

うまい分類法の見つけ方

江 副 カ

たとえそれが、どんなに複雑に人間の介入する問題であっても、そのデータに対する理想的な分類ができて、その問題の目的に対する概念の外延を、徹底的に区分することにより、完全な体系を組織することができれば、その問題はもはや解決できたも同然であろう。

このとき、二分法は形式的には完全だが抽象的であるとか、また、多分法は具体的であるが不完全であるなど、従来、分類法に関する定性的な考え方は相当よく整理されており、また問題解決のコツはデータの上手な分類法に尽きるといっても言いすぎではなからう。

しかし、具体的な問題にぶつかって、いざうまい分類を探すととなかなかむずかしい。そして、その奥義はいろいろな問題にぶつかりながら、いくたの経験を積んでゆく以外に、これを会得することは不可能であろうが、会得のための1つの近道は、グラフを使ってデータをいろいろに分けたり、並べたり、集めたり、入れ替えたりしながら、その中にあるクセを見いだす工夫をしてみることであろう。

ここで初歩的な分類法をふり返ってみよう。つぎにあげる3組の図は、4月号以降のつづきで、生活満足指標の項目別ヒストグラムを示すが、この中で使っているデータの分け方などの方法を見てみよう。

1. データを大きさで分けてみる

1つの集団のある変量に対して、その変異の様子などを簡明に記述する方法として、ヒストグラムが使われるが、これは変量をその大きさ別に少数のクラスに分けて、それらの大きさの順に並べてみるもので、これは集団の様子を手っ取り早く知るのに便利な方法である。

たとえば図1左図は、物価格差の大きさを6クラスに分けてみた場合、その値は佐賀・宮崎はもっとも好ましいクラスに入り、反対に東京は日本1不満足なクラスに属することを一目で分かるように示した図である。

ところで、データをその大きさに分けてみた成功例の中で、おもしろいものはいくつかあるが、中でも、第2次大戦中、ドイツ潜水艦Uボートに痛手を受けていた連合軍の商船団を救うために、ORワーカーが用いた過去のデータの分類法は印象的である。

すなわち当のORワーカーは、最初は商船数別や護送艦艇類別に沈没数をとってみたが、この方法では一方がふえると反対に他方は減るというような関係は見あたらず、きめ手がどうしてもつかめなかった。

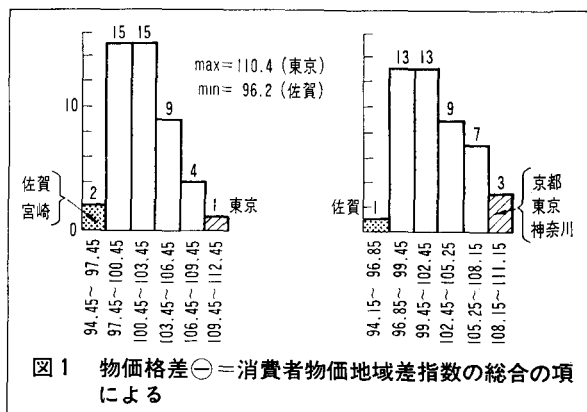


図1 物価格差 $\ominus$ ＝消費者物価地域差指数の総合の項による

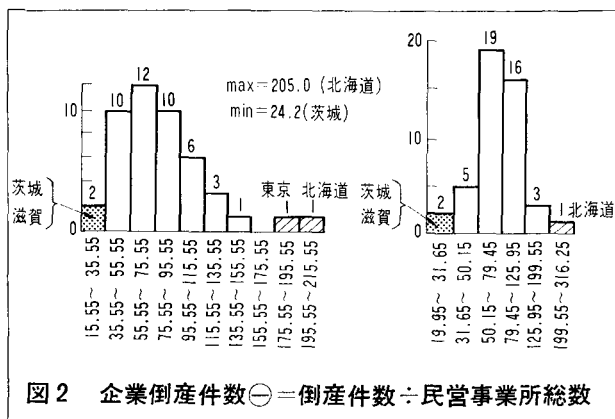


図2 企業倒産件数 $\ominus$ ＝倒産件数÷民営事業所総数

そこでつぎに、商船数や沈没数をヘミベル(Hemibel)目盛の、3, 10, 30, 80, 240, ……、というグループに分けてみたところ、グループ間にはっきりと差が出たのである。

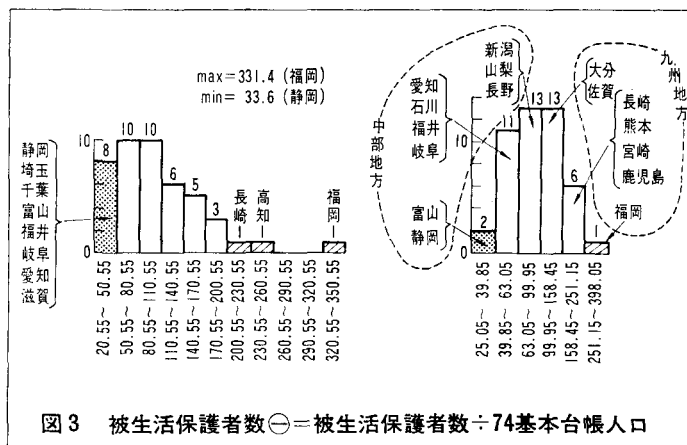
それはたとえば、各交戦ごとの沈没商船数は、護衛艦艇の数に逆比例して減少するというようなことがわかったのである。そしてこのヘミベル目盛というのは、3を何乗したかという数を目盛に使ったもので、 $3^1=3$, $3^2=9 \div 10$, $3^3=27 \div 30$, ……、のような、データを大づかみに大きき別に分けた値であり、これは実は対数目盛にはかならない。まさに対数の考え方は大づかみ思考のためのバイブルとも言えよう。

2. データの最大値と最小値の比の大きさによる分類

ある集団内におけるデータの、最大値と最小値の比率は、その集団内の競争のはげしさなどをよくあらわしていることが多い。

これは一般に統計量ではデータの最大値と最小値の差である範囲をとっているが、そのデータがもし等差的なクラスに分けるよりも、等比的スケールによるクラス分けのほうが勢力関係をよりよくあらわしている場合には、データを対数変換した後の、最大値と最小値との差をとればよく、この場合は生データでは差ではなく比になっている。

たとえば、先の図1～3の3組のデータ集団について見ると、データの最大値と最小値との比率は図1（物価格差）ではもっとも小さく1.1倍であるのに対して、図2（企業倒産件数）および図3（被生活保護者数）では、いずれも約10倍（図2は8.5倍、図3は9.9倍）となっており、図1に比べて、図2および図3のほうが競争のはげしいようである。



そして、集団内のデータの最大値と最小値との比が大きい場合には、図2および図3に示すように、生データの分布の形は指数分布に近い形になることも多く、そしてこの場合生データを対数変換することによって、これを正規分布に近づけられる。（図1～3はいずれも左側の図は、生データのままでクラス分けして書いた図であり、また右側の図は、生データを対数変換した後にクラス分けして書いたものである）

3. 現象による分類

図1～3はいずれも、46都道府県（ただし沖縄はデータがなかったため除く）のデータを用いて書いたヒストグラムであるが、各ヒストグラムの中を都道府県別にばらばらにではなく、いくつかの県をまとめて、たとえば地方別に分類してみると、その項目の中にある統計的なクセをはっきりさせることができる場合が多い。

たとえば図3右図を見てみよう。これは人口10万人あたりの被生活保護者数をヒストグラムにあらわしたもので、値は小さいほう（図の左方）ほど好ましいが、この図でたとえば中部地方の9県（新潟・山梨・長野・愛知・石川・福井・岐阜・富山・静岡）は図の左半分の好ましい側に入っているのに対して、九州地方の7県（大分・佐賀・長崎・熊本・宮崎・鹿児島）は図の右半分の、好ましくない側にあらわれている。

このように、地方別などで層別してみたときに、現象がはっきり分かれてあらわれるときには、これを手がかりにその原因を探ってゆけば、目的に対する1つの解が得られることが多い。

ところで、分類法上達のための練習問題は、日常身のまわりにごろごろがっている。そして、分類法をちょっと工夫しただけで資源や時間を新しく生み出すことも可能である。

たとえば、もらった名刺を整理するのに、苗字別のインデックスで分類しただけでは、苗字を忘れたときに役立たないが、苗字といっしょにグループ別（ゴルフ仲間とか親類など）の名刺大カードを入れておけば、苗字とグループの両方で探せるために、無駄な時間の浪費を避けることができる。

とにかく、世の中が複雑になればなるほど、分類の効果は大きく、新しい分類法を考える楽しさとこれを生かす喜びの両方を享受することができると思う。

（つづく）

（えぞえ・つとむ 日本電信電話公社）