2022 年度春学期　統計学　第 1 回

イントロダクション — 統計的なものの見方・考え方について

このたびは，私の統計学の講義に関心をくださり，ありがとうございます。今日の第 1 回では，「統計的見方」「確率的見方」「統計学と確率」と題して，これから説明していく統計学の考え方を説明します。

統計的見方

感染症と闘う統計学

世界はいまだ，コロナ禍の大混乱から抜け出せていません。

感染症を扱う医学は「公衆衛生学」とよばれます。公衆衛生学が他の医学と違うのは，「目の前のひとりの『人』」を救うのではなく，「社会を構成する『人々』」を救う，という点です。

「人々」の行動を完全にコントロールできるわけではありません。ある人が感染しているかどうかは，検査をしても完全確実にはわかりません。ワクチンを接種しても，絶対に感染しないというわけではありません。そんな条件下で，社会を全体として見たときに，感染の広がりを抑え，医療の逼迫を防がなければなりません。そのために，統計学を使って感染のようすを調べた結果得られたのが，「密閉・密集・密接の『三密』を避けよう」「大人数の会食をやめよう」などとといった指針です。

統計学は「個人」ではなく「全体」

これによって，個人個人の感染が完全に防げるかといえば，そうではありません。たまたま近くにいた人から出たウイルスを吸って，感染してしまうかもしれません。それでも，ひとりの感染者が「平均して」感染させる人数が減れば，感染の広がりは抑えられます。ひとりの感染者が平均してひとり未満にしか感染させなければ，社会全体の感染者は減っていきます。

このように，個人個人ではなく全体のようすを知るというのが，統計学の発想です。感染症の問題に限らず，統計学は「個人」「個別のケース」のことは言わない，言うことができない，というのは重要なことです。

確率的見方

確率とは

「コロナワクチン接種で重篤な副反応が出るのは 10 万人に 1 人の確率だといっても，その副反応が出た人にとっては 100%重篤な事態だ」という話を聞きます。ワクチンの件に限らず，確率にもとづく言説には必ずこういう反応があります。たしかに，「確率が小さい」ことは「事態の深刻さが小さい」こととは関係ありません。くじ引きで，「当たり確率」と「賞金の額」は別の問題なのと同じです。

このあと述べるように，統計学（統計的推測）には確率がつきものなので，この講義でも，後半のはじめとなる第 9 回で確率の説明をします。そこでは「確率とは何か」を，次のように述べています。

> 「くじの当たり確率 0.3」とは，次のような意味です。
>
> - くじを十分多くの回数引くと，そのうち 10 回に 3 回の割合で当たる
> - 十分多くの人がそれぞれ 1 回くじを引くと，その人たちのうち 10 人中 3 人が当たりをひく

どちらでも同じ意味です。どちらにしても，本来「十分多くの回数や人数」について言っていることを，「私ひとりが，1回だけくじを引く」ことに当てはめて，「当たり確率 0.3」と言っているのにすぎません。当たり確率がわかっても，「次の1回」の結果を予想することはできません。

確率は，くじ引きのような「ランダム現象」，つまり「結果に人知のおよばない現象」について持ち出されるものであり，確率を云々しても，人知がおよばないことには変わりはありません。しかし，「人知がおよばないから，わからない」で終わらせるのではなく，「どんな結果になることがどのくらい多いか」を数量的に表すために人間が考えた知恵が，「確率」という考え方です。

期待値

確率の考え方では，さきほど確率とは別だと述べた「事態の深刻さ」や「賞金の額」と確率とを合わせて，さらに「期待値」という概念が出てきます。くじ引きの例でいえば，当たり外れだけではなく当たったときの賞金も考えたときに，

- くじを十分多くの回数引いたときの，1回あたりに得られる賞金の平均
- 十分多くの人がそれぞれ1回くじを引いたとき，ひとりが得られる賞金の平均

を，賞金の「期待値」といいます。第9回の講義では，

> プロのギャンブラーは，「1回限りの大勝負」はしません。どんなギャンブラーでも，次の1回の賭けに勝てるかどうかはわかりません。
>
> ギャンブラーは日頃から多くの回数の賭けをするので，賞金の期待値の大きい賭け方をつねに見抜いて賭けることができれば，個別の賭けでは勝ち負けがあっても，多くの賭けの合計で勝つことができます。

という話もします。

「リスクとメリット」というけれど

ワクチンの話に戻ると，よく「リスクとメリットを考慮して」とよくいわれています。しかし，本当にリスクを考慮できるのは，上記のギャンブラーのように「多くの賭けをして，賞金の合計で勝ち負けを決める」ように，期待値を問題にできる場合です。確率も期待値も「十分多くのくじ引き」の話をしているわけですから，一生に1回しかしないことのリスクを考えるというのは，本当は人間にはむずかしいことで，ある意味「人間の思考の限界に達している」と言わざるをえません[1]。

統計学と確率

さて，データの一部のみを調べてデータ全体のようすを知る「統計的推測」には，「統計的見方」と「確率的見方」が密接に結びついています。

全数調査と標本調査，分布

ずいぶん昔の話ですが，1994 年にノルウェーで開かれた「リレハンメル・オリンピック」の開会式の放送で，アナウンサーが「ノルウェー人は背が高く，平均身長は男性 179cm，女性 170cm だそうです」

[1]なお，ワクチンに関していえば，私は「健康上問題のない人は（わずかな）危険を引き受けるべきだ，それが社会への貢献だ」と考えて，接種を受けました。

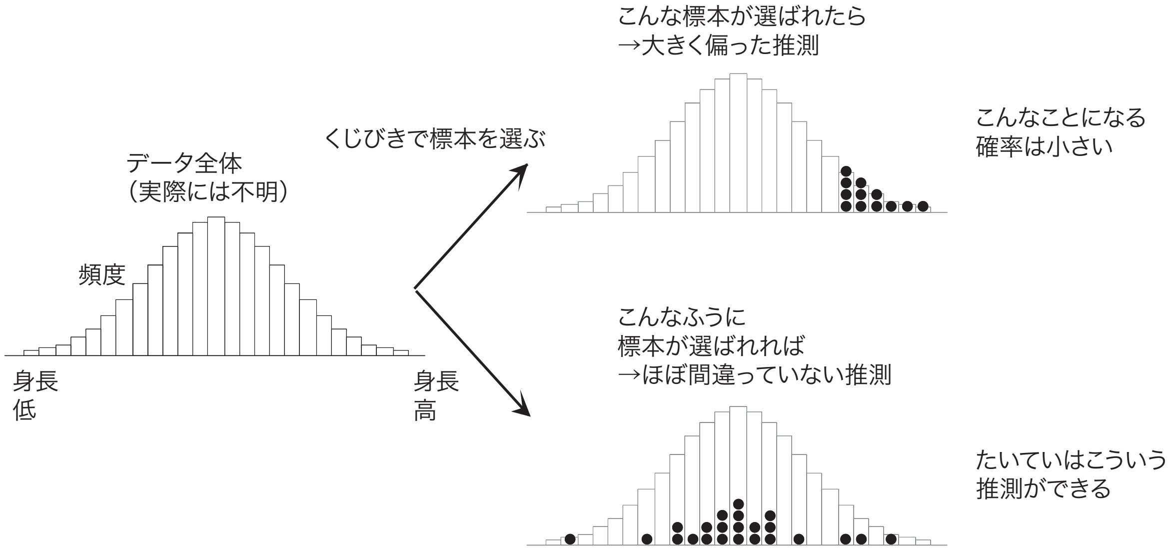

図 1: 統計的推測の原理

という話をしていたのを覚えています。それは，どうやって調べたのでしょうか？

ノルウェー人全員に，ひとりひとり身長計に乗ってもらって調べれば，確実に答えがわかるでしょう。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。平均身長を知るだけのために，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，「ノルウェー人の一部を調べて，ノルウェー人全体を調べ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。

このようなデータは，「値が大小さまざまであり，また，データ全体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「大小さまざまな値をもつデータ」を，データの**分布**といいます。

分布の推測とくじ引き

分布の一部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，「くじびき」と同じ原理が用いられています。

図 1 にある山型のグラフで，ノルウェー人の身長の分布を表しているとします。横軸で身長の高低を表し，ある範囲の身長の人の割合を縦の柱で表します。このようなグラフを**ヒストグラム**といいます。

この分布から，標本を公正なくじびきで選んだとしましょう。「公正なくじびき」とは，どの人も同じ確率で選ばれるようなくじびきです。このような選び方を**無作為抽出**といいます。

このような選び方をするとき，図 1 の右上のように，身長の極端に高い人たちだけが選ばれてしまうことが，ないとはいえません。そうやって選ばれた標本だけを見れば，ノルウェー人は「とてつもなく背の高い人たち」と誤解してしまうかもしれません。

しかし，身長の極端に高い人の割合は小さいので，右上のような偏った選ばれかたをする確率も小さ

いといえます。たいていは，右下のように，並の人は多く，極端な人は少なく選ばれます。このときは，標本だけの平均を計算すれば，それはノルウェー人全体の平均とほぼ同じになるはずです。

つまり，このように無作為抽出された標本を用いれば，ノルウェー人全体の平均身長は，ノルウェー人全員を調べなくても**たいてい，ほぼ**正確にわかります。これが，統計的推測の原理です。

「たいてい」と「ほぼ」

ここで，平均身長が「たいてい，ほぼ」正確にわかる，と述べました。図 1 の右下の場合であっても，無作為抽出で選ばれたのはあくまで一部の人ですから，標本として選ばれた人の平均と，ノルウェー人全員の平均とは，正確に同じなのではなく「ほぼ」同じであるのはしかたありません。

一方,「たいてい」の意味には注意する必要があります。図 1 の右上のような偏った標本が選ばれてしまう確率は，確かに小さいです。しかし，ノルウェー人全体の身長の分布（図中のヒストグラム）は実際には知らないわけですから，もし運悪く偏った標本が選ばれていても，その標本が偏っているのかどうかを知るすべはありません。選ばれた標本から計算された平均を，ノルウェー人の平均身長に「ほぼ」等しいと，信じるしかないのです。

つまり，平均が「たいてい」正確にわかる，というのは，間違った結果を信じて大失敗することもある，ということを意味しています。したがって，統計的推測を行う際には，大失敗の確率を計算しておく必要がます。

前節で述べたように，大失敗の確率というのは,「このような統計的推測を何度も行うとき，そのうちどのくらいの割合の推測が失敗するか」を表すものです。1 回だけ推測するときに，それが失敗するかどうかはわかりません。しかし，確率がわかっていれば，このような統計的推測を何度も行うのなら，そのうちどのくらいの割合で失敗するかも想定できますから，それに対する備えをしておく，すなわち「リスクを考える」ことができます。

統計的推測の方法のひとつで，講義の後半で説明する**区間推定**では,「ノルウェー人全体の平均身長は，179cm 〜 182cm の間にあると推測する。この推測が当たっている確率は 95% である」という答え方をします。身長の幅が「ほぼ」に相当し，当たっている確率が「たいてい」に相当します。

人間の統計学と機械学習の統計学

統計学のことを英語で statistics といい，これはドイツ語の Statistik から来たものですが，これらの言葉は「国家」を意味する state と同語源です。それは，統計学の始まりが，国という集団のようすをデータとして知り，政策に生かすことだったためです。その後も，統計学は「人間が集団の姿を把握し，行動を決める」という目的で発展してきました。

一方，近年急速に発展している「機械学習」は,「人間が」ではなく「機械（コンピュータ）が集団の姿を把握し，行動を決める」という新しい統計学と考えることができます。機械学習の結果が人間に理解できるとは限りません。例えば，将棋では，コンピュータは人間のプロよりも強くなりましたが，コンピュータは各々の手を指す理由を，人間にわかるようには教えてくれません。

この講義では，人間のための,「伝統的な」統計学を扱います。

演習問題

演習問題の回答は，提出する必要はありません。関大 LMS で出題する「小テスト」は，回答を成績評価の対象としますので，回答することをお勧めします。

次の各文は正しいかどうか，理由をつけて答えてください。

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. ある地震予知装置は，芸予地震の直前にも，福岡県西方沖地震の直前にも，警報を発した。この装置の能力は高い。
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，環境からの何かの影響があるのではないかと疑われる。