□ データの収集と整理

いろいろなデータの性質を見極めよう

情報をコンピュータに入力できる形にしたものがデータです。ここでは、いろいろなデータの種類とその収集・整理の方法を学びます。

データの種類

順序関係のない名前だけのデータを、**名義尺度のデータ**といいます。たとえば、氏名を表す「山田花子」のような文字列や、出席番号（学籍番号）、性別などがこれに当たります。

これに対して、「1年生」「2年生」「3年生」、またはこれらを数値で表した「1」「2」「3」は、量としての意味は持ちませんが、順序関係があります。このようなデータを**順序尺度のデータ**といいます。アンケートで「反対」「やや反対」「やや賛成」「賛成」またはこれらを1～4の数値で表したデータも、順序尺度のデータです。

温度は量としての意味を持ちますが、40℃が20℃の2倍熱いわけではありません。これはセ氏温度（℃）という目盛でたまたま数値が2倍になっただけです。アメリカなどではセ氏温度ではなくカ氏温度（℉）をよく使いますが、40℃と20℃はカ氏温度ではそれぞれ104℉、68℉になり、2倍ではありませんね。でも、10℃と30℃の間隔は、10℃と20℃の間隔の2倍だとはいえます。このようなデータを**間隔尺度のデータ**といいます。

これに対して、**比例尺度のデータ**は、間隔だけでなく元の数値そのものが何倍になったということができるデータです。たとえば、ものの個数や降水量がこれに当たります。

上の2つを**質的データ**、下の2つを**量的データ**ということもあります。

量的データはまた、その数値が1、2、3……のように飛び飛びの値なら**離散型データ**、そうでないものを**連続型データ**ということがあります。ものの個数は離散型データ、温度や降水量は連続型データです。

グラフの種類

棒グラフで表す量は、縦棒グラフの場合、横軸は連続型データ以外なら何でもかまいませんが、縦軸は比例尺度のデータに限られます。そして、棒は0から始め、なるべく目盛を省略しないように描きます。横棒グラフでは、縦横が縦棒グラフとは逆になります。

折れ線グラフで表す量は、横軸も縦軸も、間隔尺度または比例尺度のデータにします。

棒グラフと似ているものに**度数分布図（ヒストグラム）**があります。これは、データをいくつかの階級（ビン）に区切って、それぞれの個数を面積で示すものです。棒グラフと違って、棒と棒の間に隙間を空けません。

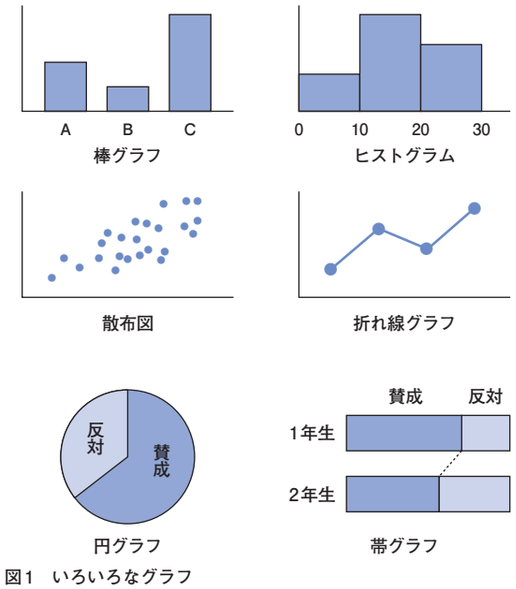

図1　いろいろなグラフ

アンケートと世論調査

1936年のアメリカ大統領選挙で、『リテラリー・ダイジェスト』という雑誌を出している出版社が、200万人以上の回答から、共和党のランドン候補の当選を予測しました。これに対して、ジョージ・ギャラップという人が設立したアメリカ世論研究所（のちのギャラップ社）は、ランダムな3000人に調査することによって、民主党のルーズベルト大統領の当選を予測します。結果は、ギャラップの予測通りでした。いくら人数が多くても、ランダムでない調査では、偏り（バイアス）が生じてしまったのです。

現在の日本の新聞社やテレビ局が行っている世論調査では、回答者は1000人程度です。これに対して、ネット調査なら何百万人もの意見を聞くことができます。しかし、実際の選挙結果と比べてみれば、1000人の世論調査の方が正しい結果が出ます。要は、人数ではなく、例えば日本の有権者の意見を知りたいなら、有権者全員から対象者をランダムに選ぶことが大切です。

世論調査では、固定電話・携帯電話の電話番号にランダムに電話をかけて調べることによって、偏りが出ないようにしています。

でも、電話を取った人だけに聞くのでは、電話を取りやすい人の意見しか集めることができません。そこで、固定電話の場合、その世帯の有権者の人数をまず聞いて、ランダムにその中から1人を選び、その人に代わってもらいます。その人が留守なら、何度もかけ直します。さらにその結果は統計学にもとづいて補正します。完全ではありませんが、ネット調査よりずっと偏りのない結果が得られます。

われわれが行うアンケート調査は、ここまでがんばることは難しいかもしれませんが、できるだけ偏りが出ないように努力しましょう。

アンケートの作り方

アンケートの回答の選択肢を作る際には、MECE（ミーシー、Mutually Exclusive, Collectively Exhaustive）の原則、すなわち「漏れなく・ダブりなく」を守ることが必要です。

たとえば、次の質問と回答を考えましょう。3つの選択肢があり、どれか1つしか選べません。このような形式を単一回答（単回答、シングルアンサー、SA）といいます。ネットアンケートではラジオボタンで回答してもらいます。

あなたの家で飼っているペットは？
○犬　○猫　○飼っていない

これで選択肢が1つしか選べないのなら、犬と猫を両方飼っている人は困ってしまいます。「漏れなく」の原則が守られてないのです。「犬と猫の両方」という選択肢も入れればいいですね。でもハムスターもペットに含めたいなら「それ以外」も含めて、あまり選択肢を増やしたくなければ、次のように複数回答（マルチプルアンサー、MA）にすればいいですね。ネットアンケートではチェックボックスで回答してもらいます。8通りの答えがあるのがわかるでしょうか。

あなたの家で飼っているペットにチェックしてください（複数回答可）
□犬　□猫　□それ以外

次の例はどうでしょうか？

勉強は好きですか？
○好き　○嫌い　○嫌いだけどする

「嫌い」と「嫌いだけどする」にダブりがありますね。最後の選択肢は不要ですが、どうしても入れたいなら、「嫌い」を「嫌いだからしない」にすればいいですね。

段階で聞く質問もあります。

選択的夫婦別姓についてどう思いますか？
○反対　○やや反対　○やや賛成　○賛成

アンケート結果のまとめかた

アンケートは、Googleフォームのようなwebのサービスを使って行えば、そのままスプレッドシートの形で取り出せます。

選択肢から1つだけ選ぶ問いでは、選択肢の番号（たとえば1～4）を入力します。

複数回答の問いでは、選択肢ごとに列を分け、チェックしてあれば1、してなければ0を入力します。

集計結果は表またはグラフで表します。単純に項目ごとに集計する**単純集計**だけでなく、複数（普通は2個）の項目をまとめて集計する**クロス集計**も必要に応じて行います。

アンケートの各項目を集計した結果は、円グラフや棒グラフで表します。

複数回答の場合、たとえば3個のチェックボックスがあれば、回答のパターンは8通りあります。この8通りのパターンを円グラフなどで表すと、読みにくいグラフになるので、チェックボックスごとのチェックの割合を棒グラフで表すことが一般的です。