はじめに

「統計解析って、統計学とどう違うんだろう？」――そんな疑問をもっている方が多いのではないでしょうか。

実際、『統計学がわかる』といった本を読んで、なんとなく理屈がわかったとしても、そのあと、現実の場で統計の知識を活かして使っている方はほとんど見あたりません。それは「統計解析の知識が不足しているから」といってよいでしょう。

統計解析というのは、**統計学の知識を応用しつつ、実際に統計データの分析を行なえるようにすること**――なのです。

ですから、統計解析を身につけることは統計の知識を実践的に使うことであり、また、特別な準備も不要です。本書では統計をイチから説明していますし、その使い方がわかるよう具体的な事例を通して伝えていきます。ただ、あなたに一つだけ用意しておいてほしいのは「統計学とは何か？」「統計学はどう利用されるのか？」という好奇心だけです。

情報化時代といわれて久しいですが、最近ではツイッター、ブログなども含め、日々新しく生まれるデータがますます巨大化し、それらのデータが互いに融合し、複雑化しています。それを「ビッグデータの時代」などとも呼んでいます。このような時代にあって、統計的分析能力の素養を身につけておくことは、ますます重要さを増しています。それには二つの理由があります。

一つ目の理由は、統計解析を活用する立場から見たものです。IT社会であふれるデータの活用法を知らないと、データは単にゴミの山にしか見えません。けれども、ほんの少しでも統計解析の素養を持っていると、それは情報の宝の山にも変容します。データに対して、「こんな見方もでき

る」「そんな解析もしてみたい」と、好奇心が刺激されます。

二つ目の理由は、統計解析を受け止める立場の話です。猛烈に発信されるデータは、現在、さまざまに解釈されながらマスコミ等で発表されています。困ったことは、その解析は必ずしも正しいとは限らないことです。しかし、ほんの少しの統計解析の素養さえ持っていれば、その誤りを見抜くことができます。

統計の扱いを評した有名な言葉があります。

There are three kinds of lies: lies, damned lies, and statistics.

（世の中には3つのウソがある。ウソと大ウソ、そして統計だ。）

これは19世紀後半のイギリスの首相ベンジャミン・ディズレーリの言葉です。ディズレーリは、「統計のウソ」はウソの中でも最大級だとしているわけですが、それだけに、「統計のウソ」を見抜くには統計解析の素養が必要なのです。

さて、その統計解析ですが、具体的にはどんなものなのでしょうか。次のA君の話からイメージが得られると思います。

工場の製品管理部門に回された新入社員のA君は次のように上司から命じられました。

「当社の人気商品のスナック菓子Sの内容量が100 gずつ正確に入っているかどうか、調べなさい」

そこでA君は製造ラインから100袋を**無作為**（むさくい）（アットランダム）に抜き出し調べました。その平均値を計算すると99.7gとなりました。この値から、A君はどうやってラインで製造される菓子の平均内容量を知ることができるでしょうか。

このようなケースに対処する統計解析法が**推定**（**統計的推定**ともいう）です。「一を聞いて十を知る」という諺がありますが、「1を調べてすべてを知る」ことが統計的推定の極意なのです。

翌日、A君は「平均値が99.7 gである」という事実を上司に報告しました。すると、今度は次のように命じられました。

「なるほど……。原因としては、製造ラインの機械が狂っているのかもしれないし、単なる誤差かもしれないな。確かめてみなさい」

確かに、たまたま検査した100袋の平均値が99.7gにすぎず、1万袋を検査してみれば100 gだったかもしれません。とすると、製造誤差の許容範囲内ともいえますが、もしこれが99.5 gだったらどうなのか……。「さて、どう対応すればよいものか」と、またまたA君は悩みました。

このような問題に応えるのが**検定**（**統計的検定**ともいう）です。得られた少ないデータから、「製品の内容量は100gで正しい」という仮定が正しいか否か、それを判定する手段を提供してくれます。

こうしてA君は上司に対してどうにか報告を済ませたところ、数か月後、再び難題が降りかかりました。スナック菓子Sの製造ラインの効率を上げるために3案X、Y、Zが出されたのですが、それらの優劣を確かめるためのチーム主任に任命されたのです。

そこで、A君は実験用ラインを設け、従来方式も含めて各案をテストすることにしました。各案を採用したラインから1分間に製造される製品数を5回に分けて計測すると、次の結果が得られました。

	1回目	2回目	3回目	4回目	5回目	平均
従来	30	29	31	33	32	31.0
X案	31	32	30	33	32	31.6
Y案	31	33	29	33	33	31.8
Z案	32	33	31	33	34	32.6

X～Zの各案はすべて、従来方式よりも1分間当たりの製造数は増えています。その中でもZ案が最も優れた結果を出しています。しかし、たった5回しかテストしていないのですから、誤差の範囲とも思えます。A君の報告書しだいでは、会社は製造ラインの変更という大きな投資を決定するかもしれないのでA君は心配です。

A君はこの場合、「改善の効果はあった」という報告書を書くべきなのでしょうか。それとも「従来方式に比べ、どの案も新規に採用するほどの効果は見いだせなかった」と報告すべきなのでしょうか。

このA君の疑問に応えるのが**分散分析**です。分散分析は得られたデータから、効果の有無を検証してくれます。「改善案の違いの効果はあった」などという結論を勘（カン）ではなく、統計的に導き出してくれるのです。

報告書作成に疲れたA君は、週末、山に行くことにしました。ホームページで週末の天気予報を調べると、雨の確率が30%と表示されています。

A君「雨の確率が30%か」

それを聞いた同僚のB子さんは、A君に質問しました。

B子「雨の確率が30% ってどう意味かしら?」

A君「同じ条件の日が100日あったなら、そのうちの30日に雨が降る、という意味だと思うけど」

と、教科書的に応えました。するとB子さんは次のように反論したのです。

B子「気象って複雑でしょ、同じ条件の日が100日もあるわけはないでしょう」

そういわれると、もっともな話です。A君は確率に関する知識が不足していることを知り、困惑してしまいました。

A君のこの困惑に応えるのが**ベイズ統計学**です。気象予報には気圧配置などの統計データとともに、予報官の経験やカンが蓄積としてあるわけです。ベイズ統計はこれらの個人的な蓄積も情報として取り入れて確率を算出できます。人間味のある統計学なのです。

以上のA君の例で、統計解析の日常性と重要性、そして面白さが垣間見えたと思います。

最初に示したように、統計学のアイデアや、それを実現する解析法をマスターするのにむずかしい準備は不要です。面倒な計算はExcel等の統計解析ツールが実行してくれるからです。大切なことは、**何が問題で、どうやってその結論が出るのか**――その過程を理解しておくことです。本書はそのために、例題を通してそれらが身につくように詳述してあります。例題の意図と解決の流れをゆっくり追っていけば、統計解析のエッセンスがつかめるはずです。

本書の解説には、中学までの数学しか利用していません。代わりに、統計学で訴えたいアイデアはグラフに示しています。掲載したグラフを眺めながら、本文の意味を確認していただければと思います。

本書によって、情報化社会においてデータの山に呑まれず、情報の海に染まらず、それらを活用する素養が提供されることを深く希望します。今後、私たちの周りはますますネットワーク化され、データ、情報が氾濫していくでしょう。それに対応するためにも、この社会を楽しめる武器として統計解析の力を身につけてください。

涌井　貞美

1. 統計学を２つに分類すると
～一部から全体を推し量る

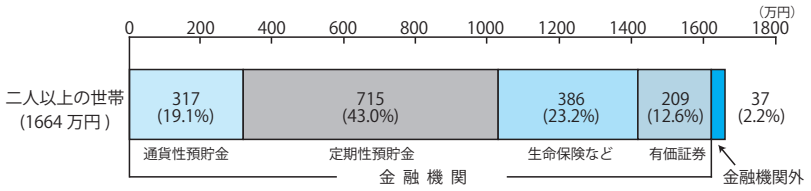
世の中では「統計学」という言葉がいろいろな意味に使われているけれど、大きく分けると記述統計学と推測統計学に。

■記述統計学は見やすくまとめること

統計学の狙いを一言でいえば、「データの裏にある本質を理解すること」にあります。そのアプローチの方法によって、統計学は**記述統計学**と**推測統計学**の２つに大きく分類することができます。

調査や実験で集めたデータをまとめて整理し、表にしたりグラフ化するのが記述統計学です。得られたデータをビジュアルにして直感的に理解できるようにすることで、データの裏にある本質に迫ろうとするわけです。

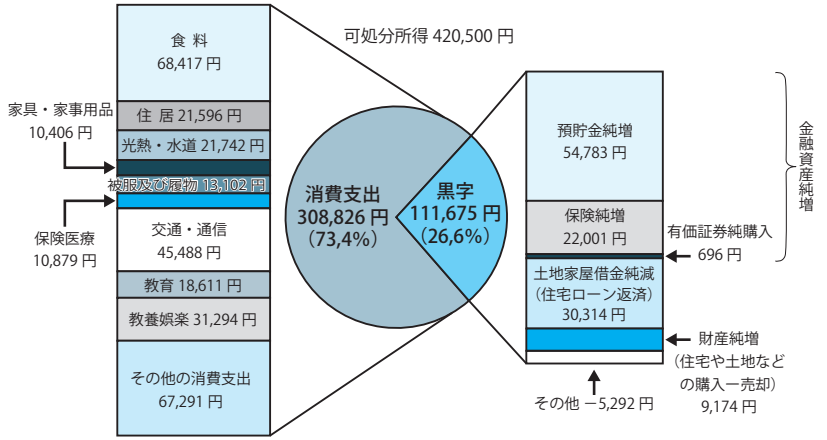
たとえば、次のグラフを見てみましょう。



（出典）総務省統計局（http://www.stat.go.jp/）

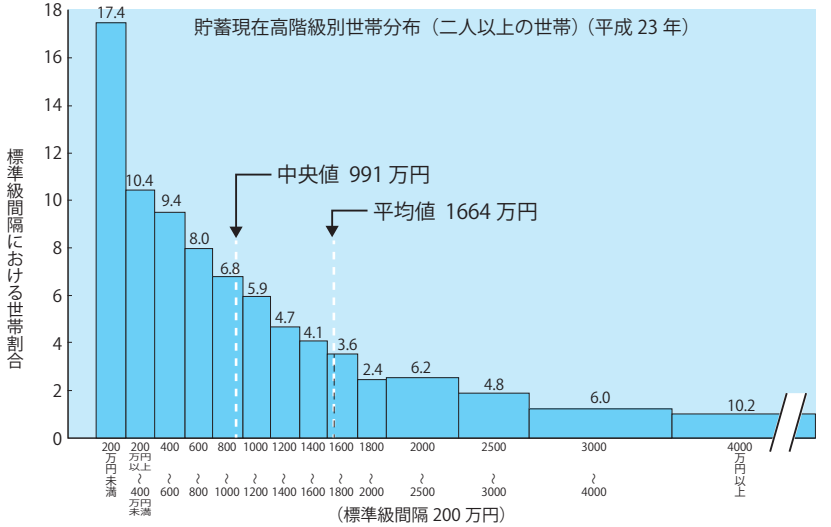
これは１世帯あたり（２人以上）の平均貯蓄額1664万円（平成23年）が預金や株など、どのような形で保有されているかを示したもので、**帯グラフ**です。帯グラフは、このように、「全体に占める構成の割合」を示すのに優れています。

次ページのグラフは、２人以上の勤労者世帯の平均可処分所得の月額420,500円（平成23年）がどのような構成かを表わしたグラフです。帯グラフに加え、中央に**円グラフ**が載せられています。円グラフも帯グラフ同様、全体に占める構成の割合を示すのに優れています。



（出典）総務省統計局（http://www.stat.go.jp/）

さらにまた、次のグラフは**棒グラフ**です。これは２人以上の世帯がどれくらいの貯蓄額があるかを示したグラフです（平成23年調べ）。帯グラフ以上に、データの特性を細かく表示するのに向いています。



（出典）総務省統計局（http://www.stat.go.jp/）

さて、この最後のグラフには中央値、平均値という言葉が記入されています。これらは資料の**代表値**と呼ばれる数です。集めた膨大なデータを整

理し、「大まかな数」として表現します。そうすることで、細部に入り込みすぎると見えにくくなる大きな全体の姿が見えるようになります。「木を見て森を見ず」という表現がありますが、そうならないために不可欠な表現法です。グラフ表示だけでなく、このような数値化も記述統計学の大切な仕事です。

■推測統計学は「一部から全体を推し量る」

統計学のもう一つの分野である**推測統計学**を見てみましょう。次の2つの統計的な記述を見てください。

- 警察庁の発表によると、2011年の女性の運転免許保有者数の割合は44%である。
- 「平成24年全国たばこ喫煙者率調査」（JT）によると、約2万人を対象にした調査の結果、日本人成人の平均喫煙率は21.1%であった。

前者の「女性の運転免許保有者数の割合は44%」という数値は、警察庁が日本全国からデータを収集して算出した結果です。日本人すべての運転免許保有者数を対象にしていますから、これを**全数調査**と呼びます。全数調査は多くの手間と時間、そして予算が必要になります。

それに対して後者の「喫煙率が21.1%」という数値は、日本人すべてを対象にした結果ではありません。1億人余りの日本人成人の中から2万人を**無作為**（むさくい）に選び出し、喫煙実態を調査した結果です。このように、たくさんの中から一部を取り出して調査する方法を**標本調査**と呼びます。標本調査の良い所は対象が小さい分、時間と手間と予算が節約できることです。

推測統計学が本領を発揮するのは、この標本調査により得られた資料の分析です。ただし、標本調査には常に、次のような疑念が伴います。

「一部から得られた結果を全体にあてはめて大丈夫か？」

たとえば、上記の喫煙率の例でいうと、「たかだか2万人から得た『喫煙率21.1%』というデータから、1億人以上の日本人成人全体の喫煙率が本当にわかるのか？」という疑問が生まれます。わずか0.02%の人から取ったアンケート結果（標本調査）なのですから、当然です。この疑念、難問に応えようとするのが推測統計学の仕事なのです。

我々の目にする資料の多くは標本調査によるものです。アンケート調査、品質調査、実験結果などは、ほとんどが全数調査ではなく、ほんの一部を抜き出して調査します。そこで、推測統計学の出番は非常に多いことがわかります。

■統計数字に惑わされてはならない

統計学というのは、資料を扱う幅広い分野を指します。その1分野に**統計解析**があります。統計解析は、最初に述べた記述統計学（グラフ表示など）ではなく、あとで説明した「推測統計学」を中心とする、実用的な統計分析の手法を提供します。統計的な推定、検定、分散分析、相関分析などが具体的なテーマなのです。

ところで、統計学の対象となるデータは人が集めるものであり、統計学の結果を発表するのも人、発表された結果を受け止めるのも人です。したがって、扱い方によって解釈はさまざまで、誤用され、意図して悪用されます。それを言い表わしたのが、「はじめに」にも示した次の言葉です。重要な言葉ですので、もう1回、掲載してみました。

> *There are three kinds of lies: lies, damned lies, and statistics.*
> （世の中には3つのウソがある。ウソと大ウソ、そして統計だ。）

統計学の分析結果は単純に数値であり、それを**解釈するのは人間**です。そのことを常に肝に銘じ、統計学の結果に対して公平無私の態度で対峙（たいじ）する習慣をつけることが最も大切なことなのです。