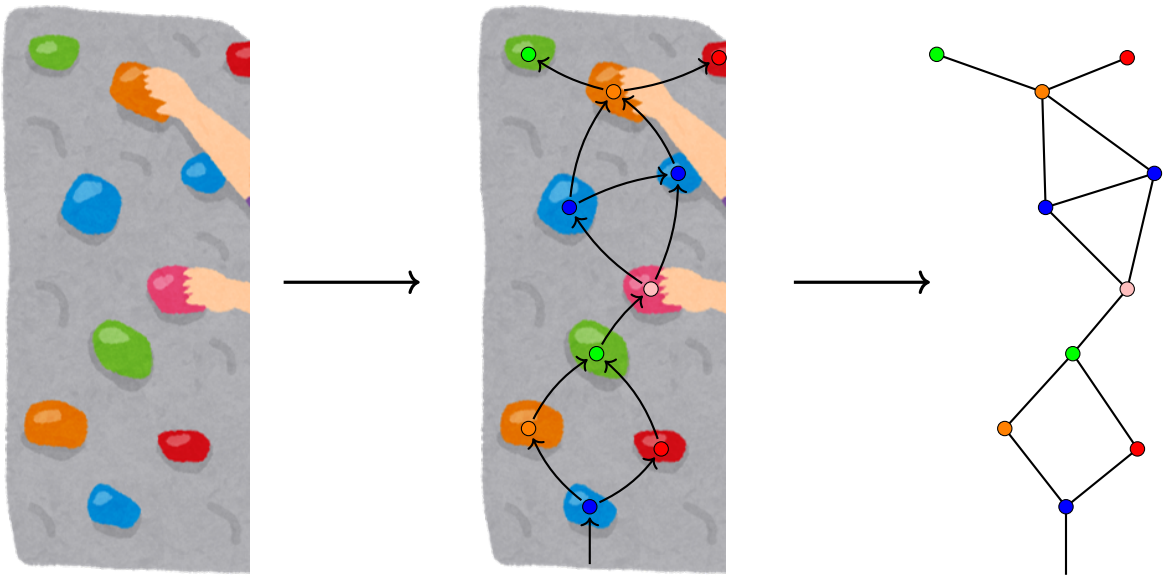
MAT120：数量的推論　第11講ハンドアウト
逐次確率

前回の講義では，不確実性を数量化する方法として確率を導入し，結果を数え上げることで古典的確率を計算する方法を学びました。今回の講義では，条件付き確率を発展させます。これは，追加の情報を得たあとでの確率を意味します。また，多段階の実験を整理するための実用的な道具である，決定木についても学びます。最後に，コイン投げにおける結果の数え上げを始めます。これは次回の講義で扱う二項実験へとつながっていきます。

1　導入：岩を登る人

岩場を登っている人の手の動きを追いかけるとしましょう。この状況は，色のついた岩どうしがどのようにつながっているかを表す一種の「グラフ」を作ることで理想化できます。



この図では，登る人が岩を上へ移動するにつれて，その手は下の青い岩から出発して，上の緑の岩または赤い岩までの経路をたどります。この過程のある段階では，登る人はどのように進むかを選ばなければなりません。たとえば岩場のいちばん最初では，青い岩からオレンジの岩へ進むこともできますし，あるいは赤い岩へ進むこともできます。他方で，ある地点では何をするかを選ぶことはできず，進み方は一意に決まっています。たとえば最初のオレンジの岩からは，次の緑の岩へ進むしかありません。

登る人が選択をするたびに，左の岩を選ぶ確率が50%，右の岩を選ぶ確率も50% であるとしましょう。実際，下から上までの異なる経路は全部で12 本あります。では，登る人がある特定の経路をたどる確率はどれくらいでしょうか。

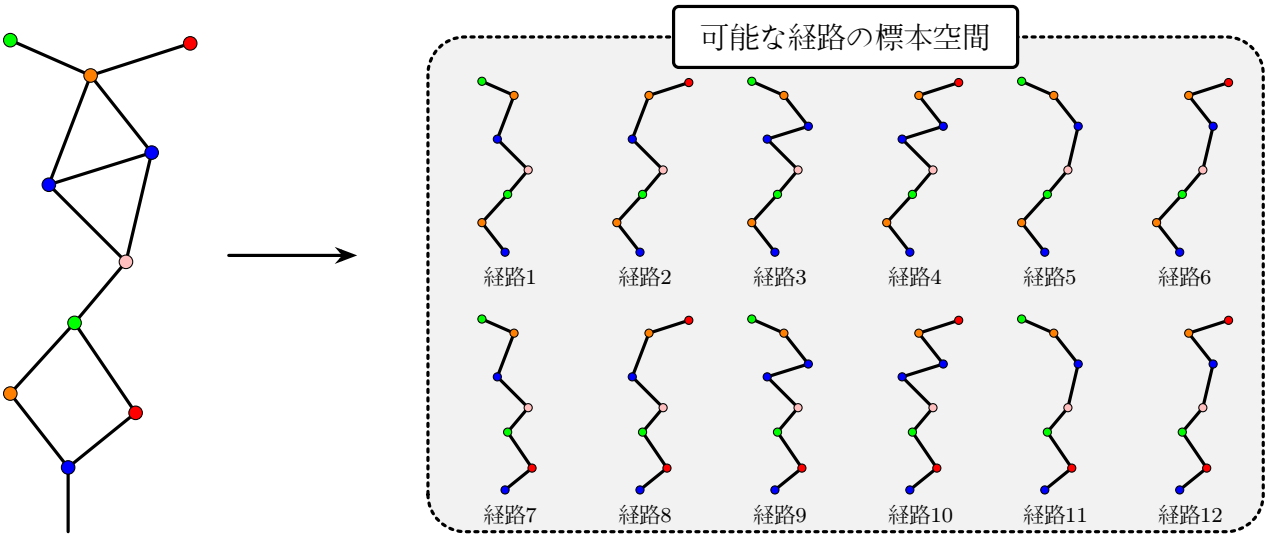
1.1　素朴な数え上げで解こうとしてみる

前回の講義では，古典的確率の設定で成り立つ基本公式を見ました。

$$P(\text{事象}) = \frac{\text{その事象に有利な結果の数}}{\text{結果の総数}}.$$

要するに，一般の事象の確率は，その事象に対応する結果の数を数え，それを可能性の総数で割ることで計算できることが多いのです。

今回の岩登りの例でも，同じことを試してみたくなります。まず標本空間を記述する必要がありますが，この場合それは下から上までのすべての経路の集合です。少し考えると，この岩場には異なる経路が12 本あることが分かります。それらを下に図示します。



したがって，この設定における標本空間は

$$S = \{\text{経路1},\ \text{経路2},\ \text{経路3}, \ldots,\ \text{経路12}\}$$

という集合です。したがって，上の公式を使ってある特定の経路の確率を計算したくなるのはもっともです。たとえば経路1 について，

$$P(\text{経路1}) \stackrel{?}{=} \frac{\text{その事象に有利な結果の数}}{\text{結果の総数}} = \frac{1}{12}$$

と推測したくなります。

しかし実際には，これは正しくありません。あとでこれを正しく計算しますが，今の段階では単に次のことを観察すれば十分です。経路全体の空間を見ると，ある経路は他の経路より複雑です。これらの異なる経路は**同様には起こりません**。したがって，上の公式はここでは正しいアプローチではないのです。

2　決定木

2.1　決定木とは何か

前回の講義では，条件付き確率という考え方を簡単に導入しました。これは，複合事象「A and B」の確率を計算する公式の中に現れていました。依存する事象A とB について，事象「A and B」の確率は

$$P(A \text{ and } B) = P(A)\, P(B \mid A)$$

で与えられました。同値な形として，

$$P(A \text{ and } B) = P(B)\, P(A \mid B)$$

とも書けます。ここで$P(B \mid A)$ という項が条件付き確率 であり，大まかには「A が起こると仮定したもとでのB の確率」を意味します。一般に条件付き確率とは，仮定によって情報を更新した結果として生じる確率の変化のことです。

前回，条件付き確率は

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} \qquad (\text{ただし} P(A) > 0)$$

によって定義できることを見ました。古典的確率では各結果が等確率なので，上の式は数え上げによって

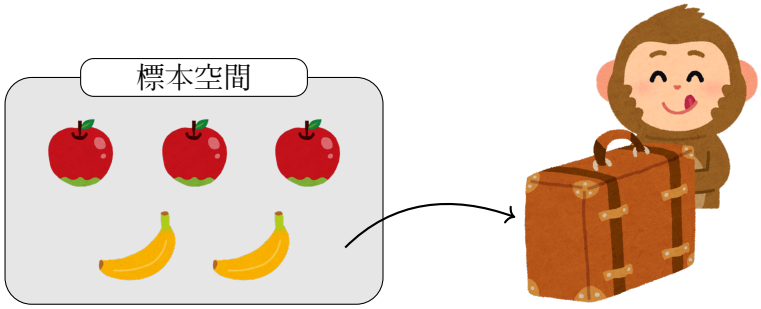
$$P(B \mid A) = \frac{\text{size of } A \cap B}{\text{size of } A}$$

と計算できます。これは集合論的な共通部分 $A \cap B$ の大きさを数える方法です。すなわち，まず事象A の中にある結果だけに注目し，そのうちどれだけがさらにB にも属しているかを数えるのです。

条件付き確率は，逐次確率，つまり複数段階からなる実験を考えるときにとくに有用になります。各段階ごとに異なる可能な結果があるような実験です。決定木とは，このような多段階実験の可能な結果を漏れなく並べた図のことです。結果の集まりは枝分かれしていき，木の枝として表されます。[1] 要するに，決定木は可能な結果の空間を整理して，各個別の結果の確率を計算しやすくするための道具なのです。

2.2 例：ジャングルと泥棒

あなたがジャングルにいて，かばんをそのまま放置したとしましょう。すると近くのサルが木から降りてきて，あなたのかばんの中を無作為にあさり始めます。たまたま，昼食用に入れていた果物が5 個あり，りんごが3 個，バナナが2 本入っていました。サルは果物を無作為に盗み，見つけたものを食べてからまたかばんの中を見ます。確率の用語では，これを復元しない抽出 といいます。



演習1

(a) サルが最初にりんごを盗む確率は何か。
(b) サルが最初にバナナを盗む確率は何か。

[1] 少し正確に言うと，枝とは木の始点から終点までを結ぶ線のことです。

解答

果物は全部で5 個なので，単純に数え上げればよいです。

(a) P(最初がりんご) $= \dfrac{3}{5}$.

(b) P(最初がバナナ) $= \dfrac{2}{5}$.

ここで，次の逐次的な問いに注目しましょう。

> サルがまずバナナを盗み，そのあとでりんごを盗む確率は何か。

2 回目の結果は1 回目の結果によって明らかに影響を受けるので，ここでは条件付き確率を使うべきです。

順序をはっきりさせるために，B_1 を「1 回目にバナナ」，A_2 を「2 回目にりんご」とします。すると「バナナ，そのあとりんご」は事象「B_1 and A_2」です。

これは条件付き確率の公式を用いて計算できます。ここでは，事象A を「サルがりんごを取る」，事象B を「サルがバナナを取る」と対応させます。すると複合事象「B and A」は「サルがバナナを取り，そのあとでりんごを取る」と解釈できます。したがって

$$P(B_1 \cap A_2) = P(B_1) \cdot P(A_2 \mid B_1)$$

を用いて計算できます。

この公式を順に見ていきましょう。

Step 1: すでに$P(B_1) = \dfrac{2}{5}$ であることを見ました。

Step 2: 次に$P(A_2 \mid B_1)$ を計算します。バナナが1 本取り除かれて（そしてサルに食べられて）いると，2 回目に果物を取ろうとするとき，サルは異なる 標本空間を持つことになります。最初の時点では果物は5 個ありました。しかしバナナが1 本すでに取り除かれているので，今では残りは4 個，すなわちりんご3 個とバナナ1 本だけです。この新しい標本空間の大きさは4 であり，サルが選べるりんごは3 個あります。したがって条件付き確率は

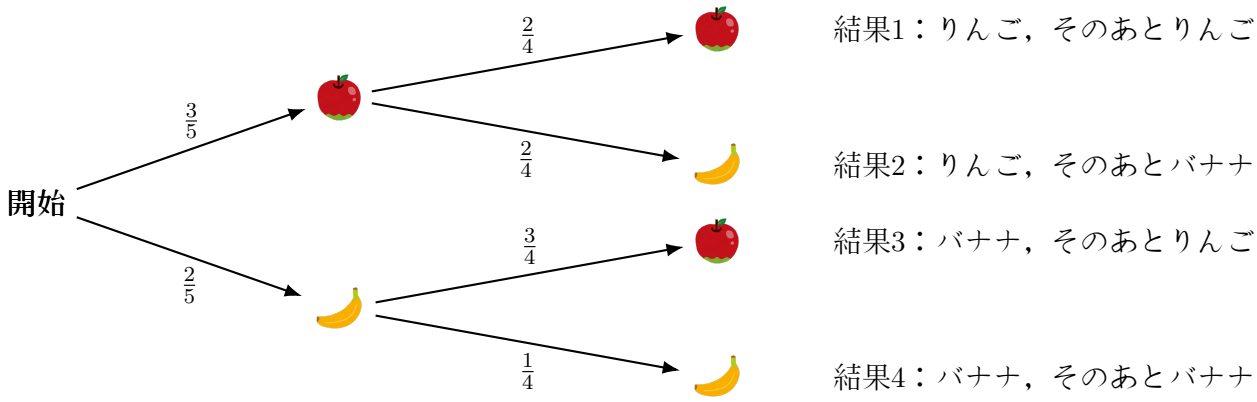
$$P(A_2 \mid B_1) = \frac{3}{4}$$

です。

Step 3: 最後に，この2 つの数を掛けて，「バナナを取ってそのあとりんごを取る」確率を得ます。

$$P(B_1 \text{ and } A_2) = P(B_1)\, P(A_2 \mid B_1) = \frac{2}{5} \cdot \frac{3}{4} = \frac{3}{10}.$$

この一連の過程は，下のような木に整理できます。



ご覧のように，先ほど行った計算

$$P(\text{バナナ，そのあとりんご}) = P(B_1) \cdot P(A_2 \mid B_1) = \frac{2}{5} \cdot \frac{3}{4} = \frac{3}{10}$$

は，上の木で結果3 に対応する枝を見れば視覚的にも確認できます。ここでは単に3 本目の枝に書かれた数を掛けているだけです。実は，ここで一般的なコツの最初の例を見たことになります。つまり，多段階実験の結果を木として整理できるなら，どの結果の確率も，その枝に現れるすべての確率を掛けることで求められるのです。

この観察を用いれば，上の木にある他の3 つの結果もすぐに計算できます。

- 結果1：$P(A \text{ then } A) = \frac{3}{5} \cdot \frac{2}{4} = \frac{3}{10}$,
- 結果2：$P(A \text{ then } B) = \frac{3}{5} \cdot \frac{2}{4} = \frac{3}{10}$,
- 結果4：$P(B \text{ then } B) = \frac{2}{5} \cdot \frac{1}{4} = \frac{1}{10}$.

これら4 つの確率の和は

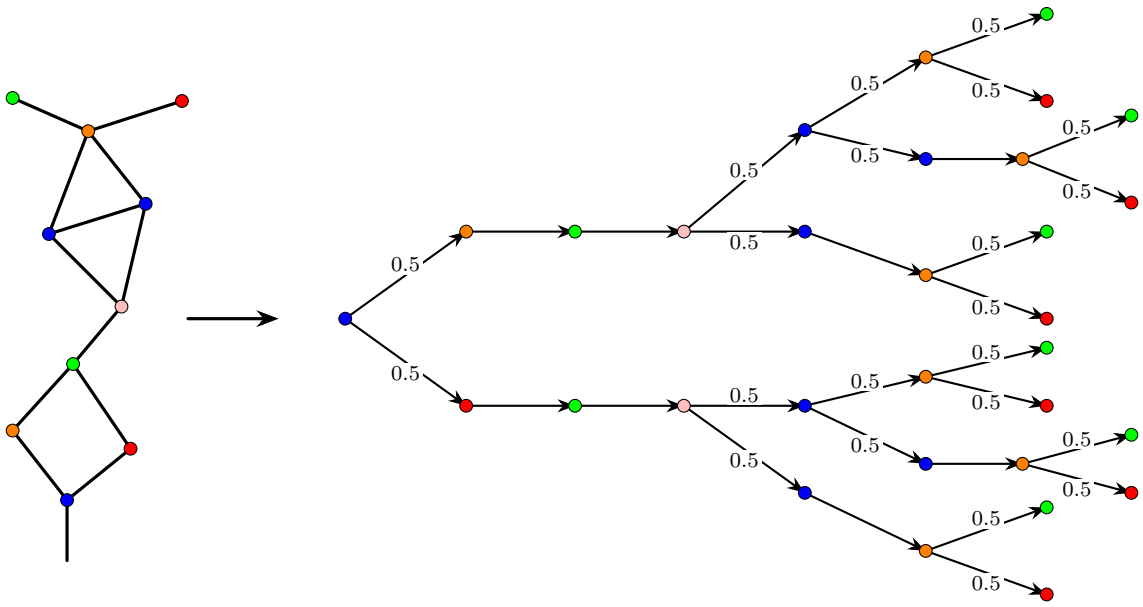
$$\frac{3}{10} + \frac{3}{10} + \frac{3}{10} + \frac{1}{10} = 1$$

になります。これは，4 つの結果のうちどれか1 つは必ず起こる，ということと同じです。

2.3　岩登りの問題に戻る

今見たように，決定木を使うと，多段階実験のすべての可能な結果を整理できます。ある特定の経路の確率を求めるには，その経路に沿って書かれた確率をすべて掛ければよいのです。この方法を使って，岩場を登る人が壁を進む各経路の確率を求めることができます。

まずこの系に対する木を書き，各「選択点」に仮定された50-50 の確率を書き込みます。完全な木は下の通りです。



この木は，各経路が順番に並ぶように整理されています。最初の枝が経路1，2 本目の枝が経路2，という具合です。ここでも，ある経路は他の経路より複雑であることに気づくはずです。たとえば経路12 は，経路9 や10 より単純に見えます。これは，壁を登るルートの中には3 回の選択だけで済むものもあれば，4 回の選択を必要とするものもあるからです。実際，ピンクの節点で左へ曲がる選択を含む経路は，右へ曲がる選択を含む経路より複雑になります。

いずれにせよ，決定木の理論を使えば，岩場を登る任意の経路の確率を計算できます。さきほどのサルの例と同様に，ここではその枝に書かれている値を掛け合わせれば全体の確率が得られます。まとめると以下の通りです。

各経路の確率

各経路で必要な選択回数を区別します。

- 経路$1, 2, 3, 4, 7, 8, 9, 10$ はすべて4 回の選択を必要とします。決定木では，これらの経路に対応する枝に0.5 が4 回現れるので，全体の確率は

$$\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{16}$$

  となります。
- 経路$5, 6, 11, 12$ は全部で3 回 の選択しか必要としません。したがって，これらの経路をたどる確率は他より大きく，

$$\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$$

  となります。

計算が正しいかどうかは，すべての答えの和が1 になることを確かめればいつでも確認できます。実際，

$$8\left(\frac{1}{16}\right) + 4\left(\frac{1}{8}\right) = \frac{1}{2} + \frac{1}{2} = 1$$

で，正しいことが分かります。ここで私たちは偶然もうひとつの事実も見つけています。登る人が単純に見える経路のどれかをたどる確率は50% であり，そうならない確率も50% です。

3　結果の数え上げ

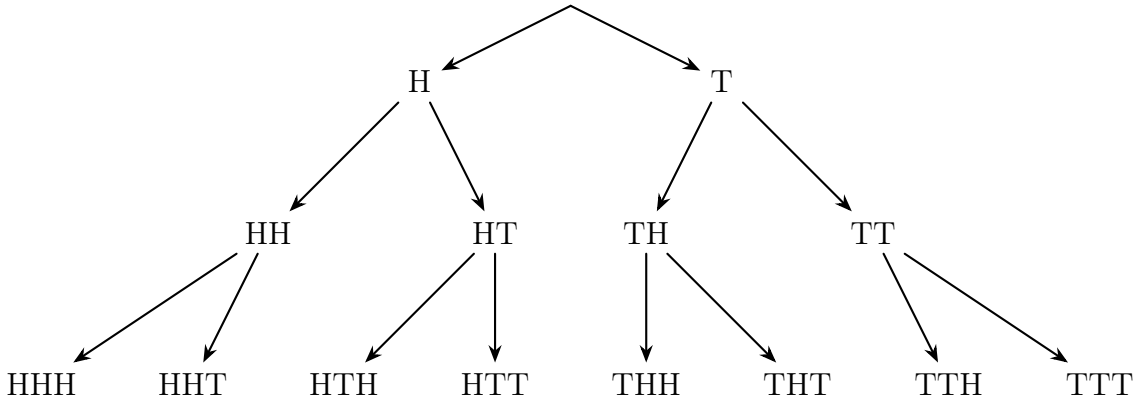
3.1　コインを3 回投げる

コインを3 回投げるとしましょう。

演習2

(a) ちょうど2 回裏が出る確率は何か。
(b) ちょうど1 回裏が出る確率は何か。

この標本空間では各結果が等確率なので，ここでは単純にそれぞれの場合について有利な結果の数を数えるのが最善の方法です。もちろん，必要ならこれも木として整理できます。あとで便利なように，この木は実験の始まりが上に来るように描いておきます。



演習2 の解答

上の木から分かるように，3 回のコイン投げの標本空間は大きさ8 の集合です。ここではすべての結果が等確率なので，直接数え上げれば確率を計算できます。

(a) ちょうど2 回裏：この事象に有利な結果は3 つ，すなわち$\{HTT, THT, TTH\}$ なので，

$$P(\text{ちょうど2 回裏}) = \frac{3}{8}.$$

(b) ちょうど1 回裏：この事象に有利な結果は3 つ，すなわち$\{THH, HTH, HHT\}$ なので，

$$P(\text{ちょうど1 回裏}) = \frac{3}{8}.$$

3.2　n 回のコイン投げで裏を数える

前の例を見ると，コイン投げに関する確率の計算はとても簡単だと思えるかもしれません。し

かし，回数が増えるにつれて状況はかなり複雑になります。まず，上の確率木の各段の大きさが時間とともに増えていくことに注意しましょう。1 段目には2 つの結果，2 段目には4 つの結果，3 段目には8 つの結果があります。もう1 回コインを投げると，すでにある各H とT の列に対してさらに2 つの結果が生じます。したがって可能な結果の数は再び2 倍になり，4 段目には16 個の結果が現れます。つまり，4 回のコイン投げのあり方は16 通りあります。一般に，コインをn 回投げると，全結果数は2^n になります。

では，n 回投げたあとに特定の枚数の裏，たとえばちょうどr 回の裏を数えたいとしましょう。原理的には，前と同じように，ちょうどr 回裏がある結果の数を数え，それを2^n で割ればよいのです。しかし，この作業はすぐに大変になります。たとえば，6 回のコイン投げの結果のうち，ちょうど2 回裏であるものはいくつあるでしょうか。これは，H とT の64 通りの結果の中から，T が2 個，H が4 個あるものを探し出す作業になります。[2]

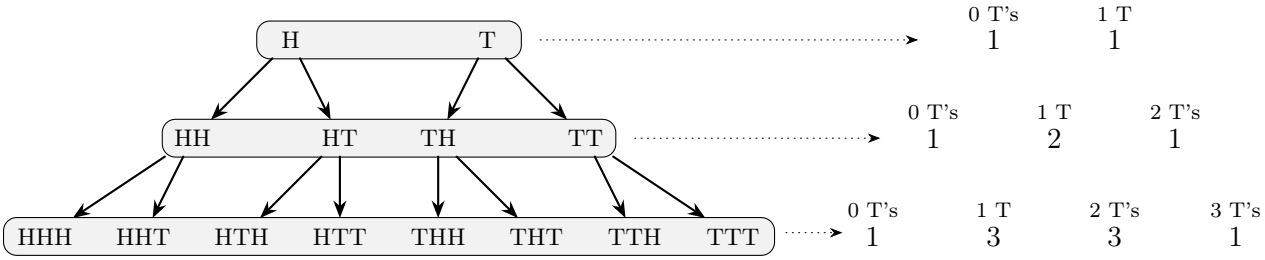
実際には，H とT を大量に書き出してそれを読み取らなくても，ちょうどr 回裏となる結果の数を数えられる巧妙な方法があると便利です。実はそのような方法はあります。しかも，それを私たちはすでに見ています。

3.2.1 n 回の投げでちょうどr 回裏を数える

この方法を説明するために，もう一度上の木を眺め，そこに現れる結果を数えてみましょう。

- 1 回投げ：コインを1 回投げると，結果集合は$\{H,T\}$ です。T を0 個含む結果が1 つ，T を1 個含む結果が1 つあります。
- 2 回投げ：2 回投げると結果は$\{HH,HT,TH,TT\}$ です。T を0 個含む結果は1 つ（HH），T を1 個含む結果は2 つ（HT とTH），T を2 個含む結果は1 つ（TT）あります。
- 3 回投げ：このとき結果集合の大きさは8 で，上の木の最下段に当たります。この数え上げはすでに行いました。T を0 個含む結果は1 つ（HHH），T を1 個含む結果は3 つ（HHT,THH,HTH），T を2 個含む結果は3 つ（HTT,THT,TTH），T を3 個含む結果は1 つ（TTT）です。

これをもっと示唆的にするために，木の横に「T の個数カウンター」を書いてみます。



ご覧のように，図の右側にある三角形はパスカルの三角形の最初の数行にとてもよく似ています。実際，これはまさに パスカルの三角形です。なぜでしょうか。コイン投げでちょうどr 回裏が出る結果を数えることは，「n 個の位置があって，そのうちr 個を裏として選ぶ方法をすべて見たい」と言っているのと同じだからです。

[2] これが楽だと感じるなら，$n=10$，$r=4$ を考えてみてください。

パスカルの三角形を，「何個の裏があり得るか」という注記つきで描き直せば，面倒な数え上げをしなくても数を読み取ることができます。

$$\begin{array}{ccccccccccccc} & & & & & & {}^{0\text{ T's}}_{1} & & & & & & \\ & & & & & {}^{0\text{ T's}}_{1} & & {}^{1\text{ T}}_{1} & & & & & \\ & & & & {}^{0\text{ T's}}_{1} & & {}^{1\text{ T}}_{2} & & {}^{2\text{ T's}}_{1} & & & & \\ & & & {}^{0\text{ T's}}_{1} & & {}^{1\text{ T}}_{3} & & {}^{2\text{ T's}}_{3} & & {}^{3\text{ T's}}_{1} & & & \\ & & {}^{0\text{ T's}}_{1} & & {}^{1\text{ T}}_{4} & & {}^{2\text{ T's}}_{6} & & {}^{3\text{ T's}}_{4} & & {}^{4\text{ T's}}_{1} & & \\ & {}^{0\text{ T's}}_{1} & & {}^{1\text{ T}}_{5} & & {}^{2\text{ T's}}_{10} & & {}^{3\text{ T's}}_{10} & & {}^{4\text{ T's}}_{5} & & {}^{5\text{ T's}}_{1} & \\ {}^{0\text{ T's}}_{1} & & {}^{1\text{ T}}_{6} & & {}^{2\text{ T's}}_{15} & & {}^{3\text{ T's}}_{20} & & {}^{4\text{ T's}}_{15} & & {}^{5\text{ T's}}_{6} & & {}^{6\text{ T's}}_{1} \end{array} \qquad \begin{array}{l} 2^0 = 1 \text{ 通り} \\ 2^1 = 2 \text{ 通り} \\ 2^2 = 4 \text{ 通り} \\ 2^3 = 8 \text{ 通り} \\ 2^4 = 16 \text{ 通り} \\ 2^5 = 32 \text{ 通り} \\ 2^6 = 64 \text{ 通り} \end{array}$$

パスカルの三角形を使うほうが，H とT のすべての組合せを書き出すよりずっと簡単です。たとえば6 回のコイン投げでちょうど2 回裏が出る確率を求めるには，三角形の$n = 6$ の行で$r = 2$ の項を見れば，ちょうど2 回裏になる結果が15 通りあることが分かります。そして全結果数は$2^6 = 64$ です。したがって，これらをすぐに数え上げの公式に代入して

$$P(\text{6 回投げてちょうど2 回裏}) = \frac{\text{6 回の投げでちょうど2 回裏になる方法の数}}{\text{6 回投げたときの結果の総数}} = \frac{15}{64} \approx 23.4\%$$

と結論できます。

パスカルの三角形の各項は$C_{n,r}$ です。一般に，この「組合せを数える」方法を使うと，n 回のコイン投げでちょうどr 回裏が出る確率の一般公式

$$P(n \text{ 回投げてちょうど} r \text{ 回裏}) = \frac{n \text{ 回の投げでちょうど} r \text{ 回裏になる方法の数}}{n \text{ 回投げたときの結果の総数}} = \frac{C_{n,r}}{2^n}$$

を得ます。

もちろん，表が出る確率と裏が出る確率は同じなので，上の公式は対称的であり，n 回の投げでちょうどr 回表が出る確率にもそのまま使えます。

4　おまけ：誕生日のパラドックス

A を「部屋の中に誕生日が同じ2 人がいる」という事象とします。このとき，確率$P(A)$ は部屋にいる人数n によって大きく変わります。たとえば$n = 2$ なら，この2 人の誕生日が同じである確率は極めて小さいはずです。しかし$n > 365$ なら，誕生日は365 通りしかないので，部屋の中に誕生日が同じ2 人がいることは必ず起こります。このように，n がとても小さいときは$P(A)$ も小さく，n が大きくなると$P(A)$ はついには1 になります。ここで次の問いを考えます。

どのn において，$P(A)$ は50% を超えるのか。

言い換えると，部屋の中に何人いれば，「2 人が同じ誕生日である」ことのほうが，そうでないことより起こりやすくなるのでしょうか。実は答えは驚くべきもので，必要なのはたった23 人です。

この驚くべき事実は「誕生日のパラドックス」と呼ばれます。これは，この講義の前のほうで扱った床屋のパラドックスのような意味でのパラドックスではありません。あちらでは，真と偽が無限ループして論理の働きそのものが破綻しました。それに対して誕生日のパラドックスは，「真なる逆説（veridical paradox）」の一例です。つまり，数学的には正しい結果なのに，あまりに意外なので間違っているように感じられる ものです。しかしこれから示すように，誕生日のパラドックスは確率の基本的な応用からきちんと導かれます。

4.1 誕生日のパラドックスを解く

誕生日のパラドックスを解くために，問題の見方をひっくり返して考えます。$P(A)$ を直接計算する代わりに，A の補事象の確率$P(A^c)$ を計算します。補事象は確率論における「not」に対応するものでした。ここでは補事象A^c は「部屋の中の全員の誕生日がすべて異なる」という意味です。

補事象の確率は単に$1 - P(A)$ であることを思い出しましょう。したがって，もし$P(A^c) < 50\%$ なら，$P(A) > 50\%$ です。なぜなら

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

だからです。各誕生日は等確率で選ばれると仮定し，どの誕生日も確率1/365 を持つとしましょう。まずは小さなn について$P(A^c)$ を計算して，どのような過程になるのかを見ます。

$n = 2$ の場合，求めたいのは2 人の誕生日が異なる確率です。明らかに1 人目の誕生日は自由に選べます。しかし2 人の誕生日が異なるためには，2 人目の誕生日は1 人目の誕生日を避けなければなりません。1 人目の誕生日と異なる日は364 日あるので，（これをランダムな事象として扱うなら）2 人目の誕生日が異なる確率は

$$P(A^c) = \frac{\text{1 人目の誕生日と異なる日の数}}{\text{1 年の日数の総数}} = \frac{364}{365} \approx 0.997 = 99.7\%$$

となります。

次に3 人目を部屋に加える，すなわち$n = 3$ の場合を考えましょう。2 人の誕生日が異なる確率はすでに$\dfrac{364}{365}$ であることが分かっています。しかしさらに，3 人目の誕生日が1 人目と2 人目の両方の誕生日と異なっていなければなりません。この次の人の誕生日をランダムに選ぶことは条件付き の事象です。なぜなら，その確率は前の選択の結果に依存するからです。このような「and」の確率を計算するには，対応する確率を掛け合わせます。

$$\begin{aligned} P(A^c) &= \frac{364}{365} \times \frac{\text{1 人目と2 人目の誕生日の両方と異なる日の数}}{\text{1 年の日数の総数}} \\ &= \frac{364}{365} \cdot \frac{363}{365} \approx 0.992 = 99.2\%. \end{aligned}$$

このパターンはさらに大きなn に対しても繰り返せます。n を1 増やすたびに，$(n+1)$ 人目が選べる誕生日の候補を少しずつ「使い切って」いくことになります。したがって，各段階でだんだん小さい数を掛けていかなければなりません。たとえば

$$n=4 \text{ のとき：} \quad P(A^c)=\left(\frac{364}{365}\right)\left(\frac{363}{365}\right)\left(\frac{362}{365}\right)\approx 0.983=98.3\%$$

$$n=5 \text{ のとき：} \quad P(A^c)=\left(\frac{364}{365}\right)\left(\frac{363}{365}\right)\left(\frac{362}{365}\right)\left(\frac{361}{365}\right)\approx 0.973=97.3\%$$

$$n=6 \text{ のとき：} \quad P(A^c)=\left(\frac{364}{365}\right)\left(\frac{363}{365}\right)\left(\frac{362}{365}\right)\left(\frac{361}{365}\right)\left(\frac{360}{365}\right)\approx 0.960=96.0\%$$

となります。

一般のn に対しては，

$$P(A^c)=\left(\frac{365}{365}\right)\left(\frac{364}{365}\right)\left(\frac{363}{365}\right)\cdots\left(\frac{365-(n-1)}{365}\right)$$

という公式を得ます。

このパラドックスは，確率$P(A^c)$ が50% を下回る最小のn を見つければ解決されます。なぜなら，それは$P(A)$ が50% を上回る ことを意味するからです。公式を使ってn を大きくしていくと，$n=23$ のとき

$$P(A^c)=\left(\frac{365}{365}\right)\left(\frac{364}{365}\right)\left(\frac{363}{365}\right)\cdots\left(\frac{342}{365}\right)\approx 0.490<0.5$$

となることが分かります。したがって，$n=23$ のとき$P(A)>50\%$ です。