2013 年度秋学期　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

一人の死は悲劇だが，数百万の死は統計にすぎない。― スターリン

このたびは，私の統計学の講義に関心をいただき，ありがとうございます。この講義では「統計・確率的思考とは何か」「確率を推定するとは」「標本調査と統計的推測とは」の 3 つを理解してもらいたいと思います。

数量的思考，微積分的思考，統計・確率的思考

2011 年の大震災にともなう原発事故で，放射線に関する報道が多数なされています。ところが，中には説明が不正確なため，混乱を招くものもあります。

百年前にハレー彗星が接近した時、尾に青酸が含まれることがわかり、さらにその尾と地球が交差することがわかりました。自転車のチューブに先に空気を貯めて尾が通過する間に吸おうとした人々がいたり，桶の水に顔をつけて息を止める練習をした人々がいたそうです。もちろん、彗星のガスは地球の大気よりはるかに薄いので影響はありませんでした。

いまの私たちは，彼らを笑えない状態になっています。例えば，「沖縄でも 480 万ベクレルの放射性ヨウ素を検出」という記事をみて大騒ぎしている人がいましたが，これは「1 平方キロメートル当たり」です。1 平方メートル当たりなら 4.8 ベクレルで，いっぽう人体は 4000 ベクレル程度の放射性物質を含んでいます。これは，数量的思考が不足している例といえます。

一方，「マイクロシーベルト」と「マイクロシーベルト毎時」がきちんと区別されないために意味が不明になってしまった報道がありました。また，「原発近くで○ミリシーベルト毎時の放射線を検出、これは 1 時間浴び続けるとレントゲン写真△枚分の被曝に相当…」と報じられると、実際にレントゲン写真△枚分の放射線を浴びたと思ってしまう人がいます。実際には○ミリシーベルト毎時の放射線は一瞬出ていただけかもしれません。これは，微積分的思考が不足している例といえます。

これらの思考を正しく理解したとしても，今回の問題を理解するには，さらに「統計・確率的思考」を理解する必要があります。放射線障害とは，放射線のエネルギーによって遺伝子にキズがつき、それが癌などの病気を引き起こすものです。キズがつくかどうかは偶然によるものですし，キズがついても修復されて病気に至らないこともあります。これらの偶然は，おきるかどうかを人が知ることはできず，おきやすさを「確率」という形で理解しているだけです。

だから，ある量の放射線を浴びたら「病気になるのか，安全なのか」と聞かれても答えられません。「健康運が少し下がる」くらいのことしかいえないのです。これは放射線に限らず，煙草の害についても同じです。世の中には，このような「偶然に依存する現象」がたくさんあり，確率でしかとらえられないのだということを理解する必要があります。

確率を推定するには

では，ある量の放射線を浴びたら，病気になる確率がどれだけ大きくなるのかは，わかるのでしょうか。また，わずかの量の放射線でも病気になる確率が大きくなるのなら，「ある量以下の放射線は安全」というのはおかしいのではないでしょうか。

これらに答えるには，確率の推定を行う必要があること，そして，データを集めて（つまり「統計によって」）それを行うことれが，そう簡単ではないことを知る必要があります。確率の推定とは，簡単にいえば「くじびきの結果から，当たり確率を推定する」ことです。そんなことが正確にできるのでしょうか？ この講義の前半では，確率を推定する方法を，下のような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？ これは，統計的推測の手法の 1 つである**仮説検定**の考え方でもあります。

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。

くじびきの結果が偶然によって決まるように，統計学では，結果が偶然のために不確実である現象を扱います。このような現象を**ランダム現象**といいます。上で述べたような仮定は，くじびきのような偶然によって結果が不確実な現象が，「どのように」不確実かを仮定したもので，**確率分布モデル**とよびます。この講義では，代表的な確率分布モデルを紹介し，それを使った確率の計算方法を説明します。

標本調査と統計的推測

上記の「確率を推定する方法」を応用すると，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」を行うことができます。

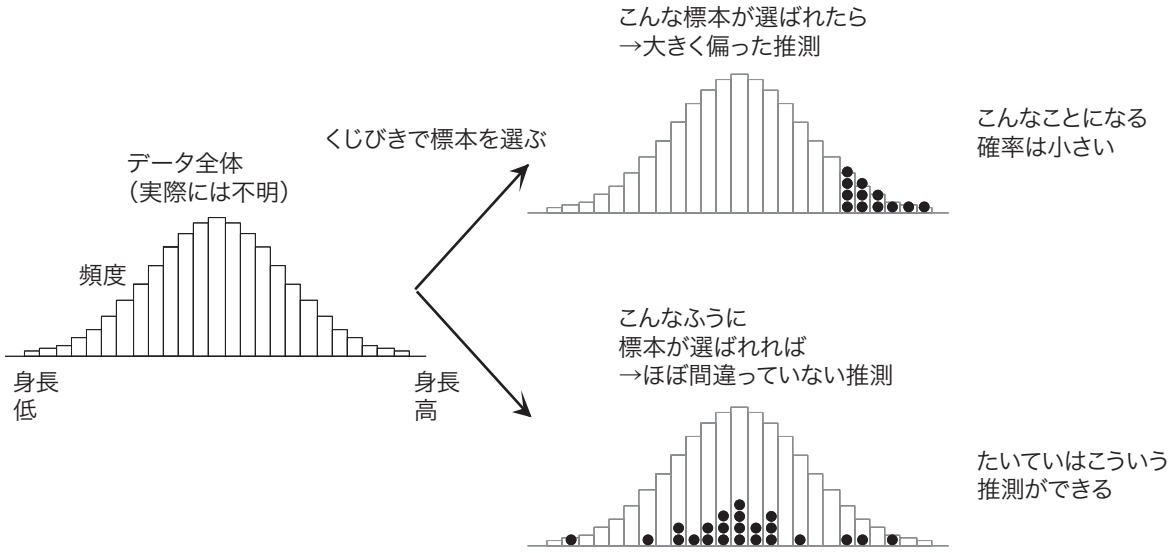

図 1: 統計的推測の原理

だいぶ前の話ですが，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」という話をしていました。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。

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。

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あります。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えば，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することができます。

統計的推測の方法のひとつである**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。

今日の演習

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，環境ホルモンか何かの影響があるのではないかと疑われる。