2009 年度前期　データ解析序説　第９回

確からしさを記述する ー 確率

なぜ統計に確率が出てくるのか？

この講義の後半は，**統計的推測**について説明します．この手法が扱うのは，次のような問題です．

日本人男性の身長について，何 cm くらいの人が何%くらいいるのか，平均身長は何 cm か，といった問題に答えるには，どうすればよいでしょうか？

日本男性 全員 の身長を測定すれば，答えは確実にわかるでしょう．このような調査を**全数調査**といい，その代表的なものが，５年に 1 回行なわれる国勢調査です．しかし，国勢調査は，国の莫大な予算と労力，それに「統計法」による強制力を用いて行われている調査です．「日本人男性の平均身長」という程度のことを知るためだけに，そのような予算と労力を使うことは，現実にはできません．

そこで行なわれるのが，「日本人男性の 一部 の身長を調べて，日本人男性全体の身長を測定したときの結果を推測する」という方法です．このとき，調査対象に選ばれた人を**標本**，標本を選んで調査する調査方法を**標本調査**といい，このような「データの一部を調べて全体を推測する」統計学の手法を**統計的推測**といいます．

当然ながら，日本人男性の身長は皆同じではなく，大小さまざまです．すなわち，「日本人男性の身長」は，分布しているデータです．分布しているデータのうち，一部のデータだけを調べて，分布全体を推測すること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，「当たりが選ばれる確率」は，箱の中の当たりくじの数の割合と同じです．

標本調査もこれと同じです．仮に，日本男性のうち，身長 170〜175cm の階級に入る人の割合が 30%だとすると，日本男性からあるひとりの人を「公正なくじびきで」選んだとき，その人が身長 170〜175cm の階級に入る確率は 30%です．このように「公正なくじびきで」標本を選び出すことを**無作為抽出**といいます．つまり，無作為抽出をすることで，統計調査の問題がくじびきの問題と同じになり，「割合」を推測することと「確率」を推測することとが同じになります．

問題は，標本調査における推測では，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．

当然ながら，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．私のような大きな人が標本に選ばれたからといって，「身長 170〜175cm の階級に入る人の割合は 0%」という結論にはならないでしょう．

では，もう少し標本を増やしてみましょう．5 人の標本を選ぶと，うち 2 人が身長 170〜175cm の階級に入る人だったとしましょう．では「身長 170〜175cm の階級に入る人は 40%程度か」といえるでしょうか？

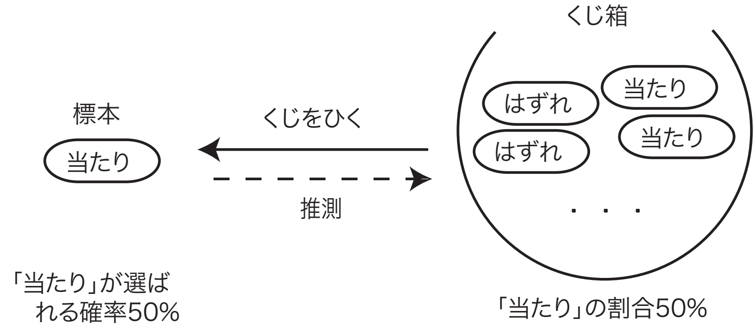

図 1: 標本とくじびき

いや，もしかすると，選ばれた 5 人以外の日本男性が，すべて全員身長 170〜175cm の階級に入るかもしれません．そんなことは常識としては考えにくいですが，調査結果だけからいえば，選ばれた 5 人以外は調べていないのだから，そんなことはないとは言い切れません．

しかし，選ばれた 5 人以外のすべての日本男性が，すべて全員身長 170〜175cm の階級に入るのに，「たまたま」こういう 5 人がくじびきで選ばれる確率は，ほぼゼロといっていいでしょう．そんなときは，たいてい，くじびきで選ばれた 5 人も，全員身長 170〜175cm の階級に入るはずです．

一方，日本男性全体のうち身長 170〜175cm の階級に入る人の割合が 40%のとき，くじびきで 5 人の標本を選ぶと，うち 2 人が身長 170〜175cm の階級に入るというのは，ごく普通におきる，つまり確率が大きいことだというのは，想像がつくと思います．

このように，「標本の中での，身長 170〜175cm の階級に入る人の割合」がわかったとき，そういう標本が選ばれる確率が，「日本男性全体の中での，身長 170〜175cm の階級に入る人の割合」がどのくらいのときにいくらになるか，を調べて，「日本男性全体の中での，身長 170〜175cm の階級に入る人の割合」はいくらぐらいである可能性が何%，と答えるのが，統計的推測の考え方です．

また，このような計算を，各々の階級について行なう，という状況を想定すると，「日本男性全体の平均身長が，いくらからいくらの範囲に入っている確率が 95%」といった推測もできます．この計算を**区間推定**といいます．実際には，各々の階級について別々に計算するのではなく，日本男性全体の身長の分布を表すヒストグラムの形が，ある数式で表されると見なしたうえで計算を行ないます．この数式を**確率分布モデル**といいます．

「可能性」の集合

では，確率とは何なのでしょうか．「確率」という言葉は日常的にもよく使われていますが，ここで，確率の意味をもう一度よく考えてみましょう．

いま，くじをひくと，当たりが出たとします．現実世界では，くじは確かに当たったのであって，それ以外の結果は現れていません．

しかし，われわれは，くじびきとはいつも当たるものではなく，いま現れている「当たり」は偶然による結果だということを知っています．「偶然による」というのは，他の可能性もあった，つまり偶然に

よって他の結果になるかもしれなかった，ということを意味しています．この例の場合ならば，「はずれ」が出るという可能性もあった，ということになります．このような「結果が偶然によって決まる現象」を**ランダム現象**といいます．

統計学の世界では，つねに，この「可能性の集合」を念頭において，考えを進めます．この例の場合ならば，「今は『当たり』という結果が現れたが，『はずれ』が現れる可能性もあった」と考えている，ということです．

そして，さらに「どの結果が，どのくらい現れやすいか」を考えます．これを数字で表したのが**確率**です．「現れやすさ」などというものを，どのように数字で表せばよいのでしょうか．ひとつの考え方は，下のようなものです．

> **ある結果が現れる確率とは，**
> **これからその結果が現れる可能性のある 十分多くの回数の 機会があるとき，**
> **そのうち本当にその結果が現れる回数の割合である．**
>
> **次にその結果が現れる確率とは，**
> **遠い将来までの十分多くの回数の機会を考えて初めて言える「結果の回数の割合」を，**
> **次の１回 の機会にあてはめて述べたものにすぎない．**

例えば，くじ引きを**十分多くの回数**行なうとき，10 回に 3 回の割合で当たりが出るとすれば，「あたりが出る確率」は 0.3 であると考えます．

ただし，次にくじを 1 回ひくとき，当たりが出るかどうかは何とも言えません．ただ，「これからもくじをひきつづけると，長い目で見れば 10 回に 3 回の割合で当たりが出るだろう」という数値で，次の 1 回の機会での当たりくじの「出やすさ」を表現しようというのが，確率の考え方です．

ギャンブルの例で言えば，プロのギャンブラーは日常的に多くの賭けをし，長い目で見た利益を考えていますから，常に確率が大きい方に賭けるほうが有利です．しかし，1 回しか賭けをしない人にとっては，「確率が大きい」ことと「次の賭けで勝てる」こととは直接は結びつかないことになります．

ここでいう「あたりが出る」などの「結果」を，確率論の言葉では**事象**といいます．また，事象が起きる機会，この例ならば「くじを引くこと」を**試行**といいます．また，このような確率の考え方を，**頻度による確率の定義**（統計的確率）といいます．つまり，確率とは「特定の結果がおきる回数の割合」ですから，その値は 0 から 1（0%〜 100%）の範囲になります．

さいころの各目が出る確率はどれも 1/6 か？

高校までの教科書で確率を学ぶ時には，「さいころの各目が出る確率は，いずれも 1/6 である」ということを前提にしていたと思います．

しかし，頻度による確率の定義から考えれば，次にさいころをふったときにある目が出る確率は，十分に多くの回数さいころを振ってみなければわからないことになります．しかも，「十分に多くの回数」振らなければなりませんが，何回なら十分なのでしょうか？ 実は，数学でいう「十分に多く」というのは，「誰も文句を言わないぐらい多く」という意味であって，何回振っても十分ではないのです．

また，さいころを 1 万回ふって，そのうち 1 の目が 1/6 の割合で出たとしても，それはあくまで「過去の実績」であって，その次に 1 万回さいころをふっても，1 の目は 1 回も出ないかもしれません．つまり，頻度による定義では，現実には確率を定めることはできないことになります．

では，なぜ「さいころの各目が出る確率は，いずれも 1/6 である」と言われているのでしょうか？ それは，

1. 各目が同じ確率で出る
2. 各目が出る確率は，いつさいころを振っても同じである

ということを皆が認めているからです．そこで「さいころには全部で 6 種類の目があって，いずれの目も常に同じ確率で出るから，各目が出る確率は 1/6」ということになります．

高校までに習った確率の問題は，このような仮定を認めたうえで，確率すなわち「特定の結果が現れる回数の割合」の問題を，「（さいころの目の種類などの）可能性のある結果の種類の割合」の問題に置き換えたものです．このような確率の考え方を**ラプラスの定義**（数学的確率）といいます．

しかし，このラプラスの定義も，よく考えるとおかしなところがあります．上で「このような仮定を認めれば」と書きましたが，これが認められるかどうかは，さいころを十分な回数振ってみないとわかりません．これでは堂々めぐりです．

つまり，確率の定義にはどのように考えてもあやしいところがあります．確率は，**遠い将来までを長い目で見てはじめて言える「特定の結果が現れる回数の割合」を，次の１回の機会にあてはめて述べたものにすぎません．**また，確率の定義には「十分多くの回数さいころを投げる」という現実には実行不可能な操作や，「各目が同じ確率で出る」という真偽を確かめられない仮定が含まれています．ですから，確率は**測定するものではなく，何らかの仮定をおいて「定義する」もの**なのです．この講義で扱う統計学では，概ね常識的に確率を理解しておけば十分ですが，ここまでに述べた確率の「あやしさ」は承知しておいてもらいたいと思います[1]．

確率のパラドックス

次の問題を考えてみましょう．

> 囚人が 3 人いて，「A, B, C の 3 人のうち，2 人は明日処刑される．あとのひとりは釈放される．誰が処刑されるかは明日朝発表される」と言い渡されています．以下は，処刑前夜の囚人 A と看守の会話です．
>
> > 囚人 A「看守さんは，明日誰が処刑されるかを知っているそうですね．私が明日処刑されるかどうかを，教えてくれとは言いません．ただ，私以外の B, C のうち，どちらが処刑されるかを教えてもらえませんか？」
> > 看守「本当は言ってはいけないのだが... 実は，C は処刑されるのだ．」
> > 囚人 A「よく教えてくださいました．これで少し安心しました．」
> > 看守「なぜだ？」
> > 囚人 A「看守さんの話を聞く前は，私が処刑される確率は 2/3 でした．ところが，看守さんに C が処刑されると教えていただきました．ということは，あとは私と B の２人にひとりが処刑されるので，私が処刑される確率は 1/2 に減りました．」
>
> さて，なにかおかしいと思いませんか？

[1]現代の数学では，確率は現実の問題から離れて，集合を測る尺度（測度）のひとつとしてとらえられています．

この問題に答えるポイントは，**「『C が処刑される』という情報は，A が処刑されるかどうかについての手がかりになるか？」**ということです．

実は，看守が同じように「C は処刑される」と答えても，その答えに至る過程によって，「A が処刑される確率」は変わってきます．それはなぜか，以下の例で考えてみましょう．

看守が「C は処刑される」と答えるに至る，以下の３通りの場合を考えます．看守は判決を知っており，その判決は「B は釈放され，A, C が処刑される」であると仮定しましょう．

1. 看守は，「A, B, C のどれについて答えるか」を確率 1/3 ずつで決めるくじをひき，C について答えることになったので，「C は処刑される」と答えた．
2. 看守は，A については何も答えないことに決めていたので，B と C のどちらについて答えるかを確率 1/2 ずつで決めるコイン投げを行ない，C について答えることになったので，「C は処刑される」と答えた．
3. （問題文と同じ例）　看守は，B が釈放されることを答えてしまうと，A が処刑されることを答えるのと同じことになってしまうので，あえて B については触れず，「C は処刑される」と答えた．

これらの３通りの例は，いったいどう違うのでしょうか．

1. の場合，看守のくじの結果は「C について答える」となりましたが，他に「A について答える」「B について答える」という結果になる可能性もありました．つまり，

- C について答える → C は処刑 → A の運命は未定
- B について答える → B は釈放 → A の処刑が確定
- A について答える → A は処刑 → A の処刑が確定

の３通りの可能性のうち，実際に起きたのはいちばん上のできごとですが，下の２通りが起きるかもしれなかったわけです．

ここで，「ある事象が起きる確率」とは，「その事象以外にすべての可能な事象を考えて，それらの事象が起きうる試行を十分に多くの回数行なうとき，その特定の事象が起きるような試行の回数の割合」であることを思い出してください．

1. の場合にも，「看守がくじをひく」ことを，仮に十分に多くの回数行なうとしましょう．このときの「起きうる事象」には，上の３通りの可能性があります．ところが，1. の文では，３通りの可能な事象のうち，いちばん上の「くじの結果，C について答えると決まった」場合しか考えていません．

ということは，「十分に多くの試行」の回数から，下の２通りの事象が起きた試行を除いたうえで，確率を計算する必要があります．全体の試行の回数が変わりますから，A が処刑される確率も，看守がくじをひく前に比べて変化します[2]．

また，2. の場合，「A については何も答えないことに決めていた」のは確かですが，この場合でも

- C について答える → C は処刑 → A の運命は未定

[2] この「全体の試行の回数を変える」計算の方法は，「条件付き確率」の計算とよばれています．

- B について答える → B は釈放 → A の処刑が確定

の 2 つの可能性があり，そのうち上の場合しか考えていません．したがって，上の例と同様に，やはりこのコイン投げは A が処刑される確率に影響しています．

ところが，3. の場合，つまり本文の問題文の例では，「C について答える」ことは初めから決まっており，他の可能性はありません．したがって，看守が「C について答えた」ことがわかっても，上の説明でいう「十分に多くの試行」の回数は変わりません．したがって，A が処刑される確率も変わらないことになります．つまり，この問題の答は「A が処刑される確率は，看守の言葉を聞いたとしても，依然 2/3」ということになります．

このように，実際に看守が「C が処刑される」と答えたのは同じでも，「他にどのような可能性があったのか」が，確率の計算に影響するのです．このことは，看守の答は同じでも，看守の「癖」や「心の中」が問題になる，ということを意味しています．しかし，心の中を客観的に調べることは，通常はできません．本文の問題文では，看守は，A の運命を答えてしまうことにならないように「C が処刑される」と答えたと思われますが，それが本当かどうかは囚人 A にはわかりません．ですから，この問題の答えは，「看守の言うことを信じれば」確率はこうなる，という意味でしかありません．「確率は，測るものではなく，定義するもの」なのです．

上の「看守と囚人」の問題は，「モンティ・ホールのパラドックス (逆説)」として知られているのと同じものです．インターネットで「モンティ・ホール」で検索すると，いろいろな解説が出てきます．講義のウェブサイトの「統計データ・ツールへのリンク」には，この問題をはじめ，ルーレットやさいころにおける確率の問題についての解説記事へのリンクを掲載しています．

付録

ところで，以前受けた質問に，次の問題について尋ねているものがありました．

> ジョーカーを除く 52 枚のトランプを伏せたままよく切り，1 枚を伏せたまま取り出して置いておく．残り 51 枚をよく切り，伏せたまま 3 枚取り出して表を見たところ，すべてダイヤであった．さきほど伏せたまま取り出した 1 枚がダイヤである確率はいくらか．

52 枚のトランプから 1 枚とりだしたとき，それがダイヤである確率は 1/4 です．問題は，この確率が「3 枚取り出すと，すべてダイヤであった」という事実によって，変化するかどうかです．

問題文では，「残り 51 枚をよく切り，3 枚取り出したところ，すべてダイヤであった」と書いてあります．つまり「すべてダイヤであった」のは偶然による結果であり，ダイヤの枚数は 0 枚，1 枚，2 枚，3 枚である可能性がありました．これらの可能性のうちから「ダイヤ 3 枚」の結果が起きたわけです．

この場合でも，「3 枚取り出す」という試行を十分多くの回数行なうと考えるならば，「ダイヤが 3 枚であった」ということが確定した時点で，「ダイヤが 0 枚あるいは 1 枚，2 枚」であるような試行は，確率の計算から除かなければなりません．ですから，この問題の答えも，1/4 とは異なることになります．