2009 年度前期　データ解析序説　第 1 回

データ解析とは ー イントロダクション

一人の死は悲劇だが，数百万の死は統計にすぎない．ー スターリン

データ解析とは

データ解析というと，データを集めて表やグラフに整理したり，平均を求めたり... というものを想像するのではないでしょうか．実は，それはこの講義では「データをまとめる」の部分でしかありません．

データ解析には，その続きがあります．ひとつは，集めたデータをもとに，そのデータがどういう仕組みでできあがったのかを調べる**記述統計学**です．もうひとつは，確率の考えを用いて，集団のうちの一部のデータを調べて集団全体の姿を推測する**統計的推測**です．

今日のイントロダクションでは，「分布」「モデル」「くじびき」「リスク」の４つのキーワードでデータ解析の考え方を説明したいと思います．

データ解析のキーワード（1）:「分布」ー 統計学が扱うもの

これまで学校で習ってきたことは，１つの問いに対して１つの結果がはっきり決まるものがほとんどでした．例えば，

- 「1+1=?」「2」
- 「２モルの水素と１モルの酸素が完全に反応すると？」「$2H_2 + O_2 \rightarrow 2H_2O$ だから，２モルの水ができる」

といったものです．しかし，現実世界では，上のような問題よりも

- 「日本男性の身長は？」「人によって違う」
- 「100cc の水素と 100cc の酸素が反応すると？」「実験条件によってできる水の量は違う」
- 「ある夫婦に次に生まれる子供は男か女か？」「生まれてみなければわからない」

という問題に出会うことのほうが多いものです．

後者の問題では，対象にしているデータが，時と場合によってばらばらになっています．人がこれを「ばらばら」と感じるのは，データが得られる仕組みを人間が完全に把握することができず，それを「神様がさいころをふって決めている」と考えているからです．このように，ある測定対象や現象から得られるデータがばらばらであることを**分布する**といい，このような分布したデータが現れる現象を**ランダム現象**といいます．統計学が扱うのは，ランダム現象によって生じた，分布しているデータです．

上の例では，「日本男性の身長」や「上の実験でできる水の量」や「次に生まれる子供の性別」は分布する，ということになります．しかし，上の問答のように「わからない」と言ってしまっては身も蓋もありません．

そこで，例えば，日本人男性の身長を何人か調べてみるとします．身長は分布していますから，165cm だったり 170cm だったり 180cm だったりすることでしょう．このとき「何 cm くらいの人が何人いたか」を表やグラフにしてみると，何 cm くらいの人が多いかを読み取ることができます．さらに，「調べた人の身長の平均は何 cm か」という，分布を特徴づける数値を求めることもできます．

データ解析のキーワード（2）:「モデル」ー 説明への欲求

人は，観察される現象が「どんな仕組みで」起きているのかを理解したいという欲求を，常にもってきました．この「仕組みの理解」こそが「科学」であり，仕組みを理解することによって未知の現象を予測することができます．そこで，「仕組み」を人間が理解できる言葉や数式で記述したモデルを仮定し，それを使って現象を説明する方法をとります．例えば，前ページでにあげた化学式もモデルのひとつです．「水素と酸素が反応すると水ができる」という観察結果だけでは，それ以上のことは何もわかりません．しかし，水素や酸素の分子が分解・結合するというモデルでこの観察結果を説明することで，他の化学反応も同様に説明でき，また未知の反応も予想することができます．

この考え方は，この講義の後半で説明する，データの組どうしの関係を記述する**多変量解析**では，さらに有効です．例えば，日本の各都市について，緯度と年平均気温というデータの組を集めたとします．このデータを並べてみても，なんとなく「北へ行けば寒くなる」ということしかわかりません．しかし，ここで「緯度－気温のグラフが直線になる」というモデルを用いると，「1 度北へ行くと何度寒くなるのか」を予想することができます．さらに，観察されたデータをこのモデルでどのくらい説明できているかを調べることで，モデルの適切さの程度や，データを説明する他の要因（「標高」など）を考えることができます．この手法は**回帰分析**とよばれ，第 5, 6 回で説明します．

さて，このようにして分布を表現することはできますが，ここまでは今調べて手元にあるデータについてのことしか述べていません．前ページの例でいえば，「今調べた日本人男性の身長の分布」を表現しただけにすぎず，他の人のことは何も言っていません．つまり，上のような分布の記述は「観察」にすぎません．「日本男性の身長の分布」という問題になると，調べるだけでも大変で，「観察」すら簡単にはできません．この場合，一部だけの観察結果から未知の集団全体のようすを推測する必要があります．この手法が**統計的推測**です．

統計的推測で用いるモデルは，**確率分布モデル**というものです．一見ランダムに分布している現象でも，「どうランダムなのか」を理解し説明する方法を，人は常に求めてきました．確率分布モデルとは，どんな分布かを決めている「神様のさいころの形」を説明する方法です．例えば，子供が産まれるとき，男児が生まれるか女児が生まれるかはランダム現象です．ですから，過去の出生をすべて調べて，男児の割合が半数よりも少し多い 0.517 であることがわかっても，次の出生で男児が生まれるか女児が生まれるかは何もわからないはずです．しかし，私たちはなんとなく「次の出生で男児が生まれる確率は 0.517 だろう」と想像しています．それは，男女どちらが生まれるかは，ただばらばらなのではなく，「男児／女児が生まれる確率は常に一定で，各出生は独立（互いに無関係）である」と想定しているからです．これは，つまり「神様はコインを投げて男女を決めている」というモデルを考えているのと同じです．このモデルが正しいかどうかを証明した人はいません．しかし，経験上このモデルを考えて問題ないことはわかっています．

コインを投げつづけたとき，表がある回数出る確率がどのくらいかは，「２項分布モデル」という数式

で表すことができます．そこで，これを使うと，将来生まれてくる子供の性別の分布を予測することができます．また，ある物の長さを何度も測ると，測定値はいつも同じではなく，分布します．この分布は「正規分布モデル」という数式で表されることが知られており，これを使って，本当の長さはどのくらいか，などを考えることができます．

データ解析のキーワード（3）:「くじびき」ー 標本抽出の原理

統計的推測の原理は，実は身近な「くじびき」と同じものです．いま，くじ箱の中にくじがたくさん入っているとして，「当たり」が全体のくじの本数のうち 50%，「はずれ」が 50%であるとします．このくじ箱の中では，くじの「当たりはずれ」が分布していると考えることができます．

このくじ箱から，公正なくじ引きで 1 本くじをひいたとしましょう．このとき，ひかれたくじが「当たり」である確率は 50%，「はずれ」である確率も 50%であることは容易に想像がつきます．つまり，「当たり／はずれが選ばれる確率」は，箱の中のくじの数の割合と同じです．

統計的推測とは，この例でいうと「箱の中の当たりくじの数の割合がわからないとき，くじをひいて，その結果から箱の中の当たりはずれの分布を推測する」という問題になります．このとき，くじ引きで当たり／はずれがそれぞれ選ばれる確率がわかれば，その確率がすなわち箱の中のくじの数の割合と同じですから，箱の中のくじの分布がわかる，つまり統計的推測ができることになります．

ところが，当然ながら，1 回だけくじをひいて例えば当たりが出たからといって，当たりが選ばれる確率は想像することもできません．そこで，くじをひく数をもう少し増やしてみます．箱の中の当たりくじの割合が 50%でも，何本かくじをひいたときの当たりの本数の割合は 50%とは限らず，いろいろな場合がある，つまり分布します．くじを何回もひいて，そのうち半分当たりが出れば，「当たりが出る確率は半分くらいだろう」→「箱の中の当たりくじの割合は半分くらいだろう」という推測ができます．この推測の確かさは，ひいたくじの本数が多いほど高まります．

ここで，モデルの考え方を使わなければ，「当たりは半分くらいだろう」以上のことはわかりません．しかし，このときの当たり本数の割合は，あるモデル（２項分布モデル）で表されると考えられます．これを用いると，そのモデルを使った計算によって「当たりくじの割合が 40%〜60%の範囲にあると，95%の確かさで言える」という表現で，推測の確かさを表現することができます．この方法が統計的推測の手法の 1 つ**区間推定**です．

また，「日本人男性の身長」の問題について考えてみましょう．「日本人男性全体のうち，身長が 170〜175cm の人の割合」が仮に 20%だとします．このとき，日本人男性全体からくじびきでひとりの人を選んだとき，その人の身長が 170〜175cm である確率は，上の説明と同様に，20%です．したがって，上と同じ考え方で，何人かをくじびきで選べば，「身長が 170〜175cm の人の割合」を推測することができます．同様に，「身長が 165〜170cm の人の割合」「身長が 175〜180cm の人の割合」なども推測することができます．したがって，身長をこのような区間に分けて表しておけば，「日本人男性全体の分布」を，くじびきで選んだ何人かの人の身長から推測することができます．さらに，「正規分布モデル」を使えば，「日本人男性の平均身長は 168〜172cm の範囲にあると，95%の確かさで言える」という区間推定をすることもできます．

統計学の言葉では，ここでいう「くじびきで選んだ何人かの人」を**標本**といい，公平なくじびきで選ぶことを**無作為標本抽出**といいます．また，「身長 170〜175cm」といった区間を階級といい，階級で表現された分布を**度数分布**といいます．これらについては，次回の講義から順に説明してゆきます．

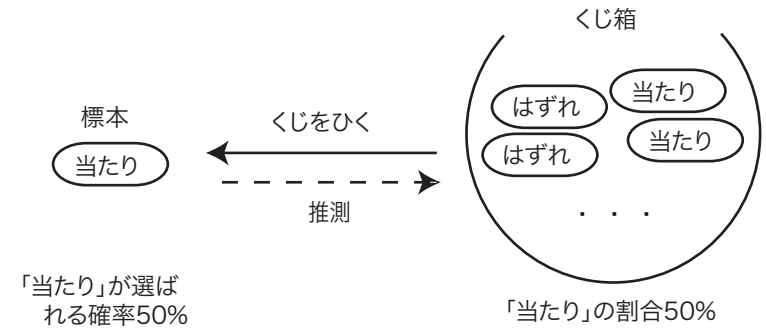

図 1: 標本とくじびき

データ解析のキーワード（4）:「リスク」ー 誤差ではなく，失敗の確率

統計的推測は，データの全体を調べずに未知のデータ全体の傾向を調べるのですから，ある程度は外れています．このことを「推測は，ほぼ当たっている」といっては誤りです．正しくは「推測は，たいてい，ほぼ当たっている」と言わなければなりません．これはどういう意味でしょうか？

それは，統計的推測の結果は，いつもほぼ正確なことを述べているのではなく，「ほとんどの場合ほぼ正確なことを言うが，大外しをする確率もわずかにある」という意味なのです．この大外しする確率が「リスク」です．

例えば，くじを 10 本ひいて，9 本が当たりだったとしましょう．おそらく，「箱の中の当たりの割合は大きいだろう」と推測することでしょう．しかし，もしかしたら，今ひいた 9 本以外は，箱の中のくじはすべてはずれかもしれません．もしそうだとしたら，「当たりの割合は大きい」という推測は大外しです．そんなことはめったにないのはその通りですが，問題は，今推測した結果が，**大外しなのかそうでないのかは，わからない**ということです．

さきほどの「区間推定」で，「当たりくじの割合が 40%〜60%の範囲にあると，95%の確かさで言える」という表現をしました．この「95%の確かさで言える」というのは，例えて言うならば「私は 95%の確率で正しいことを言いますが，5%の確率で大嘘をつきます」という予言者が，ある問題について 1 回だけ述べる予言のようなものです．今回の予言も 95%の確率で当たっているはずですが，一方今回の予言に限って大外しである確率も 5%あります．そして，今回の「40%〜60%」という推測結果が，95%の正解のほうなのか，5%の大外しのほうなのかは，今回の結果を見てもわからないのです．

これが，「5%のリスク」の意味であり，それが「確率」というものの本質です．「大嘘をつく確率が小さい」とは「予言をする機会が何度もあるとすれば，そのうち実際に大嘘をつく回数は少ない」という意味であって，ある 1 回の機会については何とも言えません．

今回の予言を「ふだん 95%の確率で当たる予言者の言うことだから，今回も信じよう」と思うかどうかは，「5%の確率で大外しする」というリスクを受け入れられるかどうかの問題になります．この予言者と長いつきあいがあって，「5%くらい嘘をつかれて損をしても，ふだん正しい予言によって儲けさせてもらっているから埋め合わせできる」と思えるのならそれでもいいでしょう[1]．しかし，1 回しかない勝負の機会にこの予言者の言うことを信じるかどうかといえば，これはほとんど「運」の問題になります．

このことは，統計的推測は「ある 1 回の機会に何が起きるか」を推測することは本当はできず，それ

[1] こういう「埋め合わせ」の考え方が，リスク管理や保険の基礎となります．

をすると大外しのリスクをともなうということを意味しています．統計的推測は，さきほどの予言者との「長いつきあいによる埋め合わせ」のような状況を考えて，はじめて正確な意味をもつのです．冒頭のスターリンの言葉に象徴されるように，統計学は個々のケースを考えるのではなく，全体としての傾向を考えるのです．

実際の統計的推測では，調べるデータが多くなるほど，大外しのリスクは小さくなります．例えば，日本人男性を 1 人調べて，その人の身長を「日本人男性全体の身長の平均の推測結果」だと言っても，大外しである確率は高いでしょう．また，くじを 1 回だけ引いて当たったからといって，「このくじ引きは必ず当たる」と推測しても，まず信用できないでしょう．しかし，さきほど「当たりくじの割合の推測の確かさは，ひいたくじの本数が多いほど高まる」と述べたように，多くの数のデータを調べれば，大外しのリスクを小さくすることはできます．そして，先に述べた確率分布モデルを用いると，データを調べた数に応じて「大外しをする確率」を見積もることができ，「当たりくじの割合が 40%〜60% の範囲にあると，95%の確かさで言える」というようにリスクを数値で述べることができます．

このように，リスクの程度とは「どの程度重大な失敗か」を表しているのではなく，「どのくらい頻繁に失敗するか」を表しています．これは，「95%の確率で当たる」予言者が今日述べたひとつの予言が当たっているかどうかは言えず，また「95%」という評価には「外した予言が，どのくらいとんでもなく外れているか」は含まれていない，ということと同じです．統計的推測の確かさは，このようなリスクの程度で測られることに注意する必要があります．