2011 年度前期　統計データ解析Ａ　第５回

20 本くじをひいて，6 本しか当たらない確率 (1) ー 中心極限定理と正規分布

計算しなければならない確率は

この講義の前半では，「『50%の確率で当たる』というくじを 20 本ひいたところ，6 本しか当たらなかった。『50%の確率で当たる』というのはウソじゃないか？」という問題を通じて，統計的推測を説明しています。前回は，「『50%の確率で当たる』というくじを 20 本ひいたとき，6 本当たる確率」を求めるために，「２項分布モデル」について説明しました。

この問題の考え方は，

> 『50%の確率で当たる』といっているのが仮に正しいとしたときに，『6 本しか当たらない』確率が小さい（つまりそんなことはめったにない）ならば，『50%の確率で当たる』というのは疑わしい，それはウソだ，と推測する

というものです。

ここで，「そんなことはめったにない」の「そんなこと」とは，正確にはどういう意味でしょうか。それは，「20 本ひいて**ちょうど 6 本だけ**当たる」ことでは**ありません。**「20 本ひいて，5 本でも 7 本でもなく，ちょうど 17 本当たる」ことは，たしかにめったにないでしょう。しかし，われわれはそれを問題にしているのではありません。

仮に，20 本のくじをひいて 1 本も当たりが出なければ，「50%の確率で当たる」というのはきわめて疑わしいでしょう。それは，「50%の確率で当たる」はずなのに 20 本中 1 本も当たらない，という確率はきわめて小さいからです。確率がより小さいはずのできごとが起きると，そもそもの「50%の確率で当たる」という前提に対する疑いは，より強くなります。

ということは，「20 本ひいて 6 本**しか**当たらない」ことが「めったにない」，という言い回しは，言外に，5 本しか当たらないことも，4 本しか当たらないことも，。。。，1 本も当たらないことも，当然すべて「めったにない」と考えている，ということを含んでいるはずです。6 本しか当たらないことが不満な人は，5 本しか当たらないことも，当然 1 本も当たらないことも，いうまでもなく不満なのです。

つまり，計算しなければならないのは，「不満をもつような現象」が起きる確率，すなわち「20 本ひいたとき，当たり本数が 6 本以下である」確率です。

この確率は，前回説明した２項分布を使って求めることができます。しかし，この計算は「階乗」の計算が入っていて結構面倒です[1]。それに，本当にもとめなければならないのは，「当選確率 50%のくじを 20 本ひいたとき，当たり本数が 6 本以下である」確率ですから，この計算を，当たる本数が 6 本，5 本，...，0 本のすべての場合について行なわなければなりません。

そこで，この計算を別の方法で簡単に行なうための，**ド・モアブル＝ラプラスの定理**というものが知られています。今回と次回の講義では，この定理と，それを理解するのに必要な**中心極限定理**と**正規分布モデル**について説明します。正規分布モデルは，この後の講義でもたびたび出てくる，非常に重要な考え方です。

[1] 2010 年度後期「情報統計学」第 5 回の講義録に，その式が出ています。演習でも取り扱いました。

中心極限定理と正規分布モデル

世の中には，さまざまなランダム現象があります。それらが「どういう理由で，どのようにランダムか」を数式で表しているのが，前回説明した２項分布モデルをはじめとする確率分布モデルです。しかし，前回のベルヌーイ試行のように，「どういう理由で，どのようにランダムか」が明確にわかる場合というのはそう多くはありません。ところが，世の中の広い範囲のランダム現象によって生じる確率変数を，ある１種類の確率分布モデルで表すことができる，という定理があります。これが**中心極限定理**で，その確率分布モデルを**正規分布モデル**といいます。

中心極限定理とは，簡単に言うと「ある確率変数が，無数の独立な確率変数の合計になっているときは，その確率変数のしたがう確率分布は概ね正規分布モデルであらわせる」ということです。この場合，「無数の独立な確率変数」は，どんなものであってもかまいません[2]。

例えば，物の長さを測定するときの測定誤差は，測定するごとに異なり，確率変数としてとらえられます。しかし，定規の熱による伸び縮み，人間の目の限界，定規を見るときの空気の乱れ，などなど，無数の独立な原因による誤差の合計になっていますから，測定誤差の確率分布は正規分布になります。あるいは，電気回路の雑音は，回路中の金属の原子が熱によって独立に振動し，その合計として現れます。ですから，雑音の瞬間瞬間の強さは正規分布にしたがいます。このように，独立な無数の原因の合計として現れる確率変数はたくさんありますから，正規分布モデルで表せる確率変数は，自然科学，社会科学の分野を問わず，世の中に無数に見つけることができます。

２項分布モデルの説明で「試行回数 n と１回の試行での成功確率 p がわかれば，どの成功回数 x についても『x 回成功する確率』は計算できる。このことを，n と p はパラメータであるという」と述べました。正規分布モデルの場合は，期待値と分散がパラメータです。つまり，期待値と分散がわかれば「確率変数がある値からある値の範囲にある確率が，いくらになるか」という計算をすることができます。しかも，実際にはいちいち計算する必要すらなく，次回の講義で説明するように，すでに計算結果が書いてある数表を使うことで，簡単に確率を知ることができます。「試行回数が n，１回の試行での成功確率が p」の２項分布を $B(n,p)$ と書くように，「期待値が μ，分散が σ^2」の正規分布を $N(\mu,\sigma^2)$ と書きます。

ド・モアブル＝ラプラスの定理

ド・モアブル＝ラプラスの定理は，中心極限定理を使って２項分布を正規分布で近似し，正規分布の数表を使って計算をする方法です。

「１回あたり確率 p で成功する，n 回のベルヌーイ試行」を別の視点から見てみましょう。ちょっと変な感じですが，「１回のベルヌーイ試行で，成功する回数」を考えて，X で表します。当然，X のとりうる値は０回または１回です。１回あたり確率 p で成功しますから，「１回のベルヌーイ試行で，成功する回数」X は２項分布 $B(1,p)$ にしたがいます。

一方，「１回あたり確率 p で成功する n 回のベルヌーイ試行で，成功する回数」を S とすると，S は２項分布 $B(n,p)$ にしたがう，すなわち，成功回数が x である確率は２項分布 $B(n,p)$ で計算できます。しかし，見方を変えれば，この S はさっきの X を n 個合計したものと考えることもできます。

n 個の X は互いに独立ですから，それらの合計である S は，n が大きいときは中心極限定理によって概ね正規分布にしたがいます。一方，S は２項分布 $B(n,p)$ にしたがうのですから，前回の説明で述べたように，S の期待値は np，分散は $np(1-p)$ です。S の分布を，「概ね正規分布」とよぼうが，「２項分

[2]本当は，まったくどんなものでもよいというわけではなく，いろいろ制約がありますが，省略します。くわしくは，2010年度後期「情報統計学」の第 13 回の講義録を参照してください。

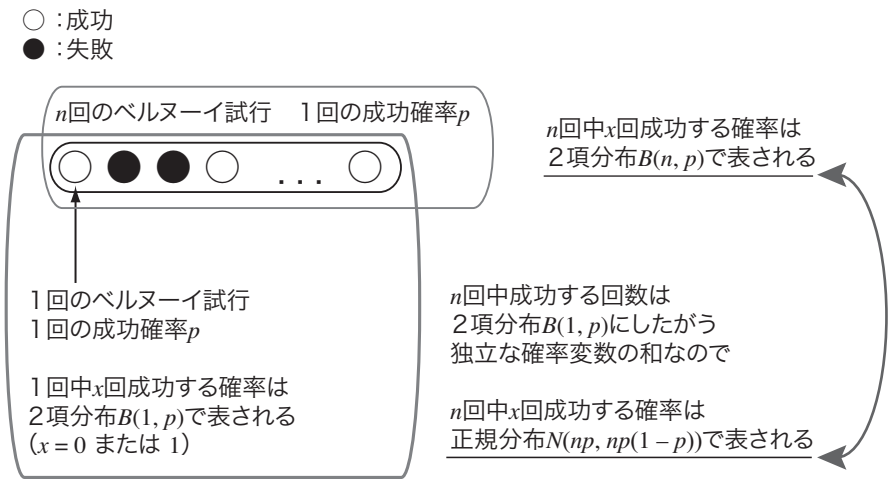

図 1: ド・モアブル＝ラプラスの定理

布」とよぼうが，同じ現象を別の名前でよんでいるだけですから，期待値や分散は同じです。したがって，n が大きいとき，２項分布 $B(n,p)$ は正規分布 $N(np, np(1-p))$ で近似できることがわかります。

ヒストグラム

今後の正規分布モデルの説明では，確率分布を目に見える形に表現した**ヒストグラム**をよく用います。そこで，ここでヒストグラムについて説明しておきます。

例えば，「１回につき 40%の確率で当たるくじを５本ひいたとき，当たる本数」を確率変数 S で表すとき，S は２項分布 $B(5, 0.4)$ にしたがいます。そこで，この確率分布モデルにしたがって，当たり回数が０回から５回までの確率をそれぞれ求めたとします。これらの確率を，横軸を当たる回数，縦軸を当たる確率としてグラフに表すと，図２のようになります。しかし，ヒストグラムはこのようには描きません。

ヒストグラムでは，図３のように，**柱の高さではなく，柱の面積で度数を表現します。**ですから，ヒストグラムの縦軸はとくに意味はなく，柱の幅が変われば，同じ高さでも表す確率は異なります。こういう表しかたをするのは，となりどうしの柱を結合したり分割したりするためです。確率を柱の面積で表しておくと，図４のように，「１回当たる確率」を表す柱と「２回当たる確率」を表す柱をくっつけるだけで，「１回または２回当たる確率」を表すことができます。また，くっつけた柱をならしてしまうと，「１回当たる確率」や「２回当たる確率」はわからなくなりますが，それでも「１回または２回当たる確率」は表されています。逆に，「１回当たる確率」や「２回当たる確率」がわかれば，柱を分割するだけでそれぞれの確率を表すことができます。

ところで，今回の例では「当たる回数」は０，１，２，３，４，５の６通りだけでした。しかし，正規分布モデルで表される例としてあげた「測定誤差」などでは，その値はさまざまな数値になり，６通りに限られる，などということはありません。こういう場合のヒストグラムはどのようになるのでしょうか？ それを考えるのに，柱の結合・分割の考え方が用いられます。これは次回に説明します。

今日の演習

「無数の独立な確率変数」の和になっているため中心極限定理が成り立ち，正規分布で表現できるような確率変数の例をあげ，それがどんな「無数の独立な確率変数」の和になっているかを答えてください。

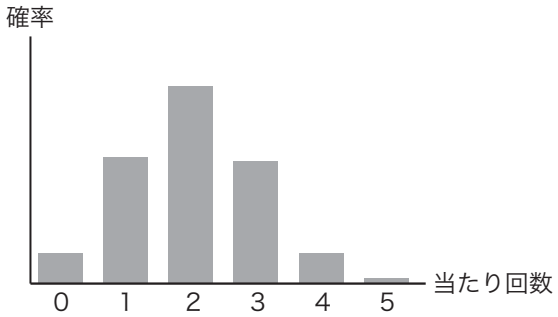

図 2: ヒストグラムはこんなふうには描かない

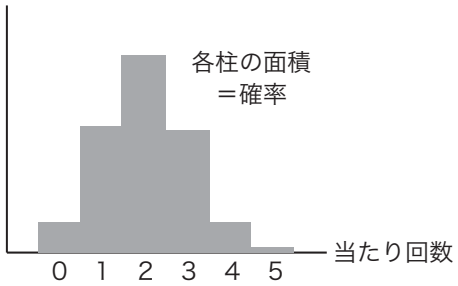

図 3: ヒストグラムはこう描く

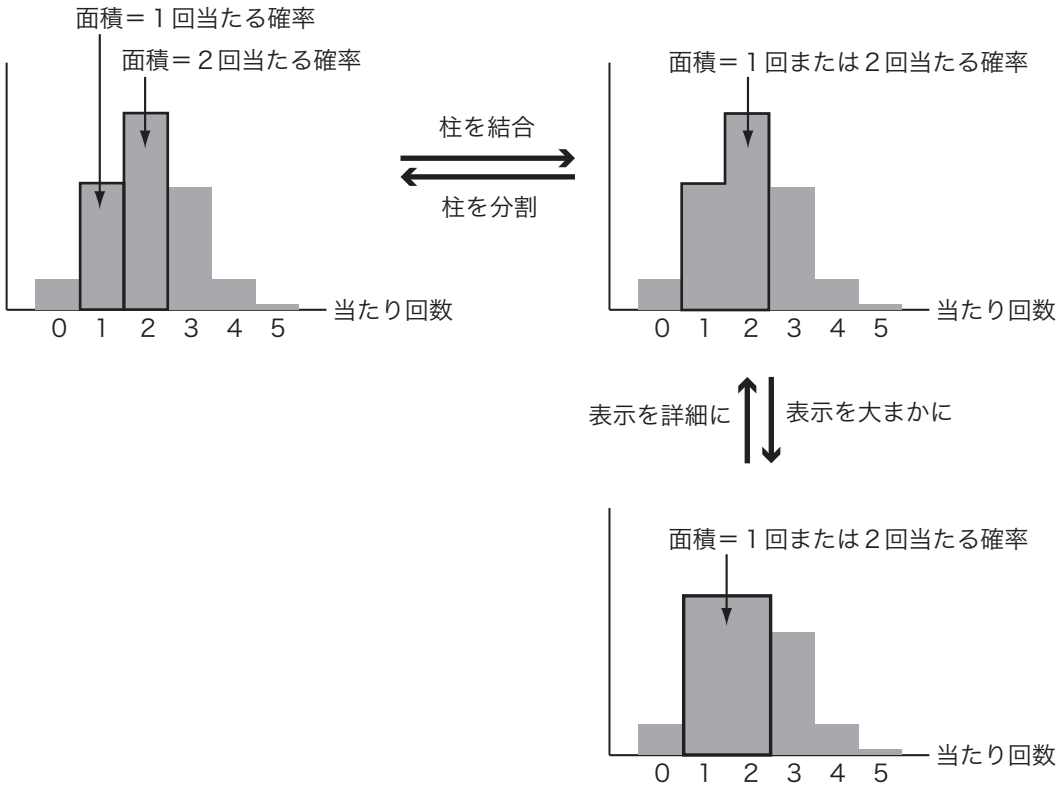

図 4: 確率を柱の面積で描く理由