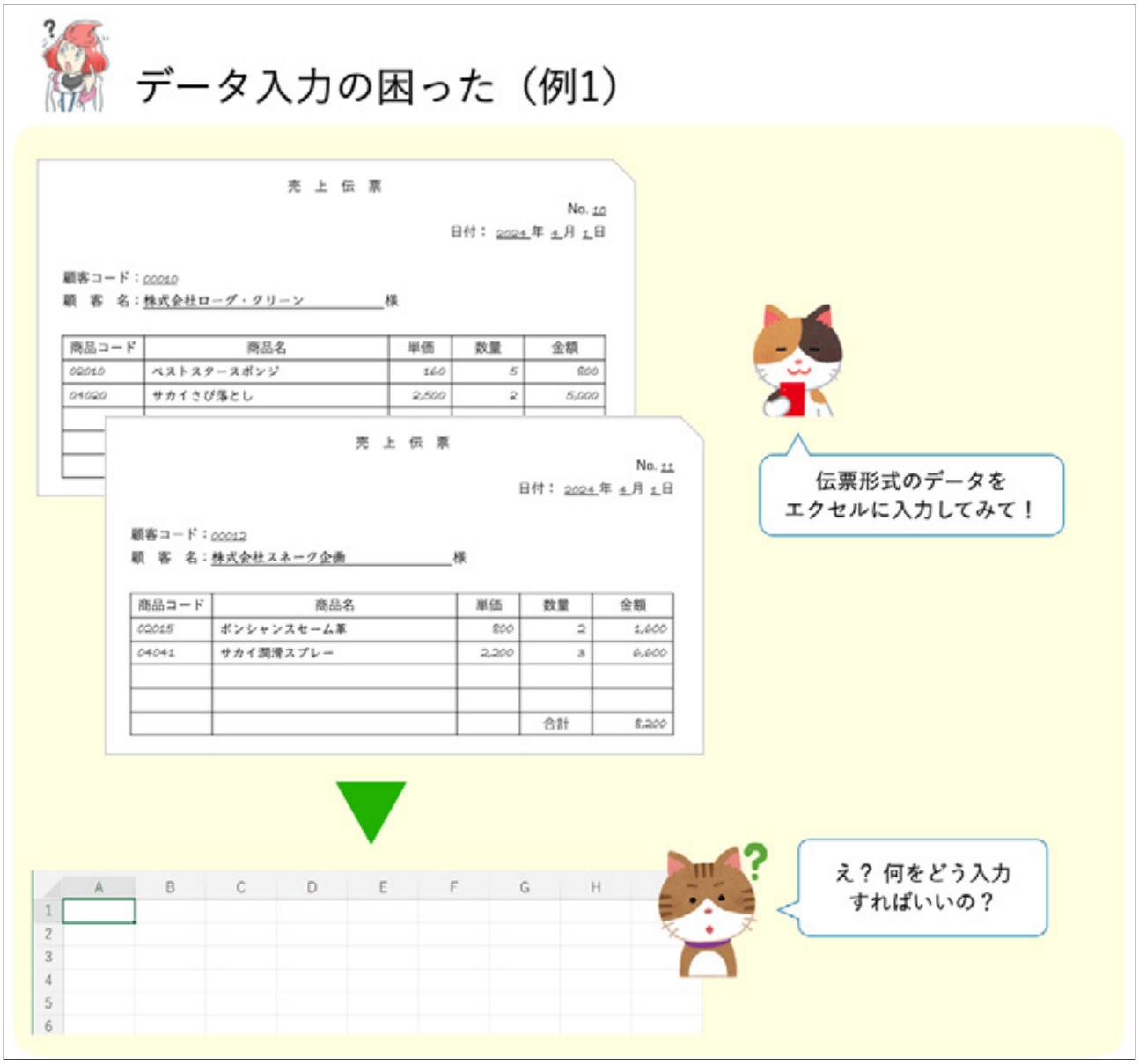


図 1　データの入力でありがちな困った例

項目が単純に並んでいるデータであればレコードにどのようなフィールドを配置すればいいのか簡単に分かるが、伝票形式のデータのように、明細に繰り返しがあるデータをどのようにレコードとして表せばいいのか悩む場合も多い。レコードやフィールドについては、単に言葉の意味だけでなく、実践的な理解が必要。

図 2　データの可視化や分析の前にありがちな困った例

オープンデータは図の左上に示したようなスタック形式で提供されているものも多い。うまく可視化できない場合は、アンスタック形式に変換したり、さらにグラフで取り扱うのに適した形式に変換する必要がある。まずは、スタック形式とアンスタック形式の違いを理解しよう。なお、このデータは IMF のオープンデータを加工したもの。

図 3　ファイルや文字コードに関してありがちな困った例

オープンデータは CSV ファイルとして提供されていることもあるが、Excel では、文字コードが UTF-8（BOM なし）だと日本語文字が文字化けする場合がある。ファイルの形式や文字コードについても理解が必要。なお、**CSV** は **Comma Separeated Values**（カンマ区切りの値）の略。

今回はデータを取り扱う上での問題点が端的に分かるように、単純な例を使って見ていくこととします。

この記事は、データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の第 16 回です。これまではケーススタディを通して、分析の手法や可視化の方法を詳しく見てきました。今回は、分析の前段階で必要となるデータの取り扱うについて見ていきます。トップページから全体の目次が参照できます。

この記事で学べること

今回は以下のようなポイントについて、分析の方法や目の付けどころを見ていきます。

- 実際のデータをレコードとして表す方法
- スタック形式とアンスタック形式の変換
- CSV ファイルからの読み込みと文字コードの取り扱い

では、伝票形式のデータをレコードとして表すところからスタートします。サンプルファイルの利用についての説明の後、本編に進みましょう。

サンプルファイルの利用について

本稿では、表計算ソフトを使って手を動かしながら学んでいきます。表計算ソフト Microsoft Excel 用の .xlsx ファイルをダウンロードできるようにしています。デスクトップ版の Excel が手元にない場合は、Microsoft アカウントで使える無料の Microsoft 365 オンライン、もしくは Google アカウントで使える無料の Google スプレッドシート（Google Sheets）をお使いください。Microsoft 365 オンラインの場合は、.xlsx ファイルを OneDrive にアップロードしてから開いてください。Google スプレッドシートの場合は、ファイルを共有して参照できるようにします。リンクを開き、メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。

繰り返しのあるデータの表し方 ～ 伝票形式のデータの取り扱い

図 1 で見たような売上伝票のデータをそのままのイメージで入力する人はさすがにいないと思います。では、レコードとしてどのように表せばいいのでしょうか。

レコードが 1 件分のデータであるということは、この連載の第 2 回で述べた通りですが、単純に 1 枚の伝票が 1 件分のデータであると考えるとうまくいきません。実は、売上伝票には 1 枚の用紙に複数件の取引データが記録されています。従って、その件数分、レコードを入力します。一方、日付や顧客コードなどの頭書き（かしらがき）の部分は、それぞれのレコードに共通した内容です。そこで、共通部分は各レコードの先頭に配置します（図 4）。

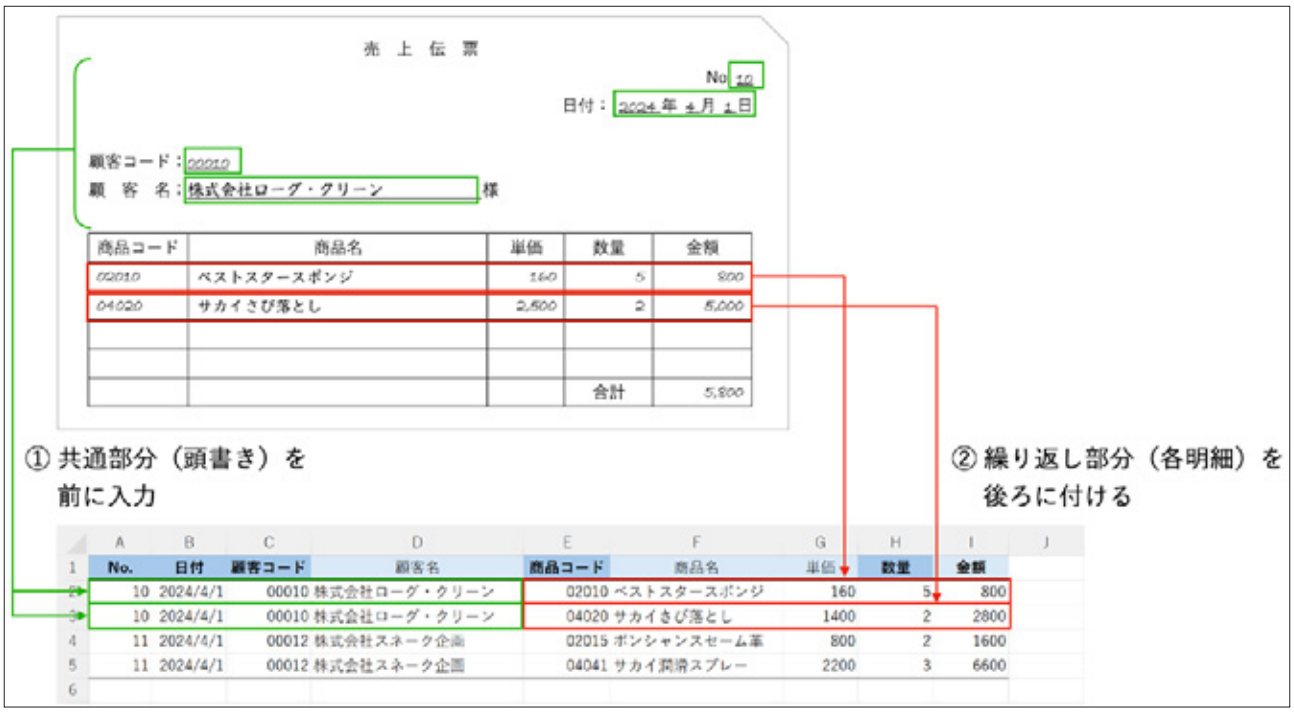


図 4　売上伝票の入力例

伝票の明細 1 行が 1 レコードになる。A ～ D 列には各レコードに共通な内容（伝票の頭書きの部分）をそのまま入力する。E ～ I 列に明細のデータを入力する。顧客名、商品名、単価については VLOOKUP 関数や XLOOKUP 関数で一覧表を検索して表示すればいいので入力の必要はない。また、金額も計算で求められるので入力しなくてもよい。下にある合計（伝票の脚書きの部分）も計算して求めればよい。

経理ソフトなどでは伝票形式の画面でデータが入力できるようになっていますが、実際には図 4 のようなレコードとしてデータベースに記録されています。ただし、もっと洗練された形式になっているのが普通です（以下のフキダシの内容を参照）。手作業での入力ではなく、POS（Point Of Sales）端末などから収集されたデータでも同様です。

データベース（リレーショナルデータベース）では、より効率良く、かつ柔軟にデータを取り扱うために**正規化**と呼ばれる操作を行い、複数の表に分けてデータを記録します。

データベースの設計時には、まず、伝票の中の明細（繰り返し部分）を分離して図 4 のように表現することから始めます。このようにして繰り返しをなくした表現を**第 1 正規形**と呼びます。

さらに、顧客コードと商品コードだけ入力しておけば、必要に応じて顧客一覧と商品一覧から顧客名や商品名、単価を引いてくれば元のデータが再現できます。この場合の顧客コードや商品コードのように元のレコードを一意に決めるのに必要なキーを**主キー**と呼び、主キーとそれに従属する顧客名や商品名、単価の表を分離した形の表現を**第 2 正規形**と呼びます。

データベースの設計では第 6 正規形まで定義されていますが、Excel でのデータの取り扱いからは話が外れるのでこれ以上は触れません。詳細については、こちらの記事などをご参照ください。

データ分析の出発点として、処理しやすい形式でデータを記録することはとても重要です。図 4 のような形式で記録しておけば、特定の日付のレコードを取り出したり、顧客や商品ごとに売上金額を集計したりすることが簡単にできるというわけです。折れ線グラフを用いて売り上げの推移を可視化したり、パレート図を用いて商品の売上金額について ABC 分析を行ったりすることも簡単にできます。

入力を容易にし、かつ、エラーデータが入らないようにする仕組み

図 4 を見ればレコードの構成が分かると思いますが、実用に際しては、入力を容易にするとともに、エラーデータが紛れ込まないようにする工夫が必要になります。そのためには**入力規則**の機能が便利です。顧客コードや商品コードをリストから選択できるようにしておけば、いちいちキーボードからコードを入力しなくても済みますし、あらかじめ登録された値以外を受け付けないようにできます。

操作の手順は図 5 の後に箇条書きにしてあります。サンプルファイルをこちらからダウンロードし、［売上データ］ワークシートを開いて試してみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。以下の操作については、動画も用意してあります。入力規則の設定とデータの入力方法を丁寧に追いかけたい方はぜひご視聴ください。

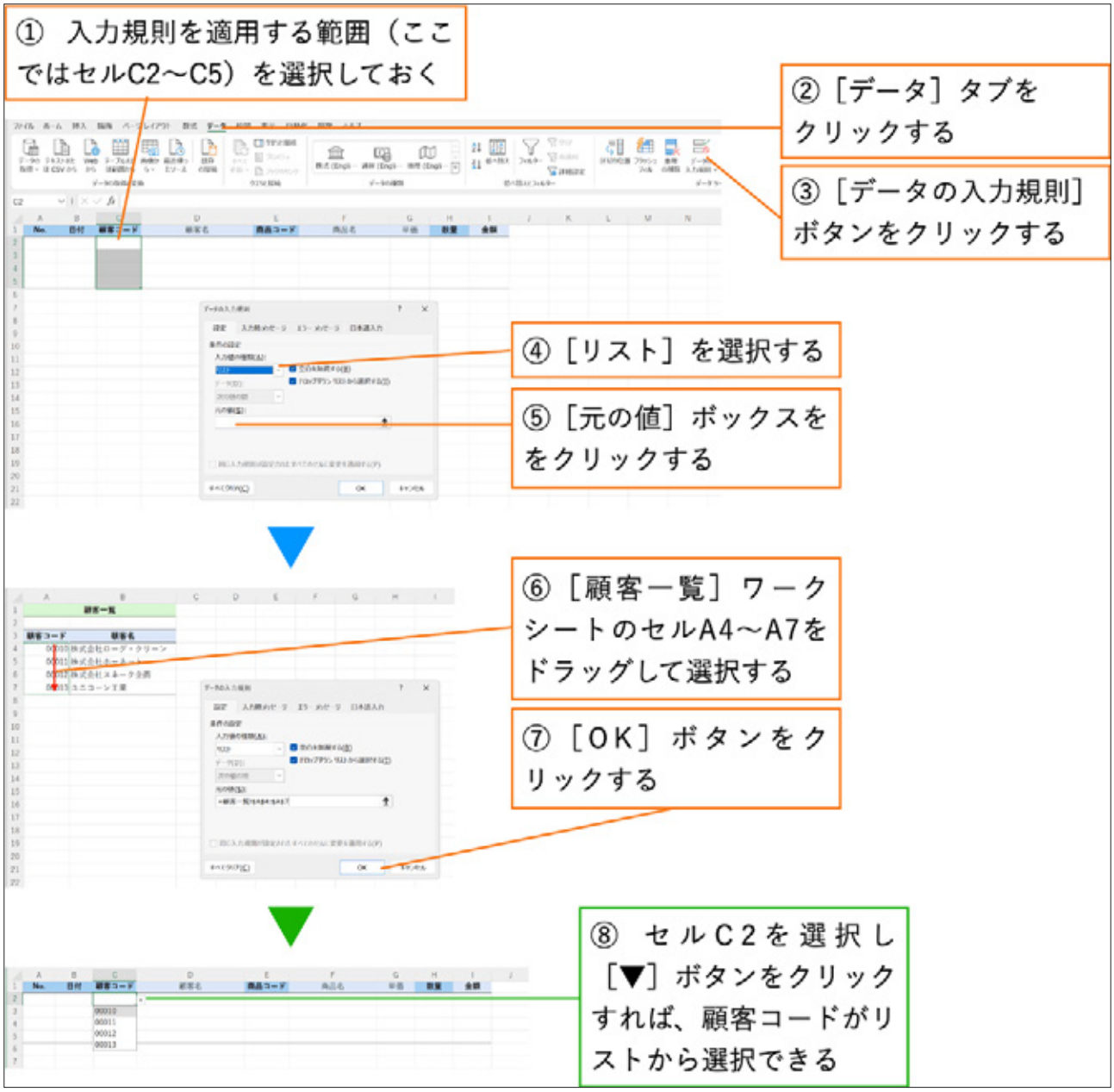


図 5　入力規則を利用し、入力を容易にするとともにエラーデータの混入を防ぐ

顧客コードや商品コードを全て記憶していて、間違いなく入力するのはきわめて難しい。そこで、入力規則を設定し、あらかじめ決められたコードの一覧から選択できるようにしておけば、効率も良くなるし、入力ミスも少なくなる。なお、標準の数値であれば先頭の 0 は表示されないが、ここでは先頭の 0 が表示されるように表示形式を変更してある。

入力規則を設定するための手順は以下の通りです。

入力規則の設定（Excel での操作手順）

- セル **C2** ～ **C5** を選択する
- ［データ］タブの［データの入力規則］ボタンをクリックする
- ［データの入力規則］ダイアログボックスの［設定］タブで［入力値の種類］から［リスト］を選択する
- ［元の値］ボックスをクリックし、［顧客一覧］ワークシートのセル **A4** ～ **A7** を選択する（`=顧客一覧!$A$4:$A$7`）
- ［OK］ボタンをクリックする

入力規則の設定（Google スプレッドシートでの操作手順）

- セル **C2** ～ **C5** を選択する
- メニューバーから［データ］－［データの入力規則］を選択する
- ［データの入力規則］作業ウインドウで［ルールを追加］ボタンをクリックする
- ［条件］のリストから［プルダウン（範囲内）］を選択する（これで散布図が作成される）
- その下の［データ範囲を選択］ボタン（田の形のボタン）をクリックする
- ［データ範囲の選択］ダイアログボックスが表示されたら、［顧客一覧］ワークシートのセル **A4** ～ **A7** を選択する
- ［OK］ボタンをクリックする
- ［データの入力規則］作業ウインドウで［完了］ボタンをクリックする

セル **E2** ～ **E5** についても同様に操作し、［商品一覧］ワークシートのセル **A4** ～ **A9** の値が選択できるようにしておきましょう。顧客名、商品名、単価については、以下のように `XLOOKUP` 関数を入力して値が自動的に表示されるようにしてあります。

- セル **C2** に「`=XLOOKUP(C2:C5, 顧客一覧!A4:A7, 顧客一覧!B4:B7,"",0,1)`」
- セル **E2** に「`=XLOOKUP(E2:E5, 商品一覧!A4:A9, 商品一覧!B4:B9,"",0,1)`」
- セル **F2** に「`=XLOOKUP(E2:E5, 商品一覧!A4:A9, 商品一覧!C4:C9,"",0,1)`」

試しにセル **C2** に「00010」と入力してみてください。セル **D2** に「株式会社ローグ・クリーン」と表示されます。

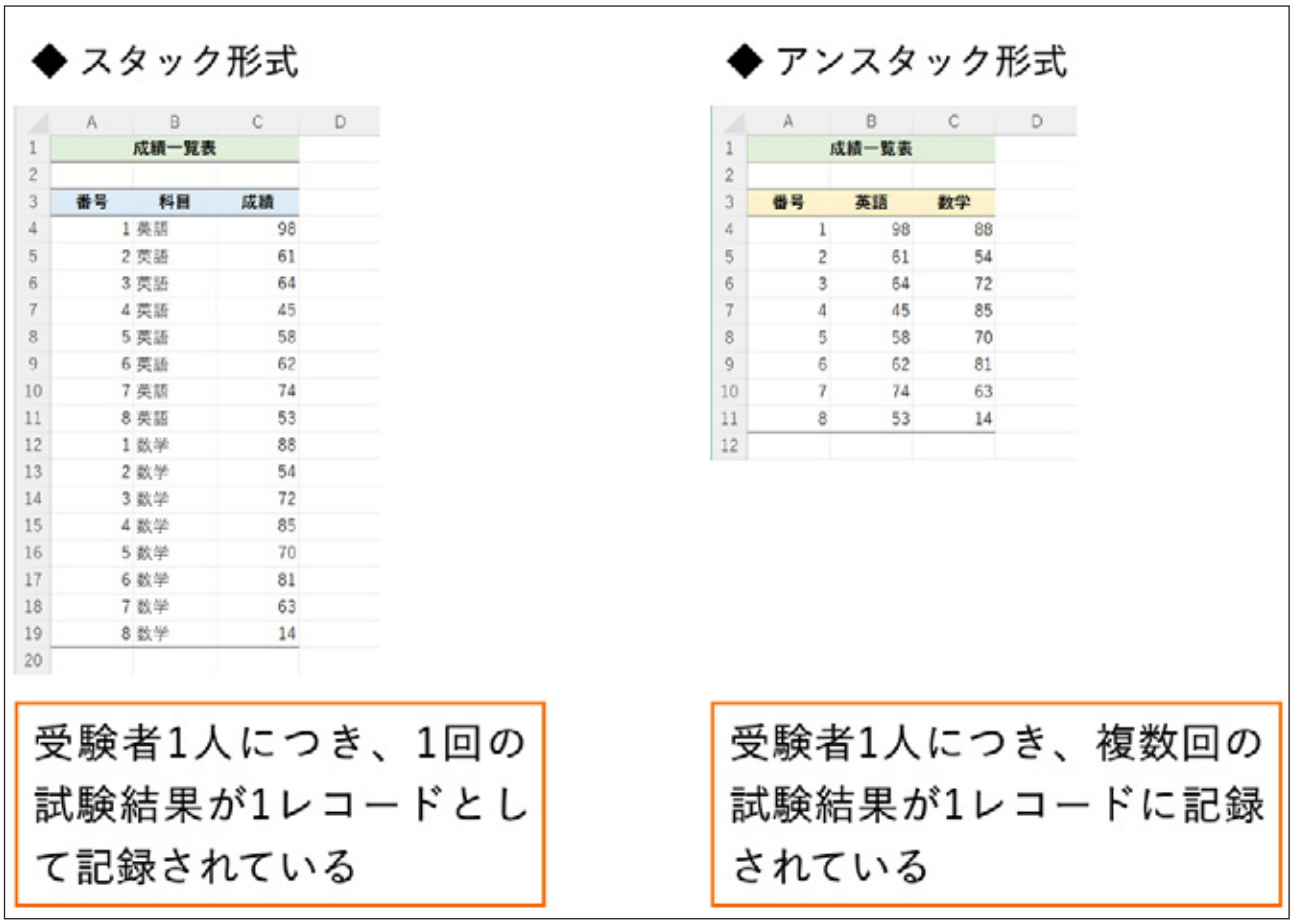
スタック形式とアンスタック形式の相互変換

グラフによる可視化を行う場合や分析のためのツールを利用する場合には、決められた形式のデータを用意しておく必要があります。そのためスタック形式とアンスタック形式の変換が必要になることもよくあります。

スタック形式のデータとは、1 つの観測データを 1 レコード（1 行）に入力したものです。一方の**アンスタック形式のデータ**では、1 レコードが複数の観測データからなっています。「スタック」とは「積み重ねる」といった意味です。

最初に見た 1 人当たり GDP の例はデータ量が多く、ちょっと見づらいので、簡単な例で見てみましょう（図 6）。スタック形式の場合、「受験番号 1 の人が英語の試験を受けたら 98 点だった」という 1 つの観測データが 1 レコードになっています。英語の試験と数学の試験は同時には行われないので、「受験番号 1 の人が数学の試験を受けたら 88 点だった」というのは別のレコードになっています。一方、アンスタック形式の例では、「受験番号 1 の人が英語の試験を受けたら 98 点で、数学の試験を受けたら 88 点だった」というように、複数の観測データが 1 レコードになっています（つまり、繰り返しがあるということですね）。

スタック形式のことを**ロングフォーマット**、アンスタック形式のことを**ワイドフォーマット**と呼ぶこともあります。



	A	B	C	D
1	成績一覧表			
2				
3	番号	科目	成績	
4	1	英語	98	
5	2	英語	61	
6	3	英語	64	
7	4	英語	45	
8	5	英語	58	
9	6	英語	62	
10	7	英語	74	
11	8	英語	53	
12	1	数学	88	
13	2	数学	54	
14	3	数学	72	
15	4	数学	85	
16	5	数学	70	
17	6	数学	81	
18	7	数学	63	
19	8	数学	14	
20				

	A	B	C	D
1	成績一覧表			
2				
3	番号	英語	数学	
4	1	98	88	
5	2	61	54	
6	3	64	72	
7	4	45	85	
8	5	58	70	
9	6	62	81	
10	7	74	63	
11	8	53	14	
12				

図 6　スタック形式とアンスタック形式の違い

スタック形式では 1 つの観測データが 1 レコードとして記録される。アンスタック形式では複数の観測データが 1 レコードとして記録される。可視化や分析のためのツールでは、どちらの形式でデータを準備しておくかが決められているので、相互に変換する方法を知っておく必要がある。

スタック形式からアンスタック形式に変換する

スタック形式のデータをアンスタック形式に変換する方法としては、WRAPCOLS 関数を利用する方法とピボットテーブルを使う方法が簡単です。ここでは関数を使ってやってみましょう。ただし、WRAPCOLS 関数は Excel 2019 以前では使えません。ピボットテーブルを利用する方法については、後のコラムにまとめておきます。

WRAPCOLS 関数は指定した個数で折り返して列を作る関数です。例えば、10 個のデータを 5 個ごとに折り返し、5 行 2 列にすることができます。15 個のデータを 5 個ごとに折り返せば 5 行 3 列になります。図 6 に示したスタック形式のデータであれば、8 個ごとに折り返せば、アンスタック形式になりますね。

では、サンプルファイルをこちらからダウンロードし、［成績（スタック形式）］ワークシートを開いて試してみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。以降の関数を使ったスタック形式／アンスタック形式相互の変換方法については、動画も用意してあります。手順を一つ一つ丁寧に追いかけたい方はぜひご視聴ください。

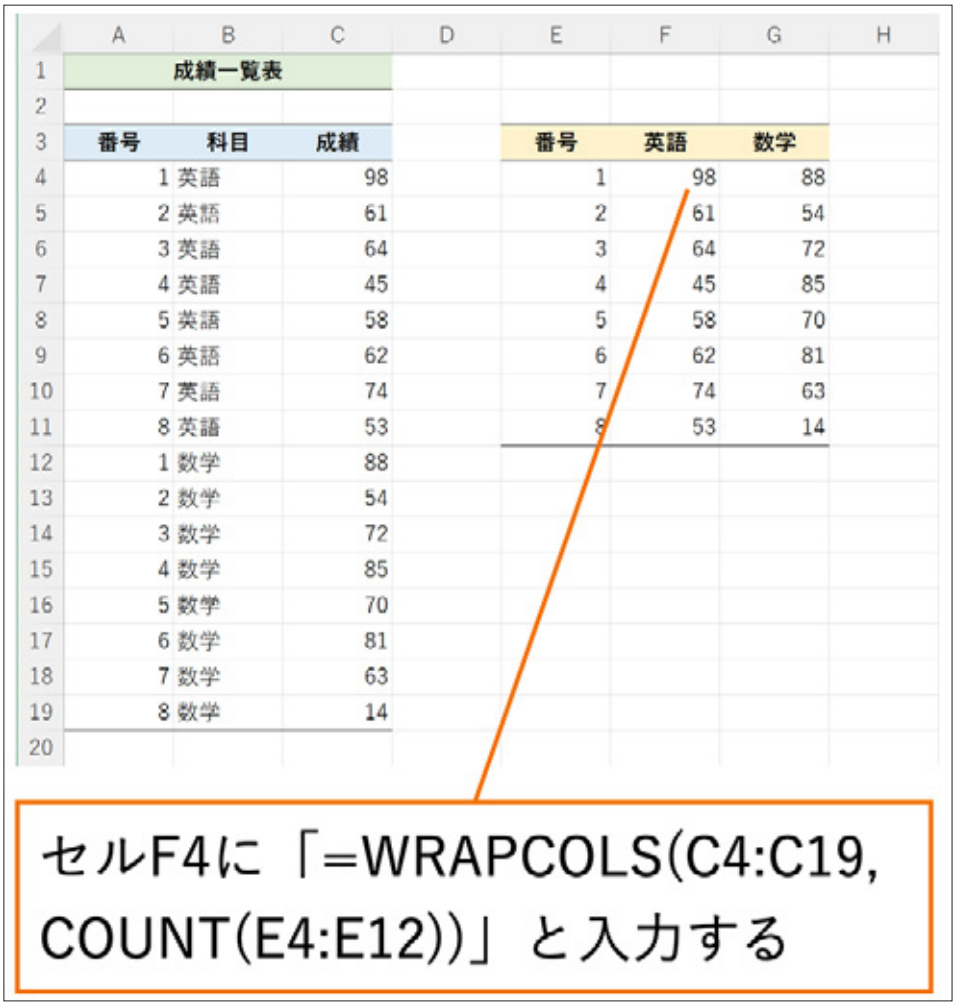


図 7　WRAPCOLS 関数を利用してスタック形式をアンスタック形式に変換する

WRAPCOLS 関数の引数には、1 列または 1 行に並んだデータの範囲として C4:C19 を指定し、折り返しの個数として、セル **E4** ～ **E12** の数値の数（＝ 8）を指定する。16 個のデータが 8 行 2 列の形に変換されて返される。

図 7 に示した操作の手順は以下の通りです。

- セル **F4** に `=WRAPCOLS(C4:C19,COUNT(E4:E11))` と入力する

番号や見出しは手作業で入力しても構いませんが、関数を使って表示することもできます。サンプルファイルには以下のような関数が入力されているので確認してみてください。

- セル **F3**：`=TRANSPOSE(UNIQUE(B4:B19))`
- セル **E4**：`=UNIQUE(A4:A19)`

最初に見た 1 人当たり GDP の例も全く同じ方法でできます。1 つの国／地域のデータが 10 件あるので、10 個ずつで折り返せばアンスタック形式にできます。後は折れ線グラフを作成するだけです。サンプルファイルに含まれる［1 人当たり GDP（US ドル作成例）］ワークシートをご参照ください。

実は、出典の IMF のデータは最初からアンスタック形式になっています。ここでは、わざわざスタック形式に変換したものを掲載しました。

コラム　ピボットテーブルを使ってアンスタック形式に変換する

ピボットテーブルを使ってスタック形式のデータをアンスタック形式に変換する場合は、図 8 のように操作します。Excel での手順は図の中に示した通りです。Google スプレッドシートでの手順は図の後に箇条書きで記しておきます。［成績（スタック形式ピボットテーブル用）］ワークシートを開いて、試してみてください（上で見たものと同じデータです）。

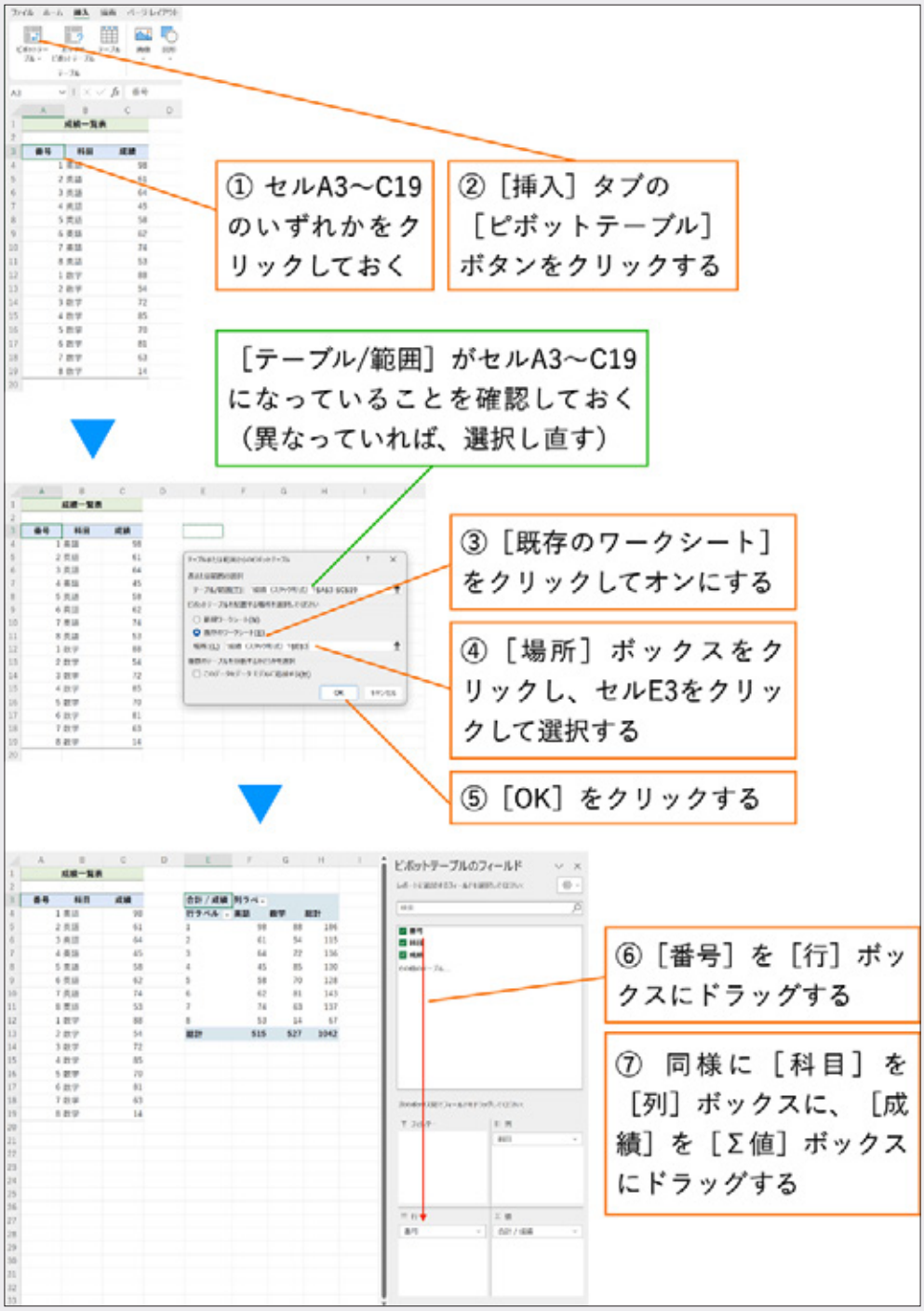


図 8　ピボットテーブルを利用してスタック形式をアンスタック形式に変換する

ここでは、元のデータと結果を同時に表示するためにピボットテーブルを既存のワークシートに作成している。各フィールドは、受験者の番号を［行］に、科目（繰り返しのある項目）を［列］に、成績を［Σ値］に指定する。結果には総計が表示されているが、不要であれば「総計」の部分を右クリックして［総計の削除］を選択すればよい。セル **E4** の「行ラベル」を「番号」に書きかえれば、図 6 のアンスタック形式のデータと同じ形になる。

→目次に戻る

Google スプレッドシートでの手順は以下の通りです。

Google スプレッドシートでの操作手順

- セル **A3** ～ **C19** のいずれかのセルをクリックしておく
- メニューバーから［挿入］－［ピボットテーブル］を選択する
- ［ピボットテーブルの作成］ダイアログボックスの［データ範囲］が「成績（スタック形式）!A3:C19」になっていることを確認する
- ［挿入先］の下の［既存のワークシート］をクリックしてオンにする
- ［データ範囲を選択］ボタン（田のマークのボタン）をクリックして、セル **E3** をクリックし、［OK］をクリックする
- ［作成］ボタンをクリックする

これで、空のピボットテーブルが作成されます。画面の右側にピボットテーブルエディタが表示されるので、以下のように操作しましょう。

- 右側に表示されている項目一覧の［番号］を、左側の［行］の下にドラッグする
- 右側に表示されている項目一覧の［科目］を、左側の［列］の下にドラッグする
- 右側に表示されている項目一覧の［成績］を、左側の［値］の下にドラッグする

結果をピボットテーブルとしてではなく、単なるデータとして利用したい場合には、他の場所にコピーしておくといいでしょう。図 8 の例であれば、セル **E4** ～ **G12** をコピーし、他の場所に貼り付けておきます。

アンスタック形式からスタック形式に変換する

続いて、アンスタック形式のデータをスタック形式に変換する方法を見ておきましょう。こちらは、ピボットテーブルではできません。TOCOL 関数を使って配列を 1 列のデータに変換します。TOCOL 関数も Excel 2019 以前では使えません。［成績（アンスタック形式）］ワークシートを開いて試してみてください。

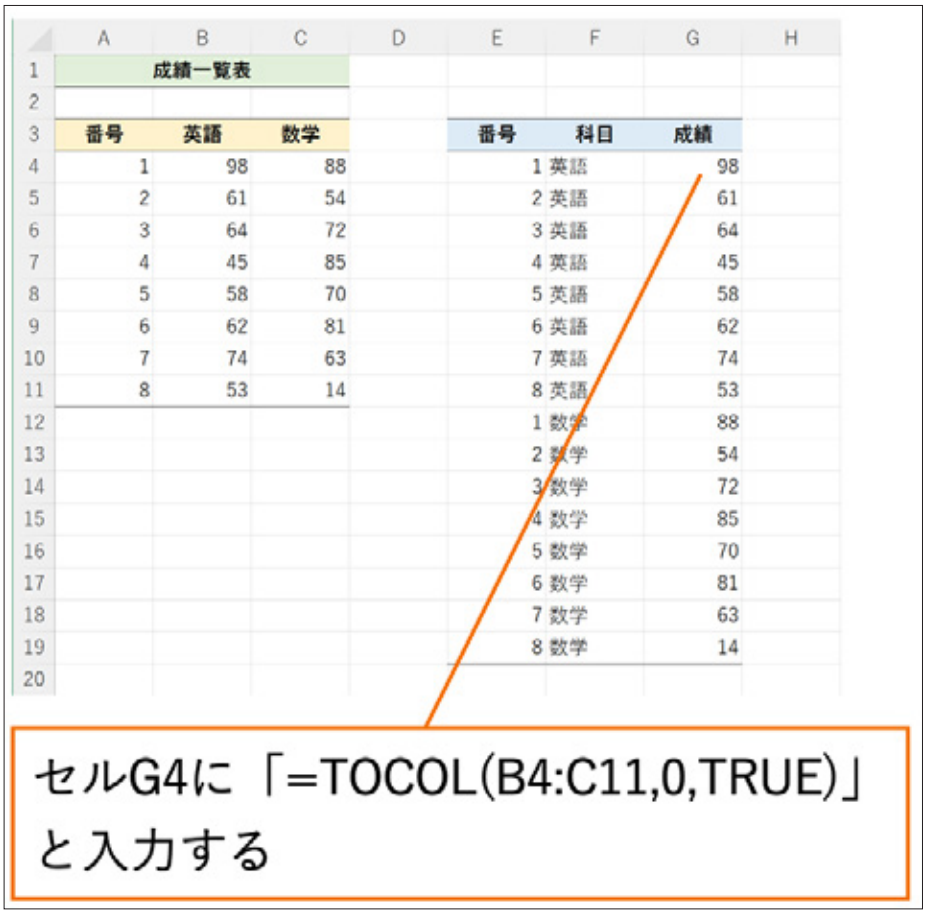


図 9　TOCOL 関数を利用してスタック形式をアンスタック形式に変換する

TOCOLS 関数の引数には、配列として B4:C11 を指定する。第 2 引数には無視する値を指定する。**0** を指定すると全ての値を対象とする（指定できる値については最後に掲載した関数の形式を参照）。最後の引数は配列から縦方向にデータを取り出して列にするか、横方向にデータを取り出して列にするかを指定する。**TRUE** を指定すると縦方向にデータを取り出すので、この例であれば、セル **B4** ～ **B11** の値を取り出し、次にセル **C4** ～ **C11** の値をつないで 1 列とする。一方、**FALSE** を指定すると横方向にデータを取り出すので、セル **B4** ～ **C4** の値を取り出し、次にセル **B5** ～ **C5** の値を、その次にセル **B6** ～ **C6** の値を……という順でセル **B11** ～ **C11** の値までをつないでいき、1 列とする。

図 9 に示した操作の手順は以下の通りです。

- セル **G4** に =TOCOL(B4:C11,0,TRUE) と入力する

E 列と F 列については手作業で入力しても構いませんが、以下のように関数を使って入力するのが効率的です。サンプルファイルには以下のような関数が入力されているので確認してみてください。

Excel の場合

- セル **E4**：`=A4:A11`
- セル **E12**：`=A4:A11`
- セル **F4**：あらかじめセル **F4** ～ **F11** を選択しておき `=B3` と入力し、入力終了時に［Ctrl］＋［Shift］+［Enter］キーを押す
- セル **F12**：あらかじめセル **F12** ～ **F19** を選択しておき `=B4` と入力し、入力終了時に［Ctrl］＋［Shift］+［Enter］キーを押す

セル **E4** とセル **E12** には、数式を配列数式として入力します。そうすれば、あらかじめ選択した範囲に同じ値を表示することができます。同じ値を繰り返して表示するには `EXPAND` 関数などを使う方法もありますが、配列数式として入力した方が手間がかからないかと思います。

Google スプレッドシートの場合

- セル **E4**：`=ARRAYFORMULA(A4:A11)`
- セル **E12**：`=ARRAYFORMULA(A4:A11)`

F 列については、`ARRAYFORMULA` 関数ではうまくいかないので、以下の方法を使います。Excel でも同じ方法が使えます。

Google スプレッドシートと Excel で利用可能な（やや高度な）方法

- セル **F4**：`=MAKEARRAY(16,1,LAMBDA(row,col,IF(row<9,B3,C3)))`

`MAKEARRAY` 関数は `LAMBDA` 関数での計算により、指定した行数、列数に配列を作成するための関数です。最初の「`16,1`」は 16 行 1 列の配列を作るという意味です。`LAMBDA` 関数に指定した `row` と `col` は行位置と列位置を表す変数で、自分で好きな変数名を付けることができます。その後に指定した数式が `row` や `col` を使った計算です。例えば、`row` の値が **9 未満**のときは「`IF(row<9,B3,C3)`」は **B3** の値（「英語」）を返し、`row` の値が **9 以上**のときには **C3** の値（「数学」）を返すというわけです。

コラム　項目の繰り返しレベルが増えた場合の取り扱い

これまで、繰り返しのレベルが 1 つの例を見てきました。さらにレベルが多くなる場合もあります。具体的には図 10 のような例です。英語と数学の試験が 2 回行われていて、第 1 回の成績と第 2 回の成績が記録されている場合ですね。このような場合でも考え方は全く同じです。

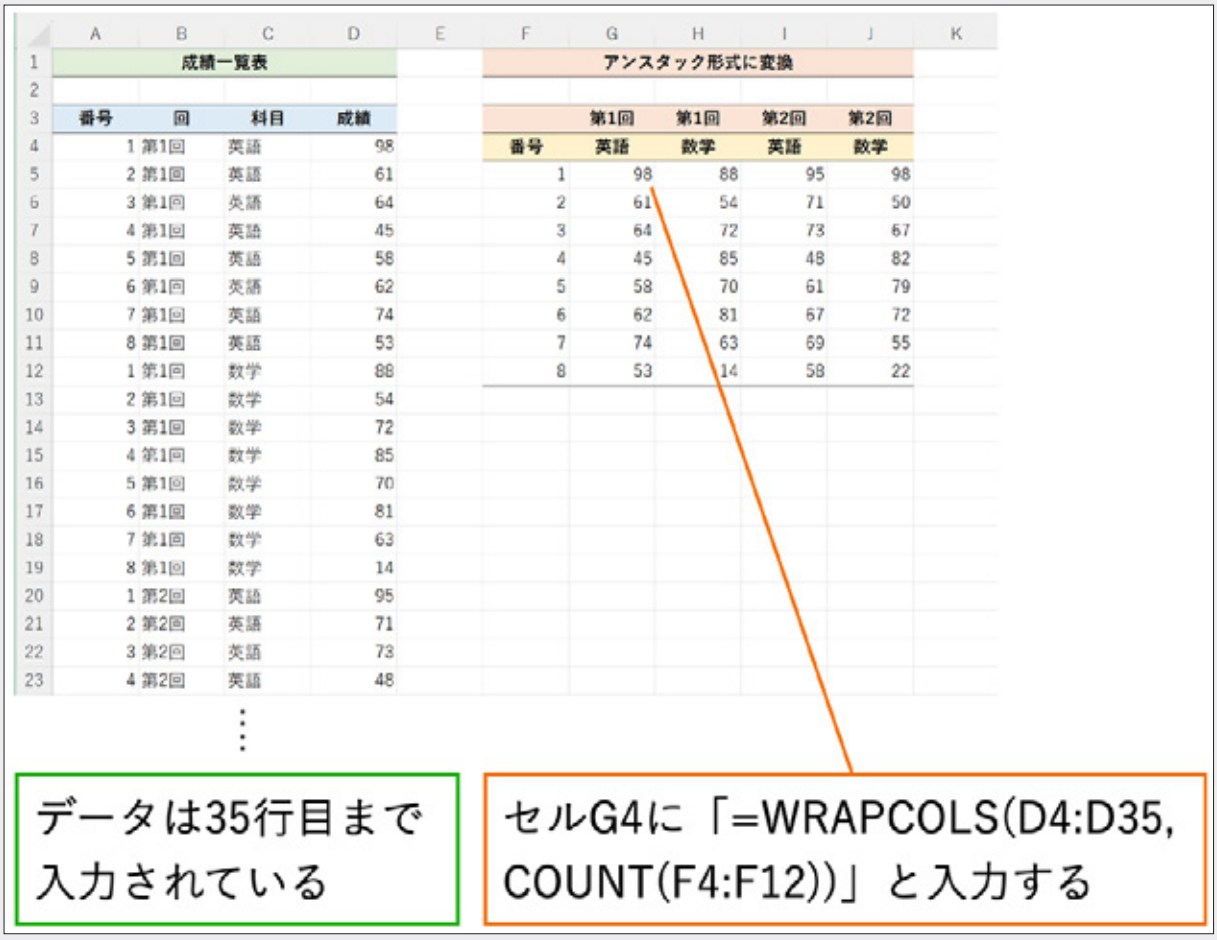


図 10　アンスタック形式に変換する（繰り返しが複数のレベルの場合）

試験の回数が繰り返しになっており、さらに科目が繰り返しになっている。WRAPCOLS 関数には成績の範囲である D4:D35 と、折り返し数を指定する。折り返し数は COUNT 関数で求めているが、目視で「**8**」と指定してもよい。

サンプルファイルの［成績（2 回）］ワークシートに図 10 の例と、ピボットテーブルを使って変換した例、さらに =TOCOL(G5:J12,0,TRUE) と入力してアンスタック形式からスタック形式に変換した例も含めてあります。［成績（2 回）］ワークシートは作成例の最後（ワークシートの右端のタブ）にあります。なお、見出しについては全て関数を使って作成していますが、この程度なら手作業で入力／コピーして作成しても構いません。

CSV ファイルの文字化けに対処する

Excel では、.xls や .xlsx などの Excel 独自のファイルだけでなく、CSV ファイルなども読み込めます。オープンデータには CSV ファイルの形式で提供されているものも数多くあります。CSV ファイルは項目がカンマで区切られた単なるテキストファイルですが、文字コードによってはうまく読み込めない（文字化けする）場合があります。以下に、よく使われる文字コードと、Excel で読み込んだ場合の状態を記しておきます。

- **シフト JIS**：文字化けしない。Windows 10 バージョン 1903 以前のメモ帳での標準的な形式（ANSI と表記されている）
- **UTF-8（BOM なし）**：ファイルの先頭に日本語文字が含まれていると文字化けする。Windows のメモ帳での標準的な形式
- **UTF-8（BOM 付き）**：正しく読み込める

現在では、インターネットで標準的に使われている UTF-8 の形式で CSV ファイルに記録されていることが多いようですが、BOM（バイトオーダーマーク）と呼ばれるコードがファイルの先頭に付いているかいないかによって結果が異なります。Excel では、BOM 付きであるという前提でデータが読み込まれるので、BOM がないと図 3 のように文字化けします（図 11 に再掲）。

図 11　ファイルや文字コードに関してありがちな困った例（図 3 を再掲）

原稿執筆時点では、文字コードが UTF-8（BOM なし）だと、ファイルの先頭に日本語文字がある CSV ファイルを Excel で読み込んだ場合に文字化けしてしまう。あらかじめ BOM 付きに変換しておく必要がある。

Excel では、上に掲載した文字コード以外のファイルを読み込むと、ほとんどの場合文字化けします。一方、Google スプレッドシートでは、上に掲載したものを含めて、ほとんどの場合、正しく読み込めます。

BOM は、Unicode と呼ばれる文字コードで、多バイト文字の下位バイトを上位アドレスに配置する（ビッグエンディアン）か、下位バイトを下位アドレスに配置する（リトルエンディアン）かを区別するなどの役割を持つコードです。UTF-8 は Unicode の一種ですが、特に BOM は必要とされていません。しかし、アプリケーションの仕様などによって、BOM が付けられる場合もあります。UTF-8 の場合、BOM は 16 進数の **EF BB BF** となっています。

UTF-8（BOM なし）のファイルを UTF-8（BOM 付き）に変換するには、Windows の場合はメモ帳で保存し直す方法が簡単です（図 12）。UTF-8（BOM なし）のサンプルファイルをこちらからダウンロードして試してみてください。Google スプレッドシートでは UTF-8（BOM なし）のファイルを正しく開けるので、サンプルファイルは用意していません。

以下の操作と、次の項の区切り位置の変更の操作については、動画も用意してあります。手順を一つ一つ丁寧に追いかけたい方はぜひご視聴ください。

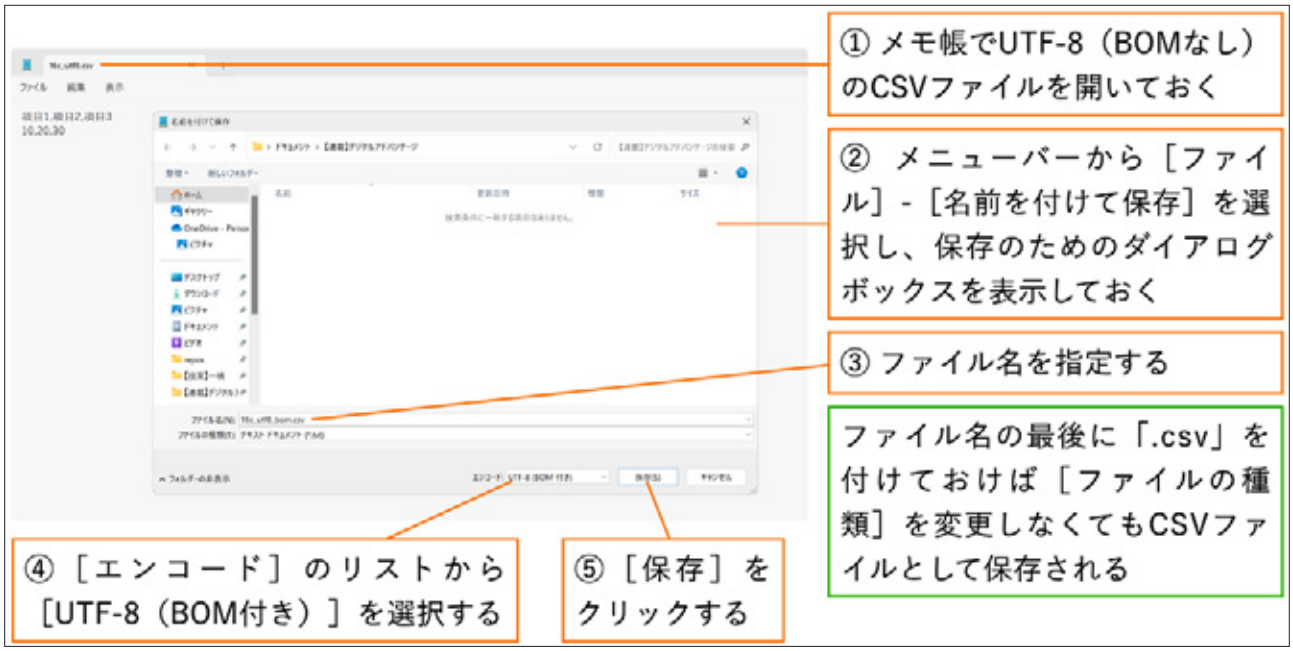


図 12　メモ帳を使ってファイルを UTF-8（BOM 付き）で保存する

サンプルファイルを開くには、メモ帳を起動して、エクスプローラーからメモ帳のウインドウにファイルをドラッグするのが手っ取り早い。ファイルが読み込めたら、メニューバーから［ファイル］-［名前を付けて保存］を選択し、［名前を付けて保存］ダイアログボックスの［エンコード］リストから［UTF-8（BOM 付き）］を選択して保存する。保存したファイルを Excel で開けば正しく表示される。

macOS では mi や Sublime Text などのテキストエディタを使って保存し直すこともできますが、いずれも標準のアプリケーションではないので、それらのアプリケーションがない場合は、ターミナルから図 13 のようにコマンドを入力するのが手っ取り早い方法です。

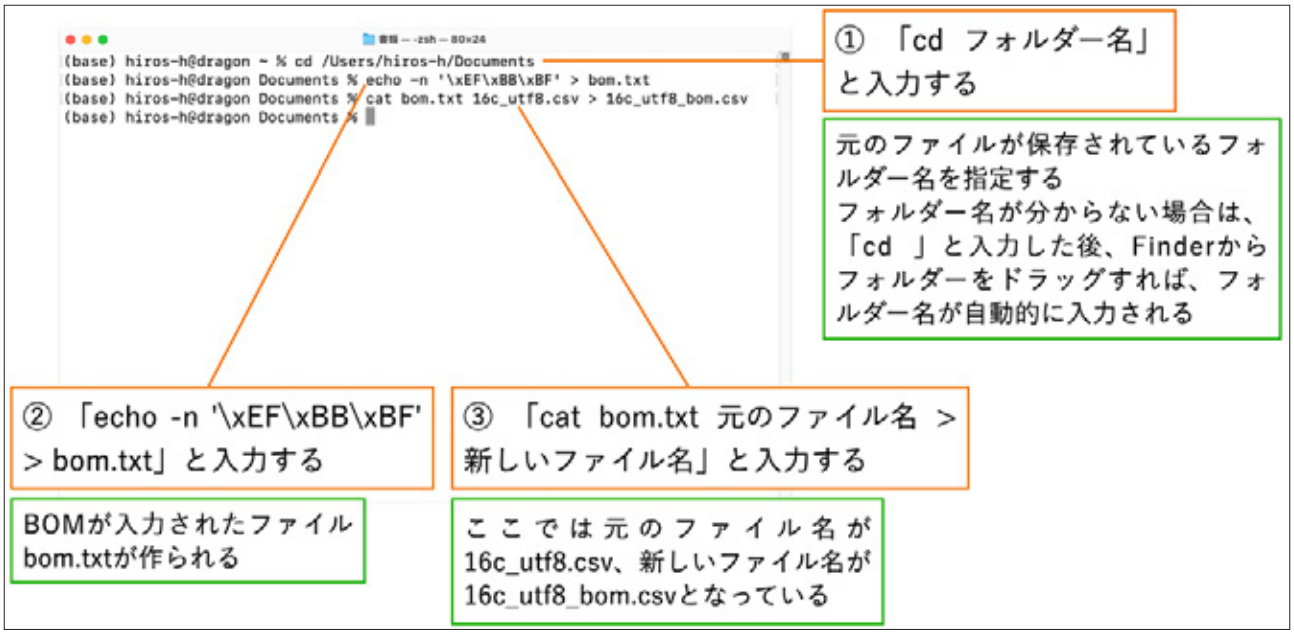


図 13　ターミナルのコマンド使ってファイルを UTF-8（BOM 付き）で保存する

「%」より左はプロンプト（入力する位置などを示す文字列）なので「%」の右からコマンドを入力する。まず、cd コマンドで、UTF-8（BOM なし）のファイルがある場所を作業フォルダーとする。次に、echo コマンドを使って、BOM だけが記録されたファイル（bom.txt）を作成する。echo コマンドは文字列をそのまま表示するコマンド。-n は、最後に改行文字を付加しないという指定。¥x は 16 進数で文字コードを指定するという意味（¥ は、macOS では［option］＋［¥］で入力する）。「>」は出力結果をファイルに保存するための指定。最後に、cat コマンドを使ってファイルを連結する。bom.txt と BOM なしの元のファイル（ここでは 16c_utf8.csv）を指定し、連結した結果を「>」を使ってファイル（16c_utf8_bom.csv）に保存する。BOM が付加された 16c_utf8_bom.csv を Excel で開けば正しく表示される。

項目がうまく区切られない場合の対処法

CSV ファイルは文字通り項目がカンマで区切られたファイルですが、場合によってはスペースやタブで区切られたテキストファイルになっていることもあります。そのような場合、行全体が 1 つのセルに読み込まれることがあります。正しく読み込むには、［データ］タブの［区切り位置］ボタンを使って、フィールド（項目）の区切りを指定します（図 14）。

サンプルファイルをこちらからダウンロードして試してみてください。Google スプレッドシートを使いたい場合は、サンプルファイルを Google ドライブにアップロードしてから、後述の手順に従って開いてください。

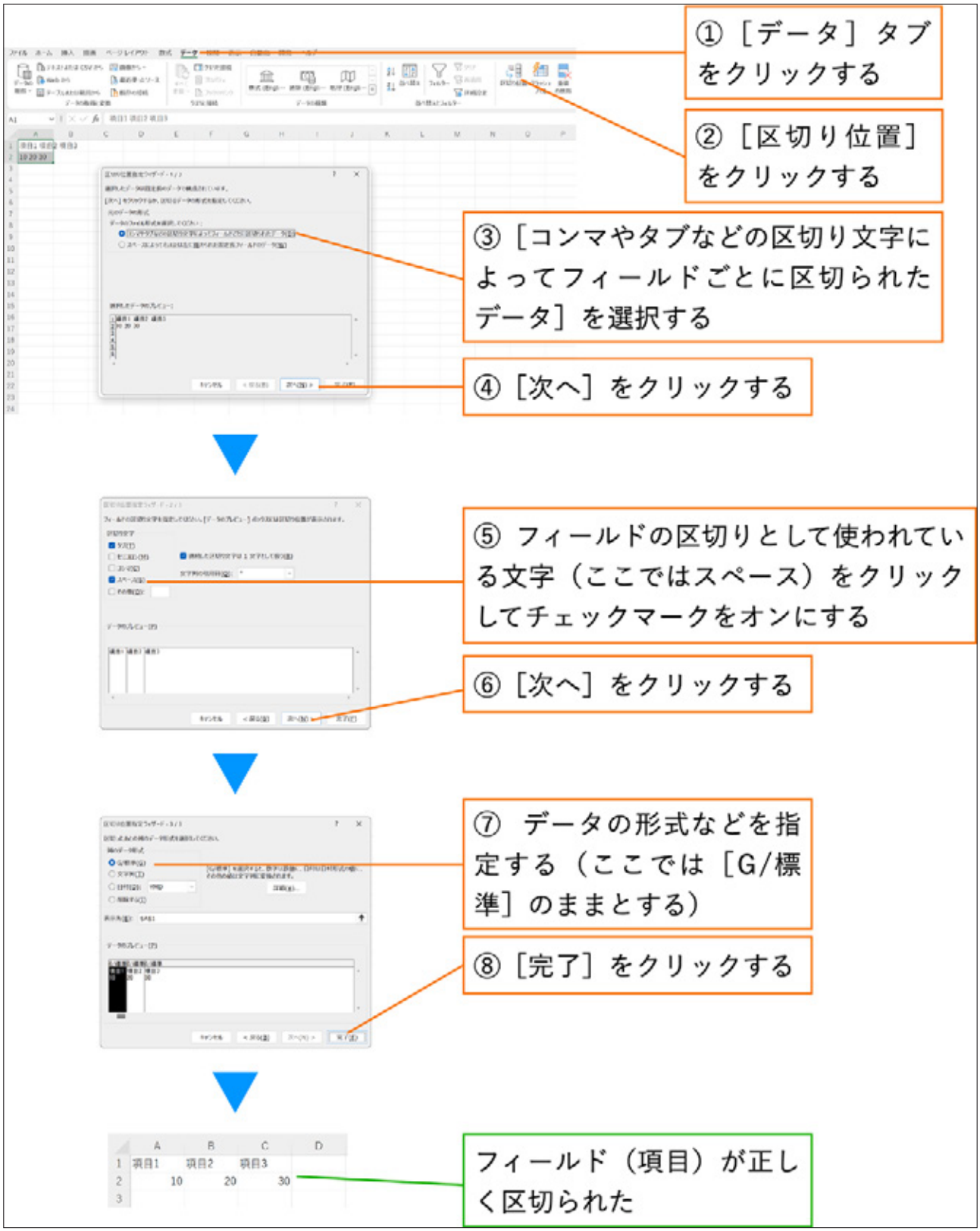


図 14　フィールドの区切りを変更する

スペースやタブでフィールドが区切られているテキストファイルを開くと、行全体が 1 つのセルに読み込まれる。［データ］タブの［区切り位置］ボタン（Microsoft 365 オンライン版では該当セルを選択した状態で［テキストを列に分割する］ボタン）をクリックして、区切り文字を指定すれば、複数の列に分けることができる。［スペースによって右または左に揃えられた固定長フィールドのデータ］は、各フィールドが同じ桁数になるように調整されているデータの場合に、桁位置を指定してフィールドを区切るのに使う。

Excel での手順は図 14 に示した通りです。Google スプレッドシートの場合は以下のように操作します。

Google スプレッドシードの場合

- Google ドライブに保存された CSV ファイルを右クリックして、[アプリで開く] - [Google スプレッドシード] を選択する
- セル **A1** ～ **A2** を選択する
- メニューバーから［データ］-［テキストを列に分割］をクリックする
- 選択範囲の右下に［区切り文字］というポップアップ表示が現れるので、[自動的に検出] と表示されているリストから［スペース］を選択する

コラム　Web ページの表を読み込む

Web ページなどで公開されているデータのうち、Excel のファイルや CSV ファイルとしてダウンロードできるものは、ダウンロードしたものをそのまま開くことができます。しかし、Web ページ上に表示されている表を Excel に取り込みたい場合もあると思います。もちろん、いずれの場合も利用に関してはそれぞれの Web サイトでの利用規約を守る必要があります。

Web ブラウザーに表示された表を選択してコピーし、Excel のワークシートに貼り付けるという方法は、あまりスマートとは言えませんね（表が崩れたりすることもよくあります）。Excel デスクトップ版の［データ］タブにはさまざまなデータを読み込むためのボタンがあるので、それを利用しましょう。なお、Google スプレッドシートや Microsofot 365 オンライン版には全く同じ機能はありません（ただし、Google スプレッドシートでは `IMPORTHTML` 関数で Web ページのデータを読み込むことができます。詳細については Google スプレッドシートのヘルプをご参照ください）。

図 15 は気象庁で提供されている毎日の気象データ（URL：https://www.data.jma.go.jp/stats/data/mdrr/synopday/data**a1**s.html）ですが、これを Excel に取り込んでみます。

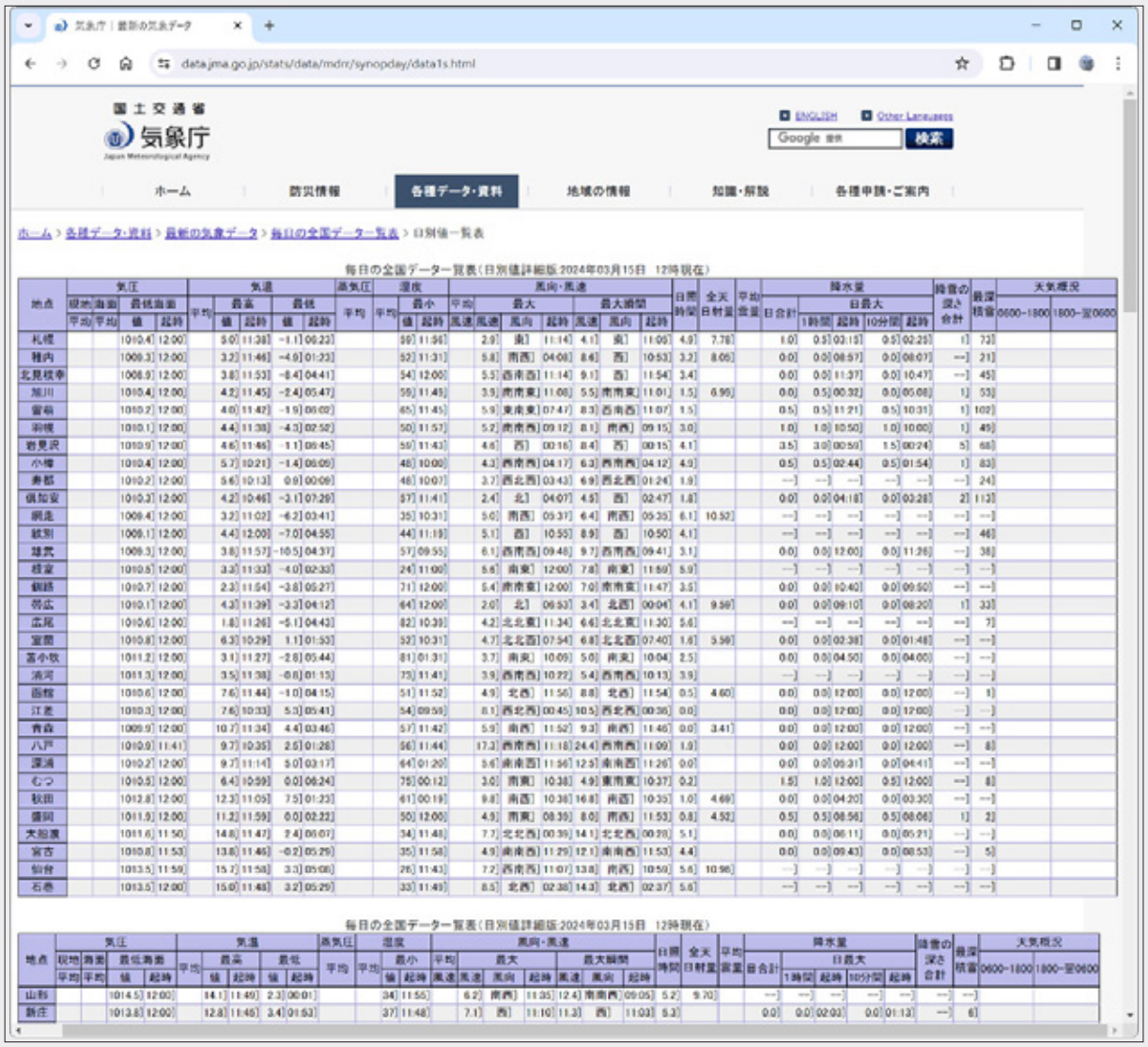


図 15　気象庁のページで提供されている毎日の気象データ

全国の気象データを一覧にしたページ（2024 年 3 月 15 日 12 時現在のデータ）。各地点の気象を記録した表が 5 つと、凡例などの表が 4 つ含まれている（見た目は表になっていないが、表として定義されているものもある）。そのうちの最初の表（札幌～石巻）を Excel に読み込んでみよう。

操作は簡単です。図 15 のように［データ］タブを開き、[Web から］ボタンをクリックします。URL を入力すれば、Web ページに含まれる表が一覧表示されるので、その中から目的の表を選択するだけです。

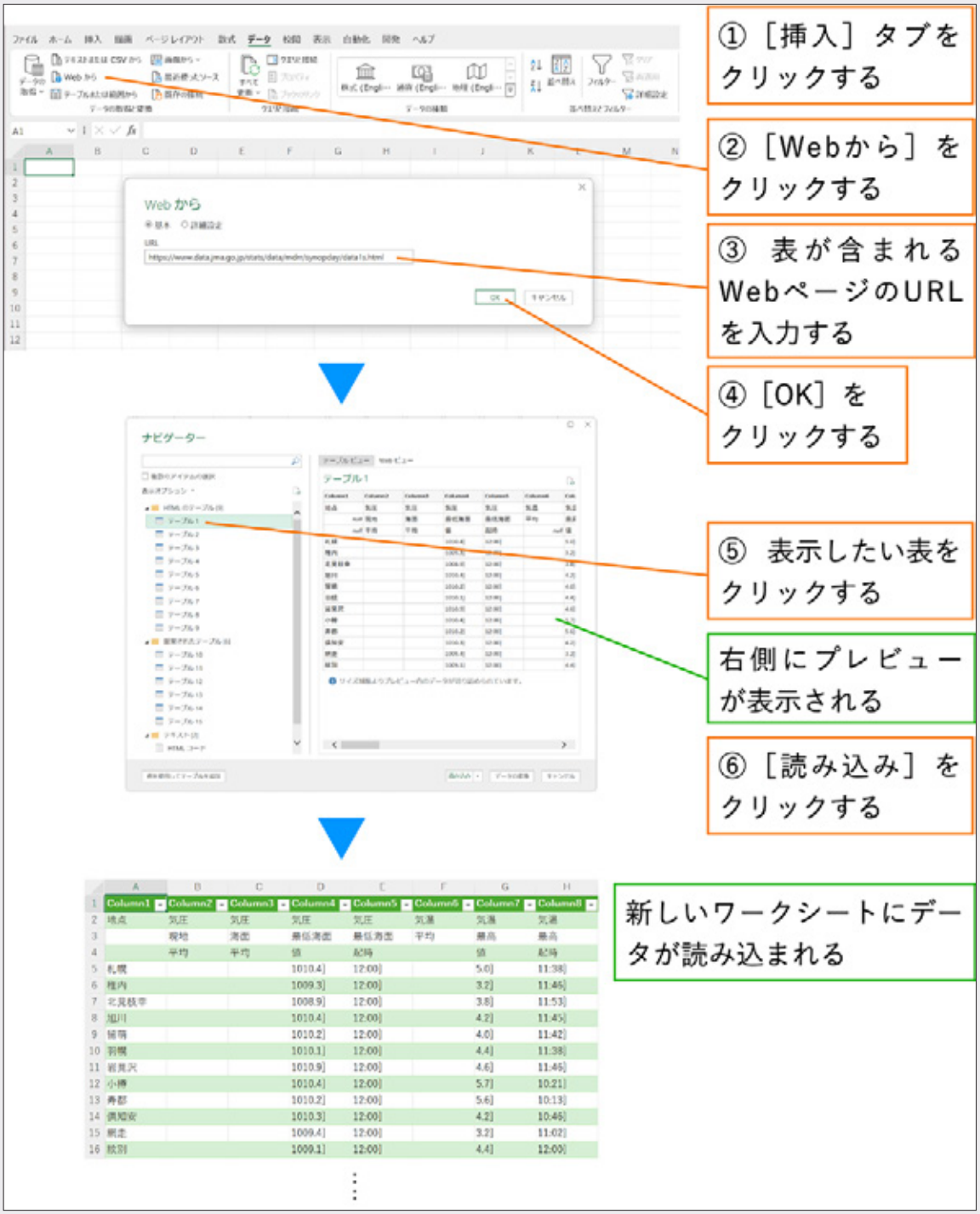


図 16　毎日の気象データを Web ページから Excel に読み込む

［ナビゲーター］ウインドウでプレビューを見ながら読み込みたい表が選択できる。右側の画面で［Web ビュー］タブを開くと、Web ページでの表示が確認できる。［読み込み］ボタンをクリックすれば、そのまま Excel のワークシートにデータが読み込まれる。また［データの変換］ボタンをクリックすれば、**Power Query** と呼ばれるウインドウが表示され、読み込む前にデータの加工ができる。

図 16 の方法は、Web ページの表だけでなく、インターネット経由で JSON ファイルや XML ファイルを読み込むのにも使えます。ただし、JSON ファイルや XML ファイルのデータは階層的な構造になっているので、Power Query の画面で階層をたどってデータを選択する必要があります。Power Query によるこれらのデータの利用については、連載のテーマから話が逸（そ）れてしまうので、こちらの公式記事などをご参照ください。

今回は、データ分析以前に考慮すべきデータの形式やファイルの形式について、幾つかの例をトラブルシューティング的に紹介しました。

実際のデータ分析に当たっては、行や列の取捨選択、並べ替えなども必要になりますが、いずれも Excel の基本的な機能で対応できます。元のデータがイレギュラーな形式であった場合に形式を整えるには、手作業での対応も必要になるかもしれませんが、Excel の機能や関数を利用すればある程度の自動化もできるかと思います。データを適切な形式にできれば、この連載で見てきた分析のための機能やツールをスムーズに適用できるようになります。

というわけで、データ分析の記述統計編は今回が最終回です。続いて、確率分布編と推測統計編を準備していますので、どうぞお楽しみに！

関数リファレンス：この記事で取り上げた関数の形式

関数の利用例については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

配列を操作するための関数

WRAPCOLS 関数：行または列のデータを指定した個数で折り返し、複数の列として返す

形式

WRAPCOLS(ベクトル , 個数 , [埋め込む値])

引数

- **ベクトル**：1 行または 1 列のデータの並び。
- **個数**：何個ごとに折り返すかを指定する。
- **埋め込む値**：元の値が存在しないときに埋め込む値を指定する。既定値は **#N/A**。

備考

折り返す個数を **3** とし、埋め込む値を既定値とした場合、以下のようになる。

	A	B	C	D	E
1	**ベクトル**		**結果**		
2	1		1	4	
3	2		2	5	
4	3		3	#N/A	
5	4				
6	5				
7					

WRAPCOLS 関数の働き

なお、この関数と似た機能を持つ WRAPROWS 関数では、行または列のデータを指定した個数で折り返して複数の行として返す。

WRAPROWS 関数は、Excel 2019 以前では使えない。

TOCOL 関数：配列を 1 列にする

形式

TOCOL(配列 , [無視する値] , [方向])

引数

- **配列**：1 列にしたい配列。
- **無視する値**：以下の値を指定する。
  - **0** または**省略** …… 全ての値を保持する
  - **1** …… 空白を無視する
  - **2** …… エラーを無視する
  - **3** …… 空白とエラーを無視する
- **方向**：横方向にデータを取得するか縦方向にデータを取得するかを以下の値で指定する。
  - **TRUE** …… 縦方向にデータを取得する
  - **FALSE** または**省略** …… 横方向にデータを取得する

備考

方向の指定により、以下のような結果になる。なお、[無視する値] に **1** を指定すると、7 行目の「**0**」は表示されない。

	A	B	C	D	E	F
1	配列			TRUEの場合	FALSEの場合	
2	1	4		1	1	
3	2	5		2	4	
4	3			3	2	
5				4	5	
6				5	3	
7				0	0	
8						

TOCOL 関数の働き

なお、この関数と似た機能を持つ TOROW 関数では、配列を 1 行にする。

TOCOL 関数は、Excel 2019 以前では使えない。

MAKEARRAY 関数：LAMBDA 関数の計算により配列を作成する

形式

MAKEARRAY（行数 , 列数 , LAMBDA(行を表す変数 , 列を表す変数 , 式）

引数

- **行数：**作成したい配列の行数。
- **列数：**作成したい配列の列数。
- **行を表す変数：**配列内の行位置を表す変数名（任意の名前）を指定する。
- **列を表す変数：**配列内の列位置を表す変数名（任意の名前）を指定する。
- **式：**行を表す変数と列を表す変数を使って計算するための式を指定する。

備考

LAMBDA 関数は、変数とその計算の方法を定義するための関数（後述）。例えば、「=LAMBDA(x, y, x+y)(1, 2)」とすると、x に **1**、y に **2** が代入され、x+y の値が返される。MAKEARRAY 関数の中に指定した場合には、配列内の行位置と列位置が LAMBDA 関数の行を表す変数と列を表す変数に渡される。

LAMBDA 関数：自作の関数を作るための関数

形式

LAMBDA（引数 1, 引数 2, ..., 式）

引数

- **引数 1, 引数 2, ...：**式に与える引数に対応する変数名を定義する。
- **式：**変数を使った計算の方法を指定する。

備考

例えば、「=LAMBDA(x, y, x+y)」であれば、最初の x に第 1 引数の値が渡され、y に第 2 引数の値が渡される。最後の x+y の値が答えとして返される。従って、セルに「=LAMBDA(x, y, x+y)(A1,A2)」と入力すると、セル **A1** の値が x に渡され、セル **A2** の値が y に渡され、x+y の値つまり **A1** と **A2** の和が返される。さらに、名前機能を利用して「=LAMBDA(x, y, x+y)」に myadd といった名前を付けておれば、「=myadd(A1, A2)」でセル **A1** とセル **A2** の和が求められる。このようにして関数を自作できる。

筆者紹介

羽山博

IT 系ライターの傍ら、非常勤講師として東大で情報・プログラミング関連の授業を、一橋大で AI 関連の授業を担当。書道、絵画を経て、ピアノとバイオリンを独学で始めるも学習曲線は常に平坦。趣味の献血は、最近脈拍が多く 99 回で一旦中断。さらにリターンライダーを目指し、大型二輪免許を取得。1 年かけてコツコツと貯金し、ようやくバイクを購入（またもや金欠）。

編集：@IT 編集部
発行：アイティメディア株式会社