2008 年度後期　情報統計学　第 1 回

「可能性」を考える ー イントロダクション

統計学とは「合理性のある偏見」

しばらく前に，ハーバード大学の学長が「女性は理系に向かない」などと発言して物議をかもしました．これは「不合理な偏見」です．それは，「女性は理系に向かない」ことが正しくないから，ではありません．

「女性は理系に向かない傾向がある」ということは，もしかしたらあるのかもしれません．しかし，理系に向く向かないというのは，本来個人の問題です．仮に「女性は理系に向かない傾向がある」ことがわかったとしましょう．だからといって，いま目の前にいる人物が理系に向くかどうかを，「男性」や「女性」という集団についての傾向を用いて判断するのは不合理です．なぜならば，その人個人の能力は，性別という情報から不確実な推測をしなくても，試験をするほうがより確実にわかるからです．

しかし，世の中の事柄は，上のように個人個人に適切な対応ができるものばかりではありません．時には，集団全体について，ひとつの対応を決めなければならない場合が多々あります．例えば，「服を作って売る」ということを考えてみましょう．オーダーメイドの洋服ならば，客ひとりひとりの好みに合わせた服を作ればよいわけですが，大量生産の既製服をたくさん売るには，「集団全体の好みの傾向」，いわば「平均的好み」を知る必要があります．

このとき，統計学をもちいた調査によって「関西は関東よりも派手好きの傾向がある」ということがわかったとしましょう．実際には，関西にも地味な人はいるでしょうし，関東にも派手な人はたくさんいることでしょう．だから，この「関西は関東よりも派手好きの傾向がある」という結論は，上の「女性は理系に向かない傾向がある」というのと本質的には変わらず，「偏見」の一種です．しかし，これは「合理的な偏見」であり，既製服を売る業者としては，関西には関東よりも派手な商品を多くするのはもっともな戦略です[1]．

このように，集団に対して「合理的な偏見」を導くのに使われるのが統計学です．この講義では，データの「さまざまな可能性」を考えて，集団の一部のデータを調べて集団全体の傾向を知る**統計的推測**を説明します．

今日のイントロダクションでは，このような統計学の考え方を，**「分布」，「モデル」，「リスク」**というキーワードで説明します．

キーワード（1）：「分布」ー 統計学が扱うもの

学校で習うことには，ひとつの問いに対して，ひとつの結果がはっきり決まることが多いものです．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

[1] 関西で派手な服の品揃えを良くするのは「合理的な偏見」ですが，「関西のおばちゃんが皆ヒョウ柄の服を着ている」と思うのは「不合理な偏見」であり，そのように思わせるテレビ番組は，間違った偏見を助長しているといえるでしょう．

- 「日本人男性の身長は？」「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，「日本人男性の身長」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．

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．

そのために人がいままでやってきたのは，おおまかな「仕組み」を仮定して，人間が理解できる言葉や数式で記述しておき，それを使って現象を説明する，という方法です．このように仕組みを表したものを**モデル**といいます．

例えば，上で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．

統計学でも，同じような考え方を用います．ただし，手元にあるデータだけを調べても，いま調べたデータのことしかわかりません．すなわち，日本人男性の身長の分布の例でいえば，「いま調べた」日本人男性の身長の分布を観察しただけにすぎず，他の日本人男性のことは何も言っていません．

これが「日本人男性**全体**の身長の分布」という問題になると，数千万人の人が対象ですから，調べるだけでも大変で，観察すら簡単にはできません．そこで，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．統計的推測でも，モデルの考え方を用います．ここで用いるモデルは，**確率分布モデル**というものです．

例えば，「日本人男性全体の身長の分布」を考えたとき，並の背の人が多く，背のとても高い人やとても低い人は少ない，ということが，経験的にわかります．身長のデータだけでなく，世の中の分布には，「並のものが多く，極端なものは少ない」というものが多いことが知られています．

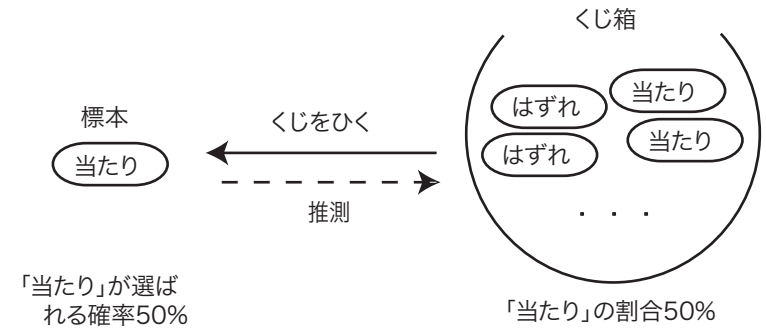

図 1: 標本とくじびき

そこで，すべての日本人男性から，何人かの人をくじびきでとりだして，その人たちの身長を測ったとします．すると，「並の人が多く，極端な人は少ない」のであれば，とりだされた人たちは「並の人」である確率が大きいことになります．そうすると，その人たちの平均は，並の人の平均で，つまり日本人男性全体の平均に近いものである確率が大きい，ということになります．

このとき，「並の人が多く，極端な人は少ない」という分布の「型」を，ある数式，すなわち確率分布モデルでと，くじびきで選ばれた人の身長の平均が，日本人男性全体の平均に近いものである確率を，計算することができるのです．

キーワード（3）:「リスク」ー 間違いの量ではなく，確率

先ほど述べた，「日本人男性全体の平均に近いものである確率」について，もう少し考えてみましょう．

くじびきで選ばれた人たちの平均は，全体の平均に近いものである確率が大きい，と上で述べました．しかし，あくまで確率が大きいというだけです．たまたまくじびきで背の高い人ばかりが選ばれてしまい，選ばれた人たちの平均が，分布全体の平均からはかけ離れたものになる可能性も，ないとはいえません．その確率は小さいですが，もしもこうなったら，そのときの推測は失敗です．

このように，統計的推測では，推測の誤差の大小を問題にするのではなく，間違える確率の大小を問題にします．この確率を，**リスク**といいます．「誤差が小さい」ことは，間違いの量が常に少ないことを意味しますが，「リスクが小さい」ことは，間違える確率が小さいのであって，間違えたときの誤差が小さいことではありません．

このような違いを気に留めていただいて，この講義を聴いてもらえれば幸いです．

くじびきと確率 ー 仮説検定

前節で述べた確率を計算する方法を，この講義の前半では，下の例を使って説明してゆきます．

> 「半分の確率であたる」と店のおじさんが言っているくじがあるとしましょう．ところが，あなたがこのくじを 10 回引いても，1 回もあたりませんでした．
>
> おじさんは「運が悪かったねー」と言っていますが，あなたはどうも納得がいきません．「おじさんの言ってる『半分の確率であたる』なんてウソじゃないの？」と思います．さて，おじさんかあなたか，どちらが正しいでしょうか？

おじさんの言っていることが正しいかどうかは，くじ箱を開けて中のくじを全部調べれば，確実にわかります．もちろん，そんなことはふつうはできません．しかし，そのようにして調べない限り，おじ

さんがウソをついているのか，それともあなたの運がものすごく悪いのか，結論は出ません．そこで，次のように考えてみます．

おじさんの説では，1 回のくじびきでは，あたりもはずれも確率は 1/2 で同じだと言っています．ならば，「10 回ひいて 1 回も当たらない」確率は $(1/2)^{10}$ すなわち 1/1024 ということになります．つまり，おじさんが言うように「半分の確率で当たる」であるとすれば，「10 回ひいて 1 回も当たらない」という結果になる確率は 1/1024 ということになります．

確率とは，「すべての可能性のうち，どの結果になりやすいか」の度合いを表すものです．ということは，「おじさんの説を正しいと受け入れる」ことは，「10 回のくじびきの結果のすべての可能性のうち，1/1024 という小さな確率でしか起きないことが，たまたま今，目の前で起きている」と考えていることになります．そんなムリのある考えを受け入れるよりも，「『半分の確率で当たる』というおじさんの言い分のほうが間違っている」と考えるほうが自然ではないでしょうか？ これが，統計的推測の手法の 1 つである**仮説検定**の考え方です．