2010 年度後期　情報統計学　第３回

確率分布とモーメント，モーメント母関数

標本調査とくじびき

統計的推測の問題には，「データの一部を調べて全体を推測する」というものがあります．例えば，

> 日本人男性 100 人の身長を調べた．このデータから，仮に日本人男性全員の身長を測ったとすればその平均は何 cm くらいになるか，を推測する．

といった問題です．このようなデータは，「値が大小さまざまであり，また，労力やコスト・時間の点で，すべてのデータ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「ある項目についての値が，大小さまざまであるようなデータの集まり」を，データの**分布**といいます．

データの分布全体のようすを，現実的な労力やコスト・時間で推測するために，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**標本調査**といい，調べたいデータの分布を**母集団**といいます．標本調査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．データの分布自体は，確かにすでに存在しているのですから，ランダム現象ではありません．それなのに，なぜランダム現象である「くじびき」と同じ原理が用いられるのでしょうか？

もう一度，簡単な「くじびき」を考えてみましょう．くじ箱の中にくじがたくさん入っているとして，「当たり」が全体のくじの本数のうち 50% ,「はずれ」が 50% であるとします．このくじ箱の中は，「当たり」というデータと「はずれ」というデータの分布であると考えることができます．

このくじ箱から，くじびきで１本くじをひいたとしましょう．このとき，第２回の「確率のラプラスの定義」を考えると，ひかれたくじが「当たり」である確率は 50% ,「はずれ」である確率も 50% であることは容易に想像がつきます．

もっとも，この想像が正しいというためには，「ラプラスの定義」のところで，さいころについて説明した

1. 各目が同じ確率で出る
2. 各目が出る確率は，いつさいころを振っても同じである

という２つの前提が成り立たなければなりません．今回の例の場合，これらの前提は，

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）

と表現することができます．言い換えれば，これらの前提が満たされるくじびきが，「公正なくじびき」です．

簡単に言えば，公正なくじびきでデータを取り出せば，「当たり（はずれ）が選ばれる確率」は，箱の中の当たり（はずれ）くじの数の割合と同じ，ということです．したがって，「箱の中の当たりくじの割合を推測する」ことは，「当たり確率を推測する」ことと同じです．

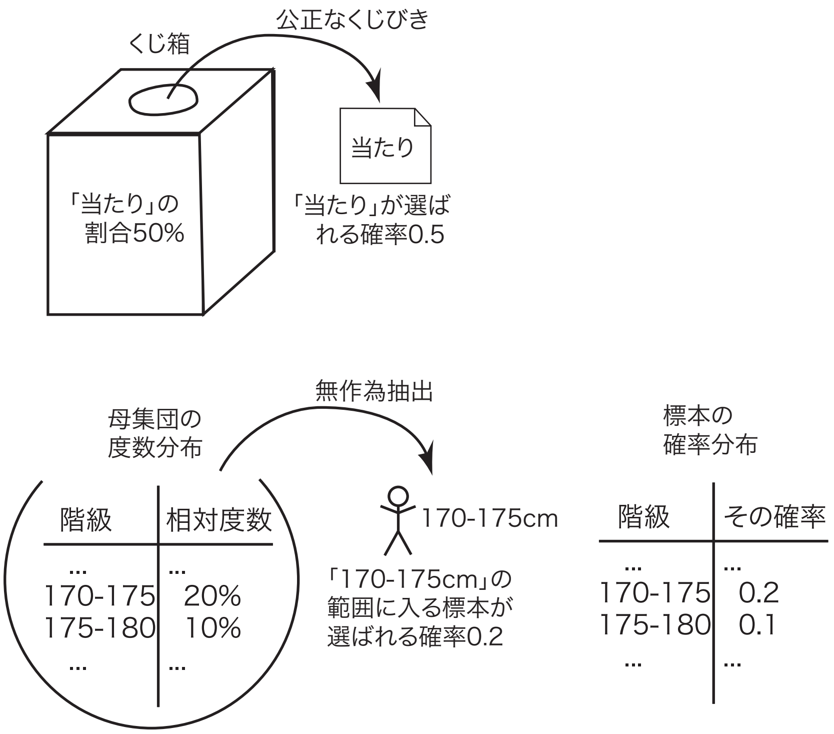

図 1: 度数分布と確率分布

数値のデータの場合も，同様に考えてみます．仮に，日本人男性全体の中での身長 170 〜 175cm の人の割合が 20%だとしましょう．日本人男性全体からあるひとりの人を「公正なくじびきで」選んだとき，その人の身長が 170 〜 175cm である確率は 20%です．つまり，「ある範囲に入るデータの割合を推測」することは，「その範囲のデータがくじびきで選ばれる確率を推測する」ことと同じです．

このように，データの集まりから，一部のデータを公正なくじびきで選び出すことを**無作為抽出**といい，選ばれたデータを**標本**といいます．上で説明した原理は，データの分布の問題をくじびきの問題と同様に考えるため，無作為抽出によってくじびきと同様のランダム現象を作り出している，というわけです．

標本調査と度数分布，確率分布

さて，上では「身長 170 〜 175cm の人の割合」といった例をあげました．この例のように，連続する数値を「170 〜 175cm」「175 〜 180cm」と段階に区切ったものを**階級**といいます．そこで，データが表す量（この場合は身長）を階級に分け，その階級に属するデータ（この場合はその範囲に入る人）の数，あるいはその数の集団全体のデータ数に対する割合を並べると，このデータの分布全体を表すことができます．各階級に属するデータ数を**度数**，割合で表したものを**相対度数**といい，このようにして表されたデータの分布を**度数分布**といいます．

ここで，仮に日本人男性全体の「身長 170 〜 175cm」の階級の相対度数が 20%だとしましょう．このとき，この集団から 1 人の受験者を標本として無作為抽出すると，前節で述べたように，その標本の身長が「170 〜 175cm」である確率は 20%です．これは，どの階級についても同じですから，

ある階級の相対度数＝その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，度数分布に対応した「確率の分布」ができます．これを**確率分布**といいます．つまり，

ある集団の度数分布 ＝ その集団から標本を無作為抽出したときの，標本の確率分布

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，それが取り出された母集団の相対度数分布と同じ確率分布にしたがう**，ということになります．

この後の講義では，この原理を使って，統計的推測のやり方を説明してゆきます．

確率分布のモーメント

データの度数分布がわかっているときに，その平均を求めることを考えます．平均とはデータの合計をデータの個数で割ったものです．一方，ある階級に入っているデータを代表する値（例えば「170〜175cm」の階級なら，172.5cm）を**階級値**といいます．こうすると，相対度数は概ね「その階級値をとるデータが，全体の何%あるか」を表していると考えることができます．そこで，

平均＝（データの合計）／（データ数）
＝（[階級値×度数］の合計）／（データ数）
＝［階級値×（度数／データ数)］の合計
＝［階級値×相対度数］の合計

ですから，**「平均 ＝［階級値×相対度数］の合計」**ということになります．

度数分布における相対度数は，対応する確率分布ではある値をとる確率に相当するので，確率分布について同様に「**［確率変数のとりうる値×その値をとる確率］の合計**」という計算をすると，確率分布から「確率変数の平均」をもとめることができます．これは，さいころを何度もふって出る目のように，確率変数が何度も何度もいろいろな値になるときの，結果の値の平均を意味します．これを確率変数の**期待値**といいます．式で書くと，確率変数 X がある値 x をとる確率を $f(x)$ であらわすとき[1]，確率変数 X の期待値を $E(X)$ は

$$E(X) = \sum_x x f(x) \tag{1}$$

と表されます[2]．$E(X)$ を短く μ と書くこともあります．

これを一般の場合に拡張して，「確率変数 X の関数 $g(X)$」の期待値 $E(g(X))$ を，

$$E(g(X)) = \sum_x g(x) f(x) \tag{2}$$

[1] $f(x)$ を**頻度関数**ということもあります．
[2] $\sum_x$ という記号は，「X がとることのできるすべての x について合計する」と読んでください．

と定義します．このとき，$E(g(X))$ の特殊な場合として，$E(X^k)$ や $E((X-\mu)^k)$ というものを考えます (k は自然数とします)．これらを**確率変数 X の k 次のモーメント（積率）**といいます．とくに，$E(X^k)$ を**原点のまわりのモーメント**といって μ'_k で表し，$E((X-\mu)^k)$ を**期待値のまわりのモーメント**といい μ_k で表します．

モーメントは，確率分布の特徴，すなわち「どんなふうに分布しているか，どんなふうにデータがばらばらか」を表現する量です．いちばん簡単なモーメントは，原点のまわりの 1 次モーメント μ'_1 で，これはすなわち上で述べた期待値です．また，期待値のまわりの 2 次モーメント μ_2 は，データのばらつきの広さを表す量で，**分散**とよばれています[3]．いつも期待値に近い値ばかりになる確率変数もあれば，期待値から遠く隔たった値にもしばしばな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．分散は $V(X)$ あるいは σ^2 という記号で表すこともあります．さらに高次のモーメントも，分布のしかたを表すのに用いられます[4]．

モーメント母関数

前節で示したモーメントの定義とは別に，1 つの関数から各次のモーメントを生成する方法があります．この関数は**モーメント母関数**とよばれ，確率変数 X のモーメント母関数 $M_X(t)$ は

$$M_X(t) = E(e^{tX}) = \sum_x e^{tx} f(x) \tag{3}$$

で定義されます．ここで t は計算の便宜上導入された助変数です．

この関数からどうやってモーメントを導きだすかを見てみましょう．ただし，(3) 式の右辺は収束するとします．

e^{tx} をテイラー展開[5]すると，(3) 式はつぎのように変形できます[6]．

$$\begin{aligned}
M_X(t) &= \sum_x \{1 + tx + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!} + \cdots + \frac{(tx)^k}{k!} + \cdots\} f(x) \\
&= \sum_x 1 f(x) + \sum_x tx f(x) + \sum_x \frac{(tx)^2}{2!} f(x) + \sum_x \frac{(tx)^3}{3!} f(x) + \cdots + \sum_x \frac{(tx)^k}{k!} f(x) + \cdots \\
&= \sum_x f(x) + t\sum_x x f(x) + \frac{t^2}{2!}\sum_x x^2 f(x) + \frac{t^3}{3!}\sum_x x^3 f(x) + \cdots + \frac{t^k}{k!}\sum_x x^k f(x) + \cdots \\
&= 1 + t\mu'_1 + \frac{t^2}{2!}\mu'_2 + \frac{t^3}{3!}\mu'_3 + \cdots + \frac{t^k}{k!}\mu'_k + \cdots
\end{aligned} \tag{4}$$

(4) 式の最後の式を見ると，各項に原点の周りの各次数のモーメントが入っています．つまり，(3) 式の右辺が収束して，モーメント母関数が上のように t のべき級数に展開できるならば，モーメント母関数から各次数のモーメントが得られます．

実際に各次数のモーメントを求めるには，次のようにします．モーメント母関数を t で k 回微分する

---

[3] 分散の平方根を**標準偏差**といいます．
[4] 歪度，尖度といった特徴量があります．今期の「統計データ解析 B」第 3 回で簡単に取り上げています．
[5] 関数を級数に展開する方法です．「微分学」あるいは同等の数学の講義のテキストを参照してください．
[6] $\sum_x f(x) = 1$ となるのは，「すべての x についての確率の和」は「何らかの値をとる確率」と同じで 100%だからです．

と，t について $(k-1)$ 次以下の項は全て消えますから，

$$\frac{d^k}{dt^k}M_X(t) = \mu'_k + t\mu'_{k+1} + \frac{t^2}{2!}\mu'_{k+2} + \cdots + \frac{t^i}{i!}\mu'_{k+i} + \cdots \tag{5}$$

が得られます．ここで $t=0$ とおくと，

$$\frac{d^k}{dt^k}M_X(t)|_{t=0} = \mu'_k \tag{6}$$

となり，k 次のモーメントが得られます．

ある確率分布が数式で与えられたときに，その期待値や分散を求める場合，モーメント母関数を求めてからそれらを導くほうが簡単になることがしばしばあります．それについては，今後の講義でいろいろな例を紹介していく際に説明します．

今日の演習

1. 「コインを1回投げて，表が出れば20円もらい裏が出れば10円払う」という賭けをします．
   (a) この賭け1回あたりにもらえる額は確率変数と考えられます．この確率変数を X で表すとき，X がとりうる値をあげ，確率変数 X がしたがう確率分布を示してください．
   (b) この賭け1回あたりにもらえる額の期待値はいくらですか．
2. 「コインを投げ続けて，はじめて表が出たらやめる．初めて表が出たのが1回目ならば 2^1 円，2回目ならば 2^2 円，…，n 回目ならば 2^n 円，…の賞金がもらえる．」という賭けをします．
   (a) この賭け1回あたりにもらえる額を確率変数 X で表すとき，X がとりうる値をあげ，X がしたがう確率分布を示してください．
   (b) この賭け1回あたりの賞金の期待値はいくらですか．また，この賭けを1回やるのに100円払わなければならないとするとき，あなたはこの賭けに参加しますか？