

『プログラミングのための確率統計』

(ISBN : 978-4-274-06775-4)

補足編

October 19, 2009 版

©2009 平岡和幸・堀玄

目次

目次	補足編 1
本編の旗マークの解説箇所	補足編 4
補足編 1 第 1 章の補足	補足編 5
1.a 三つ組 (Ω, \mathcal{F}, P) の詳細	補足編 5
1.b 面積や体積 (にあたるもの) を定められない図形	補足編 5
1.c \mathcal{F} は何者か?	補足編 5
1.c.1 エラー判定関数	補足編 5
1.c.2 確率論のための要求仕様	補足編 6
1.c.3 集合族への翻訳	補足編 6
1.c.4 \mathcal{F} の使い道	補足編 7
1.d 1192 年 6 月 6 日にここが雨だった確率	補足編 8
補足編 2 第 2 章の補足	補足編 9
2.a 同時確率と周辺確率の練習用例題	補足編 9
2.b 条件つき確率の練習用例題	補足編 12
2.c 独立性についての補足	補足編 16
2.c.1 言いかえの同値性など	補足編 16
2.c.2 無限にまつわるやっかいごと	補足編 19
2.c.3 かかわりの度合い	補足編 19
補足編 3 第 3 章の補足	補足編 20
3.a (独立な確率変数たちの) 合計値や最大値の分布	補足編 20
3.a.1 (独立な確率変数たちの) 合計値の分布	補足編 20
3.a.2 (独立な確率変数たちの) 最大値の分布	補足編 22
3.b 2 項分布のプログラムを書く際の落とし穴	補足編 23
3.c 期待値についての補足	補足編 24
3.c.1 記法	補足編 24
3.c.2 平均と期待値の違い	補足編 24
3.c.3 練習用例題	補足編 24
3.c.4 期待値が存在しない場合への深入り	補足編 26
無限和の微妙な話	補足編 26
期待値と名乗るための条件	補足編 26
3.d 分散についての補足	補足編 27
3.e 大数の法則についての補足	補足編 27
3.e.1 正確な記述 (弱法則と強法則)	補足編 27
3.e.2 成り立たない例	補足編 28
シミュレーション	補足編 28
見積り計算	補足編 29
裾の重さ	補足編 30
3.f その他の補足	補足編 30
3.f.1 すべての自然数に対する一様分布?	補足編 30
3.f.2 条件つき期待値の一般化	補足編 30
補足編 4 第 4 章の補足	補足編 31
4.a 面積ゼロが納得できない読者へ	補足編 31
4.b 累積分布関数	補足編 32
4.c 典型的でない場合	補足編 33
4.c.1 典型的でない場合の累積分布関数	補足編 33
4.c.2 典型的でない場合の確率密度関数	補足編 34
4.d 一変数の確率分布についての補足	補足編 36
4.d.1 定義と記法	補足編 36
4.d.2 累積分布関数の変換	補足編 36
4.d.3 すなおでない変数変換	補足編 37
4.e 多変数の確率分布についての補足	補足編 38
4.e.1 実数値の同時分布とは結局何か	補足編 38
4.e.2 不正確だった点	補足編 38
4.e.3 同時分布の累積分布関数	補足編 39
4.e.4 独立性への補足	補足編 39
4.e.5 すなおでない変数変換	補足編 40
折り返し	補足編 40
べちゃんこ	補足編 41
次元を減らす	補足編 41
次元を増やす	補足編 41
4.f (独立な確率変数たちの) 合計値や最大値の分布	補足編 42
4.f.1 (独立な確率変数たちの) 合計値の分布	補足編 42
4.f.2 (独立な確率変数たちの) 最大値の分布	補足編 42
4.g 正規分布についての補足	補足編 42

4.g.1	記法の濫用	補足編 42
4.g.2	正規分布の確率密度関数の積分 (誤差関数 erf)	補足編 43
4.g.3	独立な正規分布の足し算	補足編 44
4.g.4	中心極限定理	補足編 44
4.g.5	中心極限定理の証明にかえて	補足編 46
4.g.6	連続修正	補足編 47
補足編 5 第 5 章の補足		補足編 48
5.a	相関係数についてのこまごま	補足編 48
5.b	期待値ベクトルと共分散行列の演算	補足編 48
5.b.1	右から行列をかけると期待値ベクトルは.....	補足編 48
5.b.2	ベクトルを足すと共分散行列は.....	補足編 48
5.c	多次元正規分布についての補足	補足編 49
5.c.1	多次元正規分布の変換について	補足編 49
5.c.2	多次元正規分布の確率密度関数への補足	補足編 50
5.c.3	切口と影の計算の詳細	補足編 50
	切口 (条件つき分布) について	補足編 50
	(参考) ブロック正方行列の逆行列	補足編 51
	影 (周辺分布) について	補足編 51
5.c.4	正則でない行列をかけると	補足編 51
	多次元正規分布になる場合	補足編 51
	縮退してしまう場合	補足編 52
5.d	楕円に関連した話	補足編 52
5.d.1	楕円について	補足編 52
5.d.2	自乗誤差の期待値についての公式	補足編 53
5.d.3	共分散行列の非負定値性	補足編 53
	非負定値対称行列	補足編 53
	共分散行列は非負定値	補足編 54
	正定値対称行列	補足編 54
5.d.4	前提についての釈明	補足編 54
補足編 6 第 6 章の補足		補足編 55
6.a	推定論についての補足	補足編 55
6.a.1	枠組への補足	補足編 55
6.a.2	不偏分散はなぜ (サンプルサイズ - 1) で割るのか	補足編 55
	数式による説明	補足編 55
	図による説明	補足編 55
6.a.3	UMVUE の難点	補足編 56
	資格を満たす参加者が誰もいない例	補足編 56
	UMVUE への批判	補足編 56
6.b	検定論についての補足	補足編 57
6.b.1	枠組への補足	補足編 57
6.b.2	各事項について	補足編 57
	ネイマン・ピアソンの補題への細かい補足	補足編 57
	一様最強力不偏検定	補足編 57
	尤度比検定	補足編 57
補足編 7 第 7 章の補足		補足編 58
7.a	疑似乱数列についての補足	補足編 58
7.a.1	乱数の検定	補足編 58
7.a.2	モンテカルロ法の意義	補足編 58
7.b	所望の分布の生成についての補足	補足編 58
7.b.1	こまごま	補足編 58
7.b.2	確率密度関数を使う方法 (工夫版)	補足編 59
7.c	そもそも乱数列とは?	補足編 60
7.c.1	有限列に対して	補足編 60
7.c.2	無限列に対して	補足編 61
補足編 8 第 8 章の補足		補足編 63
8.a	最小自乗法についての補足	補足編 63
8.a.1	こまごました補足	補足編 63
8.a.2	チコノフの正則化の導出	補足編 63
8.b	主成分分析についての補足	補足編 64
8.b.1	全般的な補足	補足編 64
8.b.2	特異値分解の利用	補足編 64
8.b.3	自乗誤差の計算	補足編 64
8.c	ランダムウォークの練習用例題	補足編 65
8.d	多次元版のカルマンフィルタ	補足編 67
8.e	マルコフ連鎖についての補足	補足編 67
8.e.1	表記の不統一	補足編 67
8.e.2	多重マルコフ連鎖	補足編 68
8.e.3	過去も無限にしなければ	補足編 68

8.e.4	推移確率行列の性質（定常分布の存在）	補足編 68
8.e.5	極限分布と線形代数	補足編 69
8.e.6	初めて戻ってくるまでの時間と極限分布との関係	補足編 70
8.e.7	よくある誤解：マルコフ連鎖の時間を逆回しにしたら……	補足編 70
8.f	エントロピーについての補足	補足編 72
8.f.1	エントロピーのとり得る範囲	補足編 72
8.f.2	連続値版のエントロピー	補足編 73
8.g	情報源符号化定理の証明	補足編 74
8.g.1	くじ買占め問題への翻訳	補足編 74
8.g.2	最適戦略	補足編 74
8.g.3	これだけ買えば十分	補足編 74
8.g.4	少なくともこれだけは買わないと	補足編 76
8.h	通信路符号化定理の証明の概略	補足編 76
8.h.1	通信レートが容量より低い場合	補足編 77
8.h.2	通信レートが容量より高い場合	補足編 77
索引		補足編 78

本編 11 ページの [✕]	補足編 5
本編 21 ページの [✕]	補足編 5
本編 86 ページの [✕]	補足編 26
本編 106 ページの [✕]	補足編 27
本編 107 ページの [✕]	補足編 28
本編 125 ページの [✕]	補足編 31
本編 126 ページの [✕]	補足編 33
本編 129 ページの [✕]	補足編 34
本編 130 ページの [✕]	補足編 32
本編 130 ページの [✕]	補足編 36
本編 140 ページの [✕]	補足編 38
本編 166 ページの [✕]	補足編 44
本編 170 ページの [✕]	補足編 47
本編 170 ページの [✕]	補足編 46
本編 206 ページの [✕]	補足編 50
本編 207 ページの [✕]	補足編 50
本編 218 ページの [✕]	補足編 52
本編 254 ページの [✕]	補足編 60
本編 254 ページの [✕]	補足編 58
本編 260 ページの [✕]	補足編 58
本編 262 ページの [✕]	補足編 59
本編 277 ページの [✕]	補足編 63
本編 282 ページの [✕]	補足編 64
本編 282 ページの [✕]	補足編 53
本編 293 ページの [✕]	補足編 67
本編 298 ページの [✕]	補足編 68
本編 308 ページの [✕]	補足編 73
本編 312 ページの [✕]	補足編 74
本編 314 ページの [✕]	補足編 76

第1章の補足

1.a 三つ組 (Ω, \mathcal{F}, P) の詳細

本書では三つ組 (Ω, \mathcal{F}, P) の存在保証や構成法には立ち入りません。あなたが扱いたい問題をシミュレートできる (Ω, \mathcal{F}, P) が本当に存在するのか、気になる方は確率論のきちんとした数学書を勉強してください（たとえば参考文献 [35]）。

P について、「面積や体積と同じような性質を持つ」とは、具体的には次の条件を満たすという意味です。

1. P の値は常に 0 以上。特に、空集合 $\emptyset = \{\}$ に対しては $P(\emptyset) = 0$ 。
2. 集合 A_1 , 集合 A_2 , 集合 A_3, \dots がどのペアをとっても共通部分を持たない ($i \neq j$ のとき $A_i \cap A_j = \emptyset$) なら、和集合 $A_1 \cup A_2 \cup A_3 \cup \dots$ に対して

$$P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$$

こんな P のことを**測度**と呼びます（正確な定義は本格的な数学書を参照）。確率に関してはさらに、 $P(\Omega) = 1$ という前提もありました。 $P(\Omega) = 1$ であるような測度 P を**確率測度**と呼びます。

このあたりの背景思想としては、今どきのプログラマなら **duck typing** を思い浮かべてもらうのがいいかもしれません。「アヒルのように歩いてアヒルのように鳴くなら、それはアヒルだ」というのが **duck typing** のスローガンでした。これと同じ精神で、「 Ω や P の実体は何だろうとどうでもいい。 Ω の部分集合に対して P で数が得られ、ある性質が満たされるなら、それは確率だ（とみなして不都合ない）」というわけです。

\mathcal{F} については節を改めて述べます。厳密に言うとも 1.4 節 (p.12) の確率変数の定義にも \mathcal{F} にかからんだ資格制限があるのですが、本書ではそこは踏み込みません。

1.b 面積や体積（にあたるもの）を定められない図形

「 \mathcal{F} は何者か」を述べる準備として、 \mathcal{F} が必要とされる事情を先にお話しておきます。それは、面積や体積（にあたるもの）を定められない図形が存在することです。

簡単なほうから順に見ていくと、まず、パラレルワールド全体 Ω が有限集合の場合。この場合は何も問題ありません。各世界 ω の確率 p_ω が設定されれば、 Ω のどんな部分集合 A でも、

$$P(A) = \text{「} A \text{ に属する各世界 } \omega \text{ について確率 } p_\omega \text{ を調べ、それを合計した値」}$$

という形で $P(A)$ を定めることができます（この $P(A)$ は面積ではありませんが、つい先ほど述べた条件を満たすという意味で「面積や体積にあたるもの」です）。

次に、 Ω が無限集合でも、最低レベルの無限集合だった場合。具体的には、 $\Omega = \{0, 1, 2, 3, \dots\}$ や $\Omega = \{\dots, -2, -1, 0, 1, 2, \dots\}$ のような場合です（付録 A.3.2 (p.320) の可算集合）。無限集合でもこの程度なら有限集合のときと同様の扱いが通用して、どんな部分集合 A でも $P(A)$ を定められます。

問題は、 Ω がワンランク上の無限集合だった場合です。たとえば、

$$\Omega = \text{「} 0 \text{ 以上 } 1 \text{ 以下のすべての実数の集合」}$$

$$\Omega = \text{「} \text{正方形領域内のすべての点の集合」}$$

などです。こんな Ω にはたちの悪い部分集合 A が存在していて、 $P(A)$ をどう設定したとしても何かしら辻褄が合わなくなってしまう（「面積などと呼ぶなら当然満たしてほしい性質」が保てなくなるという意味）。

そういう部分集合 A の具体的な説明は、申し訳ありませんが本書の目標レベルを越えています。 A は絵には描けません。——そもそも、描けるような単純な形だったら面積を定めるのに困ったりはしません。白と黒が無限に細かく入り混じっていて、図形という言葉のイメージにそぐわないような集合だからこそ、妙な現象が生じてしまうのです。無限集合ならではの現象なので、直観的な説明はお手上げです。あきらめて信じるか、測度論のちゃんとした数学書を読むかしてください。**選択公理**がからんでくると言ったら、数学マニアはちょっとわくわくするかもしれませんね。ついでにバナッハ・タルスキのパラドックスも調べてみると、興味深い話がみつかるでしょう。

このあたりの無限にまつわる話は 4.2 節 (p.123) 「確率ゼロ」も参照。

1.c \mathcal{F} は何者か？

さて、話を戻します。三つ組 (Ω, \mathcal{F}, P) の二番手 \mathcal{F} は何者か。答は、「 P で面積を測れるような集合」の集合です。と一息で言われても飲み込みづらいでしょうから、もう少し砕いて、プログラミングにたとえながら説明します。

1.c.1 エラー判定関数

集合 A （正確には Ω の部分集合 A ）を関数 P へ入力すると、答として数値 $P(A)$ が出力されるはずでした。しかし場合によっては、 P でうまく数値を定めることができないかもしれません。前の 1.b 節 (p.補足編 5) で述べた事情からどうしてもこれは避けられないので、あきらめてください。ゼロ割り 1/0 がエラーになったり、負数の平方根 `sqrt(-1)` がエラーになったりするのと同じようなものです。

とはいえ、プログラムが実行中に突然エラーで止まってしまうのも困ります。エラーを防ぎたいければ、 A を P へ入力する前に、この A は許されるかという判定をしてやらないといけません。平方根の例でいえば、

```
if (sqrt_ok?(x))
  answer = sqrt(x)
```

```

else
  # まずい x だったときの処理をここに
end

```

のような手順をとるわけです^{*1}。そのためには、判定関数 `sqrt_ok?(x)` を定義しておく必要があります^{*2}。

```

def sqrt_ok?(x)
  if (x >= 0)
    return true
  else
    return false
  end
end

```

これと同じように、 P についても、エラーにならないかを前もって判定する関数 G を用意してもらうことにします。

- $G(A) = \text{"OK"} \rightarrow P(A)$ はちゃんと数値が得られる
- $G(A) = \text{"NG"} \rightarrow P(A)$ はエラー

この G を別の形で表現したものが、 (Ω, \mathcal{F}, P) の二番手 \mathcal{F} にあたります。そのことはまた後でくわしく。

1.c.2 確率論のための要求仕様

前項の妥協により、「 P はエラーになってもよい。そのかわり、エラーにならないかを前もって判定する関数 G を用意してくれ」という話になりました。ただ我々としては、 P を使って確率論をしたいという本来の目的を忘れるわけにはいきません。この目的のために、三点、釘をさしておくべきことがあります。

一つめは、

$$G(\Omega) = \text{"OK"}$$

という要請です。そもそも $P(\Omega) = 1$ という前提でしたから、これでエラーが出ては困ります。

二つめは、ある確率 $P(\bigcirc\bigcirc)$ が測れるのなら、

$$P(\bigcirc\bigcircでない) = 1 - P(\bigcirc\bigcirc)$$

も測れてほしいという要請です。 G で言い直せば、

$$G(A) = \text{"OK"} \quad \text{なら} \quad G(\Omega \setminus A) = \text{"OK"}$$

ということです^{*3}。

三つめは、確率 $P(\text{条件 1}), P(\text{条件 2}), P(\text{条件 3}), \dots$ が測れるときは、

$$P(\text{条件 1 または条件 2 または条件 3 または} \dots)$$

も測れてほしい。これを言い直すと、

$$G(A_1) = G(A_2) = G(A_3) = \dots = \text{"OK"} \quad \text{なら}$$

$$\text{それらの和集合についても} \quad G(A_1 \cup A_2 \cup A_3 \cup \dots) = \text{"OK"} \quad (1.1)$$

となります。

言いたいことは以上でおしまい。これだけあれば確率論ができます。残る宿題は、 G を \mathcal{F} に翻訳することだけです。

1.c.3 集合族への翻訳

ここまでは、頭の混乱を避けるために、判定関数 G という言い方で要請を述べました。でも実際の測度論では G は使いません。かわりに集合 \mathcal{F} を用いて同じことを表現します。具体的には、「 $G(A) = \text{"OK"}$ となるような A 」の集合を \mathcal{F} とおけばよい。こうすれば、「 $G(\dots) = \text{"OK"}$ 」を「 \dots は \mathcal{F} に属する」と言い直すだけで、すべての話を \mathcal{F} に翻訳できます。これが (Ω, \mathcal{F}, P) の二番手 \mathcal{F} です。

はじめから \mathcal{F} で話さなかった理由は、 \mathcal{F} が「集合の集合」という入門者泣かせなものになってしまうからです。 A というのは Ω の部分集合で、そんな A たちの集合が \mathcal{F} ですから、 \mathcal{F} は Ω の部分集合の集合。 \mathcal{F} にとっては A はただの要素（もの）なのに、 A 自身が実は集合（袋）でもある。この二つの見方がひよいひよい切りかわるせいで、どうしても入門者は混乱しがちです（→ 付録 A.3.3(p.321)「本気の数学に向けて」）。本書の目標のためには、そこまでがんばってこの方面を登らなくてもよからうと筆者は考えていますので、 \mathcal{F} について本気の説明はしません。登りたい読者は測度論を勉強してください。

そちらを狙う読者は、 **σ 加法族（シグマ加法族）** という言葉だけは気に留めておくとういでしょう。 G に関する先ほどの条件を \mathcal{F} に翻訳すれば、

- $\Omega \in \mathcal{F}$
- $A \in \mathcal{F}$ なら $\Omega \setminus A \in \mathcal{F}$
- $A_1, A_2, A_3, \dots \in \mathcal{F}$ なら $A_1 \cup A_2 \cup A_3 \cup \dots \in \mathcal{F}$

となります。この三条件を満たすような「『 Ω の部分集合』の集合」 \mathcal{F} を、一般に「 Ω の上の σ 加法族」と呼びます^{*4}。 (Ω, \mathcal{F}, P) の二番手 \mathcal{F} は σ 加法族でなくてはなりません。

余談ですが、数学では「集合の集合」のことを**集合族**と言ったりします。 \mathcal{F} も A_1, A_2, A_3, \dots も、どれも集合なのですが、外側である \mathcal{F} をあえて族という違う呼び名にすれば混乱が多少は緩和されることでしょう。

^{*1} ここではプログラミング言語 **Ruby** (<http://www.ruby-lang.org/ja/>) を説明に用いていますが、Ruby ユーザーでなくても読めば意味はつかめるでしょう。なお、Ruby のような今どきの言語には、**例外**という仕組みが備わっています。それを使えばこういう処理をもっとかつこよく扱えます。

^{*2} 本当は「`def sqrt_ok?(x); x >= 0; end`」だけでいいのですが、あえて初心者っぽく書き下しておきます。

^{*3} $\Omega \setminus A$ は、 Ω の要素のうち A に属さないものからなる集合（**差集合**）を表します。いまの場合 A は Ω の一部ですから、 $\Omega \setminus A$ とは要するに A の**補集合**です（ Ω を**全体集合**として）。

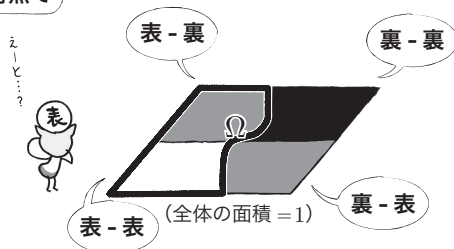
^{*4} ただし一つめの条件は、「全体集合 $\Omega \in \mathcal{F}$ 」のかわりに「空集合 $\emptyset \in \mathcal{F}$ 」を要請するほうが普通です。これはどちらでも同じこと（二つめの条件から両者は同値）。また、 σ 加法族は **σ 代数**、**可算加法族**、**完全加法族**、 **σ 集合体**とも呼ばれます。

1.c.4 \mathcal{F} の使い道

後向きな動機でしかたなく導入されたかのような \mathcal{F} ですが、せっかく導入するならもっと前向きな使い道は考えられないでしょうか。実は、情報や知識といったものを \mathcal{F} で表現するというテクニックがあります。確率過程 (8.2 節 (p.284)) についての少し進んだ教科書ではこのテクニックが多用されます。

図 1.1 のような神様視点について述べた際、人間は自分がどの世界 ω に住んでいるのか知覚できないとお話しました。

1 回目が見たと観測した時点で



► 図 1.1 人間は、自分がどの世界に住んでいるのかを知覚できない (図 1.5^(p.11) の再掲)

この状況を

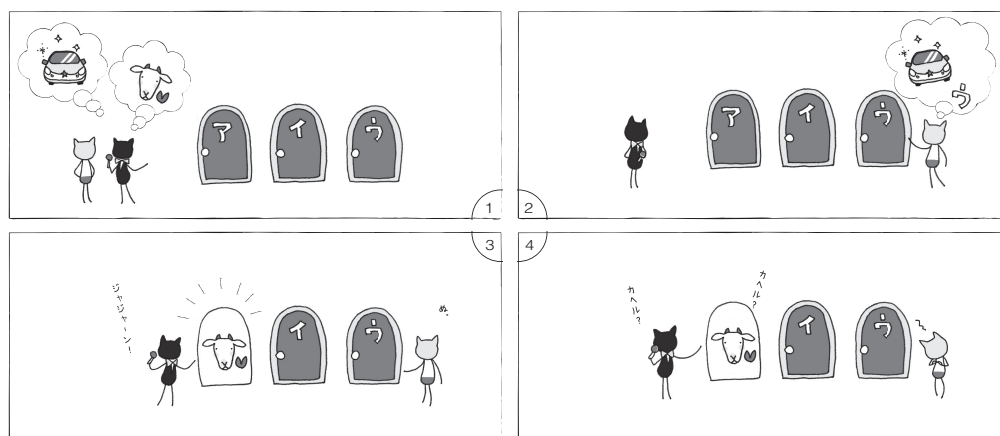
$$\mathcal{F}_0 = \{\emptyset, \Omega\}$$

で表すことにします。 Ω の部分集合 A で、「自分が A 内にいるのかいないのか」をはっきり確信できるのは、

- $A = \emptyset$ (絶対いないはず)
- $A = \Omega$ (絶対いるはず)

というあたりまえの場合しかない、という意味です。

さて、たとえば図 1.2 のモンティホール問題で、挑戦者がどの扉を選んだかを観測したとしましょう。



► 図 1.2 モンティホール問題 (図 1.1^(p.5) の再掲)。詳しくは 1.2 節 (p.4) を参照

すると、見分けられる部分集合 A のレパートリーが増えます。 \mathcal{F}_0 に加え、

- $A = \text{「挑戦者がアを選ぶ世界の集合 } B_{\text{ア}} \text{」}$
- $A = \text{「挑戦者がイを選ぶ世界の集合 } B_{\text{イ}} \text{」}$
- $A = \text{「挑戦者がウを選ぶ世界の集合 } B_{\text{ウ}} \text{」}$
- $A = B_{\text{ア}} \cup B_{\text{イ}}$
- $A = B_{\text{イ}} \cup B_{\text{ウ}}$
- $A = B_{\text{ウ}} \cup B_{\text{ア}}$

の場合について、 A 内にいるかないかが確信可能です。この状況を、

$$\mathcal{F}_1 = \{\emptyset, \Omega, B_{\text{ア}}, B_{\text{イ}}, B_{\text{ウ}}, B_{\text{ア}} \cup B_{\text{イ}}, B_{\text{イ}} \cup B_{\text{ウ}}, B_{\text{ウ}} \cup B_{\text{ア}}\}$$

で表します。

さらに、司会者がどの扉を開いたかも観測したら、見分けられる A のレパートリーがまた増えます。 \mathcal{F}_1 に加え、

- $A = \text{「司会者がアを開く世界の集合 } C_{\text{ア}} \text{」}$
- $A = \text{「司会者がイを開く世界の集合 } C_{\text{イ}} \text{」}$
- $A = \text{「司会者がウを開く世界の集合 } C_{\text{ウ}} \text{」}$
- $A = C_{\text{ア}} \cup C_{\text{イ}}$
- ...
- $A = B_{\text{ア}} \cap C_{\text{ウ}}$
- ...

といった場合が確信可能。この状況を、

$$\mathcal{F}_2 = \{\emptyset, \Omega, B_{\mathcal{A}}, \dots, C_{\mathcal{A}}, C_{\mathcal{I}}, C_{\mathcal{U}}, C_{\mathcal{A}} \cup C_{\mathcal{I}}, \dots, B_{\mathcal{A}} \cap C_{\mathcal{U}}, \dots\}$$

で表します。

情報・知識が増えるにしたがって見分けられるレパートリーが増えてゆき、 $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2$ のような包含関係となることに注目してください。いまの話によって、情報や知識の増加という現象を、集合の包含という純粋な数学の対象へ写しとれたこととなります。情報や知識という今一つはっきり定義しづかった概念が、これで数学として議論できるようになりました。

おかげで、たとえば「不正な挑戦者」という概念も数式で表現できるようになります。挑戦者が最終決断（扉を選び直すかそのままいくか）をする場面を考えましょう。このとき使える手掛りは、「挑戦者がどの扉を選んだか」「司会者がどの扉を開いてみせたか」だけしかありません。ですから、この二つの手掛りが同じなら、挑戦者は同じ決断をするはずで、そうすると、

- 挑戦者が選び直す決断する世界の集合 $D_{直}$
- 挑戦者がそのまま決断する世界の集合 $D_{ま}$

も \mathcal{F}_2 に属すはずで、もし属していないようだと、 \mathcal{F}_2 で区別できないはずの世界を区別している、つまり不正をしているということになります^{*5}。

こんなふうに数学に翻訳できれば、あとは純粋な数学に専念できます。

1.d 1192 年 6 月 6 日にここが雨だった確率

「1192 年 6 月 6 日にここが雨だった確率」を許すか許さないかは、数学ではなく立場の問題だと 1.1 節 (P³) で述べました。許す立場をとる人々は、確率は確信の度合だと考えます。確率は、観測される対象物の属性値ではなく、観測する側の属性値だというわけです。

……などと哲学に踏み込むのはおっかないので、話を数学よりに戻します。確信の度合をなぜ確率と解釈できるのか。根拠となる理屈はいろいろと提案されています。

たとえば、「もし〇〇なら 1 ドルもらえる券」の値段は何ドルが妥当と思うか、その値を $Q(\text{〇〇})$ としましょう。明日の競馬に対して誰かが仮に $Q(\text{馬アが勝つ}) < Q(\text{馬アが勝つ}) + Q(\text{馬イが勝つ})$ という値づけをした場合、この人は確実に損をしてしまいます。この人から「アが勝つ」を買って、さらに「アが勝つ」と「イが勝つ」をこの人へ売りつけることにより、結果がどう転んでもあなたのほうが得をするからです（アが勝つ場合、イが勝つ場合、他が勝つ場合、をそれぞれ考えてみてください）。一方、 $Q(\text{アが勝つ}) > Q(\text{アが勝つ}) + Q(\text{イが勝つ})$ という値づけの場合も、売り買いを逆にすれば同じことです。こんな事態を防ぐには $Q(\text{アが勝つ}) = Q(\text{アが勝つ}) + Q(\text{イが勝つ})$ とするしかありません。 $0 \leq Q(\text{〇〇}) \leq 1$ や $Q(\text{必ず起きる}) = 1$ も同様です。そういうわけで結局、確率と全く同じ格好の規則が得られます。もっと知りたい方は **Dutch book** や **コヒーレンス (coherence)** というキーワードを調べてください（たとえば参考文献 [16]）。

あるいは、次のようなまた別の理屈もあります。「1192 年 6 月 6 日が雨だった」のような命題 x, y, z, \dots に対して、「 z が真だとしたときの x の確信度」を $B(x|z)$ とおきましょう。ここで次のような前提を導入します（**Cox の公理**）。

- $B(x|z)$ は実数値である
- ある関数 f が存在し、 $B(\text{“}x \text{でない”}|z) = f(B(x|z))$ が常に成り立つ
- ある関数 g が存在し、 $B(\text{“}x \text{かつ } y\text{”}|z) = g(B(x|\text{“}y \text{かつ } z\text{”}), B(y|z))$ が常に成り立つ

びんとかなければ、 z として特に「必ず真である命題」をとった場合を考えてみてください。上の主張は、

- 確信度は実数で表される
- 「 x でない」の確信度は x の確信度から定まる
- 「 x かつ y 」の確信度は、 y の確信度と「 y が真だとしたときの x の確信度」から定まる

ということで、それなりにもっともらしく聞こえます。

実は、命題についての論理規則とこれらの公理とを組み合わせれば、確率と互換なものが得られます（**Cox の定理**）。つまり、 B を適当に変換したものが確率と解釈できるのです。詳しくは参考文献 [27] など。

^{*5} 「その場でコイントスをして決める」のような確率的な戦略は除き、確定的な戦略に限って話をしています。本当は、枠組を少し拡張して確率的な戦略も議論できるのですが、そこまでは説明しません。

第2章の補足

2.a 同時確率と周辺確率の練習用例題

例題 2.1

X, Y はどちらも、 $+1$ か -1 かの値をとる確率変数とする。周辺分布が

$$P(X = -1) = P(X = +1) = P(Y = -1) = P(Y = +1)$$

となるような同時分布の例を 2 種類以上作れ。

答

	$X = -1$	$X = +1$
$Y = -1$	1/4	1/4
$Y = +1$	1/4	1/4

	$X = -1$	$X = +1$
$Y = -1$	1/2	0
$Y = +1$	0	1/2

	$X = -1$	$X = +1$
$Y = -1$	0.3	0.2
$Y = +1$	0.2	0.3

など。

例題 2.2

2.3.4 項^(p.47)の三つの扉で、次の式は成り立っているか？

$$P(Z = \text{ア}) = P(Y = \text{ア}, Z = \text{ア}) + P(Y = \text{イ}, Z = \text{ア}) + P(Y = \text{ウ}, Z = \text{ア})$$

答

成り立つ。右辺はそれぞれこうですから、ぜんぶ合計すれば、本文で計算した $P(Z = \text{ア})$ と同じです。

$$\begin{aligned} &P(Y = \text{ア}, Z = \text{ア}) \\ &= P(X = \text{ア}, Y = \text{ア}, Z = \text{ア}) + P(X = \text{イ}, Y = \text{ア}, Z = \text{ア}) + P(X = \text{ウ}, Y = \text{ア}, Z = \text{ア}) \\ &P(Y = \text{イ}, Z = \text{ア}) \\ &= P(X = \text{ア}, Y = \text{イ}, Z = \text{ア}) + P(X = \text{イ}, Y = \text{イ}, Z = \text{ア}) + P(X = \text{ウ}, Y = \text{イ}, Z = \text{ア}) \\ &P(Y = \text{ウ}, Z = \text{ア}) \\ &= P(X = \text{ア}, Y = \text{ウ}, Z = \text{ア}) + P(X = \text{イ}, Y = \text{ウ}, Z = \text{ア}) + P(X = \text{ウ}, Y = \text{ウ}, Z = \text{ア}) \end{aligned}$$

例題 2.3

確率変数 X, Y について、 X は 1 か 2、 Y は 0 か 1 か 2 のいずれかの値をとるとし、同時分布は

$$P(X = i, Y = j) = c(2i + j), \quad (i = 1, 2 \text{ および } j = 0, 1, 2)$$

だとする。定数 c を答え、さらに周辺分布を求めよ。

答

同時確率の合計は 1 のはず。合計してみると、

$$\begin{aligned} &P(X = 1, Y = 0) + P(X = 1, Y = 1) + P(X = 1, Y = 2) \\ &\quad + P(X = 2, Y = 0) + P(X = 2, Y = 1) + P(X = 2, Y = 2) \\ &= 2c + 3c + 4c + 4c + 5c + 6c = 24c \end{aligned}$$

これが 1 になるためには $c = 1/24$ 。そのとき X の周辺分布は

$$\begin{aligned} P(X=1) &= P(X=1, Y=0) + P(X=1, Y=1) + P(X=1, Y=2) \\ &= \frac{2}{24} + \frac{3}{24} + \frac{4}{24} = \frac{9}{24} = \frac{3}{8} \\ P(X=2) &= P(X=2, Y=0) + P(X=2, Y=1) + P(X=2, Y=2) \\ &= \frac{4}{24} + \frac{5}{24} + \frac{6}{24} = \frac{15}{24} = \frac{5}{8} \end{aligned}$$

また、 Y の周辺分布は

$$\begin{aligned} P(Y=0) &= P(X=1, Y=0) + P(X=2, Y=0) \\ &= \frac{2}{24} + \frac{4}{24} = \frac{6}{24} = \frac{1}{4} \\ P(Y=1) &= P(X=1, Y=1) + P(X=2, Y=1) \\ &= \frac{3}{24} + \frac{5}{24} = \frac{8}{24} = \frac{1}{3} \\ P(Y=2) &= P(X=1, Y=2) + P(X=2, Y=2) \\ &= \frac{4}{24} + \frac{6}{24} = \frac{10}{24} = \frac{5}{12} \end{aligned}$$

なお、

$$\begin{aligned} P(X=1) + P(X=2) &= 1 \\ P(Y=0) + P(Y=1) + P(Y=2) &= 1 \end{aligned}$$

となっていることも確認してください。問われていなくても、こういう確認を習慣づけるようおすすめします。計算ミスのチェックにもなりますし、確率の意味を意識することにもつながるでしょうから。「えっ」という人は 1.8.2 項^(p.22)の「面積なんだから……」を復習。 ■

例題 2.4

0 か 1 かの値をとる確率変数 X, Y について、

$$P(X=1) = 0.7, \quad P(Y=1) = 0.6, \quad P(X=0, Y=0) = 0.1$$

のとき、 $P(X=1, Y=1)$ は？

答

書くのがめんどうなので、まず同時確率に記号をつける。

$$a \equiv P(X=0, Y=0), \quad b \equiv P(X=0, Y=1), \quad c \equiv P(X=1, Y=0), \quad d \equiv P(X=1, Y=1)$$

次のように、同時確率の合計は 1 のはず。

$$a + b + c + d = 1 \tag{2.1}$$

また、指定された条件より

$$P(X=1) = c + d = 0.7 \tag{2.2}$$

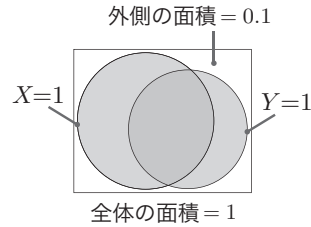
$$P(Y=1) = b + d = 0.6 \tag{2.3}$$

$$P(X=0, Y=0) = a = 0.1 \tag{2.4}$$

これらを連立 1 次方程式として解けば $d = 0.4$ が得られる^{*1}。

図 2.1 のようなベン図を描けば、もっと算数的にも解ける。全体の面積 1 から外側の面積 0.1 を引けば、アミがけの面積は 0.9。そこから左の円の面積 $P(X=1) = 0.7$ を引くと、右側の三日月型の面積 $P(X=0, Y=1) = 0.9 - 0.7 = 0.2$ が求まる。それを右の円の面積 $P(Y=1) = 0.6$ から引いて、 $0.6 - 0.2 = 0.4$ が³、求めたい領域の面積 $P(X=1, Y=1)$ 。

^{*1} $a = 0.1$ を使えば (2.1) から $b + c + d = 0.9$ がわかり、そこへ (2.2) と (2.3) を変形した $c = 0.7 - d$ と $b = 0.6 - d$ を代入することで、 $(0.7 - d) + (0.6 - d) + d = 0.9$ 。つまり $1.3 - d = 0.9$ だから、 $d = 0.4$ 。目が効く人なら「(2.2) と (2.3) と (2.4) を足して (2.1) を引けば d が出る」のほうが楽ですが。



▶ 図 2.1 ベン図を使って算数的に解く

例題 2.5

0 か 1 かの値をとる確率変数 X, Y, Z について、

$$\begin{aligned} P(X = 1) &= 0.7, & P(Y = 1) &= 0.6, & P(X + Y > 0) &= 0.9, \\ P(X + Y + Z = 2) &= 0.5, & P(Y = Z = 1) &= 0.2, & P(X = Y = Z = 1) &= 0.1 \end{aligned}$$

のとき、 $P(X = Z = 1)$ は？

答

図 2.2 のように、ベン図の各領域に相当する確率を表す。たとえば $g = P(X = 1, Y = 1, Z = 0)$ である。問題文で与えられているのは

$$b + f + g + h = 0.7 \quad (2.5)$$

$$c + e + g + h = 0.6 \quad (2.6)$$

$$b + c + e + f + g + h = 0.9 \quad (2.7)$$

$$e + f + g = 0.5 \quad (2.8)$$

$$e + h = 0.2 \quad (2.9)$$

$$h = 0.1 \quad (2.10)$$

であり、求めたいのは $f + h$ の値。(確率の前提から $a + b + c + d + e + f + g + h = 1$ も要請されるが、本問ではこれは使わなくても解ける)

まず (2.9) と (2.10) から $e = h = 0.1$ がわかる。すると残りの式は

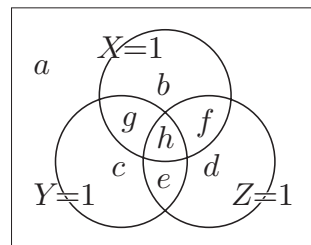
$$b + f + g = 0.6 \quad (2.11)$$

$$c + g = 0.4 \quad (2.12)$$

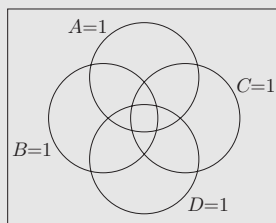
$$b + c + f + g = 0.7 \quad (2.13)$$

$$f + g = 0.4 \quad (2.14)$$

(2.11) と (2.13) とを見比べれば $c = 0.1$ 。よって (2.12) から $g = 0.3$ なので、(2.14) より $f = 0.1$ 。以上をあわせて $P(X = Z = 1) = f + h = 0.2$ 。

▶ 図 2.2 $X = 1, Y = 1, Z = 1$ に対するベン図**例題 2.6**

例題 2.5 のような問題の 4 変数版を解くために図 2.3 のような絵を描いたら怒られた。なぜこの絵はだめなのだろうか？



► 図 2.3 この絵がなぜだめなのか？

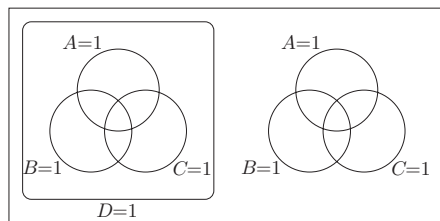
答

4 変数なら組合せは $2 \times 2 \times 2 \times 2 = 16$ 通りあるはずなのに、この絵には領域が 14 しかありません。抜けている組合せがあと 2 つあるはず。よく見ると、確かに

- $A = 1$ かつ $B = 0$ かつ $C = 0$ かつ $D = 1$
- $A = 0$ かつ $B = 1$ かつ $C = 1$ かつ $D = 0$

が抜けています。

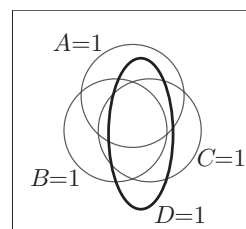
ではどう描けばよかったのか？ 一番すなおなのは、図を 2 次元から 3 次元に拡張して、4 つの球が重なりあった立体図形を想像することです。でも、この図形をしっかりと把握するには相当の空間認識能力が必要なようです（筆者には無理でした）。それよりは、図 2.4 のように $D = 1$ と $D = 0$ とを場合分けして図示するのはいいかでしょう。3 変数の図を二階建てにして、一階が $D = 1$ 、二階が $D = 0$ と思っても結構です。



► 図 2.4 4 変数の組合せの図示（二階建て方式）

? 2.a 4 変数版のベン図は描けないんですか？

巧妙にやれば図 2.5 のように描くこともできますが、あまり見やすいとは思えません。描き方は……まず 3 変数版を描いておき、「 $C = 1$ 」の円をなぞって一周しながらその中と外を行ったり来たりする、という方法が参考文献 [38] に載っています。



► 図 2.5 4 変数も描けなくはないが……

2.b 条件つき確率の練習用例題

例題 2.7

例題 2.1^(p.36) で、 $P(Y = \text{数札} | X = \text{黒})$ と $P(Y = \text{絵札} | X = \text{黒})$ を求め、

$$P(Y = \text{数札} | X = \text{黒}) + P(Y = \text{絵札} | X = \text{黒}) = 1$$

を確認せよ

答

同時分布の中で、 $X = \text{黒}$ という条件に該当する箇所は、次の二つ。

$$P(X = \text{黒}, Y = \text{数札}) = \frac{6}{16}, \quad P(X = \text{黒}, Y = \text{絵札}) = \frac{1}{16}$$

つまり、 $X = \text{黒}$ という条件を満たす領域に話を限定すれば、 $Y = \text{数札}$ な領域と $Y = \text{絵札}$ な領域との面積比は、数札 : 絵札 = $6/16 : 1/16 = 6 : 1$ 。だから、 $X = \text{黒}$ な領域のうち $6/7$ が $Y = \text{数札}$ 、 $1/7$ が $Y = \text{絵札}$ 。というわけで、 $P(Y = \text{数札} | X = \text{黒}) = 6/7$ 、 $P(Y = \text{絵札} | X = \text{黒}) = 1/7$ 、 $P(Y = \text{数札} | X = \text{黒}) + P(Y = \text{絵札} | X = \text{黒}) = 1$ が成り立ちます。 ■

例題 2.8

例題 2.1^(p.36) で、 $P(Y = \text{数札} | X = \text{赤})$ と $P(Y = \text{絵札} | X = \text{赤})$ を式 (2.7)^(p.40) により計算せよ。

答

前の算数に言い方をあわせれば、次のようになります。

- 赤黒の数札面積は $P(X = \text{赤}, Y = \text{数札}) = 3/16$ 、赤黒の絵札面積は $P(X = \text{赤}, Y = \text{絵札}) = 6/16$
- 赤黒の全面積は、

$$P(X = \text{赤}) = P(X = \text{赤}, Y = \text{数札}) + P(X = \text{赤}, Y = \text{絵札}) = \frac{3}{16} + \frac{6}{16} = \frac{9}{16}$$

- だから、赤黒全体の中で数札の割合は、

$$P(Y = \text{数札} | X = \text{赤}) = \frac{P(X = \text{赤}, Y = \text{数札})}{P(X = \text{赤})} = \frac{3/16}{9/16} = \frac{3}{9} = \frac{1}{3}$$

- 同様に、赤黒全体の中で絵札の割合は、

$$P(Y = \text{絵札} | X = \text{赤}) = \frac{P(X = \text{赤}, Y = \text{絵札})}{P(X = \text{赤})} = \frac{6/16}{9/16} = \frac{6}{9} = \frac{2}{3}$$

例題 2.9

例題 2.1^(p.36) で、 $P(Y = \text{数札} | X = \text{赤}) + P(Y = \text{絵札} | X = \text{赤}) = 1$ となる理由を数式で説明せよ。

答

$$\begin{aligned} & P(Y = \text{数札} | X = \text{赤}) + P(Y = \text{絵札} | X = \text{赤}) \\ &= \frac{P(X = \text{赤}, Y = \text{数札})}{P(X = \text{赤})} + \frac{P(X = \text{赤}, Y = \text{絵札})}{P(X = \text{赤})} \\ &= \frac{P(X = \text{赤}, Y = \text{数札}) + P(X = \text{赤}, Y = \text{絵札})}{P(X = \text{赤})} = \frac{P(X = \text{赤})}{P(X = \text{赤})} = 1 \end{aligned}$$

例題 2.10

例題 2.1^(p.36) で、

$$P(X = \text{赤}, Y = \text{絵札}) = P(X = \text{赤} | Y = \text{絵札}) P(Y = \text{絵札})$$

を確認せよ。

答

$$P(Y = \text{絵札}) = \frac{6}{16} + \frac{1}{16} = \frac{7}{16}, \quad P(X = \text{赤} | Y = \text{絵札}) = \frac{6/16}{7/16} = \frac{6}{7}$$

そのかけ算は $P(X = \text{赤} | Y = \text{絵札}) P(Y = \text{絵札}) = (6/7) \cdot (7/16) = 6/16$ で、 $P(X = \text{赤}, Y = \text{絵札})$ と一致。(全体の $7/16$ が絵札で、絵札の $6/7$ が赤だから、全体の $(6/7) \cdot (7/16)$ が「赤い絵札」) ■

例題 2.11

次の誤った主張に反論せよ。

例題 2.1^(p.36) で、 $X = \text{赤}$ となるには、 $Y = \text{絵札}$ で $X = \text{赤}$ になる場合と、 $Y = \text{数札}$ で $X = \text{赤}$ になる場合とがある。まず、 $Y = \text{絵札}$ で $X = \text{赤}$ になる確率は

$$P(X = \text{赤} | Y = \text{絵札}) = \frac{6}{7}$$

一方、 $Y = \text{数札}$ で $X = \text{赤}$ になる確率は

$$P(X = \text{赤} | Y = \text{数札}) = \frac{1}{3}$$

よって、 $X = \text{赤}$ になる確率は、その平均で

$$\frac{\frac{6}{7} + \frac{1}{3}}{2} = \frac{25}{42} \quad (2.15)$$

のはずだ。

答

$Y = \text{絵札}$ の確率と $Y = \text{数札}$ の確率とは、半々ではありません。前者が $7/16$ 、後者が $9/16$ というように偏っているのですから、(2.15) ではなく

$$\frac{6}{7} \cdot \left(\frac{7}{16}\right) + \frac{1}{3} \cdot \left(\frac{9}{16}\right) = \frac{6}{16} + \frac{3}{16} = \frac{9}{16}$$

のように計算する必要があります。こうすれば正しく

$$\begin{aligned} & P(X = \text{赤} | Y = \text{絵札}) P(Y = \text{絵札}) + P(X = \text{赤} | Y = \text{数札}) P(Y = \text{数札}) \\ &= P(X = \text{赤}, Y = \text{絵札}) + P(X = \text{赤}, Y = \text{数札}) \\ &= P(X = \text{赤}) \end{aligned}$$

を求めたことになります。

例題 2.12

次の誤った主張に反論せよ。

例題 2.1^(p.36) で、まず、 $Y = \text{絵札}$ で $X = \text{赤}$ になる確率は

$$P(X = \text{赤}, Y = \text{絵札}) = \frac{6}{16}$$

一方、 $Y = \text{数札}$ で $X = \text{赤}$ になる確率は

$$P(X = \text{赤}, Y = \text{数札}) = \frac{3}{16}$$

よって、 $X = \text{赤}$ になる確率は、その平均で

$$\frac{\frac{6}{16} + \frac{3}{16}}{2} = \frac{9}{32}$$

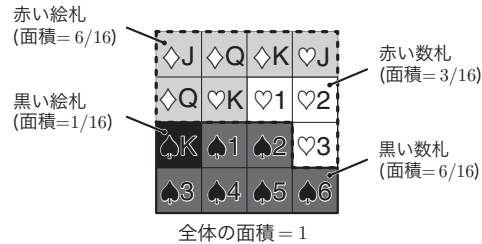
のはずだ。

答

$P(X = \text{赤}, Y = \text{絵札})$ や $P(X = \text{赤}, Y = \text{数札})$ は同時確率ですから、2 で割ったりせずただ足せば周辺確率が得られます。

$$\begin{aligned} & P(X = \text{赤}, Y = \text{絵札}) + P(X = \text{赤}, Y = \text{数札}) = P(X = \text{赤}) \\ & \frac{6}{16} + \frac{3}{16} = \frac{9}{16} \end{aligned}$$

図 2.6 を参照。



▶ 図 2.6 「絵札か数札かはどうでもいいから、とにかく赤である領域」の面積は $6/16 + 3/16$

例題 2.13

$P(X = \text{ア}) = 0.2$, $P(Y = \text{ア}) = 0.3$, $P(X = \text{ア} | Y = \text{ア}) = 0.4$, $P(Y = \text{ア} | X = \text{ア}) = 0.6$ のとき、 $P(X = \text{ア}, Y = \text{ア})$ を答えよ。

答

$$P(X = \text{ア}, Y = \text{ア}) = P(Y = \text{ア} | X = \text{ア}) P(X = \text{ア}) = 0.6 \cdot 0.2 = 0.12$$

もしくは、次式から求めてもよい。

$$P(X = \text{ア}, Y = \text{ア}) = P(X = \text{ア} | Y = \text{ア}) P(Y = \text{ア}) = 0.4 \cdot 0.3 = 0.12$$

例題 2.14

外見で見分けのつかないシュークリームが 3 個あります。そのうち 2 個は普通のクリーム入りですが、残る 1 個にはクリームでなくわさびが入っています。いま、一郎がどれか一個をでたらめに選んで食べ、残りのうちから二郎がまた一個でたらめに選んで食べ、最後の残りを三郎が食べることにしました。誰が一番不利でしょうか^{*2}。

答

まず、一郎については

$$P(\text{一郎がわさび}) = \frac{1}{3}$$

次に二郎は

$$\begin{aligned} P(\text{二郎がわさび}) &= P(\text{一郎がわさび}, \text{二郎がわさび}) + P(\text{一郎がクリーム}, \text{二郎がわさび}) \\ &= 0 + P(\text{二郎がわさび} | \text{一郎がクリーム}) P(\text{一郎がクリーム}) \\ &= \frac{1}{2} \cdot \left(1 - \frac{1}{3}\right) = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3} \end{aligned}$$

そして、 $P(\text{一郎がわさび}) + P(\text{二郎がわさび}) + P(\text{三郎がわさび}) = 1$ のはずですから、

$$P(\text{三郎がわさび}) = 1 - \frac{1}{3} - \frac{1}{3} = \frac{1}{3}$$

というわけで全員平等です。

例題 2.15

シュークリーム 5 個中わさび入りが 1 個とします。一郎、二郎、三郎、四郎、五郎の 5 人が例題 2.14 と同様に一個ずつ食べるとき、誰が一番不利でしょうか。

^{*2} 「わさび・クリーム・クリーム」「クリーム・わさび・クリーム」「クリーム・クリーム・わさび」の 3 通りが等確率で生じるはずだから、全員平等……という答でももちろん正しいのですが、念のため条件つき確率などを使って計算してみてください。同じ結論に達するでしょうか？

答

一郎のシュークリームの中身を X_1 、二郎のシュークリームの中身を X_2 、……のようににおいて、先ほどと同様に確認します。

まず、一郎については

$$P(X_1 = \text{わさび}) = \frac{1}{5}$$

次に二郎は

$$\begin{aligned} P(X_2 = \text{わさび}) &= P(X_1 = \text{わさび}, X_2 = \text{わさび}) + P(X_1 = \text{クリーム}, X_2 = \text{わさび}) \\ &= 0 + P(X_2 = \text{わさび} | X_1 = \text{クリーム}) P(X_1 = \text{クリーム}) \\ &= \frac{1}{4} \cdot \left(1 - \frac{1}{5}\right) = \frac{1}{4} \cdot \frac{4}{5} = \frac{1}{5} \end{aligned}$$

三郎は

$$\begin{aligned} P(X_3 = \text{わさび}) &= P(X_1 = \text{わさび}, X_2 = \text{わさび}, X_3 = \text{わさび}) \\ &\quad + P(X_1 = \text{クリーム}, X_2 = \text{わさび}, X_3 = \text{わさび}) \\ &\quad + P(X_1 = \text{わさび}, X_2 = \text{クリーム}, X_3 = \text{わさび}) \\ &\quad + P(X_1 = \text{クリーム}, X_2 = \text{クリーム}, X_3 = \text{わさび}) \\ &= 0 + 0 + 0 \\ &\quad + P(X_3 = \text{わさび} | X_1 = \text{クリーム}, X_2 = \text{クリーム}) P(X_2 = \text{クリーム} | X_1 = \text{クリーム}) P(X_1 = \text{クリーム}) \\ &= \frac{1}{3} \cdot \frac{3}{4} \cdot \frac{4}{5} = \frac{1}{5} \end{aligned}$$

以下同様に、

$$\begin{aligned} P(X_4 = \text{わさび}) &= \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{4} \cdot \frac{4}{5} = \frac{1}{5} \\ P(X_5 = \text{わさび}) &= \frac{1}{1} \cdot \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{4} \cdot \frac{4}{5} = \frac{1}{5} \end{aligned}$$

ですから、やっぱり全員確率は同じでした。 ■

2.c 独立性についての補足

2.c.1 言いかえの同値性など

独立性の言いかえについて、同値性の証明の一部を例題としてここに載せておきます。

例題 2.16

2.5.2 項 (p.61) 「事象の独立性 (言いかえ)」の (イ) (ウ) (エ) がどれも同値なことを数式で示せ。ただし $P(\bigcirc\bigcirc)$ は 0 でも 1 でもないとする。

答

いちいち書くのがめんどうなので、まず次のように記号をつけておく。

$$\begin{aligned} a &\equiv P(\bigcirc\bigcirc, \blacktriangle\blacktriangle) \\ b &\equiv P(\bigcirc\bigcirc, \blacktriangle\blacktriangle\text{でない}) \\ c &\equiv P(\bigcirc\bigcirc\text{でない}, \blacktriangle\blacktriangle) \\ d &\equiv P(\bigcirc\bigcirc\text{でない}, \blacktriangle\blacktriangle\text{でない}) \end{aligned}$$

すると、

- (イ) は $a/(a+b) = c/(c+d)$ と表される。両辺に $(a+b)(c+d)$ をかけて分母をはらうと $a(c+d) = c(a+b)$ 。この両辺を展開して $ac+ad = ac+bc$ 、すなわち $ad = bc$ 。
- (ウ) は $a/(a+b) = a+c$ と表される。両辺に $(a+b)$ をかけて分母をはらうと $a = (a+b)(a+c)$ 。さらに展開して移項すれば $a - a^2 - ab - ac = bc$ 、すなわち $a(1-a-b-c) = bc$ 。ここで $a+b+c+d=1$ より $1-a-b-c=d$ だから、いまの式は結局 $ad = bc$ 。
- (エ) は $a:b = c:d$ と表される。つまり $ad = bc$ 。

こうして結局、(イ) (ウ) (エ) はどれも、 $ad = bc$ という条件と同値である (前提から $a+b \neq 0$ および $c+d \neq 0$ に注意)。だから (イ) (ウ) (エ) はどれも同値。 ■

2.5.3 項 (p.63) 「確率変数の独立性」の (オ) と (オ') が同値なことも、ここで確認しましょう。「オ \rightarrow オ'」はあたりまえ (そのまま $g(a) = P(X = a)$, $h(b) = P(Y = b)$ ととればよい)。「オ' \rightarrow オ」のほうはまた例題として載せます。

例題 2.17

確率変数 X は 1, 2, 3 のどれか、 Y は 1, 2, 3, 4, 5 のどれかの値をとるとする。任意の a, b に対して $P(X = a, Y = b) = g(a)h(b)$ と表されていたとき、 $P(X = a, Y = b) = P(X = a)P(Y = b)$ を示せ。

答

$P(X = a)$ は

$$\begin{aligned} P(X = a) &= P(X = a, Y = 1) + P(X = a, Y = 2) + \cdots + P(X = a, Y = 5) \\ &= g(a)h(1) + g(a)h(2) + \cdots + g(a)h(5) \\ &= g(a)(h(1) + h(2) + \cdots + h(5)) \end{aligned}$$

と計算されます。この (\cdots) を c とおけば、要するに $P(X = a) = cg(a)$ です。同様に、 $P(Y = b)$ は

$$\begin{aligned} P(Y = b) &= P(X = 1, Y = b) + P(X = 2, Y = b) + P(X = 3, Y = b) \\ &= g(1)h(b) + g(2)h(b) + g(3)h(b) \\ &= (g(1) + g(2) + g(3))h(b) \end{aligned}$$

と計算されます。この (\cdots) を d とおけば、要するに $P(Y = b) = dh(b)$ です。さらに、確率の合計は 1 のはずですから、

$$\begin{aligned} P(X = 1) + P(X = 2) + P(X = 3) &= 1 \\ P(Y = 1) + P(Y = 2) + \cdots + P(Y = 5) &= 1 \end{aligned}$$

すなわち

$$\begin{aligned} cg(1) + cg(2) + cg(3) &= c(g(1) + g(2) + g(3)) = cd = 1 \\ dh(1) + dh(2) + \cdots + dh(5) &= d(h(1) + h(2) + \cdots + h(5)) = dc = 1 \end{aligned}$$

というわけで $cd = 1$ なことが保証されます。すると、

$$P(X = a)P(Y = b) = cg(a) \cdot dh(b) = cdg(a)h(b) = g(a)h(b)$$

こうして、 $P(X = a, Y = b) = P(X = a)P(Y = b)$ が示されました。これはまさしく (オ) そのもの。

わかってから見れば、結局 $P(X = a) \propto g(a)$ かつ $P(Y = b) \propto h(b)$ なのであって^{*3}、あとは合計が 1 になるよう比例係数を調節してやっただけの話です。 ■

独立性から言えることについても、数式による説明を例題として以下に挙げます。

例題 2.18

X, Y はどちらも整数値をとる確率変数とする。 X と Y が独立なら、その絶対値 $|X|$ と $|Y|$ も独立なことを示せ。

答

意味を考えれば $|X|$ と $|Y|$ にかかわりがないことは当然なのですが、数式で示すと以下のとおりです。

任意の正整数 a, b に対して、

$$\begin{aligned} P(|X| = a)P(|Y| = b) &= \left(P(X = +a) + P(X = -a) \right) \left(P(Y = +b) + P(Y = -b) \right) \\ &= P(X = +a)P(Y = +b) + P(X = -a)P(Y = +b) \\ &\quad + P(X = +a)P(Y = -b) + P(X = -a)P(Y = -b) \end{aligned}$$

つまり、

$$P(|X| = a)P(|Y| = b) = \sum_{|x|=a, |y|=b} P(X = x)P(Y = y) \quad (2.16)$$

ただしこの \sum は、 $|x| = a$ かつ $|y| = b$ となるような (x, y) の組合せすべてについて合計するとします。 $a =$

^{*3} \propto は比例を表す記号です。くどく言えば、(a によらない) 定数 r が何かあって、 $P(X = a) = rg(a)$ と書けるということ。

0 や $b = 0$ の場合も、同じようにしてやはり (2.16) が言えます。ここで、 X と Y が独立だという前提により、2.5.3 項^(p.63)の (オ) から

$$\sum_{|x|=a, |y|=b} P(X=x)P(Y=y) = \sum_{|x|=a, |y|=b} P(X=x, Y=y)$$

が成り立つことに注意。上式の右辺は、要するに $P(|X|=a, |Y|=b)$ です。まとめると結局 $P(|X|=a)P(|Y|=b) = P(|X|=a, |Y|=b)$ ですから、 $|X|$ と $|Y|$ に対しても (オ) が導かれました^{*4}。よって $|X|$ と $|Y|$ は独立。 ■

例題 2.19

X, Y はどちらも正の整数値をとる確率変数とする。 X と Y が独立なら

$$P(X \leq a, Y \leq b) = P(X \leq a)P(Y \leq b)$$

が成り立つことを示せ。

答

任意の a, b に対して、

$$\begin{aligned} & P(X \leq a)P(Y \leq b) \\ &= \left(P(X=1) + P(X=2) + \cdots + P(X=a) \right) \left(P(Y=1) + P(Y=2) + \cdots + P(Y=b) \right) \\ &= P(X=1) \left(P(Y=1) + P(Y=2) + \cdots + P(Y=b) \right) \\ &\quad + P(X=2) \left(P(Y=1) + P(Y=2) + \cdots + P(Y=b) \right) \\ &\quad + \cdots \\ &\quad + P(X=a) \left(P(Y=1) + P(Y=2) + \cdots + P(Y=b) \right) \\ &= \sum_{1 \leq x \leq a, 1 \leq y \leq b} P(X=x)P(Y=y) \end{aligned}$$

ただし、この \sum は「 $1 \leq x \leq a$ かつ $1 \leq y \leq b$ となるような (x, y) の組合せすべて」について合計するとします。ここで、 X と Y が独立だという前提により、

$$\sum_{1 \leq x \leq a, 1 \leq y \leq b} P(X=x)P(Y=y) = \sum_{1 \leq x \leq a, 1 \leq y \leq b} P(X=x, Y=y)$$

この右辺は要するに $P(X \leq a, Y \leq b)$ です。

(別解) 以下のようにして、前の話に帰着させる。

$$g(x) = \begin{cases} 1 & (x \leq a) \\ 0 & (\text{他}) \end{cases}, \quad h(y) = \begin{cases} 1 & (y \leq b) \\ 0 & (\text{他}) \end{cases}$$

とおけば

- $X \leq a$ という条件は $g(X) = 1$ と同値
- $Y \leq b$ という条件は $h(Y) = 1$ と同値

だから

$$P(X \leq a, Y \leq b) = P(g(X) = 1, h(Y) = 1) = P(g(X) = 1)P(h(Y) = 1) = P(X \leq a)P(Y \leq b)$$

途中の式変形には、 $g(X)$ と $h(Y)$ が独立なことを利用しました。 ■

^{*4} 「 a, b がともに ≥ 0 」でないときには、 $P(|X|=a, |Y|=b)$ も $P(|X|=a)P(|Y|=b)$ も 0 なので、やはりこれらは等しい。

2.c.2 無限にまつわるやっかいごと

無限個の確率変数の独立性については、ちょっとやっかいな事情があります（初心者は気にしないで結構です）。無限個の確率変数 X_1, X_2, X_3, \dots の独立性は、そのまますなおに

$$P(X_1 = a_1, X_2 = a_2, X_3 = a_3, \dots) = P(X_1 = a_1) P(X_2 = a_2) P(X_3 = a_3) \cdots$$

が成り立つこととは定義できません。この式だと、 a_1, a_2, a_3, \dots の値にかかわらず両辺とも 0 になってしまったりして、意味のある比較が一般にはできないからです。そこでこんなふうに定義します：「 X_1, X_2, X_3, \dots から任意の有限個をピックアップしたら、それらが常に独立」となっているとき、「 X_1, X_2, X_3, \dots は独立」と呼ぶ。確率ゼロのはらむ問題については 4.2 節 (p.123) 「確率ゼロ」も参照。8.e.3 項 (p. 補足編 68) も同類の話です。

また、先ほどの例題 2.16 のようなことをきちんと一般的に議論しようとする、本当は無限個の足し算というデリケートな話からできます。何がデリケートなのかは後ほど 3.c.4 項 (p. 補足編 26) で。

2.c.3 かかわりの度合い

8.3.2 項 (p.308) 「二変数のエントロピー」では、情報量という観点から独立性について再考しています。情報量に着目すれば、単に独立か従属かだけでなく、どれくらい強くかかわりあっているかを数値化することができます。例題 2.4^(p.43) (トランプの色あわせ) では、一枚目の色 X を見ても二枚目の色 Y の条件つき確率が若干違ってくるだけ。例題 2.14^(p. 補足編 15) (わさびシュークリーム) だと、もし一郎がわさびなら残りはすべてクリームなことが 100% 保証される。前者より後者のほうが強くかかわっているよねという話です。似た話題として 5.1 節 (p.174) 「共分散と相関係数」も挙げられますが、これについては重要な注意が 5.1.2 項 (p.176) 「共分散の性質」にありますから読み逃さないように。

第3章の補足

3.a (独立な確率変数たちの) 合計値や最大値の分布

一般論として、確率変数 X, Y の同時分布がわかっているならば、 $(2X + 9Y + 4)(7X + 5Y + 3)$ だろうが $\sin(\cos(\pi X)) + \log(X + Y)$ だろうが確率分布を求めることはできます。前の例題 3.1^(p.73) のようにがんばって地道に計算すれば良いだけです。でもいくつかの場合には、地道に計算するよりもすっきり求められる方法があります。本節ではそんな話を 2 つ紹介します。

3.a.1 (独立な確率変数たちの) 合計値の分布

確率変数の合計値の分布を知りたくなる場面があります。たとえば、あの仕事とこの仕事の合計所要時間の分布や、昼と夜の合計売上げの分布などです。他にも、図形の回転やデータの平均といった足し算を含む処理に確率がからむと、その解析には合計値の分布が顔を出します。

確率変数たちの合計値の分布を計算すること自体は、ここまでの知識でできるはずですが。まずはそれをやって感触をつかみましょう。その後で、こういうときはこういうきれいな話になっている、という指摘をします。

例題 3.1

次のようないかさまサイコロを二回ふるとき、出た目の合計が 5 になる確率は？ また、合計が 6 になる確率は？

値	その値が出る確率
1	0.4
2	0.1
3	0.1
4	0.1
5	0.1
6	0.2

答

一回目の値 X と二回目の値 Y とは独立だから、任意の a, b に対して $P(X = a, Y = b) = P(X = a)P(Y = b)$ となることに注意する。

$$\begin{aligned}
 P(X + Y = 5) &= P(X = 1, Y = 4) + P(X = 2, Y = 3) + P(X = 3, Y = 2) + P(X = 4, Y = 1) \\
 &= P(X = 1)P(Y = 4) + P(X = 2)P(Y = 3) + P(X = 3)P(Y = 2) + P(X = 4)P(Y = 1) \\
 &= 0.04 + 0.01 + 0.01 + 0.04 = 0.1 \\
 P(X + Y = 6) &= P(X = 1)P(Y = 5) + P(X = 2)P(Y = 4) + P(X = 3)P(Y = 3) + P(X = 4)P(Y = 2) \\
 &\quad + P(X = 5)P(Y = 1) \\
 &= 0.04 + 0.01 + 0.01 + 0.01 + 0.04 = 0.11
 \end{aligned}$$

計算自体は地道にやればよいだけのことですが、規則性に気づいたでしょうか。 $P(X + Y = 5)$ の計算は、こんなふうでした。

X の値 a	...	-1	0	1	2	3	4	5	6	...
Y の値 b	...	6	5	4	3	2	1	0	-1	...
確率 $P(X = a)$...	0	0	0.4	0.1	0.1	0.1	0.1	0.2	...
確率 $P(Y = b)$...	0.2	0.1	0.1	0.1	0.1	0.4	0	0	...
確率のかけ算 $P(X = a)P(Y = b)$...	0	0	0.04	0.01	0.01	0.04	0	0	...

→ 合計 0.1

後の応用も考えて、確率 0 のところもあえて書いておきました。同様に、 $P(X + Y = 6)$ の計算は、こうです。

X の値 a	...	0	1	2	3	4	5	6	7	...
Y の値 b	...	6	5	4	3	2	1	0	-1	...
確率 $P(X = a)$...	0	0.4	0.1	0.1	0.1	0.1	0.2	0	...
確率 $P(Y = b)$...	0.2	0.1	0.1	0.1	0.1	0.4	0	0	...
確率のかけ算 $P(X = a)P(Y = b)$...	0	0.04	0.01	0.01	0.01	0.04	0	0	...

→ 合計
0.11

次の規則性に注目してください。

- X の確率と Y の確率とを逆順に並べる
- それぞれの確率を縦にかけ算して、結果を合計したものが答
- $X + Y$ の値を 1 増やすごとに、 X と Y との対応が 1 つずれる

この規則性を数式でまとめると、次のようになります。最後に出てくる \sum の特徴的な格好にご注目。

$$\begin{aligned}
 P(X + Y = c) &= \cdots + P(X = 0)P(Y = c) + P(X = 1)P(Y = c - 1) + P(X = 2)P(Y = c - 2) \\
 &\quad + \cdots + P(X = c - 1)P(Y = 1) + P(X = c)P(Y = 0) + \cdots \\
 &= \sum_{i=-\infty}^{\infty} P(X = i)P(Y = c - i)
 \end{aligned} \tag{3.1}$$

最後の式で \sum の範囲が $-\infty$ から ∞ までと大風呂敷になっていますが、余分に広げたところはどうせ値が 0 なので不都合ありません^{*1}。

一般に、整数 k に対して定義された関数 $f(k), g(k)$ があつたら、それらから作られる関数 $h(c) = \sum_{i=-\infty}^{\infty} f(i)g(c - i)$ のことを、 f と g の畳み込み (**convolution**) と呼びます。記号では $h = f * g$ などと書いたりします (人によって流儀がいろいろ)。式 (3.1) はまさにこの格好ですね。要するに、 $j + k = c$ となるすべての組合せについて f と g のかけ算 $f(j)g(k)$ を合計するのが畳み込みです。たとえば $c = 2$ なら、図 3.1 のアミがけの部分で足し合わせることになります。

				\vdots			
	$f(-1)g(4)$	$f(0)g(4)$	$f(1)g(4)$	$f(2)g(4)$	$f(3)g(4)$	$f(4)g(4)$	
	$f(-1)g(3)$	$f(0)g(3)$	$f(1)g(3)$	$f(2)g(3)$	$f(3)g(3)$	$f(4)g(3)$	
...	$f(-1)g(2)$	$f(0)g(2)$	$f(1)g(2)$	$f(2)g(2)$	$f(3)g(2)$	$f(4)g(2)$...
	$f(-1)g(1)$	$f(0)g(1)$	$f(1)g(1)$	$f(2)g(1)$	$f(3)g(1)$	$f(4)g(1)$	
	$f(-1)g(0)$	$f(0)g(0)$	$f(1)g(0)$	$f(2)g(0)$	$f(3)g(0)$	$f(4)g(0)$	
	$f(-1)g(-1)$	$f(0)g(-1)$	$f(1)g(-1)$	$f(2)g(-1)$	$f(3)g(-1)$	$f(4)g(-1)$	
				\vdots			

▶ 図 3.1 $h(2)$ は要するに、 $j + k = 2$ となるすべての組合せについての $f(j)g(k)$ の合計

畳み込みは $h(c) = \sum_{i=-\infty}^{\infty} f(c - i)g(i)$ と書けます。こちらも結局は、 $j + k = c$ となるすべての組合せについて $f(j)g(k)$ の合計を求めたことになるからです。

まとめると、

整数値をとる独立な確率変数 X, Y に対して、合計値 $X + Y$ の確率分布は、「 X の確率分布と Y の確率分布との畳み込み」になる。

これが本項で言いたかったことです。ついでに引き算 $X - Y$ も、 $X + (-Y)$ と解釈すれば合計値の分布に帰着させることができます。

? 3.a やけにアピールしていますが、それほどすごい話には見えません。素朴な計算の結果にただ名前をつけただけなのでは？

はい。本書の中ではそれだけです。それでもあえて一項を割いたわけは、次の二点です。

- 合計値の分布の計算法を確認しておきたい。—— この連続値版 (4.f.1 項 (p. 補足編 42)) が後ほど必要になります (→ 4.g.3 項 (p. 補足編 44))。
- この機会に畳み込みという言葉的印象づけたい。—— 畳み込みという演算は他の分野 (物理や工学) でも出てきます。

実は (離散) フーリエ変換というテクニックを使えば、畳み込みの結果をもっとかつこよく求めることができます。畳み込みの真骨頂はそこにあります。アマチュア向けの優先順位を考慮して本書では割愛しましたが、確率論の本気の教科書ならまず載っているはずですから調べてみてください (付録 C.2^(p.350) 「特性関数」)。

^{*1} $\sum_{i=-\infty}^{\infty} \cdots$ は、 $\sum_{i=a}^b \cdots$ に対する $a \rightarrow -\infty, b \rightarrow \infty$ の極限を表します。 $i =$ と書いたからといって、 i に $-\infty$ なんていう謎な数を代入するという意味ではありません。 $\int_{-\infty}^{\infty} \cdots$ や $[\cdots]_{-\infty}^{\infty}$ などと同様です。

なお、独立でないときの合計値の分布は、変数どうしのからみ具合に応じてケースバイケースとしか言えません。一般には同時確率を地道に合計するしかないでしょう。たとえば $P(X + Y = 3) = \dots + P(X = 0, Y = 3) + P(X = 1, Y = 2) + P(X = 2, Y = 1) + \dots$ 。一般的に書けば $P(X + Y = c) = \sum_{i=-\infty}^{\infty} P(X = i, Y = c - i)$ です。

3.a.2 (独立な確率変数たちの) 最大値の分布

部品 A と部品 B のどちらか片方でも壊れたら機能停止する、という製品が壊れるまでの日数は、「A が壊れるまでの日数」と「B が壊れるまでの日数」との小さいほうで決まります。また、部品 A と部品 B のどちらか片方でも生きていれば OK、という製品が壊れるまでの日数は、「A が壊れるまでの日数」と「B が壊れるまでの日数」との大きいほうで決まります。そういった話では、確率変数たちの最小値や最大値の分布が問題になるでしょう。

では、またいつものいかさまサイコロで説明します。

値	その値が出る確率
1	0.4
2	0.1
3	0.1
4	0.1
5	0.1
6	0.2

こんなサイコロを二回ふることにして、一回目の値を X 、二回目の値を Y とおきます。 X と Y は独立で、どちらも確率分布は上の表のようになります。ここで、 X, Y の最大値

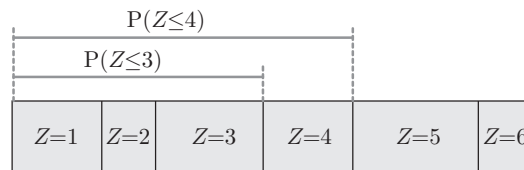
$$Z = \max(X, Y) \quad \dots\dots X, Y \text{ の大きいほうの値 (付録 A.2 (p.319))}$$

を考えましょう。どうすれば Z の確率分布を上手に求められるでしょうか。

ヒントを出します。「○○になる確率」を直接考えるよりも、「○○以下になる確率」に着目してください。ポイントは二つ。一つは、「大きいほうが a 以下」という条件は「どちらも a 以下」と言い換えられることです：

$$P(Z \leq a) = P(X \leq a, Y \leq a)$$

もう一つは、 $P(Z \leq a)$ の一覧があればそこから $P(Z = a)$ の一覧を図 3.2 のように求められることです。



全体の面積 = 1

▶ 図 3.2 $P(Z = 4) = P(Z \leq 4) - P(Z \leq 3)$ 。各長方形の面積が確率を表す

ではヒントをふまえて、 $P(Z \leq a)$ の一覧をまず作りましょう。初手は、

$$P(Z \leq a) = P(X \leq a, Y \leq a) = P(X \leq a) P(Y \leq a)$$

という変形です。 X と Y が独立なことから、最後のように分解されてしまいます (→ 2.5.3 項 (p.63) 「確率変数の独立性」)。分解されればもう簡単。 $P(X \leq a)$ は

$$P(X \leq 1) = P(X = 1) = 0.4$$

$$P(X \leq 2) = P(X = 1) + P(X = 2) = 0.4 + 0.1 = 0.5$$

$$P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = 0.4 + 0.1 + 0.1 = 0.6$$

という調子で計算できます。残りも書いておくと、

$$P(X \leq 4) = 0.4 + 0.1 + 0.1 + 0.1 = 0.7$$

$$P(X \leq 5) = 0.4 + 0.1 + 0.1 + 0.1 + 0.1 = 0.8$$

$$P(X \leq 6) = 0.4 + 0.1 + 0.1 + 0.1 + 0.1 + 0.2 = 1$$

$P(Y \leq a)$ のほうも同様ですから、結論は

$$P(Z \leq 1) = 0.4 \times 0.4 = 0.16$$

$$P(Z \leq 4) = 0.7 \times 0.7 = 0.49$$

$$P(Z \leq 2) = 0.5 \times 0.5 = 0.25$$

$$P(Z \leq 5) = 0.8 \times 0.8 = 0.64$$

$$P(Z \leq 3) = 0.6 \times 0.6 = 0.36$$

$$P(Z \leq 6) = 1.0 \times 1.0 = 1.00$$

これで $P(Z \leq a)$ の一覧が得られました。

あとはそこから $P(Z = a)$ を求めればよい。

$$P(Z = 1) = P(Z \leq 1) - P(Z \leq 0) = 0.16 - 0 = 0.16$$

$$\begin{aligned}
P(Z=2) &= P(Z \leq 2) - P(Z \leq 1) = 0.25 - 0.16 = 0.09 \\
P(Z=3) &= P(Z \leq 3) - P(Z \leq 2) = 0.36 - 0.25 = 0.11 \\
P(Z=4) &= P(Z \leq 4) - P(Z \leq 3) = 0.49 - 0.36 = 0.13 \\
P(Z=5) &= P(Z \leq 5) - P(Z \leq 4) = 0.64 - 0.49 = 0.15 \\
P(Z=6) &= P(Z \leq 6) - P(Z \leq 5) = 1.00 - 0.64 = 0.36
\end{aligned}$$

こうして、素朴に全組合せを考えるよりは楽に $P(Z=a)$ が求められました。念のため次の検算もする習慣をつけてください。

$$P(Z=1) + \cdots + P(Z=6) = 0.16 + 0.09 + 0.11 + 0.13 + 0.15 + 0.36 = 1$$

サイコロに限らず、独立な確率変数たちの最大値の分布は同様の考え方で計算できます。3 個とか 4 個とかの最大値でも同様です。また、**最小値**については後の例題 3.3 を参照してください。ただしこれらはあくまで、確率変数たちが独立な場合です。独立でない場合、一般には、該当する同時確率を地道に合計するしかないでしょう。

例題 3.2

(いかさまではない普通の) サイコロを 3 回ふったとき、その最大値が 5 となる確率は？

答

3 回ふった結果をそれぞれ X, Y, Z とし、その最大値を $W = \max(X, Y, Z)$ と置く。 $P(W=5) = P(W \leq 5) - P(W \leq 4)$ である。 X, Y, Z が独立なことから、右辺の各項は

$$\begin{aligned}
P(W \leq 5) &= P(X \leq 5, Y \leq 5, Z \leq 5) = P(X \leq 5) P(Y \leq 5) P(Z \leq 5) = \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} = \frac{125}{216} \\
P(W \leq 4) &= P(X \leq 4, Y \leq 4, Z \leq 4) = P(X \leq 4) P(Y \leq 4) P(Z \leq 4) = \frac{4}{6} \cdot \frac{4}{6} \cdot \frac{4}{6} = \frac{64}{216}
\end{aligned}$$

と計算される。よって

$$P(W=5) = \frac{125}{216} - \frac{64}{216} = \frac{61}{216} \approx 0.28$$

■

例題 3.3

(いかさまではない普通の) サイコロを 2 回ふったとき、その最小値が 3 となる確率は？

答

2 回ふった結果をそれぞれ X, Y とし、その最小値を $Z = \min(X, Y)$ と置く。 $P(Z=3) = P(Z \geq 3) - P(Z \geq 4)$ である。 X, Y が独立なことから、この右辺の各項は

$$\begin{aligned}
P(Z \geq 3) &= P(X \geq 3, Y \geq 3) = P(X \geq 3) P(Y \geq 3) = \frac{4}{6} \cdot \frac{4}{6} = \frac{16}{36} \\
P(Z \geq 4) &= P(X \geq 4, Y \geq 4) = P(X \geq 4) P(Y \geq 4) = \frac{3}{6} \cdot \frac{3}{6} = \frac{9}{36}
\end{aligned}$$

と計算される。よって

$$P(Z=3) = \frac{16}{36} - \frac{9}{36} = \frac{7}{36} \approx 0.19$$

(別解) $\tilde{X} = -X, \tilde{Y} = -Y, \tilde{Z} = -Z$ と置けば、 \tilde{X} と \tilde{Y} は独立であり、 $\tilde{Z} = \max(\tilde{X}, \tilde{Y})$ となる*2。これで話を最大値の問題に変換できたので、あとは本文のようにして $P(\tilde{Z} = -3)$ を求めればよい。—— このように符号をひっくり返して \min と \max を入れ替える技は、思考を節約するための常套手段です。いつでも両者を変換できるので、片方だけ検討しておけば十分。

■

3.b 2 項分布のプログラムを書く際の落とし穴

2 項分布 $\text{Bn}(n, p)$ がらみのプログラムを書く際には、ひっかかりやすい落とし穴があります。

$n!$ を素朴に計算したのでは、すぐオーバーフローしてしまうでしょう。実は、大学レベルの解析学を勉強すると、 **Γ 関数** (ガンマ関数) というものを習います。

$$\Gamma(x) \equiv \int_0^\infty t^{x-1} e^{-t} dt, \quad (x > 0)$$

この関数は、整数 $n \geq 0$ に対して $\Gamma(n+1) = n!$ という性質を持ちます*3。そして多くのプログラミング言語には、この関数

*2 具体例を見れば仕掛けがわかります。たとえば、 $X=2, Y=4$ のとき、 $\max(-2, -4) = -2$ 。

*3 整数に限らず一般の $x > 0$ でも $\Gamma(x+1) = x\Gamma(x)$ が成り立ちます (部分積分してみましょう。 $t \rightarrow \infty$ のとき $t^x e^{-t} \rightarrow 0$ に注意)。また、 $\Gamma(1/2) = \sqrt{\pi}$ となることも有名です ($t \equiv u^2$ において置換積分し、付録 A.5.2(p.328) のガウス積分と見比べましょう)。5.3.6 項 (p.211) のカイ自乗分布も参照。

$\Gamma(x)$ を計算するライブラリが用意されています。ご使用の言語のマニュアルを調べてみてください。たいてい `gamma` のような関数名になっているはず^{*4}。 $n!$ の近似計算については付録 B.1(p.337)「Stirling の公式」も参照。

もう一つの落とし穴は、べき乗 $p^k q^{n-k}$ を素朴に計算したらアンダーフローが起きてしまうことです。そんなときは、 $P(X = k)$ を直接計算せず、 $c \equiv \log P(X = k)$ を求めてから $P(X = k) = e^c$ と計算するのがよいでしょう (→ 付録 A.5(p.326)「指数と対数」)。2 項分布の話に限らず、極端に大きい数や極端に小さな数を扱う際には、 \log を使うのが一つの定石です。上述の Γ 関数についても、 $\log \Gamma(x)$ を計算する関数が、`lgamma` といった名前¹でたいてい用意されています。

また、 n が大きいなら正規分布で近似するのも手です。これについては 4.6.3 項 (p.167)「中心極限定理」で。

3.c 期待値についての補足

3.c.1 記法

期待値の記法にはいろいろな流儀があつて、括弧を $E(X)$ や $E\{X\}$ で書く人も多くいます。 E を \mathbb{E} や \mathcal{E} と書いた本もあります。また、 $\langle X \rangle$ や \bar{X} のような記号で示す流儀も見かけます。書き方によって意味やニュアンスを使いわけられる場合もありますが、ここでは深入りしません。筆者の好みは $E[X]$ です。神様視点では確率変数 X とは関数 $X(\omega)$ のことでしたから、 E は「関数を入れたら数が出てくるもの」です。このようなものを一般に **汎関数** と呼びます。角括弧を使って $E[X]$ と書くと、汎関数であることが強調されます。

3.c.2 平均と期待値の違い

本書では平均と期待値とをきっちり使い分けます。平均のほうは、「ここ 10 試合の来場者数の平均」のように使います。期待値のほうは、「サイコロをふって出る値の期待値」のように使います。違いはなんでしょう？——端的に言えば、「何個かの値の平均」「確率変数の期待値」という違いです。平均とは、値が何個あるときに、その合計値を個数で割ったもののこと。確率は出てきません。上の例でも、10 個の値に対して、単純に「その合計値を 10 で割った答」を指しています。一方、期待値は、確率変数 (ゆらぐ値) X に対する概念です。 X はゆらぐので単一の数値では言い表せません。それでもあえて、ゆらがなく一つの目安がほしい。そんな目安の一案が期待値です。

なお、世の中では「平均」「期待値」という言葉をそれほど区別せず使う場合が多くあります。平均のことを期待値と言う人はまずいませんが、期待値のことを平均と呼ぶ例は専門書でもよく見かけます。たとえば、基準値 a と実際の値 X との自乗誤差 $(X - a)^2$ について、その期待値 $E[(X - a)^2]$ は **平均自乗誤差** と呼ばれます。自乗誤差の期待値は応用上も頻出するので、頭に留めておいてください (→ 3.4.6 項 (p.98)「自乗期待値と分散」や 6.1.5 項 (p.235)「期待罰金」など)。

3.c.3 練習用例题

例题 3.4

確率 $1/2$ で「1」、確率 $1/3$ で「2」、確率 $1/6$ で「5」が出るような確率変数 X に対し、 $E[1/(X - 3)^2]$ を求めよ。

答

$$\begin{aligned} E[1/(X - 3)^2] &= \text{「各場合の } 1/(X - 3)^2 \text{ の値} \times \text{（その場合が起きる確率）} \text{」の合計} \\ &= \frac{1}{(1-3)^2} \cdot P(X=1) + \frac{1}{(2-3)^2} \cdot P(X=2) + \frac{1}{(5-3)^2} \cdot P(X=5) \\ &= \frac{1}{4} \cdot \frac{1}{2} + 1 \cdot \frac{1}{3} + \frac{1}{4} \cdot \frac{1}{6} = \frac{3+8+1}{24} = \frac{1}{2} \end{aligned}$$

例题 3.5

3.3.2 項 (p.79)「期待値の基本性質」の不等式と $E[X + c]$, $E[cX]$ のところを式で証明せよ。

答

$$\begin{aligned} E[X + c] &= \sum_i (i + c) P(X = i) = \sum_i i P(X = i) + \sum_i c P(X = i) \\ &= E[X] + c \sum_i P(X = i) = E[X] + c \cdot 1 = E[X] + c \\ E[cX] &= \sum_i (ci) P(X = i) = c \sum_i i P(X = i) = c E[X] \end{aligned}$$

また、 X が必ず c 以上なのであれば、

$$E[X - c] = \sum_i (i - c) P(X = i)$$

を観察することにより以下がわかる。 $i - c < 0$ については $P(X = i) = 0$ なので、上式右辺は「0 以上の値

^{*4} 同じ `gamma` という名前¹で中身は別物の場合もありますから、ちゃんと調べましょう。この混乱を避けるため、C 言語の標準規格 (C99) では `tgamma` という名前になっています。

の合計」だから $E[X - c] \geq 0$ 。これを $E[X - c] = E[X + (-c)] = E[X] - c$ とあわせれば $E[X] \geq c$ が言える。 ■

例題 3.6

X と Y の同時分布が次の表のようになっていたとする。(1)(2) それぞれについて、 $E[X]E[Y] \neq E[XY]$ を確かめよ。

(1)

	$Y = -1$	$Y = +1$
$X = -1$	0.5	0
$X = +1$	0	0.5

(2)

	$Y = -1$	$Y = +1$
$X = -1$	0.1	0.4
$X = +1$	0.4	0.1

答

(1) 周辺分布は

$$P(X = -1) = P(X = +1) = 0.5, \quad P(Y = -1) = P(Y = +1) = 0.5$$

ですから、それぞれの期待値は

$$E[X] = (-1) \cdot 0.5 + (+1) \cdot 0.5 = 0, \quad E[Y] = (-1) \cdot 0.5 + (+1) \cdot 0.5 = 0$$

一方、かけ算の期待値は

$$E[XY] = (-1) \cdot (-1) \cdot 0.5 + (+1) \cdot (+1) \cdot 0.5 = 1 \neq E[X]E[Y]$$

(2) 同様に、 $E[X] = E[Y] = 0$ 。一方、

$$\begin{aligned} E[XY] &= (-1) \cdot (-1) \cdot 0.1 + (+1) \cdot (+1) \cdot 0.1 + (-1) \cdot (+1) \cdot 0.4 + (+1) \cdot (-1) \cdot 0.4 \\ &= -0.6 \neq E[X]E[Y] \end{aligned}$$

例題 3.7

あなたが経営するスーパー SHOUJIKI の隣に、ライバルのスーパー USOTSUKI が店を出しました。ライバルは、「くじびきで 10 人に 1 人返金」という目玉企画を大々的に宣伝し、客を集めています。あなたは考えました。「この企画は、実質的には全商品を 10% 値引きするようなものだ。いくらなんでも採算がとれるはずはない。おかしい」……しかし、いくら偵察してみても、確かにちゃんと 10 人に 1 人は買い物代金を全額返してもらっているようです。どうしたことでしょう？

答

あなたがにらんだとおり、ライバルはいんちきをしています。くじに細工して、少額の客にばかり当たりが偏るようにしていたのです。これなら、店側の負担は少なくて済みます。

確率に焼きなおせば以下のとおり。買い物の金額を確率変数 X 、当たりはずれを確率変数 Y で表しましょう (確率 10% で $Y = 1$ 、確率 90% で $Y = 0$)。そのかけ算 XY が値引き額になります。ここで、 X と Y が独立でなかったのがポイントです。このため、あなたが考えた「値引き額の期待値」 $E[X]E[Y] = 0.1E[X]$ と、実際の「値引き額の期待値」 $E[XY]$ とが一致しなかったのです。(もちろんこの例はフィクションであり、実在の人物・団体とは一切関係ありません) ■

例題 3.8

「 $E[XY] = E[X]E[Y]$ なら X と Y は必ず独立である」という主張は正しいだろうか？ 正しいか否かを答え、それを証明せよ。

答

正しくない。たとえば、 X, Y の同時確率が以下の表のようになっていたら、

- $E[XY] = E[X]E[Y]$ が成り立っている。
- しかし、 X, Y は独立でない。

(「必ず〇〇である」という相手の主張を崩すには、反例を一つ挙げれば十分です)

	$Y = -1$	$Y = 0$	$Y = +1$
$X = -1$	0	$1/4$	0
$X = 0$	$1/4$	0	$1/4$
$X = +1$	0	$1/4$	0

■

？ 3.b 「 X が大きいと Y も大きい」という傾向があれば $E[XY] > E[X]E[Y]$ になって、「 X が小さいと Y が大きい」という傾向があれば $E[XY] < E[X]E[Y]$ になる、と理解してよい？ ここまでの話を聞いたら、そんな気がします。

するどいですね。5.1 節 (p.174) 「共分散と相関係数」はこの性質を活用しています。「そういう傾向があるか、あるならどれくらい強いかな」といったことをデータから測るために、かけ算の期待値を活用する話です。例題 5.4 (p.178) も参照。

3.c.4 期待値が存在しない場合への深入り

無限和の微妙な話

3.3.4 項 (p.84) 「期待値が存在しない場合」の Z について、 $E[Z]$ を求める公式をとりあえずそのままあてはめてみると、

$$(-2) \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} + (-8) \cdot \frac{1}{8} + 16 \cdot \frac{1}{16} + \cdots = (-1) + 1 + (-1) + 1 + \cdots$$

こういった無限個の数の足し算を無造作に取り扱っていると、次のようにおかしなことが生じます。

- A さん：上の式は $(-1 + 1) + (-1 + 1) + \cdots = 0$ じゃないの？
- B さん：上の式は $(-1) + (1 - 1) + (1 - 1) + \cdots = -1$ じゃないの？

ここはもっときっちり考えなくてはいけません。

基本に戻って順に見ていきましょう。2 個の数の足し算はすでに定義されているものとしてください。3 個の数の足し算 $a + b + c$ は、2 個の足し算を使って $(a + b) + c$ と定義します。この値が $b + (c + a)$ などとも同じになることは皆さんご存知のとおりです。どう並べかえようと括弧をつけようと、答は変わりません。

同様に、4 個なら $((a + b) + c) + d$ のように、5 個なら $((a + b) + c) + d + e$ のように定義していきます。こうして、任意の有限個の足し算が定義されました。この場合も、

$$a + b + c + d + e = c + ((a + d) + (e + b))$$

等々、どう並べかえても括弧をつけても答は変わりません。

では、無限個（正確には可算無限個 → 付録 A.3.2 (p.320)）の数 a_1, a_2, a_3, \dots の足し算はどう定義しましょうか。現段階で我々が手にしているのは有限個の足し算ですから、とりあえずこんなふうに途中までの足し算を考えていきます。

$$\begin{aligned} s_1 &= a_1 \\ s_2 &= a_1 + a_2 \\ s_3 &= a_1 + a_2 + a_3 \\ &\vdots \end{aligned}$$

すると無限数列 s_1, s_2, s_3, \dots ができました。この数列の極限をもって無限和 $a_1 + a_2 + a_3 + \cdots$ を定義するのが自然でしょう。ここでまず注意してほしいのは、収束しない数列もあるという事実です。たとえば、数列 $1, 2, 4, 8, \dots$ は無限大へ発散してしまいます。また、数列 $1, -1, 1, -1, \dots$ はいつまでも振動するので、こちらも極限は存在しません。数列 s_1, s_2, s_3, \dots がこんなふうになった場合には、無限和 $a_1 + a_2 + a_3 + \cdots$ は存在しないことになります。先ほどの $(-1) + 1 + (-1) + 1 + \cdots$ はまさにこれでした。

小うるさい話はまだ続きます。無限和では、並べかえたり括弧をつけたりすると答が変わることがあります。事実なのだから文句は言いつこなし。上のように途中までの足し算の極限で無限和を定義する限り、そうなるってしまうのです。実際、上の定義に従えば、

$$\begin{aligned} (-1) + 1 + (-1) + 1 + (-1) + 1 - \cdots &\text{—— 存在しない} \\ (-1 + 1) + (-1 + 1) + (-1 + 1) + \cdots &= 0 \\ (-1) + (1 - 1) + (1 - 1) + \cdots &= -1 \end{aligned}$$

はどれも正しいことが確認できます。

期待値と名乗るための条件

計算順序によって答が変わってしまうようでは、期待値を計算するとき困りますね。標準の順序を決めておけば良いと思うかもしれませんが、それだと公式 $E[g(X)] = \sum_i g(i) P(X = i)$ が保証できなくなります。これは辛すぎ。

そこで、期待値という言葉は、計算順序の問題が生じないとき限って使うことにします。つまり、計算順序をどう変えても答が同じになる場合にだけ、「期待値は〇〇だ」と言うことを許します。

ではこの厳しい条件をどうやって確認すればよいかな。実はこんなテスト法が知られています。いま、 $a_1 + a_2 + a_3 + \cdots$ が有限値として存在するとしましょう。その場合、

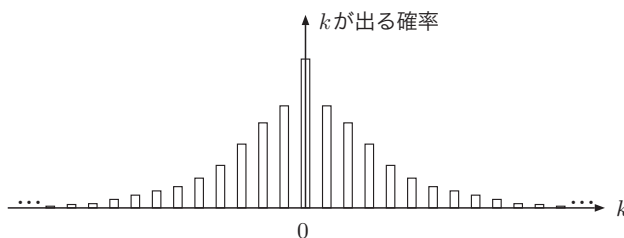
- $|a_1| + |a_2| + |a_3| + \cdots$ が有限値として存在するなら、 $a_1 + a_2 + a_3 + \cdots$ は計算順序をどう変えても答が同じ。
- さもなくば、 $a_1 + a_2 + a_3 + \cdots$ は計算順序を変えると答が変わってしまう。

前者の場合を絶対収束、後者の場合を条件収束と呼びます。詳しくは解析学のきちんとした教科書（たとえば参考文献 [7]）を参照してください。

期待値とは $\sum_i i P(X = i)$ のことだと本文では述べましたが、それだけだと本当は不十分です。正確には、この足し算が絶対

収束するという条件がつきます。(だから ? 3.4(p.87) の説明も厳密に言えば不十分でした)

? 3.c この図のように分布が左右対称だったら、無条件に「期待値は 0」と言ってよいのでは？



いいえ。言葉の定義を自分勝手に変えてはいけません。期待値という言葉を使ってよいのは、あくまで絶対収束するときだけです。

たとえば、3.3.4 項 (p.84) の Y に対し、それと独立に $+1$ か -1 を確率半々でとる確率変数 S を考えて、 $W \equiv SY$ とおきましょう。分布を具体的に書けば、 $P(W = 2^k) = P(W = -2^k) = 1/2^{k+1}$ ($k = 1, 2, 3, \dots$) です。この分布は左右対称ですが、期待値は存在しません。

期待値が存在する場合と比べて W がいかに違った挙動を示すか、後ほど 3.e.2 項 (p. 補足編 28) や例題 C.2(p.352) で吟味します。

3.d 分散についての補足

3.4.2 項 (p.90) で分散を定義する際の外れ具合の測り方は、単純に引き算して $x - \mu$ としたくなるかもしれませんが、でもこれはだめです。たとえば、 $\mu = 100$ に対して $x = 3$ が出たら、大外れだから指標が大きくなってほしい。なのに、 $3 - 100 = -97$ という「小さい」(0 よりもさらに小さい) 値になってしまいます。これでは、 μ からの外れ具合の激しさを測るという趣旨に合いません。(実際、本文でやったように「外れ具合の期待値」を求めると、常に $E[X - \mu] = E[X] - \mu = \mu - \mu = 0$ となってしまう、差がつきません)

期待値と同様、分散もやはり存在しないことがあります。まずそもそも、期待値が存在しなければ分散は定義されません。分散の定義の中で期待値を使っていたからです。逆に言うと、「分散が存在する」と書いたら、それだけでも期待値の存在も暗黙に主張することになります。また、たとえ期待値が存在しても、分散がいつも存在するとは限りません。具体例は次の例題 3.9 で。(分散についても、 $V[X] = \infty$ の場合は「分散は存在しない」と言うことにします)

例題 3.9

$P(X = k) = 2^{-k}$ ($k = 1, 2, \dots$) として、 $W \equiv \sqrt{2}^X$ という確率変数を考える。 W には期待値が存在するものの、分散は存在しない。これを確かめよ。

答

$$E[W] = \sum_{k=1}^{\infty} \sqrt{2}^k \cdot 2^{-k} = \sum_{k=1}^{\infty} \frac{1}{\sqrt{2}^k} = \frac{1/\sqrt{2}}{1 - (1/\sqrt{2})} < \infty \quad \text{等比級数 (付録 A.4.4(p.325))}$$

$$E[W^2] = \sum_{k=1}^{\infty} (\sqrt{2}^k)^2 \cdot 2^{-k} = \sum_{k=1}^{\infty} 1 \quad \dots\dots \text{発散}$$

よって、期待値 $E[W]$ は存在するが分散 $V[W] = E[W^2] - E[W]^2$ は存在しない (→ 3.4.6 項 (p.98) 「自乗期待値と分散」)。

定数 c に対して $V[c]$ という式があったら、やはり「確率 1 で c が出る確率変数」の分散だと考えてください (→ 1 章脚注*7(p.18))。 $E[c] = c$ より、答は $V[c] = E[(c - c)^2] = E[0] = 0$ です。

3.e 大数の法則についての補足

3.e.1 正確な記述 (弱法則と強法則)

3.5.3 項 (p.105) 「大数の法則」で述べた $V[Z_n] \rightarrow 0$ という結果から、任意の $\epsilon > 0$ に対して

$$P(|Z_n - \mu| > \epsilon) \rightarrow 0 \quad (n \rightarrow \infty)$$

が導かれます*5。正確にはこれを大数の法則 (あるいは大数の弱法則) と呼びます。

上式が導かれる理由は付録 C.1(p.347) 「確率変数・確率分布の収束」を参照。その用語で言えば、本文で述べたのは 2 次平均

*5 ギリシャ文字 ϵ (イブシロン) は、「微小な値」というニュアンスでよく使われます。

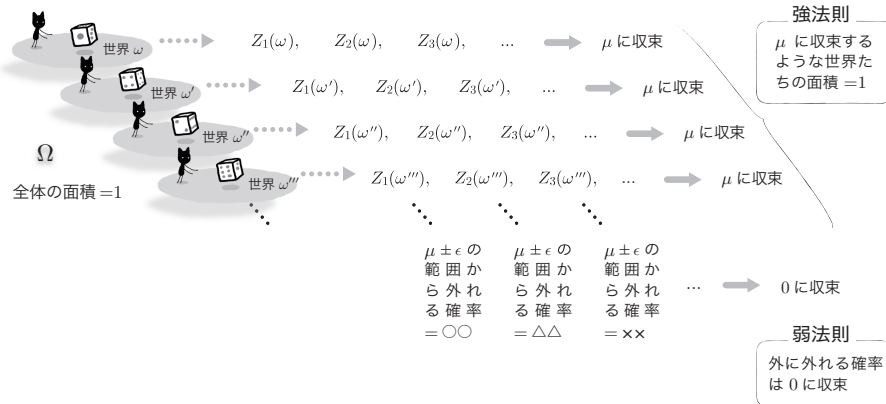
収束、上式は確率収束です。大数の（弱）法則は、正確には確率収束のほうを指します。

わざわざ弱法則と呼ぶのは、強法則というものもあるからです。期待値 μ が（有限値として）存在するときには、実は

$$P\left(\lim_{n \rightarrow \infty} Z_n = \mu\right) = 1$$

が証明されます。確率 1 で、 Z_n は μ に収束するわけです。これを大数の強法則と呼びます。付録 C.1(p.347)「確率変数・確率分布の収束」の用語で言えば、 μ に概収束することです。

両者の違いを理解するためには、人間視点を離れ、図 3.3 のように神様視点から世界たちを見下ろしたところを想像してください。強法則のほうが弱法則よりも強い主張となっています。



▶ 図 3.3 大数の弱法則と大数の強法則

? 3.d 強法則はなぜそんな回りくどい言い方をするの？ 要するに、必ず $\lim_{n \rightarrow \infty} Z_n = \mu$ だってことでしょ？

いいえ。そう書くと、すべての世界 ω で $\lim_{n \rightarrow \infty} Z_n(\omega) = \mu$ だという意味になります。これは言いすぎです。なぜなら、サイコロでたとえば永遠に 6 が出続けるような世界もありえて、そこでは $\lim_{n \rightarrow \infty} Z_n(\omega) = 6 \neq \mu$ だからです。強法則の結論は、そういう「 $Z_n(\omega)$ が μ に収束しない世界 ω 」たちの面積は 0 ということ。こういった面積ゼロの話については、4.2 節 (p.123)「確率ゼロ」も参照。

3.e.2 成り立たない例

大数の法則が成り立たない状況は直感的に想像しづらいかもしれませんが、期待値が存在しようがしまいが、たくさんくり返し試して平均を求めれば、ゆらぎは必ず小さくなるのではないかと。——これが実は保証できないことを本項で例示します。

シミュレーション

3.3.4 項 (p.84) や ? 3.c (p. 補足編 27) で紹介した「期待値の存在しない例」について、シミュレーション結果をご覧にしましょう。 X_1, \dots, X_{1000} を、 $P(X_n = k) = 1/2^k$ ($k = 1, 2, \dots$) という分布に従う i.i.d. な確率変数とします。この X_n には、ちゃんと期待値 $E[X_n] = 2$ が存在していました ($n = 1, \dots, 1000$)。一方、そこから作った $Y_n \equiv 2^{X_n}$ や $W_n \equiv S_n Y_n$ には、期待値は存在しないのでした (S_1, \dots, S_{1000} は i.i.d. で、 X_1, \dots, X_{1000} とは独立に ± 1 を確率半々でとる)。では、それぞれの平均はどんな振舞いをするでしょうか。

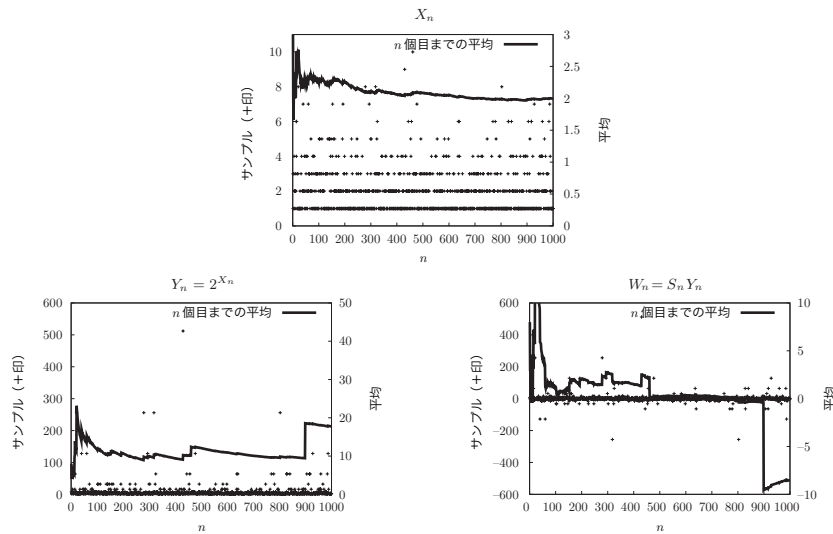
雰囲気を感じてもらうために、途中経過も含めてプロットしてみます。 n 個目までの平均を

$$\bar{X}_n \equiv \frac{X_1 + \dots + X_n}{n}, \quad \bar{Y}_n \equiv \frac{Y_1 + \dots + Y_n}{n}, \quad \bar{W}_n \equiv \frac{W_1 + \dots + W_n}{n}$$

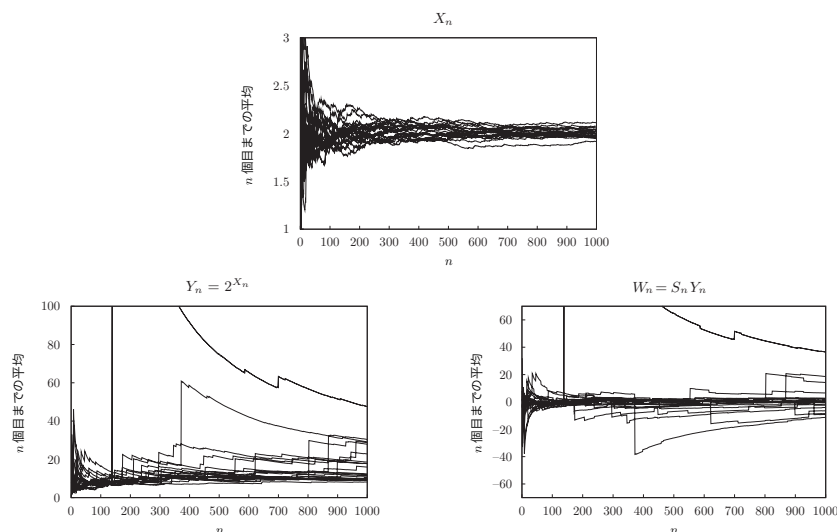
と置きましょう^{*6}。計算機実験で得られた結果の一例を図 3.4 に示します。 X_n と比べて Y_n や W_n では、512 といったとんでもない値が出ており、そういうとんでもない値に平均値がひっぱられてびくびくしている様子がグラフからうかがえます。

こんな実験を 20 回くり返し、結果を重ねたのが図 3.5 です。神様から見た「20 個の異なる世界 ω における結果」を模倣したものだと思ってください。 \bar{X}_{1000} はどれもほぼ 2 前後の値になりました。一方、 \bar{Y}_n や \bar{W}_n はいつまでもびくびくしてしまい、ばらつきがおさまらずにありまません。

^{*6} \bar{X}_n は、 X_1, \dots, X_n から作られた新しい確率変数です。 \bar{X} という記号で X の期待値を表す流儀もありますが、ここでは単に「 X とは別の記号」として \bar{X} を使っているだけ。混乱するようなら X' や \bar{X} などと書きかえてもらっても構いません。



▶ 図 3.4 平均の推移。上が X_n 、左下が $Y_n = 2^{X_n}$ 、右下が $W_n = S_n Y_n$ のグラフ。横軸はサンプルの個数 n 。折線はその個数までの平均 $\bar{X}_n, \bar{Y}_n, \bar{W}_n$ (目盛は右辺)。+ はサンプル X_n, Y_n, W_n (目盛は左辺)。 $Y_{461} = -W_{461} = 1024$ と $Y_{899} = -W_{899} = 8192$ はグラフからはみ出てしまいました



▶ 図 3.5 平均の推移。上が X_n 、左下が $Y_n = 2^{X_n}$ 、右下が $W_n = S_n Y_n$ のグラフ。横軸はサンプルの個数 n 。折線はその個数までの平均 $\bar{X}_n, \bar{Y}_n, \bar{W}_n$ 。20 回の実験結果を重ねて示す

見積り計算

いまのシミュレーションは $n = 1000$ でしたが、 n を 1 万とか 10 万とかもっと増やしていても、やはり収束はしません。個数を非常に増やすと、その分だけ非常にとんでもない値がまたどこかで出てしまい、平均が引っ込まれるからです。 Y_n について、そのことを数式で見えていきましょう。

何か正の数 a を指定し、平均 \bar{Y}_n が a 以上になる確率を井勘定で見積ります：もし Y_1, \dots, Y_n のどれかが an 以上なら、その一人の分だけでも平均 \bar{Y}_n が a 以上になってしまいます。ですから、 \bar{Y}_n が a 以上になる確率は、少なくとも

$$\begin{aligned} P(\bar{Y}_n \geq a) &\geq P(Y_1, \dots, Y_n \text{ のうちどれかが } an \text{ 以上}) \\ &= 1 - P(Y_1, \dots, Y_n \text{ がどれも } an \text{ より小さい}) \end{aligned}$$

だけはあるはずです。 Y_1, \dots, Y_n は i.i.d. なので、右辺は

$$1 - P(Y_1 < an) P(Y_2 < an) \cdots P(Y_n < an) = 1 - (P(Y_1 < an))^n$$

と書き直せます (→ 3.a.2 項 (p. 補足編 22) 「独立な確率変数たちの」最大値の分布)。ここで、 $k = 2, 3, 4, \dots$ に対して

$$P(Y_1 < 2^k) = P(X_1 < k) = \frac{1}{2^1} + \frac{1}{2^2} + \cdots + \frac{1}{2^{k-1}} = \frac{1/2 - 1/2^k}{1 - 1/2} = 1 - \frac{2}{2^k}$$

となっていたことを思い起こしてください (付録 A.4.4(p.325)「等比級数」)。 2^k 以外の u に対しても、 $2^k \sim 2^{k+1}$ の範囲で考えれば、 $P(Y_1 < u) \leq (1 - 2/2^{k+1}) \leq (1 - 1/u)$ となることがわかります。そうすると、 $P(\bar{Y}_n \geq a)$ は

$$1 - \left(1 - \frac{1}{an}\right)^n$$

以上だと見積られます。 $n \rightarrow \infty$ のとき、これは 0 にはならず、 $1 - e^{-\frac{1}{a}} (> 0)$ に収束するのです (A.2(p.328))。だから、たとえば「 \bar{Y}_n が 10 以上になる確率」は、 n をいくら大きくしたところで 0 には収束してくれず、どうしても一定以上は確率が残ってしまいます。「100 以上」でも「1000 以上」でも同様です。もし \bar{Y}_n が一定値に収束する (ゆらがなくなる) としたら、こんなことはありません。

W_n のほうは例題 C.2(p.352) を参照。

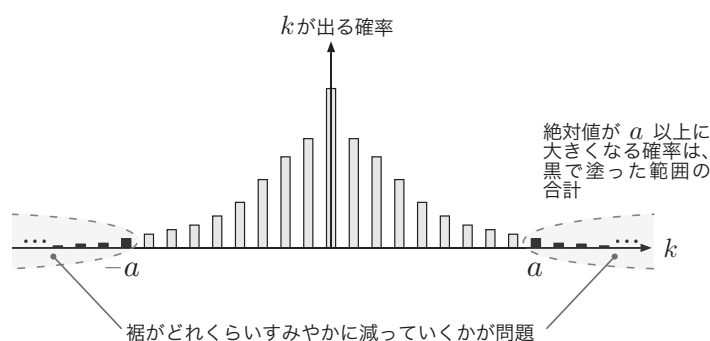
裾の重さ

X_n では大数の法則が成り立って、 Y_n や W_n では成り立たないという違いの源は何でしょうか。 X_n や Y_n や $|W_n|$ がある値 a 以上に大きくなる確率は、 $a \rightarrow \infty$ でどれも 0 に収束します (もしそうでないと、「確率の合計は 1」というおきてが破られる)。

$$P(X_n \geq a) \rightarrow 0, \quad P(Y_n \geq a) \rightarrow 0, \quad P(|W_n| \geq a) \rightarrow 0 \quad (a \rightarrow \infty)$$

問題はその収束の速さです。 X_n と比べて、 Y_n や $|W_n|$ のほうは収束が遅いのです。収束が遅いということは、とんでもない値がそれなりの確率で出てしまうことになります。

この収束の遅さを、統計業界では「分布の裾が重い」と表現します。図 3.6 のように確率分布をグラフで表したときの裾がなかなか減らないというイメージです。 X_n と比べて、 Y_n や W_n の分布は裾が重い。そのせいで期待値が存在せず、大数の法則も成り立ちません。



► 図 3.6 裾の重さ

3.f その他の補足

3.f.1 すべての自然数に対する一様分布？

(離散値で) 無限とおりの値をとる場合、一様分布 (3.1 項 (p.71)) は定義されません。ですからたとえば、「すべての自然数に対する一様分布」も定義されません*7。4.3.2 項 (p.131) の連続値版「一様分布」は、同じ名前でも意味がちよっと違います。

3.f.2 条件つき期待値の一般化

余談ついでに、さらに進んだ教科書向けの話も少しだけ。3.6 節 (p.107) の条件つき期待値 $E[Y|X]$ でやったことを端的に言えば、「 X で見分けられない範囲はならしてしまえ」という処理でした。これをもっと一般化した概念として、1.c.4 項 (p. 補足編 7) 「 \mathcal{F} の使い道」で述べたような集合族 (正確には σ 加法族) \mathcal{F}_1 を使った、 $E[Y|\mathcal{F}_1]$ という条件つき期待値もあります。「 \mathcal{F}_1 で見分けられない範囲はならしてしまえ」という処理で得られる確率変数が $E[Y|\mathcal{F}_1]$ です。 \mathcal{F}_1 は、パラレルワールドどうしの見分けがどこまでつくかを表すのでしたね。

*7 本書の目標レベルを全く逸脱しますが、この答だけでは不満な読者のために参考文献 [17] を挙げておきます (5.4 (a) 「 \mathcal{N}^+ 上の一様分布」)。

第4章の補足

4.a 面積ゼロが納得できない読者へ

面積ゼロの話はデリケートで、追究しだすとかなり本気の数学が必要になります。だから一般の読者は「そんなものか」と流していただければ結構です。以下の質疑は、それでもどうしても気になって仕方がない読者のためのものです。

？ 4.a 面積 0 というのは図形がないことを意味するはずだ。線分なら図形はあるのだから面積 0 はおかしい。

あなたの主張は「集合 B の面積が 0 なら B は空集合である」なわけですが、主張の根拠は何でしょうか。そんな気がしたという以上の根拠はないはずですよ。

？ 4.b 図 4.17^(p.124) が納得いきません。なぜ $0 + 0 + 0 + \cdots = 1$ になるの？ ゼロを何個足してもゼロなのに、これは矛盾でしょう。

矛盾ではありません。そこで言っていることは、 $0 + 0 + 0 + \cdots = 1$ という主張ではないからです。整理しましょう。

- 0 を何個足しても 0
- 点の面積は 0
- 正方形 Ω は点の集合
- 正方形 Ω の面積は 1

これらはどれも確かです。矛盾とを感じる理由は、点の面積を足したものが正方形の面積なはずだと素朴に思ったからでしょう。しかし、上で挙げたことがすべて正しいのなら、その素朴な思い込みは否定せざるをえません。いくら直感に合わないと不平を言ったところで、論理的に否定されてしまいます。

？ 4.c まだ納得いきません。 $0 + 0 + 0 + \cdots = 0$ と正方形の話とは何が違うのですか？

3.c.4 項 (p. 補足編 26) の無限和の説明をよく読むと、定義されたのはあくまで、無限数列の足し算でした。それをいまの例に適用しようと思ったら、正方形の話を数列の格好に翻訳する必要があります。そのためには、1 個目はこれ、2 個目はこれ、3 個目はこれ、……という調子で点たちを一行に並べてあげないとはいけません。順序はどうあれ、正方形に含まれる点たちをこんなふうに一個一個並べつくるものでしょうか。

実は、並べつくせないことが知られています。並べつくせるというのは、「通し番号 $0, 1, 2, 3, \dots$ をつけられる」、つまり「自然数 $0, 1, 2, 3, \dots$ の集合と一対一対応がつけられる」というのと同値。しかし、「正方形に含まれる点の集合」と「自然数の集合」とでは、一対一対応はつけられないことが証明されています (→ 付録 A.3.2^(p.320) 「無限集合の大小」)。

並べつくせないのでは、足し算を定義することができません。ざっくり言うと、点が多すぎるせいで「足す」という概念自体がナンセンスなのです。点の面積の合計がうんぬんという素朴な直感は、定義されない言葉を使っていて、真偽以前に意味を成していない。そんなわけで、「点の面積は 0、正方形の面積は 1」に何も矛盾はない。これが結論です。

？ 4.d だからやっぱり点の面積は無限小という答が合理的なのは？ 無限小が無限個あって、無限小 \times 無限大 $= 1$ 。これで問題ないように思いますが？

定義していない言葉を使ってはいけません。あなたはまだ「無限小」も「無限大」も定義していないし、もちろんそのかけ算も定義していません。意味の決まっていない言葉で主張を述べられても、反応しようがありません^{*1}。

じゃあ定義してやる、と奮起するかもしれませんが……不都合の出ないうまい定義を与えることは、簡単じゃないと思いますよ。もしすっきりうまく定義できるのだったら、それが教科書にのりそうなものです。でも、現実に教科書にのっているのは「 ϵ - δ 論法 (イプシロン-デルタ論法)」^{*}。これは、無限大や無限小という謎な数を導入したりせずに極限を扱うためのテクニックです。

^{*1} 筆者も面積や極限などを定義せず直感で話を進めていて、同罪に見えるかもしれません。でも実際は、結論が正しいこと

❓ 4.e でも高校の微積分では無限小が出てきましたよ。極限

$$\lim_{x \rightarrow 0} \frac{\sin x}{x}$$

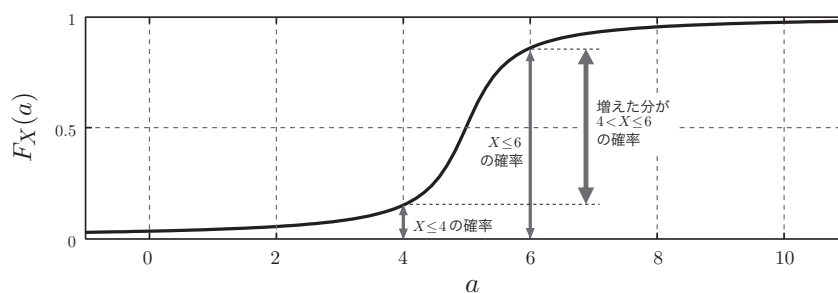
は無限小/無限小の格好だし、微分の説明も「 x が無限に小さな数 ϵ だけ変化したときの $f(x)$ の変化の割合」という調子でした。

それはいずれも非公式な説明で気持ちを述べただけです。オフィシャルな数学では上述の ϵ - δ 論法が使われます。

4.b 累積分布関数

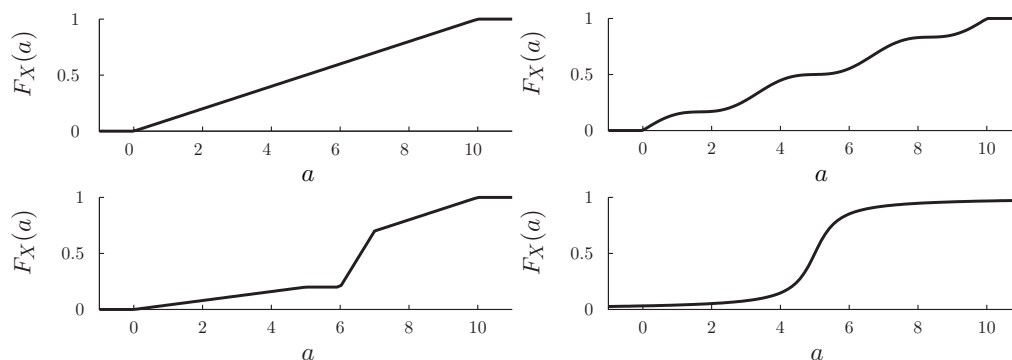
4.3.1 項 (p.126) で紹介した累積分布関数 $F_X(x) = P(X \leq x)$ について、次の性質が成り立ちます。4.1.1 項 (p.114) の消費インク量 F に戻って考えれば、どれもきつと納得いただけるでしょう。

- 左端での値は 0。 ($\lim_{a \rightarrow -\infty} F_X(a) = 0$)
- 右端での値は 1。 ($\lim_{a \rightarrow \infty} F_X(a) = 1$)
- 単調非減少 (広義単調増加) ……右へ行くにしたがって値が上がっていく。グラフが水平になることはあっても、下がることはない。 ($a \leq b$ なら $F_X(a) \leq F_X(b)$)
- $P(a < X \leq b) = F_X(b) - F_X(a)$ (ただし $a < b$ とする) ……図 4.1 のとおり*2。



► 図 4.1 累積分布関数 F_X から区間の確率を求める

いろいろな累積分布関数の例を図 4.2 に挙げます。



► 図 4.2 累積分布関数 F_X の例

例題 4.1

4.2.1 項 (p.123) と同じ Ω で、 $Z(u, v) = u + v$ のとき F_Z は？

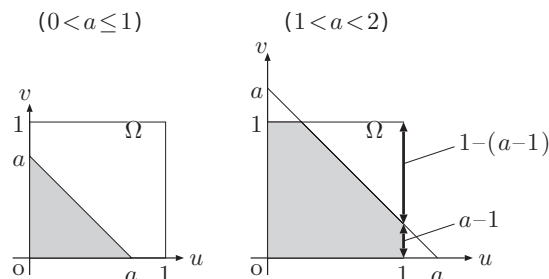
を (直感でなく) 別途確かめた上でそうしています。

*2 $a < X \leq b$ と $a \leq X \leq b$ との違いは、今はこだわらなくて結構です。4.3 節 (p.126) 冒頭で想定したように典型的な場合なら、「 a ぴったり」が出る確率はどうせゼロだから、 $P(a < X \leq b)$ でも $P(a \leq X \leq b)$ でも値は同じ。

答

図 4.3 より次のとおり。 $1 < a < 2$ の場合の答は、正方形の面積 1 から右上の白い三角形の分を引けば求められます。

$$F_Z(a) = P(Z \leq a) = \begin{cases} 0 & (a \leq 0) \\ a^2/2 & (0 < a \leq 1) \\ -(a^2/2) + 2a - 1 & (1 < a < 2) \\ 1 & (2 \leq a) \end{cases}$$

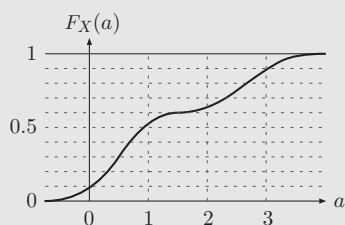


▶ 図 4.3 $Z(u, v) = u + v \leq a$ を満たす、 Ω 上の領域

■

例題 4.2

F_X のグラフが下の図のようになっていたとして、 $p_1 \equiv P(0 < X \leq 1)$, $p_2 \equiv P(1 < X \leq 2)$, $p_3 \equiv P(2 < X \leq 3)$ の中で一番大きいのはどれか？



答

$p_1 = F_X(1) - F_X(0)$ と $p_2 = F_X(2) - F_X(1)$ と $p_3 = F_X(3) - F_X(2)$ でどれが一番大きい。答は p_1 。 ■

4.c 典型的でない場合

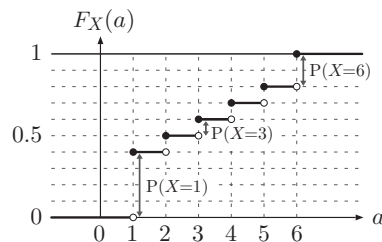
ここまでは、ある値びつたりが出る確率はことごとくゼロ（どんな a でも $P(X = a) = 0$ ）という、連続値では典型的な場合を想定してきました。本節では、この想定が成り立たない一般の場合について補足します。初心者は本節をスキップして構いません。

4.c.1 典型的でない場合の累積分布関数

整数は実数の一部ですから、整数値の確率変数 X も、実数値の確率変数の一種とみなせます。では、4.b 項 (p. 補足編 32) の累積分布関数をこの X へ適用したらどうなるでしょうか。

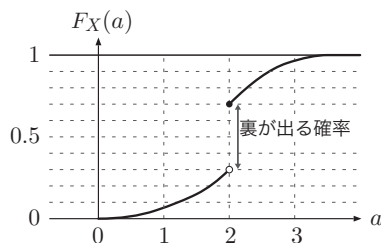
たとえばいかさまサイコロを例にとって、定義 $F_X(a) = P(X \leq a)$ のとおり忠実にグラフを描けば、図 4.4 が得られます。 $F_X(0.7)$ も $F_X(0.9)$ も $F_X(0.99999)$ も 0 ですが、 $F_X(1)$ は $P(X \leq 1) = 0.4$ 。その先はまた $F_X(1) = F_X(1.2) = F_X(1.9) = F_X(1.99999)$ ですが、 $F_X(2)$ は $P(X \leq 2) = 0.4 + 0.1 = 0.5$ 。こんな調子ですから、 $F_X(1)$ のところで $P(X = 1)$ の分だけジャンプして、 $F_X(2)$ のところで $P(X = 2)$ の分だけジャンプして、……という階段状のグラフになります。

値	その値が出る確率
1	0.4
2	0.1
3	0.1
4	0.1
5	0.1
6	0.2



► 図 4.4 整数値の確率分布（いかさまサイコロ）の累積分布関数 F_X

ついでにもっとひねくれた例も挙げておきましょう。ゲタをでたために放り投げて、表が出たら X はその飛距離、裏が出たら $X = 2$ とします。これは実数値と離散値のハイブリッドのような確率変数になって、その累積分布関数は図 4.4 のように不連続になります。



► 図 4.5 不連続な累積分布関数 F_X

いまの例のように、ある値ぴったりが出る確率が > 0 だと、累積分布関数はそこでジャンプすることになります。つまり、左から近づけた極限 $\lim_{u \uparrow a} F_X(u)$ と右から近づけた極限 $\lim_{u \downarrow a} F_X(u)$ とが一致しないわけです^{*3}。実はこれらは、

$$\text{左から: } \lim_{u \uparrow a} F_X(u) = \lim_{u \uparrow a} P(X \leq u) = P(X < a)$$

$$\text{右から: } \lim_{u \downarrow a} F_X(u) = \lim_{u \downarrow a} P(X \leq u) = P(X \leq a)$$

となっています^{*4}。だから、 a ぴったりが出る確率がもし > 0 だと、両者は一致しません。

例題 4.3

$a \leq X \leq b$ となる確率を累積分布関数 F_X で表せ ($a \leq b$)。

答

$P(X < a) + P(a \leq X \leq b) = P(X \leq b)$ より、 $P(a \leq X \leq b) = P(X \leq b) - P(X < a) = F_X(b) - \lim_{u \uparrow a} F_X(u)$ 。

4.c.2 典型的でない場合の確率密度関数

確率密度関数 f_X というのは、累積分布関数 F_X を微分したものでした (4.3.1 項 (p.126))。しかし F_X が必ず微分できるとは限りません。実際、前項のように F_X がジャンプして不連続になっていたら微分なんてできません。だからそんな分布は確率密度関数では表せません。とはいえ、応用場面で扱う実数値の確率分布は確率密度関数で表せるものがほとんどですから、あまり心配しなくて結構です。

? 4.f 確率密度関数で表せない分布でも、累積分布関数でなら表せますか？

はい。実数値の確率分布はすべて累積分布関数で表せます。ただし、それがきれいな式で書けるかや、そのグラフが絵に描けるようなたちのいい姿をしているかは、また別の話^{*5}。実はここに、累積分布関数のもう一つの不満点があります。最も重要な実数値分布である正規分布 (→ 4.6 節 (p.161)) の累積分布関数は、加減乗

^{*3} $\lim_{u \uparrow a}$ は、「 $u < a$ という制限のもとで u を a に近づけた極限」の意味です。 $\lim_{u \rightarrow a-0}$ と書くこともあります。 $\lim_{u \downarrow a}$ はその反対で、「 $u > a$ という制限のもとで u を a に近づけた極限」の意味になります。単に $\lim_{u \rightarrow a}$ と書いたら、

$\lim_{u \rightarrow a+0}$ も $\lim_{u \rightarrow a}$ も同じ値になるという意味です。

^{*4} 右からのほうは要するに $\lim_{u \downarrow a} F_X(u) = F_X(a)$ です (累積分布関数の右連続性)。このあたりの議論は本書のレベルを越えるので、深入りしません。なお、4 章脚注*4(p.127) のように累積分布関数の定義が違う本だと、本項の内容もそれぞれ連動して変わります。

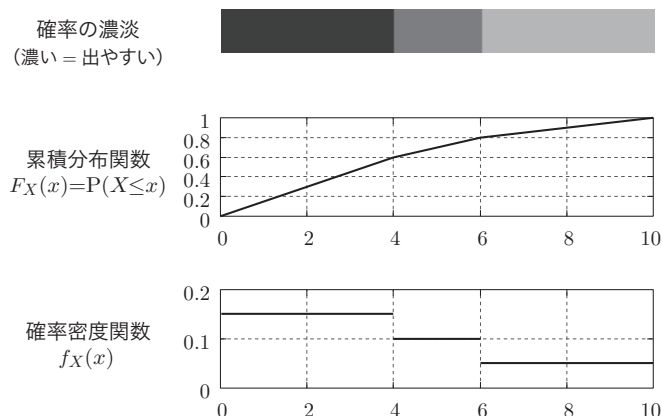
除や指数対数などのおなじみの関数では書き下せません (→ 付録 A.5.2^(p.328)「ガウス積分」)。一方、確率密度関数なら、数式としてきれいに正規分布を表すことができます。

一般の方はここまで読んでいただければひとまず十分です。この先は話がどんどんマニアックになっていきますから、いつでも離脱して先へ進んで構いません。

さて、 F_X を微分したものが確率密度関数 f_X だ、という本文の定義は実は不正確です。正しくは、「積分すると F_X になるような (非負の) 関数 f_X 」を確率密度関数と呼びます。つまり、

$$\int_{-\infty}^b f_X(x) dx = F(b) \quad (4.1)$$

となるような関数 $f_X(x) \geq 0$ があれば、それが確率密度関数です^{*6}。この定義なら、図 4.6 のように累積分布関数 F_X が折れているときでも確率密度関数が使えます。



▶ 図 4.6 累積分布関数 F_X が折れている例

ただしちょっとやっかいなことに、「積分すると F_X になるような関数」は一つに決まりません。実際、いまの図 4.6 では、 $f_X(4) = 0.15$ と定めても $f_X(4) = 0.1$ と定めても式 (4.1) は成り立ちます。それどころか、折れとぜんぜん関係ないところもいじれてしまいます。たとえば、 $f_X(3)$ の値だけは特例で 999 ということにして、それ以外の x では $f_X(x)$ はグラフどおりとすれば、やはり式 (4.1) は満たされます。結局、一点や二点いじったところで積分には影響が出ないということです。だから、そういう積分に影響しない程度の違いは無視するという立場で議論をする必要があります。詳しくは測度論を勉強して、「ほとんど至るところ」(almost everywhere, 略して a.e.) という概念について調べてください^{*7}。実は 4.3.2 項 (p.131) の一様分布の説明は、以上の話をこっそりふまえたものでした。区間 $[a, b]$ 上の一様分布の累積分布関数は、 a や b において折れています。

そんなわけで、図 4.22^(p.129) での $f_X(x) \geq 0$ の説明も厳密には不適切です。 $f_X(x) \geq 0$ を本文のように導くことは本当はできません。だから、 $f_X(x) \geq 0$ は確率密度関数の性質ではなく定義 (→ 2 章脚注*13^(p.61)) とされます。でも読者はとりあえず本文のように思っておいてください。道の真中を歩きましょう (→ 4.3 節 (p.126)「確率密度関数」冒頭の引用)。

? 4.g じゃあ、累積分布関数が連続だったら、必ず確率密度関数で表せると思っていいですか？

厳密に答えると、いいえ。実はさらにひねくれた病的な例 (カントールの「悪魔の階段」) が知られています。連続で、ほとんど至るところ水平なのに、たどるとなぜか登っているという奇妙な関数です。累積分布関数がこんなものだと、確率密度関数では表せません。

? 4.h 離散値も連続値も統一的に扱いたいのですが、累積分布関数だと見つらいです。サイコロみたいな離散値の場合も含めて、確率密度関数で何とかありませんか？

図 4.7 のような確率密度関数 $f(x)$ を作れば、サイコロの目に近い連続値の確率分布が表せます。しかも、柱の幅をせまくすればするほど、サイコロの本当の分布に近づきます。

そこで、図 4.8 のような確率密度関数の「極限」を仮想的に考えて、それをデルタ関数 $\delta(x)$ と呼びます。

^{*5} 関数という概念は、「加減乗除や log や sin などの組合せで書けるもの」には限りません。たとえば「今日の正午の各地の気温」も、地球上の各点で定義された関数です。

^{*6} 「 f_X の積分が F_X なんだから、やっぱり F_X を微分したら f_X じゃないのか？」と混乱してしまった読者は、解析学の教科書をもう一度確認してください。きちんとした本なら、「 f_X が連続な場合」といった但し書きがきつとどこかについているはずですが。

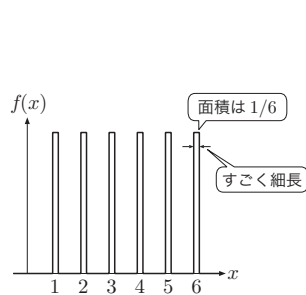
^{*7} 「ほとんど確実に」(almost surely, 略して a.s.) や「ほとんど全ての」(almost all, 略して a.a.) という言い回しもあります。

デルタ関数を使えば、サイコロの目 X の分布を

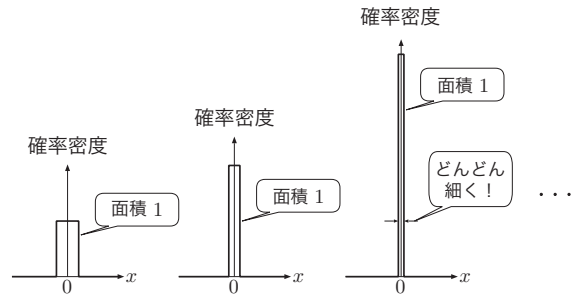
$$f_X(x) = \frac{1}{6} \left(\delta(x-1) + \delta(x-2) + \delta(x-3) + \delta(x-4) + \delta(x-5) + \delta(x-6) \right)$$

のように確率密度関数風に表すことができます。 $\delta(x-c)$ は、 $x=c$ のところに細い針が立つ格好です。

ただし、いまの説明は厳密な数学としては意味をなしていません。「極限」という言葉が何のことなのか定義されていないからです。きちんとした説明が知りたければ**超関数**という用語を調べてください。



▶ 図 4.7 サイコロの確率密度関数？



▶ 図 4.8 デルタ関数のイメージ

4.d 一変数の確率分布についての補足

4.d.1 定義と記法

確率変数 X が離散値でも連続値でも、**確率分布**の本来の定義は、^{P.4.4(p.130)} のとおり「 X の値が集合 A に属する確率」の一覧だと思ってください。離散値のとき（特に X のとり得る値が有限通りのとき）は、 $P(X=a)$ を全部の a について特定しておけば、そこから

$$P(X \text{ の値が } A \text{ に属する}) = \text{「} A \text{ の各要素 } a \text{ に対する } P(X=a) \text{ の合計」}$$

も特定できます。だから離散値のときは「 $P(X=a)$ の一覧」だけでもよかったのです。

ただし^{P.4.4(p.130)} の説明には不正確なところがあります。それは、確率の一覧が満たすべき条件を明示していないことと、「あらゆる集合 A 」と言っていることです。前者については 1.a 節 (p. 補足編 5) 「 (Ω, \mathcal{F}, P) の詳細」を参照。後者については、「区間に分解できる話と、その極限などにより導ける話」だけを扱います。こちらは 1.c 節 (p. 補足編 5) 「 \mathcal{F} は何者か？」にからむ話です。

$P(X=x)$ という記法にあわせて確率密度関数 $f_X(x)$ も $f(X=x)$ と書きたくなるかもしれませんが、それはおすすめしません。確率密度については、 $X=2Y$ だったとしても $f_X(6) \neq f_Y(3)$ だからです (4.3.3 項 (p.132))。確率の計算 $P(X=6) = P(2Y=6) = P(Y=3)$ と混同しないためには、記法も変えておくのが安全でしょう。

1.6 節 (p.18) 「現場流の略記法」の連続値版、すなわち

- 流儀 I: X の確率密度関数を g 、 Y の確率密度関数を h などと書く
- 流儀 II: $f(x)$ と書いたら $f_X(x)$ の意味、 $f(y)$ と書いたら $f_Y(y)$ の意味

は、慣れてきたら使っても結構です。前者を使う場合は、何の確率密度関数を何で表すのか、そのつどちゃんと宣言してください。後者については、 $f(3)$ と書いたときどちらの意味かわからなくなってしまうので、混乱しないように注意が必要です。

ちなみに筆者の周囲では、「確率は大文字の P 、確率密度は小文字の p 」なんていう流儀も見かけます。このあたりは業界によっていろいろ慣習があるかもしれません。

4.d.2 累積分布関数の変換

4.3.3 項 (p.132) のような確率密度関数の変換と比べて、累積分布関数

$$F_X(a) = P(X \leq a)$$

の変換にはあまり大した話はありません。変換公式としてパターンを覚えようなんて思うとかえって混乱しそうですから、定義に戻って地道に計算するほうがおすすめです。

たとえば、 $Y = 3X - 5$ として、 $P(Y \leq 4)$ を X の式に書き直してみてください。答は、

$$P(Y \leq 4) = P(3X - 5 \leq 4) = P(3X \leq 9) = P(X \leq 3)$$

というだけのこと。こんな調子で、 Y の累積分布関数が

$$F_Y(b) = P(Y \leq b) = P(3X - 5 \leq b) = P\left(X \leq \frac{b+5}{3}\right) = F_X\left(\frac{b+5}{3}\right)$$

と計算されます。

もう一例。 $Y = X^3$ として、 $P(Y \leq 8)$ はどうなるでしょうか。同じようにやれば

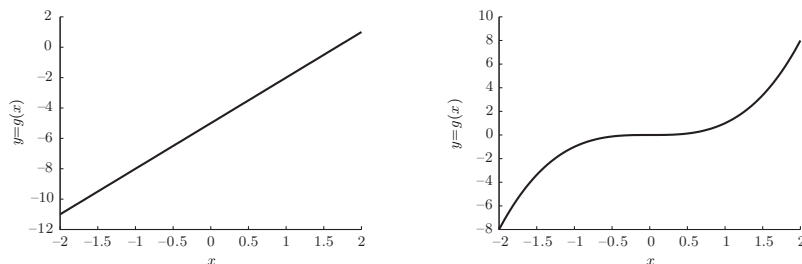
$$P(Y \leq 8) = P(X^3 \leq 8) = P(X \leq 2)$$

です。累積分布関数も、

$$F_Y(b) = P(Y \leq b) = P(X^3 \leq b) = P\left(X \leq \sqrt[3]{b}\right) = F_X\left(\sqrt[3]{b}\right)$$

で問題ありません。

連続で、しかも「 X が増えれば Y も増える」というすなおな変換なら、悩みどころは特にありません。上の変換 $g(x) = 3x - 5$ や $g(x) = x^3$ がそうになっていること（連続で、「 $u < v$ なら $g(u) < g(v)$ 」が成り立つこと）を図 4.9 で確認しておいてください。



▶ 図 4.9 連続で、しかも「 x が増えれば y も増える」というすなおな変換 g 。左は $g(x) = 3x - 5$ 、右は $g(x) = x^3$

4.d.3 すなおでない変数変換

本文の 4.1.3 項 (p.119) 「印刷したものを伸縮させるとインクの濃さはどうなるか」や 4.3.3 項 (p.132) 「確率密度関数の変数変換」では、すなおな変換だけを扱いました。すなおでない変換だと何が起きるのかを本項で補足します。

$y = x^2$ みたいに「二枚重ね」となるような変換を施す場合は、密度も二枚分それぞれ計算して重ねあわせないとはいけません。以下の例題 4.4 を参考にしてください。

例題 4.4

次の場合に、変換後の位置 $y = 4$ および $y = -4$ におけるインク密度をそれぞれ求めよ。

$$\text{インクの濃さ } f(x) = \begin{cases} 0.0008(x-5)^4 & (0 \leq x \leq 10) \\ 0 & (\text{他}) \end{cases}$$

$$\text{変換 } y \equiv g(x) = (x-5)^2$$

答

$y = 4$ に対応するのは $x = 3$ と $x = 7$ 。 $g'(x) = 2(x-5)$ より $g'(3) = -4$, $g'(7) = 4$ 。よって $y = 4$ における変換後の密度は

$$\left| \frac{f(3)}{g'(3)} \right| + \left| \frac{f(7)}{g'(7)} \right| = 0.0008 \cdot \frac{16}{4} + 0.0008 \cdot \frac{16}{4} = 0.0064$$

また、 $y = -4$ に対応する x はないから、 $y = -4$ における密度は 0。 ■

なお、いまの例題 4.4 で、変換後の位置 $y = 0$ においてはちょっと困りごとが生じます。 $y = 0$ に対応するのは $x = 5$ 。そのとき $g'(5) = 0$, $f(5) = 0$ なので、式 (4.1) (p.121) にそのまま代入すると $0/0$ になる。これをどう解釈するかは……こだわらないことにしましょう。どうせ一点だけのことですから (→ 4.c 節 (p. 補足編 33) 「典型的でない場合」)。

累積分布関数についても、すなおでない変換だといろいろうさぎ問題が生じます。以下の例題を参考にしてください。

例題 4.5

$Y = -2X$ のとき、 $F_Y(4)$ を F_X で表せ。

答

不等号の向きに注意して、

$$F_Y(4) = P(Y \leq 4) = P(-2X \leq 4) = P(X \geq -2) = 1 - P(X < -2) = 1 - \lim_{u \uparrow -2} F_X(u)$$

最後の変形については 4.c 節 (p. 補足編 33) 「典型的でない場合」を参照。 ■

例題 4.6

$Y = X^2$ のとき、 $F_Y(4)$ と $F_Y(-4)$ をそれぞれ F_X で表せ。

答

X の値と Y の値とが一対一対応ではないため、前問とはまた別の注意が必要です。

$$F_Y(4) = P(Y \leq 4) = P(X^2 \leq 4) = P(-2 \leq X \leq 2) = F_X(2) - \lim_{u \uparrow -2} F_X(u)$$

$$F_Y(-4) = P(Y \leq -4) = P(X^2 \leq -4) = 0 \quad \cdots \cdots (X^2 \leq -4 \text{ はおこりえないから})$$

■

4.e 多変数の確率分布についての補足

4.e.1 実数値の同時分布とは結局何か

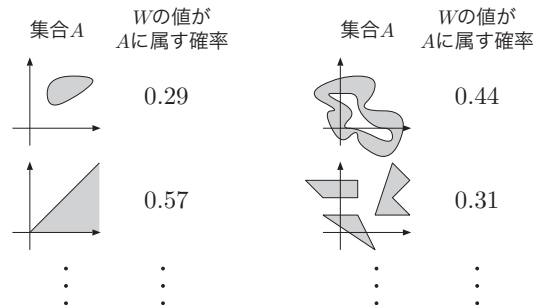
4.4.1 項 (p.136) では同時分布をさらっと導入しました。イメージはあれでいいとしても、実数値の同時分布とは結局何のことでしょうか。? 4.4 (p.130) で述べた一変数の話をすなおに拡張して考えてください。すなわち、

- $X \leq Y$ である
- $1000 \sin(X^3 + e^Y)$ の百の位が 7 である
- 北緯 X 度・東経 Y 度が海である
- ……

のようなあらゆる条件に対して「 (X, Y) がその条件を満たす確率」を特定した一覧が、確率変数 X, Y の同時分布です*8。

同時分布の定義はこうも言い換えられます。実数値の確率変数 X, Y に対して、それを組にした $W \equiv (X, Y)$ を考えてください。 W は、2 次元ベクトルを値とする確率変数と解釈できます。この W の確率分布は、

2 次元空間内のあらゆる領域 (集合) A に対して、確率 $P(W \text{ の値が } A \text{ に属する})$ を特定したものとして定義するのが自然でしょう。図 4.10 のようなイメージです。こうして定義された W の確率分布をもって、 X と Y の同時分布と呼びます。「条件」を「その条件を満たす値の集合」に翻訳しただけなので、どちらの言い方でも結局は同じこと。



▶ 図 4.10 確率分布 = あらゆる集合 A に対する確率の一覧

3 個についても同様です。 $W \equiv (X, Y, Z)$ という 3 次元ベクトル値の確率変数を考えて、3 次元空間内のあらゆる領域 (集合) A に対して $P(W \text{ の値が } A \text{ に属する})$ を特定したものが、 X, Y, Z の同様分布です。

? 4.i 混乱してきた。いまの絵と「パラレルワールド全体の集合 Ω 」の絵とは何が違うんですか？

確率変数と確率分布との違いをよく思い出してください。 Ω のほうは、一点一点がそれぞれ一つのパラレルワールドでした。確率変数 X, Y の値は、パラレルワールドごとに違ってきます。一方、いまの絵の点、たとえば点 $(5.8, 3.1)$ は、「 $X = 5.8$ かつ $Y = 3.1$ 」という条件を表しているだけです。この条件に対応するパラレルワールド (その世界での X, Y の値がこうなっているようなパラレルワールド) は、いくつもあるかもしれないし、あるいは一つもないかもしれません。——まだ納得がいかなければ、1.7.1 項 (p.19) 「 Ω の正体にはこだわらない」もふり返ってみましょう。 Ω の形にはあまり意味はなく、円板だろうと球体だろうと構わないのでした。

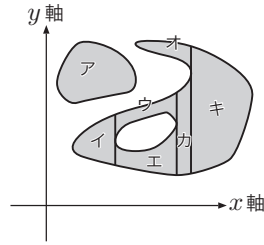
4.e.2 不正確だった点

4.4.3 項 (p.139) で同時分布から周辺分布を求める際に、「出やすさを集めて」などと言っているのは、説明としてはインチキです。あの説明では確率と確率密度との区別をすつとぼけました。また、数式による説明も不正確でした。正確には「積分に影響しない程度の違いは無視する前提のもとで」という但し書きが必要なのですが、本書レベルでは特にこだわらなくて結構です。この断りは以後いちいち述べません (→ 4.c.2 項 (p. 補足編 34))。

4.4.4 項 (p.142) の条件つき確率でもし $f_X(a) = 0$ だったときには、 $f_{Y|X}(b|a)$ は不定として、? 2.5 (p.42) と同様「てきとうに」扱うこととします。

図 4.42 (p.149) のように積分で確率を求める方法は、凹みや空洞があっても使えます。そんな場合は、図 4.11 のようにパーツに分けてそれぞれに入る確率を計算し、結果を合計してください。

*8 厳密には、「あらゆる」は言いすぎですが (→ 4.d.1 項 (p. 補足編 36))。以後も同様。



► 図 4.11 (アミがけ内に入る確率) = (アに入る確率) + ... + (キに入る確率)

4.e.3 同時分布の累積分布関数

4.b 項 (p. 補足編 32) の累積分布関数を同時分布へ拡張するのは簡単です。

$$F_{X,Y}(x,y) \equiv P(X \leq x \text{ かつ } Y \leq y)$$

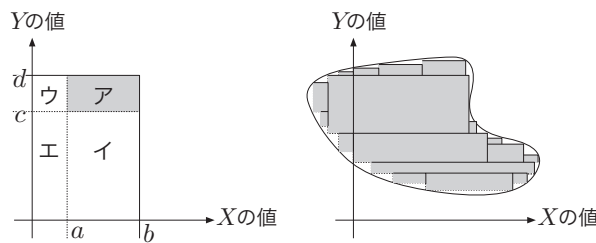
というのを考えてやればよい。こんな $F_{X,Y}$ が与えられていれば、図 4.12 左のアミがけ部分のような長方形領域に入る確率も

$$P(a < X \leq b \text{ かつ } c < Y \leq d) = F_{X,Y}(b,d) - F_{X,Y}(b,c) - F_{X,Y}(a,d) + F_{X,Y}(a,c)$$

と表せます。実際、上式の右辺は図でいうと

$$(\text{ア} + \text{イ} + \text{ウ} + \text{エ}) - (\text{イ} + \text{エ}) - (\text{ウ} + \text{エ}) + \text{エ} = \text{ア}$$

になっています (式中の「ア」は「 (X,Y) がアに入る確率」の意味です)。長方形領域の確率が得られれば、たとえば図 4.12 右のような領域に入る確率も、極限として表すことができます。ですから、 $F_{X,Y}$ でちゃんと確率分布を表せて (さまざまな領域の確率を特定できて) います。……と一応説明しましたが、実際には、同時分布の表現法としてこれはあまり使われません。主役はやはり確率密度関数です。



► 図 4.12 いろいろな領域に入る確率を $F_{X,Y}$ で表す

同時分布に対する累積分布関数と確率密度関数との関係は、次のとおり。まず、定義から

$$F_{X,Y}(x,y) = \int_{-\infty}^y \left(\int_{-\infty}^x f_{X,Y}(u,v) du \right) dv = \int_{-\infty}^x \left(\int_{-\infty}^y f_{X,Y}(u,v) dv \right) du \quad (4.2)$$

です⁹。多変数の解析学を勉強した方は、このとき

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$$

となることもおわかりでしょう。 $\frac{\partial^2 \square}{\partial x \partial y}$ は $\frac{\partial}{\partial x} \left(\frac{\partial \square}{\partial y} \right)$ と解釈してください。

3 変数以上でも全く同様です。

4.e.4 独立性への補足

4.4.6 項 (p.146) で最初に紹介した独立性の条件 $f_{Y|X}(b|a) = f_Y(b)$ では、「 $f_X(a) \neq 0$ の a について」というただし書きが本当は必要です (ゼロ割り防止のため。他も同様)。一方、分母をはらった $f_{X,Y}(a,b) = f_X(a)f_Y(b)$ ではそんなただし書きはいりません。この事情は離散値のときの 2 章脚注*15(p.64) と同じです。また、 $f_{X,Y}(x,y) = g(x)h(y)$ の形に書けることが独立性を意味する理由も、離散値のときの例題 2.17(p. 補足編 17) と同様です (足し算のところを積分におきかえればよい)。

確率密度関数が存在しない場合の独立性については、次のように解釈してください。確率変数 X と Y が独立であることの本当の定義は、

$$P(\text{「}X\text{に関する条件」 かつ 「}Y\text{に関する条件」}) = P(X\text{に関する条件}) P(Y\text{に関する条件}) \quad (4.3)$$

が常に成り立つことです。たとえば

$$P(X > 7 \text{ かつ } Y \leq 6) = P(X > 7) P(Y \leq 6)$$

$$P(\text{「}\sin X \text{ の小数点以下 3 桁目が奇数」 かつ 「北緯 } Y \text{ 度・東経 } (Y^2 + Y + 1) \text{ 度が海」})$$

$$= P(\sin X \text{ の小数点以下 3 桁目が奇数}) P(\text{北緯 } Y \text{ 度・東経 } (Y^2 + Y + 1) \text{ 度が海})$$

⁹ 積分範囲の指定「 $-\infty$ から x まで」で文字 x がすでに使われているので、積分の中身 (走る変数) には何か別の空いている文字 (ここでは u) を使わないといけません。付録 A.4(p.322) 「総和 \sum 」も参照。

⋮

など。これなら離散値でも連続値でも、確率密度関数で書いても書けなくても、そんなこと関係なく独立性を定義できます。

本文で述べた独立性の定義は、実は (4.3) の特別な場合でした。もし X, Y が離散値だったら、「(4.3) が常に成り立つことは

$$P(X = a, Y = b) = P(X = a)P(Y = b) \quad (\text{すべての } a, b \text{ で成立})$$

と同値」です（正確には、とり得る値が高々可算個だったら→付録 A.3.2^(p.320)）。また、もし X, Y の分布が確率密度関数で表されるなら、「(4.3) が常に成り立つことは

$$f_{X,Y}(a, b) = f_X(a)f_Y(b) \quad (\text{すべての } a, b \text{ で成立}) \quad (4.4)$$

と同値」です。この二通りがカバーできれば実用的には十分なので、本文ではそのように説明しました。

ちなみに、累積分布関数を用いれば、独立性は

$$F_{X,Y}(a, b) = F_X(a)F_Y(b) \quad (\text{すべての } a, b \text{ で成立})$$

とも表現できます。

最後にもう一言、「 X と Y が独立なら $g(X)$ と $h(Y)$ も独立だ」を指摘しておきます。 g, h は任意の関数です。上の「本当の定義」にしたがえばこれは自明でしょう。離散値版の 2.5.3 項^(p.63)とも見比べてください。

? 4.j なぜ (4.3) と (4.4) とが同値なのか？

もし (4.3) が常に成り立つなら、

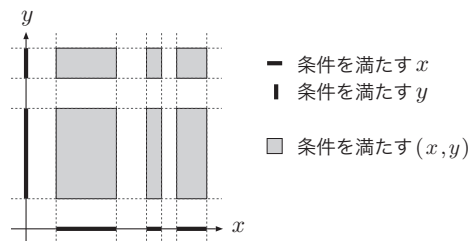
$$\begin{aligned} P(a_1 \leq X \leq a_2 \text{ かつ } b_1 \leq Y \leq b_2) &= P(a_1 \leq X \leq a_2)P(b_1 \leq Y \leq b_2) \\ &= \left(\int_{a_1}^{a_2} f_X(x) dx \right) \left(\int_{b_1}^{b_2} f_Y(y) dy \right) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f_X(x)f_Y(y) dy dx \end{aligned}$$

つまり、 $f_X(x)f_Y(y)$ を積分すると確率が出るのですから、この $f_X(x)f_Y(y)$ がまさに $f_{X,Y}(x, y)$ であるはず。よって (4.4) が導かれます。

逆に、もし (4.4) が常に成り立つなら、

$$\begin{aligned} &P(\text{「}X\text{に関する条件」 かつ 「}Y\text{に関する条件」}) \\ &= \iint f_{X,Y}(x, y) dy dx = \iint f_X(x)f_Y(y) dy dx \\ &\quad (\text{積分範囲は「条件を満たす } (x, y) \text{ 全体」}) \\ &= \left(\int f_X(x) dx \right) \left(\int f_Y(y) dy \right) \\ &\quad (\text{積分範囲は「条件を満たす } x \text{ 全体」と「条件を満たす } y \text{ 全体」}) \\ &= P(X \text{ に関する条件})P(Y \text{ に関する条件}) \end{aligned}$$

だから、(4.3) が導かれます^{*10}。積分範囲については図 4.13 を参照。



▶ 図 4.13 式 (4.3) のための積分範囲（各範囲の確率を合計）

4.e.5 すなおでない変数変換

4.4.7 項^(p.148)への補足として、すなおでない変数変換だとどんなことになるかを少しお話します。

折り返し

確率変数 X, Y に対して $Z \equiv X, W \equiv |Y|$ という変換を例にしましょう。これはシートを x 軸で折り返すような変換です。面積の拡大縮小はありませんが、シートに重なりが生じることに注意しないといけません。変換後の位置 (z, w) に対応する元の位置は、

1. $w \geq 0$ の場合、 $(x, y) = (z, \pm w)$ 。

^{*10} 本書では、こんなふうに積分範囲を別途明記したら、 \int だけでも定積分を表すことにさせていただきます。

2. $w < 0$ の場合、どこも対応せず
したがって、確率密度関数は

$$f_{Z,W}(z, w) = \begin{cases} f_{X,Y}(z, w) + f_{X,Y}(z, -w) & (w \geq 0) \\ 0 & (w < 0) \end{cases}$$

となります ($w = 0$ の扱いについては、4.c.2 項 (p. 補足編 34) のとおりではありません)。

伸縮と折り返しが入り混じったもっとややこしい例も、考えようと思えば考えられます。でも要領はもうつかめたでしょうか、本書ではここまでにしておきます。

ぺちゃんこ

行列式 $\det A = 0$ の場合、正方行列 A をかける変数変換をしたら、変換後の同時分布は確率密度関数では表せません。 A が正則行列でない、これはぺちゃんこにつぶすという変換になるからです。ぺちゃんこのイメージが湧かない方は、参考文献 [32] やそのサポートページの「アニメーションで見る線形代数」をご覧ください。URL は次のとおりです (2009 年 10 月現在)。
<http://www.ohmsha.co.jp/data/link/4-274-06578-2/anime/index.html>

次元を減らす

たとえば、 X, Y の確率密度関数 $f_{X,Y}(x, y)$ から $Z \equiv X^2 + Y^2$ の確率密度関数 $f_Z(z)$ を求めたいとします。これは 2 次元を 1 次元へ移す話なので、本文の変数変換の議論をあてはめられません。

手はいくつか考えられますが、最も素朴なのは、

$$f_Z(z) = \lim_{\epsilon \downarrow 0} \frac{P(z \leq Z \leq z + \epsilon)}{\epsilon}$$

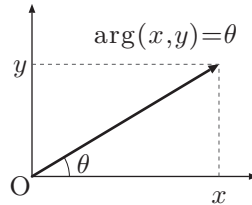
という計算法でしょう。そのバリエーションで、累積分布関数 $F_Z(z) = P(Z \leq z)$ を求めて

$$f_Z(z) = \frac{d}{dz} F_Z(z)$$

を使う手もあります。また、一対一変換と周辺分布とを組合せる手も考えられます。いまの例なら、

- $W \equiv \arg(X, Y)$ のようにもう一つ確率変数を用意する (図 4.14 を参照)
- (X, Y) から (Z, W) への変換は ($0 \leq W < 2\pi$ など適当に範囲制限すれば) 一対一になるので、本文のように確率密度関数 $f_{Z,W}(z, w)$ を計算する
- $f_{Z,W}(z, w)$ から、周辺分布の確率密度関数 $f_Z(z) = \int_0^{2\pi} f_{Z,W}(z, w) dw$ を計算する

すぐ後の 4.f.1 項 (p. 補足編 42) 「独立な確率変数たちの 合計値の分布」でもこの手を使います。例題 5.15 (p. 213) も参照。



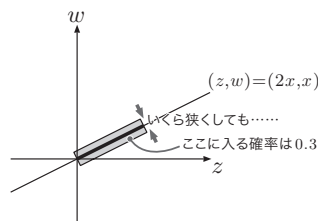
► 図 4.14 ベクトル (x, y) の偏角をここでは $\arg(x, y)$ と表す

❓ 4.k 混乱してきた。前の「ぺちゃんこ」と今の「次元を減らす」とは何が違うの？

前のは、行き先の空間が 2 次元なのにその中の 1 次元領域へとすべてを移していました。だから 2 次元空間としての密度 (つまり面積あたりの密度) は発散してしまいました。一方、今のは行き先の空間自体が 1 次元です。だから計算した答は、1 次元空間としての密度 (つまり長さあたりの密度) です。

次元を増やす

さっきと反対に次元を増やしてしまったら、こんどは確率密度関数では表せなくなります。たとえば $P(0 \leq X \leq 1) = 0.3$ だったとして、 $(Z, W) \equiv (2X, X)$ のように次元を増やしてみましょうか。すると図 4.15 のように、面積ゼロの線分上へ確率 0.3 が集中しています。したがって、平面上の密度 (= 確率/面積) としては発散してしまいます。



► 図 4.15 確率密度関数では表せない

4.f (独立な確率変数たちの) 合計値や最大値の分布

3.a 節 (p. 補足編 20) の整数値のときの話を、実数値の場合にも翻訳しておきます。整数値版をスキップした方も本節は読んでみるようお勧めします。実は整数値版より実数値版のほうが話がすっきりしているし、しかも後で使うからです (→ 4.g.3 項 (p. 補足編 44) の正規分布の足し算)。

4.f.1 (独立な確率変数たちの) 合計値の分布

先に結論を書いておきます。

実数値の確率変数 X, Y の確率密度関数を $f_X(x), f_Y(y)$ とする。 X, Y が独立なら、その合計値 $Z \equiv X + Y$ の確率密度関数は、

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(s)f_Y(z-s)ds = \int_{-\infty}^{\infty} f_X(z-s)f_Y(s)ds \quad (4.5)$$

結論はいつものように、整数版 3.a.1 項 (p. 補足編 20) の \sum が \int に化けただけ。(4.5) の積分も、整数値のときと同様に畳み込みと呼びます。ですから標語としては、整数値でも実数値でも同じです：独立な確率変数の合計値の分布は畳み込みで計算される。

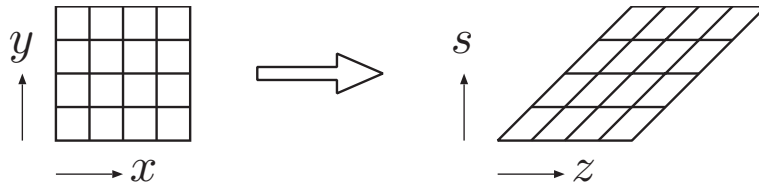
この結論を導くには、変数変換を活用する方法がとっと早いでしょう。 X, Y を $Z \equiv X + Y, S \equiv Y$ へと変数変換してやります。変換の様子は図 4.16 のとおりで、この変換の面積拡大率は 1 です^{*11}。したがって、 Z, S の同時分布の確率密度関数は

$$\begin{aligned} f_{Z,S}(z,s) &= f_{X,Y}(x,y) \quad \text{ただし } z = x + y, s = y \\ &= f_{X,Y}(z-s,s) \end{aligned}$$

するとその周辺分布として、 Z の確率密度関数が計算できます。

$$f_Z(z) = \int_{-\infty}^{\infty} f_{Z,S}(z,s)ds = \int_{-\infty}^{\infty} f_{X,Y}(z-s,s)ds$$

ここで、 X, Y は独立という前提でしたから $f_{X,Y}(z-s,s) = f_X(z-s)f_Y(s)$ と書きかえられます。よって (4.5) が得られました。(もう一方の表現も、 $S \equiv X$ において同様の計算をすれば得られます)



▶ 図 4.16 この変換の面積拡大率は 1

4.f.2 (独立な確率変数たちの) 最大値の分布

実数値の独立な確率変数 X, Y に対して、今度は両者の大きいほうの値

$$W \equiv \max(X, Y)$$

を考えましょう。 W の確率分布を求めることが目標です。

3.a.2 項 (p. 補足編 22) の整数値版の筋を思い出すと、 $\circ\circ$ 以下になる確率を考えるのがミソでした。本章の用語で言えば累積分布関数を考えたわけです。実数値でも同様に、累積分布関数で表現すると分布がすぐ求められます。

$$F_W(w) = P(W \leq w) = P(X \leq w \text{ かつ } Y \leq w) = P(X \leq w)P(Y \leq w) = F_X(w)F_Y(w)$$

途中で「かつ」がかけ算になったのは、 X と Y が独立という前提だからです。もし確率密度関数がほしいなら、この累積分布関数から計算すればよい。計算法は前にもやりました。

$$f_W(w) = F'_W(w) = F'_X(w)F_Y(w) + F_X(w)F'_Y(w) \quad (\text{かけ算の微分を使用})$$

累積分布関数と確率密度関数との関係を思い出せば次のようにも表せます。

$$f_W(w) = f_X(w)F_Y(w) + F_X(w)f_Y(w) = f_X(w) \int_{-\infty}^w f_Y(y)dy + f_Y(w) \int_{-\infty}^w f_X(x)dx$$

(この式をわざわざ覚える必要はありません。最大値の分布は累積分布関数で考えるとよい、ということだけ頭に入れておけば、本項の計算なんてその場でささっと自作できるはずです。)

4.g 正規分布についての補足

4.g.1 記法の濫用

正規分布を $N(\mu, \sigma^2)$ で表すと 4.6.2 項 (p.164) で述べましたが、だからといって $X \sim N(\mu, \sigma^2)$ のことを $X = N(\mu, \sigma^2)$ や $f_X = N(\mu, \sigma^2)$ とはふつう書きません。 X は確率変数、 f_X は確率密度関数、 $N(\mu, \sigma^2)$ は確率分布を表しますから、等号で結ぶのは不自然です (→ 1.5 節 (p.16) 「確率分布」)。

^{*11} この変換は行列で次のように表されますから、確かに $\det A = 1$ です。

$$\begin{pmatrix} z \\ s \end{pmatrix} = A \begin{pmatrix} x \\ y \end{pmatrix}, \quad A \equiv \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

4.g.2 正規分布の確率密度関数の積分（誤差関数 erf）

たとえば標準正規分布で -3 以上 3 以下の値が出る確率を求めようと思ったら、確率密度関数の定積分を求める必要があります。しかし $\exp(-z^2/2)$ の積分は、おなじみの関数（ $\sqrt{\quad}$ や \log や \sin など）ですばつとは表せません。だから昔は数表を使っていた。今なら、付録 A.5.2(p.328)「ガウス積分」で紹介している誤差関数 $\text{erf}(t) \equiv (2/\sqrt{\pi}) \int_0^t \exp(-x^2) dx$ に帰着させてコンピュータで数値計算します。

$$\begin{aligned} & \int_{-3}^3 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz \quad (u^2 = z^2/2 \text{ となるよう } u = z/\sqrt{2} \text{ において置換積分}) \\ &= \sqrt{2} \int_{-3/\sqrt{2}}^{3/\sqrt{2}} \frac{1}{\sqrt{2\pi}} \exp(-u^2) du \\ &= \frac{1}{\sqrt{\pi}} \int_{-3/\sqrt{2}}^{3/\sqrt{2}} \exp(-u^2) du \quad \text{ここで } \exp(-u^2) \text{ の対称性を利用して……} \\ &= \frac{2}{\sqrt{\pi}} \int_0^{3/\sqrt{2}} \exp(-u^2) du = \text{erf}(3/\sqrt{2}) \approx 0.997 \end{aligned}$$

より一般には、直後の例題 4.7 で求める累積分布関数 $F_Z(a)$ を使って、

$$\begin{aligned} P(Z \leq a) &= F_Z(a) \\ P(Z \geq a) &= 1 - F_Z(a) \\ P(a \leq Z \leq b) &= F_Z(b) - F_Z(a) \quad (\text{ただし } a \leq b) \end{aligned}$$

により計算してください。いまは不等号に $=$ が入っても入らなくても確率は同じです（ \rightarrow 4.3.1 項 (p.126) 「確率密度関数」）。

例題 4.7

標準正規分布に従う Z の累積分布関数 $F_Z(a)$ は、誤差関数 erf を用いて

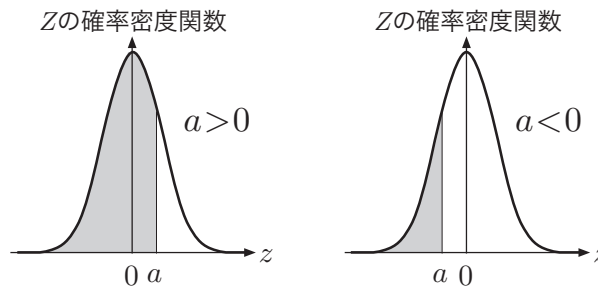
$$F_Z(a) = \frac{1}{2} \left(1 + \text{erf} \left(\frac{a}{\sqrt{2}} \right) \right)$$

と書ける。これを示せ。

答

$$\begin{aligned} F_Z(a) &= P(Z \leq a) = \int_{-\infty}^a \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz \\ &= \int_{-\infty}^0 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz + \int_0^a \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz \end{aligned}$$

もし a が負だったときも、 $\int_0^a = -\int_a^0$ を思い出せば上式で大丈夫なことがわかります（図 4.17 を見ながら確認してください）。上式の第 1 項は、正規分布の左右対称性から $1/2$ 。また、第 2 項は先ほどと同様の置換積分により $(1/2) \text{erf}(a/\sqrt{2})$ となります。よって、 $F_Z(a) = (1/2)(1 + \text{erf}(a/\sqrt{2}))$ です。 ■



► 図 4.17 累積分布関数の計算

例題 4.8

$X \sim N(50, 100)$ のとき、 $P(X \leq 60)$ を誤差関数 erf で表せ。

答

$Z \equiv (X - 50)/\sqrt{100}$ は標準正規分布に従う。また、 $X \leq 60$ は $Z \leq 1$ と同値。すると本問は例題 4.7 に帰

着する。

$$P(X \leq 60) = P(Z \leq 1) = \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{1}{\sqrt{2}} \right) \right)$$

4.g.3 独立な正規分布の足し算

4.6.2 項 (p.164) では、独立な正規分布の足し算がまた正規分布になるという結果だけを述べました。この結果は、付録 C.2(p.350) の特性関数というものを經由すれば簡単に導けます。でもここでは特性関数を使わず、力まかせに計算して見せます。

確率変数 X, Y が独立で、どちらも正規分布に従うとしましょう (期待値や分散は同じでなくても構いません)。 $W \equiv X + Y$ の確率密度関数を求めたいのですが、式がごちゃつくのをやわらげるために、興味のない定数はすべて \square で表すことにします。いまは格好が正規分布になることさえ示せばよいので、こんな手抜き計算でも十分です。

まず、独立な確率変数の合計値の分布 (→ 4.f.1 項 (p. 補足編 42)) ですから、

$$f_W(w) = \int_{-\infty}^{\infty} f_X(w-s) f_Y(s) ds = \square \int_{-\infty}^{\infty} \exp \left(-\frac{(w-s-\square)^2}{\square} \right) \exp \left(-\frac{(s-\square)^2}{\square} \right) ds$$

この積分の中身を s について整理し、平方完成してやると

$$\begin{aligned} \exp \left(-\frac{(w-s-\square)^2}{\square} \right) \exp \left(-\frac{(s-\square)^2}{\square} \right) &= \exp \left(-\frac{(w-s-\square)^2}{\square} - \frac{(s-\square)^2}{\square} \right) \\ &= \exp \left(\square s^2 + (\square w + \square)s + (\square w^2 + \square w + \square) \right) \\ &= \exp \left(\square (s - (\square w + \square))^2 \right) \exp(\square w^2 + \square w + \square) \end{aligned}$$

これにより、

$$f_W(w) = \square \left\{ \int_{-\infty}^{\infty} \exp \left(\square (s - (\square w + \square))^2 \right) ds \right\} \exp(\square w^2 + \square w + \square)$$

と変形されます (s によらないものは積分の外に出しました)。さらに、残った積分は実は w によらない定数になるので*12、結局

$$f_W(w) = \square \exp(\square w^2 + \square w + \square) = \text{定数} \cdot \exp(w \text{ の } 2 \text{ 次式})$$

という格好。これでもう正規分布だと判断できることは式 (4.7)(p.166) で指摘しました。

4.g.4 中心極限定理

ちゃんと断わりませんでしたでしたが、4.6.3 項 (p.167) 「中心極限定理」では期待値も分散も有限値として存在することを前提に話を進めています。

細かいゆらぎが無数に合わさって目に見えるゆらぎが起きるという想定について、もっと詳しい説明が読みたければ、参考文献 [10] が丁寧におすすめです。

中心極限定理の主張である式 (4.8)(p.168) から、

$$P(W_n < a) \rightarrow \text{「標準正規分布 } N(0, 1) \text{ で } a \text{ より小さい値が出る確率」} \quad (n \rightarrow \infty)$$

$$P(a \leq W_n \leq b) \rightarrow \text{「標準正規分布 } N(0, 1) \text{ で } a \text{ 以上 } b \text{ 以下の値が出る確率」} \quad (n \rightarrow \infty)$$

なども言っていることに注意してください*13。また、中心極限定理が成り立つときには、任意の有界連続関数 g に対して*14

$$E[g(W_n)] \rightarrow E[g(Z)] \quad (n \rightarrow \infty) \quad \text{ただし } Z \sim N(0, 1)$$

となることも保証されます (付録 C.1.4(p.349) 「法則収束」)。

中心極限定理を根拠にするなどで話を正規分布に限定してしまえば、いろいろな議論が非常に単純になります。それはうれしいことなのですが、単純すぎて退屈な面もあります。より一般の分布に目を向けると、正規分布では見られなかった豊かな現象があふれています。参考文献 [32] の 0.2 節でほのめかされているような「線形の退屈さ」「非線形ならではの豊かな現象」とも似ていますね。



4.1 中心極限定理はよくわかりました。なるほど。だから 3.2 節 (p.74) の 2 項分布 $Bn(n, p)$ は、図 4.18 のように n を増やすと正規分布になっていくんですね？

そのセリフでは本当に理解できているのか心配になります。次の二点をしっかり確認してください。

一点目は、 \sqrt{n} で割ることを忘れていないかです。生の 2 項分布を見るのでは、 n が増えるにつれて分散がどんどん大きくなりますから、特定の分布には収束しません。

二点目は、確率と確率密度との区別です。2 項分布は離散値なので、確率密度関数では表せません。

? 4.h (p. 補足編 35) のデルタ関数を使って強引に確率密度関数風に表してみても、べったり 0 の途中にぽつぽつと針が立っている格好です。だから、確率密度の値そのものが正規分布の確率密度へ収束するというイメー

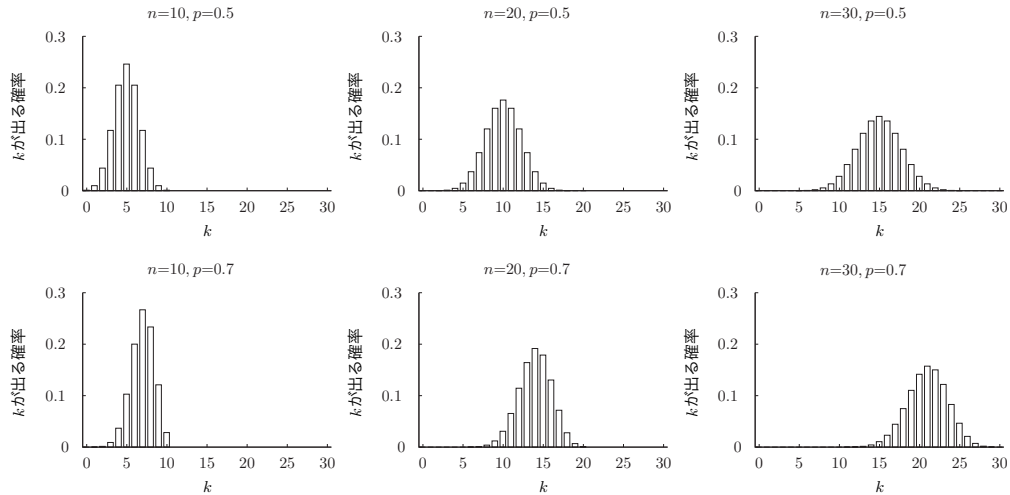
*12 $t = s - (\square w + \square)$ において t の積分に直してみる (置換積分) と、 w を含まない式になります。

*13 第一式は $-X_1, \dots, -X_n$ に中心極限定理をあてはめれば導ける。第二式は $P(W_n \leq b) - P(W_n < a)$ により第一式から得られる。

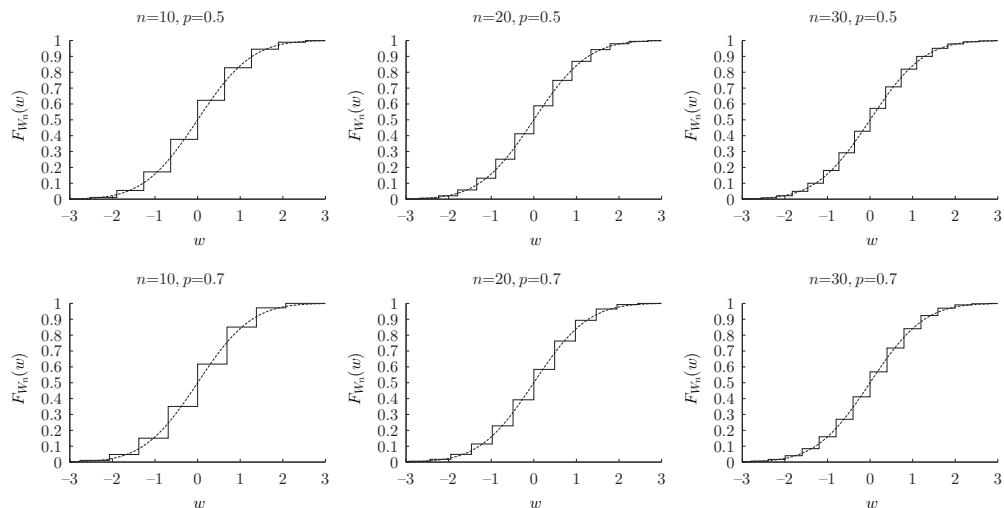
*14 関数 g が有界とは、「 $|g(x)| \leq c$ が必ず成り立つような定数 c が存在する」という意味です。要はある有限範囲に値がおさまっているということ。

ジは不適切です。

では正しくは何が何に収束するのか。各 X_i を「確率 p で 1、確率 $(1-p)$ で 0」が出る i.i.d. な確率変数としたとき、 $S_n \equiv X_1 + \cdots + X_n$ は 2 項分布 $\text{Bn}(n, p)$ に従うのでした ($0 < p < 1$)。中心極限定理をあてはめるには、生の S_n でなく $W_n \equiv (S_n - np)/(\sqrt{n}\sigma)$ を考えないといけません ($\sigma \equiv \sqrt{V[X_1]} = \sqrt{p(1-p)}$)。この W_n の累積分布関数 $F_{W_n}(x)$ が、図 4.19 のように各点 x で標準正規分布 $N(0, 1)$ の累積分布関数へと収束します。それが中心極限定理の主張です。



► 図 4.18 いろいろな 2 項分布 $\text{Bn}(n, p)$ (図 3.2^(p.74) の再掲)



► 図 4.19 2 項分布 $\text{Bn}(n, p)$ に対する中心極限定理。実線は $S_n \sim \text{Bn}(n, p)$ から作った $W_n \equiv (S_n - np)/\sqrt{np(1-p)}$ の分布、点線は標準正規分布 (いずれも累積分布関数)

? 4.m 前の? 4.l で、 S_n が $N(np, n\sigma^2)$ へ収束すると言ってはいけませんか?

($n \rightarrow \infty$ のとき) $\bigcirc\bigcirc$ へ収束すると言いながらその $\bigcirc\bigcirc$ が n に応じて動くのでは、収束先を明示したことになります。あなたが言いたかったことはこうですか?

「 S_n の累積分布関数」と「 $Y_n \sim N(np, n\sigma^2)$ の累積分布関数」との差 $F_{S_n}(x) - F_{Y_n}(x)$ が $n \rightarrow \infty$ で 0 に収束する

これなら確かに成り立ちます。でもこれはあまり有意義な主張ではありません。なぜなら、(x を固定したとき) $n \rightarrow \infty$ では $F_{S_n}(x) \rightarrow 0$, $F_{Y_n}(x) \rightarrow 0$ だからです。それならわざわざ Y_n なんて持ち出さず、 $\lim_{n \rightarrow \infty} F_{S_n}(x) = 0$ と言っても同じこと。そしてこの式は、分布が結局どこまでも広がっていくことしか述べていません。一方、中心極限定理は、分布の広がり具合や形状についてもっと具体的な情報を与えてく

れます。

4.g.5 中心極限定理の証明にかえて

中心極限定理の証明は本書の目標レベルを越えるので省きます。ここでは証明のかわりに、どんなことが起きているかの一端を観察してみましょう。4.6.3 項 (p.167) と同様に、 X_1, \dots, X_n (i.i.d.) の期待値と標準偏差がそれぞれ 0 と $\sigma > 0$ だとして、 $W_n \equiv (X_1 + \dots + X_n)/(\sqrt{n}\sigma)$ を考えます。

作り方から、1 乗や 2 乗の期待値 $E[W_n] = 0$, $E[W_n^2] = 1$ は常に一定でした (期待値が 0 なので $E[W_n^2] = V[W_n]$ です)。それなら 3 乗の期待値 $E[W_n^3]$ はどうなるでしょうか。

$$\begin{aligned} E[W_n^3] &= \frac{1}{\sqrt{n}^3 \sigma^3} E[(X_1 + X_2 + \dots + X_n)(X_1 + X_2 + \dots + X_n)(X_1 + X_2 + \dots + X_n)] \\ &= \frac{1}{\sqrt{n}^3 \sigma^3} (E[X_\bigcirc^3] \text{ たちの合計}) + \frac{1}{\sqrt{n}^3 \sigma^3} (E[X_\bigcirc^2 X_\Delta] \text{ たちの合計}) \\ &\quad + \frac{1}{\sqrt{n}^3 \sigma^3} (E[X_\bigcirc X_\Delta X_\square] \text{ たちの合計}) \end{aligned}$$

のように展開して、グループごとに調べていきます (○△□は互いに異なる値を表す)。第一グループは、

$$\begin{aligned} \frac{1}{\sqrt{n}^3 \sigma^3} (E[X_\bigcirc^3] \text{ たちの合計}) &= \frac{1}{\sqrt{n}^3 \sigma^3} (E[X_1^3] + \dots + E[X_n^3]) \\ &= \frac{1}{\sqrt{n}^3 \sigma^3} \cdot n E[X_1^3] \quad \dots \text{i.i.d. なので } E[X_1^3] = \dots = E[X_n^3] \\ &= \frac{1}{\sqrt{n} \sigma^3} E[X_1^3] \rightarrow 0 \quad (n \rightarrow \infty) \end{aligned}$$

のように 0 へ収束します。一方、第二グループと第三グループはもともと 0 です。なぜなら、独立性より「かけ算の期待値は期待値のかけ算」となるので、

$$\begin{aligned} E[X_\bigcirc^2 X_\Delta] &= E[X_\bigcirc^2] E[X_\Delta] = E[X_\bigcirc^2] \cdot 0 = 0 \\ E[X_\bigcirc X_\Delta X_\square] &= E[X_\bigcirc] E[X_\Delta] E[X_\square] = 0 \cdot 0 \cdot 0 = 0 \end{aligned}$$

だからです。以上より、 $n \rightarrow \infty$ のとき $E[W_n^3] \rightarrow 0$ だとわかりました。この収束先は、 $Z \sim N(0, 1)$ に対する $E[Z^3] = 0$ と一致しています。

もう一段やってみます。 $E[W_n^4]$ はどうなるでしょうか。

$$\begin{aligned} E[W_n^4] &= \frac{1}{n^2 \sigma^4} E[(X_1 + \dots + X_n)(X_1 + \dots + X_n)(X_1 + \dots + X_n)(X_1 + \dots + X_n)] \\ &= \frac{1}{n^2 \sigma^4} (E[X_\bigcirc^4] \text{ たちの合計}) + \frac{1}{n^2 \sigma^4} (E[X_\bigcirc^2 X_\Delta^2] \text{ たちの合計}) \\ &\quad + \frac{1}{n^2 \sigma^4} (E[X_\bigcirc^3 X_\Delta] \text{ たちの合計}) + \frac{1}{n^2 \sigma^4} (E[X_\bigcirc^2 X_\Delta X_\square] \text{ たちの合計}) \\ &\quad + \frac{1}{n^2 \sigma^4} (E[X_\bigcirc X_\Delta X_\square X_\star] \text{ たちの合計}) \end{aligned}$$

のように展開して、またグループごとに調べていきます (○△□☆は互いに異なる値を表す)。第一グループは先ほどと同様の理屈で 0 に収束します。第三・四・五グループも先ほどと同様の理屈でもともと 0 です。残るは第二グループだけになりました。個々の $E[X_\bigcirc^2 X_\Delta^2]$ は、独立性より「かけ算の期待値は期待値のかけ算」がまた効いて、

$$E[X_\bigcirc^2 X_\Delta^2] = E[X_\bigcirc^2] E[X_\Delta^2] = \sigma^2 \cdot \sigma^2 = \sigma^4$$

それが何個あるかという (→ 3.2.2 項 (p.75) 「補足：順列 $n P_k$ ・組合せ $n C_k$ 」) 、

- ○と△に何が入るかで $n C_2$ とおり (○ < △ とした)
- $X_\bigcirc, X_\bigcirc, X_\Delta, X_\Delta$ の並べかえで $4 C_2$ とおり

のように数えれば

$$n C_2 \cdot 4 C_2 = \frac{n(n-1)}{2!} \cdot \frac{4 \cdot 3}{2!} = 3n(n-1) \text{ 個}$$

よって第二グループは結局

$$\frac{1}{n^2 \sigma^4} (E[X_\bigcirc^2 X_\Delta^2] \text{ たちの合計}) = \frac{1}{n^2 \sigma^4} \cdot 3n(n-1) \sigma^4 = 3 \left(1 - \frac{1}{n}\right) \rightarrow 3 \quad (n \rightarrow \infty)$$

以上より、 $n \rightarrow \infty$ のとき $E[W_n^4] \rightarrow 3$ だとわかりました。この収束先は、またもや $E[Z^4]$ と一致しています^{*15}。

^{*15} 付録 A.5.2(p.328) 「ガウス積分」より $E[Z^4] = \int_{-\infty}^{\infty} z^4 \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz = 3$ 。なお実は、こんなにかんばって具体的に数えあげなくても、次のようにショートカットして $E[Z^4]$ との一致を主張することもできます：本文の議論から、 $a \equiv \lim_{n \rightarrow \infty} E[W_n^4]$ は X_i の分布の詳細によらずに定まるはず。ならば、 X_1, \dots, X_n のかわりに、i.i.d. な標準正規分布に従う $\tilde{X}_1, \dots, \tilde{X}_n \sim N(0, 1)$ を考えても答は一致するはず。つまり、対応する $\tilde{W}_n \equiv (\tilde{X}_1 + \dots + \tilde{X}_n)/\sqrt{n}$ の $\tilde{a} \equiv \lim_{n \rightarrow \infty} E[\tilde{W}_n^4]$ は a と同じ値になる。ところが \tilde{W}_n の分布はふたたび標準正規分布 $N(0, 1)$ になるのだから、 \tilde{a} は $E[Z^4]$ と当然等しい。

あとはもう同様^{*16}。こんなふうに、 W_n に関するいろいろな期待値が

- $n \rightarrow \infty$ のときある値に収束する。
- しかも、収束先は X_i の分布の詳細によらない。
- そして、この収束先が標準正規分布 $N(0, 1)$ のそれと一致する。

……という現象が生じていると思ってください。いくらイメージがわいたでしょうか。

なお、上の議論では $E[X_i^3]$ や $E[X_i^4]$ が (有限値として) 存在することを暗黙に仮定していました。付録 C.2(p.350) の特性関数を用いた本当の証明では、そんな仮定は必要ありません。

4.g.6 連続修正

2 項分布 $Bn(100, 1/2)$ で 60 以下の値が出る確率を、近似でいいから求めたいとしましょう。こんなときは中心極限定理を使って正規分布に換算するのが定石ですが、その換算の際に一工夫することで近似精度を上げる小技が知られています。**連続修正**と呼ばれる技です。これは**不連続修正**や**連続補正**とも呼ばれます。

具体的にやってみましょう。 S が 2 項分布 $Bn(100, 1/2)$ に従うとき、

$$W \equiv \frac{S - 50}{\sqrt{100 \cdot 1/4}} = \frac{S}{5} - 10$$

の分布は標準正規分布 $N(0, 1)$ で近似されるのでした (例題 4.23(p.170))。ここで、 $P(S = k)$ を $P(k - 1/2 < S \leq k + 1/2)$ と解釈し直すのが連続修正のアイデアです^{*17}。そうすると、

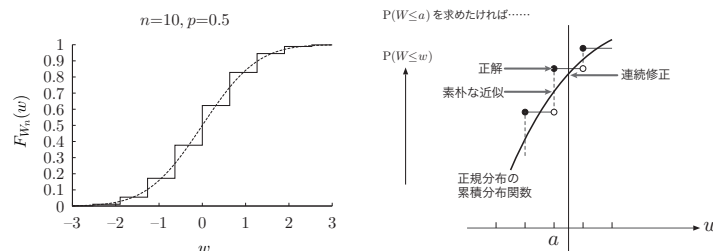
$$P(S \leq 60) = P(S = 0) + \cdots + P(S = 60)$$

は $P(-0.5 < S \leq 60.5)$ と解釈し直されます。あるいは、どうせ $S \geq 0$ しか出ないのだから、単に $P(S \leq 60.5)$ と言っても同じことです。ここまでは言葉の遊びみたいなもので、 S は整数値しかとらないので、もちろん結局は $P(S \leq 60) = P(S \leq 60.5)$ です。

$S \leq 60$ と $S \leq 60.5$ をそれぞれ W に換算したら、 $W \leq 2$ と $W \leq 2.1$ 。これを正規分布で近似する段階で、両者の違いが効いてきます。

- 素朴な近似: 「 $N(0, 1)$ において 2 以下の値が出る確率 (約 0.9772)」を答える
- 連続修正: 「 $N(0, 1)$ において 2.1 以下の値が出る確率 (約 0.9821)」を答える

実は後者のほうが正解 (約 0.9824) に近い値を与えます。そのことは図 4.20 から察せられるでしょう (白丸と黒丸にとまどった方は 4.c.1 項 (p. 補足編 33) 「典型的でない場合の確率密度関数」を参照)。



▶ 図 4.20 連続修正を施したほうが正解に近くなる (左は図 4.19(p. 補足編 45) の再掲)

例題 4.9

標準正規分布 $N(0, 1)$ において w 以下の値が出る確率を $F(w)$ とおく。「サイコロを 60 回ふったときに 1 が 15 回以上出る確率」を中心極限定理で見積り、この F で表せ。ただし連続修正を施すこと。

答

i 回目に 1 が出たら $X_i \equiv 1$ 、他の目が出たら $X_i \equiv 0$ とおく ($i = 1, \dots, 60$)。 X_1, \dots, X_{60} は i.i.d. であり、

$$E[X_i] = \frac{1}{6}, \quad V[X_i] = E[X_i^2] - E[X_i]^2 = \frac{1}{6} - \left(\frac{1}{6}\right)^2 = \frac{5}{36}$$

だから、 $W \equiv \sum_{i=1}^{60} (X_i - 1/6) / \sqrt{60 \cdot 5/36}$ の分布は $N(0, 1)$ で近似される。連続修正を施すと、問題の条件は $\sum_i X_i \geq 14.5$ 、つまり $W \geq (14.5 - 60/6) / \sqrt{60 \cdot 5/36} = 0.9\sqrt{3}$ と解釈される。したがって、その確率は $1 - F(0.9\sqrt{3}) \approx 0.060$ と見積られる (正確な確率は 0.06478... である)。

^{*16} たとえば $E[W_n^8]$ だと、 $E[X_{\square}^3 X_{\triangle}^3 X_{\square}^2]$ のような形も出てきます。でもこの形の項の個数は、多めにみても n^3 に比例する程度しかありません ($\square \triangle \square$ がそれぞれせいぜい n とおり)。だから、 $n \rightarrow \infty$ では分母の $\sqrt{n^8}$ に負けて消える運命です。結局残るのは $E[X_{\square}^2 X_{\triangle}^2 X_{\square}^2 X_{\star}^2] = E[X_{\square}^2] E[X_{\triangle}^2] E[X_{\square}^2] E[X_{\star}^2] = \sigma^8$ の項だけ。

^{*17} $< \text{か} \leq \text{か}$ はこだわらなくて構いません。整数 k に対して $S = k - 1/2$ だの $S = k + 1/2$ だのという半端な値が出る確率はどうせ 0。また、正規分布についても、特定の値ぴったりが出る確率なんてどうせ 0。

第5章の補足

5.a 相関係数についてのこまごま

$V[X]$ や $V[Y]$ が 0 だった場合、相関係数 ρ_{XY} は定義されません (ゼロ割りになってしまう)。しかしそもそも、 $V[X] = 0$ ということは X がゆらがないわけですから、こんな分析をするまでもありません。本書で相関係数を扱う際は、何も言わなくても分散 = 0 の場合は除くことにさせていただきます。なお、分散 < 0 はありえませんから、 $\sqrt{\text{分散}}$ で虚数が出るような心配は無用です。

相関係数が -1 から $+1$ までになる理由については、とり得る値が有限通りの場合を例題 5.10^(p.183) で調べました。一般の場合もその無限次元版だと思っていたらとてあえず結構です。関数を無限次元ベクトルとして扱う理論の詳細は、関数解析という分野を勉強してください。参考文献 [32] では付録 D でこの考え方に言及しています。

例題 5.1

U を区間 $[0, 2\pi]$ 上の一様分布とし、 $(X, Y) \equiv (\cos U, \sin U)$ とおく。 X と Y の相関係数が 0 であることを示せ。また、 X と Y が独立でないことを示せ (確率密度関数が存在しない場合の独立性の定義は、^{P.4.3}(^{p. 補足編 39}) を参照)。

答

$$E[X] = \int_0^{2\pi} \frac{\cos u}{2\pi} du = \frac{1}{2\pi} [\sin u]_0^{2\pi} = 0, \quad E[Y] = \int_0^{2\pi} \frac{\sin u}{2\pi} du = \frac{1}{2\pi} [-\cos u]_0^{2\pi} = 0$$

なので^{*1}、 $\text{Cov}[X, Y]$ とはいまの場合 $E[XY]$ のこと。すると

$$\text{Cov}[X, Y] = E[XY] = \int_0^{2\pi} \frac{(\cos u)(\sin u)}{2\pi} du = \frac{1}{2\pi} \left[\frac{1}{2} \sin^2 u \right]_0^{2\pi} = 0$$

だから、相関係数 $\rho_{XY} = 0$ 。また、 $P(X \geq 0.9 \text{ かつ } Y \geq 0.9) = 0$ (\because 常に $X^2 + Y^2 = \cos^2 U + \sin^2 U = 1$ のはず) と $P(X \geq 0.9)P(Y \geq 0.9) > 0$ とが一致しないから、 X と Y は独立でない^{*2}。 ■

5.b 期待値ベクトルと共分散行列の演算

5.b.1 右から行列をかけると期待値ベクトルは……

R をゆらぐ行列、 B をゆらがないただの行列とすると、 $E[RB] = E[R]B$ が成り立つことを 5.2.3 項 (p.187) 「ベクトル・行列の演算と期待値」で述べました。これを示すには、地道に成分を計算してもよいのですが、行列に慣れた人なら転置して前の話に帰着させるほうがお手軽です：まず $E[R]^T$ と $E[R^T]$ が等しいことは説明済。同じ理屈で、 $E[RB]^T$ も $E[(RB)^T]$ と等しいはず。ここで $(RB)^T = B^T R^T$ なのがポイント (「えっ」という人は線形代数を復習してください)。 $B^T R^T$ は、「ゆらがないただの行列 B^T 」かける「ゆらぐ行列 R^T 」という格好です。その期待値ならもう知っています。

$$E[(RB)^T] = E[B^T R^T] = B^T E[R^T]$$

これが「 $E[RB]$ の転置」に等しいというのですから、 $E[RB]$ 自身を知りたいればこれをもう一度転置すればよい。

$$E[RB] = (B^T E[R^T])^T = E[R^T]^T (B^T)^T = E[R]B$$

最後の等号には、転置の転置が元に戻ることに $(\circ^T)^T = \circ$ を使いました。

なお、 R, S がゆらぐ行列、 A がゆらがないただの行列のとき、 $E[RAS] = E[R]A E[S]$ と変形するのは一般にはダメです。一つの E の中に入っていたゆらぐ量 R, S を、勝手にそれぞれの E に分離してはいけません。もしあなたの教科書がこんな式変形をしているなら、どこかに何か前提条件が書かれていたはず (R と S は独立だ、など)。

5.b.2 ベクトルを足すと共分散行列は……

5.2.5 項 (p.191) 「変数変換すると共分散行列がどう変わるか」では、定数や行列をかけるという変換について調べました。ここでは補足として、ベクトルを足したときにどうなるのかをお話しておきます。

\mathbf{X} がゆらぐ縦ベクトルで \mathbf{a} がゆらがないただの縦ベクトルのとき、

$$V[\mathbf{X} + \mathbf{a}] = V[\mathbf{X}]$$

が成り立ちます。また、 \mathbf{Y} がゆらぐ縦ベクトルのときは、 \mathbf{X} と \mathbf{Y} が独立なら

$$V[\mathbf{X} + \mathbf{Y}] = V[\mathbf{X}] + V[\mathbf{Y}]$$

^{*1} わざわざ定積分を計算しなくても、三角関数のグラフの対称性から 0 なのは見え見えですが。

^{*2} 独立の定義が「任意の $\circ \circ$ に対して $\times \times$ 」だったから、それを否定するには、「 $\circ \circ$ なのに $\times \times$ でない」という反例を一つ挙げれば十分。

が成り立ちます。独立でない場合には一般にこれは成り立ちません。いずれも、3.4.4 項 (p.94) や 3.4.5 項 (p.97) で述べた数のときの結果がベクトルにもすなおに拡張された格好です。

$V[\mathbf{X} + \mathbf{a}]$ のほうは定義からすぐわかるので、以下、 $V[\mathbf{X} + \mathbf{Y}]$ がなぜそうなるのかを説明します。 $V[\mathbf{X}]$ の実体は分散 $V[X_i]$ や共分散 $\text{Cov}[X_i, X_j]$ をただ並べたものだったことを思い出しましょう。すると対角成分については、 X_i と Y_i が独立なら $V[X_i + Y_i] = V[X_i] + V[Y_i]$ 、というおなじみの事実を述べているだけ。一方、非対角成分については、 $E[\mathbf{X}] \equiv (\mu_1, \dots, \mu_n)^T$, $E[\mathbf{Y}] \equiv (\nu_1, \dots, \nu_n)^T$ とおくと

$$\begin{aligned} & \text{Cov}[X_i + Y_i, X_j + Y_j] \\ &= E\left[\{(X_i - \mu_i) + (Y_i - \nu_i)\}\{(X_j - \mu_j) + (Y_j - \nu_j)\}\right] \\ &= E[(X_i - \mu_i)(X_j - \mu_j)] + E[(Y_i - \nu_i)(Y_j - \nu_j)] \\ &\quad + E[(X_i - \mu_i)(Y_j - \nu_j)] + E[(Y_i - \nu_i)(X_j - \mu_j)] \\ &= \text{Cov}[X_i, X_j] + \text{Cov}[Y_i, Y_j] + \text{Cov}[X_i, Y_j] + \text{Cov}[Y_i, X_j] \end{aligned}$$

のように展開されます。ここでもし \mathbf{X} と \mathbf{Y} が独立なら、 X_i と Y_j は独立だし、 Y_i と X_j も独立です。すると後の 2 項は 0 になるので、確かに $\text{Cov}[X_i + Y_i, X_j + Y_j] = \text{Cov}[X_i, X_j] + \text{Cov}[Y_i, Y_j]$ 。

5.c 多次元正規分布についての補足

5.c.1 多次元正規分布の変換について

^{P.5.4(p.200)} では、 $Q^T = Q^{-1}$ から自動的に $\mathbf{q}_i \neq \mathbf{o}$ が保証されています。なぜなら、もし仮に Q のどこかの列が \mathbf{o} だったら、 Q の逆行列なんて存在しないはずだからです。なお、重複固有値があった場合は、同じ固有値に対する固有ベクトルたちが互いに直交するように \mathbf{p}_i を選んでください。具体的には、線形独立な固有ベクトルを重複度の本数だけ仮にまずとり、それらにグラム・シュミットの直交化 (ないしは QR 分解) を適用します。こうすれば互いに直交する固有ベクトルが得られます。このあたり、「えっ」という人は線形代数の教科書を参照ください。

例題 5.2

次の行列 V に対し、 $Q^T V Q$ が対角行列となるような直交行列 Q を一つ求めよ。

$$V = \begin{pmatrix} 4 & 1 & -2 \\ 1 & 4 & -2 \\ -2 & -2 & 7 \end{pmatrix}$$

答

V の特性方程式 $\det(\lambda I - V) = 0$ は

$$\lambda^3 - 15\lambda^2 + 63\lambda - 81 = (\lambda - 9)(\lambda - 3)^2 = 0$$

となるので、固有値は 9 と 3 (3 のほうは重複固有値)。固有値 3 の固有ベクトル、すなわち $V\mathbf{r} = 3\mathbf{r}$ となるベクトル $\mathbf{r} \neq \mathbf{o}$ は、 $\mathbf{r}_1 \equiv (1, 1, 1)^T$ や $\mathbf{r}_2 \equiv (2, 0, 1)^T$ など。この $\mathbf{r}_1, \mathbf{r}_2$ に対して

$$\mathbf{p}_1 \equiv \mathbf{r}_1 = (1, 1, 1)^T, \quad \mathbf{p}_2 \equiv \mathbf{r}_2 - \frac{\mathbf{r}_1 \cdot \mathbf{r}_2}{\|\mathbf{r}_1\|^2} \mathbf{r}_1 = (1, -1, 0)^T$$

のように直交化を施すと、固有値 3 の直交する固有ベクトル $\mathbf{p}_1, \mathbf{p}_2$ が得られる。あとは、残る固有値 9 の固有ベクトル $\mathbf{p}_3 \equiv (1, 1, -2)^T$ とあわせて $\mathbf{q}_i \equiv \mathbf{p}_i / \|\mathbf{p}_i\|$ を並べればよい ($i = 1, 2, 3$)。

$$\mathbf{q}_1 = \begin{pmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{pmatrix}, \quad \mathbf{q}_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \\ 0 \end{pmatrix}, \quad \mathbf{q}_3 = \begin{pmatrix} 1/\sqrt{6} \\ 1/\sqrt{6} \\ -2/\sqrt{6} \end{pmatrix} \rightarrow Q \equiv \begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & -1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & 0 & -2/\sqrt{6} \end{pmatrix}$$

そうすれば $Q^T V Q = \text{diag}(3, 3, 9)$ と対角化できる。 ■

5.3.2 項 (p.196) 「一般の多次元正規分布」では、伸縮と回転という 2 ステップで単位円を楕円に変換しました。別の方法として、作りたい斜め楕円の主軸方向を直接伸縮しても同じ分布を得ることができます。その事情は以下のとおり：斜め伸縮という操作は、次の 3 つの操作を続けて行うことと同値です。

- ア. 主軸が座標軸の向きになるよう回転する
- イ. 座標軸にそって伸縮する
- ウ. 回転を元に戻す

しかし分布の観点からは、アの操作はしてもしなくても変わりません。多次元標準正規分布をいくら回転しても、やっぱり多次元標準正規分布のままだからです^{*3}。ということは、アを省いてイウだけを施しても同じ分布が得られるはず。それがまさに、本文で説明した 2 ステップの構成法でした。

特異値分解というものを知っている人は、多次元正規分布を線形変換したらまた多次元正規分布になることを、以下のようにたやすく確かめられます。 \mathbf{X} を多次元正規分布に従う確率変数、 A はゆらがない正則行列とし、 $\mathbf{Y} = A\mathbf{X}$ と変数変換してみま

^{*3} 確率変数としては違うものになりますが、確率分布は同じです。混乱した方は 1.5 節 (p.16) 「確率分布」を復習してください。

しょう。本文の定義を思い出すと、 \mathbf{X} 自体が、 $\mathbf{X} = \mathbf{B}\mathbf{Z}$ のように「多次元標準正規分布に従う \mathbf{Z} 」と「ゆらが無い正則行列 \mathbf{B} 」とから作られているはずです。すると結局、 $\mathbf{Y} = \mathbf{A}(\mathbf{B}\mathbf{Z}) = (\mathbf{A}\mathbf{B})\mathbf{Z}$ なわけで、 \mathbf{Y} も「 \mathbf{Z} に正則行列 $\mathbf{C} \equiv \mathbf{A}\mathbf{B}$ をかけたもの」です。ここで $\mathbf{C} = \mathbf{Q}\mathbf{D}\mathbf{P}$ (\mathbf{Q}, \mathbf{P} は直交行列、 \mathbf{D} は対角行列、 \mathbf{D} の対角成分はすべて > 0) と特異値分解をすれば、「 \mathbf{C} をかける」は「 \mathbf{P} をかけて、その結果に \mathbf{D} をかけ、さらに \mathbf{Q} をかける」という 3 ステップに分解されます。しかし、多次元標準正規分布に第一ステップ「 \mathbf{P} をかける」を施しても、分布は変わりません（直交行列 \mathbf{P} をかけることは空間の回転や裏返しに対応していました。多次元標準正規分布にそんな操作をしてもやっぱり多次元標準正規分布のまま）。したがって、最終的に得られる分布は、対角行列 \mathbf{D} （座標軸にそった伸縮）と直交行列 \mathbf{Q} （回転や裏返し）をかけた場合と同じ。だから \mathbf{Y} も多次元正規分布に従います。

5.c.2 多次元正規分布の確率密度関数への補足

多次元正規分布の確率密度関数の等高線は「目安の楕円」と相似だ、と本文で主張しました（5.3.2 項 (p.196) 「一般の多次元正規分布」）。このことは、式の上からも次のようにわかります。関数 f の等高線や等位面とは、 $f(\mathbf{x})$ の値が一定な \mathbf{x} の集合のことでした。 f が $N(\boldsymbol{\mu}, \mathbf{V})$ の確率密度関数の場合、それは「 $(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ が一定な \mathbf{x} の集合」と言いかえられます。式の格好がこうだったからです。

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det \mathbf{V}}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) \quad (\text{式 (5.3)(p.202) の再掲})$$

一方、5.3.3 項 (p.201) 「多次元正規分布の確率密度関数」の計算をふり返ると、単位円の方程式 $\|\mathbf{z}\|^2 = 1$ は $(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = 1$ という方程式に変換されることがわかります。これが「目安の楕円」を表す方程式です。以上を見比べて、両者が相似だと結論されます。

多次元正規分布の共分散行列の格好からは、いろいろなことが読みとれます。たとえば、

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{pmatrix} \sim N(\boldsymbol{\mu}, \mathbf{V}), \quad \mathbf{V} = \begin{pmatrix} * & * & & & & \\ * & * & & & & \\ & & * & * & & \\ & & * & * & & \\ & & & & * & * \\ & & & & * & * \end{pmatrix} \quad \text{空欄はすべて 0}$$

のように共分散行列がブロック対角行列だったなら、

$$\mathbf{S} \equiv \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \mathbf{T} \equiv \begin{pmatrix} X_3 \\ X_4 \end{pmatrix}, \quad \mathbf{U} \equiv \begin{pmatrix} X_5 \\ X_6 \end{pmatrix}$$

は独立です。なぜなら確率密度関数が $f_{\mathbf{S}, \mathbf{T}, \mathbf{U}}(\mathbf{s}, \mathbf{t}, \mathbf{u}) = f_{\mathbf{S}}(\mathbf{s}) f_{\mathbf{T}}(\mathbf{t}) f_{\mathbf{U}}(\mathbf{u})$ と分解されるからです。独立なので、 \mathbf{S} や \mathbf{T} の値を聞いても \mathbf{U} を当てるための手がかりにはなりません。

5.c.3 切口と影の計算の詳細

5.3.5 項 (p.204) 「切口と影」で省略した計算を以下で示します。

切口（条件つき分布）について

多次元正規分布の条件つき分布が多次元正規分布になることは比較的容易にわかりました。でも、どんな多次元正規分布になるかの計算は少しがんばりが必要です。

多次元正規分布に従う

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim N \left(\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\nu} \end{pmatrix}, \begin{pmatrix} \text{あ} & \text{い} \\ \text{い}^T & \text{え} \end{pmatrix} \right)$$

に対し、 $\mathbf{X} = \mathbf{c}$ が与えられたときの \mathbf{Y} の条件つき分布を求めてみましょう。太字はベクトル、かなは行列です。（ブロック行列については参考文献 [32] 1.2.9 項などを参照）

いま

$$\Delta \mathbf{c} \equiv \mathbf{c} - \boldsymbol{\mu}, \quad \Delta \mathbf{y} \equiv \mathbf{y} - \boldsymbol{\nu}, \quad \begin{pmatrix} \text{か} & \text{き} \\ \text{き}^T & \text{け} \end{pmatrix} \equiv \begin{pmatrix} \text{あ} & \text{い} \\ \text{い}^T & \text{え} \end{pmatrix}^{-1}$$

とおくとき、 $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{c})$ を書き下すと次のようになります^{*4}。

$$\begin{aligned} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{c}) &\propto f_{\mathbf{X}, \mathbf{Y}}(\mathbf{c}, \mathbf{y}) \\ &\propto \exp \left(-\frac{1}{2} (\Delta \mathbf{c}^T, \Delta \mathbf{y}^T) \begin{pmatrix} \text{か} & \text{き} \\ \text{き}^T & \text{け} \end{pmatrix} \begin{pmatrix} \Delta \mathbf{c} \\ \Delta \mathbf{y} \end{pmatrix} \right) \\ &= \exp \left(-\frac{1}{2} (\Delta \mathbf{y} + \text{け}^{-1} \text{き}^T \Delta \mathbf{c})^T \text{け} (\Delta \mathbf{y} + \text{け}^{-1} \text{き}^T \Delta \mathbf{c}) + (\Delta \mathbf{y} \text{ によらない定数}) \right) \end{aligned}$$

上式の格好から、 $\mathbf{X} = \mathbf{c}$ が与えられたときの \mathbf{Y} の条件つき分布は $N(\tilde{\boldsymbol{\nu}}, \tilde{\mathbf{W}})$ だとわかります。ここに、

$$\tilde{\boldsymbol{\nu}} \equiv \boldsymbol{\nu} - \text{け}^{-1} \text{き}^T (\mathbf{c} - \boldsymbol{\mu}) = \boldsymbol{\nu} + \text{い}^T \text{あ}^{-1} (\mathbf{c} - \boldsymbol{\mu}) \quad (5.1)$$

$$\tilde{\mathbf{W}} \equiv \text{け}^{-1} = \text{え} - \text{い}^T \text{あ}^{-1} \text{い} \quad (5.2)$$

最後の式変形には、すぐ後で述べるブロック逆行列の公式を使いました。

同様に、 $\mathbf{Y} = \mathbf{d}$ が与えられたときの \mathbf{X} の条件つき分布は $N(\tilde{\boldsymbol{\mu}}, \tilde{\mathbf{V}})$ となります。ここに

$$\tilde{\boldsymbol{\mu}} \equiv \boldsymbol{\mu} - \text{か}^{-1} \text{き} (\mathbf{d} - \boldsymbol{\nu}) = \boldsymbol{\mu} + \text{い} \text{え}^{-1} (\mathbf{d} - \boldsymbol{\nu})$$

$$\tilde{\mathbf{V}} \equiv \text{か}^{-1} = \text{あ} - \text{い} \text{え}^{-1} \text{い}^T$$

^{*4} \mathbf{y} (もしくは $\Delta \mathbf{y}$) によらない比例係数はここではいちいち明示しません。それはどうせ「積分が 1」という掟によって自動的に定まるものだからです。

(参考) ブロック正方行列の逆行列

先ほどの式 (5.1) と (5.2) で使ったのはこんな公式です：ブロック正方行列に対して一般に、

$$\begin{pmatrix} \text{あ} & \text{い} \\ \text{う} & \text{え} \end{pmatrix}^{-1} = \begin{pmatrix} \text{カ} & \text{キ} \\ \text{ク} & \text{ケ} \end{pmatrix}$$

ここに、

$$\begin{aligned} \text{カ} &= (\text{あ} - \text{い}\text{え}^{-1}\text{う})^{-1} = \text{あ}^{-1} + \text{あ}^{-1}\text{い}\text{ケ}\text{う}\text{あ}^{-1} \\ \text{キ} &= -\text{カ}\text{い}\text{え}^{-1} = -\text{あ}^{-1}\text{い}\text{ケ} \\ \text{ク} &= -\text{ケ}\text{う}\text{あ}^{-1} = -\text{え}^{-1}\text{う}\text{カ} \\ \text{ケ} &= (\text{え} - \text{う}\text{あ}^{-1}\text{い})^{-1} = \text{え}^{-1} + \text{え}^{-1}\text{う}\text{カ}\text{い}\text{え}^{-1} \end{aligned}$$

ひらがなやカタカナはそれぞれ行列を表し、特に「あ」「え」「カ」「ケ」は正方行列を表します。なお、式中でてくる逆行列がすべて存在することが前提です^{*5}

この公式が成り立つことは、実際に元の行列にかけて単位行列になることを計算すれば確かめられます。あるいは、「基本変形による逆行列の筆算」(参考文献 [32] 2 章など) のブロック版を使ってこの公式を導くこともできます：

$$\begin{aligned} \left(\begin{array}{cc|cc} \text{あ} & \text{い} & I & O \\ \text{う} & \text{え} & O & I \end{array} \right) &\rightarrow \left(\begin{array}{cc|cc} \boxed{I} & \text{あ}^{-1}\text{い} & \text{あ}^{-1} & O \\ \text{う} & \text{え} & O & I \end{array} \right) \rightarrow \left(\begin{array}{cc|cc} I & \text{あ}^{-1}\text{い} & \text{あ}^{-1} & O \\ \boxed{O} & \text{え} - \text{う}\text{あ}^{-1}\text{い} & -\text{う}\text{あ}^{-1} & I \end{array} \right) \\ &\rightarrow \left(\begin{array}{cc|cc} I & \text{あ}^{-1}\text{い} & \text{あ}^{-1} & O \\ O & \boxed{I} & -\text{ケ}\text{う}\text{あ}^{-1} & \text{ケ} \end{array} \right) \quad \text{ケ} \equiv (\text{え} - \text{う}\text{あ}^{-1}\text{い})^{-1} \\ &\rightarrow \left(\begin{array}{cc|cc} I & \boxed{O} & \text{あ}^{-1} + \text{あ}^{-1}\text{い}\text{ケ}\text{う}\text{あ}^{-1} & -\text{あ}^{-1}\text{い}\text{ケ} \\ O & I & -\text{ケ}\text{う}\text{あ}^{-1} & \text{ケ} \end{array} \right) \end{aligned}$$

最終結果の縦棒の右側にできあがったのが逆行列です^{*6}。もう一方の表示も、「あ」と「え」の役割を交代して同様の操作をすれば得られます^{*7}。

影 (周辺分布) について

多次元正規分布の周辺分布がやはり多次元正規分布になることを確認します。 $\mathbf{X} = (X_1, X_2, \dots, X_n)^T \sim N(\mathbf{o}, V)$ に対し、 $\tilde{\mathbf{X}} \equiv (X_2, \dots, X_n)^T$ の周辺分布の確率密度関数は

$$\begin{aligned} f_{\tilde{\mathbf{X}}}(x_2, \dots, x_n) &= \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, x_2, \dots, x_n) dx_1 \\ &= \int_{-\infty}^{\infty} \square \exp \left(-\frac{1}{2} (x_1, x_2, \dots, x_n) \begin{pmatrix} \square & \square & \cdots & \square \\ \square & \square & \cdots & \square \\ \vdots & \vdots & \ddots & \vdots \\ \square & \square & \cdots & \square \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \right) dx_1 \\ &= \int_{-\infty}^{\infty} \square \exp \left(-\square x_1^2 + \boxed{x_2, \dots, x_n \text{ の 1 次式}} x_1 + \boxed{x_2, \dots, x_n \text{ の 2 次式}} \right) dx_1 \\ &= \int_{-\infty}^{\infty} \square \exp \left(-\square \left(x_1 - \boxed{x_2, \dots, x_n \text{ の 1 次式}} \right)^2 + \boxed{x_2, \dots, x_n \text{ の 2 次式}} \right) dx_1 \\ &= \square \left(\exp \boxed{x_2, \dots, x_n \text{ の 2 次式}} \right) \int_{-\infty}^{\infty} \exp \left(-\square \left(x_1 - \boxed{x_2, \dots, x_n \text{ の 1 次式}} \right)^2 \right) dx_1 \end{aligned}$$

となります。しかし、最後の積分の値は x_2, \dots, x_n に依存しません^{*8}。したがって結局、

$$f_{\tilde{\mathbf{X}}}(x_2, \dots, x_n) = \square \exp(x_2, \dots, x_n \text{ の 2 次式})$$

これは $(n-1)$ 次元正規分布の格好です。いまのは期待値 \mathbf{o} の例でしたが、期待値が \mathbf{o} でない場合も同様です。

図 5.18(p.207) のように「目安の楕円体」の影が周辺分布の「目安の楕円」となることは、図 5.34 や図 5.35(p.219) からわかります。

5.c.4 正則でない行列をかけると

多次元正規分布になる場合

$\mathbf{X} \sim N(\boldsymbol{\mu}, V)$ に正則な正方行列をかけても多次元正規分布になることは、5.3.4 項 (p.203) 「多次元正規分布の性質」でお話しました。その拡張として、横長フルランク行列 A をかけた $\mathbf{Y} \equiv A\mathbf{X}$ も多次元正規分布になることが示せます^{*9}。どんな多次元

^{*5} 先ほどの (5.1) や (5.2) の場合は、共分散行列全体の正定値性 (→ 5.d.3 項 (p. 補足編 53)) からこの前提が保証されます。

^{*6} 各ステップでは、順に $\begin{pmatrix} \text{あ}^{-1} & O \\ O & I \end{pmatrix}$, $\begin{pmatrix} I & O \\ O & \text{え}^{-1} \end{pmatrix}$, $\begin{pmatrix} I & O \\ O & \text{ケ}^{-1} \end{pmatrix}$, $\begin{pmatrix} I & -\text{あ}^{-1}\text{い} \\ O & I \end{pmatrix}$ を左からかけました。

^{*7} より正確には、導出過程から次のことがわかります：「あ」が正則だという前提で、ブロック行列全体が正則となるための必要十分条件は「え - うあ⁻¹い」が正則なこと。その場合、この両者の逆行列を使ったほうの表現により、ブロック行列全体の逆行列はやはりカキケと表される（「え」が正則でなくても OK）。行列「え - うあ⁻¹い」は「あ」のシュアアの補元 (Schur complement) と呼ばれます。なお、「カ」や「ケ」の二通りの表示が等しいという結果は、いわゆる逆行列の補題 (Sherman-Morrison-Woodbury の公式) そのものの (参考文献 [32] 1.2.13 の「腕だめし (2)」など)。

^{*8} $u \equiv x_1 - \boxed{x_2, \dots, x_n \text{ の 1 次式}}$ において、 x_1 による積分を u による積分に変数変換すれば、 $\int_{-\infty}^{\infty} \exp(-\square u^2) du$ となります。この値は x_2, \dots, x_n に依存しません。

^{*9} $m \times n$ の横長行列 A がフルランクであるとは、 $\text{rank } A = m$ という意味です ($m < n$)。これは A の各行ベクトルが線形独立なことと同値です。また、 A の像 $\text{Im } A$ が m 次元空間全体になることとも同値です。rank (ランクまたは階数) な

正規分布かは、期待値ベクトルと共分散行列がどうなるかを考えればわかります。

$$E[\mathbf{Y}] = E[A\mathbf{X}] = A E[\mathbf{X}] = A\boldsymbol{\mu}, \quad V[\mathbf{Y}] = V[A\mathbf{Y}] = A V[\mathbf{Y}] A^T = A V A^T$$

なので、答は $\mathbf{Y} \sim N(A\boldsymbol{\mu}, A V A^T)$ 。

では、 $\mathbf{Y} \equiv A\mathbf{X}$ が本当に多次元正規分布になるのか確かめましょう。それを見るには、基本変形で A を

$$A = BDC, \quad B, C \text{ は正則}, \quad D = (I|O) = \left(\begin{array}{c|c} 1 & \\ \hline & \ddots \\ & 1 \end{array} \right) \quad (\text{空欄は } 0)$$

と表すのが近道です。 A が横長フルランクという前提から、 D のところは必ず上の形にできます。このとき $\mathbf{Y} = BDC\mathbf{X}$ ですから、「 \mathbf{X} に C をかけて、さらに D をかけて、そこへさらに B をかけた結果」が \mathbf{Y} です。

- C は正則だから、 \mathbf{X} に C をかけてもやはり多次元正規分布。
- D をかけるという操作は、要するに、前のほうの成分はそのまま残して後のほうの成分を捨てること^{*10}。だから、多次元正規分布に D をかけると、周辺分布を求めたことになってやはり多次元正規分布。
- そこへさらに正則行列 B をかけても、結局は多次元正規分布。

こうして \mathbf{Y} も多次元正規分布だと確かめられました。

縮退してしまう場合

残りの場合、つまり A が

- 正則でない正方行列
- フルランクでない横長行列
- 縦長行列

という場合だと、 $\mathbf{Y} \equiv A\mathbf{X}$ の分布はべちゃんこなものになってしまいます^{*11}。これは確率密度関数では表せませんし、多次元正規分布ともふつうは呼びません。4.e.5 項 (p. 補足編 40) 「すなおでない変数変換」の「べちゃんこ」や「次元を増やす」でお話したとおりです。

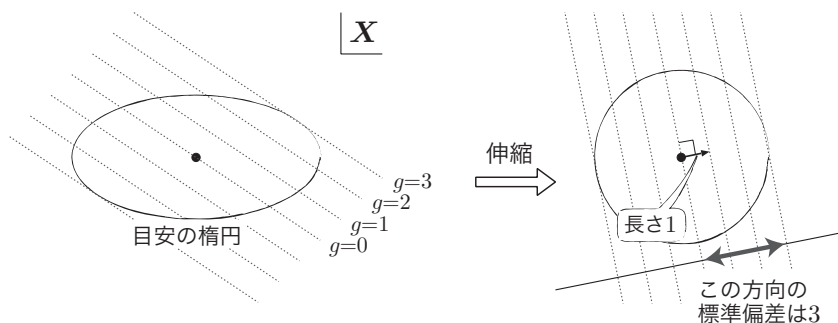
5.d 楕円に関連した話

5.d.1 楕円について

図 5.34(p.218) のように目安の楕円から任意方向のばらつき具合が読み取れる理由は、以下のとおりです。長さ 1 のゆらが無い定ベクトル \mathbf{u} に対して $g(\mathbf{x}) \equiv \mathbf{u}^T \mathbf{x}$ とおくと、「 \mathbf{u} 方向の標準偏差」とは要するに「 $g(\mathbf{X})$ の標準偏差」のことでした (5.2.6 項 (p.192) 「任意方向のばらつき具合」)。関数 $g(\mathbf{x})$ の等高線は、図 5.1 左のような縞模様になります。この図に対し、

- 楕円が円になるように縦横伸縮
- さらに、縞模様の間隔が 1 になるように全体を拡大縮小

という変換をしましょう (同図右)。変換後の図において、矢印方向の標準偏差が 3 であることは一目瞭然。すると $g(\mathbf{X})$ の標準偏差も 3 だとわかります。それを左側の原図に戻って解釈したのが、「影の長さからその方向のばらつきが読みとれる」という主張でした。いまの話のミソは、「等間隔の平行線」を線形変換してもやはり「等間隔の平行線」だという事実です^{*12}。もちろん接線なり影の長さなりを地道に計算しても同じ結果は得られるのですが、ごちゃつくので省略します。



▶ 図 5.1 伸縮で円に変換して考える

ベクトル値の確率変数 \mathbf{X} に対する目安の楕円は、多次元正規分布のときと同様に、次の方程式で表されます。

$$(\mathbf{x} - \boldsymbol{\mu})^T V^{-1} (\mathbf{x} - \boldsymbol{\mu}) = 1, \quad \boldsymbol{\mu} = E[\mathbf{X}], \quad V = V[\mathbf{X}]$$

つまり、この方程式を満たすベクトル \mathbf{x} の集合が、目安の楕円です。これは、5.4.3 項 (p.219) ケース 3 で、変換先 \mathbf{W} の空間における楕円の方程式

$$(\mathbf{w} - \boldsymbol{\nu})^T \Lambda^{-1} (\mathbf{w} - \boldsymbol{\nu}) = 1, \quad \boldsymbol{\nu} = E[\mathbf{W}], \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$$

へ変換式 $\mathbf{w} = Q^T \mathbf{x}$ と $\boldsymbol{\nu} = Q^T \boldsymbol{\mu}$ を代入することにより得られます。

^{*10} D のサイズを $m \times n$ として (横長だから $m < n$)、 $\mathbf{z} \equiv (z_1, \dots, z_n)^T$ に D をかけてみると、 $D\mathbf{z} = (z_1, \dots, z_m)^T$ 。

^{*11} 参考文献 [32] の表現でいうと、要するに「 $\text{Im } A$ が行き先の空間全体をカバーしきれない場合」です。

^{*12} 右側の図中の矢印 (等高線に垂直なベクトル) は、縦ベクトル \mathbf{u} の移り先とは別物です。一般の線形変換では、平行は平行のままですが、直角は直角でなくなってしまうからです。詳しくは参考文献 [4][5]などを参照。参考文献 [32] のサポートページにも、関連する解説を掲載しています (2009 年 10 月現在)。→ 「転置行列の図形的イメージ」
<http://wiki.fdiary.net/lacs/?Transpose>

5.d.2 自乗誤差の期待値についての公式

ベクトルの長さにつながる次の公式が、統計的手法にしばしば現れます。誤差を長さの 2 乗で測ることにして、その期待値を考える場面がよくあるからです（付録 A.6 脚注*13(p.334)）。

\mathbf{X} を n 次元ベクトル値の確率変数とし、 $\boldsymbol{\mu} \equiv \mathbb{E}[\mathbf{X}]$ をその期待値ベクトルとすると、

$$\mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|^2] = \text{Tr } \mathbb{V}[\mathbf{X}]$$

$$\mathbb{E}[\|\mathbf{X}\|^2] = \|\boldsymbol{\mu}\|^2 + \text{Tr } \mathbb{V}[\mathbf{X}]$$

より一般には、ゆらがない任意の定ベクトル \mathbf{a} に対し、

$$\mathbb{E}[\|\mathbf{X} - \mathbf{a}\|^2] = \|\boldsymbol{\mu} - \mathbf{a}\|^2 + \text{Tr } \mathbb{V}[\mathbf{X}]$$

(例題 3.12(p.100) も参照)。

右辺の Tr はトレース (trace) と読んでください。これは正方行列の対角成分の合計を表します。たとえば

$$\text{Tr} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} = 1 + 5 + 9 = 15$$

です。実は $\text{Tr } A$ は、 A の固有値すべての合計に一致します。また、行列 A, B のかけ算 AB が正方行列なら、 $\text{Tr}(AB) = \text{Tr}(BA)$ が成り立ちます。特に、同じ次元の縦ベクトル \mathbf{a}, \mathbf{b} に対して $\text{Tr}(\mathbf{b}\mathbf{a}^T) = \text{Tr}(\mathbf{a}^T\mathbf{b}) = \mathbf{a}^T\mathbf{b}$ が言えます (\mathbf{a}^T や \mathbf{b} を 1 行や 1 列の行列とみなした)。 $\mathbf{b}\mathbf{a}^T$ が正方行列なことや、 $\mathbf{a}^T\mathbf{b}$ は 1×1 行列なので数と同一視できることについては、線形代数の教科書を参照ください。

例題 5.3

上の公式を示せ。

答

ゆらぐ正方行列 R に対して一般に $\text{Tr } \mathbb{E}[R] = \mathbb{E}[\text{Tr } R]$ となることは Tr の定義から明らか。そこで上述の性質を活用して

$$\begin{aligned} \text{Tr } \mathbb{V}[\mathbf{X}] &= \text{Tr } \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] = \mathbb{E} \left[\text{Tr} \left\{ (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T \right\} \right] \\ &= \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})^T(\mathbf{X} - \boldsymbol{\mu})] = \mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|^2] \end{aligned}$$

が得られる。また、例題 5.11(p.190) の公式をあてはめると

$$\begin{aligned} \text{Tr } \mathbb{V}[\mathbf{X}] &= \text{Tr}(\mathbb{E}[\mathbf{X}\mathbf{X}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T) = \mathbb{E}[\text{Tr}(\mathbf{X}\mathbf{X}^T)] - \text{Tr}(\boldsymbol{\mu}\boldsymbol{\mu}^T) \\ &= \mathbb{E}[\mathbf{X}^T\mathbf{X}] - \boldsymbol{\mu}^T\boldsymbol{\mu} = \mathbb{E}[\|\mathbf{X}\|^2] - \|\boldsymbol{\mu}\|^2 \end{aligned}$$

から $\mathbb{E}[\|\mathbf{X}\|^2] = \|\boldsymbol{\mu}\|^2 + \text{Tr } \mathbb{V}[\mathbf{X}]$ が示される。より一般には、 $\Delta\mathbf{X} \equiv \mathbf{X} - \boldsymbol{\mu}$, $\Delta\mathbf{a} \equiv \mathbf{a} - \boldsymbol{\mu}$ とおき、 $\mathbb{E}[\Delta\mathbf{X}] = \mathbf{0}$ に注意して

$$\begin{aligned} \mathbb{E}[\|\mathbf{X} - \mathbf{a}\|^2] &= \mathbb{E}[\|\Delta\mathbf{X} - \Delta\mathbf{a}\|^2] = \mathbb{E}[(\Delta\mathbf{X} - \Delta\mathbf{a})^T(\Delta\mathbf{X} - \Delta\mathbf{a})] \\ &= \mathbb{E}[\Delta\mathbf{X}^T\Delta\mathbf{X}] - \mathbb{E}[\Delta\mathbf{X}^T\Delta\mathbf{a}] - \mathbb{E}[\Delta\mathbf{a}^T\Delta\mathbf{X}] + \mathbb{E}[\Delta\mathbf{a}^T\Delta\mathbf{a}] \\ &= \mathbb{E}[\Delta\mathbf{X}^T\Delta\mathbf{X}] - \mathbb{E}[\Delta\mathbf{X}^T]\Delta\mathbf{a} - \Delta\mathbf{a}^T\mathbb{E}[\Delta\mathbf{X}] + \Delta\mathbf{a}^T\Delta\mathbf{a} \\ &= \mathbb{E}[\|\Delta\mathbf{X}\|^2] - \mathbf{0}^T\Delta\mathbf{a} - \Delta\mathbf{a}^T\mathbf{0} + \|\Delta\mathbf{a}\|^2 \\ &= \text{Tr } \mathbb{V}[\mathbf{X}] + \|\mathbf{a} - \boldsymbol{\mu}\|^2 = \|\boldsymbol{\mu} - \mathbf{a}\|^2 + \text{Tr } \mathbb{V}[\mathbf{X}] \end{aligned}$$

5.d.3 共分散行列の非負定値性

共分散行列は対称行列ですが、ただの対称行列ではなくて、非負定値という特別な性質を持つ対称行列になっています。

非負定値対称行列

対称行列 H が非負定値であるとは、 $\mathbf{r}^T H \mathbf{r} \geq 0$ が任意の (実) ベクトル \mathbf{r} に対して成り立つという意味です*13。「非負定値対称行列」とつなげて呼ぶことが多いので、いま 5 回ほど唱えてこの言葉を覚えておくとよいでしょう。成分がすべて ≥ 0 だという意味ではないことを注意してください*14。たとえば、次の行列はどれも非負定値対称行列です。

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}, \quad \begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

次の行列はそうではありません。

$$\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 3 \\ 3 & 2 \end{pmatrix}, \quad \begin{pmatrix} 1 & -3 \\ -3 & 2 \end{pmatrix}, \quad \begin{pmatrix} -2 & -1 \\ -1 & -3 \end{pmatrix}$$

対称行列 H の非負定値性は、 H の固有値 $\lambda_1, \dots, \lambda_n$ がすべて ≥ 0 であることと同値です。理由は、 $\textcircled{P}5.4(\text{p.200})$ で述べた「直交行列 Q による対角化」 $Q^T H Q = \text{diag}(\lambda_1, \dots, \lambda_n)$ を行えばすぐわかります*15。

*13 人によって半正値や半正定値や準正定とも呼びます。

*14 「成分がすべて ≥ 0 の行列」は、単純に非負行列と呼んだりします。

*15 $Q^T \mathbf{r} \equiv \mathbf{s} = (s_1, \dots, s_n)^T$ とおけば $\mathbf{r} = Q\mathbf{s}$ であり、

$$\mathbf{r}^T H \mathbf{r} = (Q\mathbf{s})^T H (Q\mathbf{s}) = \mathbf{s}^T (Q^T H Q) \mathbf{s} = \lambda_1 s_1^2 + \dots + \lambda_n s_n^2$$

共分散行列は非負定値

共分散行列は必ず非負定値です。実際、ベクトル値確率変数 \mathbf{X} とゆらがないベクトル \mathbf{r} に対する $\mathbf{r}^T V[\mathbf{X}] \mathbf{r}$ は、

$$\mathbf{r}^T V[\mathbf{X}] \mathbf{r} = V[\mathbf{r}^T \mathbf{X}]$$

のように「実数値確率変数 $\mathbf{r}^T \mathbf{X}$ の分散」と等しいので、必ず ≥ 0 になります (→ 5.2.5 項 (p.191) 「変数変換すると共分散行列がどう変わるか」)。

また逆に、任意の非負定値対称行列 V が与えられたら、ベクトル値確率変数 \mathbf{X} をうまく作って $V[\mathbf{X}] = V$ とすることが出来ます。具体的には、 $Q^T V Q = \text{diag}(\lambda_1, \dots, \lambda_n)$ のようにうまい直交行列 Q で V を対角化し、 $V[\mathbf{Z}] = I$ という確率変数 \mathbf{Z} から $\mathbf{X} \equiv Q \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n}) \mathbf{Z}$ を作ればよい。 \mathbf{Z} 自身は、分散が 1 の独立な実数値確率変数を並べるだけで簡単に作れます。

例題 5.4

いま作った \mathbf{X} が $V[\mathbf{X}] = V$ となっていることを確認せよ。

答

$D \equiv \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n})$ とおく。 $D^T = D$ と $Q^T V Q = D^2$ に注意して、

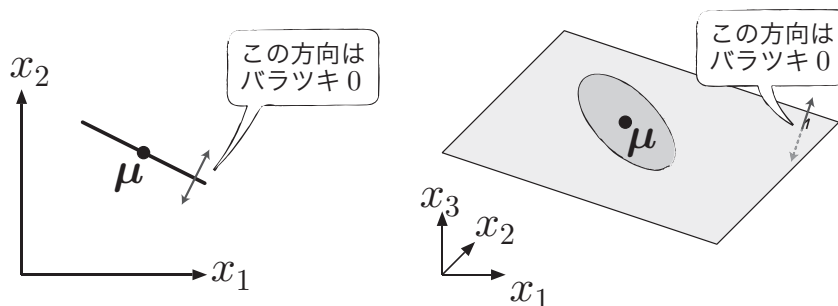
$$\begin{aligned} V[\mathbf{X}] &= V[Q D \mathbf{Z}] = (Q D) V[\mathbf{Z}] (Q D)^T = Q D V[\mathbf{Z}] D^T Q^T = Q D I D Q^T = Q D^2 Q^T \\ &= Q (Q^T V Q) Q^T = (Q Q^T) V (Q Q^T) = I V I = V \end{aligned}$$

■

正定値対称行列

5.4.3 項 (p.219) ケース 3 について、 $V[\mathbf{X}]$ がもし正則行列でない場合 ($\lambda_i = 0$ の場合) はどうなるか。正則な場合から類推して、正則でない場合は、楕円のどちらかの方向がつぶれてぺちゃんこになったものを目安の図形だと思ってください。これは要するに、図 5.2 左のような線分ですね。つぶれる方向は、固有値 0 に対応する固有ベクトルの方向です。3 次元なら、楕円体のどちらかの方向がつぶれてぺちゃんこになったもの。つまり同図右のように、(空間内のある平面上に描かれた) 中身のつぶった楕円です。

こんなふうにつぶれないための必要十分条件は、共分散行列の固有値 $\lambda_1, \dots, \lambda_n$ がすべて > 0 なことです。一般に、対称行列 H の固有値がすべて > 0 であることは、次の条件と同値です：「 $\mathbf{r}^T H \mathbf{r} > 0$ が任意の (実) ベクトル $\mathbf{r} \neq \mathbf{o}$ に対して成り立つ」。この条件を満たす対称行列を正定値対称行列と呼びます^{*16}。ですから、つぶれないための必要十分条件は、共分散行列が正定値であることと言い直されます。



▶ 図 5.2 目安の図形 (共分散行列 $V[\mathbf{X}]$ が正則でない場合。 $\mu = E[\mathbf{X}]$)

5.d.4 前提についての釈明

(玄人の方へ) 本章では計量とそれに関する正規直交基底を暗黙に想定しています。だから対称行列の固有値・固有ベクトルという概念も正当化されます。

この右辺がどんな実数 s_1, \dots, s_n でも ≥ 0 かどうか、話が翻訳される。

^{*16} 対称行列 H が正定値なことは、「 H が非負定値かつ正則」とも同値です (対角化してみれば一目瞭然)。

第6章の補足

6.a 推定論についての補足

6.a.1 枠組への補足

統計学の立場は記述統計と推測統計に分けられると 6.1.1 項 (p.227) で述べました。本当はもう一つ、6.1.9 項 (p.240) の策ウではのめかしたように、**Bayes 統計**をこれらに並ぶ第三の立場と位置づける専門家も多くいます。

本文の罰金に 2 乗をつけたのは、3.4.2 項 (p.90) で分散を定義したときと同様、数式上扱いやすいという利点からです。一方、2 乗の欠点としては、大外れな値 (**外れ値**) に影響を受けすぎることが問題とされます。それが気になる場面では、次のように絶対値を使うのも一つの定石です。

正解 $a = (a_1, \dots, a_k)$ に対して $b = (b_1, \dots, b_k)$ と推定したら、罰金は $|b_1 - a_1| + \dots + |b_k - a_k|$ 外れ値の影響を受けにくいという意味で、2 乗評価よりも絶対値評価のほうが**ロバスト (robust; 頑健)**であると言われます。6.1.2 項 (p.228) 「記述統計」の平均値と中央値の対比も思い出してください。2 乗評価と絶対値評価の対比もこれと同様です。

6.1.8 項 (p.238) の策イ (最尤推定) では、一致性や漸近有効性の雑な説明を書きました。一致性のところで言っていた「収束」とは、正確には確率収束のことです (→ 付録 C.1 (p.347) 「確率変数・確率分布の収束」)。漸近有効性の厳密な定義は参考文献 [14] などを参照ください。

6.a.2 不偏分散はなぜ (サンプルサイズ - 1) で割るのか

不偏分散について、サンプルサイズ n そのものでなく $(n - 1)$ で割ってはじめて不偏になることを、例題 6.2 (p.237) で計算しました。しかしあれだけだと心情的に納得しづらいかもしれないので、何が起きているのかももう少しお話しします。

数式による説明

データの平均 \bar{X} と本当の期待値 μ との違いをよく注意してください。からくりが見えやすいように、 $n = 2$ という露骨な場合を考えてみましょうか。 $X_1, X_2 \sim N(\mu, \sigma^2)$ (i.i.d.) のとき、

$$T^2 \equiv \frac{(X_1 - \mu)^2 + (X_2 - \mu)^2}{2}$$

についてなら、確かに期待値は $E[T^2] = \sigma^2$ となります。でも

$$\tilde{S}^2 \equiv \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2}{2}, \quad \bar{X} = \frac{X_1 + X_2}{2}$$

はこれとはちよつと違ってしますね。

T^2 も \tilde{S}^2 も、「 X_1, X_2 がある基準点からどれくらい離れているか」という誤差を測っているのだとは解釈できます。違いは基準点の選び方です。 T^2 では、固定された点 μ を基準としています。一方 \tilde{S}^2 では、 X_1, X_2 を見てからそれに応じた都合のよい基準点 \bar{X} を決めています。つまり、 \tilde{S}^2 では、誤差が最も小さくなるような基準点を後出ししているわけです。そのせいで、 \tilde{S}^2 は T^2 よりも小さな値になってしまいます。

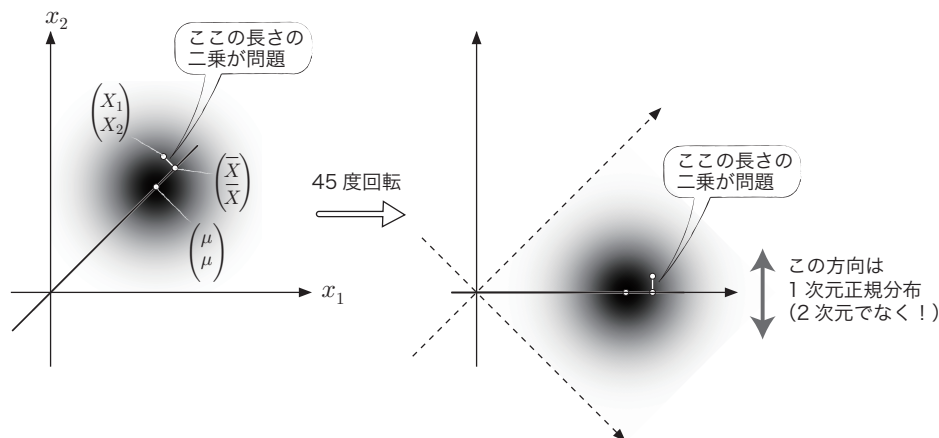
実際、与えられた数 x_1, x_2 に対して $(x_1 - c)^2 + (x_2 - c)^2$ という式の c をいろいろ動かしてみると、 $c = (x_1 + x_2)/2$ のとき最も小さな値になります。 \tilde{S}^2 はまさにこのように基準点 c を選んだ格好です。

というわけで、後出しにより誤差が小さく見える分を補正するために、 n ではなく $(n - 1)$ で割る必要があるのです。

図による説明

まだびんどこなければ、図 6.1 でいかがですか。 $\mathbf{X} \equiv (X_1, X_2)^T$ は 2 次元正規分布であり、その期待値は $\boldsymbol{\mu} \equiv (\mu, \mu)^T$ 、共分散行列は $\sigma^2 I$ 。そして $\bar{\mathbf{X}} \equiv (\bar{X}, \bar{X})^T$ はこんな位置にきます。 \tilde{S}^2 の分子 $(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2$ が「 $(\mathbf{X} - \bar{\mathbf{X}})$ の長さの 2 乗」であることに注意しましょう。

さて、左の図を 45 度回転すると右のようになります。すると、いま興味のあるところが、2 次元ではなく 1 次元正規分布の話になっていることがわかるはず。それが、 $n = 2$ でなく $n - 1 = 1$ で割らないといけない理由です。



▶ 図 6.1 なぜ n ではなく $(n-1)$ で割るのか ($n=2$ の例)。濃淡模様はデータ X_1, X_2 の同時分布の確率密度関数 $f_{X_1, X_2}(x_1, x_2)$ を表す

6.a.3 UMVUE の難点

資格を満たす参加者が誰もいない例

UMVUE (一様最小分散不偏推定量) について、資格を満たす参加者が誰もいない状況すらあると 6.1.7 項 (p.237) で述べました。その具体例の一つ挙げます。

作例的な例ですが、「確率 θ^3 で表、確率 $1 - \theta^3$ で裏が出るコイントス一回の結果 X から θ を当てろ ($0 \leq \theta \leq 1$)」などがそうです。実際、 X から作られた推定量 $\hat{\theta}(X)$ に対し、 $E[\hat{\theta}(X)] = \hat{\theta}(\text{表}) \cdot \theta^3 + \hat{\theta}(\text{裏}) \cdot (1 - \theta^3)$ 。この式の $\hat{\theta}(\text{表})$, $\hat{\theta}(\text{裏})$ をどんなに調節してもしょせんは $\square\theta^3 + \square$ の格好なので、 $E[\hat{\theta}(X)] = \theta$ とはできません。もっとおもしろい例は参考文献 [14] などを参照。

UMVUE への批判

UMVUE には、さらに以下のような批判もあります (話が込み入ってしまうので、ライトユーザは読み飛ばして結構です)。このあたりも詳しくは参考文献 [14] などを参照ください。

- パラメータを変換したら話がかわってしまう。……たとえば、6.1.7 項 (p.237) の S^2 は分散 σ^2 の UMVUE ですが、その平方根 S は標準偏差 σ の UMVUE ではありません。「分散が σ^2 だ」も「標準偏差が σ だ」も同じことを言っているはずなのに、表現のしかたによって UMVUE かな否かが変わってしまうわけです。
- 露骨に不合理な推定が UMVUE となってしまうこともある。……また作例的な例ですが、「松、竹、梅 という三通りの値をそれぞれ確率 $\theta/2$, $\theta^2/2$, $1 - (\theta + \theta^2)/2$ でとる確率変数 X を見て、 θ を当てろ ($0 \leq \theta \leq 1$)」などがそうです。実際、 $E[\hat{\theta}(X)] = \theta$ が成り立つような推定量 $\hat{\theta}(X)$ は、 $\hat{\theta}(\text{松}) = 2$, $\hat{\theta}(\text{竹}) = \hat{\theta}(\text{梅}) = 0$ しかありません (理由は先ほどと同様)。でもこれは $0 \leq \theta \leq 1$ の範囲を逸脱していますし、「竹が出たから竹の確率は 0 だろう」と推定するのも妙な話です。もっとおもしろい例は参考文献 [14] などを参照。
- 参加資格外に、「全種目で UMVUE に勝る選手」がいたりする。……実際、上の例で $\hat{\theta}(\text{松}) \equiv 1$, $\hat{\theta}(\text{竹}) = \hat{\theta}(\text{梅}) \equiv 0$ とでもとれば、 $0 \leq \theta \leq 1$ の全範囲で $E[(\hat{\theta}(X) - \theta)^2] \leq E[(\hat{\theta}(X) - \theta)^2]$ となります (等号成立は $\theta = 0$ のみ)。もっと興味深い例は、次の例題 6.1 や参考文献 [2] などを参照。さらに、びっくりするような例 (Stein のパラドックス) も知られています。詳しくは参考文献 [33]。

例題 6.1

$N(\mu, \sigma^2)$ に従う i.i.d. な X_1, \dots, X_n の平均を $\bar{X} \equiv \sum_{i=1}^n X_i/n$ 、不偏分散を $S^2 \equiv \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ とおく。与えられた定数 a に対し、 aS^2 と分散 σ^2 との期待罰金 $E[(aS^2 - \sigma^2)^2]$ を求めよ。

答

$E[S^2] = \sigma^2$ は例題 6.2 (p.237) ですでに述べました。さらに、 $V[S^2] = 2\sigma^4/(n-1)$ です*1。すると、例題 3.12 (p.100) で示した公式より、期待罰金は

$$E[(aS^2 - \sigma^2)^2] = (E[aS^2] - \sigma^2)^2 + V[aS^2] = (a\sigma^2 - \sigma^2)^2 + a^2 \frac{2\sigma^4}{n-1} = \left((a-1)^2 + \frac{2}{n-1} a^2 \right) \sigma^4$$

この結果から、 a を 1 よりちよつと小さくすれば、 $a=1$ のときよりも期待罰金を下げられることがわかります (どんな $\sigma^2 > 0$ でも!)。

*1 がんばつて展開しても求められますが、前の図 6.1 (p. 補足編 56) を思い出して次のようにするほうが見通しが効きます：
 $N(0, \sigma^2)$ に従う i.i.d. な Y_1, \dots, Y_{n-1} を使って、 $V[S^2] = V[\sum_{i=1}^{n-1} Y_i^2/(n-1)]$ と書ける。あとは $V[Y_i^2] = E[(Y_i^2)^2] - (E[Y_i^2])^2 = 3\sigma^4 - (\sigma^2)^2 = 2\sigma^4$ (→ 4.g 節脚注*15 (p. 補足編 46)) より……

6.b 検定論についての補足

6.b.1 枠組への補足

検定ではなぜわざわざ二つ仮説を用意するのか。それは、 H_1 とよく合うデータが出たからといって、直接「 H_1 だ」とは言い切りにくいからです。一般に、ある結果を説明できる理論は無数にあります。本文の例なら、「第一戦に勝った側が勢いづいて、残りの勝負も勝ちやすくなる」のような仮説 H_2 でも 61 勝 39 敗を説明できそうです。だから比較相手を明示しないとほつきりした議論になりません。

検定の論法は、自分の本音 (H_1) を棚に上げて他人 (H_0) の悪口ばかり言っているように見えたかもしれません。「もし H_0 だとしたら……」というセリフになったのは、6.2.2 項 (p.246) のように、我々があわてものの誤りについて品質保証を求めたからです。6.2.3 項 (p.247) 「単純仮説」の内容をふり返れば、悪口ばかりでもないことがわかるでしょう。

6.b.2 各事項について

ネイマン・ピアソンの補題への細かい補足

6.2.3 項 (p.247) 「単純仮説」で紹介したネイマン・ピアソンの補題において、 $g_1(x) > 0$ かつ $g_0(x) = 0$ となる場所 x がもしあったら、そこはすべて棄却とします。また、 $g_1(x) = g_0(x) = 0$ の場所は、どちらにしろ確率 0 なので何を答えてもどうせ効きません。

もし閾値 c をどう調節しても「あわてものの誤り」の確率をぴったり α にできない場合、理論的には次のような確率化検定 (ランダム検定) を考えます： $g_1(x)/g_0(x)$ が $> c$ なら棄却、 $< c$ なら受容、 $= c$ なら別途ルーレットを回して確率 r で棄却し確率 $(1-r)$ で受容する。 r は、上の「ぴったり α 」が達成されるように定めます。あまり実用的な話ではありませんが (参考文献 [2])。

一様最強力不偏検定

6.2.4 項 (p.249) 「複合仮説」で挙げた一様最強力不偏検定とは、次のようなものです：データと無関係にルーレットを回して確率 α で「棄却」と出力する、なんていうひどい検定法 δ_0 をまず考えます。そして、この δ_0 よりは一様に良い検定法だけに出場資格を制限し、その中で文句なしのチャンピオンを探します。もしみつければ、チャンピオンを一様最強力不偏検定と呼びます。いつもみつかるとは限りません。

尤度比検定

ネイマン・ピアソンの補題で尤度の比が活躍したことをふまえ、

$$L \equiv \frac{H_1 \text{ のもとでの尤度 } f_X(x) \text{ の最大値}}{H_0 \text{ のもとでの尤度 } f_X(x) \text{ の最大値}}$$

にもとづいて同様の検定を行うのが尤度比検定です (6.2.4 項 (p.249) 「複合仮説」でも名前だけ挙げました)。これは次のように解釈すると飲み込みやすいでしょう：「 H_0 も H_1 も複数の分布を含むかもしれないのが悩みの種だった。ならばまず、それぞれから代表となる分布を選出して、代表どうしの一騎打ちで勝負をつければいいではないか」。つまりこうです。

1. 仮説 H_0 を前提として、最尤推定により未知パラメータ θ を推定。推定結果を θ_0 とおく。
2. 同様に、仮説 H_1 を前提とした最尤推定で θ_1 を得る。
3. それらを用いて単純化した仮説 $\tilde{H}_0 : \theta = \theta_0$ および $\tilde{H}_1 : \theta = \theta_1$ を作る。
4. \tilde{H}_0, \tilde{H}_1 は単純仮説なので、あとはネイマン・ピアソンの補題にもとづいて検定を行う。

尤度比検定には、

- 機械的な計算により、とにかく何らかの検定方式が得られる
- しかも、良さげな検定方式が得られることが多い

という利点があります。さらに、

- サンプルサイズが大きくなれば、もし H_0 だとしたときの $2 \log L$ の分布はカイ自乗分布 (→ 5.3.6 項 (p.211)) に収束する

という便利な性質もそなえています^{*2}。そのおかげで、 L がいくらより大きければ棄却すべきかという閾値を近似的に計算することができます。コンピュータが普及していなかったころは、この性質が特に重宝されたようです。ただし、6.1.8 項 (p.238) の最尤推定の漸近有効性に相当するようないきなり根拠づけは、残念ながら尤度比検定にはありません。詳細は参考文献 [14]などを参照ください。

^{*2} もう少しだけ詳しく言うと、次のような話です： X_1, \dots, X_n が i.i.d. で、各 X_i の確率密度関数が未知パラメータ θ に応じて定まるとせよ (たとえば正規分布なら、期待値と分散とを並べた 2 次元ベクトルを θ ととるなど)。 θ の住む r 次元のパラメータ空間 Θ 内に、 q 次元の部分空間 Θ_0 を指定して、

- $H_0 : \theta \text{ が } \Theta_0 \text{ に属す}$
- $H_1 : \theta \text{ が } \Theta_0 \text{ に属さない (つまり、} \Theta \text{ のうち } \Theta_0 \text{ を除いた範囲に属す)}$

という格好の尤度比検定を考える。このとき、もし H_0 だとしたら、サンプルサイズ $n \rightarrow \infty$ の極限で $2 \log L$ の分布は自由度 $(r-q)$ のカイ自乗分布に弱収束 (→ 付録 C.1.4 脚注^{*3}(p.349)) する。正確には、パラメータが縮退していないことなどの前提条件を課す必要はありますが。

第7章の補足

7.a 擬似乱数列についての補足

7.a.1 乱数の検定

どうなっていたら擬似乱数列と呼ぶのか、決定版の定義はないと 7.1.2 項 (p.254) で述べました。ではせめて何か指針はないものか。実は、いろいろな検定を用いて、これは乱数列っぽいとは言えないだろうというアルゴリズムを排除するような話はありません。あまりすっきりした話ではないし、6.2 節 (p.244) 「検定論」で述べた枠組と合わない使い方で混乱を招きそうだから、本書では乱数の検定は説明しません。

7.a.2 モンテカルロ法の意義

7.1.3 項 (p.255) のモンテカルロ法のときに、確率変数列を擬似乱数列で代用して本当に大丈夫なのか、気になった読者もいることでしょう。でもまず実績として、こんな方法が広く使われて世の中の役に立っているのは事実です。また、この懸念に本気で取り組む理論的研究もなされています (参考文献 [28])。そこでは、「モンテカルロ法がうまく機能すること」をもって安全な擬似乱数という概念が定義されます。「ランダムな短い bit 列」を種として生成された長い擬似乱数列と、本当にランダムな長い bit 列とで、モンテカルロ法の性能に差が出るかどうか議論の鍵です。

7.b 所望の分布の生成についての補足

7.b.1 こまごま

一様分布に「相当する」擬似乱数列、という 7.2 項 (p.259) の微妙な表現には、次の注意が込められています。まず、あくまで「擬似」乱数列であること。そしてもう一つ、厳密に言うとは連続値ではないこと。コンピュータ内では数値の精度が有限ですから、実際に生成されるのは、 $[0, 1)$ 上の細かいとびとびの値です。

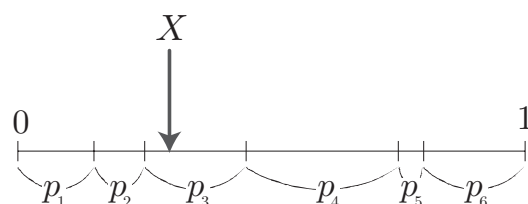
ちなみに ? 7.3 (p.259) の区間の記法は、昔のスタイルの本だと次のような流儀もあります。

$$]a, b[\rightarrow (a, b) \text{ と同じ, } [a, b[\rightarrow [a, b) \text{ と同じ, }]a, b] \rightarrow (a, b] \text{ と同じ}$$

図 7.1 のようにして離散値の一般の分布を作る方法を 7.2.1 項 (p.259) で述べましたが、実際にプログラムを書くときは、もっと工夫したほうが処理が速くなります。

- 値を並べかえる：順番に比較していくなら、確率の高い順に並べかえたほうが、比較回数の期待値を小さくできます。
 - もし $X < p_4$ なら $Y \equiv 4$
 - さもなくば、もし $X < p_4 + p_3$ なら $Y \equiv 3$
 - さもなくば、もし $X < p_4 + p_3 + p_6$ なら $Y \equiv 6$
 - ……
- 分岐させる：順番に比較するのではなく、まず大きく場合分けしてからだんだん細かく分けていくほうが比較回数の最悪値を小さくできます。期待値も小さくなるかもしれません。
 - もし $X < p_1 + p_2 + p_3$ なら…
 - もし $X < p_1 + p_2$ なら…
 - もし $X < p_1$ なら $Y \equiv 1$
 - さもなくば、 $Y \equiv 2$
 - さもなくば、 $Y \equiv 3$
 - さもなくば…

最適な比較のしかたにこだわりたい方は、ハフマン (Huffman) 法について調べてください (本によってハフマン木やハフマン符号という名前で載っているかもしれません)。また、テーブルルックアップを使ったさらに巧妙な方法 (Walker の別名法) は参考文献 [19] や [30] を参照。



► 図 7.1 確率 p_i で i が出るサイコロ (図 7.7 (p.260) の再掲)

なお、離散値でも規模がものすごく大きくなると、本文のやり方が現実的でなくなることもあります。たとえば、 10×10 ピクセルの白黒 (二値) 画像をランダムに、しかも指定された分布で生成したい場合を考えてください。こんな小さな画像でさえ、とり得る値が $2^{10 \times 10} \approx 10^{30}$ 通りもあるので、本文のような足し算なんてやってられません (「全ピクセルの同時分布」がどんなふうに指定されるかにもよりますが)。こういった場合に対しては、マルコフ連鎖モンテカルロ法 (Markov chain Monte Carlo method; MCMC) と呼ばれる手法が近年注目を集めています (→ 参考文献 [37])。

7.a

7.2.2 項 (p.261) 「連続値の場合」で累積分布関数を使った方法について、 F が平坦になったり不連続にジャンプしたりする場合には逆関数 F^{-1} が存在しませんよね？

はい。そんな場合は、 $F(y) \geq X$ となるような y の最小値を Y ととります (参考文献 [19])。

例題 7.2(p.264) の別解として、偏角と長さに着目した解き方も書いておきます。ベクトル $\mathbf{Y} \equiv (Y_1, Y_2)^T$ は、ベクトル $\mathbf{Z} \equiv (\sqrt{-2 \log X_1}, 0)^T$ を $2\pi X_2$ ラジアンだけ反時計回りに回転したのになっています。このことから、ベクトル \mathbf{Y} の方向 (偏角) は、360 度一様に分布することがわかります。また、ベクトル \mathbf{Y} の長さの 2 乗は $U \equiv \|\mathbf{Y}\|^2 = Y_1^2 + Y_2^2 = -2 \log X_1$ となっています。この U の確率密度関数は、例題 7.1(p.262) でやったとおり

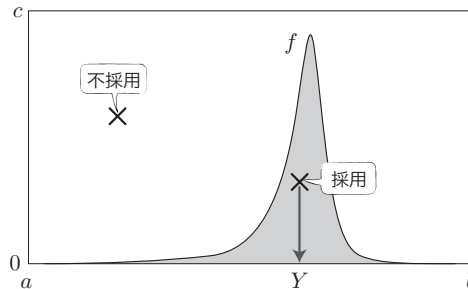
$$f_U(u) = \begin{cases} \frac{1}{2} \exp\left(-\frac{u}{2}\right) & (u \geq 0 \text{ のとき}) \\ 0 & (\text{他}) \end{cases}$$

です。さらに、作り方から偏角と長さは独立です。以上を例題 5.15(p.213) と見比べれば、 $\mathbf{Y} \sim N(\mathbf{o}, I)$ が確かめられます。

本文では図 7.14(p.267) 左の三角形上で一様分布する (Y_1, Y_2, Y_3) を構成してみせましたが、この分布は確率密度関数では表せません (→ 4.e.5 項 (p. 補足編 40))。あそこで言う「一様に」とは、次の意味です：3 次元空間内の領域 A が何か指定されたとき、 (Y_1, Y_2, Y_3) が A に属す確率は、図 7.14 左の三角形のうち A に入っている範囲の面積に比例する。

7.b.2 確率密度関数を使う方法 (工夫版)

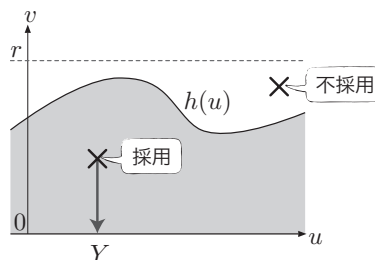
指定された確率密度関数を持つ乱数を作るには、原理的には 7.2.2 項 (p.262) の素朴版でもよいのですが、実用的にはもっと工夫する余地があります。素朴版だと、所望の確率密度関数 f の格好によっては、採用されるまでに何度もランダムな点を作り直さないといけません。たとえば、図 7.2 のように f が極端に大きいところがあると、アミがけ部分に点が入る確率はとても小さくなってしまいます。



► 図 7.2 f が極端に大きいところがあるようだと、全体の採用率が低くなる

そこで採用率を上げるために、こんな工夫をします。

1. 所望の分布と似た分布を持つ確率変数をどうにかして用意し、その確率密度関数を g とおく。
2. $h(u) = f(u)/g(u)$ のグラフを図 7.3 のように描く^{*1}。
3. このグラフ上にランダムに点を打つ。ただし横軸は 1. で用意した確率変数とし、縦軸はそれと独立に一様分布で 0 から r までの値をとる。 $(r$ はグラフより上ならどう設定しても OK。できるだけグラフぎりぎりに設定するほど効率が良くなる)
4. 点が h のグラフより下であればそれを採用し、横軸の値を Y として出力する。点がグラフより上にあった場合は、その点を捨てて、採用されるまで新しい点を同様に打ち直す。



► 図 7.3 $h(u) = f(u)/g(u)$ のグラフを描き、ランダムに点を打つ

この方法でも所望の分布が得られることは、以下のように確かめられます。点の座標を (U, V) とおくと、

^{*1} ゼロ割りを避けるために、 $f(u) > 0$ の範囲では必ず $g(u) > 0$ だとします (そういう g を使うようにしてください)。

という数列を出力するには、「1 を 40 回出力して停止せよ」のようなプログラムを書けば済みます。一方、

2, 7, 0, 2, 1, 3, 5, 4, 1, 5, 8, 9, 3, 7, 3, 1, 0, 7, 9, 4, 7, 2, 4, 6, 3, 9, 2, 7, 2, 1, 7, 8, 9, 6, 8, 8, 1, 9, 5

のようなでたらめっぽい数列だと、「2, 7, 0, ..., 9, 5 をこの順に出力して停止せよ」みたいにその数列自体を実質的に書き下してやるしかない（それ以上短くは書けない）かもしれません。そこで、後者のような数列のことを乱数列と呼ぶ、というのが一つの立場です。

ただしこの定義は理論的なものに過ぎず、実際に「使える」ものではありません。すぐ目につく使いづらさは、計算機の種類に依存する点です。どの万能計算機を考えるかによって、複雑度もいくらか変わってしまいます。さらに、もっと深刻な使いづらさとして、この複雑度は計算不可能なことが知られています。計算機科学を勉強した方は、停止性問題（与えられたプログラムがいつかは停止するかどうか）も思い出してください。それを有限時間で判定することは一般には不可能なのでした。チューリングマシンがどうのこうの……といった話です。

? 7.b 上の数列を短く書くアイデアを思いついた。「三十六進数 3cclyhh0epc8tg7e50we7dettv を十進法で一桁ずつ出力して停止せよ」。どう？

本文ではきちんと述べませんでしたが、数列もプログラムも bit 列に直して長さを比べるルールにすれば、その手の細工（文字の種類が多いことの活用）は効きません。文字の種類が多ければその分だけ一文字の bit 数も多いので、得にはならないからです。

7.c.2 無限列に対して

無限列に対しては、有限列の話を拡張するという方針も考えられますが、以下ではそれと少し違うアプローチを議論したいとします。……といっても、本気でやると深みにはまってしまうので、妥当な定義をめざしての模索の様子をざわりだけお見せします。詳しくは参考文献 [30] を参照ください。

簡単のため、数列のかわりに 0 と 1 からなる無限長の bit 列 x_1, x_2, x_3, \dots があつたとしましょう（各 x_i は 0 か 1）。小文字で書いていることから伺われるように、確率変数ではなく確定した「ただの bit 列」を考えています。こんな bit 列の中には、110110110110... や 01001000100001... のように規則性が見えるものもあるし、001011111010011100100110101100... のようにでたらめっぽく見えるものもあります。では、印象でなく客観的には、どんな bit 列をランダムであると呼ぶべきか。きちんと数学的に定式化できてしかも我々の感覚とも合うような、うまい条件がみつかるでしょうか。いろいろ検討してみます*2。

気持ちとしては、「確率半々で 0 か 1 が出る i.i.d. な確率変数の列 X_1, X_2, X_3, \dots 」っぽい無限列のことを乱数列と呼びたい。それが、我々が何となく持っている乱数列というもののイメージです。このイメージを数式でうまく表現することを本項の目標とします。イメージにあわせるためには、先ほどのような露骨にきっちりした規則だけでなく、「こういうときは 1 のほうがよく出がちだ」みたいな偏った傾向も持たないことを要求しないといけません。

そういう方向でまず思いつく条件は、0 と 1 の割合が半々であることでしょう。つまり

$$\lim_{n \rightarrow \infty} \frac{x_1, \dots, x_n \text{ 中の } 0 \text{ の個数}}{n} = \lim_{n \rightarrow \infty} \frac{x_1, \dots, x_n \text{ 中の } 1 \text{ の個数}}{n} = \frac{1}{2}$$

という条件です。しかしこれだと、

010101010101010101010101...

などという、ランダムにはほど遠い bit 列でも合格します。我々が何となく持っている乱数列というもののイメージを表現するには、この条件では弱すぎました。

では、「00, 01, 10, 11 の割合がそれぞれ 1/4 であること」、すなわち

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{(x_1, x_2), (x_2, x_3), \dots, (x_n, x_{n+1}) \text{ 中の } (0, 0) \text{ の個数}}{n} &= \frac{1}{4} \\ \lim_{n \rightarrow \infty} \frac{(x_1, x_2), (x_2, x_3), \dots, (x_n, x_{n+1}) \text{ 中の } (0, 1) \text{ の個数}}{n} &= \frac{1}{4} \\ \lim_{n \rightarrow \infty} \frac{(x_1, x_2), (x_2, x_3), \dots, (x_n, x_{n+1}) \text{ 中の } (1, 0) \text{ の個数}}{n} &= \frac{1}{4} \\ \lim_{n \rightarrow \infty} \frac{(x_1, x_2), (x_2, x_3), \dots, (x_n, x_{n+1}) \text{ 中の } (1, 1) \text{ の個数}}{n} &= \frac{1}{4} \end{aligned}$$

という条件ならどうか。これも、

001100110011001100110011...

が合格してしまうから弱すぎです。

さらに、「000, 001, 010, ..., 111 の割合がそれぞれ 1/8 であること」でも、

00010111 00010111 00010111 ...

という同じパターンのくり返し合格してしまうから弱すぎです。

じゃあもう、任意の正整数 k と任意の k bit パターンに対して、「連続する k bit がそのパターンになる割合は $1/2^k$ 」であることまで要求してみましょうか。実はこれでもまだ弱すぎます。たとえば、 $x_7 = x_{77} = x_{777} = x_{7777} = \dots = 1$ なんていう規則（7 のゾロ目の回はすべて 1）を隠し持った bit 列が合格できてしまうからです。こういうまばらな規則性は、上の要求では検

*2 乱数についてすでに知識を持っている読者は、7.a.1 項 (p. 補足編 58) で触れた乱数の検定と混同しないようご注意ください。いまは確定したただの bit 列について話をしています。確率変数列ではないし、検定でもありません。

出できません。

だからといって、任意の $i_1 < i_2 < i_3 < \dots$ に対し、部分列 $x_{i_1}, x_{i_2}, x_{i_3}, \dots$ 中の 0 と 1 の割合が半々なこと

$$\lim_{m \rightarrow \infty} \frac{x_{i_1}, x_{i_2}, \dots, x_{i_m} \text{ 中の 0 の個数}}{m} = \lim_{m \rightarrow \infty} \frac{x_{i_1}, x_{i_2}, \dots, x_{i_m} \text{ 中の 1 の個数}}{m} = \frac{1}{2} \quad (7.1)$$

まで要求すると、こんどは強すぎます。この要求には誰も合格できません。実際、どんな bit 列 x_1, x_2, \dots でも、それにあわせて後出しで

$$i_k \equiv (x_1, x_2, \dots \text{ 中で } k \text{ 回目に } 1 \text{ が出た位置})$$

という意地悪なとり方をされたら、 $x_{i_1} = x_{i_2} = \dots = 1$ 。よって不合格となってしまいます*3。

そろそろ「どないせえちゅうねん」と言いたくなってきたかもしれません。でもこんな議論がまだもつと続きます。任意の $i_1 < i_2 < i_3 < \dots$ だときつすぎた。じゃあどんな i_1, i_2, i_3, \dots について (7.1) を要求するべきか。それにはまた計算可能性の概念がからんできます (チューリングマシンがどうのこうの……)。本書ではこれ以上の説明はしませんが、一筋縄でいかないことだけはきつと感じていただけたでしょう。

*3 なお、1 が有限回しか出ないような bit 列 x_1, x_2, \dots は、どこから先がずっと 0 なので乱数列としては論外です。

第8章の補足

8.a 最小自乗法についての補足

8.a.1 こまごました補足

5章 (p.173) 「共分散行列と多次元正規分布と楕円」に対する 5.d.4 項 (p. 補足編 54) の釈明と同じく、8.1 節 (p.271) 「回帰分析と多変量解析から」でも暗黙に正規直交基底を想定しています。何を言っているのかびんとこない方は気にしないでください。

8.1.1 項 (p.271) 「最小自乗法による直線あてはめ」の冒頭の例は、温度設定をいろいろ変えてという話に合うよう $n \geq 2$ だと想定しています。また、 $x_1 = \dots = x_n$ という場合 (温度設定を一回も変えなかった場合) は除くことにします。

自乗誤差にもとづく方法に対しては、正規分布とはまた別の解釈もしばしば可能です。正規分布を仮定せず、そのかわり推定のしかたを「データの 1 次式」に限定して、その中でのベストを追究した結果だ、などです。

8.a.2 チコノフの正則化の導出

8.1.1 項 (p.271) で述べたような解釈からチコノフの正則化が導かれることを、実際に計算してみましょう。

まず、 \mathbf{A} と \mathbf{W} とが独立なので、それを並べた長い縦ベクトルも次のような $(m+n)$ 次元正規分布に従います (P.5.8(p.210) のベクトル版)。

$$\begin{pmatrix} \mathbf{A} \\ \mathbf{W} \end{pmatrix} \sim N\left(\mathbf{o}, \begin{pmatrix} \tau^2 I & \mathbf{O} \\ \mathbf{O} & \sigma^2 I \end{pmatrix}\right)$$

さらに、ブロック行列を使って

$$\begin{pmatrix} \mathbf{A} \\ \mathbf{Y} \end{pmatrix} = \begin{pmatrix} I & \mathbf{O} \\ \mathbf{C} & I \end{pmatrix} \begin{pmatrix} \mathbf{A} \\ \mathbf{W} \end{pmatrix}$$

と書けるのですから、5.3.4 項 (p.203) で述べた「多次元正規分布に正則行列をかけても多次元正規分布」が適用できます*1。すなわち、

$$\begin{pmatrix} \mathbf{A} \\ \mathbf{Y} \end{pmatrix} \sim N(\mathbf{o}, S) \\ S \equiv \begin{pmatrix} I & \mathbf{O} \\ \mathbf{C} & I \end{pmatrix} \mathbf{V} \left[\begin{pmatrix} \mathbf{A} \\ \mathbf{W} \end{pmatrix} \right] \begin{pmatrix} I & \mathbf{O} \\ \mathbf{C} & I \end{pmatrix}^T = \begin{pmatrix} I & \mathbf{O} \\ \mathbf{C} & I \end{pmatrix} \begin{pmatrix} \tau^2 I & \mathbf{O} \\ \mathbf{O} & \sigma^2 I \end{pmatrix} \begin{pmatrix} I & \mathbf{C}^T \\ \mathbf{O} & I \end{pmatrix}$$

この多次元正規分布 $N(\mathbf{o}, S)$ の確率密度関数は、すでに学んだとおり

$$f_{\mathbf{A}, \mathbf{Y}}(\mathbf{a}, \mathbf{y}) = \square \exp\left(-\frac{1}{2}(\mathbf{a}^T, \mathbf{y}^T) S^{-1} \begin{pmatrix} \mathbf{a} \\ \mathbf{y} \end{pmatrix}\right)$$

となります (興味がでない定数は□で省略します)。さて、いま調べたいのは、 $\mathbf{Y} = \tilde{\mathbf{y}}$ のときの \mathbf{A} の条件つき分布

$$f_{\mathbf{A}|\mathbf{Y}}(\mathbf{a}|\tilde{\mathbf{y}}) = \frac{f_{\mathbf{A}, \mathbf{Y}}(\mathbf{a}, \tilde{\mathbf{y}})}{f_{\mathbf{Y}}(\tilde{\mathbf{y}})}$$

です。それはグラフでいうと、 $f_{\mathbf{A}, \mathbf{Y}}(\mathbf{a}, \mathbf{y})$ のグラフを $\mathbf{y} = \tilde{\mathbf{y}}$ で切って、切口を (面積が 1 になるよう) 定数倍したものでした (→ 図 4.33(p.141))。しかし、みんながいつせいに定数倍されたところで、どこが一番高いかは変わりません。ですから、本文の方針アは、 $f_{\mathbf{A}, \mathbf{Y}}(\mathbf{a}, \tilde{\mathbf{y}})$ が最大となる \mathbf{a} を求めるのと同じです。さらに、多次元正規分布の切口がまた多次元正規分布になることも我々はすでに知っています (5.3.5 項 (p.204) 「切口と影」)。多次元正規分布なら、確率密度の最大点が期待値に一致。したがって方針イでもやはりこの \mathbf{a} が選ばれます。

では、 $f_{\mathbf{A}, \mathbf{Y}}(\mathbf{a}, \tilde{\mathbf{y}})$ が最大となる \mathbf{a} をがんばって求めましょう。式の格好からして、要は

$$h(\mathbf{a}) \equiv (\mathbf{a}^T, \tilde{\mathbf{y}}^T) S^{-1} \begin{pmatrix} \mathbf{a} \\ \tilde{\mathbf{y}} \end{pmatrix}$$

が最小となる \mathbf{a} を求めればよい。 S^{-1} を具体的に計算すると

$$\begin{aligned} S^{-1} &= \begin{pmatrix} I & \mathbf{C}^T \\ \mathbf{O} & I \end{pmatrix}^{-1} \begin{pmatrix} \tau^2 I & \mathbf{O} \\ \mathbf{O} & \sigma^2 I \end{pmatrix}^{-1} \begin{pmatrix} I & \mathbf{O} \\ \mathbf{C} & I \end{pmatrix}^{-1} \\ &= \begin{pmatrix} I & -\mathbf{C}^T \\ \mathbf{O} & I \end{pmatrix} \begin{pmatrix} \tau^{-2} I & \mathbf{O} \\ \mathbf{O} & \sigma^{-2} I \end{pmatrix} \begin{pmatrix} I & \mathbf{O} \\ -\mathbf{C} & I \end{pmatrix} \\ &= \begin{pmatrix} \tau^{-2} I & -\sigma^{-2} \mathbf{C}^T \\ \mathbf{O} & \sigma^{-2} I \end{pmatrix} \begin{pmatrix} I & \mathbf{O} \\ -\mathbf{C} & I \end{pmatrix} = \begin{pmatrix} \tau^{-2} I + \sigma^{-2} \mathbf{C}^T \mathbf{C} & -\sigma^{-2} \mathbf{C}^T \\ -\sigma^{-2} \mathbf{C} & \sigma^{-2} I \end{pmatrix} \end{aligned}$$

なので*2、

$$h(\mathbf{a}) = \mathbf{a}^T (\tau^{-2} I + \sigma^{-2} \mathbf{C}^T \mathbf{C}) \mathbf{a} - \sigma^{-2} \mathbf{a}^T \mathbf{C}^T \tilde{\mathbf{y}} - \sigma^{-2} \tilde{\mathbf{y}}^T \mathbf{C} \mathbf{a} + \sigma^{-2} \tilde{\mathbf{y}}^T \tilde{\mathbf{y}}$$

*1 この行列が正則なことは、次のようにすぐわかります：正方な下三角行列なので、その行列式は対角成分のかけ算。対角成分はすべて 1 だから、行列式は 1。ということは、行列式が 0 じゃないんだから正則。

*2 かけ算 $\triangle \diamond$ の逆行列が $\diamond^{-1} \triangle^{-1} \triangle \diamond^{-1}$ のように逆順になることや、 $\begin{pmatrix} I & \mathbf{O} \\ \mathbf{C} & I \end{pmatrix}$ の逆行列が $\begin{pmatrix} I & \mathbf{O} \\ -\mathbf{C} & I \end{pmatrix}$ であることなどは、実際に両者をかけてみればすぐ確かめられます (ちゃんと単位行列になりますね)。ちなみに、目の効く人なら、ここで S^{-1} を次のように整理しておくとか後がスマートです。でも慣れないと思いつきにくいでしょうから、本文ではひとまず力まかせに展開します。

これは次のように整理できます。

$$h(\mathbf{a}) = \tau^{-2} \mathbf{a}^T \mathbf{a} + \sigma^{-2} (\tilde{\mathbf{y}} - C\mathbf{a})^T (\tilde{\mathbf{y}} - C\mathbf{a}) = \frac{1}{\sigma^2} \left(\|\tilde{\mathbf{y}} - C\mathbf{a}\|^2 + \frac{\sigma^2}{\tau^2} \|\mathbf{a}\|^2 \right)$$

ここまでくれば答えは見えましたね。 \mathbf{a} を調整して $h(\mathbf{a})$ を最小化するには、 $\|\tilde{\mathbf{y}} - C\mathbf{a}\|^2 + \alpha \|\mathbf{a}\|^2$ を最小化すればよい ($\alpha = \sigma^2/\tau^2$)。まさにチコノフの正則化です。

8.b 主成分分析についての補足

8.b.1 全般的な補足

暗黙に正規直交基底を想定することはすでに述べたので繰り返しません。

8.1.2 項 (p.278) の主成分分析の説明では、記述統計の立場を取っています。つまり、 $\mathbf{x}_1, \dots, \mathbf{x}_n$ の背後に「見えない仕組み」を想定してそれを当てようというのではなく、 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 自体がどんな様子になっているか探ろう、というお話にしています (6.1.1 項 (p.227) 「記述統計と推測統計」)。

主成分分析やそれに類する話題では、文字式での行列計算がばりばり使われます。線形代数というタイトルの講義や教科書だとういう計算の練習機会が意外と少ないので、苦手としている方が多いかもしれません。それを意識して、参考文献 [32] の第 1 章では、この手の計算に必要なポイントを手厚く解説しました。計算が追いつらい方にはご一読をお勧めします。

主成分の意味づけや解釈を考えるという使い方には、デリケートな問題がからんでいます。場面場面でこの使い方が妥当なのかという議論は、数学だけでは結着が付きません。詳しくは、**多変量解析**や**因子分析**といったキーワードで、統計を専門に扱った教科書を勉強してください。

単位の違う成分が混ざっているとき、応用場面では、分散が 1 になるよう各軸を正規化 (3.4.4 項 (p.94)) した上で PCA をかけるという簡便な手も使われます。ただし 2 次元の場合は、主成分ベクトルが ± 45 度方向にしか出なくなるので PCA としての意味はあまりありません。

8.b.2 特異値分解の利用

特異値分解という技を知っていれば、8.1.2 項 (p.278) の手順よりもっと直接に主成分を求めることもできます。 n 本の縦ベクトル $\mathbf{x}_1, \dots, \mathbf{x}_n$ を並べて行列 $\Xi \equiv (\mathbf{x}_1, \dots, \mathbf{x}_n)$ を作り、 $\Xi = QDP$ (Q, P は直交行列、 D は対角行列) のように特異値分解したとしましょう (Ξ はギリシャ文字 ξ (グザイ) の大文字です)。ただし、 D の対角成分 (特異値) は大きい順に並べておきます。すると、 $E[\mathbf{X}] = \mathbf{o}$ という前提から

$$\begin{aligned} V[\mathbf{X}] &= E[\mathbf{X}\mathbf{X}^T] = \frac{1}{n} (\mathbf{x}_1\mathbf{x}_1^T + \dots + \mathbf{x}_n\mathbf{x}_n^T) = \frac{1}{n} \Xi\Xi^T \\ &= \frac{1}{n} (QDP)(QDP)^T = \frac{1}{n} QDPP^T D^T Q^T \end{aligned}$$

ここで P は直交行列だから $PP^T = I$ だし、 D は対角行列なので $D^T = D$ です。よって、 $V[\mathbf{X}] = (1/n)QD^2Q^T$ 。ですから、 D の第 i 対角成分が $\sqrt{n\lambda_i}$ 、 Q の第 i 列が主成分ベクトル \mathbf{q}_i になります。さらに、 $Q^T\Xi = Q^TQDP = DP$ という関係により、行列 DP の (i, j) 成分はデータ \mathbf{x}_j の第 i 主成分となっています。

8.b.3 自乗誤差の計算

式 (8.2) (p.281) の計算をいまからやってみせます。 $E[\mathbf{X}] = \mathbf{o}$ から $E[\mathbf{Y}] = RR^T E[\mathbf{X}] = \mathbf{o}$ なので、 $E[\mathbf{X} - \mathbf{Y}]$ も \mathbf{o} です。それを念頭に置き、 $\mathbf{X} - \mathbf{Y}$ に 5.d.2 項 (p. 補足編 53) の公式をあてはめると、

$$\begin{aligned} E[\|\mathbf{X} - \mathbf{Y}\|^2] &= \text{Tr } V[\mathbf{X} - \mathbf{Y}] = \text{Tr } V[(I - RR^T)\mathbf{X}] \\ &= \text{Tr} \left((I - RR^T) V[\mathbf{X}] (I - RR^T) \right) \\ &= \text{Tr} \left((I - RR^T)(I - RR^T) V[\mathbf{X}] \right) \dots \dots \text{Tr の性質より} \end{aligned}$$

ここで、 $R^T R = I$ に注意し⁴³、

$$\begin{aligned} (I - RR^T)(I - RR^T) &= I - 2RR^T + RR^T RR^T = I - 2RR^T + R(R^T R)R^T \\ &= I - 2RR^T + RR^T = I - RR^T \end{aligned}$$

であることを使って式変形を進めます。

$$\begin{aligned} E[\|\mathbf{X} - \mathbf{Y}\|^2] &= \text{Tr} \left((I - RR^T) V[\mathbf{X}] \right) \\ &= \text{Tr } V[\mathbf{X}] - \text{Tr} \left(RR^T V[\mathbf{X}] \right) \dots \dots \text{定義から } \text{Tr}(\bigcirc + \bigtriangleup) = \text{Tr } \bigcirc + \text{Tr } \bigtriangleup \text{ は明らか} \\ &= \text{Tr } V[\mathbf{X}] - \text{Tr} \left(R^T V[\mathbf{X}] R \right) \dots \dots \text{ふたたび Tr の性質より} \end{aligned}$$

さて、 $Q^T V[\mathbf{X}] Q$ は固有値が並んだ対角行列になるのです。つまり、 $\Lambda \equiv \text{diag}(\lambda_1, \dots, \lambda_n)$ とおけば $Q^T V[\mathbf{X}] Q = \Lambda$ 。よって $V[\mathbf{X}] = Q\Lambda Q^T$ です。これを上式へ代入して、

$$E[\|\mathbf{X} - \mathbf{Y}\|^2] = \text{Tr} \left(Q\Lambda Q^T \right) - \text{Tr} \left(R^T Q\Lambda Q^T R \right) \quad (8.1)$$

$$S^{-1} = \tau^{-2} \begin{pmatrix} I & O \\ O & O \end{pmatrix} + \sigma^{-2} \begin{pmatrix} -C^T \\ I \end{pmatrix} (-C, I)$$

⁴³ $\mathbf{q}_1, \dots, \mathbf{q}_k$ の長さが 1 で互いに直交することから、 $R^T R = I$ が言えます。やってみてください。

(8.1) の右辺第 1 項は、次のように計算されます ($Q^T Q = I$ という性質を利用)。

$$\mathrm{Tr}(Q\Lambda Q^T) = \mathrm{Tr}(Q^T Q\Lambda) = \mathrm{Tr}\Lambda = \lambda_1 + \cdots + \lambda_m \quad (m \text{ は } \mathbf{x} \text{ の次元数})$$

一方、(8.1) の右辺第 2 項については、

$$Q^T R = \begin{pmatrix} \overbrace{1}^k & & \\ & \ddots & \\ & & 1 \end{pmatrix} \quad (\text{空欄はすべて } 0)$$

であることから

$$\mathrm{Tr}(R^T Q\Lambda Q^T R) = \mathrm{Tr}((Q^T R)^T \Lambda (Q^T R)) = \mathrm{Tr} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_k \end{pmatrix} = \lambda_1 + \cdots + \lambda_k$$

となります。差引すると、(8.1) は結局 $E[\|\mathbf{X} - \mathbf{Y}\|^2] = \lambda_{k+1} + \cdots + \lambda_m$ になることがわかりました。

8.c ランダムウォークの練習用例題

8.2.1 項 (p.286) 「ランダムウォーク」に関連した練習用例題をいくつか置いておきます。

例題 8.1

ランダムウォーク X_t について、以下がそれぞれ独立か否かを答えよ。

1. X_1 と X_{100}
2. X_3 と $X_5 - X_4$
3. $X_3 - X_2$ と $X_5 - X_4$
4. X_5 と $X_8 - X_2$
5. $X_4 - X_2$ と $X_5 - X_3$
6. はじめて $X_t = 0$ へもどってくる時刻 $t = T_1 (> 0)$ と、二度目に $X_t = 0$ へもどってくる時刻 $t = T_2 (> T_1)$ 。(「永遠にもどってこない確率」は後述のとおり 0 なので無視してよい)
7. T_1 と $T_2 - T_1$

答

独立なのは 2 と 3 と 7。それ以外は、片方がどんな値になったかしだいで、他方がどんな値になりやすいかが変わってしまいます。ぴんとこなければ、8.2.1 項 (p.286) 冒頭の定義に戻って Z_i で書き直してみてください。たとえば 3 については、 $X_3 - X_2 = Z_3$ と $X_5 - X_4 = Z_5$ だから、前提より両者は独立。5 については、 $X_4 - X_2 = Z_3 + Z_4$ と $X_5 - X_3 = Z_4 + Z_5$ だから、 Z_4 が重複している。そのせいでたとえば、もし $X_4 - X_2 = 2$ なら ($Z_4 = 1$ が確定するから) $X_5 - X_3 = -2$ は決して起きない。

$$\text{不一致} \quad \begin{cases} P(X_5 - X_3 = -2 | X_4 - X_2 = 2) = 0 \\ P(X_5 - X_3 = -2) = 1/4 \end{cases}$$

この不一致は、 $X_4 - X_2$ と $X_5 - X_3$ とが独立でないことを意味します。 ■

例題 8.2

ランダムウォーク X_t において、 X_1, \dots, X_{2n} がすべて ≥ 0 となる確率 u_{2n} を求めよ (n は正整数)。

答

X_1, \dots, X_{2n} のどれかが < 0 となる確率は $1 - u_{2n}$ 。これを X_{2n} の値で場合分けして計算しよう。

$$\begin{aligned} 1 - u_{2n} &= P(X_1, \dots, X_{2n} \text{ のどれかが } < 0) \\ &= \sum_{i < 0} P(X_{2n} = i) + \sum_{i \geq 0} P(X_{2n} = i \text{ かつ途中どこかで } < 0) \\ &= \sum_{i < 0} P(X_{2n} = i) + \sum_{i \geq 0} P(X_{2n} = -(i+1)) \quad \cdots \cdots \text{鏡像原理 (例題 8.4 (p.287))} \\ &= \sum_{i < 0} P(X_{2n} = i) + \sum_{i \geq 0} P(X_{2n} = i+1) \quad \cdots \cdots \text{対称性より} \\ &= \sum_{i \neq 0} P(X_{2n} = i) = 1 - P(X_{2n} = 0) \end{aligned}$$

したがって結局、 $u_{2n} = P(X_{2n} = 0) = {}_{2n}C_n / 2^{2n}$

例題 8.3

ランダムウォーク X_t において、 X_1, \dots, X_{2n} がすべて $\neq 0$ となる確率を求めよ (n は正整数)。

答

この確率も先ほどの例題 8.2 の u_{2n} に一致します：「すべて $\neq 0$ 」とは要するに「すべて > 0 」か「すべて < 0 」かであり、対称性からどちらのケースも確率は同じのはず。そして、「すべて > 0 」という条件は、「 $X_1 = 1$ かつ $X_2 - X_1, \dots, X_{2n} - X_1$ がすべて ≥ 0 」と言い直せる。ここで、 $Y_s \equiv X_{s+1} - X_1$ とおこう ($s = 0, 1, 2, \dots$)。 Y_s もランダムウォークになっていることに注意せよ。また、「 Y_1, \dots, Y_{2n-1} がすべて ≥ 0 」という条件は「 Y_1, \dots, Y_{2n} がすべて ≥ 0 」と同じことである ($\because Y_{\text{奇数}}$ は必ず奇数なので、 $Y_{2n-1} \geq 0$ は $Y_{2n-1} \geq 1$ を意味する)。以上から、

$$\begin{aligned} P(X_1, \dots, X_{2n} \text{ がすべて } \neq 0) &= 2P(X_1, \dots, X_{2n} \text{ がすべて } > 0) \\ &= 2P(X_1 = 1)P(Y_1, \dots, Y_{2n} \text{ がすべて } \geq 0) \quad \cdots \cdots X_1 \text{ と } (Y_1, \dots, Y_{2n}) \text{ とは独立} \\ &= 2 \cdot \frac{1}{2} \cdot u_{2n} = u_{2n} \end{aligned}$$

例題 8.4

ランダムウォーク X_t において、「原点へ永遠に戻らない確率」が 0 であることを示せ。すなわち、 $P(X_1 \neq 0, X_2 \neq 0, \dots) = 0$ を示せ。

答

例題 8.3 をふまえて $u_{2n} \rightarrow 0$ ($n \rightarrow \infty$) を直接確認してもよいのですが、ここでは別解を述べます。

「いま位置 $x \neq 0$ にいたとすると、その後いつかは位置 0 へ戻る確率」を $r(x)$ とおきましょう。そんな確率は、 x から次に $x-1$ へ進んだ場合と $x+1$ へ進んだ場合とに場合分けして、

$$r(x) = \frac{1}{2}r(x-1) + \frac{1}{2}r(x+1), \quad x \neq 0 \quad (8.2)$$

と表されます。要するに、両どりの平均値が自分だという格好です。ただし、 $x = \pm 1$ のときにつじつまを合わせるため、 $r(0) \equiv 1$ と定義しておきます。さて、このときもし $r(1) < 1$ だったら、 $r(2)$ はその分さらに小さくなり、 $r(3)$ はさらにさらに小さくなり……^{*4}。すると先のほうでは確率 $r(i) < 0$ という不合理な結論が出てしまいます。それはありえないから、背理法より $r(1) = 1$ 。同じ理屈で $r(-1)$ も 1 です。よって、最初の一步がどちらだったとしても、確率 1 でいつかは $X_t = 0$ に戻ってきます。

例題 8.5

ランダムウォーク X_t において、原点にはじめて復帰する時刻 (つまり、はじめて $X_t = 0$ となる $t > 0$) を T とおく^{*5}。 T には期待値が存在しないことを示せ。(この結果は、とんでもなく大きな T がそれなりの確率で出てしまうことを意味しています → 3.e.2 項 (p. 補足編 28))

答

例題 8.5^(p.287) みたいな方法で $P(T = k)$ を具体的に求めて $\sum_k k P(T = k)$ を計算する手もありますが、ここでは別解を述べます。

背理法で示しましょう。「いま位置 $x \neq 0$ にいたとすると、その後はじめて位置 0 へ戻るまでの所要時間」の期待値が仮に存在したとして、それを $\mu(x)$ とおきます。そんな期待値は、 x から次に $x-1$ へ進んだ場合と $x+1$ へ進んだ場合とに場合分けして、

$$\mu(x) = \frac{1}{2} \left(1 + \mu(x-1) \right) + \frac{1}{2} \left(1 + \mu(x+1) \right) = 1 + \frac{\mu(x-1) + \mu(x+1)}{2}, \quad x > 0$$

と表されます。ただし、 $x = 1$ のときにつじつまを合わせるため、 $\mu(0) \equiv 0$ と定義しておきます。さて、上式が

^{*4} びんとこなければ、式 (8.2) を $r(x) - r(x-1) = r(x+1) - r(x)$ と変形して考えましょう。

^{*5} T は連しだいでゆらぐ確率変数です。びんとこなければ神様視点で考えてください。 $T(\omega) = \min\{t > 0 \mid X_t(\omega) = 0\}$ です。なお、永遠に戻ってこない確率 (「 $X_1(\omega), X_2(\omega), \dots$ が一度も 0 にならないような世界 ω 」たちの面積) は、前問のとおり 0 なので、無視して構いません。

$$(\mu(x+1) - \mu(x)) = (\mu(x) - \mu(x-1)) - 2, \quad x \neq 0$$

と変形できることに注目。 x を増やしていったときの $\mu(x)$ の増え方は、だんだんと鈍っていき、いずれは減少に転じるわけです。そこからさらに x を増やしていくと、先のほうでは $\mu(x)$ が負に落ち込んでしまいます。しかしそれは不合理。よって背理法から、少なくともどこかの位置 c では「そこから0へ戻るまでの期待所要時間」が発散しているはず。すると $E[T]$ も発散してしまいます（∵ある時刻 t （たとえば $t=c$ ）に運悪く $X_t=c$ となる確率がゼロではない。そして、 $X_t=c$ となってしまう後は……。3.6節 (p.107)「おまけ：条件つき期待値と最小自乗予測」も参照）。 ■

8.d 多次元版のカルマンフィルタ

8.2.2項 (p.289) の最後でほめかした多次元版のカルマンフィルタを本節で導出します。時刻 $t=1, 2, 3, \dots$ における状態 \mathbf{X}_t と観測値 \mathbf{Y}_t が次のように定まるとします。

$$\begin{aligned}\mathbf{X}_t &= A\mathbf{X}_{t-1} + \mathbf{W}_t \\ \mathbf{Y}_t &= C\mathbf{X}_t + \mathbf{Z}_t\end{aligned}$$

ただし、独立性などの仮定は8.2.2項 (p.289) と同様とし、

$$\mathbf{W}_t \sim N(\mathbf{o}, U), \quad \mathbf{Z}_t \sim N(\mathbf{o}, V), \quad \mathbf{X}_0 \sim N(\boldsymbol{\mu}_{0|0}, P_{0|0})$$

のように設定します。 $\mathbf{X}_t, \mathbf{Y}_t, \mathbf{W}_t, \mathbf{Z}_t$ はゆらぐベクトル、 $\boldsymbol{\mu}_{0|0}$ はゆらがない定ベクトル、 $A, C, U, V, P_{0|0}$ はゆらがない定行列です（ベクトルはすべて縦ベクトルとします）。 $\mathbf{Y}_1, \dots, \mathbf{Y}_t$ の値を見て \mathbf{X}_t を当てましょう。

$\mathbf{Y}_1, \dots, \mathbf{Y}_{t-1}$ が与えられたときの \mathbf{X}_{t-1} の条件つき分布を $N(\boldsymbol{\mu}_{t-1|t-1}, P_{t-1|t-1})$ 、 \mathbf{X}_t の条件つき分布を $N(\boldsymbol{\mu}_{t|t-1}, P_{t|t-1})$ とおきます（これらが多次元正規分布になることは上の前提からわかります）。両者の関係は、

$$\boldsymbol{\mu}_{t|t-1} = A\boldsymbol{\mu}_{t-1|t-1}, \quad P_{t|t-1} = AP_{t-1|t-1}A^T + U \quad (8.3)$$

です。

では、同じく $\mathbf{Y}_1, \dots, \mathbf{Y}_{t-1}$ が与えられたときの

$$\begin{pmatrix} \mathbf{X}_t \\ \mathbf{Y}_t \end{pmatrix} = \begin{pmatrix} I \\ C \end{pmatrix} \mathbf{X}_t + \begin{pmatrix} O \\ I \end{pmatrix} \mathbf{Z}_t$$

の条件つき分布はどうなるでしょうか。答は、次の期待値ベクトルと共分散行列を持つ多次元正規分布です。

$$\text{期待値ベクトル: } \begin{pmatrix} I \\ C \end{pmatrix} \boldsymbol{\mu}_{t|t-1}$$

$$\text{共分散行列: } \begin{pmatrix} I \\ C \end{pmatrix} P_{t|t-1} \begin{pmatrix} I & C^T \end{pmatrix} + \begin{pmatrix} O \\ I \end{pmatrix} V \begin{pmatrix} O & I \end{pmatrix} \equiv \begin{pmatrix} \text{あ} & \text{い} \\ \text{い}^T & \text{え} \end{pmatrix}$$

すると、ここへさらに $\mathbf{Y}_t = \mathbf{y}_t$ を見せられたときの \mathbf{X}_t の条件つき分布は、5.c.3項 (p. 補足編 50) より

$$\begin{aligned}\boldsymbol{\mu}_{t|t} &= \boldsymbol{\mu}_{t|t-1} + \text{い}\text{え}^{-1}(\mathbf{y}_t - C\boldsymbol{\mu}_{t|t-1}) \\ &= \boldsymbol{\mu}_{t|t-1} + P_{t|t}C^TV^{-1}(\mathbf{y}_t - C\boldsymbol{\mu}_{t|t-1})\end{aligned} \quad (8.4)$$

$$\begin{aligned}P_{t|t} &= \text{あ} - \text{い}\text{え}^{-1}\text{い}^T \\ &= P_{t|t-1} - P_{t|t-1}C^T(CP_{t|t-1}C^T + V)^{-1}CP_{t|t-1}\end{aligned} \quad (8.5)$$

となります*6。

以上の更新式 (8.3) (8.4) (8.5) をあわせると、

$$(\boldsymbol{\mu}_{0|0}, P_{0|0}) \longrightarrow (\boldsymbol{\mu}_{1|0}, P_{1|0}) \xrightarrow{\text{観測値}\mathbf{y}_1} (\boldsymbol{\mu}_{1|1}, P_{1|1}) \longrightarrow (\boldsymbol{\mu}_{2|1}, P_{2|1}) \xrightarrow{\text{観測値}\mathbf{y}_2} (\boldsymbol{\mu}_{2|2}, P_{2|2}) \longrightarrow \dots$$

のように順次計算していく逐次アルゴリズムが得られます。たとえば、観測値 $\mathbf{y}_1, \dots, \mathbf{y}_{10}$ を見て \mathbf{X}_{10} を当てろと言われたら、期待値ベクトル $\boldsymbol{\mu}_{10|10}$ を答えればよい。さらに、その推定にどれくらい自信があるかと問われたら、共分散行列 $P_{10|10}$ を答えればよいわけです。共分散行列から各方向のばらつき具合を読み取る方法は5章 (p.173) でお話しました。

8.e マルコフ連鎖についての補足

8.2.3項 (p.294) へのいろいろな補足を本節で述べます。

8.e.1 表記の不統一

推移確率の表記で i, j のどちらがどちらか混乱しないよう、本書では $p_{i \leftarrow j}$ という我流の記法を使いました。ふつうの本には p_{ij} や p_{ji} と書いてあるはずですが、ただ困ったことに、 $p_{i \leftarrow j}$ のことを p_{ij} と書くか p_{ji} と書くかは本によって違います。そのつ

*6 ここまでの計算では、共分散行列が対称行列であること (5.2.1項 (p.184)) をいちいち断わずに使っています。また、(8.4) では次の関係を使いました：(8.5) の両辺に右から C^T をかけて、

$$P_{t|t}C^T = P_{t|t-1}C^T(CP_{t|t-1}C^T + V)^{-1}((CP_{t|t-1}C^T + V) - CP_{t|t-1}C^T) = \text{い}\text{え}^{-1}V$$

ど確認してください。特に後者の場合は、行列 P が本書の P の転置になっていたり、 \mathbf{u}_t が縦ではなく横ベクトルだったりしますから、驚かないように。

8.a \mathbf{u}_t が縦ベクトルか横ベクトルかなんていう基本的なことがなぜ不統一なんでしょうねえ？

理工系と数学系とで流儀が違っているようです (→ 参考文献 [9] など)。参考文献 [32] の 1.1.1 でも述べているとおり、縦ベクトルにも横ベクトルにも一長一短があります。一般に、縦ベクトルのうれしい点として、見なれた $y = f(x)$ と同じ語順で $\mathbf{y} = A\mathbf{x}$ と書けることが挙げられます。一方、横ベクトルのうれしさは、「 \mathbf{x} に A した結果をさらに B して、それをさらに C する」がそのままの語順で $\mathbf{x}ABC$ と書けることです。Ruby 等の言語の得意技であるメソッドチェーン `x.split(…).select(…).sort.join(…)` に通じるものがありますね。

8.e.2 多重マルコフ連鎖

明日の状態の分布は今日と昨日と一昨日の状態に依存して決まる、といった場合、そのままだとマルコフ連鎖ではありませんが、読みかえによってマルコフ連鎖へ翻訳することが可能です。この場合の前提条件は、「今日と昨日と一昨日の状態を指定すれば、それより過去の状態には依存せず明日の状態 (の条件つき分布) が定まる」ということです。

$$P(\text{明日} \mid \text{今日, 昨日, 一昨日, それ以前}) = P(\text{明日} \mid \text{今日, 昨日, 一昨日})$$

これを 3 重マルコフ連鎖と呼びます。

いま、 Y_0, Y_1, Y_2, \dots が 3 重マルコフ連鎖だったとしましょう。これをマルコフ連鎖へ翻訳するには、 (Y_{t+2}, Y_{t+1}, Y_t) という組を考えてください。各 Y_t のとり得る値が n とおりなら、この組がとり得る値のバリエーションは $n \cdot n \cdot n = n^3$ とおりです。そこで、バリエーションに 1 番から n^3 番までの番号をつけておき、その番号を X_t と読みかえます。そうすれば X_t は式 (8.5)(p.294) を満たすので、マルコフ連鎖になっています。8.2.2 項 (p.289) 「カルマンフィルタ」の最後で述べた「車は急に止まれない」の話も、これと似た発想でした。

8.e.3 過去も無限にしたければ

本文では、マルコフ連鎖の時刻 t を 0 から始めました。過去も無限にして $\dots, X_{-1}, X_0, X_1, X_2, \dots$ という系列を考えたいほうがより一般的に見えますが、これにはちょっとしたワナがあります。

ワナとは何か。単純に

$$P(X_{t+1} = x_{t+1} \mid X_t = x_t, X_{t-1} = x_{t-1}, \dots) = P(X_{t+1} = x_{t+1} \mid X_t = x_t)$$

が成り立つときをマルコフ連鎖だと言ってみたくなるでしょうが、そこにワナがあります。条件つき確率を

$$P(X_{t+1} = x_{t+1} \mid X_t = x_t, X_{t-1} = x_{t-1}, \dots) = \frac{P(X_{t+1} = x_{t+1}, X_t = x_t, X_{t-1} = x_{t-1}, \dots)}{P(X_t = x_t, X_{t-1} = x_{t-1}, \dots)}$$

のように計算しようとしたら、この分母や分子が「どんな x_t, x_{t-1}, \dots をもってきても 0 になってしまう」という事態が起こります。それは、無限個の確率変数 X_t, X_{t-1}, \dots を考えているせいです。

状況は 4.2 節 (p.123) 「確率ゼロ」と同じ。特定の値がびったり出る確率はすべてゼロ、という事態です。このやっかいごとを避けるためには、ちょっととまどろっこしい言い方でマルコフ性を定義しないといけません。

任意の整数 $m \geq 3$ と任意の $t_1 < \dots < t_m$ に対して

$$P(X_{t_m} = x_m \mid X_{t_{m-1}} = x_{m-1}, \dots, X_{t_2} = x_2, X_{t_1} = x_1) = P(X_{t_m} = x_m \mid X_{t_{m-1}} = x_{m-1})$$

が常に成り立つとき、確率過程 X_t をマルコフ過程と呼ぶ。

ちなみに連続時間のマルコフ過程も、こんなふうにして定義されます。無限個の確率変数の独立性についても似たような話がありました (→ 2.c 節 (p. 補足編 16))。

8.e.4 推移確率行列の性質 (定常分布の存在)

本文で述べたとおり、どんな推移確率行列 P にも定常分布が必ず存在します。いまから線形代数の言葉を使ってそのあたりの説明をします。 $\mathbf{u}_t = P^t \mathbf{u}_0$ と表せることをまず思い出しましょう。こんなふうに行列 P を何度も何度もかけていくとき、未来のはての運命は P の固有値で決まります^{*7}。実は推移確率行列 P には、

- 固有値 1 を必ず持つ。しかもその固有ベクトルで、成分がすべて実数かつ ≥ 0 なものが存在する。
- 絶対値が 1 を越える固有値は持たない。

という性質があります。まず前者は、初期分布 \mathbf{u}_0 をうまく設定すれば $P\mathbf{u}_0 = \mathbf{u}_0$ とできることを意味しています^{*8}。一方、後者は、 \mathbf{u}_t がどんどん大きくなって爆発するようなことはないという保証を与えてくれます (まあもともと、 \mathbf{u}_t の各成分は ≥ 0 だし合計は必ず 1 なので、爆発の余地はないのです)。

上のようなことがなぜ保証されるのか。ポイントは、正方行列 P の各成分が非負なことで、各列の合計が 1 なことです (こんな行列を確率行列と呼びます)。

まず P の固有値の絶対値が 1 を越えないことについて。いま固有値 λ の固有ベクトル $\mathbf{v} = (v_1, \dots, v_n)^T$ があつたとしましょう。ベクトル \mathbf{x} の各成分の絶対値の合計を $\|\mathbf{x}\|_1$ と書くことにしたら、

$$\|P\mathbf{v}\|_1 = |p_{1\leftarrow 1}v_1 + \dots + p_{1\leftarrow n}v_n| + \dots + |p_{n\leftarrow 1}v_1 + \dots + p_{n\leftarrow n}v_n|$$

^{*7} 線形代数の教科書を参照 (たとえば参考文献 [32] の 4 章)。

^{*8} そういう固有ベクトル $\mathbf{v} = (v_1, \dots, v_n)^T$ を何か一本もってきて、 $\mathbf{u}_0 \equiv \mathbf{v}/(v_1 + \dots + v_n)$ ととればよい。固有ベクトルの定義より $\mathbf{v} \neq \mathbf{o}$ だから、分母がゼロになる心配はありません。こうすると、 \mathbf{u}_0 の各成分は ≥ 0 だし成分の合計は 1 になるので、確率分布の資格を満たします。また、 $P\mathbf{v} = \mathbf{v}$ から $P\mathbf{u}_0 = \mathbf{u}_0$ も OK です。

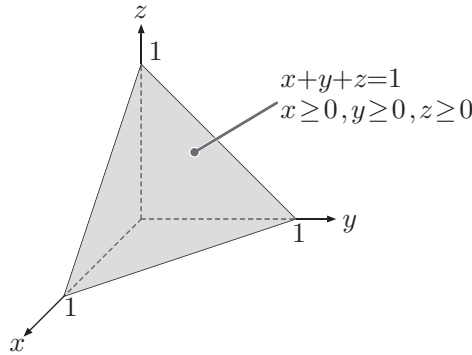
$$\begin{aligned}
&\leq (|p_{1\leftarrow 1}v_1| + \cdots + |p_{1\leftarrow n}v_n|) + \cdots + (|p_{n\leftarrow 1}v_1| + \cdots + |p_{n\leftarrow n}v_n|) \\
&\quad \cdots \cdots \text{一般に } |a+b| \leq |a| + |b| \text{ だから} \\
&= (p_{1\leftarrow 1}|v_1| + \cdots + p_{1\leftarrow n}|v_n|) + \cdots + (p_{n\leftarrow 1}|v_1| + \cdots + p_{n\leftarrow n}|v_n|) \\
&\quad \cdots \cdots P \text{ の成分はすべて } \geq 0 \text{ だから} \\
&= (p_{1\leftarrow 1} + \cdots + p_{n\leftarrow 1})|v_1| + \cdots + (p_{1\leftarrow n} + \cdots + p_{n\leftarrow n})|v_n| \\
&\quad \cdots \cdots |v_{\square}| \text{ ごとにまとめなおした} \\
&= |v_1| + \cdots + |v_n| \quad \cdots \cdots P \text{ の各列の合計は } 1 \text{ だから} \\
&= \|\mathbf{v}\|_1
\end{aligned}$$

また、固有ベクトルなのでから $P\mathbf{v} = \lambda\mathbf{v}$ のはずで、

$$\|P\mathbf{v}\|_1 = \|\lambda\mathbf{v}\|_1 = |\lambda|\|\mathbf{v}\|_1$$

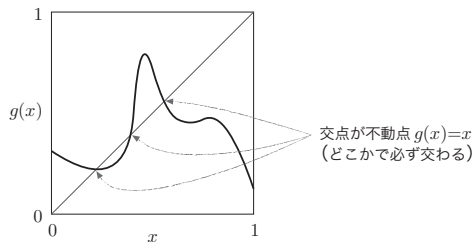
よって $|\lambda| \leq 1$ が導かれます (固有ベクトルという前提から $\|\mathbf{v}\|_1 > 0$ に注意)。

次は固有値 1 が必ず存在することについて。イメージしやすいように、状態数 $n=3$ の場合を説明します。確率分布に対応するような \mathbf{u}_0 とはこの場合、すべての成分が ≥ 0 で、しかもその合計が 1 な 3 次元縦ベクトルのことです。この条件を図示すると、図 8.1 のような三角形領域になります。さて、参考文献 [32] で強調されているとおり、行列 P をかけることは、(点 \mathbf{x} を点 $P\mathbf{x}$ へ移すような) 3 次元空間上の変換とみなせるのです。いまの話では、この三角形上の点が P によりまた三角形上のどこかへと移されることに注目してください。 $\mathbf{u}_1 = P\mathbf{u}_0$ も確率分布だ、という理解でもいいし、先ほど冒頭で述べたポイントから地道に確かめても結構です。さらに、 P による写像が連続なことも異論はないでしょう。 \mathbf{x} を連続的に動かしたら、その行き先 $P\mathbf{x}$ も連続的に動くはずで、いきなりジャンプしたりはしません。すると実は、三角形上に不動点が存在することが保証されます。つまり、 $P\mathbf{u}_0 = \mathbf{u}_0$ となる \mathbf{u}_0 が三角形上のどこかにあると保証されるのです。この \mathbf{u}_0 はまさに固有値 1 の固有ベクトル。こうして固有値 1 の存在が示されました。



► 図 8.1 確率分布に対応するのは三角形領域

不動点の存在の保証 (不動点定理) は本書の範疇を越えるので説明しませんが、次のような話の一般化だと思ってください：
 $0 \leq x \leq 1$ に対して $0 \leq g(x) \leq 1$ であるような連続関数 g が与えられたとしましょう。この g は、区間 $[0, 1]$ 上の各点を $[0, 1]$ 上のどこかへ移す連続写像だと解釈できます。そして実は、こういう g は必ず不動点を持つこと、つまり $g(x) = x$ となる点 x が $[0, 1]$ 上に必ず存在することが保証されます。理由は図 8.2 のとおりです。



► 図 8.2 不動点の存在 (g が連続なら、必ずどこかで対角線と交わる)

8.e.5 極限分布と線形代数

線形代数を勉強した方のために、あちらの言葉でも、極限分布について何が起きているのかを述べておきます。 \mathbf{u} が極限分布だということは、すべての初期分布 \mathbf{u}_0 に対して $\lim_{t \rightarrow \infty} P^t \mathbf{u}_0 = \mathbf{u}$ であることを意味しています。するとたとえば、確率 1 で状態 3 にいるという初期分布 $\mathbf{u}_0 = (0, 0, 1, 0, \dots, 0)^T$ をとれば、 P^t の第 3 列は \mathbf{u} に収束することがわかります (「えっ」という人は参考文献 [32] などを参照)。それは第 1 列でも第 2 列でも第何列でも同様。つまりこうです。

$$\lim_{t \rightarrow \infty} P^t = \left(\mathbf{u} \mid \cdots \mid \mathbf{u} \right)$$

行列慣れた人なら、もっと端的に

$$\lim_{t \rightarrow \infty} P^t = \mathbf{u} \mathbf{1}^T, \quad \mathbf{1}^T \equiv (1, \dots, 1)$$

と書くこともできます^{*9}。逆に、こうなっていれば \mathbf{u} が極限分布になることも、次のとおりすぐ確かめられます：どんな初期分布も $\mathbf{1}^T \mathbf{u}_0 = 1$ を満たすことに注意して、

$$\lim_{t \rightarrow \infty} P^t \mathbf{u}_0 = (\mathbf{u} \mathbf{1}^T) \mathbf{u}_0 = \mathbf{u} (\mathbf{1}^T \mathbf{u}_0) = \mathbf{u} \quad \cdots \cdots \text{確かに } \mathbf{u} \text{ は極限分布}$$

さて、前項で説明した P の固有値の性質を思い出してください。ベクトル \mathbf{u}_0 に P を何度もかけていくと、 $|\lambda| < 1$ の固有値 λ に対応する固有ベクトル方向の成分がどんどん縮みます。残るのは $|\lambda| = 1$ の固有値 λ に対応する方向だけです。特に、 $|\lambda| = 1$ となる固有値が $\lambda = 1$ だけで、しかもそれが単根だったら、 $t \rightarrow \infty$ のとき $\mathbf{u}_t = P^t \mathbf{u}_0$ はその固有ベクトルへと収束します。この収束先が極限分布です。

? 8.b 定常分布が一つしか存在しなかったら、それは必ず極限分布ですね？

いいえ。たとえば図 8.25^(p.300) の左側の例だと、定常分布は $\mathbf{u}_0 = (1/6, \dots, 1/6)^T$ だけです。でも、その本文で述べたとおり、極限分布は存在しません。

8.e.6 初めて戻ってくるまでの時間と極限分布との関係

一般に、「元と同じ状態に初めて戻ってくるまでの時間」は、極限分布と密接に関係します。本項でそのあたりを観察してみます。

状態 j からスタートして状態 i に初めて到達する時刻の期待値を $m_{i \leftarrow j}$ と書くことにして、それがすべて有限値だったとしましょう^{*10}。図 8.27^(p.302) のミニミニすごろくの議論を一般的に書けば、

$$\begin{aligned} m_{i \leftarrow j} &= 1 \cdot p_{i \leftarrow j} + \sum_{k \neq i} (1 + m_{i \leftarrow k}) p_{k \leftarrow j} \\ &= 1 + \sum_{k \neq i} m_{i \leftarrow k} p_{k \leftarrow j} \quad \cdots \cdots \left(\because \sum_k p_{k \leftarrow j} = 1 \text{ より} \right) \end{aligned}$$

でした。これはまとめて行列で表すこともできます。たとえば状態数 $n = 3$ なら、

$$\begin{pmatrix} m_{1 \leftarrow 1} & m_{1 \leftarrow 2} & m_{1 \leftarrow 3} \\ m_{2 \leftarrow 1} & m_{2 \leftarrow 2} & m_{2 \leftarrow 3} \\ m_{3 \leftarrow 1} & m_{3 \leftarrow 2} & m_{3 \leftarrow 3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} + \begin{pmatrix} 0 & m_{1 \leftarrow 2} & m_{1 \leftarrow 3} \\ m_{2 \leftarrow 1} & 0 & m_{2 \leftarrow 3} \\ m_{3 \leftarrow 1} & m_{3 \leftarrow 2} & 0 \end{pmatrix} \begin{pmatrix} p_{1 \leftarrow 1} & p_{1 \leftarrow 2} & p_{1 \leftarrow 3} \\ p_{2 \leftarrow 1} & p_{2 \leftarrow 2} & p_{2 \leftarrow 3} \\ p_{3 \leftarrow 1} & p_{3 \leftarrow 2} & p_{3 \leftarrow 3} \end{pmatrix}$$

両辺の各成分を見ると、確かに前の式と同じです。 $p_{i \leftarrow j}$ の並んだ行列がまさに推移確率行列 P であることに注目。

さて、いま $\mathbf{u} = (u_1, u_2, u_3)^T$ が P の定常分布だったとします。つまり、 $P\mathbf{u} = \mathbf{u}$ かつ $u_1 + u_2 + u_3 = 1$ です。すると、上式の両辺に右から \mathbf{u} をかけて得られる等式

$$\begin{aligned} &\begin{pmatrix} m_{1 \leftarrow 1} & m_{1 \leftarrow 2} & m_{1 \leftarrow 3} \\ m_{2 \leftarrow 1} & m_{2 \leftarrow 2} & m_{2 \leftarrow 3} \\ m_{3 \leftarrow 1} & m_{3 \leftarrow 2} & m_{3 \leftarrow 3} \end{pmatrix} \mathbf{u} \\ &= \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \mathbf{u} + \begin{pmatrix} 0 & m_{1 \leftarrow 2} & m_{1 \leftarrow 3} \\ m_{2 \leftarrow 1} & 0 & m_{2 \leftarrow 3} \\ m_{3 \leftarrow 1} & m_{3 \leftarrow 2} & 0 \end{pmatrix} \begin{pmatrix} p_{1 \leftarrow 1} & p_{1 \leftarrow 2} & p_{1 \leftarrow 3} \\ p_{2 \leftarrow 1} & p_{2 \leftarrow 2} & p_{2 \leftarrow 3} \\ p_{3 \leftarrow 1} & p_{3 \leftarrow 2} & p_{3 \leftarrow 3} \end{pmatrix} \mathbf{u} \end{aligned}$$

は、

$$\begin{pmatrix} m_{1 \leftarrow 1} u_1 \\ m_{2 \leftarrow 2} u_2 \\ m_{3 \leftarrow 3} u_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

と整理されます（やってみてください）。すなわち、

$$m_{i \leftarrow i} u_i = 1 \quad (i = 1, \dots, n)$$

なのです。なお、 $u_i = 0$ の場合は、 $m_{i \leftarrow j}$ を定義する際の仮定によりはじめから除外されています。それは具体的には、図 8.24^(p.299) のように「到達できない」という状態が存在する場合です^{*11}。そんな場合は極限分布が存在しないのでした。ですから、極限分布が存在するときには、 $u_i = 0$ の心配はいりません。

極限分布で i にいる確率が $1/10$ なら、 i から i へ戻ってくるのに期待値としては 10 回かかっている。極限分布のところでお話しした「 i にいる時間の割合」とも話があっていますね。

8.e.7 よくある誤解：マルコフ連鎖の時間を逆回しにしたら……

推移確率行列が P であるようなマルコフ連鎖の時間を逆回しにしたら、推移確率行列が P^{-1} であるようなマルコフ連鎖になる、と誤解する人がときどきいます。でもよく考えたら、そもそも逆行列 P^{-1} が存在するとは限らないし、存在しても P^{-1} は一般には推移確率行列になりません（成分が 0 から 1 までにおさまらない）。だから P^{-1} うんぬんは全く外れです。

じゃあ正解は？ 一般には、時間的に一様でないマルコフ連鎖になります。

吟味しやすいように単純な例を観察しましょう。状態はアカイかの二通りとし、推移確率は

^{*9} 太字の $\mathbf{1}$ はベクトルを、細字の 1 はただの数を表します。横ベクトルかける縦ベクトルは数、縦ベクトルかける横ベクトルは行列、という区別にも注意。

^{*10} もし「永遠にたどりつかない確率」が少しでもある場合には、 $m_{i \leftarrow j}$ は定義されず、本項の議論も成立しません。

^{*11} こう言える理由の概略を念のため：もし状態 j から状態 i へ到達可能だったら、適当な t をとれば $P(X_{s+t} = i | X_s = j) > 0$ のはず。この左辺は一般に、行列 P^t の (i, j) 成分に一致していました。すると、 $P^t \mathbf{u} = P \cdots P \mathbf{u} = \mathbf{u}$ により、 $u_j > 0$ なら $u_i > 0$ 。これがすべてのペア i, j で言えるときには、必然的にすべての i で $u_i > 0$ が結論されます。

$$p_{j \leftarrow i} = \begin{cases} 2/3 & (i = j) \\ 1/3 & (i \neq j) \end{cases}, \quad \text{つまり } P = \begin{pmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix}$$

初期分布は

$$P(X_0 = \text{ア}) = 2/3, \quad P(X_0 = \text{イ}) = 1/3, \quad \text{つまり } \mathbf{u}_0 = \begin{pmatrix} 2/3 \\ 1/3 \end{pmatrix}$$

だとします^{*12}。このときたとえば、

$$P(X_2 = \text{イ}, X_1 = \text{ア}, X_0 = \text{ア}) = p_{\text{イ} \leftarrow \text{ア}} p_{\text{ア} \leftarrow \text{ア}} P(X_0 = \text{ア}) = \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{2}{3} = \frac{4}{27}$$

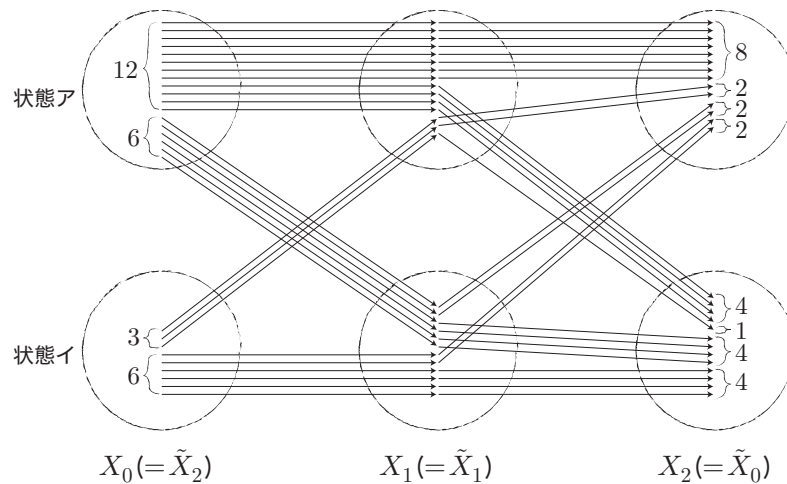
同じように計算していくと、 X_0, X_1, X_2 の同時分布は

	$(X_1, X_2) = (\text{ア}, \text{ア})$	$(X_1, X_2) = (\text{ア}, \text{イ})$	$(X_1, X_2) = (\text{イ}, \text{ア})$	$(X_1, X_2) = (\text{イ}, \text{イ})$
$X_0 = \text{ア}$	8/27	4/27	2/27	4/27
$X_0 = \text{イ}$	2/27	1/27	2/27	4/27

と求められます。これを絵で表したのが図 8.3 です。

- 最初の X_0 では、全体の 2/3 が状態ア、1/3 が状態イにいる。
- $X_0 \rightarrow X_1$ の遷移では、いずれも 1/3 が元と異なる状態へ移り、2/3 が元と同じ状態に留まる。
- $X_1 \rightarrow X_2$ の遷移でも、どこからその X_1 へ来たのかとは無関係に、1/3 が異なる状態へ移り、2/3 が同じ状態に留まる。

となっていることを確かめてください。たとえば、 $X_0 = \text{ア}$ から $X_1 = \text{ア}$ へやってきた 12 本のうち、1/3 にあたる 4 本が $X_2 = \text{イ}$ へ移り、2/3 にあたる 8 本が $X_2 = \text{ア}$ に留まっています。また、 $X_0 = \text{イ}$ から $X_1 = \text{ア}$ へやってきた 3 本についても、1/3 にあたる 1 本が $X_2 = \text{イ}$ へ移り、2/3 にあたる 2 本が $X_2 = \text{ア}$ に留まっています。「どこから $X_1 = \text{ア}$ へやってきた一群も、それぞれが同じ比率で $X_2 = \text{ア}$ と $X_2 = \text{イ}$ とに別れる」という事実が、この時系列のマルコフ連鎖っぷりを示しています。



► 図 8.3 マルコフ連鎖の時間を逆転すると？

では、 $(\tilde{X}_0, \tilde{X}_1, \tilde{X}_2) = (X_2, X_1, X_0)$ のように逆順に並べた時系列がマルコフ連鎖になっているか調べます。上のようなマルコフ連鎖っぷりが成り立っているでしょうか。

- $\tilde{X}_0 = \text{ア}$ から $\tilde{X}_1 = \text{ア}$ へやってきた 10 本のうち、1/5 にあたる 2 本が $\tilde{X}_2 = \text{イ}$ へ移り、4/5 にあたる 8 本が $\tilde{X}_2 = \text{ア}$ に留まる
- $\tilde{X}_0 = \text{イ}$ から $\tilde{X}_1 = \text{ア}$ へやってきた 5 本についても、1/5 にあたる 1 本が $\tilde{X}_2 = \text{イ}$ へ移り、4/5 にあたる 4 本が $\tilde{X}_2 = \text{ア}$ に留まる
- つまり、どこから $\tilde{X}_1 = \text{ア}$ へやってきた一群も、それぞれが同じ比率で $\tilde{X}_2 = \text{ア}$ と $\tilde{X}_2 = \text{イ}$ とに別れる

式で言えば

$$P(\tilde{X}_2 = i | \tilde{X}_1 = \text{ア}, \tilde{X}_0 = \text{ア}) = P(\tilde{X}_2 = i | \tilde{X}_1 = \text{ア}, \tilde{X}_0 = \text{イ}) = \begin{cases} 4/5 & (i = \text{ア}) \\ 1/5 & (i = \text{イ}) \end{cases}$$

これは、 k がアでもイでも

$$P(\tilde{X}_2 = i | \tilde{X}_1 = \text{ア}, \tilde{X}_0 = k) = P(\tilde{X}_2 = i | \tilde{X}_1 = \text{ア}) = \begin{cases} 4/5 & (i = \text{ア}) \\ 1/5 & (i = \text{イ}) \end{cases}$$

となることを意味します。同様に、

- $\tilde{X}_0 = \text{ア}$ から $\tilde{X}_1 = \text{イ}$ へやってきた 4 本のうち、1/2 にあたる 2 本が $\tilde{X}_2 = \text{ア}$ へ移り、1/2 にあたる 2 本が $\tilde{X}_2 = \text{イ}$ に留まる
- $\tilde{X}_0 = \text{イ}$ から $\tilde{X}_1 = \text{イ}$ へやってきた 8 本についても、1/2 にあたる 4 本が $\tilde{X}_2 = \text{ア}$ へ移り、1/2 にあたる 4 本が $\tilde{X}_2 = \text{イ}$ に留まる

^{*12} ちなみに、 $P^{-1} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ は推移確率行列の資格をまるで満たしていません。ですから、「時間を逆転した系列の推移確率行列が P^{-1} になる」という予想は、この例だけ見ても大はずれです。

から、 k がアでもイでも

$$P(\tilde{X}_2 = i | \tilde{X}_1 = \text{イ}, \tilde{X}_0 = k) = P(\tilde{X}_2 = i | \tilde{X}_1 = \text{イ}) = \begin{cases} 1/2 & (i = \text{ア}) \\ 1/2 & (i = \text{イ}) \end{cases}$$

以上より $P(\tilde{X}_2 = i | \tilde{X}_1 = j, \tilde{X}_0 = k)$ が $P(\tilde{X}_2 = i | \tilde{X}_1 = j)$ と常に一致するので、系列 $\tilde{X}_0, \tilde{X}_1, \tilde{X}_2$ もマルコフ連鎖です。ただし、もともとは対称 ($p_{\text{イ} \leftarrow \text{ア}} = p_{\text{ア} \leftarrow \text{イ}}$) だった推移確率が、時間を逆にすると対称でなくなっていました。よく見ると、もっと妙なことに気づきます。

$$P(\tilde{X}_1 = \text{ア} | \tilde{X}_0 = \text{イ}) = \frac{5}{13}$$

$$P(\tilde{X}_2 = \text{ア} | \tilde{X}_1 = \text{イ}) = \frac{1}{2}$$

のように、イからアへの推移確率が時刻によって変わってしまっています。ですから、マルコフ連鎖とはいっても、時間的に一様でないマルコフ連鎖になっているのです^{*13}。

以上の観察は次のように一般化できます。 X_0, X_1, \dots, X_s がマルコフ連鎖だったとしましょう^{*14}。現在時刻を t とし、そこから過去をまとめて $Y_t = (X_0, \dots, X_{t-1})$ 、未来をまとめて $Z_t = (X_{t+1}, \dots, X_s)$ と書くことにします。マルコフ連鎖という前提から、

$$P(X_{t+1} = \bigcirc | X_t = \triangle, Y_t = \square) = P(X_{t+1} = \bigcirc | X_t = \triangle)$$

…… 明日（の条件つき確率）は、過去によらず現在だけから定まる

です。さらに、その先の未来がどうなるかは X_{t+1} から決まるので、

$$P(Z_t = \bigcirc | X_t = \triangle, Y_t = \square) = P(Z_t = \bigcirc | X_t = \triangle)$$

…… 未来（の条件つき確率）は、過去によらず現在だけから定まる

も言えます。すると、

$$\begin{aligned} P(Z_t = \diamond, Y_t = \square | X_t = \triangle) \\ &= P(Z_t = \diamond | X_t = \triangle, Y_t = \square) P(Y_t = \square | X_t = \triangle) \quad \dots\dots \text{式 (2.9)}^{(p.50)} \text{より常に成立} \\ &= P(Z_t = \diamond | X_t = \triangle) P(Y_t = \square | X_t = \triangle) \quad \dots\dots \text{未来は現在だけから} \end{aligned}$$

が成り立つことになります。 $X_t = \triangle$ という条件のもとでは、 Z_t と Y_t との同時確率が、それぞれの周辺確率のかけ算になる。つまり、現在の状態を与えたとき、過去と未来とは条件つき独立になっているのです。この条件つき独立性から、反対に

$$\begin{aligned} P(Y_t = \square | X_t = \triangle, Z_t = \diamond) &= \frac{P(Y_t = \square, Z_t = \diamond | X_t = \triangle)}{P(Z_t = \diamond | X_t = \triangle)} \\ &= \frac{P(Z_t = \diamond | X_t = \triangle) P(Y_t = \square | X_t = \triangle)}{P(Z_t = \diamond | X_t = \triangle)} \\ &= P(Y_t = \square | X_t = \triangle) \end{aligned}$$

も導かれます。また標語的に言えば、「過去（の条件つき確率）は、未来によらず現在だけから定まる」。だから、時間を逆転した時系列もマルコフ連鎖になるのです。

ただし、一般には時間的に一様でないマルコフ連鎖になってしまうことは、先ほどの例でも述べたとおり。各時刻の推移確率は、初期分布 \mathbf{u}_0 に依存して決まります。

特例として、初期分布を定常分布 $\mathbf{u} \equiv (u_1, \dots, u_n)^T$ にとった場合には、時間を逆転してもちゃんと時間的に一様なマルコフ連鎖になります。実際、

$$\begin{aligned} P(X_t = j | X_{t+1} = i) &= \frac{P(X_{t+1} = i | X_t = j) P(X_t = j)}{P(X_{t+1} = i)} \quad \dots\dots \text{Bayes の公式} \\ &= \frac{p_{i \leftarrow j} u_j}{u_i} \end{aligned}$$

ですから、逆転版の $j \leftarrow i$ の推移確率は時刻 t によりません。

8.f エントロピーについての補足

8.f.1 エントロピーのとり得る範囲

例題 8.6

8.3.1 項 (p.305) 「エントロピー」で述べた不等式 $0 \leq H[X] \leq \log m$ を証明せよ (X のとり得る値が m とおりのとき)。付録 B.2 (p.337) 「Jensen の不等式」や付録 B.3 (p.339) 「Gibbs の不等式」を使ってよい。

^{*13} ぴんとこなかった方は、さらに極端に、 $P(X_0 = \text{ア}) = 1$ という初期分布の場合を考えてみてください。この場合、 $\tilde{X}_1 \rightarrow \tilde{X}_2$ については「必ず状態アへ移る」という格好ですから、明らかに $\tilde{X}_0 \rightarrow \tilde{X}_1$ のときの推移確率とは違っています。

^{*14} 無限の未来まで考えようとする、時間を逆転した際に「無限の過去」が生じてしまい、扱いがめんどどうになります (→ 8.e.3 項 (p. 補足編 68))。だからここでは有限時間で打ち切った話をします。

答

X の分布を $P(X = i) = p(i)$ とおく。確率なのだからもちろん $p(i) \leq 1$ のはず。すると、エントロピーの定義式 $H[X] = E[\log(1/p(X))]$ は「 ≥ 0 なものの期待値」なので $H[X] \geq 0$ 。また、一様分布を $q(i) = 1/m$ とおいて p と q に Gibbs の不等式を適用すれば、

$$\begin{aligned} D(p||q) &= \sum_i p(i) \log \frac{p(i)}{q(i)} = \sum_i p(i) \log(mp(i)) = \sum_i p(i) (\log m + \log p(i)) \\ &= (\log m) \sum_i p(i) + \sum_i p(i) \log p(i) = \log m - H[X] \end{aligned}$$

が 0 以上だとわかる。よって $H[X] \leq \log m$ 。等号成立は $p(i) = q(i)$ のとき、つまり X が一様分布のとき。

(別解) 関数 \log が上に凸なことから、Jensen の不等式 (付録 B.2^(p.337)) を使って

$$H[X] = E \left[\log \frac{1}{p(X)} \right] \leq \log E \left[\frac{1}{p(X)} \right] = \log \sum_i p(i) \frac{1}{p(i)} = \log \sum_i 1 = \log m$$

(ただし \sum_i は X のとり得る値すべてについての合計を表すとする) ■

8.f.2 連続値版のエントロピー

実数値の確率変数については、総和 \sum を積分 \int に読みかえてエントロピーを定義します。すなわち、 X の確率密度関数 $f_X(x)$ からびつくり度を

$$h(x) \equiv \log \frac{1}{f_X(x)}$$

とおき、エントロピーを

$$H[X] \equiv E[h(X)] = \int f_X(x) \log \frac{1}{f_X(x)} dx$$

(積分範囲は「 $f_X(x)$ がゼロでないところすべて」)

により定義します。

離散値版エントロピーの性質の多くは実数値版でも成り立ちます。ただし次の点には注意が必要です。

- 実数値版では $H[X] < 0$ がありえます。確率密度関数 $f_X(x)$ の値は 1 を越えることもあるからです (例題 4.2^(p.128))。
- 変数変換するとエントロピーは変わってしまいます。たとえば $Y = 3X$ と変換した場合、 $f_Y(y) = (1/3)f_X(x)$ のように密度が薄まるので ($y = 3x$)、

$$\log \frac{1}{f_Y(y)} = \log \frac{3}{f_X(x)} = \log \frac{1}{f_X(x)} + \log 3$$

のようにその分のおつりがつきます。したがって $H[Y] = H[X] + \log 3$ となります^{*15}。

また、次の事実は有名です。

- 分散が一定値 σ^2 であるような実数値の確率分布のうちで、エントロピーが最大になるのは、正規分布 $N(\mu, \sigma^2)$ です^{*16}。
- 変分法を習った方なら、以下のようにしてこの事実を導くことができます (微積分の取扱いなどが荒っぽいのはご容赦ください)。 $f(x)$ を確率密度関数としましょう。エントロピー H の定義から、定数 c のシフトはどうせ効きません ($g(x) = f(x+c)$ という確率密度関数を考えても H は同じになる)。そこで、はじめから期待値が 0 の場合に限定しておきます。解くべき問題は、

$$H = \int_{-\infty}^{\infty} f(x) \log \frac{1}{f(x)} dx \text{ を最大化せよ}$$

ただし $\int_{-\infty}^{\infty} f(x) dx = 1, \quad \int_{-\infty}^{\infty} x f(x) dx = 0, \quad \int_{-\infty}^{\infty} x^2 f(x) dx = \sigma^2$

そこで、制約を崩さないように f をちょっとずらしてみ、 H の変化を調べます。具体的には、

$$\int_{-\infty}^{\infty} u(x) dx = 0, \quad \int_{-\infty}^{\infty} x u(x) dx = 0, \quad \int_{-\infty}^{\infty} x^2 u(x) dx = 0$$

を満たす関数 u に対して $f(x; s) \equiv f(x) + s u(x)$ とおき、

$$r(s) \equiv \int_{-\infty}^{\infty} f(x; s) \log \frac{1}{f(x; s)} dx = - \int_{-\infty}^{\infty} f(x; s) \log f(x; s) dx$$

の微分 r' を計算してみます^{*17}。

$$r'(0) = - \int_{-\infty}^{\infty} u(x) \log f(x) dx - \int_{-\infty}^{\infty} f(x) \cdot \frac{u(x)}{f(x)} dx$$

^{*15} 「 X も Y も実質的には同じ情報のはずなのに、エントロピーが増えるなんて不合理だ。情報の量を正しく測れていないじゃないか」という文句が出そうですね。これは「ノイズへの耐性が増えたのだ」と思ってください。ざわついた喫茶店で音楽を流しても、音量が小さいと、「バイオリンだな」という程度しかわかりません。音量をもっと上げれば、演奏の細部までが聞きとれるようになります。デジタル情報と違い、アナログ情報の量はこんなふうに測るほうが自然です。(もしノイズを考えず無限精度の測定ができるとしたら、連続値一個だけでも無限の情報を持つてしまう。たとえば A さんの体重は 67.8644883816546072912291434934592153539385881... kg だ、など)

^{*16} もし分散の制限をつけないと、エントロピーはいくらでも大きくできるので、意味のある問いかけになりません。

^{*17} (別解) $u(x)$ をはじめから制限するのではなく、ラグランジュ未定係数法を使って解くこともできます。

$$\begin{aligned}
&= - \int_{-\infty}^{\infty} u(x) \log f(x) dx - \int_{-\infty}^{\infty} u(x) dx \\
&= - \int_{-\infty}^{\infty} u(x) \log f(x) dx - 0
\end{aligned}$$

最適な f だったら、この $r'(0)$ が常に 0 のはず。どんな u に対しても $r'(0) = 0$ となるためには、(u の制限と見比べて)

$$\log f(x) = \alpha x^2 + \beta x + \gamma, \quad (\alpha, \beta, \gamma \text{ は定数})$$

とおくしかありません。言い換えれば $f(x) = \exp(2 \text{ 次式})$ の形で^{*18}。これは f が正規分布の確率密度関数であることを意味するのです (→ 4.6.2 項 (p.164) 「一般の正規分布」)。

8.g 情報源符号化定理の証明

8.3.3 項 (p.311) で紹介した情報源符号化定理のより正確な記述とその証明を以下で述べます。設定はこうです。

- 長さ n の文字列を長さ m の bit 列に圧縮したい
- ただし圧縮失敗の確率は、ある定数 δ より小さくしたい
- n が大きいという想定で、圧縮率 m/n はどこまで下げられるだろうか

8.g.1 くじ買占め問題への翻訳

文字列圧縮という言い方は抽象的でびんこないかもしれませんから、別のたとえ話にまず翻訳します。

こんどのくじ引きの一等賞品（非売品）を、あなたはどうしても当てたいとしましょう。絶対確実に当てようと思ったらくじをすべて買占めるしかありませんが、そこまでのお金は出せません。そこでほんの少し (δ) だけ妥協して、確率 $(1 - \delta)$ 以上で当ててをめざします。くじを何枚買占めれば確率 $(1 - \delta)$ 以上を確保できるでしょうか？

各くじには文字列が書かれていて、これが当選文字列 X と一致したら当たりです。 X の設定は 8.3.3 項 (p.311) と同じ。 k 種類の文字からなる長さ n の文字列で、1 文字目 X_1 、2 文字目 X_2 、……はそれぞれ独立にルーレットで決められます。

実は本文の文字列圧縮は、いまの設定で 2^m 枚のくじを買占めることに翻訳されます。理由は次のように考えてください。買占めたくじの裏に、整理コードとして通し番号をふります。ただし、1 番、2 番、と数字を書くかわりに、長さ m の bit 列「000...00」「000...01」……「111...11」を割り当てます。そんな bit 列は 2^m とおりですから、これでちょうど買占めた枚数分びつたりです (通し番号を二進数で書いたと思っても結構です)。

では買占めたくじ一式を使って、文字列の圧縮と復元をやってみせます。それには、圧縮したい文字列 X を当選文字列と思えばよい。

- もしくじが当たっていたら (つまり、 X と一致するくじが買占めた中にあれば)、その整理コードを圧縮 bit 列 Y として答える。
- もしくじがはずれていたら、「圧縮できませんでした」と答える。

これで確かに、くじの当選確率が文字列圧縮の成功確率へ翻訳されました。圧縮 bit 列 Y から元の文字列 X を復元する際は、整理コード Y がついたくじの文字列を答えるだけです。

8.g.2 最適戦略

文字列圧縮とくじ買占めが等価だと確かめられたので、ここからは主に後者で話をしていきます。

当選文字列を決めるルーレットは、どの文字がどれくらい出やすいか前もってわかっている設定でした。だから、どのくじ (文字列) がどれくらい当たりやすいかも前もってわかることになります。それならもちろん、当たりやすいくじを買うほうが得です。つまり、

- 当たりやすいくじから順番にどんどん買っていく
- 買ったくじの当選確率の合計が $(1 - \delta)$ 以上になったところで買占め終了
- あとは当選発表を待つ

というのが最適戦略です。こうすれば最少の買占め枚数で当選確率 $(1 - \delta)$ を確保できます。そのとき結局何枚買うはめになるか。それがいま議論したい問題です。

前に図 8.31 (p.312) でお見せした数値例は、上述の最適戦略を用いたときの結果でした。

8.g.3 これだけ買えば十分

最適戦略は確かに最適で結構なのですが、解析しにくくて話が進みません。そこで、最適ではないけれど計算しやすい、以下のような「典型くじ戦略」を考えます。これは、極端に当たりにくいくじや当たりやすいくじを避けて、典型的な当たりやすさのくじだけ買占めという戦略です。もちろん最適戦略ならもっと少ない枚数で済むはずですから、典型くじ戦略を調べることににより、必要枚数の上限 (これだけ買えば十分) が得られます。

典型くじ戦略をきちんと説明するには、少し準備がいります。当たりにくさ・当たりやすさは、(もし当たったときの) びつくり度の大小と言いかえられることに着目してください。文字 x_i のびつくり度は $\log(1/r(x_i))$ でした。文字列 $x \equiv (x_1, \dots, x_n)$ のびつくり度は、独立性から、

$$\log \frac{1}{r(x_1) \cdots r(x_n)} = \log \frac{1}{r(x_1)} + \cdots + \log \frac{1}{r(x_n)}$$

のように各文字のびつくり度の合計となります。これを文字数 n で割って、文字列 x の一文字あたりのびつくり度

^{*18} 丁寧に途中を書けば、

$$f(x) = 2^{\alpha x^2 + \beta x + \gamma} = (e^{\log_e 2})^{\alpha x^2 + \beta x + \gamma} = e^{(\log_e 2)(\alpha x^2 + \beta x + \gamma)}$$

上式は結局 $e^{(2 \text{ 次式})}$ 、つまり $\exp(2 \text{ 次式})$ の格好。式変形にとまどった方は付録 A.5 (p.326) 「指数と対数」を参照してください。

$$h(x) \equiv \frac{1}{n} \left(\log \frac{1}{r(x_1)} + \cdots + \log \frac{1}{r(x_n)} \right)$$

という量を考えます。文字列 x には、出にくい（びつくり度 $h(x)$ が高い）ものもあるし、出やすい（びつくり度 $h(x)$ が低い）ものもあります。典型くじ戦略は、両極端を避けて平均的なびつくり度のくじを買い集める戦略です。そのために、平均的なびつくり度、すなわち「びつくり度の期待値」を計算しておかないといけません^{*19}。びつくり度の期待値とはエントロピーのことでした。当選文字列 $X \equiv (X_1, \dots, X_n)$ の各文字 X_i のエントロピー $H[X_i]$ を、 η という記号で表します^{*20}。具体的には

$$\eta \equiv H[X_i] = E \left[\log \frac{1}{r(X_i)} \right] = \sum_{j=1}^k r(a_j) \log \frac{1}{r(a_j)}$$

と計算されます（i.i.d. なので答は i によらず一定）。そして、文字列 X のエントロピーは $H[X] = n\eta$ 。だから結局、 $E[h(X)] = \eta$ です。

以上をふまえて、 $h(x)$ の値が η に近いくじ（平均的な当たりやすさのくじ^{*21}） x を典型くじ^{*22}と呼ぶことにします。この典型くじを買占めるのが典型くじ戦略です。「近い」の幅は正定数 ϵ で表すことにします。つまり、

$$\eta - \epsilon < h(x) < \eta + \epsilon \quad (8.6)$$

を満たすくじ x をすべて買占めるわけです。 ϵ の具体的な値は、当選確率 $(1 - \delta)$ を確保できるように後ほど調節します。

買占め幅 ϵ と買占め枚数 2^m との関係は、次のように見積られます。買占めたくじ x のびつくり度の範囲 (8.6) を当選確率に直すと、

$$2^{-n(\eta+\epsilon)} < P(X = x) < 2^{-n(\eta-\epsilon)} \quad (8.7)$$

です。つまり、一枚一枚が少なくとも $p_{\min} \equiv 2^{-n(\eta+\epsilon)}$ の当選確率を持っています。そんなくじを 2^m 枚買占めたら、合計で少なくとも $2^m p_{\min}$ だけの当選確率を確保したことになります。ところが、当選確率は 1 を越えないはずなので、 $2^m p_{\min} \leq 1$ 。よって、

$$2^m \leq \frac{1}{p_{\min}} = 2^{n(\eta+\epsilon)}$$

要するに $m \leq n(\eta + \epsilon)$ です。

次は当選確率の見積り。Chebyshev の不等式というものを使いたいので、準備として分散 $V[h(X)]$ を計算します。 X の各文字 X_1, \dots, X_n が独立なことから、

$$\begin{aligned} V[h(X)] &= V \left[\frac{1}{n} \log \frac{1}{P(X=x)} \right] = \frac{1}{n^2} V \left[\log \frac{1}{r(X_1) \cdots r(X_n)} \right] \\ &= \frac{1}{n^2} V \left[\log \frac{1}{r(X_1)} + \cdots + \log \frac{1}{r(X_n)} \right] \\ &= \frac{1}{n^2} \left(V \left[\log \frac{1}{r(X_1)} \right] + \cdots + V \left[\log \frac{1}{r(X_n)} \right] \right) \end{aligned}$$

ここで、 $V[\log(1/r(X_i))]$ はどの i でもすべて等しい（各文字は i.i.d.）から、

$$\sigma^2 \equiv V \left[\log \frac{1}{r(X_1)} \right] = \cdots = V \left[\log \frac{1}{r(X_n)} \right]$$

と共通の名前をつけましょう。 σ^2 の具体的な値は、 $r(a_1), \dots, r(a_k)$ から計算されます。すると結局、

$$V[h(X)] = \frac{\sigma^2}{n}$$

これが求まれば、付録 B.4^(p.340) の Chebyshev の不等式から

$$P(|h(X) - \eta| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2 n}$$

が保証されます。ということは、

$$P(\text{当選文字列 } X \text{ が買占めたくじの中にある}) = P(|h(X) - \eta| < \epsilon) \geq 1 - \frac{\sigma^2}{\epsilon^2 n}$$

のように買占め幅 ϵ に応じた当選確率が保証されます。

あとは当選確率が希望の $(1 - \delta)$ 以上になるよう、範囲 ϵ を設定すればよい。それは

$$\frac{\sigma^2}{\epsilon^2 n} \leq \delta$$

ということだから、

$$\epsilon \geq \sqrt{\frac{\sigma^2}{n\delta}}$$

^{*19} 当選文字列はランダムなんだから、どれぐらいびつくりな当選文字列が出るかもランダム。びつくりしない（ありがちな）当選文字列が出る場合もあるだろうし、びつくりな（めずらしい）当選文字列が出る場合もあるだろうし。

^{*20} エントロピーには大文字の H をあてるのがふつうですが、ゆらがいないただの数であることを忘れないよう、ここではギリシャ文字 η （イータ）を使うことにします。

^{*21} 「そのくじがもし当たったときのびつくり度」が平均的。ということは結局、そのくじの当たりやすさもある意味で平均的。

^{*22} ちゃんとした用語は typical sequence。「典型的な文字列」ということ。

と設定すれば大丈夫です。

まとめると、当選確率 $(1 - \delta)$ を確保するには、 $m \geq n(\eta + \epsilon)$ と設定して 2^m 枚買占めれば十分。文字列圧縮の話に戻せば、圧縮率が

$$\frac{m}{n} \geq \eta + \epsilon \quad \text{ただし } \epsilon = \sqrt{\frac{\sigma^2}{n\delta}}$$

なら OK ということです。つまり、圧縮率を η よりちょっとだけ甘くしておけば、所望の圧縮成功確率 $(1 - \delta)$ が実現可能です。しかも、文字列長 n を大きくすれば、そのちょっとはいくらでも小さくすることができます。

見た目には長い文字列でも、(一文字あたりの) エントロピーが小さければ、短い bit 列に圧縮できる。エントロピーが本質的な情報の量をとらえていると感じていただけたでしょうか。

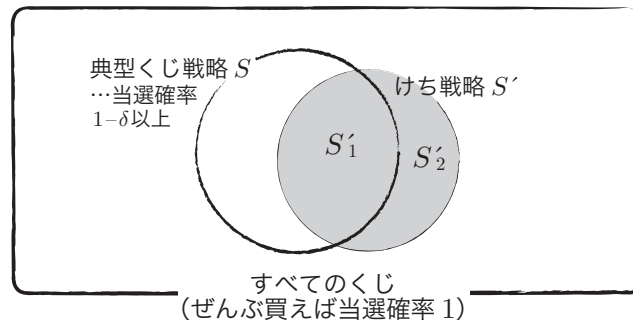
8.g.4 少なくともこれだけは買わないと

ここまでで、当選確率 $(1 - \delta)$ を確保するには何枚買占めれば十分かを評価しました。結果は、不正確な表現でざっくり言えば、 $2^{n\eta}$ 枚 (よりちょっとだけ多く) 買えということでした。ここからは逆に、少なくとも何枚は買占める必要があるかを評価し、次の意味でやはり $2^{n\eta}$ 枚が限界なことを示します。

(意味)

資金をけちって $2^{n(\eta - \epsilon')}$ 枚以下しか買占めないことにする (ϵ' は任意の小さな正定数)。その場合、 n がある程度大きくなると、当選確率 $(1 - \delta)$ は決して確保できない²³。

この「けち戦略」で買占めるくじの集合を S' と置きます。けち戦略に対抗して、われらが典型くじ戦略で $0 < \epsilon \leq 0.01\epsilon'$ ととったときに買占めるくじの集合を S と置きましょう。 n が十分大きければ、この小さな幅 ϵ でも当選確率 $(1 - \delta)$ を確保できます。それが前の結論でした。さて、図 8.4 を見てください。こんなふうに、 S' を「典型くじ戦略 S との共通部分」 $S'_1 = S' \cap S$ とそれ以外 S'_2 とに分けて、おののお当選確率を評価してみます。



▶ 図 8.4 共通部分とそれ以外とに分けて評価

S'_1 については、式 (8.7) (p. 補足編 75) より一枚一枚の当選確率が高々 $2^{-n(\eta - \epsilon)}$ しかなくて、枚数も元々の前提から $2^{n(\eta - \epsilon')}$ 以下です。だから、 S'_1 内のくじ全部をあわせても、当選確率はせいぜい

$$P(X \in S'_1) \leq 2^{-n(\eta - \epsilon)} 2^{n(\eta - \epsilon')} \leq 2^{-0.99n\epsilon'}$$

しかありません。 n が大きくなれば、この当選確率はどんどん小さくなってしまいます。一方、 S'_2 については、当選確率なんてはじめから

$$P(X \in S'_2) \leq \delta$$

しかないのです。それは図 8.4 から明らかでしょう。 S だけで当選確率 $(1 - \delta)$ を確保しているのですから、その外はどんなにがんばってもせいぜい確率 δ しかありません。そうすると、 S'_1 と S'_2 の当選確率を合計しても

$$P(X \in S') \leq \delta + 2^{-0.99n\epsilon'}$$

しかありません。 n が大きくなれば、希望の当選確率 $(1 - \delta)$ を実現することは不可能です。

文字列圧縮の話に戻せばこうなります：正定数 ϵ' がどんなに小さくても、圧縮率 $m/n \leq \eta - \epsilon'$ を実現することは不可能 (n が大きくなると圧縮成功確率 $1 - \delta$ を確保できない)。この意味で、一文字あたりのエントロピー η は圧縮率の限界を与えます。

雑な表現を許してもらえらるなら、本項の結論は次のとおり。

文字列を長くしていけば、圧縮率をエントロピーにいくらでも近づけることが可能。しかし、エントロピーより低い圧縮率は実現不可能。

エントロピーが本質的な情報の量をとらえていることを味わってください。

付録 C.3 (p. 352) 「Kullback-Leibler divergence と大偏差原理」にも関連する話題があります。また、7.2 節 (p. 259) 「一様分布以外の乱数の生成」で触れたハフマン法も、情報源符号化に用いられます。

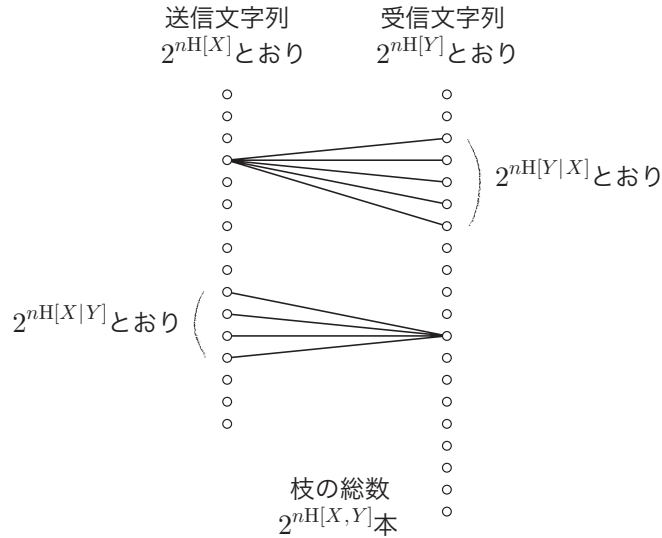
8.h 通信路符号化定理の証明の概略

8.3.4 項 (p. 313) で述べた通信路符号化定理について、証明の概略を粗っぽく紹介します。

²³ ただし、 $0 < \delta < 1/2$ とします。ほんの少しの妥協 δ という設定だったのだから、この仮定に文句はないはず。

8.h.1 通信レートが容量より低い場合

いま仮に、送信文字 X の分布 $P(X=x)$ を何か設定し、その分布で i.i.d. に n 文字の送信文字列を生成したとします。 n が大きければ、この送信文字列ではおよそ $2^{nH[X]}$ とおりがほぼ等確率で生じること、そしてそれ以外が生じる確率は無視できることが、8.g.3 項 (p. 補足編 74) の典型的くじ戦略から言えます。受信文字 Y についても同様に、 $2^{nH[Y]}$ とおりの受信文字列が等確率で生じ、さらに各受信文字列には $2^{nH[X|Y]}$ とおりの送信文字列が対応します (それ以外の確率は無視できる)。まとめると図 8.5 のような具合です²⁴。



► 図 8.5 典型的な送受信文字列。枝の総数は $2^{nH[Y|X]}2^{nH[X]} = 2^{nH[X|Y]}2^{nH[Y]} = 2^{nH[X,Y]}$ 本

さて、通信レートが r なら、もとのメッセージは長さ rn の bit 列ですから 2^{rn} とおり。符号化方式を決めることは、これらのメッセージそれぞれに対して送信文字列を何か割り当てることだと言いかえられます (何を何に割り当てるかはあらかじめ送信者と受信者として打合せておいてから、実際の通信を行います)。ここでは、上述の $2^{nH[X]}$ とおりの中からランダムに割り当てた場合を考えてみましょう。いまあるメッセージ M を符号化して送信し、受信文字列を得たとします。受信文字列がこれになる可能性のある送信文字列候補は、上のように $2^{nH[X|Y]}$ とおり。もしその中に割り当てられているメッセージが M ただ一つだったら、受信者は自信をもって「メッセージは M ですね」と答えられますから、通信成功です。そんなふうに通信が成功する確率は、「確率 $2^{nH[X|Y]}/2^{nH[X]}$ で表が出るコインを $(2^{rn}-1)$ 回投げた後、すべて裏になる確率」と見積れます (無視できる誤差は除いて)。この確率は

$$\left(1 - \frac{2^{nH[X|Y]}}{2^{nH[X]}}\right)^{(2^{rn}-1)} = \left(1 - 2^{nH[X|Y]-nH[X]}\right)^{(2^{rn}-1)} = \left(1 - 2^{-nI[X;Y]}\right)^{(2^{rn}-1)}$$

と計算され、 $r < I[X;Y]$ なら $n \rightarrow \infty$ で右辺は 1 に収束します (式 (A.6)(p.328) を参照)。したがって、通信レート r を通信路の容量より低くしておけば、通信成功の確率をいくらでも 1 に近づけられることがわかりました。

以上で言えたのは、「いろいろな符号化方式の中からランダムにどれかを使うとしたら、通信成功の確率は……」という結果です。だから、定理の主張そのもの (通信成功確率が……であるようなうまい符号化方式が存在する) とはまだ少し隔りがあります。そこを埋めるには、次のことに気づけばよい: 「上の結果は、いろいろな符号化方式の成功確率の平均値が……だということを意味する。そして一般に、何かの平均値がある値 a となるためには、少なくとも一個は $\geq a$ のものがないといけない」。こうして、うまい符号化方式の存在が保証されます。

8.h.2 通信レートが容量より高い場合

逆に通信レート r が通信路の容量 c より高い場合の不可能性は、おおよそ以下のような筋道で証明されます: もし仮に通信誤りの確率をほぼ 0 にできたとして、それは、受信文字列を見て送信文字列がほぼ当てられることだ。つまり、 $H[\text{送信文字列} | \text{受信文字列}]$ がほぼ 0 ということだ。言い換えれば、 $I[\text{送信文字列}; \text{受信文字列}]$ が $H[\text{送信文字列}]$ にほぼ一致している。ここで実は、 $I[\text{送信文字列}; \text{受信文字列}] \leq nc$ が成り立つ²⁵。また、メッセージが圧縮という前提から $H[\text{送信文字列}] = rn$ (さもなくば、情報源符号化によりもっと圧縮して通信レートを上げられるはず)。こういった計算により $r \leq c$ が導かれます。

²⁴ ここまでを数え終わったら、もう仮の分布 $P(X=x)$ は忘れてください。この分布は上述の文字列集合を定めるためのものです。この分布のとおりに通信するというわけではありません。

²⁵ 1 文字の送信文字と受信文字との間で高々 c しか情報を共有できないのだから、 n 文字なら高々 nc だという話。式で書くとうです: 送信文字列の i 文字目を X_i 、受信文字列の i 文字目を Y_i とおけば、図 8.29(p.309) の性質から $H[\text{受信文字列}] \leq \sum_{i=1}^n H[Y_i]$ 。また、脚注 30(p.314) で述べた無記憶性より $H[\text{受信文字列} | \text{送信文字列}] = \sum_{i=1}^n H[Y_i | X_i]$ 。これらを辺々引いて $I[\text{送信文字列}; \text{受信文字列}] \leq \sum_{i=1}^n I[X_i; Y_i] \leq nc$ 。

記号・数字・ギリシア文字

*	(畳み込み)	補足編 21
\Leftrightarrow	(同値)	127
\approx	(およそ)	43
\cap	(共通部分)	320
\cdot	(内積)	333
\cup	(和集合)	320
\emptyset	(空集合)	320
\in	(属す)	320
∂	(偏微分)	154
\propto	(比例)	343, 補足編 17
\setminus	(差集合)	補足編 6
\sim	(従う)	164
\subset	(包含)	320
\vee	(最大値)	320
\wedge	(最小値)	320
$\binom{n}{k}$	(組合せ)	76
$\ \cdot\ $	(長さ)	334
$ \cdot $	(絶対値)	319
${}_nC_k$	(組合せ)	76
${}_nP_k$	(順列)	75
\equiv	(定義)	14
T	(転置)	185
!	(階乗)	75
1 次式		167
2 項係数		76
2 項分布	74, 82, 98, 230, 245, 286, 341, 補足編 23, 補足編 44, 補足編 47	
2 次式		166
2 次平均取束		349, 350, 補足編 28
3 重マルコフ連鎖		補足編 68
Γ 関数	\rightarrow G の項目にある Γ 関数	
ρ	\rightarrow 相関係数	
σ	\rightarrow 標準偏差	
χ^2 分布	\rightarrow カイ自乗分布	

A

a.a.	補足編 35
a.e.	補足編 35
a.s.	補足編 35
accept	244
AIC	243
Akaike's information criterion	243
almost all	補足編 35
almost everywhere	補足編 35
almost surely	補足編 35

B

Baum-Welch アルゴリズム	303
Bayesian filter	243
Bayesian information criterion	243
Bayesian network	243
Bayes 推定	240, 277
Bayes 統計	補足編 55
Bayes の公式	30, 56, 145, 155, 240, 241, 補足編 72
BIC	243
binomial distribution	\rightarrow 2 項分布
bit	306
Bn	\rightarrow 2 項分布
bootstrap 法	243
Brown 運動	304

C

${}_nC_k$ (組合せ)	76
Cauchy-Schwarz の不等式	\rightarrow シュワルツの不等式
Cauchy の不等式	\rightarrow シュワルツの不等式
Cauchy 分布	351
Chebyshev の不等式	341, 342, 補足編 75
Chernoff 限界	341
χ^2 分布	\rightarrow カイ自乗分布
Cholesky 分解	265
coherence	補足編 8
combination	76
convolution	\rightarrow 畳み込み

covariance	\rightarrow 共分散
Cox の公理	補足編 8
Cox の定理	補足編 8
cross validation	243
CV	243

D

∂ (偏微分)	154
δ 関数	\rightarrow デルタ関数
det	\rightarrow 行列式
diag (対角行列)	199
duck typing	補足編 5
Dutch book	補足編 8

E

E	\rightarrow 期待値
$E[\cdot]$ (条件つき期待値)	107
e (自然対数の底)	327
EM アルゴリズム	304
ϵ - δ 論法	補足編 31
erf	\rightarrow 誤差関数
error function	\rightarrow 誤差関数
estimator	234
Euler の公式	328
exp (指数関数)	327
expectation	\rightarrow 期待値
exponential function	327

F

F_X	\rightarrow 累積分布関数
f_X	\rightarrow 確率密度関数
$f_{X,Y}$ (同時分布の確率密度関数)	136
$f_{Y X}$ (条件つき分布の確率密度関数)	144
false accept	246
false reject	246
Fibonacci 数列	254
Forward-Backward アルゴリズム	303
Fourier 変換	\rightarrow フーリエ変換

G

Γ 関数	213, 補足編 23
Gauss 積分	\rightarrow ガウス積分
Gauss 分布	\rightarrow 正規分布
Gibbs の不等式	339, 補足編 73
Google	294

H

H	\rightarrow エントロピー
$H[\cdot, \cdot]$ (同時エントロピー)	308
$H[\cdot]$ (条件つきエントロピー)	308
Halton 列	258
hidden Markov model	303
HMM	303
Hölder の不等式	343
Huffman 法	\rightarrow ハフマン法

I

I (単位行列)	196
I	\rightarrow 相互情報量
i (虚数単位)	320
i.i.d.	103
ICA	\rightarrow 独立成分分析
independent component analysis	\rightarrow 独立成分分析

J

Jensen の不等式	337, 340, 345, 補足編 73
-------------	-----------------------

K

Kac の定理	351
Kalman フィルタ	\rightarrow カルマンフィルタ
Kolmogorov complexity	補足編 60
Kullback-Leibler divergence	340, 342, 353

Kullback-Leibler 距離.....→ Kullback-Leibler divergence

L

Lagrange 未定係数法.....→ ラグランジュ未定係数法
 λ199
 Lebesgue 積分.....→ ルベグ積分
 Lisp.....199
 logarithm.....331
 low-discrepancy sequence.....257
 LU 分解.....265

M

Markov chain.....→ マルコフ連鎖
 Markov chain Monte Carlo method
→ マルコフ連鎖モンテカルロ法
 Markov process.....→ マルコフ過程
 Markov 過程.....→ マルコフ過程
 Markov の不等式.....340, 341, 342, 349
 Markov 連鎖.....→ マルコフ連鎖
 Markov 連鎖モンテカルロ法.....→ マルコフ連鎖モンテカルロ法
 max.....→ 最大値
 MCMC.....→ マルコフ連鎖モンテカルロ法
 MDL.....243
 Mersenne Twister.....255
 min.....→ 最小値
 minimum description length.....243
 Minkowski の不等式.....342
 Monte Carlo 法.....→ モンテカルロ法
 Monty Hall problem.....→ モンティホール問題

N

N.....→ 正規分布
 Napier 数.....327
 natural logarithm.....331
 Neyman-Pearson の補題.....249
 nonparametric.....234
 nonparametric Bayes 法.....243
 normal distribution.....→ 正規分布
 n 次元標準正規分布.....195

P

P.....→ 確率
 $P(\cdot, \cdot)$→ 同時確率
 $P(\cdot | \cdot)$→ 条件つき確率
 ρ (ピーではなくロー).....→ 相関係数
 $n P_k$ (順列).....75
 p -value.....→ p 値
 PageRank.....294
 parametric.....234
 PCA.....→ 主成分分析
 Pearson の補題.....249
 permutation.....75
 Petersburg のパラドックス.....85
 π→ 円周率
 Pr.....→ 確率
 principal component analysis.....→ 主成分分析
 Prob.....→ 確率
 probability.....→ 確率
 p 値.....244, 249

Q

QR 分解.....補足編 49

R

random variable.....16
 rank (ランク, 階数).....補足編 52
 reject.....244
 ρ→ 相関係数
 Riemann 積分.....114
 Riemann 面.....326
 robust.....補足編 55
 Ruby.....101, 262, 補足編 6, 補足編 68
 r.v. (確率変数).....16

S

Scheme.....199
 Schur complement.....補足編 51

Schwarz の不等式.....→ シュワルツの不等式
 seed.....254
 Shannon 情報量.....306
 Sherman-Morrison-Woodbury の公式.....補足編 51
 σ→ 標準偏差
 σ 加法族.....補足編 6, 補足編 30
 σ 集合体.....→ σ 加法族
 σ 代数.....→ σ 加法族
 Simpson のパラドックス.....45
 s.pr. (確率過程).....285
 St. Petersburg のパラドックス.....85
 standard deviation.....92
 Stein のパラドックス.....補足編 56
 Stirling の公式.....337, 353, 補足編 24
 stochastic process.....→ 確率過程

T

Tikhonov の正則化.....→ チコノフの正則化
 Tr.....→ トレース
 trace.....→ トレース
 typical sequence.....補足編 75

U

UMP.....→ 一様最強力不偏検定
 UMVUE.....237, 補足編 56
 uniformly minimum variance unbiased estimator
→ UMVUE
 uniformly most powerful unbiased test
→ 一様最強力不偏検定
 utility function.....→ 効用関数

V

V.....→ 分散
 V (共分散行列).....186
 Var.....→ 分散
 variance.....→ 分散
 Viterbi アルゴリズム.....303

W

Walker の別名法.....補足編 58
 web ページ.....294
 Woodbury の公式.....補足編 51

X

χ^2 分布 (エックスではなくカイ).....→ カイ自乗分布

ア

青虫.....248
 赤池情報量規準.....243
 アクシデント.....→ 事故
 悪魔の階段.....補足編 35
 圧縮.....311
 誤り訂正.....313
 あわてものの誤り.....246
 暗号論的擬似乱数列.....257
 安全な擬似乱数.....補足編 58

イ

イェンセンの不等式.....→ Jensen の不等式
 一次式.....→ 1 次式
 一様最強力不偏検定.....250, 補足編 57
 一様最小分散不偏推定量.....→ UMVUE
 一様分布 (実数値).....131, 134, 補足編 35
 一様分布 (多次元).....150
 一様分布 (離散値).....58, 73, 308, 補足編 30, 補足編 73
 一致性.....238, 補足編 55
 伊藤積分.....304
 イプシロン-デルタ論法.....補足編 31
 因果関係.....45, 51, 57, 184
 インク.....114, 136
 印刷.....114
 因子分析.....補足編 64

ウ

ウェブページ.....294
 宇宙人.....54

裏返し 121, 151, 201, 補足編 50

エ

絵書き歌 52
円グラフ 228
円周率 162, 255
エントロピー 306, 312, 補足編 72, 補足編 73, 補足編 75

オ

オイラーの公式 328
帯グラフ 228
オブジェ 77, 110, 157
折線グラフ 272

カ

カイ 2 乗分布 → カイ自乗分布
回帰分析 271
カイ自乗分布 212, 補足編 57
買占め 補足編 74
概収束 347, 348, 350, 補足編 28
階乗 75
階数 補足編 51
回転 198, 201, 219, 補足編 49, 補足編 50, 補足編 55
ガウス積分 162, 328, 補足編 23
ガウス分布 → 正規分布
カオス 304
学食 → カレー
拡張ピタゴラスの定理 355
確率 3, 17, 22, 補足編 8
確率化検定 補足編 57
確率過程 284, 補足編 7
確率行列 補足編 68
確率空間 4
確率収束 348, 349, 350, 352, 補足編 28, 補足編 55
確率測度 補足編 5
確率微分方程式 304
確率分布 16, 130, 補足編 36
確率変数 12
確率密度関数 127, 136, 190, 補足編 34, 補足編 37, 補足編 41
隠れマルコフモデル 303
影 194, 207, 218, 222, 291, 補足編 51
可算 294, 320, 補足編 5, 補足編 26
可算加法族 → σ 加法族
片側検定 246
偏り 100
可付番 → 可算
神様視点 9
カルバック・ライブラー情報量
..... → Kullback-Leibler divergence
カルマンフィルタ 243, 292, 294, 303, 補足編 67
カレー 184, 209
頑健 補足編 55
関数解析 補足編 48
完全加法族 → σ 加法族
カントールの悪魔の階段 補足編 35
ガンマ関数 → Γ の項目にある Γ 関数

キ

幾何平均 344
棄却 244
記述統計 227, 228, 272, 補足編 64
疑似乱数列 25, 106, 254, 補足編 58
基礎空間 12
期待値 77, 156, 186, 補足編 24
期待値ベクトル 186, 補足編 48
ギブスの不等式 → Gibbs の不等式
基本変形 補足編 52
帰無仮説 244
逆行列の補題 補足編 51
逆正弦法則 288
逆問題 51
共役事前分布 242
狭義に下に凸 339
共通部分 320
共分散 174
共分散行列 185, 補足編 48, 補足編 53
行列式 153, 補足編 41, 補足編 63
極限分布 298, 補足編 69, 補足編 70
虚数単位 320

切口 139, 142, 146, 205, 補足編 50, 補足編 63
金庫 57

ク

クーゲル 294
空集合 320
区間推定 234
くじ買占め問題 補足編 74
国 28, 77
組合せ 76
グラデーション 114
グラム・シュミットの直交化 補足編 49
車 → モンティホール問題

ケ

経験分布 231
計算可能性 補足編 62
計算尺 332
計算不可能 補足編 61
ケーキ 171
桁数 332, 337
結合エントロピー 308
結合確率 → 同時確率
県 → 国
検定 244, 補足編 57, 補足編 58

コ

高級車 → モンティホール問題
合計値の分布 166, 補足編 20, 補足編 42, 補足編 44
交差検証法 243
交差検定法 243
高次相関 222
格子点 258
工場 28
効用関数 85, 89
コーシー・シュワルツの不等式 → シュワルツの不等式
コーシーの不等式 → シュワルツの不等式
コーシー分布 351
国勢調査 227
黒板ボード体 185
誤差関数 330, 補足編 43
コヒーレンス 補足編 8
固有値 200, 203, 219, 278, 282, 補足編 49, 補足編 53, 補足編 54, 補足編 64, 補足編 68, 補足編 70
固有ベクトル 200, 219, 278, 補足編 49, 補足編 54, 補足編 68, 補足編 70
コルモゴロフ複雑性 補足編 60
コレスキー分解 265
根源事象 12

サ

最小記述長 243
最小自乗法 271, 補足編 63
最小値 319, 補足編 23
最大値 319, 補足編 22, 補足編 42
採択 245
最尤推定 238, 242, 273, 補足編 57
差集合 補足編 6
三角行列 補足編 63
三角不等式 334
サンクトペテルブルクのパラドックス 85
三重マルコフ連鎖 補足編 68
算術平均 344
散布図 183, 228
サンプルサイズ 234, 239, 242, 276, 352, 補足編 57

シ

シーソー 87
司会者 → モンティホール問題
時間的に一様 294
シグマ加法族 → σ 加法族
シグマ集合体 → σ 加法族
シグマ代数 → σ 加法族
時系列解析 285
次元の呪い 224
事故 70, 112
事後確率 52
事後分布 52, 240
事象 12, 19, 46, 59, 68, 補足編 16

自乗誤差	90, 100, 235, 276, 277, 283, 293, 355, 補足編 24, 補足編 53, 補足編 63, 補足編 64
指数関数	327
事前確率	52
自然数	319, 320
自然対数	331
自然対数の底	327
事前分布	52, 240, 277
下三角行列	補足編 63
下に凸	337
視聴率調査	227
実現値	231, 238, 253, 254, 277, 286
実数	319, 321
シナリオ	6, 9
四分位数	228
弱収束	349
シャノン	306
シュアの補元	補足編 51
シュークリーム	補足編 15, 補足編 19
集合族	補足編 6, 補足編 30
重心	87, 186
従属	59
住宅	28
自由度	212
周辺確率	28, 35, 42, 補足編 9
周辺分布	35, 42, 139, 補足編 9
主軸	208, 220, 278, 補足編 49
主成分	279
主成分分析	278, 補足編 64
主成分ベクトル	278
シュヴァルツの不等式	→ シュワルツの不等式
受容	245
シュワルツの不等式	182, 334, 343
瞬間移動装置	271
準正定	→ 非負定値
順問題	51
準モンテカルロ法	258
順列	75
条件収束	補足編 26
条件つきエントロピー	308
条件つき確率	19, 29, 39, 補足編 12
条件つき期待値	107, 補足編 30
条件つき分散	110
条件つき分布	39, 144
情報幾何	354
情報源符号化定理	311, 補足編 74
情報理論	305
常用対数	331
食堂	→ カレー
真の分布	231
真の乱数列	253
シンプソンのパラドックス	45

ス

推移確率	294
推移確率行列	295
遷移行列	295
推測統計	227, 272
推定	227, 補足編 55
推定量	234
酔歩	→ ランダムウォーク
すごろく	269
裾が重い	351, 補足編 30
スターリングの公式	→ Stirling の公式

セ

正規化	95, 101, 165, 168, 169, 179, 補足編 64
正規直交基底	333, 補足編 54, 補足編 63, 補足編 64
正規分布	161, 194, 補足編 42, 補足編 73
正規方程式	274
整数	319, 320
正則行列	202, 204, 219, 265, 補足編 41, 補足編 49, 補足編 54, 補足編 63
正定値	補足編 51, 補足編 54
正の相関	175
聖ペテルスブルクのパラドックス	85
世界	9
積雪	→ 雪
絶対収束	補足編 26
絶対値	121, 151, 319, 320, 補足編 55
遷移確率	294

遷移行列	295
漸近有効性	238, 補足編 55
全体集合	補足編 6
選択公理	補足編 5

ソ

相加平均	344
相関	175
相関係数	179, 310, 補足編 48
相互情報量	309, 314, 355
相乗平均	344
相対エントロピー	→ Kullback-Leibler divergence
測度	補足編 5
測度論	24
素数	319

タ

ダーツ	255
第一種の過誤	246
第二種の過誤	246
第一逆正弦法則	288
対角行列	198, 204, 216
対角成分	185
対角線論法	321
対称行列	185, 199, 補足編 53, 補足編 54
対数	331
大数の強法則	補足編 28
大数の弱法則	補足編 27
大数の法則	106, 170, 231, 256, 349, 351, 352, 補足編 27
対数尤度	239
大偏差原理	170, 354
対立仮説	244
楕円	197, 217, 219, 補足編 52
高々可算	71, 321, 補足編 40
宝箱	51, 59
多次元正規分布	194, 補足編 49
多次元標準正規分布	195
足し算の分布	→ 合計値の分布
多重マルコフ連鎖	補足編 68
畳み込み	補足編 21, 補足編 42
ダックタイピング	補足編 5
盾	55
種	254
田畑	28
多変量解析	271, 補足編 64
多変量正規分布	194
多目的最適化	236
単位円	196, 補足編 50
単位行列	196
単純仮説	247

チ

チェビシェフの不等式	→ Chebyshev の不等式
チェルノフ限界	341
チコノフの正則化	242, 277, 補足編 63
中央値	228, 233
中心極限定理	168, 234, 264, 352, 補足編 44, 補足編 46, 補足編 47
チューリングマシン	補足編 61, 補足編 62
超一様分布列	257
超関数	補足編 36
挑戦者	→ モンティホール問題
重複固有値	補足編 49
調和平均	345
直交行列	199, 201

ツ

通信レート	314
通信路の容量	314
通信路符号化定理	314, 補足編 76

テ

定義	14, 61, 335
低くい違い列	257
定常分布	297, 補足編 68
定常無記憶通信路	314
てこの原理	87
デルタ関数	補足編 35, 補足編 44

転移確率	294
点推定	234
伝送レート	314
転置	185, 185
ト	
等位面	195, 205, 345, 補足編 50
等高線	195, 215, 218, 343, 345, 補足編 50, 補足編 52
同時エントロピー	308
同時確率	28, 35, 42, 補足編 9
同時分布	35, 42, 136, 155, 190, 補足編 38
動的計画法	303
等比級数	72, 112, 130, 325, 補足編 27, 補足編 30
特異値	補足編 64
特異値分解	補足編 49, 補足編 64
特性関数	170, 350, 補足編 21, 補足編 44
特性方程式	221, 280, 補足編 49
独立	32, 59, 63, 146, 191, 補足編 16, 補足編 39
独立成分分析	58, 170
独立同一分布	103
土地利用	27
扉	→ モンティホール問題
ともえ戦	250
トレース	補足編 53
ナ	
内積	333
内包的記法	320
ニ	
二項係数	→ 2 項係数
二項分布	→ 2 項分布
二次式	→ 2 次式
二次平均収束	→ 2 次平均収束
ネ	
ネイピア数	327
ネイマン・ピアソンの補題	249, 補足編 57
ノ	
濃度	321
ノンパラメトリック	234
ノンパラメトリックベイズ法	243
ハ	
パイアス	100
バスケットボール	167
外れ値	228, 補足編 55
バターン	315
鳩の巣原理	311
バナッハ・タルスキのパラドックス	補足編 5
ハフマン木	→ ハフマン法
ハフマン符号	→ ハフマン法
ハフマン法	補足編 58, 補足編 76
ばらつき	100
パラメトリック	234
パラレルワールド	9
汎関数	308, 補足編 24
半正值	→ 非負定値
半正定値	→ 非負定値
ヒ	
ピアソンの補題	249
ビール	71
飛行船視点	8
非対角成分	185
びっくり度	305
非負行列	補足編 53
非負定値	265, 補足編 53
微分幾何	354
標準正規分布	161, 195
標準偏差	92, 160
標本	12, 234
標本空間	12
標本点	12
標本分散	237

フ

ファイル圧縮	311
フィボナッチ数列	254
ブートストラップ法	243
フーリエ変換	285, 350, 補足編 21
複合仮説	249
複素数	320
符号化レート	314
不動点	補足編 69
負の相関	175
不偏分散	101, 237, 240, 補足編 55
不変分布	→ 定常分布
ブラウン運動	304
フリースロー	167
プリンタ	114
フルランク	補足編 51
不連続修正	→ 連続修正
ブロック対角行列	補足編 50
分散	90, 160, 228, 補足編 27
分散共分散行列	→ 共分散行列
分散行列	→ 共分散行列
分布	→ 確率分布
分布関数	→ 累積分布関数
分布収束	→ 法則収束

ヘ

平均	104, 228, 補足編 24
平均自乗誤差	補足編 24
平均収束	→ 2 次平均収束
平衡分布	→ 定常分布
ベイジアンネットワーク	243
ベイジアンフィルタ	243
ベイズ情報量規準	243
ベイズ推定	→ Bayes 推定
ベイズ統計	補足編 55
ベイズの公式	→ Bayes の公式
別名法	補足編 58
ペテルスブルクのパラドックス	85
ヘルダーの不等式	343
偏差値	95
ベン図	補足編 10, 補足編 11, 補足編 12
偏微分	154

ホ

法則収束	168, 349, 351
ポートフォリオ	111
ホールケーキ	171
補集合	補足編 6
ほとんど至るところ	補足編 35
ほとんど確実に	補足編 35
ほとんど全ての	補足編 35
ぼんやりものの誤り	246

マ

マッドサイエンティスト	271
魔法	51, 59
マルコフ過程	294, 補足編 68
マルコフの不等式	→ Markov の不等式
マルコフ連鎖	294, 補足編 67
マルコフ連鎖モンテカルロ法	243, 補足編 58
マルチンゲール	289

ミ

三つ組 (Ω, \mathcal{F}, P)	9
三つの扉	→ モンティホール問題
密度	116
ミンコフスキーの不等式	342

ム

無記憶通信路	314
無相関	175
無理数	12

メ

メソッドチェイン	補足編 68
メルセンヌツイスター	255

モ

文字列圧縮問題	311
モデル選択	243
モンスター	51, 89
モンティホール問題	4, 21, 25, 38, 48, 補足編 7
モンテカルロ法	25, 256, 258, 補足編 58

ヤ

ヤギ	→ モンティホール問題
ヤコビアン	154

ユ

有意	246
有意水準	244
尤度	238
尤度比	249
尤度比検定	250, 補足編 57
有理数	319, 321
雪	77, 340

ヨ

容量	314
----	-----

ラ

ラグランジュ未定係数法	344, 補足編 73
ランク	補足編 51
乱数の検定	補足編 58
乱数列	253, 補足編 60
ランダムウォーク	286, 補足編 65
ランダム検定	補足編 57

リ

リーマン積分	114
リーマン面	326
両側検定	246
リンク	294

ル

累積寄与率	282
累積分布関数	127, 261, 補足編 32, 補足編 33, 補足編 36, 補足編 37, 補足編 39
ルベグ積分	24, 114

レ

例外	補足編 6
レシビ	200, 219
連続修正	補足編 47
連続補正	→ 連続修正

ロ

ロールケーキ	171
ロールプレイングゲーム	51, 55, 89
ロバスト	補足編 55

ワ

わさび	補足編 15, 補足編 19
和集合	320
忘れもの	184
罫の気配	51, 59
和の分布	→ 合計値の分布