現代人の統計

統計解析法の原理

林 知己夫 編

1

鈴木 義一郎 著

朝倉書店

統計解析法の原理

鈴木義一郎

朝倉書店

著者略歴

1937年　山形県に生れる
1960年　東北大学理学部卒業
現　在　文部省統計数理研究所附属統計技術員
　　養成所所長　理学博士

現代人の統計 1

統計解析法の原理　　定価 2500 円

1977年 8 月 30 日　初版第1刷
1984年 5 月 20 日　第9刷

著　者　鈴木義一郎
発行者　朝倉邦造
発行所　株式会社 朝倉書店
東京都新宿区新小川町6-29
郵便番号 162
電話 03 (260) 0141 (代)
振替口座 東京 6-86733番

〈検印省略〉

政弘印刷・渡辺製本

はじめに

この書の主たる目的は，**統計解析のための各手法の原理を明らかにすること**にある．“原理”などというと一見いかめしそうな印象をもたれるかもしれないが，いわばカラクリである．統計的手法が活かされて使われ得るには，第一に正しい使われ方がなされていること，第二に巧い使われ方がなされていることである．そのためには，各手法のカラクリをよく理解しておくことが前提となってくる．

ある種の病気を患ってお医者さんのところへかけこむと，適当な医薬品を調合してもらえる．与えられた薬品の成分がどんなものかぐらいまでなら，化学についての知識をある程度もっている人なら判るかもしれない．だが，その成分がどんな効き方をするのかに関しては，医学ないし薬学的な知識をもっていないとできないだろう．薬品を正しく使う，巧く使うためには，その薬品の成分の“効き方”を知る必要があるからである．医師はそのような知識をもっているが，患者のほうはもってない．だから患者はお医者さんを頼りにせざるをえないということになる．統計的方法を正しく使う，巧く使うという場合でも，その方法の“効き方”をよく認識しておくことが必要である．“成分表”に関する知識だけでは不十分なのである．効き方が認識できるようになるための第一ステップとして，その方法のカラクリをよく理解せよ，ということになる．

最近，統計学に関する入門書はかなり多く出まわっているが，どちらかといえば統計薬の“成分表”の羅列にしかすぎないようなものがほとんどのようである．この書により，成分表に関する知識を増強することはもとより可能であるが，できることなら統計に関する基本概念や各手法の“カラクリ”までをも

理解されることをお勧めしたい．統計薬の成分表の記載よりも，統計薬服用の際の処方箋を叙述することが，この書の眼目とするところだからである．かなり初歩的と思われるような概念でさえも，カラクリまで認識されてる向は存外少ないように思う．

この書を手にされる方で「統計はどうも判りにくくて……」といった感慨をもたれてる向が，結構いるものと推察される．統計の概念の中には，確かに判りにくいものも少なからずあると思う．これは何も統計の場合だけに限ったことではないが，一般に判りにくい概念を明解に解説したとしても，決して判りやすくはならないのである．これが判るようになるためには，書き手が明解な論理を展開することも必要であるが，読み手の方も相応の努力をしなければならない．ある程度基本的と思われるような概念を習得するためのエネルギー投与を，惜しんではならないということである．“苦学”とか“学問に王道なし”といったような言葉が何故に生じたのかを，再考願いたい．

この書の構成内容に関しての概観を与えておこう．

第1章は，いわゆる**記述統計**と呼ばれている分野の内容である．数学系出身者の記述した入門書ではこの部分の“手薄な”ものが多いので，意識的に重きを置いてみた．特に統計データの“背景”を読むという態度が身につけば，記述統計という分野が決して無味乾燥なものではないことを認識されると思う．“統計とは「平均」の学なり”といった表現もあるように，**“平均”**という概念を中心に据えて**データ簡略化の原理**を弁えることが，統計に関する諸概念を理解する上での基本であるという事実を，胆に命ずるべきである．

第2章と第3章は，推測統計にとっての基本的概念である**確率**と**確率変数**に関する話題が論議されている．特に確率変数という概念の認識が厄介である．この辺でどうもよく判らなくなった，という読者が多くなるという“統計的法則”もある位だから，気落ちせずに次章以降へ読み進まれたい．この部分に関する理解度が完全ではないような読者層をも考慮して，第4章以降の叙述を行ったつもりだからである．推測統計のメイン・パートは何といっても，第4章以降にあるからである．

第4章では，統計理論の“エース格”的存在である，**2項分布**と**正規分布**に関する内容をかなり詳しく記述してみた．統計理論の展開面で中心的役割を果たすこのような分布が，どのような“現象”を通じて導出されたものであるかに関して，特に熟読願いたい．無論“＊”印の付されている部分を読みとばされても，次章以降の読解に支障を来たすことはない．この代表的なもの以外の種々の分布に関しては，第5章，第6章で扱われている．ここでも，各種確率分布と呼応する現象の対応関係を明示してある．確率分布を単に形式的に認識しているだけでは，実際面で統計を利用する能力が身についたとは言いきれない．

第7章から第10章までは，推測統計に関する標準的内容のものが敍述される．母集団と標本との対応関係を把握しながら**標本抽出の原理**を認識すること，**統計的推定・検定の論理**を明確に認識しておくことがまずは基本的なことである．ここでも**標本分布**という概念が判りにくい．この概念を正しく認識できてるとはとても考えられないような人達まで統計の入門書を出したりする御時世だから，判らなかったからとて気落ちすることもない．

この書を執筆するにあたっては，筆者の及び得る範囲内で次のようなことを留意したつもりである．第一に，諸概念の導入部分では，“何故そのような概念の生ずる必要性があったのか”に関する叙述を行って，天降り的な定義を極力排除したこと．第二に，各章の構成を統計的な意味合いという観点から行い，形式的な整合性を主眼とした従来の類書にみられるような分類形式を踏襲してないこと．第三に，実際問題に即応した内容の例題を多くとり入れて，実際問題に対する処理能力が養成できるよう配慮したこと．かかる筆者の留意点が当を得たものであるかどうか，また筆者のそのような配慮が効を奏しているかどうかに関しては，疑問なしとはいえない．願わくば“肯定的”な答が返ってくることを期待しつつ，この書を世に出す次第である．

1977年6月

鈴木義一郎

目　　次

1 データ簡略化の原理

1.1 平均は何のために求めるか

「1と4の平均は？」と問うと，大抵の人は「2.5」と答える．「足し算には馴れている，平均ぐらいは計算できらあ」ということらしいので，「ではなぜ平均を求めるの？」とか「それをどう使うの？」と重ねて聞くと，今度は「さあ…」と首をかしげる人が多くなる．このように，“平均”という言葉は非常に親しまれている割には，その意味を正しく理解して使われてるケースが存外少ないように思われる．

1と4とを足して2で割れば，確かに2.5という数値が得られる．これを**算術平均**というわけだから，「1と4の平均が2.5である」と答えて間違いではない．だが**幾何平均**を求めよという場合だったら

$$\sqrt{1\times 4}=2$$

のようになるし，また**調和平均**だと

$$\frac{2}{\frac{1}{1}+\frac{1}{4}}=1.6$$

のように算出されるから，どのような意味での平均を求めればよいのかについて，もう少し慎重な態度で対処すべきなのである．特に，データの意味から解釈して納得のいく数値を対応させることが肝心である．たとえば，算数の成績が1で国語が4という生徒の成績は，総合でどう処理すべきか．5段階評価では“2.5”という数値を対応させることが許されないので，総合成績は2か3ということにでもなろう．

そもそも何のためにデータの平均を求めるかというと，データをひとまとめにして表現したいということである．データを何らかの判断の足しに用いよう

とすると，生のデータのままでは繁雑すぎて非常に判断しにくい．だから，判断しやすいような形にデータを**縮約**することが必要となる．

ところでデータが与えられたら算術平均を採るというのは，データの縮約の仕方のひとつの方法にすぎない．平均だけがいつもベストであるとは限らない．いろいろな縮約法を考えてみて，これがこの場合には最も良さそうだという判断を加味して求めた答と，ただ機械的に答を出したものとでは，データからの情報の吸収の仕方がかなり違ってくる．なぜ平均を求めるかという問に，ある程度答えられるような考え方をしておいて，なおかつ算術平均を計算したり，それを基に判断したりという考え方が必要になってくる．

結局，データの平均を求めるというプロセスは，基本的にはデータを判断のため都合のいいレベルまで簡略化するということである．一般に，あまり簡略化しすぎると肝心の情報を見失うし，かといって情報の損失を懸念しすぎると，いつまでたっても肝心の判断しやすいというレベルまでは簡略化できない．

1.2 データの代表値

1.2.1 (算術)平均

与えられた数量データを，$x_1, x_2, \cdots, x_n$ とするとき，データの総和をデータ数で割算した値のことを，**(算術)平均**といい，通例“$\bar{x}$”と表わす($\bar{x}$は“エックス・バー”と読む)．つまり

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}$$

さらに，紙数を節約する目的からは

$$x_1+x_2+\cdots+x_n=\sum_{i=1}^{n} x_i$$

というシグマ記号“$\sum$”を導入するのが便利で，この記号を用いることにすれば

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i$$

のように表わされる．

a, b を勝手な定数とすると

$$\sum_{i=1}^{n}(ax_i+b)=a\sum_{i=1}^{n}x_i+nb$$

であるから，辺々 n で割算して

$$\overline{ax+b}=a\bar{x}+b$$

という関係が得られる．この関係式は，“$\{ax_i+b\}$ というデータの平均が，$\{x_i\}$ というデータの平均に a を掛けて b を加えた値に等しくなる”という事実を示している．

このように，平均を求めるという手続き自体は簡単であるが，その意味を理解することは存外むずかしい．たとえば，年収 300 万円の人が 4 人，800 万円の人が 1 人いれば，5 人の年収の平均は

$$\frac{300\times4+800}{5}=400\text{(万円)}$$

のようになる．この“400”という値が，必ずしも具体的な意味をもってない．人数で割る前の 2000 万円，これなら 5 人の年収の合計値ということで比較的わかりやすい．

平均に対する物理的意味づけとして，次のように解釈することもできる．いま目盛のついたモノサシ状の棒を用意し，300 の目盛のところに同じ重さの玉を 4 個固定し，目盛 800 の位置には 1 個の玉を固定する．棒の重さは，玉の重さに比べて無視し得る程度のものとみなせば，左右ちょうどバランスするような位置の目盛が 400 ということになる．つまり，データの全体を“質点系”の分布とみなしたときの“バランス・ポイント”の位置が「平均値」ということになる．

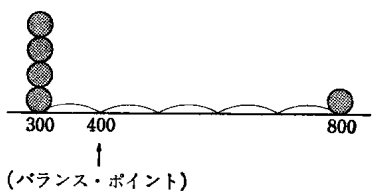

図 1

例 1 データ $(x_1, x_2, \cdots, x_m)$ の平均を $\bar{x}$
データ $(y_1, y_2, \cdots, y_n)$ の平均を $\bar{y}$
とするとき，

2 組のデータを合わせたデータの平均 $\bar{z}$ を，$\bar{x}, \bar{y}$ を用いて表現せよ．

解　$$\bar{z}=\frac{1}{m+n}\left\{\sum_{i=1}^{m}x_i+\sum_{j=1}^{n}y_i\right\}=\frac{1}{m+n}\{m\bar{x}+n\bar{y}\}$$

$$=\frac{m}{m+n}\bar{x}+\frac{n}{m+n}\bar{y}$$

例2　{3, 4, 4, 4, 6, 6, 7, 8, 13, 15}というデータの平均値は「7」であるが，なんとはなしに不公平な感じがするのはなぜか．

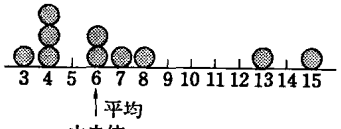

図 2

解　図を描いてみると，確かにバランス・ポイントは“7”であるが，平均値より小さなデータは6個もあるのに，逆に平均以上のデータは3個しかない．そこで，データを大きさの順番に並べてみたときに，ちょうど中央にくるものを考えてやる．この例だと「6」である．これを**中央値**（または**メディアン**）という．これだと，6より小さなデータと大きなデータとがそれぞれ4個ずつとなり，データのよい代表値とみなすことができる．

一般に，分布形のすそが片側のほうだけに尾を引いてるようなデータ（たとえば賃金分布のようなもの）では，平均値よりも中央値を考えたほうが無難である．平均値や中央値以外の代表値としては，**最頻値**（または**モード**）や**ミド・レンジ**のようなものも考えられる．前者は，最大頻度の観測値で，いわゆる“トップ・モード”のモードのことである．後者は，データの最大値と最小値との平均である．先の例では，モードは4，ミド・レンジ9となり，あまりよい代表値とはいえない．

1.2.2　平均のいろいろ

2組の正数 a, b の**幾何平均**は

$$\sqrt{a\cdot b}=e^{\frac{\log a+\log b}{2}}$$

で与えられる．図3をみれば

$$\sqrt{a\cdot b}\leqq\frac{a+b}{2}$$

という不等式の成立することがわかる．

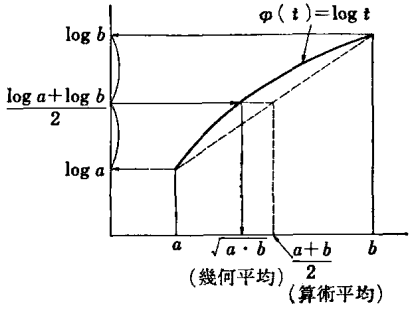

図 3

たとえば，10年後に元金の2倍になる定期預金に100万円預けておいた人が，5年後に解約を申し入れたら，150万円は要求できない．つまり金の増え方は幾何級数的だから

$$\sqrt{100万\times 200万}=141(万)$$

しか要求できないことになる．

一般に n 個の正数 $x_1, x_2, \cdots, x_n$ の幾何平均は

$$\sqrt[n]{x_1\times x_2\times\cdots\times x_n}=e^{\frac{1}{n}\sum_{i=1}^{n}\log x_i}$$

で与えられる．

また，2組の正数 a, b に対して

$$\frac{1}{\left(\frac{1}{a}+\frac{1}{b}\right)\Big/2}$$

のことを a, b の**調和平均**という．調和平均と算術平均との間にも

$$\frac{1}{\left(\frac{1}{a}+\frac{1}{b}\right)\Big/2}\leqq\frac{a+b}{2}$$

という不等式の成立することが，図4よりながめることができる．

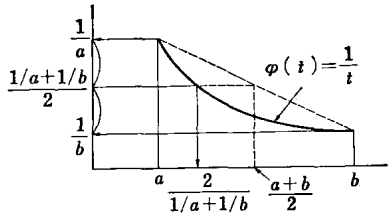

図 4

たとえば，行きは時速60 km，帰りは時速40 kmで往復したときの平均時速は50 kmではない．出発地から目的地までを仮りに120 kmとする．往復の所要時間は

$$\frac{120}{60}+\frac{120}{40}=5(h)$$

であり，全走行距離は

$$120\times 2=240(km)$$

であるから，平均速度は

$$\frac{240}{5}=48(km/h)$$

となる．これはちょうど60と40との調和平均になっている．

一般に n 個の $x_1, x_2, \cdots, x_n$ の調和平均は次のようになる.

$$\frac{1}{\frac{1}{n}\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\right)}$$

平均の概念をさらに一般化すると，次のようなものも考えられる. φ を勝手な(狭義)単調関数とする. あるyに対して

$$y=\varphi(x)$$

という関係を満足するとき

$$x=\varphi^{-1}(y)$$

という記号を用いる. つまりyがxの関数ならば，xの方もyの関数になる. ただxとyとは互いに“逆”の関係にあるから，φ という関数に対して φ^{-1} の方を**逆関数**と名づける. これは狭義単調な関数 φ に対しては一通りに定まる. たとえば

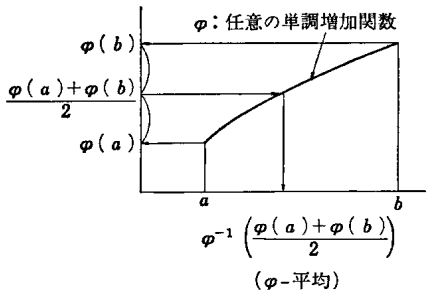

図 5

$$y=x \quad \text{ならば} \quad x=y$$

$$y=\log x \quad \text{ならば} \quad x=e^y$$

$$y=\frac{1}{x} \quad \text{ならば} \quad x=\frac{1}{y}$$

といった具合である. いま2つの正数 a, b に対して

$$\varphi^{-1}\left(\frac{\varphi(a)+\varphi(b)}{2}\right)$$

という数値を考え，これを a, b の **φ-平均**と呼ぶことにしよう. 一般に n 個の $x_1, x_2, \cdots, x_n$ に対しては

$$\varphi^{-1}\left(\frac{1}{n}[\varphi(x_1)+\varphi(x_2)+\cdots+\varphi(x_n)]\right)$$

が φ-平均である. 特に

$\varphi(t)=t$　　の場合の φ-平均が算術平均

$\varphi(t)=\log t$ の場合の φ-平均が幾何平均

$\varphi(t)=\frac{1}{t}$　　の場合の φ-平均が調和平均

ということになる.