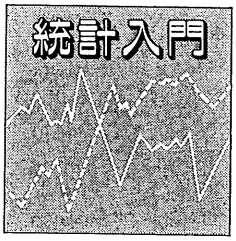

〔第3回〕

統計用語の基礎知識-1

中井里史*　新田裕史*　高木廣文**

はじめに

前号では統計解析についての基本的な考え方について解説したが，本号では実際に統計解析を進めるうえで必要となるさまざまな用語について，データを要約する尺度という点から解説する。

I. 量的データについて

統計解析を進めるうえで，まず第一に重要なことは，データを眺めること，たとえば一峰性であるのか二峰性であるのか，**外れ値**（分布からとび離れている値）はあるのか，といったことである。そのうえで，なんらかの尺度をもってデータを要約することも必要とされる。以下で，データの特徴を要約するためのさまざまな尺度を解説する。

1. 分布の位置の尺度

データの分布の中心的位置を表わす値を**代表値**(average)と呼ぶ。まず代表値について解説する。

1) 平均値

最もよく用いられる代表値であるが，**平均値**(mean)にも種類がある。まず第一は**算術平均**(arithmetic mean)で，普通平均値といえばこの算術平均を指し，数学的取り扱いも容易であるためもっともよく用いられる。算術平均は，データ数をn，それぞれのデータを，$x_1, x_2, \cdots\cdots, x_n$とすると，以下のように，データの総計をデータ数で割ったものとして求めることができる。

$$\text{算術平均} = \left(\sum_{i=1}^{n} x_i\right)/n$$

この他に平均値には**幾何平均**(geometric mean)，**調和平均**(harmonic mean)といったものもあり，前者は比率の平均値を求めたい場合などに用い，対数をとった場合の算術平均に相当し，後者は逆数に意味がある場合の平均値を求めたい場合に用いる。

$$\text{幾何平均} = \sqrt[n]{x_1 \times x_2 \times \cdots\cdots \times x_n}$$

$$\text{調和平均} = n/\sum_{i=1}^{n}(1/x_i)$$

なお3つの平均値の間には，

算術平均≧幾何平均≧調和平均

の関係が成立する。

2) 最小値と最大値

データの中で，最も小さい値を**最小値**と呼び，最も大きい値を**最大値**と呼ぶ。これらは代表値とはいわないが，以下に解説する中央値などを考える際に必要となってくる尺度である。データ解析の場面では，これらの値を調べることによって，得られたデータにミスがないかどうか，たとえば身長5mの人がいるといったデータの入力ミスをチェックすることができる。

3) パーセンタイル，分位数，中央値

最小値から最大値まで小さい順番にデータを並べ替え（データを一定の順番に並べ替えることを**ソート**と呼ぶ），小さい方から眺めてどの位の位置にあるかを示す尺度がある。全体を100%として何パーセント目に当たるかを示したものが**パーセンタイル**(percentile)と呼ばれる。このうち25%値，50%値，75%値を**四分位数**(quartile)と呼

* 東京大学医学部保健学科疫学
〔〒113 東京都文京区本郷7-3-1〕
** 文部省統計数理研究所

び，それぞれ**第1四分位，第2四分位，第3四分位**と呼ばれる。データを眺める場合によく用いられる。四分位数のうち，第2四分位（50%値）はちょうどデータの真ん中の値で，**中央値**(median)と呼ばれる。中央値の求め方は，データ数が奇数の場合は真ん中の値を用いればよいが，偶数の場合は中間の2つの値を平均して求める。なお，中央値は，データの分布が歪んでいる場合など，平均値よりも要約値としてすぐれている。

2. 分布の散布度

分布の位置の尺度としていくつかの代表値を解説してきたが，**図1**のような2つの分布を考えてみよう。この場合，平均値をとってみた場合同じ値となるが，一方の分布は幅が狭く，他方の分布はかなり裾を引いており，同じ分布をしているといえないのは明らかであろう。このような場合，位置の尺度だけでなく，分布がどれくらいの拡がりを持っているのかを知ることが重要となる。そのために，いくつか分布の**散布度**を表わす尺度が考えられている。以下で，それらについて解説する。

1）範囲，四分位範囲，四分位偏差

散布度を表わすのに最も基本的なものとして，**範囲**(range)があり，前述の最大値から最小値を引いたものとして表わされる。これはデータから容易に計算でき，統計学に不慣れな人であっても，直感的に理解しやすい。しかし外れ値があった場合，影響をうけやすいといった欠点もある。そのため，前述の第3四分位から第1四分位を引いて求めた**四分位範囲**も散布度の尺度として用いられる。なお，四分位範囲を2で割ったものを**四分位偏差**（quartile deviation）と呼ぶ。

2）分散，標準偏差

範囲はデータが全体としてどのくらいの拡がりを持つかという点ではよい指標ではあるが，平均値などの分布の位置を示す尺度との関連でデータを記述するためには不十分である。そのため平均値の周りにどれだけ散布しているかという尺度が必要となる。それが**分散**（variance），**標準偏差**(standard deviation)といった尺度である。分散は平均値（$\bar{x}$）からの**偏差**，すなわち$(\bar{x}-x_i)$をもとに計算される尺度で，

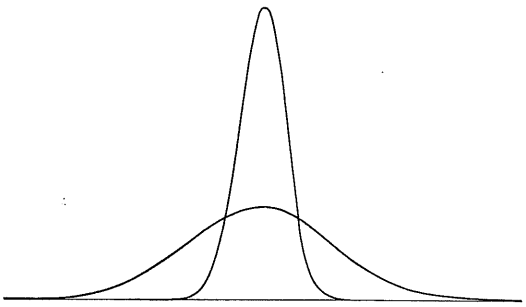

図1 平均値が等しく，分散の異なる分布

$$\text{分散}=\frac{1}{n}\sum_{i=1}^{n}(\bar{x}-x_i)^2$$

で表わされる。ここで各偏差の平方和を求めているのは，単純に偏差を足しあわせると合計が0になってしまうため，これを避けるための工夫である。この他に偏差の絶対値を足しあわせる方法（**平均偏差**，mean deviation）も考えられるが，数学的取り扱いの容易さから偏差の平方和を用いたものが用いられる。この偏差平方和の平均をとったものを分散と呼ぶが，分散は2乗して求めてあるため，平均値とは次元が異なる。そのため，分散の平方根をとった**標準偏差**が実際にはよく用いられる。なお，分散，標準偏差はデータ数nで割って求めているが，この特集の後で解説する推定，検定といった手法を用いる場合は，データ数nではなく(n−1)で割った**不偏分散**，**不偏標準偏差**といったものを用いることを申し添えておく。

3）変動係数

2群以上のデータで，どちらの方が散らばっているかを知りたい場合がある。この場合，同じ単位で測定されたものであれば，分散，標準偏差を比較すればよいが，異なるもの，例えば身長と体重の散布度を比較したい場合は，単純に標準偏差を比較するだけでは不適当である。このような場合は，以下に示す**変動係数**（coefficient of variation）を求めて比較することが可能となる。

$$\text{変動係数}=\frac{\text{標準偏差}}{\text{平均値}}$$

ただし，変動係数は比尺度で測られたデータにのみ有効であり，間隔尺度では用いてはならないことに注意されたい。

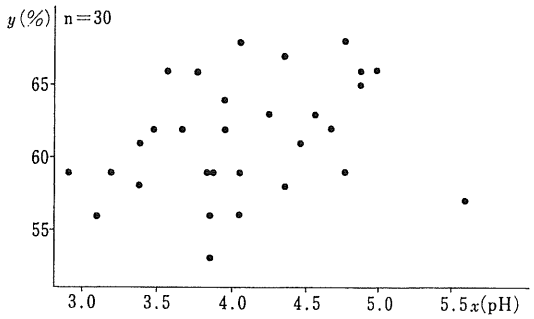

図 2　x と y の散布図

3. 相　　関

これまでは，1変数についての尺度を解説してきたが，実際にデータ解析を行う場合は，一つの変数について分布を調べるばかりでなく，いくつかの変数についての関係を調べること，例えば身長と体重は関係しているかといったこと，が必要となることが多い。このような場合，**ピアソンの積率相関係数** (Pearson's product moment correlation coefficient) と呼ばれる尺度を用いて2変数間の関係を調べることができる。なお通常は**相関係数** (correlation coefficient) と略称することが多い。変数 x と y の相関係数を求めるためには，まずそれぞれの変数について，標準偏差 s_x, s_y を求める。さらに，以下に示す**共分散**と呼ばれるものを計算する。

$$s_{xy}=\frac{1}{n}\sum_{i=1}^{n}(\bar{x}-x_i)(\bar{y}-y_i)$$

これらより，相関係数は，

$$\text{相関係数}=\frac{s_{xy}}{s_x s_y}$$

で求めることができる。相関係数は，－1から1までの値をとり，正の値のときは正の相関が，負の値の場合は負の相関があるという。0のときは相関がないことを示す。絶対値が大きいほど強い相関があることを示すが，相関の程度については厳密な定義はない。一般的にデータ数が多いとき，絶対値で以下のように扱われている。

0～0.2　ほとんど相関がない
0.2～0.4　やや相関がある
0.4～0.7　かなり相関がある
0.7～1.0　強い相関がある

相関係数は，平均値と同様に外れ値の影響を受けやすい。そのため，全般的には相関はないにも関わらず，強い相関を示すことがある。したがって，相関係数を求める際は，併せて**散布図**（**相関図**，scattergram）と呼ばれる図を描いてデータを眺めてみることが必要となる（**図2**参照）。散布図は，x_i それぞれについて，対応する y_i を x－y 平面上にプロットしたものである。また，相関係数が大きいからといって因果関係があることを示しているわけではないことに注意してほしい。

II. 質的データについて

量的なデータの場合，平均のような代表値を求めることが可能であるが，質的なデータについては代表値を求めることはできない。質的なデータを考える場合は，例えば男が何人で，女が何人であるかといったように，各値にどのくらい頻度があるかに関心がある。1変数についてはこれで十分であるが，量的データの場合と同様に，2変数間の関連を調べたい場合も生じてくる。このような場合は，**クロス表**と呼ばれる表を作る必要がある。

		喫煙 有	喫煙 無	計
性別	男	a	b	a＋b
	女	c	d	c＋d
計		a＋c	b＋d	n

（n＝a＋b＋c＋d）

これはクロス表の最も簡単で基本的な形で，各変数のとる値（**カテゴリ**と呼ぶ）がそれぞれ2つ，つまり男―女，はい―いいえ，のような場合で，**四分表** (four-fold table) と呼ばれる。なお上表は，例として性と喫煙の関連を調べている四分表であり，男性の数はa＋b人であり，男性でかつタバコを吸う人はa人である，というように読みとる。

クロス表での関連性を示す尺度はいろいろなものがあるが，ここでは四分表について医学分野で最もよく用いられる，**オッズ比** (odds ratio) について解説することにする。オッズ比は，**見込み比**，

交差積比とも呼ばれ，以下のように求めることができる。

$$\text{オッズ比}=\frac{a\,d}{b\,c}$$

オッズ比は，上表の例でいうと，男は女に比べ何倍の見込みでタバコを吸うかを表わすものとなる。なおa/b，c/dがそれぞれオッズ（見込み）と呼ばれ，男女それぞれにおいてタバコを吸うかどうかが五分五分のとき値は1，それより大きいと1を超え，小さいと1未満となる。

おわりに

以上，データを要約する際によく用いられる統計用語について解説を行った。しかし紙面の関係もあり，説明は十分ではなく，これ以外にも尺度は存在する。各尺度についての詳細は，成書によって補ってほしい。なお，どの尺度であってもただ一つでデータを要約できることができるわけではないことに注意されたい。

次号では，統計を考える際に必要となってくるいくつかの確率分布を中心に解説する。

参考文献

1) 豊川裕之，柳井晴夫，編者：医学・保健学の例題による統計学，現代数学社，1982.
2) 高木廣文：ナースのための統計学，医学書院，1984.
3) 新田裕史，佐藤俊哉：パソコンBASIC 生物統計学入門，現代数学社，1989.

*　　*　　*