2010 年度後期　統計データ解析Ｂ　第４回

分布の推測とは ー 標本調査と確率分布

統計的推測

ここまでで，データの集まりを度数分布という形式で整理する方法と，さらに平均や分散を計算することで度数分布を要約する方法を説明しました．

しかし，度数分布を求めるには，すべてのデータを調べなければなりません．しかし，ここまでの例で，日本男性の身長の分布といった例をあげてきましたが，すべての日本男性の身長を調べるのは，現実問題として不可能です．

そこで，すべてのデータを調べることがむずかしいとき，その一部のデータを調べて，その結果から度数分布を推測したり，あるいはせめてデータ全体の平均あるいは分散だけでも推測する方法を考えます．

これが**統計的推測**というものです．この手法は「くじびき」の考え方が基本になっています．

無作為抽出

統計的推測では，すべてのデータを調べずに，データの集まり全体のようすを調べようというのですから，調べた結果は間違っている可能性があります．

たとえば，日本男性全体の身長の平均を，１０人だけを調べて，その平均で推測するとしましょう．背の高い人・低い人，いろいろな人を１０人取り出せば，１０人の平均は日本男性全体の平均に近いものになるでしょう．しかし，身長 180cm 以上のひとばかりを取り出してしまったら，「日本男性全体の身長の平均は，185cm ぐらいだろう」という，誤った結論を出してしまうことになります．

もちろん，「わざわざ」背の高い人ばかりを選んで，わざわざ間違った推測を行なう必要はありません．しかし，１０人を取り出すときには，まだ身長を調べていないわけですから，「背の高い人・低い人，いろいろな人」を選ぶこともできません．

そこで，この１０人を「公平なくじびき」で選ぶことにします．「公平なくじびき」とは，「どの人も同じチャンスで選ばれる」というくじです．公平なくじびきで選んだとしても，背の高い人ばかりが選ばれて，誤った結論を出してしまう可能性はあります．しかし，もし日本男性に身長 180cm 以上の人が少ないのなら，１０人選んだときにその人たちが 180cm 以上である可能性は小さいですから，この方法で誤った結論を出す可能性は少ないことになります．

可能性の多少を測るのが，**確率**という考え方です．統計的推測と確率がどのように結びつくのか，次節で説明します．

なお，統計的推測の言葉では，このようなくじびきを**無作為標本抽出**（無作為抽出）といいます．また，「日本男性の身長全体」のような，調べたいデータの集まりを**母集団**，調べるために取り出したデータを**標本**，取り出したデータの数を**標本サイズ**といいます．

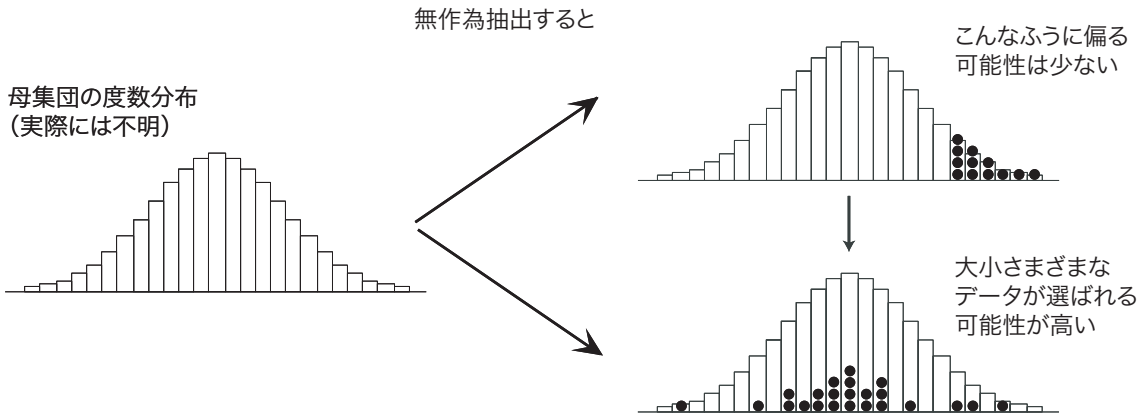

図 1: 無作為抽出の考え方

度数分布と確率分布

くじ箱の中の当たりくじの割合が 20%のとき，当たる確率は 20%である，ということは，当たり前のように思われています．本当でしょうか？

それが本当であるためには，箱の中の特定のくじが選ばれやすかったり，あるいは当たりが出たら次ははずれが出やすい，といったことがなく，「どのくじもつねに同じチャンスで選ばれる」くじでなければなりません．これが「公平なくじびき」で，前節の「無作為抽出」と同じです．

つまり，公平なくじびきでは，

1. どのくじも，同じ確率で選ばれる
2. 各くじが選ばれる確率は，いつくじを選んでも同じである
   （他にどんなくじが選ばれたかには影響されない）

ということになっています（２番目の条件を，各くじは**独立**であるといいます）．このとき，

どのくじも選ばれる確率は同じ

→ ひとつのくじが選ばれる確率は，1/(くじの総数)

→ くじ箱の中の当たりくじが 20%入っているのなら，当たりくじの総数は 20% × (くじの総数)

→ 当たりくじが選ばれる確率は，1/(くじの総数) × 20% × (くじの総数)，すなわち 20%

という常識的な考えがなりたちます．これを，**ラプラスの確率の定義**といいます．

これを，当たりはずれのくじびきではなく，度数分布の場合で考えてみましょう．日本人男性全体の度数分布において，階級値 172.5cm の相対度数が 20%だとしましょう．そうすると，上の原理から，日本人男性全体からあるひとりの人を無作為標本抽出したとき，選ばれた人が階級値 172.5cm の階級に属している確率は 20%です．これは，どの階級についても同じです．つまり，

母集団のある階級の相対度数＝その母集団から無作為抽出された標本が，その階級に属する確率

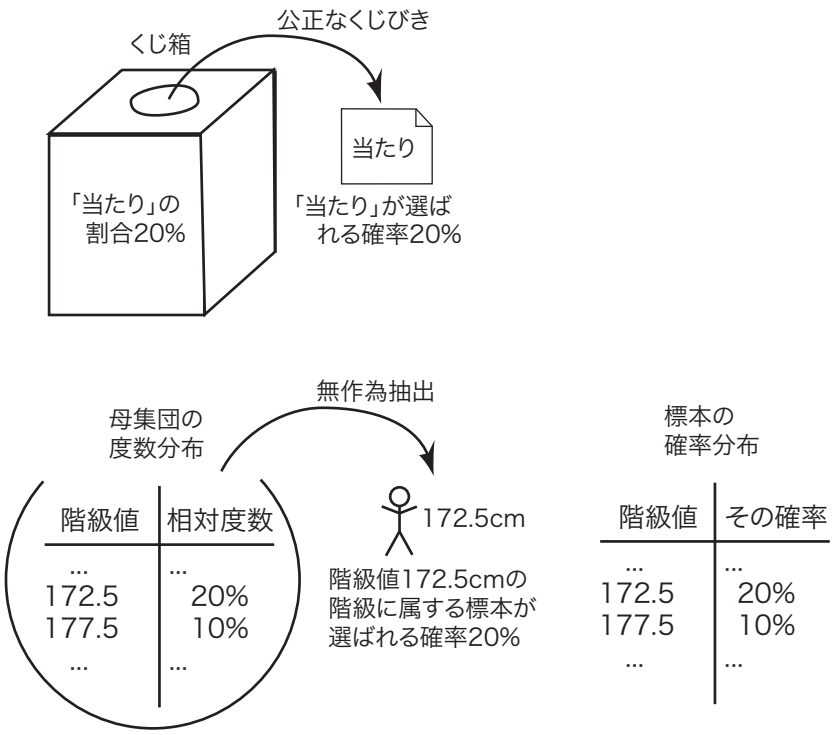

図 2: 度数分布と標本の確率分布

となります．これを度数分布全体でみると，度数分布とまったく同じ「確率の分布」ができます．これを標本の**確率分布**といいます．つまり，

母集団の度数分布（母集団分布） ＝ その母集団から標本を無作為抽出したときの確率分布

となります．

なお，この場合の標本のように，「どんな値かは決まっていないが，とりうる可能性のある値とその値をとる確率，つまり確率分布は決まっている」ような数を，**確率変数**といいます．さらに，確率変数と対応する確率分布の関係を，「（何々という）確率変数は，（これこれという）確率分布に**したがう**」といいます．この表現を使うと，**標本という確率変数は，母集団分布と同じ確率分布にしたがう**，ということになります．

注・復元抽出と非復元抽出

上記のように「母集団のどのデータも同じ確率で取り出され，各データが取り出される確率は他にどんなデータが選ばれたかには影響されない」ことが正確に実現されるには，標本はいつも同じ状態の母集団から取り出されなければなりません．母集団をいつも同じ状態に保つには，取り出した標本を母集団に戻し，それから次の標本を取り出さねばなりません．このような抽出のしかたを**復元抽出**といいます．しかし，実際には取り出した標本を戻さずに次の標本を取り出さざるを得ないことも多く，これを**非復元抽出**といいます．母集団の個体数が標本の数よりも十分に多い場合は，非復元抽出であっても復元抽出とほとんど変わりませんが，母集団の個体数が小さい場合は補正が必要です（この講義では扱いません）．

標本平均と母平均

「無作為抽出」の節で,「日本男性全体の身長の平均を，１０人だけを調べて，その平均で推測するとしましょう」という例をあげました．このような，標本として取り出したデータの平均を，**標本平均**といいます．一方,「日本男性全体の平均」，すなわち母集団全体のデータの平均のほうは，**母平均**といいます．

やはりその節で述べたように，標本平均は，母平均からかけ離れた値になってしまう可能性があり，そのときに標本平均を母平均の推測結果としてしまったら，まちがった推測をしてしまったことになります．

では，標本を無作為抽出した場合は，標本平均は母平均からかけはなれてしまう可能性がどのくらいあるのでしょうか？ これを，図 3 で考えます．この図で，母集団分布の平均（母平均）を μ，母集団分布の分散（母分散）を σ^2 で表しています．この母集団から，n 個のデータからなる標本を取り出したとしましょう．これを $X_1, ..., X_n$ で表します．これらの標本平均が $\bar{X}_n$ です．

図 3 で，破線の上が，現実に抽出された標本を表しています．しかし，標本は無作為抽出されているのですから，いま標本として取り出されているデータは「偶然」取り出されただけで，もしかしたら他のデータが取り出されたかもしれません．そういう「可能性」を，破線の下に描いています．

例えば，X_1 について，他のいろいろな可能性を考えてみましょう．標本は，母集団分布と同じ確率分布にしたがう，と前節で述べました．ということは，その確率分布の平均は，母集団分布の平均と同じで，μ です．この「確率分布の平均」を，**期待値** といいます．また，確率分布の分散も，母集団分布の分散と同じで，σ^2 です．標本 X_1 の期待値は，X_1 はさまざまな値になる可能性がある（確率変数である）が，その値は平均していくらか，ということを表しています．また，分散は，そのさまざまな値が，期待値からみてどのくらいばらついているかをあらわしています．

さて，標本平均 $\bar{X}_n$ は，標本 $X_1, ..., X_n$ がみな確率変数ですから，やはり確率変数で，いろいろな値になる可能性があります[1]．しかし，$X_1, ..., X_n$ の中に極端に大きなあるいは小さな値があっても，平均することで他の値と相殺されますから，標本平均は，ひとつひとつの標本に比べて，極端な値にはなりにくくいつもあまり変わらない値になります．これは,「標本平均の分散は，σ^2 にくらべて小さい」ことを意味しています．

詳しい説明は省略しますが[2]，標本平均の期待値は μ，分散は σ^2/n になります．このことは，

標本サイズが大きければ，標本平均の分散は小さい

→ 標本平均がその期待値から大きくかけはなれた値になることは少ない

→ いま 1 回だけ計算して標本平均が，その期待値から大きくかけはなれた値である可能性は小さい

→ 標本平均の期待値とは母平均であるから，いま計算した標本平均が，母平均から大きくかけはなれた値である可能性は小さく，ほぼ母平均に近い値であると思ってよい

ということを意味しています．したがって，標本平均を計算して，それを母平均の推測結果とするのは，そうおかしなことではない，ということがわかります．

ただ,「いま計算した標本平均が母平均から大きくかけはなれた値である可能性は小さい」とはいって

[1]標本平均のように，標本をまとめて一つの量に要約したものを**統計量**といい，統計量がしたがう確率分布を**標本分布**といいます．

[2]私の講義「情報統計学」(2008 年度後期）第 4 回を参照してください．

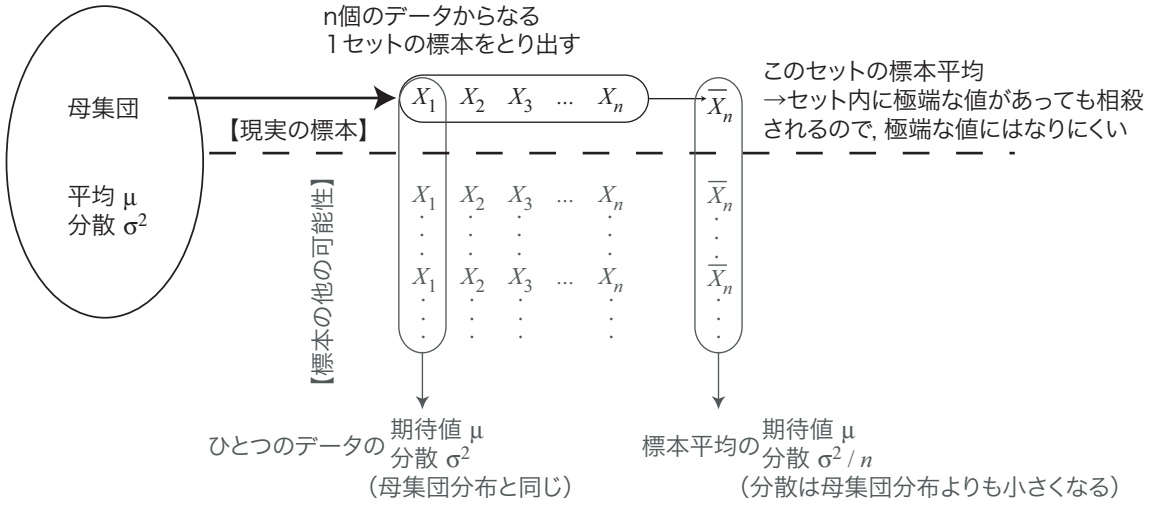

図 3: 標本平均のしたがう確率分布

も，それはゼロではありません．もしかしたら，いま計算した標本平均は，たまたま（非常に運が悪くて）母平均とはまったく違う値で，大きくまちがった推測をしてしまっているかもしれません．

母平均がいくらなのかは，データを全部調べない限りわからないのですから，いま計算した標本平均が母平均に近いかどうかは，わかりません．ですから，**統計的推測は，大きく間違った推測をしてしまう危険を常にはらんでいる**ということになります．ただし，その危険の度合は，間違った推測をする確率という形で，計算することができます．これについては，次回と次々回で説明します．

今日の演習

無作為標本抽出は，考え方は簡単ですが，実行するのはそう簡単ではありません．下の各項は，適切な無作為標本抽出になっているかどうかを理由をつけて答えてください．

1. 学籍番号 000 番から 999 番の 1000 人の学生から 10 人を抽出するため，赤・青・黄の乱数さい（０～９の数字を正２０面体の各面に１つずつ書き込み，０～９の数字のどの数字も２面ずつに書かれているようにしたさいころ）１個ずつを同時にふって，赤のさいの目を１００の位，青の目を１０の位，黄の目を１の位とした数を作ってその番号の学生を選び出す，という作業を１０人を抽出するまで繰り返した．
2. 1 と同様に学籍番号 000 番から 999 番の学生から 10 人を抽出するため，目を閉じて五十音別電話帳を開き，右ページの一番初めに載っている電話番号の末尾の３桁をとってその番号の学生を選び出す，という作業を 10 回繰り返した．
3. 広島市の職業構成を調べるため，選挙人名簿から標本を無作為抽出して調査した．
4. 広島市の繁華街にいる高校生の趣味を調査するため，繁華街でグループで歩いている高校生をみつけ，グループ全員に質問票を渡して回答してもらった．これをいくつかのグループに対して行った．