２０本くじをひいて，６本当たる確率 － ２項分布モデル，期待値と分散

講義第１部で取り扱う問題

この講義の前半では，

> 「50%の確率で当たる」と称するくじを 20 本ひいたところ，6 回しか当たらなかった．「50%の確率で当たる」というのはウソではないかと感じられる．どういう方法を用いれば，「ウソだ」と主張できるのだろうか？

という問題を通じて，統計的推測の基礎を説明します．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 回しか当たらない」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．そこで，この問題に答えるには，まず「『50%の確率で当たる』というくじを 20 本ひいたとき，6 回当たる確率」を求める必要があります．今回は，そのための知識として，「確率分布モデル」の考え方と，そのひとつである「２項分布モデル」を説明します．

確率分布モデルと２項分布

今回の問題を考えるには，「くじを何回かひいたときに，そのうち当たりがある回数出る確率」を求めなければなりません．当然ながら，くじの当たりはずれは，偶然によって決まるランダム現象です．ですから，この確率を求めるには，本当は「どのようなメカニズムによって，くじの当たりはずれが決まるのか？」がわからなければなりません．

しかし，ランダムとか偶然という言葉には「人知のおよばないもの」という意味が含まれているように，「くじの当たりはずれのメカニズム」を正確に知ることは，実際には不可能です．そこで，そのメカニズムを，われわれの納得のゆく範囲で仮定してしまうことにします．このように，「今問題にしているランダム現象は，どのようなメカニズムによって生じているのか」，すなわち「どうランダムか」を仮定し，ある現象が起きる確率を求める数式で表現したものを，確率分布モデルといいます．

今回の問題のような「当たり・はずれを選ぶくじびき」の場合，次のようなことを仮定しても，そう問題はないと思います．細かい点ではいろいろ問題があるのは承知していますが，あまり細かいことを言うと，確率の計算ができなくなり，モデルの意味をなさなくなってしまいます．

- １回のくじびきで，「当たり」または「はずれ」のどちらかが必ず出る．
- 各回のくじびきで，「当たり」「はずれ」それぞれの確率はつねに一定である．
- ある回のくじびきの結果は，他の回の結果には影響を及ぼさない．

これを，くじびきの例から離れて一般的に表すと，

- 「当たり・はずれ」,「成功・失敗」など，２種類の結果のどちらかが必ず起きるような実験（あるいは観測，検査）があり，
- この実験を何度も繰り返すとき，各実験での「成功」「失敗」それぞれの確率はつねに一定で，
- １つの実験の結果が他の実験に影響を及ぼさない（各実験が独立である），

となります．このような仮定ができるような実験をベルヌーイ試行といいます．

ベルヌーイ試行で起こりうる２種類の結果を「成功」,「失敗」で表し，成功する確率が p，失敗する確率が $1 - p$ とします．このとき，「n 回のベルヌーイ試行を行うとき，成功する回数」を S とします．

S は，何かの数字のように見えますが，実際にはその数字，つまり「成功する回数」は，偶然によって決まるもので，はじめからひとつの数字に決まっているものではありません．このような「偶然によって決まる数」を確率変数といいます．

S の値は「偶然によって決まるので，わからない」というだけでは話は進みません．そこで，「S がある値になる確率」を，いろいろな値について求めることを考えます．これを，確率変数 S のしたがう確率分布といいます．上のベルヌーイ試行について，「n 回の試行を行うとき，x 回成功する確率」— これを $P(S = x)$ と書きます — を，$x = 0$ から $x = n$ までのすべての可能な x について求めれば，S のしたがう確率分布を表したことになります．

$P(S = x)$ が実際にどういう式で表されるかは，この講義では省略します[1]．重要なのは，S のしたがう確率分布が，ベルヌーイ試行の仮定から導かれた式で表されるということで，これは確率分布モデルのひとつです．この確率分布モデルには２項分布という名前があり，$B(n, p)$ という記号で書くこともあります．ここで，わざわざ「(n, p)」と書いてあるのは，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，という意味です．このことを，n と p はパラメータであるといいます．

期待値と分散

ここまでで述べたように，「n 回の試行のうち，成功する回数」S は，偶然によって決まる確率変数で，ひとつの決まった数ではありません．仮に，いま n 回の試行をしてある回数成功したとしても，それは，さまざまな可能性のうちのひとつの結果でしかありません．

そこで，「n 回の試行」を「１セット」という単位で表すことにし，「試行結果のあらゆる可能性を考えたときの，１セットあたりの成功回数の平均」を考えます．

図１のように，「１セットの試行」の現実の結果以外に，十分多くの「可能性」を考えたとしましょう．これは，「十分多くのセットの，試行を行なった」と，仮に考えてみることに相当します．すると，「１セットのうち x 回成功する確率」は，これらの可能性のうち「１セットあたり x 回成功した」というセットの数の割合になります．

[1] 総合科学部専門科目「情報統計学」(2006 年度後期）第５回のプリントに掲載されています．講義のサイトからリンクされていますから参照してください．

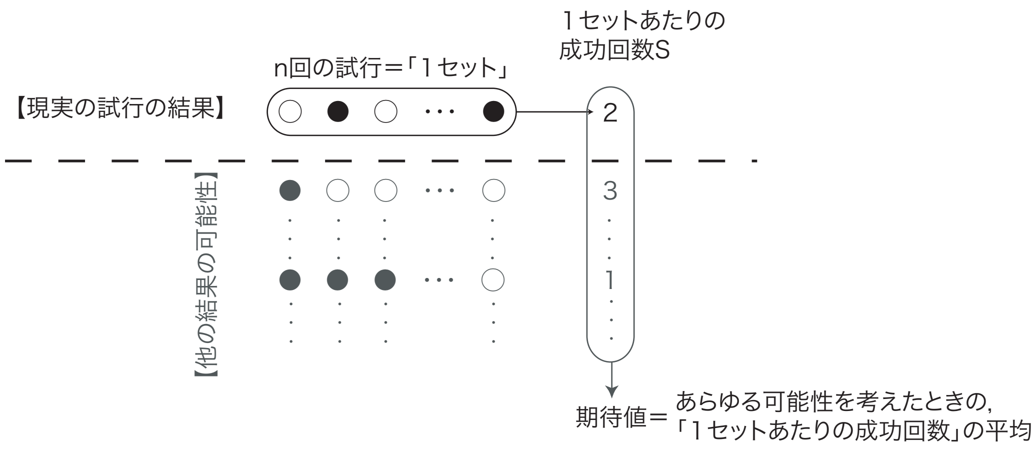

図 1: 確率変数の期待値

したがって，「１セットあたりの成功回数」の平均は，

$$
\begin{aligned}
&\text{「１セットあたりの成功回数」の平均}\\
=\ &\frac{\text{「１セットあたりの成功回数」の合計}}{\text{セット数（十分多い）}}\\
=\ &\frac{\text{［「１セットあたりの成功回数の例（0 から } n \text{ まで）」} \times \text{その回数だけ成功したセットの数］の合計}}{\text{セット数}}\\
=\ &\text{［「１セットあたりの成功回数の例」} \times \frac{\text{その回数だけ成功したセットの数}}{\text{セット数}}\text{］の合計}\\
=\ &\text{［「１セットあたりの成功回数の例」} \times \text{「その回数成功したセットの割合」］の合計}\\
=\ &\text{［「１セットあたりの成功回数の例」} \times \text{「その回数成功する確率」］の合計}
\end{aligned}
$$

となります．これを確率変数 S の期待値といいます．

式で書くと，確率変数 S の期待値を $E(S)$ で表すとき，

$$E(S) = \sum_x xP(S = x) \tag{1}$$

となります[2]．

確率変数 S が２項分布 $B(n, p)$ にしたがうときは，その期待値 $E(S)$ は np となります[3]．

また，「ある１セットでの成功回数と，成功回数の期待値との差」を偏差といいます．さらに，「試行のあらゆる可能性を考えたときの，偏差の２乗の平均」，すなわち「偏差の２乗の期待値」を分散といい，分散の平方根を標準偏差といいます（図 2）．分散は，各セットでの成功回数の，偶然によるばらつきの大きさを表しています．

分散は「偏差の２乗の期待値」ですから，「［(１セットあたりの成功回数の例 − １セットあたりの成功回数の期待値）^{2} × その回数成功する確率］の合計」として求められます．

[2] "$\sum_x$" とは，「可能なすべての x について合計する」という意味です．
[3] この計算は，総合科学部専門科目「情報統計学」(2006 年度後期）第５回のプリント（付録）に掲載されています．講義のサイトからリンクしてあります．

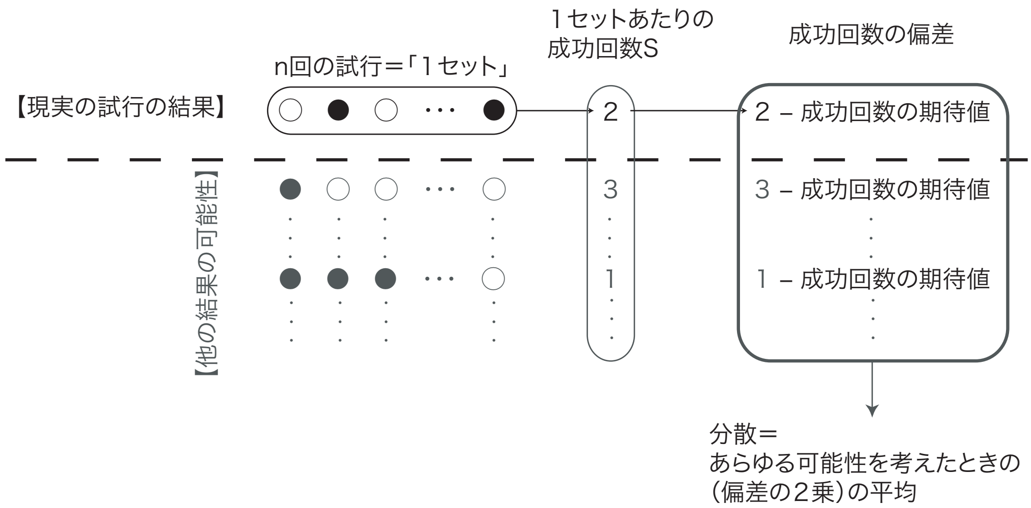

図 2: 確率変数の分散

式で書くと，確率変数 S の分散を $V(S)$ で表すとき，

$$V(S) = \sum_x (x - E(S))^2 P(S = x) \tag{2}$$

となります．

確率変数 S が２項分布 $B(n, p)$ にしたがうときは，その分散 $V(S)$ は $np(1-p)$ となります[4]．

1 回の試行での成功の確率が p で，それを n 回行うわけですから，成功回数の期待値，すなわち「1 セットで n 回試行するときの，成功する回数の平均」が np であることは直観的にわかります．また，$p = 1/2$ の時に分散が最大になることは，「成功する回数」の予測は $p = 1/2$ の時が一番しにくいことを示しています．

サンクト・ペテルブルクのパラドックス

ちょっと２項分布から離れて，「期待値」について，次の問題を考えてみましょう．

「コインを表が出るまで投げ続けて，はじめて表が出たときにやめる．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．

1. 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．
2. 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？

[4]期待値の場合と同様に，講義のサイトからリンクされている講義録を参照してください．

∧∧
≡・・≡　この賭けは，どうみてもやりたいとは思わないんですけど ...
(　)〜

∧◆∧
まあ，ふつうはそうやな．そやけど，期待値はどうなるやろか？　三 o- o 三
(　)〜

可能な X の値は $2, 2^2, \ldots, 2^n, \ldots$ となり，無限の種類があります．賞金が 2^n（円）である確率は $1/2^n$ ですから $(n = 1, 2, ...,)$，X がしたがう確率分布は $P(X = 2) = 1/2, P(X = 2^2) = 1/2^2, \ldots, P(X = 2^n) = 1/2^n, \ldots$ と表されます．

したがって，期待値 $E(X)$ は $2^1 \times (1/2^1) + 2^2 \times (1/2^2) + \cdots + 2^n \times (1/2^n) + \cdots = 1 + 1 + \cdots + 1 + \cdots \to \infty$ となります．

「賞金の額の期待値が無限大」ということは，この賭けを永遠にし続ければ無限に儲かる，つまり，客はいくら損をし続けても，いつかはその損をすべて取り返せる，というのが統計学や確率論の結論です．ウェブ上にこの賭けのシミュレータがあり（後述)，これまでの賭けの収支も掲載されています．これを見ると，長い目で見ると確かに胴元が損をしていることが記録されています．

しかし，では 1 回 100 円の賭け料でこの賭けをするかといわれれば，やらないでしょう．それは，「人は期待値の大小ではなく，期待効用の大小にしたがって行動する」からです．効用とは，満足度のことです．効用は金額に比例して多くはなりません．人は，この賭けに対して，確率の小さい，自分が生きているうちにめぐってくるかどうかもわからない大きな収益よりも，大きな確率で生じる小さな損失を重く見て，「この賭けは損だ，思うような満足度が得られない」と判断するのです．

この例は，数学者ダニエル・ベルヌーイによって提示されたもので，「サンクト・ペテルブルクのパラドックス」と言われています．「サンクト・ペテルブルクのパラドックス」については，講義のサイトの「統計データ・ツールへのリンク」から，解説のページやシミュレータのページにリンクしてありますので，参照してください．