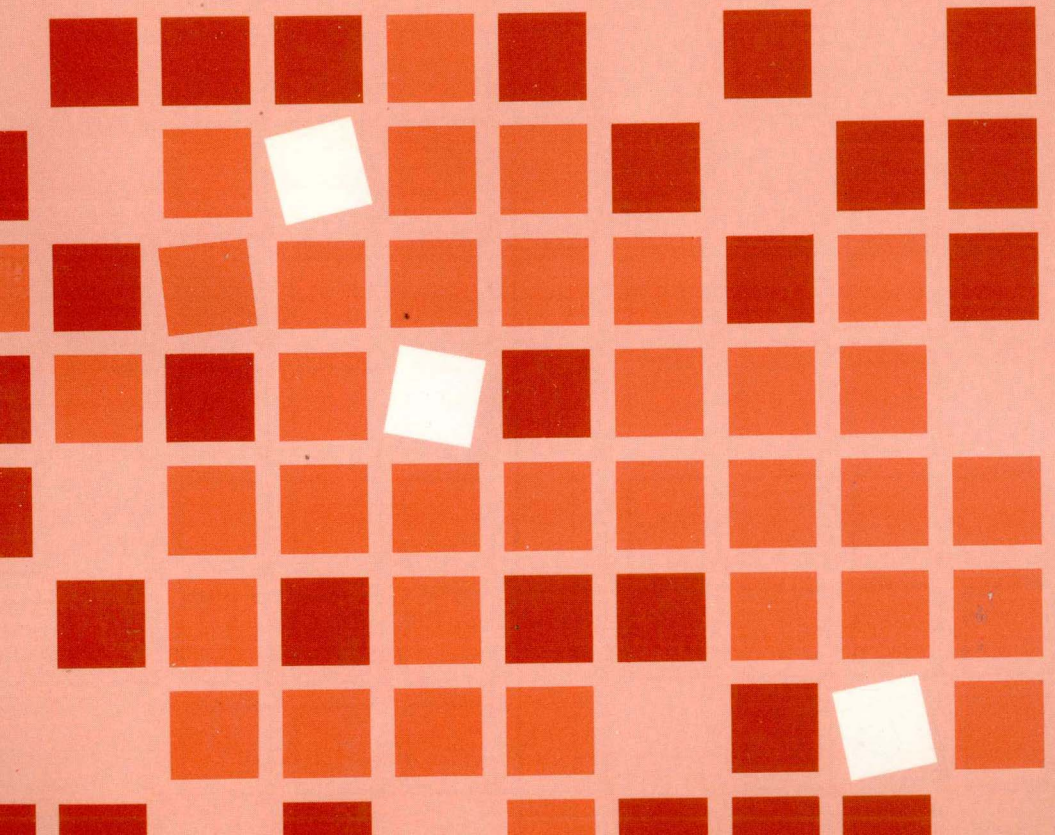
# 現代統計解析

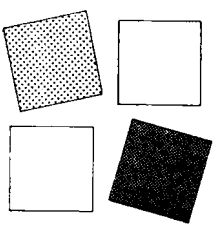
吉野　紀／後藤儀一郎

芦書房



# 現代統計解析

吉野　紀／後藤儀一郎



芦書房

**著者略歴**

**吉野　紀**（よしの・おさむ）

1940年　東京都に生まれる

1971年　早稲田大学大学院商学研究科博士課程修了
（経済統計学専攻）

現　在　駒沢大学経済学部教授

**後藤儀一郎**（ごとう・ぎいちろう）

1934年　山形県に生まれる

1971年　駒沢大学大学院商学研究科博士課程修了
（統計学専攻）

現　在　駒沢大学経営学部教授

**現代統計解析**

昭和57年5月1日　初版第1刷発行

著　者　吉　野　　紀
　　後　藤　儀一郎
発行者　中　山　一二三

発行所　株式会社 芦 書 房
〒101 東京都千代田区神田司町2—17
電話(03)293-0556　振替東京7-66145

小林印刷　三栄社製本

# はしがき

用具としての情報処理機器が発達し，処理能力の拡大が飛躍的に向上する中で情報化社会の到来が喧伝されて既に久しい。しかし，そのような社会で情報をどのように位置づけるべきかを積極的に吟味，検討する機会は未だ不十分としか言いえないように思われる。確かに，情報の処理能力を高める技術面の進歩は瞠目に足るものがあり，したがって，伝達される情報の量は着実に増加し続けているけれども，量の拡大をもって情報化社会と呼ぶことは決して正鵠を得ていない。むしろ量の拡大が著しいとき程，選択の重みが増してくるであろう。正しい情報と価値のある情報の取捨選択こそが要求されるべきであり，そのためには人それぞれの主体性と個々人の独自な判断能力と叡知とが前提として強く求められてくるであろう。

情報は徒手で坐したまま手に入れられるものではない。自ら自然や人間あるいは社会に能動的に問いかけ，回答が得られるように努力を傾けたときに始めて獲得できるものである。そこで，情報収集には発信と受信の感度いかんが常に問われる。何を対象に何をどのように尋ねるか，そしてどのような回路を経て受信されたものなのか，これらをすべて詳細に前提条件と明示した上で集められた情報源がデータと呼ぶに相応しい。個々のデータがすべてその背後にもつこれらの諸条件が忘れられては，真の情報を選別することは極めて難しい。データは，一定の目的と意図とをもって一定の方式で自然現象や社会・経済現象に問いかけをし，それに照応して返ってきた回答を数量標識で集計したものである。したがって，集計結果たるデータは問いかけを行った，すなわち発信を行った人の目的と意図を反映した内容をもつに過ぎず，また，その内容も自ら問いかけた方式に制約されており，事柄の本質を自動的に表明しうる構造となっていない。

さらに，データは，事柄のわずかな側面，しかも数量化可能な側面に限ってその姿をのぞかせているに過ぎない。データを通して事柄の本来もっている姿を透視するには，数量化しえない別の側面からもその事柄の本質に迫ってゆく能力が不可欠である。換言すれば，データから事柄の本質を把握しようとすればする程，数量でデータに表しえない質的側面から本質に近づく姿勢がより強く求められるということである。

統計学とは一般にデータの処理方法とそこから一定の結論を導き出す統計的帰納法とを学ぶ学問と考えられている。データを活かすためには確かに正しい処理方法を適用しデータのもつ潜在的な知識を引き出すことが不可欠であり，そのための統計的手法の重要性は増加こそすれ減退することは決してありえない。しかしながら，データのもつ知識を真に価値ある情報という高品位のものに高めるためには，データの必ずしも語りえない質的側面との連結環を常に意識しながら数量的処理に当たることがまた避けて通れない道筋でもある。この道筋を踏みはずさなければ，統計学がその意義を現在認識されている以上にさらに高めることは間違いない。

統計学を学ぶ上では，統計的方法あるいはより広い意味での統計科学自体の発展の歴史の中に，1つの顕著な傾向ないし特質が見い出せることをわきまえておくことも大切である。すなわち，統計科学の発展は，この科学自身に内在する芽から結実した部分が無いわけではないが，自然，社会，人文諸科学のあらゆる分野の発展に負うところが決して少なくないのである。これら諸科学はその発展過程で新たな課題を次々と発生させてきたが，その解決に取り組む過程で次々とデータの新しい処理方法の開発・展開をまた統計学に強く求めてきたのである。この刺激を受けることから統計学の発達が急速に進んだという事実は否定しえない歴史的特徴の1つであった。

このことを想起するとき，統計的方法を学ぶことは，一方でまた，それぞれの諸科学分野で扱われるデータの性癖にも深い理解を示すことでなければならないといえるであろう。この意味で，本書は，経済学ないし経営学との関わり

を強く意識して執筆されたことを1つの特色としていると言ってよい。したがって，扱われるデータは現実的な経済・経営現象を表すように選択されているはずである。このことは，経済学や経営学を学ぶ過程で出合う種々のデータから統計学への接近が図られる一方で，本書の例や演習問題に散見されるデータから逆に経済学や経営学への近しさを汲みとるよすがとなることも期待しうると考えている。

本書の構成は全8章から成っているが，全体を通じて一貫する流れにも意を尽した。それは，管理，制御された実験の行いえない社会科学で宿命的ともいえる標本観察値の意義と限界を繰り返し強調し，なおかつ，それらを一定の意志決定のための判断素材として有効に活用するという姿勢をとり続けたことである。

すなわち，第1章ではデータのまとめ方が述べられるが，そのデータも天賦のものとして野路にころがっているものをひろい上げてくればよいといった印象を与えることは極力避けた。そのことは，この段階で標本と母集団の関係に若干の言及が図られていることにも表れている。

したがって，第2章の確率も，母集団に関する知識を標本に基づく推測に求めるという統計学の基本姿勢に則って導入されたのであり，数学的な確率計算を主眼には決しておいていない。

第4章でも再び標本観察値の占める位置が概観され，標本データの特性とその背後にある母集団の特性との関わり方がより具体的に示されて，以下，第5章，第6章の推定，検定へと連結されてゆく。

第7章と第8章は経済分析等で多用される回帰・相関分析に当てられるが，ここでも標本データを回帰関係の上でとりまとめる段階と，母集団回帰関係の推定・検定とが対応させられるような構成となっている。前者が第7章で，後者が第8章で具体的に扱われるはずである。

本書で使用される数学的な道具箱は最小なもので足りる。そのため，定理等の証明はほとんど扱われていないが，数式で展開することをすべて除外するこ

とは，一方で，問題と結果をつなぐ部分をすべてブラック・ボックス化してしまうことにもなりかねないので，若干の部分については数理的な扱いを敢えて残した。

本書を著わす過程で御好意に与り得た財団法人・日本規格協会の大西正宏氏，飯泉貢氏には特に心からの感謝の意を表したい。また，芦書房の中山元春氏には著者達の遅筆にもかかわらず御寛容を頂いた点，記して謝意を表す次第である。

1982年3月

吉野　紀

後藤儀一郎

# 目　　次

# 第1章　記述統計 ——標本データの整理——

## 1.1 度数分布表

各種の経済分析を進め，具体的な経済政策を推進してゆく上で，家計がストックとしてどれだけの資産をどのような形態で保有しているかが大きな関心を呼ぶことは多言を要しないであろう。そこで，いま，昭和54年における全国の勤労者世帯の金融資産の保有状況を把握することに関心を持っているものとしよう。最も素朴には，当面の関心事たる全国の勤労者世帯が保有している資産額の数値全体が知られていることを期待するであろうが，現実には，それらを調査するための時間や労力と費用の制約からほとんど不可能に近い。実際には，全国の勤労者世帯の全体の中から，適当な方法で抜き出されたいくつかの世帯について調査して得た資産保有高の数値の集まりを入手する方途が採られる。統計学では前者の数値の集合体を母集団，後者のそれを標本(標本データ)と呼んでいる。したがって，母集団に関する情報，すなわち，当面の問題にとっては，全国勤労者世帯の資産保有高の分布を得ることを終局の目的とするが，それには，標本に関する情報から推測するという道を選ぶことが現実的となる。その方法を明らかにすることが統計学の主要な役割となっているのである。

最初に問題となるのは，標本となる勤労者世帯をどのように選んで調査対象とし，どのような方法で資産に関する回答を集計してゆくかである。基本的には，母集団たる全国の勤労者世帯を代表し，資産保有状況に関して母集団全体の状況をよりよく反映するような標本を選び出すことが肝要となろう。しかし

ながら，資産に関する母集団の状態は未だ不明であり，それを知ろうとすることからこの問題が出発してきているのであるから，どのような標本が全体の母集団を代表しているかを予め知ることはできない。そこで考えられるのは，種種の既知の情報をもとに，当面問題となっている資産保有状況に影響を与えると考えられる各種の属性が，とられた標本について片よりがないように配慮することである。

このように，片よりが生じないように標本を抜き出すことを，無作為に標本を抽出するといい，以下の統計的方法の適用に当たっては無作為標本がその前提になっていることを理解することが大切である。

いま，無作為抽出を基本的認識として全国の勤労者世帯から約4,000世帯が調査対象に抜き出されたとしよう。この対象世帯について資産保有高が質問され，虚偽のない回答が得られたとすれば，全国勤労者世帯の資産保有高に関する大きさ約4,000の標本データが入手されたことになる。それらは，具体的には，金融資産保有高を表す215（千円），1,836（千円），5,132（千円），829（千円），… 等々といった約4,000個の数値であるに違いない。この標本を構成している数値の個数を標本の大きさと呼ぶが，ある程度以上の大きさの標本になると，標本値を一べつしただけでは標本の性質を捉えるのに苦労するであろう。そこに標本値（データ）を整理する必要が生まれてくる。

**表 1-1　貯蓄現在高階級別貯蓄の1世帯当たり現在高**（勤労者世帯，昭和54年）

| 貯蓄現在高階級（万円） | 貯蓄額（千円） | 世帯数 |
|---|---|---|
| ～ 50 | 259 | 264 |
| 50～ 100 | 760 | 387 |
| 100～ 150 | 1,249 | 444 |
| 150～ 200 | 1,734 | 391 |
| 200～ 250 | 2,233 | 366 |
| 250～ 300 | 2,747 | 314 |
| 300～ 400 | 3,490 | 465 |
| 400～ 500 | 4,458 | 327 |
| 500～ 700 | 5,894 | 461 |
| 700～1,000 | 8,304 | 326 |
| 1,000～ | 15,356 | 309 |
| 計 | | 4,054 |

（出所）　総理府統計局『昭和54年貯蓄動向調査』

その最初の手続きとなるのは，データをいくつかの階級に分類することである。表1-1は，昭和54年における全国勤労者世帯の貯蓄動向を示

す分類表の具体的な例である。この例のように，多くの経済データの分類では小さい値をもつ階級に比較的データが集中してしまい，高い値の階級には少ないデータしかまとめられない場合がある。このようなときには，高い値をもつ階級では階級の幅を広げて，同一階級内に入るデータの個数を極端に小さくしないよう工夫がこらされている。

仮設的な実験データを次にとり上げて，階級分類の手順を追ってみよう。ある電球製造業者がいま，1つの製造工程から作り出された電球の品質に関心を寄せているものとしよう。この場合，品質は電球の連続的耐用時間のみで表されるものとしておく。彼にとっての関心事は，この工程で製造されたすべての電球（母集団）が規準値ないしそれ以上の耐用時間をもつか否かにかかるわけであるが，この母集団を構成する電球をすべて検査にかけてしまうと商品として市場に出すことができなくなることは自明であろう。そこで母集団から一定の方法に従って無作為標本を抽出し，その標本を破壊実験にかけて得たデータから母集団の特性を推定するという方法が有用となろう。いま，大きさ100の無作為標本から得られた電球の耐用時間に関する測定値が表1-2の通りであったとしよう。

**表 1-2 電球の耐用時間**

| | | | | |
|---|---|---|---|---|
| 1,126 | 1,380 | 1,254 | 995 | 1,252 |
| 995 | 1,187 | 1,412 | 1,072 | 1,312 |
| 1,222 | 1,258 | 1,316 | 1,133 | 1,341 |
| 1,243 | 1,358 | 1,014 | 1,402 | 1,363 |
| 1,134 | 1,013 | 1,147 | 1,266 | 1,056 |
| 1,398 | 1,221 | 1,185 | 1,551 | 1,023 |
| 1,261 | 1,294 | 1,067 | 1,226 | 1,213 |
| 1,216 | 1,265 | 1,447 | 1,315 | 1,296 |
| 1,269 | 1,280 | 1,303 | 1,202 | 1,227 |
| 1,216 | 1,043 | 1,293 | 1,293 | 1,444 |
| 1,415 | 1,287 | 1,356 | 1,285 | 1,319 |
| 1,203 | 1,510 | 1,272 | 1,388 | 1,185 |
| 1,455 | 1,497 | 1,234 | 1,245 | 1,287 |
| 1,350 | 1,031 | 1,368 | 1,461 | 1,021 |
| 1,001 | 1,257 | 1,253 | 1,342 | 1,291 |
| 1,223 | 1,078 | 1,178 | 1,172 | 1,278 |
| 1,252 | 1,270 | 1,257 | 1,253 | 1,119 |
| 1,327 | 1,299 | 1,247 | 1,179 | 1,549 |
| 1,246 | 1,174 | 1,161 | 979 | 1,377 |
| 1,195 | 1,063 | 1,259 | 1,129 | 1,110 |

この測定値の系列から標本データの特性を見やすくするために，これらをいくつかの階級に分類しておくことが必要である。そのためには，まず，階級の数を設定しなければならない。経験則からそ

の数は10から20の間に求めることが最も望ましいとされている。この前提の上に立つと，データの最大値と最小値の差を範囲(レインジ)として求めておいて，その1/10と1/20の間から階級の幅が決定されてくる。次いで，階級の境を定めておく。この際，各階級の境の中央に位する値となる階級値を考慮に入れながら設定することが望ましい。なぜならば，階級分類されたデータはすべてこの階級値によって代表されることとなり，以後計算に際してこの値が多用されるから計算上の便宜も階級値に要求される場合があるからである。一般的には，階級の境は，測定値の最小有効桁から0.5単位小さくとることによって，測定値と境界値との重なりを避けるように決められる。表1-3はこのような手続きを経て表1-2のデータを分類した一例である。

**表 1-3　電球の寿命の度数分布表**

| 階　　級 | 階級値 | 度数 | 相対度数 | 累積相対度数 |
|---|---|---|---|---|
| 942.5～ 997.5 | 970 | 3 | 0.03 | 0.03 |
| 997.5～1,052.5 | 1,025 | 7 | 0.07 | 0.10 |
| 1,052.5～1,107.5 | 1,080 | 5 | 0.05 | 0.15 |
| 1,107.5～1,162.5 | 1,135 | 8 | 0.08 | 0.23 |
| 1,162.5～1,217.5 | 1,190 | 13 | 0.13 | 0.36 |
| 1,217.5～1,272.5 | 1,245 | 24 | 0.24 | 0.60 |
| 1,272.5～1,327.5 | 1,300 | 18 | 0.18 | 0.78 |
| 1,327.5～1,382.5 | 1,355 | 9 | 0.09 | 0.87 |
| 1,382.5～1,437.5 | 1,410 | 5 | 0.05 | 0.92 |
| 1,437.5～1,492.5 | 1,465 | 4 | 0.04 | 0.96 |
| 1,492.5～1,547.5 | 1,520 | 2 | 0.02 | 0.98 |
| 1,547.5～1,602.5 | 1,575 | 2 | 0.02 | 1.00 |
| 計 | | 100 | 1.00 | |

具体的には，まず，表1-2から最大値1,551と最小値979をとり出して範囲を求めれば572となるので，階級の数を10とすれば階級の幅は57.2であり，20階級とすれば同じく階級の幅は28.6が目やすとなる。ここでは階級の幅 $c=55$ を選んでみた。こうして階級値を970とすることによって第1階級は(942.5～997.5)となり，最小値を含む2個のデータがここに入る。続いて第2階級以下，階級値と階級の境は55ずつずれた値をとり，最大値1,551を含む第12階級が設定されたところで分類表の枠組みができ上る。各階級の境界値は観測データの測定値単位より1桁下回っているので，すべてのデータは必ずいずれかの階級に自動的に所属させられるはずであり，境界値と重なり合うことはない。

所属データの個数を数え上げ，その値を度数として記録すれば表1-3が完成するわけである。なお，同表には相対度数および累積相対度数も併せて埋めてあるが，前者は各階級に属する度数をデータの総数で除した値であり，後者は，ある特定の階級に至るまで下位の階級の相対度数を順次加算して求めた値である。

この分類表を度数分布表と呼ぶが，それは表中の度数，すなわち，各階級に属する測定値の個数が標本データの集合の特徴を最もよく代表すると見なされるからである。

**表 1-4 度数分布表の一般型**

| 階級番号 | 階　級 | 階級値 | 度　数 | 相対度数 | 累積度数 |
|---|---|---|---|---|---|
| 1 | $x_1-\frac{c}{2}\sim x_1+\frac{c}{2}$ | $x_1$ | $f_1$ | $f_1/n$ | $F_1$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $j$ | $x_j-\frac{c}{2}\sim x_j+\frac{c}{2}$ | $x_j$ | $f_j$ | $f_j/n$ | $F_j$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $k$ | $x_k-\frac{c}{2}\sim x_k+\frac{c}{2}$ | $x_k$ | $f_k$ | $f_k/n$ | $F_k$ |
| 計 | | | $n$ | 1.00 | |

度数分布表の一般型を表1-4に示すごときものとすれば，そこから引き出される自明の事柄は次のようにまとめることができる。

標本の大きさ　$n=f_1+\cdots+f_k$

$$=\sum_i f_i \qquad i=1,\cdots,k$$

累積度数　$F_j=f_1+\cdots+f_j$

$$=\sum_i f_i \qquad i=1,\cdots,j,\ j\leqq k$$

相対度数　$\dfrac{f_i}{n}$

階級の幅　$c$

表1-3の度数分布表をもとにデータの分布の有様をより直観的にみるために

グラフに表したものが図1-1である。このような柱状グラフをヒストグラムと呼ぶ。このようなヒストグラムを描くことは，直観的また視覚的ながら，しばしば分布に関する有用な情報を与えてくれる。図1-2は表1-1から相対度数を用いて描いたヒストグラムであるが，電球の寿命に関する図1-1に比べてデータの分布の左側に片よる程度が相対的に強いことを読みとることは容易である。なお，このヒストグラムは，表1-1で階級幅が不統一となっている階級に

**図 1-1　耐用時間のヒストグラム**

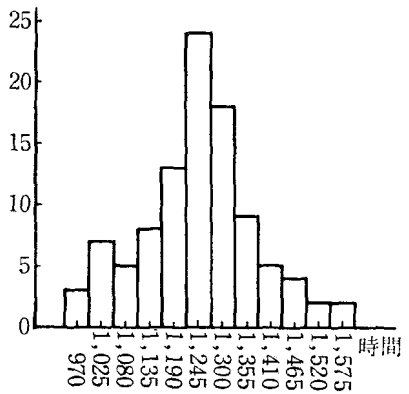



**図 1-2　貯蓄保有高のヒストグラム**（勤労者世帯，昭和54年末）