

MAT220 第6回講義：確率論の続き

Contents

1	数え上げを用いる問題	2
1.1	形式の簡単な復習	2
1.2	例1：シャツとネクタイを選ぶ	3
1.3	例2：果物を並べる	5
1.4	例3：ロイヤルフラッシュ	8
2	古典的確率を越えて	9
2.1	古典的確率に隠れている仮定	9
2.2	単純事象に確率を割り当てる	10
2.3	おなじみの公式は引き続き成り立つ	11
2.4	例：不公平なサイコロ	12
2.5	例：不公平なコイン	12
3	離散確率分布	14
3.1	確率の重みから分布へ	14
3.2	確率変数	14
3.3	離散確率分布	15
3.4	分布の形	16

第5回講義では、確率の基本的な用語を導入した。すなわち、単純事象、標本空間、事象、補事象、共通部分、和事象、条件付き確率、独立性である。また、古典的確率ではすべての結果が同様に起こりやすいため、Golden Formula を用いて数え上げによって確率を計算できることも見た。今日は、この理論を実際に使い、古典的確率と第4回講義の数え上げの技法に関する問題を考える。その後、古典的確率を越えて、結果が必ずしも同様に起こりやすいとは限らない一般的な設定へ進む。これにより、確率分布へと自然につながり、最終的には記述統計へと戻ってくる。

1 数え上げを用いる問題

今日はまず古典的確率の練習問題をいくつか復習し、その後、結果が同様に起こりやすいとは限らない状況へ移る。その前に、前回の理論を簡単に復習しておこう。

1.1 形式の簡単な復習

第5回講義では、確率論の基本的な形式を導入した。そこで出てきた重要な概念をいくつか思い出しておこう。

1. 結果とは、実験や操作の結果として起こりうるものである。
2. すべての結果を集めたものを標本空間という。形式的には、これはすべての結果 ω を含む集合 S である。
3. 事象とは、大まかには、あるシステムについて述べられる命題のことである。形式的には、事象は標本空間 S の部分集合 A として定義される。
4. 単純事象とは、ただ一つの結果だけを含む特別な事象である。つまり、ある結果 $\omega \in S$ に対して $A = \{\omega\}$ となるものである。
5. 事象の確率は $P(A)$ と書かれ、0 から 1 までの数である。ここで 0 は不可能性を表し、1 は確実性を表す。

第5回講義では、古典的確率を扱った。これは、すべての結果が同様に起こりやすく、標本空間 S が有限個の結果からなる、という追加の仮定を置くものである。標本空間の大きさが $|S| = n$ ならば、各結果の確率は $\frac{1}{n}$ である。実際、これはいくつかの同値な形で書ける。厳密には、 P は結果ではなく事象を入力に取る関数である。したがって、ある結果が起こる確率は、単純事象を用いて次のように書ける。

$$P(\{\omega\}) = \frac{1}{n}.$$

前回の講義の中心的な公式は、古典的確率の Golden Formula であった。最初は直感的に導入し、その後、集合の性質に基づいて丁寧に調べた。Golden Formula は次の通りである。

Golden Formula

標本空間が有限で、すべての結果が同様に起こりやすいならば、

$$P(A) = \frac{\text{都合のよい結果の数}}{\text{結果の総数}} = \frac{|A|}{|S|}.$$

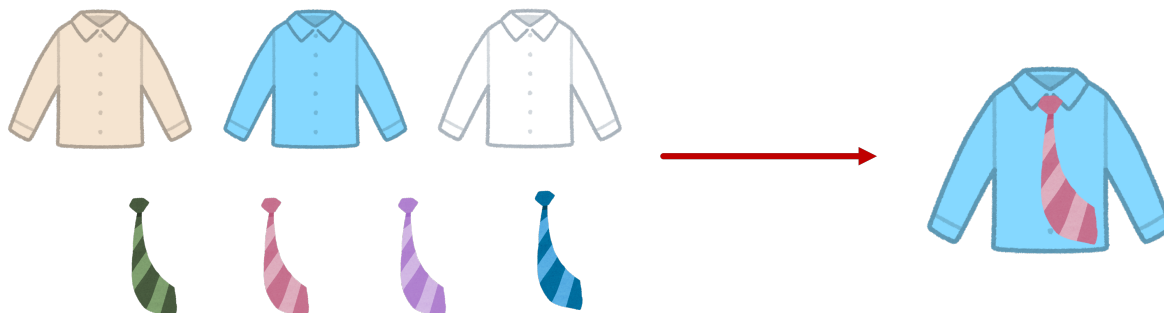
特に、単純事象 $A = \{\omega\}$ を取ると、Golden Formula を適用することで、各結果の確率 $\frac{1}{n}$ を復元できる。

Golden Formula が使えるのは古典的確率の場合だけである。つまり、標本空間 S が有限であり、 S のすべての結果が同様に起こりやすい場合である。多くの場合、主な難しさは計算そのものではなく、すべてを数え上げることである。これから、Golden Formula を異なる文脈でどのように使うか、三つの例を丁寧に見ていく。これらは第4回講義の議論に基づいている。

- いくつかのカテゴリから選ぶ場合、
- 物を異なる順序に並べる場合、
- 大きな集まりから無作為に選ぶ場合。

1.2 例1：シャツとネクタイを選ぶ

第4回講義の例を思い出そう。ワードローブには3枚のシャツと4本のネクタイがあり、その中から一つずつ選んで服装を作る。例えば、そのような組み合わせの一つを下に示す。



この設定を使うと、物を無作為に選ぶことができるので、服装の選択を確率の問題として扱える。次の三つの問いを考えよう。

演習 1

シャツを一枚選び、その後ネクタイを一本無作為に選ぶとする。ただし、それぞれのシャツとそれぞれのネクタイは同じ確率で選ばれる。

- (a) 茶色のシャツを選ぶ確率はいくらか。
- (b) 白いシャツと赤いネクタイを選ぶ確率はいくらか。
- (c) 青色を含む服装を選ぶ確率はいくらか。

この問いに答えるために、まず標本空間を作る。この設定では、結果は服装全体の選択である。第4回講義で見たように、これらは格子状に並べることができる。



見て分かるように、可能な服装は $12 = 3 \times 4$ 通りある。したがって、このシステムの標本空間 S の大きさは 12 であり、 $|S| = 12$ である。

- (a) 最初の問いには、いくつかの方法で答えられる。まず、非形式的な数え上げとして Golden Formula を適用できる。 A を「茶色のシャツと任意のネクタイを選ぶ」という事象とする。このとき

$$P(A) = \frac{\text{茶色のシャツを含む服装の数}}{\text{服装の総数}} = \frac{4}{12} = \frac{1}{3}.$$

別の実用的な見方として、シャツの選択はネクタイの選択から完全に独立しているので、シャツの選択だけを一つの確率システムとして考えることもできる。新しい標本空間を

$$S' = \{\text{白いシャツ, 青いシャツ, 茶色のシャツ}\}$$

と書ける。シャツは三枚あり、それぞれ同じ確率で選ばれるので、茶色のシャツを選ぶ確率は単に $\frac{1}{3}$ である。その後にネクタイを選んでも、茶色のシャツを選ぶ確率には影響しない。

- (b) 次の問いでは、個別の結果、すなわち白いシャツと赤いネクタイの確率を求めている。Golden Formula を直接使うと、

$$P(\text{白いシャツと赤いネクタイ}) = \frac{\text{白いシャツと赤いネクタイの服装の数}}{\text{服装の総数}} = \frac{1}{12}.$$

別の方法として、事象 A を「白いシャツを選ぶ」、事象 B を「赤いネクタイを選ぶ」と定義する。このとき「白いシャツと赤いネクタイを選ぶ」という文は

A and B : 白いシャツと赤いネクタイを選ぶ

と解釈できる。この二つの事象は独立なので、以前の掛け算の公式を使うと、

$$P(A \text{ and } B) = P(A) \cdot P(B) = \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{12}.$$

- (c) 最後に、何か青いものを含む服装を選ぶ確率を計算する。ここでも、いくつかの方法がある。おそらく最も単純な方法は、標本空間を見て、青色を含む服装をすべて強調することである。



見て分かるように、そのような服装は 6 通りある。したがって、Golden Formula を素直に適用すると、

$$P(\text{何か青いもの}) = \frac{\text{青色を含む服装の数}}{\text{服装の総数}} = \frac{6}{12} = \frac{1}{2}.$$

別の方法として、二つの事象

$$B = \{\text{青いシャツを含む服装}\} \quad \text{and} \quad C = \{\text{青いネクタイを含む服装}\}$$

を定義してもよい。上の図では、部分集合 B が赤で、部分集合 C が青で示されている。この二つの事象は互いに排反ではない。なぜなら、青いシャツと青いネクタイからなる服装が存在するからである。したがって、or 文を計算するときには、この重なりを引かなければならない。

$$P(B \text{ or } C) = P(B) + P(C) - P(B \text{ and } C) = \frac{4}{12} + \frac{3}{12} - \frac{1}{12} = \frac{6}{12} = \frac{1}{2}.$$

もちろん、これらの問いはすべて、事象を記号で実際に書くことで、より形式的に答えることができる。このシステムは、シャツとネクタイの選択、およびそれによって得られる服装に対応する順序対の集まりとして書ける。



上の図によれば、茶色のシャツは三番目であり、 s_3 に対応する。標本空間は、上の格子にあるすべての順序対の集まりとして書ける。

$$S = \{(s_1, t_1), (s_1, t_2), \dots, (s_3, t_4)\},$$

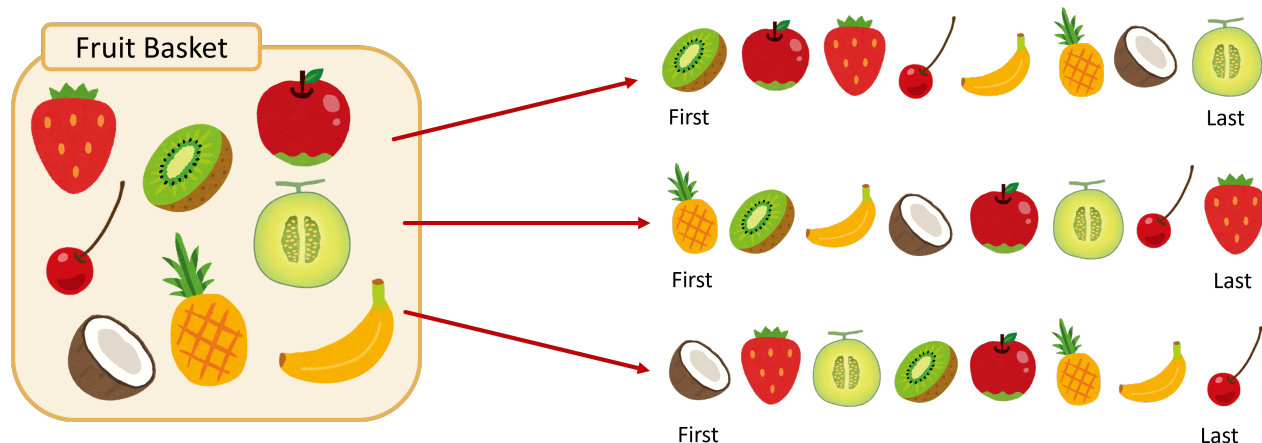
また、茶色のシャツを含むすべての服装からなる事象は

$$A = \{(s_3, t_1), (s_3, t_2), (s_3, t_3), (s_3, t_4)\}.$$

公式 $P(A) = \frac{|A|}{|S|}$ を使うと、先ほどと同じく $P(A) = \frac{1}{3}$ が得られる。

1.3 例2：果物を並べる

第4回講義の果物かごの例を思い出そう。そこでは、かごの中に 8 個の異なる果物があり、それらを一つずつ選んでいった。このとき、その方法は $8! = 40320$ 通りあることを見た。そのうち三つを下に示す。



演習 2

果物を順番に無作為に取り出すとする。言い換えると、かごの中の 8 個の果物の置換を無作為に一つ選ぶとする。

- (a) キウイを最初に選ぶ確率はいくらか。
- (b) イチゴが三番目にある並びを選ぶ確率はいくらか。
- (c) メロンが二番目にあり、サクランボが三番目にある並びを選ぶ確率はいくらか。

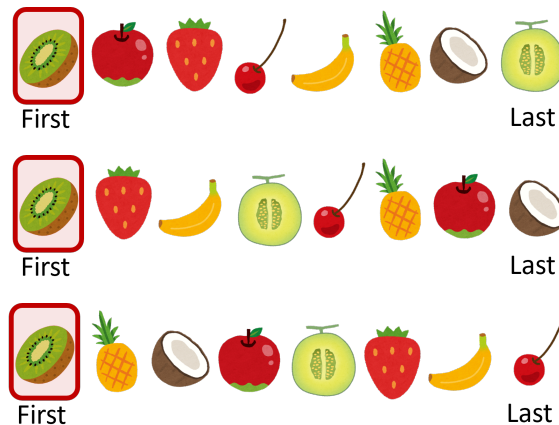
- (a) 二通りの方法で考える。まず、おそらく最も簡単な方法は、並びの最初の位置について、かごの中の 8 個の果物がそれぞれ同じ確率で選ばれると観察することである。果物は 8 個あり、そのうちキウイは一つだけなので、キウイを最初に選ぶ確率は

$$P(\text{キウイが最初}) = \frac{\text{キウイの数}}{\text{果物の総数}} = \frac{1}{8}.$$

これは、完全な置換のレベルで考えても確認できる。問題文にあるように、それぞれの並びは同じ確率で起こるので、キウイを最初に選ぶ確率は、次の比として計算できる。

$$P(\text{キウイが最初}) = \frac{\text{キウイが最初にある置換の数}}{\text{置換の総数}}.$$

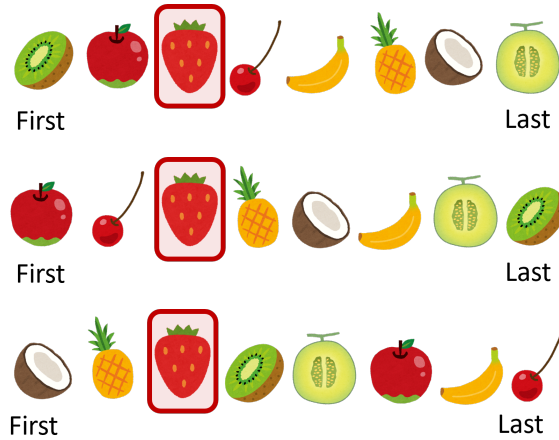
分母については、第4回講義で、8 個の果物を並べる方法が $8! = 40320$ 通りあることを見た。分子については、キウイを位置に固定することで計算できる。例えば、キウイを固定した可能な並びの三つは次のようになる。



これらの並びの総数を計算するには、残りの 7 個の果物を 7 個の場所に並べればよいことに注意する。したがって、可能な並びの総数は $7!$ であり、これを約分すると先ほどと同じ確率が得られる。

$$\begin{aligned} P(\text{キウイが最初}) &= \frac{\text{キウイが最初にある置換の数}}{\text{置換の総数}} \\ &= \frac{7!}{8!} = \frac{\cancel{7} \times \cancel{6} \times \cancel{5} \times \cancel{4} \times \cancel{3} \times \cancel{2} \times \cancel{1}}{8 \times \cancel{7} \times \cancel{6} \times \cancel{5} \times \cancel{4} \times \cancel{3} \times \cancel{2} \times \cancel{1}} = \frac{1}{8}. \end{aligned}$$

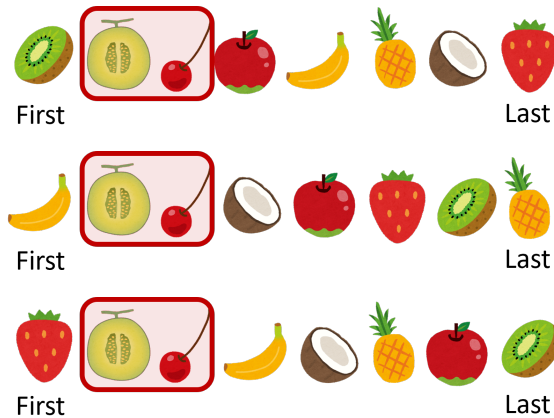
- (b) 同じ考え方で、イチゴを固定して、その周りに他のものを並べることができる。例えば、そのような並びの三つは次の通りである。



前の問いと同様に、イチゴの周りに並べる必要がある果物は 7 個であり、その方法は 7! 通りある。したがって、確率は再び

$$P(\text{イチゴが三番目}) = \frac{\text{イチゴが三番目にある置換の数}}{\text{置換の総数}} = \frac{7!}{8!} = \frac{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} = \frac{1}{8}.$$

- (c) 最後に、メロンが二番目にあり、サクランボが三番目にある置換を選ぶ確率を求める。二つの方法で考える。まず、これまでと同じように純粋な数え上げを行う。ただし、最初の二つの問いとは異なり、今回は二つの場所がすでに埋まっている。例えば、いくつかの置換は次のようになる。



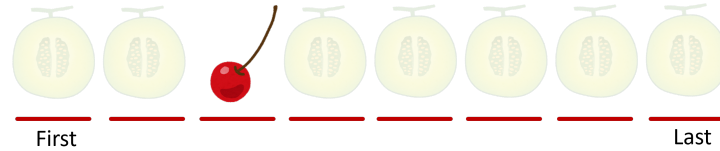
この場合、実質的には、メロンとサクランボの周りに残りの 6 個の果物を並べる方法を数えている。残りの果物は 6 個なので、並べ方の総数は 6! である。したがって、確率は

$$P(\text{メロンが二番目、サクランボが三番目}) = \frac{\text{メロンが二番目、サクランボが三番目にある置換の数}}{\text{置換の総数}} = \frac{6!}{8!} = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} = \frac{1}{8 \times 7} = \frac{1}{56}.$$

別の方法として、この問いを複合事象として解釈できる。もし

$$A = \{\text{メロンが二番目}\}, \quad B = \{\text{サクランボが三番目}\},$$

と置くなら、求めたい確率は $P(A \text{ and } B)$ である。「and」文の公式によれば、条件付き確率 $P(A|B)$ を計算する必要がある。(a) と (b) の議論から、サクランボが三番目にある確率は $\frac{1}{8}$ である。なぜなら、8 個の位置があり、それぞれが同じ確率でサクランボを含むからである。一方、条件付き確率 $P(A|B)$ を計算するには、サクランボが三番目にあることを知っているという条件のもとで、メロンが二番目にある確率を求める必要がある。サクランボが三番目にあると分かっているので、利用可能な位置の数は 8 から 7 に減る。



メロンはこれらの位置のどれに入ることも同じ確率である。特に、メロンが二番目にある確率は $\frac{1}{7}$ である。これをまとめると、この条件付き確率と $P(B)$ を掛けることで複合確率を計算できる。

$$P(A \text{ and } B) = P(A|B) \cdot P(B) = \frac{1}{7} \cdot \frac{1}{8} = \frac{1}{56}.$$

1.4 例3：ロイヤルフラッシュ

ポーカーにはいくつかの遊び方がある。一つの簡単な種類はファイブカードドローと呼ばれ、その名前が示すように、各プレイヤーには 5 枚のカードが配られる。プレイヤーは手札のカードを好きなように並べ替えられるので、カードの順序はまったく重要ではない。ルールによれば、最も強い手は「ロイヤルフラッシュ」であり、これは同じスートの 10、ジャック、クイーン、キング、エースからなる。

演習 3

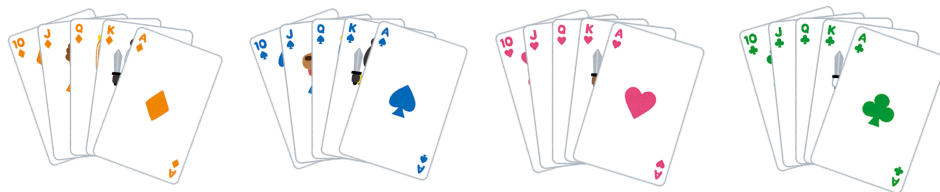
プレイヤーがロイヤルフラッシュを受け取る確率はいくらか。

この問いに答えるために、Golden Formula と組合せ論を使う。確率は

$$P(\text{ロイヤルフラッシュ}) = \frac{\text{可能なロイヤルフラッシュの数}}{\text{可能な手札の数}}$$

となるはずである。

幸い、分子は非常に簡単に計算できる。スートは 4 種類あるので、可能なロイヤルフラッシュは 4 通りである。



分母を計算するために、標準的なトランプのデッキには 52 枚のカードがあることを思い出す。ファイブカードドローのルールでは、一つのポーカーの手札は 5 枚のカードからなり、手札の中でのカードの順序は重要ではない。したがって、52 枚から 5 枚を、順序を気にせず

選ぶ方法の数を数えればよい。これを行うために、第4回講義の $C_{n,r}$ の公式を使う。ここでは $n = 52$ 、 $r = 5$ である。

$$C_{52,5} = \frac{52!}{5!47!} = 2,598,960.$$

これで、確率は次のように計算できる。

$$P(\text{ロイヤルフラッシュ}) = \frac{\text{可能なロイヤルフラッシュの数}}{\text{可能な手札の数}} = \frac{4}{2,598,960} = \frac{1}{649,740}.$$

したがって、ロイヤルフラッシュを得る確率は 649,740 分の 1 である。この例は、Golden Formula そのものは非常に単純であることをよく示している。難しいのはしばしば分母である。確率を計算する前に、標本空間の構造を十分に理解し、それを数え上げられるようにしなければならない。

2 古典的確率を越えて

2.1 古典的確率に隠れている仮定

ここまで、ほとんどすべての確率計算で Golden Formula を使ってきた。

$$P(A) = \frac{|A|}{|S|}.$$

使える場合には、この公式は非常に有用である。しかし、その適用範囲は限られている。なぜなら、この公式は n 個の結果がすべて同じ確率、つまり $\frac{1}{n}$ で起こると仮定しているからである。この仮定は、扱っているシステムにある種の対称性があるために正当化されることが多い。例えば、公平なコインには二つの反対側の面があり、6 面のサイコロは立方体である。しかし、現実には多くのシステムはこのように完全には対称ではない。そのような場合でも確率について議論できるが、すべての結果が同じ確率で起こるという仮定を緩めなければならない。

例えば、コインが表または裏に出る確率が等しくなく、次のようであるとする。

$$P(H) = 0.4, \quad P(T) = 0.6.$$

ここで $P(H)$ は $P(\{H\})$ を意味し、 T についても同様である。標本空間は依然として $S = \{H, T\}$ であるが、確率はもはや 50% ずつではない。言い換えれば、二つの結果は同様に起こりやすいわけではない。この場合、単純な Golden Formula を使って事象の確率を求めることはできない。

$$P(\text{表が出る}) \neq \frac{S \text{ 中の表の数}}{S \text{ の大きさ}} = \frac{1}{2}.$$

これは単純な例にすぎないが、より一般的な確率の扱いが必要であることを動機づけている。これから、この点を詳しく議論する。この節では、確率論の規則はすべて引き続き成り立つと仮定するが、すべての結果が同様に起こりやすいとは仮定しない。

2.2 単純事象に確率を割り当てる

S を有限な標本空間とし、例えば

$$S = \{\omega_1, \omega_2, \dots, \omega_n\}$$

と書く。ここで ω_i は、このシステムの n 個の異なる結果である。各 ω_i が固定された確率 $\frac{1}{n}$ を持つと仮定する代わりに、その仮定を緩め、各個別の結果の確率を単に

$$p(\omega_i)$$

と書くことにする。これは $P(\{\omega_i\})$ の略記である。

確率論の他の規則と整合するために、次のことを要求する必要がある。

1. 各確率 $p(\omega_i)$ は 0 から 1 までの数である。
2. それらすべての確率の和は 1 である。

$$\sum_{i=1}^n p(\omega_i) = p(\omega_1) + p(\omega_2) + \dots + p(\omega_n) = 1.$$

不公平なコインの例がこれらの性質を満たしていることは確認できる。この設定を使うと、より一般的な場合の新しい「Golden Formula」を作ることができる。

有限確率モデル

$S = \{\omega_1, \omega_2, \dots, \omega_n\}$ を大きさ n の有限標本空間とする。確率モデルは、各結果 ω_i に数 $p(\omega_i)$ を割り当てる。ただし、

$$0 \leq p(\omega_i) \leq 1$$

かつ

$$\sum_{\omega \in S} p(\omega) = 1.$$

このとき、任意の事象 $A \subseteq S$ に対して、

$$P(A) = \sum_{\omega \in A} p(\omega)$$

と定義する。

この最後の公式が、Golden Formula の一般的な置き換えである。これは、事象 A の確率を計算するには、 A に含まれるすべての結果 ω を見て、それらの確率を足し合わせればよい、ということ述べている。

実際、古典的確率の Golden Formula は、このより一般的な考え方の一つの例にすぎない。これを見るために、事象 A が m 個の結果を含む、すなわち

$$A = \{\omega_1, \dots, \omega_m\}$$

であるとする。ここで、標本空間 S の大きさを n とし、各 ω_i が同じ確率 $\frac{1}{n}$ を持つと仮定すると、簡単な代数から次が分かる。

$$P(A) = \sum_{\omega \in A} p(\omega) = p(\omega_1) + \dots + p(\omega_m) = \underbrace{\frac{1}{n} + \dots + \frac{1}{n}}_{m \text{ 個}} = \frac{1 + \dots + 1}{n} = \frac{m}{n} = \frac{|A|}{|S|}.$$

したがって、古典的確率の Golden Formula は、より大きな枠組みに含まれていることが分かる。

2.3 おなじみの公式は引き続き成り立つ

古典的確率を越えたとしても、第5回講義の主要な確率公式は引き続き成り立つ。念のため、もう一度まとめておく。

確率の公式

事象 A と B について、

$$P(\text{not } A) = 1 - P(A).$$

and 文について、

$$P(A \text{ and } B) = \begin{cases} P(A) \cdot P(B), & A \text{ と } B \text{ が独立である場合,} \\ P(A|B) \cdot P(B), & A \text{ と } B \text{ が従属である場合 (かつ } P(B) \neq 0 \text{) .} \end{cases}$$

or 文について、

$$P(A \text{ or } B) = \begin{cases} P(A) + P(B), & A \text{ と } B \text{ が互いに排反である場合,} \\ P(A) + P(B) - P(A \text{ and } B), & \text{一般の場合.} \end{cases}$$

なぜこれが自然なのか、簡単に見ておこう。

補事象については、標本空間は重ならない二つの部分に分かれる。

$$S = A \cup A^c.$$

すべてのものは A に属するか、 A に属さないかのどちらかである。したがって、

$$P(A) + P(A^c) = 1,$$

なので

$$P(A^c) = 1 - P(A).$$

or 文については、問題は重なりである。 $P(A)$ と $P(B)$ を足すと、 $A \cap B$ に入っているものを二回数えてしまう。したがって、一回分を引く。

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

and 文については、一般的な規則は条件付き確率から来る。 $P(A) \neq 0$ ならば、

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}.$$

これを変形すると、

$$P(A \text{ and } B) = P(A) \cdot P(B|A).$$

もし A と B が独立ならば、 A が起こったと知っても B の確率は変わらない。この場合、

$$P(B|A) = P(B),$$

したがって

$$P(A \text{ and } B) = P(A) \cdot P(B).$$

2.4 例：不公平なサイコロ

次のような確率を持つ不公平なサイコロがあるとする。

$$p(1) = 0.1, \quad p(2) = 0.1, \quad p(3) = 0.1, \quad p(4) = 0.1, \quad p(5) = 0.5, \quad p(6) = 0.1.$$

これは公平なサイコロではない。なぜなら、5 が出る確率が他の各結果よりもかなり高いからである。しかし、これは完全に有効な確率モデルである。各個別の確率は 0 と 1 の間にあり、それらの総和が正確に 1 だからである。

$$0.1 + 0.1 + 0.1 + 0.1 + 0.5 + 0.1 = 1.$$

これから、新しい公式を使って、このシステムに関する確率をいくつか計算する。

演習 4

上の不公平なサイコロを使って、次を計算しなさい。

- (a) 5 以外の数が出る確率はいくらか。
- (b) 偶数が出る確率はいくらか。
- (c) 3 より小さい数が出る確率はいくらか。

解答

- (a) A を「5 が出る」という事象とする。このとき、「5 とは異なる数が出る」という事象は A の補事象である。したがって、

$$P(5 \text{ ではない}) = 1 - p(5) = 1 - 0.5 = 0.5.$$

- (b) 偶数の結果は 2、4、6 である。したがって、

$$P(\text{偶数}) = p(2) + p(4) + p(6) = 0.1 + 0.1 + 0.1 = 0.3.$$

- (c) 3 より小さい結果は 1 と 2 である。したがって、

$$P(3 \text{ より小さい}) = p(1) + p(2) = 0.1 + 0.1 = 0.2.$$

2.5 例：不公平なコイン

不公平なコインの例に戻ろう。ここでは

$$P(H) = 0.4, \quad \text{and} \quad P(T) = 0.6.$$

演習 5

この不公平なコインを独立に何回か続けて投げるとする。

- (a) 二回連続で表が出る確率はいくらか。
- (b) 一回は表、一回は裏が出る確率はいくらか。ただし順序は問わない。
- (c) 三回連続で裏が出る確率はいくらか。

解答

- (a) A を「一回目に表が出る」という事象、 B を「二回目に表が出る」という事象とする。コイン投げは独立なので、単に掛け算をすればよい。

$$P(HH) = P(H) \cdot P(H) = 0.4 \times 0.4 = 0.16.$$

- (b) 一回は表、一回は裏が任意の順序で出るには、二つの可能な列がある。

$$HT \quad \text{or} \quad TH.$$

これらは互いに排反なので、その確率を足す。

$$P(HT \text{ or } TH) = P(HT) + P(TH).$$

コイン投げは独立なので、

$$P(HT) = 0.4 \times 0.6 = 0.24,$$

また

$$P(TH) = 0.6 \times 0.4 = 0.24.$$

したがって、

$$P(HT \text{ or } TH) = 0.24 + 0.24 = 0.48.$$

- (c) 三回連続で裏が出る確率は

$$P(TTT) = P(T) \cdot P(T) \cdot P(T) = 0.6 \times 0.6 \times 0.6 = 0.216.$$

(b) の拡張として、二回投げたときの完全な標本空間を書ける。

$$S = \{HH, HT, TH, TT\}.$$

四つの確率は掛け算によって得られる。

結果	HH	HT	TH	TT
確率	0.16	0.24	0.24	0.36

これらは等しくなく、全体として確率は HH から離れ、 TT の方へ偏っている。確認すると、これらの確率は当然ながら合計して 1 になる。

$$0.16 + 0.24 + 0.24 + 0.36 = 1.$$

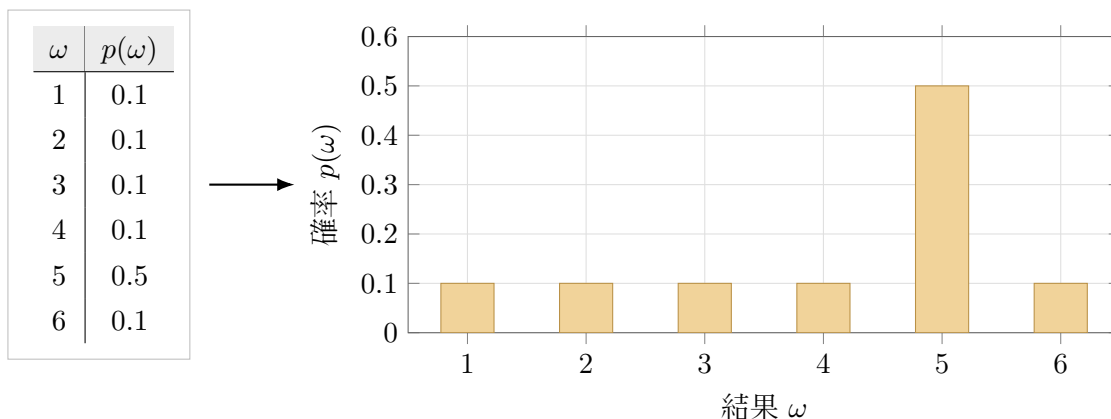
3 離散確率分布

3.1 確率の重みから分布へ

第2.4節の不公平なサイコロの例では、各結果に確率を割り当てた。

$$p(1) = 0.1, \quad p(2) = 0.1, \quad p(3) = 0.1, \quad p(4) = 0.1, \quad p(5) = 0.5, \quad p(6) = 0.1.$$

ここで概念的な飛躍をする。この情報の集まりは、データのように整理できる。つまり、表に書き出し、ヒストグラムとして描くことができる。



これは第2回講義での議論を思い出させる。ただし、ここでは確率的な文脈が加わっている。棒の高さは単なる度数ではなく、確率である。したがって、このヒストグラムは、システム全体の確率が個々の結果にどのように分布しているかを示している。こうして、確率分布に出会ったことになる。

確率分布

確率分布とは、確率がランダムな実験の可能な結果や値の間にどのように広がっているかを示すものである。

次に、確率分布の横軸の性質について簡単に触れる。

3.2 確率変数

上の例では、ヒストグラムの x 軸は標本空間、すなわち 1 から 6 までの数によってラベル付けされている。通常の数学では、この軸を変数でラベル付けすることがある。これは、実数直線上の値を取ることができる謎の文字 x である。普通のグラフは、 x 変数を数直線上で動かし、それに対応する出力を計算することで記述される。確率分布の場合には、単に変数を「動かす」のではなく、変数が標本空間のある値をランダムに取ると考える。そのため、 x -軸は確率変数の可能な値を示していると解釈する。このような確率変数はしばしば X のような大文字で書く。

確率変数については、後の講義、特に第8回講義の二項分布の扱いで詳しく議論する。今のところ、次の定義を理解しておけば十分である。

定義：確率変数

確率変数とは、ランダムな実験の各結果に数値を割り当てる規則である。

ここでの考え方は、与えられた標本空間が、ヒストグラムに描けるような数量的データからなるとは限らない、ということである。そこで、確率変数は結果を数値に変換する方法を与え、その結果として得られる確率分布に対して記述統計を使えるようにしてくれる。
例として、公平なコインを二回投げるとする。標本空間は質的なものである。

$$S = \{HH, HT, TH, TT\}$$

そして対応する確率は

$$P(HH) = P(HT) = P(TH) = P(TT) = \frac{1}{4}.$$

この情報から、例えば数を数えることで確率変数 X を定義できる。

$$X = \text{表の数.}$$

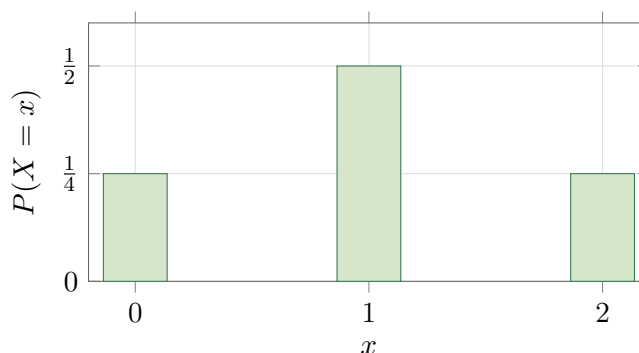
これにより、同じシステムを数値を使って記述できる。

$$X(HH) = 2, \quad X(HT) = 1, \quad X(TH) = 1, \quad X(TT) = 0.$$

ここから、確率を次のように再評価できる。

$$P(X = 0) = P(TT) = \frac{1}{4}, \quad P(X = 1) = P(HT) + P(TH) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}, \quad P(X = 2) = P(HH) = \frac{1}{4}.$$

この情報から、真の確率分布が得られる。



3.3 離散確率分布

通常の数、例えば3は、離散的にも連続的にも解釈できることを思い出そう。3を何かの個数を数えるものとして考えることもできるし、連続的な区間の長さを表すものとして考えることもできる。この講義では、標本空間が有限サイズなので、離散確率変数だけを扱う。

定義：離散確率分布

X が有限な離散確率変数であり、その可能な値が

$$x_1, x_2, \dots, x_n$$

であるとする。このとき、**離散確率分布**は各値に確率を割り当てる。

$$P(X = x_1), \quad P(X = x_2), \quad \dots, \quad P(X = x_n).$$

これらの確率は

$$0 \leq P(X = x_i) \leq 1$$

および

$$P(X = x_1) + P(X = x_2) + \dots + P(X = x_n) = 1$$

を満たさなければならない。

分布が得られれば、確率の問題は足し算の問題になる。例えば、

$$P(X \geq 1) = P(X = 1) + P(X = 2) = \frac{2}{4} + \frac{1}{4} = \frac{3}{4}.$$

言葉で言えば、これは少なくとも一回は表が出る確率である。

同様に、

$$P(X < 2) = P(X = 0) + P(X = 1) = \frac{1}{4} + \frac{2}{4} = \frac{3}{4}.$$

離散分布の使い方

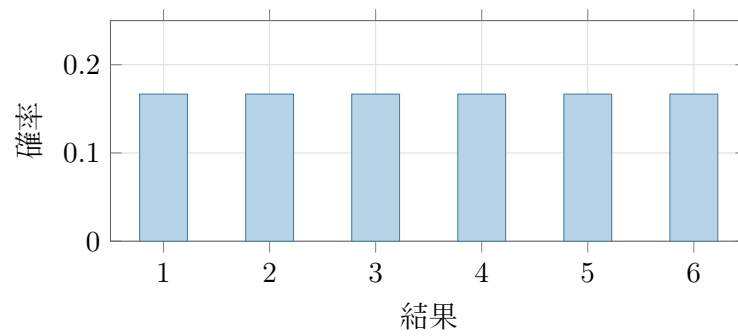
X に関する条件の確率を計算するには、その条件を満たすすべての値 x の確率を足し合わせる。

例えば、

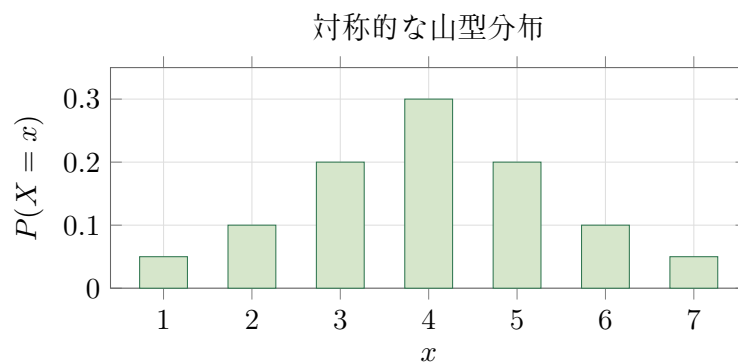
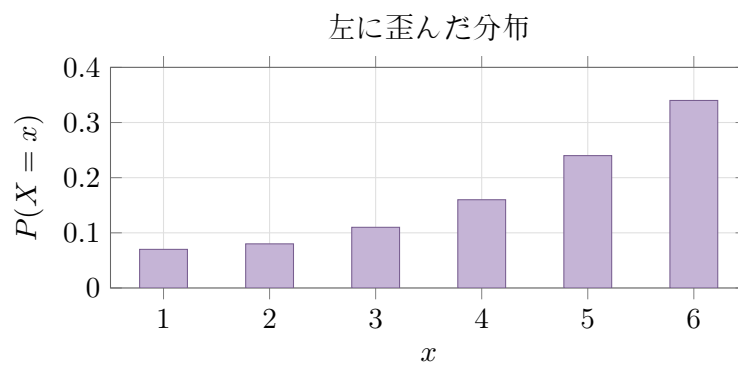
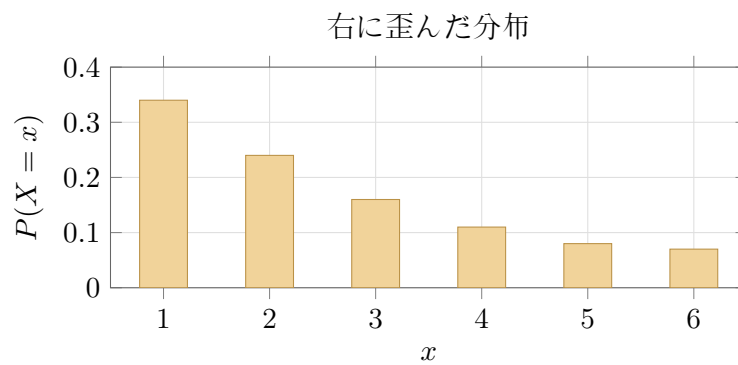
$$P(X \geq 1) = \sum_{x \geq 1} P(X = x).$$

3.4 分布の形

第3回講義で見たように、ヒストグラムの形は分布全体の構造を効率よく知らせてくれる。同じ考え方は確率分布にも当てはまる。おそらく最も単純な場合は古典的確率である。この場合、各結果はまったく同じ確率を持つので、得られる分布は一様になる。



しかし、他のシステムでは異なる形を取ることがある。例えば、確率が左に歪んでいたたり、右に歪んでいたたり、あるいは対称的な山型であったりする。



第8回講義と第9回講義では、離散確率分布について詳しく話す。そこで見ると、二項分布は確率における特に興味深い設定である。

わー！期待値についての部分をタイプする時間がなくなりました！週末に追加します。