2008 年度前期　統計データ解析Ａ　第１回

統計的推測とは，何をすることか ー イントロダクション

一人の死は悲劇だが，数百万の死は統計にすぎない．ー 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．

統計学というと，データを集めて表やグラフに整理したり，平均を求めたり... というものを想像される人も多いのではないでしょうか．この講義で説明する統計学には，その続きがあります．それは，集めたデータをもとに，そのデータがどういう仕組みでできあがったのかを調べる記述統計と，確率の考えを用いて，集団のうちの一部のデータを調べて，集団全体の姿を推測する統計的推測です．これらの考え方は，無秩序にばらついているかのような大小さまざまのデータが，なぜ，どのような仕組みで，そのようにばらばらなのかを知りたいという欲求を満たすため，先人達が作り上げてきたものです．

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

- 「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」— 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．

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」— 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．

今日の演習

次の各文は正しいか考えてみてください．

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．