2006 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．

「可能性」を考える ー 統計的推測

いま，あなたが福引きのくじをひくと，めでたく当たりが出ました．

しかし，当然のことですが，くじ引きというのはいつも当たりが出るわけではありません．くじびきの結果は，偶然に左右されます．このような「偶然によって結果が決まる現象」のことを，**ランダム現象**といいます．ランダム現象を取り扱うには，「いま得られている結果のほかに，どんな結果が起きる可能性があるか」を考え，さらに「すべての可能な結果のうち，どの結果になりやすいか」を考えます．「どの結果になりやすいか」を数値で表したものが**確率**です．

この考えをもとに，次の例を考えてみましょう．

> 「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．

[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．

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 回のくじ引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した．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？

上で述べたように，くじ箱を開けないかぎり，本当のところはわかりません．ですから，これはおじさんの言い分に対する，「おじさん，どうせウソついてるんじゃないの？」という偏見です．しかし，合理的根拠にもとづく偏見です．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．

「モデル」の考え方

では，この問題が「このくじを 10 回ひいても 1 回もあたらなかった」ではなく，「50 回ひいて 17 回しかあたらなかった」だったとしたらどうでしょうか？

こうなると，上のように簡単には計算できなくなります．それに，そもそも，上の「$(1/2)^{10}$ 」という計算だって，1/2 を 10 回かければよいのはなぜなのでしょうか．

それは，「各回のくじびきで，当たる確率は一定」「ある回のくじびきの結果が，別の回の結果に影響しない（独立）」などと考えているからです．これらのことは，けっして当たり前ではないにもかかわらず，正しいと仮定しています．このような仮定をすることで，上の確率の計算が可能になります．

統計学では，このようなランダム現象を扱います．上で述べたような仮定は，偶然によって決まるランダム現象の結果が，「どのような」偶然によって決まるのかを仮定したもので，確率分布モデルとよびます．

統計学は「誤差」ではなく「リスク」

このくじびきの例における統計的推測の結論は，今ひいた 10 本なり 50 本なりのくじだけを調べて，くじ箱全体を推測するのですから，確実な答えではなく，ある程度間違っている可能性があります．このことを，「統計的推測は，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？

さきほど，

> おじさんが言っている「このくじが 50%の確率で当たる」というのが正しいとすると，「10 回ひいて 1 回も当たらない」という結果になる確率は $(1/2)^{10}$ でしかありません．こんな小さな確率でしか起こらないことが現実に起こっていると考えるよりも，「『半分の確率であたる』というおじさんの説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？

と述べました．しかし，おじさんが言っていることが正しいとしたとき，「10 回ひいて 1 回も当たらない」確率は非常に小さいのであって，ゼロではありません．ですから，10 回ひいて 1 回も当たらなかったからといって，おじさんに「あなたはウソツキだ」と言い放ってしまっても，実は自分が運が悪いだけで，おじさんにとってはとんでもない濡れ衣であることもあるわけです．

このような「大外し」をする確率は，上で計算したとおり非常に小さいものです．ですが，問題なのは，今回計算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．それが「確率」というものの本質です．「間違える確率が小さい」とは「計算をする機会が何度もあるとすれば，そのうち実際に間違える回数は少ない」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．

このような「大失敗をする危険」をリスクといいます．統計的推測の確かさはリスクの程度で測られますが，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，「99%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「99%」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．

データからの推測と「くじびき」

統計的推測の問題には，「データの一部を調べて全体を推測する」というものがあります．例えば，

- 今までの出生についての男女の比率を調べた．このデータから，今後 100 人が出生したときに男あるいは女の割合はいくらぐらいになるか，を推測する．
- 日本人男性 100 人の身長を調べた．このデータから，仮に日本人男性全員の身長を測ったとすればその平均は何 cm くらいになるか，を推測する．

このようなデータは，「値が大小さまざまであり，また，すべてのデータを調べることはできない」という性質をもっています．このような「大小さまざまなデータの集まり」を，データの分布といいます．分布の一部のデータだけを調べて分布全体を推測するのを可能にするために，実は「くじびき」と同じ原理が用いられています．

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．

データからの推測もこれと同じです．仮に，日本男性全体の中での身長 170 〜 175cm の人の割合が 20%だとすると，日本男性全体からあるひとりの人を「公正なくじびきで」選んだとき，その人の身長が 170 〜 175cm である確率は 20%です．このようにして一部のデータを選び出すことを無作為抽出と

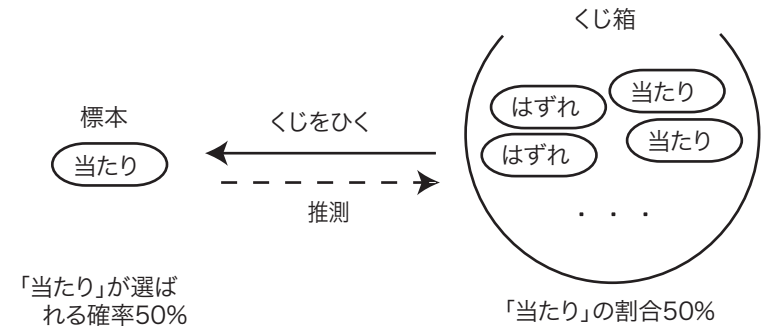

図 1: 標本とくじびき

いい，選ばれたデータを標本といいます．つまり，無作為抽出によってランダム現象を作り出すことで，データの問題がくじびきの問題と同じになります．

問題は，データからの推測では，上の例とは逆で，選ばれた標本だけからデータの集まり全体を推測しなければならないことです．これは，箱の中の当たりくじの数の割合がわからないときに，くじびきの結果をみて当たりの割合を推測することに相当します．

1 回だけくじをひいて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．「当たりが選ばれる確率」は，「いま得られている『当たり』という結果のほかに，どんな結果が起きる可能性があるか」を考え，さらに「すべての可能な結果のうち，どの結果になりやすいか」を考えなければわからないからです．

そこで，くじをひく数をもう少し増やしてみます．くじを何回もひくと，「当たり」「はずれ」という可能な結果のうち，どれがよく起きるかがわかってきます．例えば，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．統計的推測の場合でも，「日本男性の平均身長は 170cm くらいだろう」くらいのことしかわかりません．しかし，日本男性の身長の分布を表すモデルを仮定すると，このモデルを使った計算によって「平均身長が 168cm 〜 173cm 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．