2026 年度秋学期　統計学　第１０回

分布の推測とは — 標本調査，度数分布と確率分布

統計的推測

ここまでで，データ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，データの中のすべての数値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，データ全体を調べることがむずかしいとき，そのデータの一部を調べて，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

「まんべんなく」抽出することはできるのか

統計的推測では，データ全体を調べていないのに，データ全体のようすを知ろうというのですから，推測し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。

「公平なくじびき」で抽出する

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は，前回説明した「確率」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数値

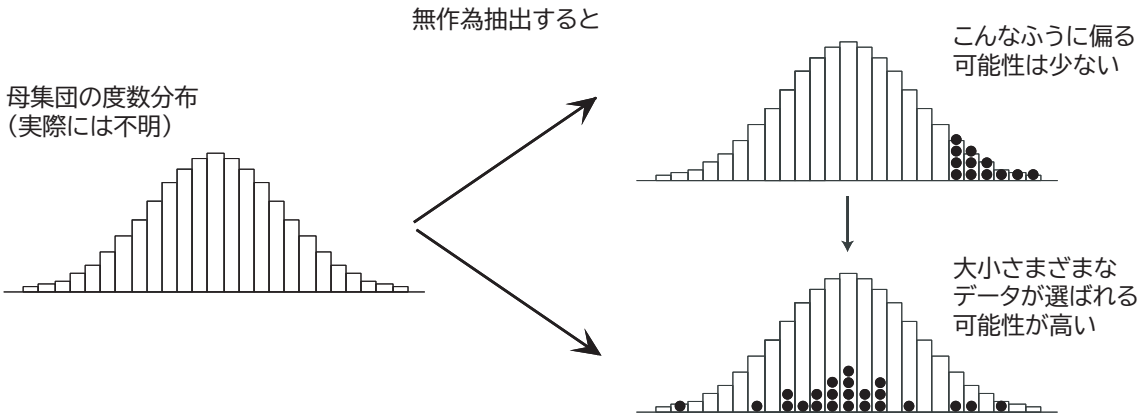

図 1: 無作為抽出の考え方

の集まりを**標本**，取り出した数値の個数を**標本の大きさ**あるいは**標本サイズ**といいます[1]。

度数分布と確率分布

「当たりくじの割合」と「当たる確率」

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，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 番目の条件は，各くじが（前回説明した）「独立」であることを意味しています。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。これは，前回説明した「ラプラスの確率の定義」に相当します。

[1]「標本」という言葉は数値の集まりをさすので，「標本の数」とはいいません。

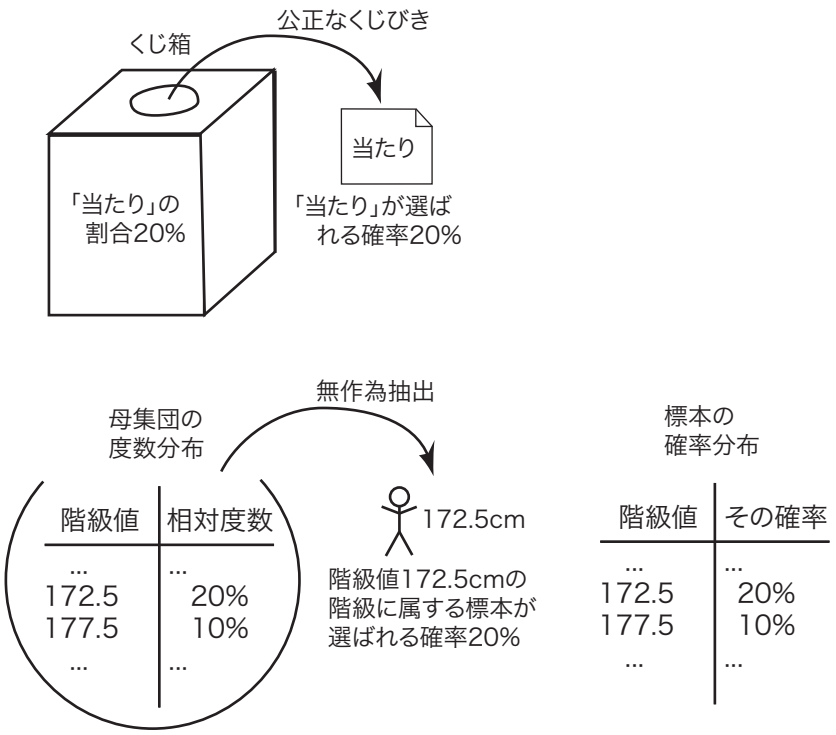

図 2: 度数分布と標本の確率分布

「母集団の度数分布」と「標本の確率分布」

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

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。

確率変数

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[2]。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

[2]中国語では「随機変数」といいます。その意味や英語 random variable からして，私は「随機変数」のほうがよい表現だと思います。

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。

復元抽出と非復元抽出

上記のように「母集団のどの数値も同じ確率で取り出され，各数値が取り出される確率は他にどんな数値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

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，取り出された標本の平均を，**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数値は「偶然」取り出されただけで，もしかしたら他の数値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。

標本の期待値と分散

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。

標本平均の期待値と分散

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[3]。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

[3]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。

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, 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, 互いに独立な確率変数がいくつかあるとき,

- それらの確率変数の和の期待値は，元の各確率変数の期待値の和
- それらの確率変数の和の分散は，元の各確率変数の分散の和

となる性質から来ています[4]。標本 $X_1, ..., X_n$ は（さきほどのくじ引きの考えから）互いに独立で，いずれも期待値は μ，分散は σ^2 ですから，上記の性質から，それらの合計の期待値は $n\mu$ で分散は $n\sigma^2$ となります。

一方，講義第５回の「標準得点」のところで説明したように，確率変数に定数 a をかけると，その期待値は a 倍，分散は a^2 倍になります。標本 $X_1, ..., X_n$ の平均である標本平均 $\bar{X_n}$ は，$X_1, ..., X_n$ の合計を $1/n$ 倍したものですから，$\bar{X_n}$ の期待値は $n\mu$ の $1/n$ 倍で μ，分散は $n\sigma^2$ の $(1/n)^2$ 倍で σ^2/n となります。

標本サイズと母集団サイズ

上で述べたとおり，母分散が σ^2/n のとき，標本平均の分散は，標本サイズを n として，σ^2/n となります。このことは，「標本平均の分散に関係しているの標本サイズであって，母集団の大きさ（母集団サイズ）は関係ない」ことを表しています。つまり，標本を使って推測を行うとき，推測の確かさに影響するのは標本サイズ自身であって，「標本の大きさの，母集団の大きさに対する割合」ではない，ということになります。

「10 人からなる標本」の意味は，母集団サイズが 1,000 人でも 100,000 人でも変わらない，というのは，ちょっと不思議な気がします。これは，先に理想的な無作為抽出として述べた「復元抽出」を考えてみるとわかります。「10 人からなる標本」というと，10 人が一度に母集団から取り出されるような印象を与えます。しかし，理想的な無作為抽出では，母集団からどの人を取り出すときにも母集団分布が変わらないようにしなければなりませんから，1 人ずつを復元抽出することになります。つまり，標本サイズは「一度に取り出される人数」ではなくて「1 人 1 人取り出す回数」のことであり，つねに母集団分布は変わらないので，回数に母集団サイズは関係ない，ということになります。

標本平均は，母平均に「たいてい，ほぼ」近い

さて，ここまでで導かれたことは,

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，「たいてい，ほぼ」母平均に近い値であると思ってよい

[4] この性質を証明するには，多次元確率分布の知識が必要です。私の講義「解析応用」(2013 年度後期) 第１３回を参照してください。また，私の著書「挫折しない統計学入門」では，証明に必要な多次元確率分布の知識を，例を使って説明しています。

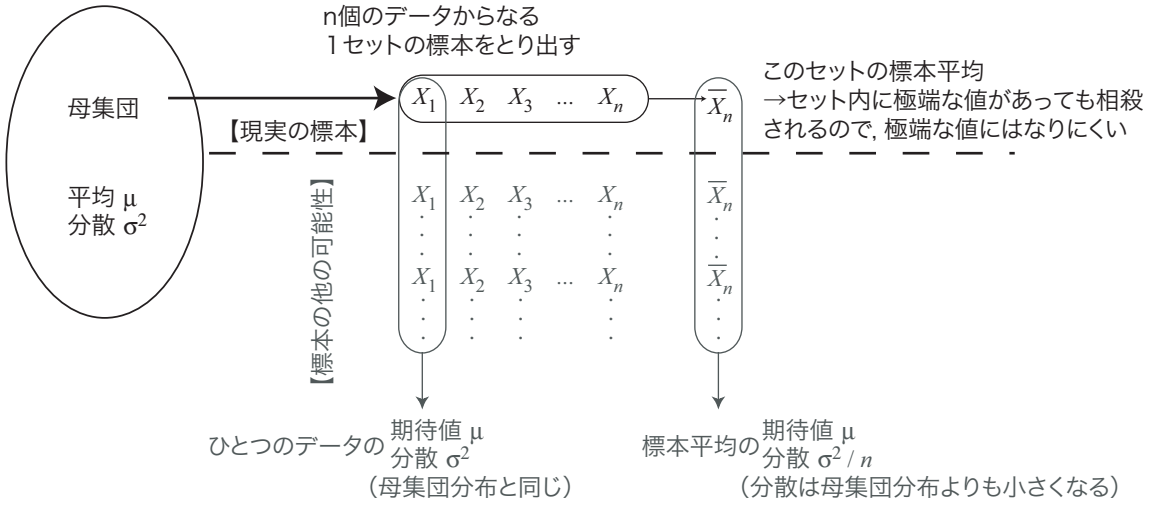

図 3: 標本平均のしたがう確率分布

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，母集団全体を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，次回以降で説明します。

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，正しい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 1 番から 100 番の 100 人の学生から 10 人を抽出するために，新井式回転抽選器に「1」から「100」までの番号をひとつずつ書いた 100 個の玉を入れ，よく混ぜたあと，抽選器を 10 回続けて回して 10 個の玉を取り出し，その番号の学生を抽出した。

2. 学籍番号 000 番から 999 番の 1000 人の学生から 10 人を抽出するために，乱数さいを用いた。乱数さいとは図 4 に示されるもので，正 20 面体の各面に，0 〜 9 の数字が，それぞれ 2 回ずつ刻まれているさいころである。ここでは，赤・青・黄の乱数さい 1 個ずつを同時にふって，赤のさいの目を百の位，青の目を十の位，黄の目を一の位とした数を作ってその番号の学生を選ぶ，という作業を 10 回繰り返した。



図 4: 乱数さい

3. 選挙での出口調査のため，投票を済ませて投票所から出てきた人に，何人かごとに適当な間隔をおいて声をかけ，質問して回答を得た。