2009 年度前期　データ解析序説　第１０回

分布を推測する ー 度数分布と確率分布，確率分布モデル

統計的推測とは，何をすることか

統計的推測とは，ひとことでいうと，調べたい集団のデータ（例えば，日本男性全体の身長のデータ）が，数がものすごく多いなどの理由ですべてのデータを調べることができないときに，限られた数のデータだけを抜き取って調べて，対象のデータ全体の（相対）度数分布を推測しようとするものです．

このようなことが可能なのは，いわゆる「くじびきの原理」によります．つまり，「当たりくじの割合が 50%であるくじ箱から１本くじをひくと，そのくじが当たる確率は 50%」というものです．これを，「度数分布」と「確率」の関係で考えてみましょう．

くじびきと標本調査

もう一度，前回とりあげたさいころの例を考えます．確率の「ラプラスの定義」によれば，

「１，２，３のいずれかの目が出る確率」
＝「『１，２，３，４，５，６の６通り』に対する，『１，２，３の３通り』の割合」
＝３／６

です．もうすこし一般的にいえば，

「ある性質をもつ目が出る確率」
＝「全部の種類の目の数に対する，その性質をもつ目の種類の数の割合」

ということができます．この関係を度数分布にあてはめてみると，

「あるデータが入っている階級の相対度数」
＝「集団全体のデータの個数に対する，その階級に入るデータの個数の割合」
＝「集団全体から１つデータを取り出したとき，
取り出されたデータがその階級に属する確率」

という関係があることがわかります．たとえば，階級値 162.5cm の階級の相対度数は 0.08，すなわち「身長が（概ね）162.5cm の人の割合が 8%」であるならば，確率を使って言うと「身長が（概ね）162.5cm の人が取り出される確率は 0.08」となります．

ただし，この関係がなりたつためには，前回説明した２つの前提が成り立たなければなりません．この場合，前提１，２は，

1. **どのデータも，同じ確率で取り出される**
2. **各データが取り出される確率は，いつデータを取り出しても同じである**
（他にどんなデータが取り出されたかに影響されない）

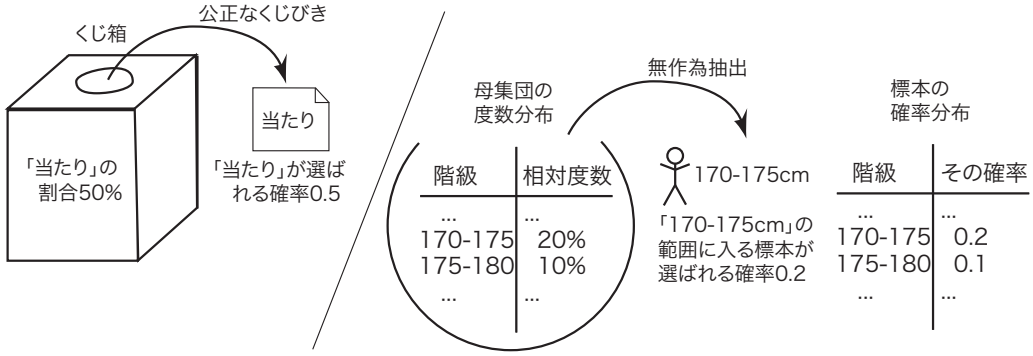

図 1: 度数分布と確率分布

と言いかえることができます．このようなデータの取り出しかたを**無作為抽出**といいます．

以上のように考えるとき，母集団の度数分布では各階級値に相対度数が対応しているのと同様に，各階級値に「無作為抽出によって取り出したデータがその値である確率」が対応していると考えることができます．このように，「大小さまざまな値に対して，その値が取り出される確率が対応している」ものを**確率分布**とよびます．

統計的推測では，未知のデータの分布を，その一部から調べます．このために，集団からデータをいくつか無作為抽出して，上で述べた相対度数分布と確率分布が実は同じものであることを手がかりに，分布の特徴を推測します．このとき，調べたい集団を**母集団**，そこからいくつか無作為抽出されるデータを**標本**といいます．また，この場合の「標本として取り出される値」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，それが取り出された母集団の相対度数分布（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．

度数分布における相対度数は「集団の中での，ある特徴を持つデータの数の割合」であるのに対して，確率変数の確率は「将来何度もデータを抽出したときの，ある特徴を持つデータが出てくる機会の数の割合」であり，本質的に違うものです．しかし，無作為抽出という仮定によって，度数分布と確率分布は同等のものとして扱うことができます．そこで，度数分布の平均・分散を求めるのと同じ方法で，確率変数の平均・分散を求めることができます．ただし，確率変数の平均はとくに**期待値**とよびます．期待値は，確率変数から何度も何度も値を取り出したときの，出てくる値の平均を意味します．また，期待値が等しい確率変数であっても，いつも期待値に近い値ばかりが出る確率変数もあれば，期待値から遠く隔たった値もしばしば出る確率変数もあります．このような「散らばり具合」を表すのが分散です．

ここまでの話は，すなわち「くじびきの原理」で，一見当たり前のことです．ただ，統計的推測がくじびきと異なるのは，統計的推測では「標本を調べて，確率分布を推測する」，すなわち「くじびきの結果を見て，当たる確率を推測する」ことが要求されている，ということです．ふつうのくじ引きでそんなことは不可能であり，統計的推測を実現するにはもう少し工夫が必要です．その技術は，この講義の後半で順に説明してゆきます．

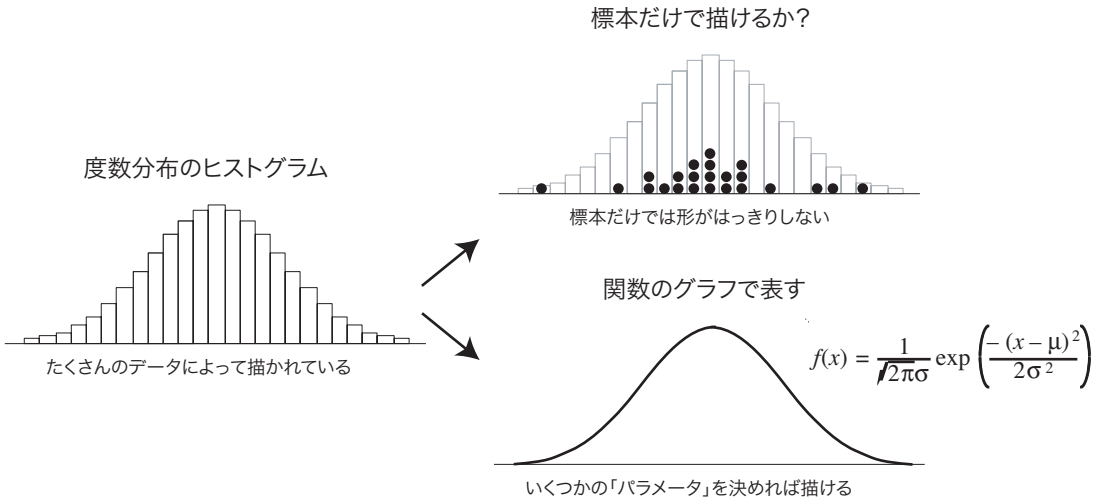

図 2: 確率分布モデル

確率分布モデル

ところで,「母集団分布を推測する」といっても，母集団分布の何を推測すれば「相対度数分布が推測できた」と言えるのでしょうか．標本というわずかな手がかりから，母集団分布の全体を完全に誤りなく推測しようというのは，無理な話です．

そこで用いるのが，**モデル**の考え方です．この考えでは，母集団分布（標本がしたがう確率分布と同じ）が，ある数式で表されるものだと仮定してしまいます．いわば，母集団分布のヒストグラムが，ある関数のグラフになっていると考えるのです．このようにヒストグラムの形を決めてしまえば，あとはそのグラフの縦横の大きさや位置などを，母集団分布に合うように推測すればよいわけです．この数式を**確率分布モデル**といい，母集団のなりたちに合わせていろいろなものが考えられています．また，これから推測する，グラフの大きさや位置を決める数値を**パラメータ**といいます．

連続型確率分布

前節では，確率分布モデルは「数式」で表すと述べました．一方，ここまで度数分布→確率分布→確率変数という順に進めてきた説明では，確率変数は 162.5cm のつぎは 167.5cm というように，「とびとび」の値をとると考えてきました．とびとびの値をとる数式は，複雑です．それよりも，連続なグラフになるような数式のほうがずっと簡単です．では,「とびとびではなく連続的な値をとる」確率変数というのは考えられないでしょうか.

ここで，第２回で説明した「ヒストグラムの柱はなぜ間隔が空いていないのか」を思い出してください．ヒストグラムの横軸の量（点数，など）は，本来とびとびの値に限らずどんな値でもとることができます．ヒストグラムの各柱の面積は，それをいくつかの階級に区切って考えたときに，各階級にあたる区間に入っているデータの個数あるいは割合を表しています.

そこで，ヒストグラムの階級の区切りかたをものすごく細かくしたとしましょう．このような確率分布は，値がとびとびにならない,「ある範囲内のどんな値にでもなることができる」確率分布と考えることができます（図 3）．このような確率分布を**連続型確率分布**といいます．これに対し，ここまでで説明した，確率変数がとびとびの値（例えば，階級値）をとる確率分布を**離散型確率分布**といいます.

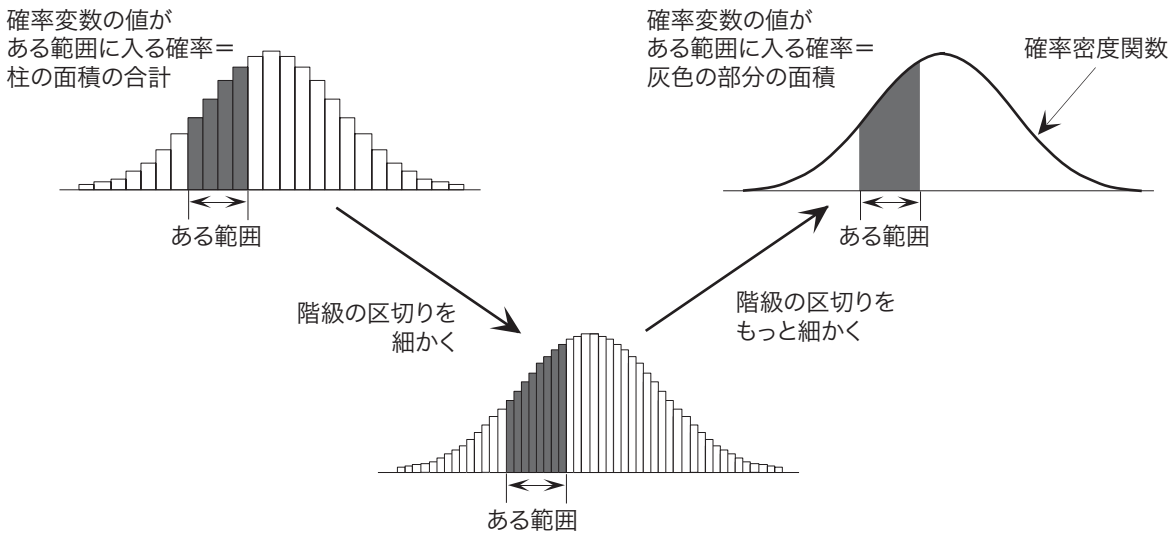

図 3: 連続型確率分布

連続型確率分布では，確率変数が「ある 1 つの値」をとる確率ではなく，「ある範囲の値」をとる確率を考えます．離散型確率分布で確率変数が「ある範囲の値」をとる確率は，確率変数のある範囲内の値に対応する確率を合計したものです．ヒストグラム上でこれを見ると，ある範囲内にある「柱」の面積を合計したものになります（図 3 の左）．「ヒストグラムで度数を表しているのは柱の高さではなく柱の面積」であるからです．

これを，階級の区切りが見えないほど細かくなったヒストグラムで考えると，柱の境目は見えなくなっているので，灰色の部分の面積がそれに相当します（図 3 の右）．この面積は，数学では「『ヒストグラムの上端をつないだグラフで表される関数』の『ある範囲』での積分」といいます．この「ヒストグラムの上端をつないだグラフで表される関数」を**確率密度関数**といいます．

ところで，現実のデータは必ず何桁かの数字で表されるわけですから，どんなに細かく表現してても必ず「デジタル」，すなわち「とびとび（離散的）」です．それなのにわざわざ「連続型」というものを考えるのは，確率分布モデルは数式で表されるからです．数学では，とびとびの値をとる数式よりも，連続なグラフになるような数式のほうがずっと簡単なのです．この講義では積分の計算をすることはありませんが，特定の確率分布モデルでの積分の値を計算してまとめた数表はよく用います．

「確率変数がある範囲の値に入る確率」
＝「確率密度関数のグラフの下の部分のうち，この範囲にあたる部分の面積」
＝「確率密度関数のこの範囲での積分」

という関係は，今後の講義でよく出てきますので，よく理解してください．

確率密度関数は確率変数がとりうる各値の「現れやすさ」を表してはいますが，確率そのものではないことに注意してください．「連続型確率変数がある 1 つの値をとる確率」は，確率密度関数の値ではありません．「連続型確率変数がある 1 つの値をとる確率」は，範囲の幅が 0 ですからその範囲に対応するグラフの下の部分の面積も 0 で，すなわち 0 であることに注意しましょう．また，グラフの下の部分全体の面積は，「確率変数の値が，とりうる値の範囲全体のどこかにある確率」ですから 1 (100%) となります．