

2014 年度秋学期 統計学 第3回

クロス集計とデータの可視化（解説付き）

今日は、「クロス集計」と「データの可視化」の2つのトピックを扱います。統計調査によって集められるデータには、次の「尺度水準」で述べるように、質的データと量的データに大きく分けられます。大雑把に言えば、量的データとは平均（算術平均）に意味のあるデータで、質的データはそうではないデータです。次回以降の講義では、平均から出発してデータ解析の手法を説明していきますが、今回は平均ができない質的データの簡便な整理法として、クロス集計を説明します。また、データをグラフによって直感的に把握できるようにする「可視化」についても説明します。

尺度水準

調査によって集めたデータは、ふつう数値で表されています。というよりも、統計学は、集めたデータに対して計算をすることで、データの集まりから情報を取り出そうとするものですから、数値で表されたデータを用いるのがふつうです。

ただ、データが数値で表されているからといって、必ずしも「数量」を表しているとは限りません。例えば、三択問題で「1番・2番・3番さあどれ？」というとき、1,2,3は選択肢の名前に過ぎず、a,b,cでもイ・ロ・ハでも同じですから、数量を表してはいません。

そこで統計学では、数値で表されたデータを、それが数量としての意味をどの程度持っているかによって、4つのレベルに分類しています。これを**尺度水準**といいます。

一番レベルが低いのは、**名義尺度**です。これは、さきほどの「三択問題の1番・2番・3番」や「男性：1、女性：2」のような数値で、数値は選択肢を区別するためだけにあり、2番が1番より「大きい」という意味はありません。

次のレベルにあたるのが**順序尺度**です。これは、「この講義に満足しましたか？ 1) 非常に不満・2) 不満・3) 満足・4) 非常に満足」といった調査で得られる数値です。この例では、番号の順序に意味があり、2番には「満足度が1番より大きい」という意味合いがあります。しかし、「1番と2番の満足度の差」と「2番と3番の満足度の差」が同じということはありませんし、ましてや4番が2番の2倍満足しているということもありません。

名義尺度と順序尺度にあたるデータを、**質的データ**といいます。質的データは、足し算引き算をすることができません。

一方、さらに上のレベルのデータは、足し算引き算ができるデータで、これを**量的データ**といいます。量的データは、さらに次の2つのレベルに分けられます。

間隔尺度は、数値の間の順序だけでなく、数値の間隔にも意味のあるデータです。例えば、摂氏温度は間隔尺度で、「0℃と10℃の差」と「10℃と20℃の差」はどちらも10度で、同じ意味があります。しかし、20℃が10℃の2倍暖かいという意味はありません。もしそうなら、20℃は-10℃の何倍暖かいのか？ということになってしまいます。

間隔尺度の性質を持ち、さらに「データ間の比率」にも意味があるのが、最上位のレベルである**比例尺度**です。例えば、40歳の人と20歳の人と2倍の年数を生きていますから、年齢は比例尺度です。温

	好き	嫌い	合計
男性	25	25	50
女性	35	15	50
合計	60	40	100

表 1: クロス集計

	本当に病気である	本当は病気ではない
検査で陽性	A	B
検査で陰性	C	D
合計	$A + C$	$B + D$

表 2: 感度・特異度

度でいえば、絶対温度（それ以上冷やすことのできない「絶対零度」をゼロ度とした温度）は比例尺度で、絶対温度が2倍であれば2倍のエネルギーを表しています。

データの整理の方法として、平均（算術平均）がよく知られていますが、算術平均はデータを足し算してデータの数で割ることですから、量的データでなければ意味がありません。¹

クロス集計

クロス集計とは

次回以降の講義では、平均を計算できる量的データを対象として、記述統計学によるデータ解析の手法、さらに標本調査を用いた統計的推測の方法を説明していきます。今日は、質的データに対する解析の手法として、**クロス集計**について簡単に説明します。

例えば、「商品 A が好きか嫌いか」を調査し、「好き：60%、嫌い：40%」というデータを得たとします²。これだけでは、「好きな人のほうがいくぶん多い」ということ以上の情報は得られません。そこで、この調査のさいに、回答者が男性か女性かも調べておいて、男女別にデータを整理します。その結果を、表 1 のようにまとめます。

このように、ひとつのデータ群を2つの項目から見て、それらの関係を表に表すのが、クロス集計です。この表によると、「商品 A が好きな人のほうがいくぶん多い」のは女性についてであり、男性ではあまり差がないことがわかります。

「感度」について

クロス集計と同様の方法を使って、ここで「検査の感度」についてお話ししておきます。ある病気の新しい検査法が開発されたとして、本当に病気であるかどうかわかっている人に対して、その検査法を適用して、本当にその検査法が有効かどうかを調べる実験を行います。実験結果を、「本当に病気の人／本当は病気でない人」の人数と「検査で陽性／陰性」の人数とでクロス集計の形で表したものが表 2 です。

検査の**感度**とは、「本当に病気の人のうち、検査で陽性になった人の割合」で、表では $A/(A + C)$ にあたります。もちろん感度は高いほうがよいのですが、それだけではその検査法が優秀だとはいえません。なぜなら、「病気の有無にかかわらず、いつも陽性と答える」検査なら、 $C = 0$ ですから、感度は

¹上で順序尺度の例としてあげた「授業評価」でも平均点を出していることがありますが、厳密には意味がないこととなります。ただ、このような調査では「各番号の満足度の間隔が概ね等しい」つまり近似的に間隔尺度であると仮定して、平均にも意味があると考える考え方もあります。

²ふつうは「どちらでもない」という選択肢もあると思いますが、今日は説明を簡単にするためにこのようにしておきます。

100%になるからです。もちろん、こんな検査には意味はありません。

そこで、検査の能力を表すのには、感度以外に**特異度**というものも用いられます。特異度とは「本当は病気でない人のうち、検査で陰性になった人の割合」で、表では $D/(B + D)$ にあたります。感度は病気の人について、特異度は病気でない人について、それぞれ検査が正解を答えている割合です。どちらも高いほうがよいのですが、両方を同時に高くすることは難しく、実際には「特異度 90%のときの感度がいくら」といった表現で、検査の能力を表します。

データの可視化

収集したデータをまとめて図に描いてみることは、データの傾向をつかみどのような解析手法を使うかを考える、データ解析の第一歩になります。また、データを用いて人を説得するためには、グラフによる表示は説得力を飛躍的に高める道具になります。しかし、グラフを見るほうとしては、グラフから作者の意図を見抜き、だまされないようにする必要があります。今日は、皆さんがおそらくよくご存じの「グラフ」を例にとり、見た目に対する注意を考えてみます。なお、この講義では、この先にも「ヒストグラム」「散布図」という可視化の方法が出てきます。

棒グラフ

各データの大きさを棒の長さで表して比較する「棒グラフ」は、小学生の時から知っているおなじみのグラフですが、慣れてるだけに、よく注意して見ないとだまされるおそれがあります。次の設問について考えてみましょう。

1. 図1は、いずれも同じデータをグラフにしたものです。差が際立って見えるのはどれでしょうか。

右のグラフは、棒の長さが値と比例しておらず、棒の長さどうしの比率を誇張することで、各棒の値の差を際立たせています。左のグラフは正しく書かれていますが、各棒で差がある部分を小さくなるので、差がないかのように見えます。真ん中のように切れ目（ブレイクといいます）を入れるのが、まあまあフェアなやりかたでしょう。

2. 棒グラフの棒を、いろいろな形で表すと、親しみが持てるグラフにはなりますが、誤解を生みやすくなるもあります。図2の例は、棒グラフの棒を缶の形にしたものですが、このような描き方は正当でしょうか？

真ん中のグラフは、縦軸を表示して、棒の高さが値を示すことを明らかにしてはいます。しかし、棒の高さだけでなく面積も拡大されていますので、面積が値を示しているかのように見えます。高さが2倍になれば面積は4倍になりますから、より差が際立って見えるようにして、読者を欺こうとしています。

右のグラフは、体積で値を示しているかのように見せています。高さが2倍になれば体積は8倍になるので、さらに差が際立って見えます。しかも、このグラフには縦軸がないので、高さで値を示していることがわからず、体積で値を示しているかのように錯覚させています。これは完全に反則です。

折れ線グラフ

折れ線グラフは、量の変化を表すために用いられます。これもよく知られた方法ですが、やはり「変化を誇張する」細工がされていることがあります。次の例を見てみましょう。

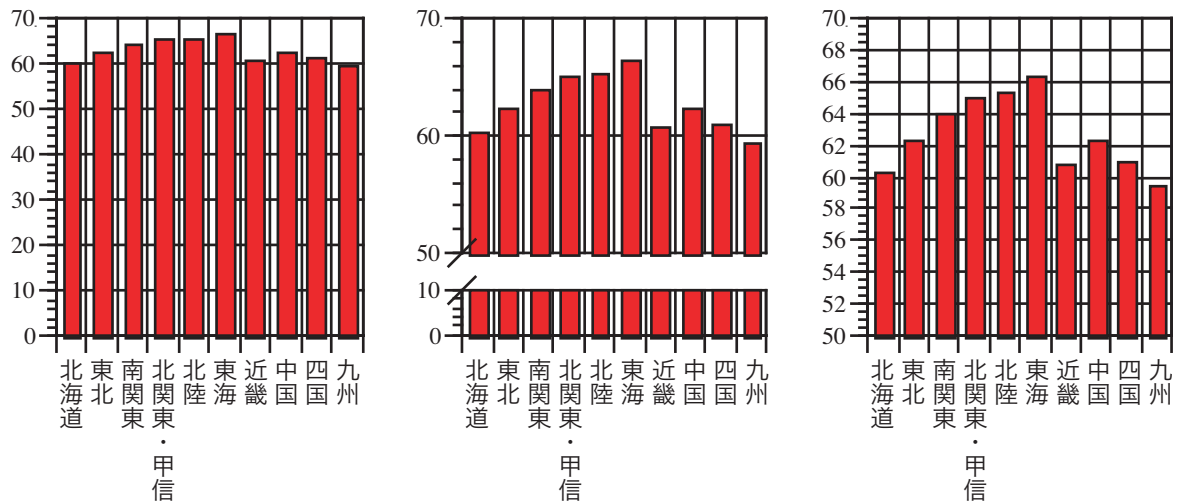


図 1: 棒グラフの例 (平成 9 年就業構造基本調査より)

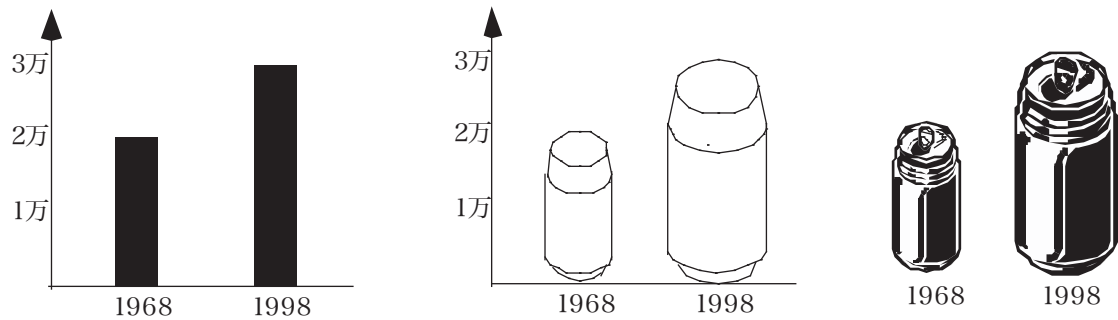


図 2: 怪しいグラフ (架空のデータ)

図 3 を使って、著者は「大恐慌時代の銀行への資本投入は大部分が返済されたといわれているが、実際にはインフレ政策がとられ物価が大きく上昇したので、実質的にはそれほど返済されたわけではない」と主張しています (横軸が年、上段が物価指数 (上: 卸売, 下: 消費者), 下段が資本注入額 (上: 実績, 下: 残高))。このグラフには、著者の主張を強調するための「細工」がされています。それは何でしょうか。

グラフに遠近感を持たせて立体風に見せており、後の年代の方が同じ値を大きく描いています。「資本注入残高」は後の年代ではほぼ一定となっているので、この効果はあまり影響がないのに対して、「物価」は後の年代で大きく変化しているので、その上昇が大きく誇張されています。

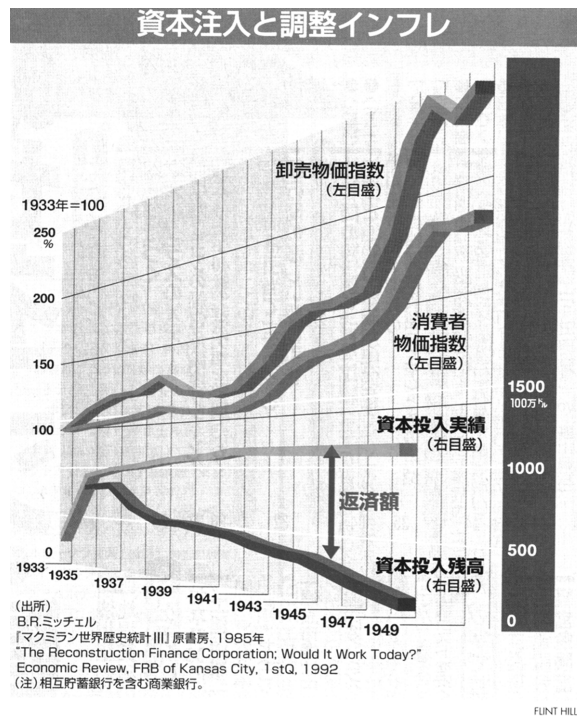


図 3: 細工入りのグラフ (週刊ダイヤモンド 1998 年 10 月 17 日号 p. 5.)