2010 年度後期　統計データ解析Ｂ　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，もとの気温のばらつきに比べて，直線のグラフからみた気温のばらつきのほうがずっと小さくなっていれば，このモデルは気温のばらつきをかなり説明できている，適切なモデルということになります．また，まだ十分適切なモデルでないのならば，データを説明する他の要因（「標高」など）を考えることもできます．この手法は**回帰分析**とよばれ，第 10, 11 回で説明します．

データ解析のキーワード（3）：「くじびき」ー 標本抽出の原理

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．世の中のデータの分布には，「並のデータが多く，極端なデータは少ない」という傾向を持つものが多くあります．例えば，「日本男性の身長の分布」を考えれば，極端に背の高い人や低い人は少なく，並みの人が多いことは，誰でも知っています．

そこで，日本男性全体から何人かの人をくじびきで取り出すと，その人たちは並みの人である確率が大きく，極端に背の高い人や低い人が選ばれる確率は小さいことがわかります．そこで，取り出された人たちの平均身長を求めると，並の人が多く，たとえ極端な人がいても平均することで相殺されますから，「取り出された人たちの平均身長≒日本男性全体の平均身長」である確率が大きくなります．

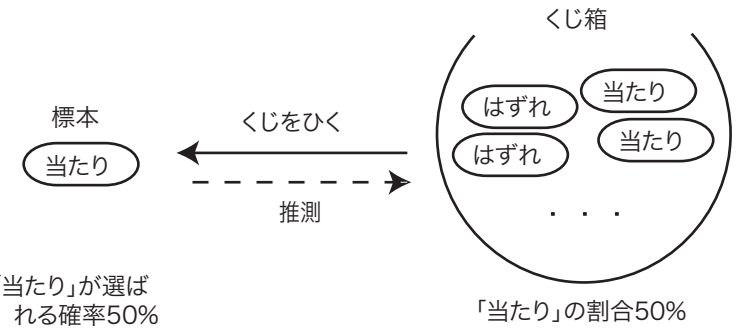

図 1: 標本とくじびき

このときの，「並の人が多く，極端な人は少ない」という傾向を数式で表したものが，確率分布モデルのひとつである**正規分布モデル**です．また，「日本男性の身長の分布」という一見雑多なものが，ある数式で表されるというのは，**中心極限定理**という定理で保証されています．そして，正規分布モデルを用いると，「取り出された人たちの平均身長」と「日本男性全体の平均身長」との食い違いがいくら以内である確率がいくら，といった計算ができます．これを**区間推定**といい，この講義の前半の最後で説明します．また，この「くじびきで取り出す」という操作を**無作為標本抽出**といい，取り出されたデータを**標本**といいます．

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

前節で，「取り出された人たちの平均身長≒日本男性全体の平均身長」である確率が大きくなります，と述べました．ずいぶん持って回った言い方ですが，これはどういう意味でしょうか？

確かに，日本男性全体からは並の人が選ばれる確率が大きいので，たいていは「取り出された人たちの平均身長≒日本男性全体の平均身長」でしょう．しかし，例えば極端に背の高い人ばかりが「たまたま」選ばれる確率は，ゼロではありません．そんなときは，「取り出された人たちの平均身長」は，日本男性の平均身長とは大幅に異なることになってしまいます．

つまり，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」のです．この大外しする確率が**リスク**です．前節では，「取り出された人たちの平均身長」と「日本男性全体の平均身長」との食い違いがいくら以内である確率がいくら，といった計算ができます，と述べましたが，これは「食い違いがある程度以上大きくなるリスクの程度」を求めています．

このように，統計学では，誤差の大小ではなく，大外しをしている確率，つまりリスクの大小を問題にします．「誤差が小さい」とは，間違いの量が常に少ないことですが，「リスクが小さい」とは，間違える確率が小さいことであって，どの程度の大間違いかはまた別の問題です．

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，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．

今日の演習

次の各文は正しいか考えてみてください．（まず直感で答えてください．）

1. 百発百中の大砲一門は，百発一中の大砲百門に匹敵する．（明治の軍人・東郷平八郎の言葉）
2. 国民所得と酒の消費量には相関関係がある．つまり，酒の消費量が増えれば国民所得が増える．
3. ある地域では，女子の出生数が男子の 5 倍に達した．これは異常で，環境ホルモンか何かの影響があるのではないかと疑われる．

```
  ∧∧
≡×××≡
  (  )〜
```

博士，これからいっぱい数式をおぼえなくちゃいけないんですか... ？

別に式を覚える必要なんかない．仕事で統計を使うときに，本をみながらやったらあかん，なんてことはないやろ？ それよりも，統計学というのが何をすることなのか，どういう考え方をするのかをちゃんと理解して，それからどうやって計算するかを数式で考えればええんや．

```
 _∧◆∧_
Ξ o- o Ξ
  (   )〜
```