絵本と統計表

中平一郎

わが国の産業界は二年続いた不況からようやくたち直り，景気の回復も本格化して来たとみられている。この景気回復の牽引力が輸出の増であると一般に言われているように，わが国の景気回復は世界的景気回復によって裏打ちされているわけであり，今後のわが国の経済情勢の見通しは，世界経済の動向を抜きにしては語ることができない。それどころか，現在では，各県が樹立する県勢振興計画すらも，県内企業の輸出の見通しを通じて世界経済の動向と密接に関連しているのであって，県勢振興計画の策定段階で，県の統計課長さんが，上層部から「米国の景気の見通しはどうか」などと，ご下問を受けることがあるそうである。ちょっと答え方に苦しむ事も多かろうと拝察申し上げる次第である。

ところで，国際連合が毎年発行している刊行物の一つに，世界統計年鑑がある。約九百ページの部厚い本で，その中には，百五十を超える国々を網羅した二百を超える項目の統計表が，ビッシリ詰っている。勿論，わが国の統計も，殆んどの項目について掲載されている。載っていないのは，ダイヤモンドの生産量や，燐鉱石の生産量といった項目で，これは天然資源が無いのだから仕方がない。

この年鑑を見ていると，いろいろ面白いことが発見できる。例えば，ヨーロッパ諸国の多くで，わが国と同様な人口の都市集中化現象が起っていたり，ネパール国の会計年度が七月十六日から翌年の七月十五日までだったりする。恐らく，ネパールにおける国家的記念日が，区切りとなっているのであろう。又，この年鑑によって考えさせられることも多い。例えば，平均寿命が男女とも七十歳を超えて世界最高水準に達したわが国に対し，特定の地域では今でも平均寿命が三十歳台の国が多数存在することも事実である。稀には平均寿命が二十歳台の国すらも存在している。これらは，0歳から五歳までの乳幼児死亡率が異常に高いためであり，成人後の寿命は他の国と大差ないのであろうと推察されるが，いづれにしても衛生状態の早急な改善と，そのための先進国の援助の重要さが痛感される。

更に，この年鑑の中には，世界の中での日本の立場が浮彫りにされている。百五十余りの世界の国々の中で，第五十位くらいの国土面積に，世界第五位の人口が生活し，乏しい天然資源にも拘らず，世界第三位の鉄鋼生産高を挙げ，世界第二位の自動車生産量を示し，世界第一位のテレビジョン生産台数を記録しているわが国の現実が，この年鑑の中には，いきいきと示されている。

統計表は，原則として0から9までの十個の数字によって構成される単純さが特徴であり，使用言語の壁を乗り越えて，どの国の人が見ても簡単にその概要を把握できる点で，世界中の共有財産である。恐らく世界統計年鑑も，あらゆる国で，多くの人達によって盛んに利用されていることであろう。世界中の共有財産という点で，統計表は絵本に軍配を挙げざるを得ないだろう。数字の行列よりも，色彩と表情を通してその内容を視覚に訴える絵の方が，全体的な印象を把え易いのは当然であるから。とすれば，統計表を国民全体に親しみ易いものとするための今後の課題は，統計表の持つ豊かな内容をグラフやシンボルで表現し，謂わば絵のない絵本に変身させていくことであろうか。

私が勤務している行政管理庁国際統計課には，国際機関及び諸外国から，多くの統計表が送られて来ている。これらをわかり易くグラフ化し，簡単な説明をつけて，常時皆様方にお届けしたい，そうすることによって，国際統計課を各県の統計課ひいては全国民に身近かな存在にしたい，というのが私の抱いている夢である。この夢が実現するまでには，まだ相当な期間が必要かもしれない。しかし，統計表のままでならば，今すぐにでもコピーを取ってお送りすることができる。

県の統計課長さん方，そして皆様方，もしもお困りの節は，遠慮なくご一報を。　（行政管理庁国際統計課長）

●隔月シリーズ「統計を考える」

「平均に」強くなろう……………………

右を見ても左を見ても，この世は数字だらけです。野球を見れば9対8，オリンピックは10.0，昼飯食えば400円。こんなに数字だらけの現代に生き，さらに統計を扱うとなると少しは数字の性質や扱い方を知っていた方が便利な時があります。と言ってもむずかしいことは書く方も困難ですから，書く方も読む方も困難をきたさないように進むことにします。統計の数字の基本は，和と平均と比率です。シリーズ「統計を考える」今月は平均を追ってみましょう。

集団を表す代表値には，**平均**の他に**モード**（mode）と**メヂアン**（median）というのが有ります。この３つが代表値の３羽ガラスですが，３羽のうち最も多く飛びまわっているのが御存知「平均カラス」です。平均にも２種あって，**算術平均**カラスと**幾何平均**カラスがいます。この２種を比べてみると前者の方が後者よりも大きく，また人によく好かれます。しかし算術カラスが人に好かれるのは大きいからではなく扱いやすいからなのです。幾何カラスは性格がこむずかしくて扱いにくいために嫌われてしまうのです。算術平均の扱いやすい点は何よりもまず第１に計算が簡単なことです。たとえば「２・３・４・５」の算術平均は

$$\frac{2+3+4+5}{4}=3.5$$

です。これを一般に，

$$\bar{x}=\frac{\sum xi}{n}$$

で表します。扱いやすい２番目の点は先の式からもわかるとうり，平均値 $\bar{x}$ に項数 n をかければ総和が得られるために利用価値が非常に大きいことです。一方幾何平均というのはグループの n 個の値を全部掛け合わせてそれを n で開いた値です。たとえばここに８人の人がいてその給料が，１万円が２人，２万円が５人，16万円が１人だったとします。この人達の「平均給料」は算術平均では3.5万円ですが，これが幾何平均では

$$\sqrt[8]{1\times1\times2\times2\times2\times2\times2\times16}=\sqrt[8]{572}$$

となり平均2.2万円です。実際に皆さん計算してみて下さい。算術平均を出すように簡単にできましたでしょうか。平均給料は3.5万円より2.2万円の方がおだやかです。１人16万円の高額者のために算術平均がひき上げられてしまったのです。

３羽ガラスのうち１羽は以上のような性質ですが，他の２羽についてもちょっとその姿を観察してみます。

メヂアンは和名で中央値と呼ばれるように，ある集団のまん中に位置する値です。ひとつの集団を構成している連中を小さい順から大きい順に並べてその中央にきたのをメヂアンというのですから計算も何もいりません。「５・５・６・９・９」のメヂアンは「６」，「１・３・６・６・６」のメヂアンも中央の「６」です。

モードは和名で最頻値又は並み値と呼ばれます。最も頻ぱんに出て来る値だからです。たとえば，「２・３・３・３・５」のモードは「３」，「１・３・６・６・６」のモードは「６」です。

さて，これで３羽ガラスの説明は終わりですが，集団を表すこれら３つの代表値にもそれぞれ欠点があります。算術平均は確かにその性質が優れていますが，誰もが気楽に利用し，またどこにでも顔を出して「私がこの集団の唯一の代表者であります。」というデカイ顔をしているので欠点が隠れてしまい，その集団の姿を誤解させたり，また，算術平均値と自分のナマ身を比べて我と我が身を嘆いたり喜んだりする人を絶やしません。それは，その代表値の性質をよく知らないからです。また，そのグループの代表値に何を使うかよく吟味しないで何でもかんでも算術平均で表してしまうからです。ミソもクソも一緒では困るのです。代表者を選ぶときには十分注意しないと後で大変なのはどこの世界でも同じです。オリンピックの体操競技やスキーのジャンプ競技の飛型点の採点方法は，何人かの審判員が出したそれぞれの点数の一番高い点数と一番低い点数を除いた残りの点数の算術平均になっているようです。スポーツの世界でもこのように注意しているのですから、統計の世界ではなおさら注意しなければなりません。

さて，それでは次に３つの集団を見て下さい。

(ア)　７・７・７・７・７
(イ)　５・６・７・８・９
(ウ)　１・４・７・10・13

上の３つのグループの特徴を見るために先の３羽ガラスに登上してもらうと，平均もメディアンも(ア)～(ウ)それぞれ同じく７です。どの集団も７で代表されるとはいえ，明らかにこの集団の間には差異が有るのですが，その差異が３羽ガラスでは明らかにならないのです。この場合代表値としての３羽は失格しました。どこが違うのでしょうか？違うのはバラツキなのです。ではそのバラツキ方を調べて表現するにはどんな方法が有るでしょうか。それにはまず**レンヂ**があります。レンヂはRangeで範囲のことです。

(ア)の場合は７－７＝０
(イ)の場合は９－５＝４
(ウ)の場合は13－１＝12

冨永重己

です。計算といえばひき算だけのすこぶる簡単な方法です。しかし計算が簡単な分，いかにも荒っぽいやり方だという感じがします。たとえば，

(エ)　1・1・1・2・9

(オ)　1・3・5・7・9

の2つのグループはバラツキがともに「8」のレンヂで表わされますが，この2つのグループのバラツキ方には明らかに差異が有ると見るのが普通です。そのバラツキ方の差異がレンヂで表しきれないといえます。レンヂは簡単に出せてある程度バラツキを表現できて便利ですが，その計算はほんのわずかな値だけ取り上げるだけで，その他大勢のデータを無視している分，荒っぽいのです。これは前の3羽ガラスにもいえることです。算術平均が代表値としてモードやメヂアンよりも優れている点は，算術平均が集団の中のデータすべてを計算に入れていることなのです。それと同じように，バラツキを調べる上でもデータをすべて考慮して表現する良い方法が有ればこれは便利です。そしてそんな方法が実際有るのです。それが**「標準偏差」**なのです。どうして標準偏差と呼ばれるのかわかりませんがそう呼ばれています。これは重宝です。

標準偏差はシグマで表されますが，このシグマは大文字の$\sum$ではなく（御存知のとおり大文字は総和を表します）小文字の「σ」です。このやり方は，ある集団の算術平均値（$\bar{x}$）から個々の単位（変量x）がどのくらい離れた距離に有るか調べる方法です。つまり「$x-\bar{x}$」です。個々の単位についてですから，ひとつひとつ「$x-\bar{x}$」を出してそれらすべてを合計します。つまり

$\sum(x-\bar{x})$

で表されます。統計学ではこの$\bar{x}$からのxの距離を**偏差**と呼んでいるのです。しかしちょっと待って下さい。実は，算術平均の性質を考えて頂ければおわかりのとおり，$\sum(x-\bar{x})$はいつの場合でも「0」になってしまうのです。例えば先の(オ)の例でやってみますと，

1・3・5・7・9の平均$\bar{x}$は5ですから$1-5=-4$，$3-5=-2$，$5-5=0$，$7-5=2$，$9-5=4$これを合計すると，$(-4)+(-2)+(0)+(2)+(4)=0$です。そこで考えられる方法としては，＋・－の符号を取りはずして絶対値で計算するやり方で，上の例では12になります。これをxの数nで割って$\bar{x}$からの平均的な距離を出すと2.4でこれならまあ良さそうですが，これは実は平均偏差と呼ばれているものなのです。標準偏差に行き着くためにはもうひと工夫必要です。＋・－をはずず方法は，絶対値を使うよりも後々の計算のことを考えてもっと優れた方法を使った方が良いのです。それは2乗して，－の符号を＋に変えてしまうやり方です。(オ)の例を使って早速やってみましょう。式は

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

となるわけです。

$(-2)^2$は4，$(-4)^2$は16になりますから$\Sigma(x-\bar{x})^2$は40です。これをnで割って平均的な距離を出すと「8」と出ますが，実はこれは2乗して計算したためσ^2の値になっています。σを出すには$\sqrt{\ }$で開いて元に戻してやらなければなりません。

$\sqrt{8}=2.83$，これがσです。

こうして，標準偏差を出す式は

$$\sigma=\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} \quad \cdots\cdots\cdots\cdots\cdots ①$$

と決めます。この式は覚えておいた方が良いでしょう。実際の例で標準偏差を出す場合には平均値$\bar{x}$が小数点以下の細かい数字になることが多く，いちいち$(x-\bar{x})^2$を出さなければならない①式のやり方では手間がかかります。そのため，もっと簡便なやり方として①式を変形した公式，

$$\sigma=\sqrt{\frac{\Sigma x^2}{n}-\bar{x}^2} \quad \cdots\cdots\cdots\cdots\cdots ②$$

を使います。御用とお急ぎでない方でもこの方が便利です。

σの値は0に近いほど，分布の広がりが狭いことを示します。先の(エ)の例を使ってσの値を出してみて下さい。

x(エ)	x^2	$x-\bar{x}$	$(x-\bar{x})^2$
1	1	−1.8	3.24
1	1	−1.8	3.24
1	1	−1.8	3.24
2	4	−0.8	0.64
9	81	6.2	38.44

$n=5$，$\Sigma x^2=88$，$\Sigma(x-\bar{x})^2=48.80$

$\bar{x}=2.8$　$\bar{x}^2=7.84$

①のやり方では，

$$\sigma=\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}=\sqrt{\frac{48.80}{5}}=\sqrt{9.76}\fallingdotseq 3.12$$

②のやり方では

$$\sigma=\sqrt{\frac{\Sigma x^2}{n}-\bar{x}^2}=\sqrt{\frac{88}{5}-7.84}=\sqrt{9.76}\fallingdotseq 3.12$$

(オ)の標準偏差は2.83でしたので(オ)の方が(エ)の集団より分布の広がりが狭いことがわかりました。（県消費統計係）