情報理論

Information Theory

今井　浩

2010 年 4 月 13 日

目次

1　情報量とエントロピー

情報とは、私たちに何かを伝え、不確実だった知識をより確実にしてくれるものであると考えられる。つまり

「情報の量」は、その情報を得た事によって知識の不確実性がどのくらい減ったかという量で定式化する。

今、サイコロを振って、何の目が出たかを知りたいという事を考えよう。$A_1, A_2, \cdots, A_6$ をそれぞれ「1 の目が出た」、「2 の目が出た」,・・・「6 の目が出た」という事象を表すとする。また、B_1, B_2 をそれぞれ「偶数の目が出た」、「奇数の目が出た」という事象を表すとする。

この時、「偶数または奇数が出た」という情報よりも、「ある特定の目（例えば 1 の目）が出た」という情報の方が「情報を得る前の不確実性は大きく減った」事になる。つまり、情報の量は、可能な事象の数（$m = 2$、$n = 6$）と密接に関係している。

言い換えれば、起こりえる確率が低い事象がおこったという情報の方が多くの情報を持っている事を表しており、これは一般的な感覚とも一致する。ありふれたできごとが起こったことを知ってもそれはたいした「情報」にはならないが、逆に珍しいできごとが起これば、それはより多くの「情報」を含んでいると考えられる。

1.1　情報量の定式化

■情報の加法性という特徴を使った定式化　この情報の量を $f(x)$ と書くことにし、この関数 $f(x)$ を定めることを考える。そのために情報を小出しにした場合を想定し、図 1 ように n 個の事象 $A_1, A_2, \cdots, An$ が k 個ずつの m 組に組み分けされていたとする。

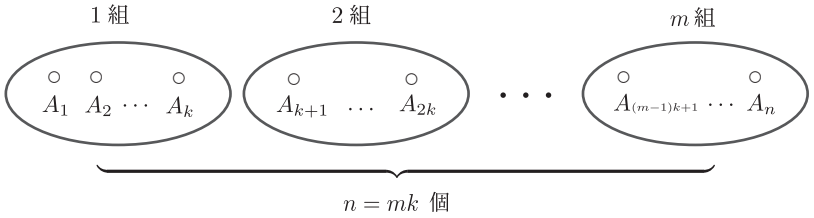


図 1　小出しにした情報の量

この時、先のサイコロの場合のように、先に偶数か奇数かをを知り、次にその組の中の 3 つのうちのどの目が出たかを知らせるというように、情報を小出しにする場合を想定する。

まず「サイコロのどの目が出た」というように、どの A_i が起こったかを、ズバリ直接答えてくれる情報の量を考えよう。情報の量は事象の数 n と関係するので、それを $f(n)$ と表そう。また組み分けした場合に、m 個のどの組に入っているかを教えてくれる情報を $f(m)$ とし、さらにその組の中の k 個のどれであるかを教える情報の量を $f(k)$ とする。この時、情報をズバリと直接出そうと、ちょっとずつ小出しにしていったとしても、最終的には得られる情報の量は変わらないと考えられる。つまり

$$f(n) = f(m) + f(k)$$

ここで $n = mk$ なので

$$f(mk) = f(m) + f(k) \tag{1.1}$$

この式は、小出しにした情報の量を加え合わせれば、全体の情報量になるという「情報の加法性」を表す。この式 (1.1) を満たす関数は、適当な連続性を仮定すれば式 (1.2) のように一意に表す事ができる。この式の導出は 18 ページの式 (1.32) に記載している。

$$f(x) = a \log_e x \tag{1.2}$$

■ビットという情報量を定式化する　次にこの式の a と対数の底 e を求めよう。情報の最も基本的なものは 2 つの事項のうちの一つを教えてくれるものであると考えられる。そこで、yes/no であるとか、右か左か、偶数か奇数かという二者択一の情報を情報量の単位とすることにする。つまり

$$f(2) = 1$$

とするという事である。先の式 (1.2) に $x = 2$ の場合を当てはめれば

$$\begin{aligned} a \log_e 2 &= 1 \\ \log_e 2 &= \frac{1}{a} \\ e^{\frac{1}{a}} &= 2 \end{aligned}$$

両辺を a 乗すると

$$e = 2^a$$

この両辺の対数 $\log_2$ をとって

$$a = \log_2 e$$

これで a を求める事ができた。この a を式 (1.2) に当てはめるのだが、その前に対数関数の底の変換公式より

$$\log_e x = \frac{\log_2 x}{\log_2 e}$$

なので $f(x) = a \log_e x$ は、

$$f(x) = \log_2 e \cdot \frac{\log_2 x}{\log_2 e}$$

つまり、

$$f(x) = \log_2 x \tag{1.3}$$

この式 (1.3) が 1 ビット (bit:binary digit) の情報量の定義となる。この変数 x はひとつひとつ事象の数を意味し、この式は x 個の事象の中から一つが起こった事を知らせる情報の情報量を意味している。例えば、今 8 個の事象（例えば 8 面体のサイコロ）があったとする。この場合の事象の数は $2^3 = 8$ であり、$f(8) = 3$ となる。

この 3 という数は、一つの事象を特定するまでに必要な回数である。つまり、全体を半分さらに半分というように二分割していって最後に一つの事象に特定するまでの回数を意味している。また、$x = 8$ 個の事象から一つを特定するには、3 ビットの情報が必要であるという言い方も可能である。

■確率を変数として情報量を定義する　次に確率を変数 p とした時の情報量 $f(p)$ を考える。いま、図 2 のように、n 個の等確率な事項を考え、その中の k 個をひとまとめにしたのが事象 A であるとする。

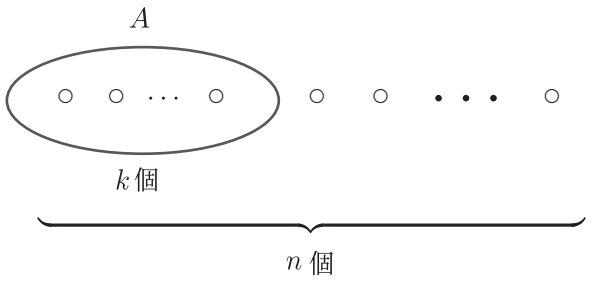


図 2　確率による情報量の定義

この事象 A の起こる確率は

$$p = \frac{k}{n}$$

である。この時に、この A が起こったということを教えてくれる情報があったとして、この情報の量はどのぐらいかを定式化しよう。まず、A が起こった事を知るための情報の量を I とする。その時

- n の内どれかが起こったかを知るための情報量は $\log n$
- A が起こった時に、さらにその中の k 個のうちのどれが起こったかを知るための情報量は $\log k$

ここから、

$$\log n = I + \log k$$

$$-(\log k - \log n) = I$$

対数の差は割り算になるので

$$I = -\log \frac{k}{n}$$

定義 1.1. 確率 p の事象を観測した時に得られる情報量 (I) は

$$I = -\log_2 p \qquad (bit) \tag{1.4}$$

上の定義に従うと、n 個の等確率な事象の中から 1 つが実際に起きたときの情報量 (I) は

$$I = -\log_2 \frac{1}{n} = -(\log_2 1 - \log_2 n) = \log_2 n$$

となり、情報量の定義式（1.3）と一緒になる。

1.2　エントロピーの定義とその性質

■エントロピーは状況の不確実性を意味する　今までは「ある事象が起こった」という情報の情報量を考えた。今度は、事象が起こる前に得られるであろう情報量の期待値について計算してみる。

いま、$A_1, A_2, \cdots, A_n$ の n 個の事象があって、それぞれ $p_1, p_2, \cdots, p_n$ の確率で生じる場合を考えよう。また当然ながら

$$\sum_{i=1}^{n} p_i = 1$$

である。この時に得られるであろう情報量の期待値は A_1 から A_n までのそれぞれの情報量 $-\log p_1$ から $-\log p_n$ の平均で得られる。[*1]

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

これをエントロピーという。

定義 1.2. n 個の事象がそれぞれ確率 $p_1, p_2, \cdots, p_n$ で発生するとき、得られる情報の期待値をエントロピーと呼び、以下の式で表す。

$$H(p_1, p_2, \cdots, p_n) = -\sum_{i=1}^{n} p_i \log p_i \tag{1.5}$$

エントロピーは不確定な状況を確定するのに必要な平均情報量であるが、見方を変えると状況の特徴（つまり状況の不確定度合）を表す量と考えられる。

このエントロピーには以下のような性質がある

性質 1.1. エントロピー H は非負

$$H \geqq 0 \tag{1.6}$$

であり $H = 0$ が成立するのは、どれかひとつの p_i が 1 で、他がすべて 0 の時に限る。

この事を確認していこう。まず確率の定義から

$$0 \leqq p_i \leqq 1$$

となり、図 3 のように対数関数 $\log p_i$ はゼロ以下の負の数になる。なので、$-p_i \log p_i \geqq 0$ となり、

$$H \geqq 0$$

[*1] 確率の合計が 1 なので、平均は確率と値とをかけ合わせて合計した値となる

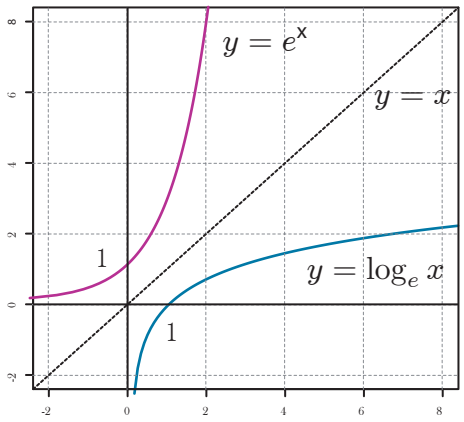


図 3　指数関数と対数関数のグラフ

また、$H = 0$ となるのは $H \geqq 0$ なので、すべての i について

$$-p_i \log p_1 = 0$$

となる場合であり、$p_i = 0$ または $\log p_1 = 0$ のどちらかでなければならない。$\log p_i = 0$ になるためには $p_i = 1$ でなければならないので、結局「すべての i について $p_i = 0$ または $p_i = 1$ のどちらかでなければならない」事になる。さらに確率の合計が 1（$\sum p_i = 1$）であることを考慮するなら、一つの p_i が 1 で、残りは 0 となる以外には存在しない。

エントロピーがゼロというのは、このように一つの事象が確実に起こり、その他の事象は起こらない事を意味している。つまり不確実性がない事を意味している。

性質 1.2. n 個の事象の表すエントロピーの最大値 $H(n)$ は

$$H(n) = \log n \tag{1.7}$$

であり、すべての確率が等しい時、つまり

$$p_i = \frac{1}{n}$$

の時に最大になる。

ラグランジュの未定乗数法を用いてこの性質を確認する。ラグランジュの未定乗数法については 1.4 節（21 ページ）を参照。まず、エントロピーは 1 から n 個のそれぞれの事象の確率を p_i とすると、以下の H で計算できる。

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

また確率の合計は 1 なので以下の条件がつく、つまりエントロピーを最大化する問題とは、以下の制約条件の

もとで関数 H を最大化する問題となる。

$$\sum_{i=1}^{n} p_i = 1$$

ラグランジュの未定乗数を λ とした時に、以下の L をそれぞれの p_i と λ について偏微分して 0 とおいた時に極値が得られる。

$$L = \left(-\sum_{i=1}^{n} p_i \log_2 p_i \right) - \lambda \left(\sum_{i=1}^{n} p_i - 1 \right) \tag{1.8}$$

この L を展開すると

$$L = -\left(p_1 \log_2 p_1 + p_2 \log_2 p_2 + \cdots + p_n \log_2 p_n\right) - \lambda\{(p_1 + p_2 + \cdots + p_n) - 1\}$$

これをある特定の p_i について偏微分すると

$$\frac{\partial L}{\partial p_i} = -(p_i \log_2 p_i)' - \lambda$$

この時、積の微分公式

$$\{f(x)g(x)\}' = f'(x)g(x) + f(x)g'(x)$$

と対数関数の微分の公式

$$(\log_a x)' = \frac{1}{x} \log_a e$$

を活用すると

$$\begin{aligned}
\frac{\partial L}{\partial p_i} &= -(p_i \log_2 p_i)' - \lambda \\
&= -\{p_i' \log_2 p_i + p_i (\log_2 p_i)'\} - \lambda \\
&= -\log_2 p_i - p_i \left(\frac{1}{p_i} \log_2 e\right) - \lambda \\
&= -\log_2 p_i - \log_2 e - \lambda
\end{aligned}$$

この偏微分を 0 とおくと以下のように変形できる。[*2]

$$\begin{aligned}
\log_2 p_i &= -\lambda - \log_2 e \\
\log_2 p_i &= \log_2 2^{-\lambda} - log_2 e \\
\log_2 p_i &= log_2 \frac{2^{-\lambda}}{e} \\
p_i &= \frac{2^{-\lambda}}{e}
\end{aligned}$$

ここからわかるのは、エントロピー H が最大をとる場合、各事象の確率 p_i は、i によらず全てが同じ値 $(\frac{1}{e}\, 2^{-\lambda})$ であるという事である。また確率の総和 $\sum p_i = 1$ なので、もし事象が n 個ならば、すべてが同じ値ならば、それぞれの確率は $\frac{1}{n}$ でなければならない。

[*2] 変形の途中で対数関数の差を商に変換する以下の公式を使っている。

$$\log \frac{M}{N} = \log M - \log N$$

エントロピーが一番大きい（つまり状況の不確定度が一番大きい）のは、すべての事象が等確率で起こり、どれが起こりそうかがまったく予想できない場合である。

2 つの事象がそれぞれ確率 p、q（$p+q=1$）で起こるとした場合のエントロピー $H(p,1-p)$ のグラフは図 4 ようになる。$p=q=1/2$ の時がエントロピー H が最大であり、その値は以下のように 1 ビットになる[*3]

$$H = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

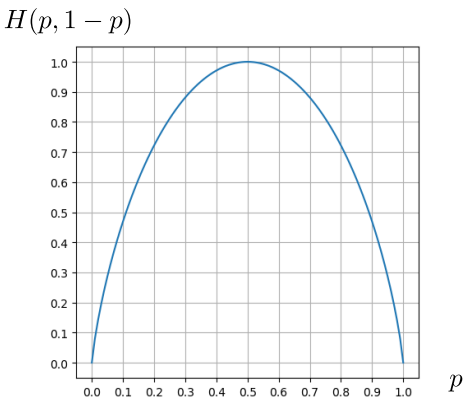


図 4　$n=2$ の場合のエントロピー

ソースコード 1　n＝2 の場合のエントロピーを描くプログラム

```
import numpy as np
import matplotlib.pyplot as plt

x = np.arange(0, 1.01, 0.01)
y = −x∗np.log2(x) − (1−x)∗np.log2(1−x)
y[0]=0 ; y[100]=0

# figure インスタンスを生成し、figure インスタンスの axes を生成
fig = plt.figure()
ax = plt.axes()
ax.grid(True)
ax.xaxis.set_ticks(np.arange(0, 1.1, 0.1))
ax.yaxis.set_ticks(np.arange(0, 1.1, 0.1))
ax.plot(x, y)
plt.show()
```

[*3] この計算は、$\frac{1}{2} = 2^{-1}$ より $\log_2\frac{1}{2} = -1$ なので H は 1 となる。

定義 1.3. いまある情報を得る事によって、状況のエントロピーが H から $H^{'}$ へ変わったとする。この時この情報の持つ情報量 I は以下になる。

$$I = H - H^{'} \tag{1.9}$$

事例で説明しよう。ある地方の年間の平均の天候が表 1 のようになっているとする。

表 1　天気の年間平均確率

晴れ	曇り	雨
40 %	40 %	20 %

この時、明日の天気について以下のように予報が出た場合、この天気予報の情報量を求める事を考える。

表 2　明日の天気予報

晴れ	曇り	雨
80 %	15 %	5 %

予報を知る前のエントロピーは[*4]

$$H = -0.4\log_2 0.4 - 0.4\log_2 0.4 - 0.2\log_2 0.2 = 1.52$$

予報を知った後のエントロピーは

$$H' = -0.8\log_2 0.8 - 0.15\log_2 0.15 - 0.05\log_2 0.05 = 0.88$$

したがって、この天気予報の情報量は

$$I = H - H' = 0.64(bit)$$

このように状況の不確実度合をエントロピーとして定量的にとらえることで、さまざまな情報の性質を調べる事が可能になる。

[*4] 3つともに等確率の場合のエントロピーは 1.58 となる。平均を知っていてもあまり情報量は得られていない事がわかる。

■複合事象の確率の性質

２つの事象の組み合わせについての確率を考える。例えば以下のように病院でまず熱を測り、その後に詳細な診断をするような事例を考えると、以下の２つの事象によって４通りの複合事象 (A_i, B_j) が出来上がる。

$$
\begin{aligned}
A \text{ 事象} \quad &: (A_1, \text{熱がある} \quad , A_2 \text{熱がない}) \\
B \text{ 事象} \quad &: (B_1, \text{風邪ひいている}, B_2 \text{風邪ひいてない})
\end{aligned}
$$

これら４つの複合事象の確率について調べたところ、以下の表 3 のような値になっているとする。

表 3　病院での診断確率

	B_1 風邪ひいている	B_2 風邪ひいてない	$p(A_i)$
A_1 熱がある	0.55	0.05	0.60
A_2 熱がない	0.10	0.30	0.40
$p(B_j)$	0.65	0.35	

この時、この表の各セルの確率を以下のように呼ぶ。なお、以下の条件付き確率 $p_{A_i}(B_j)$ は、A_i が起こったという条件で B_j が起こる確率の事で、$p(B_j|A_i)$ とも書く場合もある。

$$
\begin{aligned}
\text{同時確率} \quad &: p(A_i, B_j) \\
\text{周辺確率} \quad &: p(A_i), \quad p(B_j) \\
\text{条件付確率} \quad &: p(B_j|A_i), \quad p(A_i|B_j)
\end{aligned}
$$

まず最初に、これらの確率の関係式についてまとめておく。

性質 1.3. 【確率の合計が 1 になるという性質】

$$\sum_{i,j} p(A_i, B_j) = 1 \tag{1.10}$$

$$\sum_{i} p(A_i) = \sum_{j} p(B_j) = 1 \tag{1.11}$$

$$\sum_{j} p(B_j|A_i) = \sum_{i} (A_i|B_j) = 1 \tag{1.12}$$

この三番目の式 $\sum_j p(B_j|A_i) = 1$ はちょっと注意が必要。表 3 をみると、例えば A_1 が起きて B_1 が起きる確率と B_2 が起きる確率の和は、$p(B_1|A_1) + p(B_2|A_1) = 0.55 + 0.05 = 0.60$ となりそうだが、実は図 5 のように A_1 が起きたという前提なので $p(A_1) = 0.60$ を分母として $p(B_1|A_1) = \dfrac{0.55}{0.60}$ であり $p(B_2|A_1) = \dfrac{0.05}{0.60}$ と計算する。条件付き確率の式 5.1 を参照。

	B_1 風邪ひいている	B_2 風邪ひいてない	$p(A_i)$
A_1 熱がある	0.55	0.05	0.60
A_2 熱がない	0.10	0.30	0.40
$p(B_j)$	0.65	0.35	

$$p(A_1, B_1) = p(A_1)\ \ p_{A_1}(B_1)$$

図 5　条件付確率の和

性質 1.4. 【同時確率と周辺確率】

$$p(A_i) = \sum_j p(A_i.B_j) \tag{1.13}$$

$$p(B_j) = \sum_i p(A_i, B_j) \tag{1.14}$$

これは縦と横の周辺確率が、その縦と横の同時確率の和になっているだけで集計表そのもの。

性質 1.5. 【同時確率と条件付確率】

$$p(A_i, B_j) = p(A_i)\ p(B_j|A_i) \tag{1.15}$$

$$p(A_i, B_j) = p(B_j)\ p(A_i|B_j) \tag{1.16}$$

これは以下の図 6 のような同時確率と条件付確率の関係をしめしている。例えば、

$$\begin{aligned} p(A_1, B_2) &= p(A_1)\ p(B_2|A_1) \\ &= 0.60 \times \frac{0.05}{0.60} = 0.05 \end{aligned}$$

	B_1 風邪ひいている	B_2 風邪ひいてない	$p(A_i)$
A_1 熱がある	0.55	0.05	0.60
A_2 熱がない	0.10	0.30	0.40
$p(B_j)$	0.65	0.35	

$$p(A_1, B_2) = p(A_1)\ p(B_2|A_1)$$

図 6　条件付確率と同時確率の関係

以下のベイズの公式については節 5（51 ページ）を参照。

性質 1.6. 【ベイズの公式】

$$p(B|A) = \frac{p(A|B)\,p(B)}{p(A)} = \frac{p(A,B)}{p(A)} \tag{1.17}$$

$$p(A|B) = \frac{p(B|A)\,p(A)}{p(B)} = \frac{p(A,B)}{p(B)} \tag{1.18}$$

もし事象 A と事象 B とが独立、つまり A の何が起こったかは B の何が起こるかとは関係ない（その逆も）という場合なら

性質 1.7. 【2つの事象が独立している場合】

$$p(A,B) = p(A)\,p(B) \tag{1.19}$$

$$p(B|A) = p(B) \tag{1.20}$$

$$p(A|B) = p(A) \tag{1.21}$$

■条件付きエントロピー

A が何であるかを知った時の、B が何であるかについての不確定度を調べよう。いま A が A_1 である事がわかったとする。この時に B の中の $B_1, B_2, \cdots, B_m$ が起こる確率は、それぞれの条件付き確率を求めればよいので

$$p(B_1|A_1), \quad p(B_2|A_1), \quad \cdots, p(B_m|A_1)$$

A が A_1 である事がわかった時のエントロピーは式（1.5）[*5]より、確率と情報量の積和、つまり情報量の期待値を求めればよいので以下のようになる。

$$H(B \mid A_1) = -\sum_{j=1}^{m} p(A_1, B_j) \log p(A_1 \mid B_j)$$

A が A_1 である事がわかった時のエントロピーをすべての A_i について平均したものが、**条件付きエントロピー**と呼ばれるものである。

定義 1.4. A が何であるかを知った後の B についての不確定度を表す条件付きエントロピーは以下で定義される。

$$\begin{aligned} H(B \mid A) &= \sum_{i=1}^{n} p(A_i)\ H(B \mid A_i) \\ &= -\sum_{i=1}^{n} \sum_{j=1}^{m} p(A_i)\ p(B_j \mid A_i)\ \log\ p(B_j \mid A_i) \end{aligned} \tag{1.22}$$

先ほどの事例で確認してみよう。先ほどの事例は次の表。

	B_1 風邪ひいている	B_2 風邪ひいてない	$p(A_i)$
A_1 熱がある	0.55	0.05	0.60
A_2 熱がない	0.10	0.30	0.40
$p(B_j)$	0.65	0.35	

熱を測る前の B についてのエントロピー $H(B)$ は

$$H(B) = -(0.65 \times \log 0.65 + 0.35 \times \log 0.35) = 0.93$$

次に、熱を測った後の B についての条件付きエントロピーを求める。

[*5] エントロピーの定義式を再掲すると以下。

$$H(p_1, p_2, \cdots, p_n) = -\sum_{i=1}^{n} p_1 \log p_1$$

まず、各状態のエントロピーが

$$p(B_1 \mid A_1) = \frac{0.55}{0.6} = 0.92 \qquad p(B_2 \mid A_1) = \frac{0.05}{0.6} = 0.08$$
$$p(B_1 \mid A_2) = \frac{0.10}{0.4} = 0.25 \qquad p(B_2 \mid A_2) = \frac{0.30}{0.4} = 0.75$$

なので

$$H(B \mid A_1) = -\left(0.92 \times \log 0.92 + 0.08 \times \log 0.08\right) = 0.41$$
$$H(B \mid A_2) = -\left(0.25 \times \log 0.25 + 0.75 \times \log 0.75\right) = 0.81$$

条件付きエントロピーは式（1.22）のように各 $p(A_i)$ の期待値 $p(A_i)H(B \mid A_i)$ の合計なので、

$$\begin{aligned} H(B \mid A) &= p(A_1)\, H(B \mid A_1) + p(A_2)\, H(B \mid A_2) \\ &= 0.60 \times 0.41 + 0.4 \times 0.81 = 0.57 \end{aligned}$$

検温を済ませる前の B についてのエントロピー $H(B)$ は、0.93 に対して、検温を済ませた事で不確定度が 0.57 になったという事を意味している[*6]。また以下のように、検温によって得られる情報量の平均値 I は、エントロピーの減少分と同じである。

$$I = H(B) - H_A(B) = 0.36 \qquad (bit)$$

[*6] ※注意　検温したらエントロピーが必ず減るというわけではなく、検温後の結果によっては検温前より増える可能性はある。つまり i によって異なるが、その平均値をとったものが $H_A(B)$ となる。

■同時エントロピーと条件付きエントロピー

同時エントロピー（joint entropy）は、2 つ以上の事象が同時にどうなるか、ということについての不確かさ（エントロピー）を測る量で、以下のように定義される。

性質 1.8. 【同時エントロピー】

$$H(A,B) = -\sum_{i,j} p(A_i, B_j) \log p(A_i, B_j) \tag{1.23}$$

先の事象 $A(A_1$熱がある　$, A_2$熱がない$)$ と事象 B $(B_1$風邪ひいている$, B_2$風邪ひいてない$)$ の例ならば、それぞれは 2 通りだが、組み合わせ（A, B）としては全部で 4 通りあり、同時エントロピーは、この 4 通りのパターンがどれくらい予測しにくいかを測る量である。次に条件付きエントロピーとの関係を整理しておく。

性質 1.9. 【複合事象のエントロピー】

$$\begin{aligned} H(A,B) &= H(A) + H(B \mid A) \\ &= H(B) + H(A \mid B) \qquad (1.24) \\ H(B \mid A) &= H(A,B) - H(A) \qquad (1.25) \end{aligned}$$

式（1.24）の $H(A,B)$ は同時エントロピーである。これら式はその意味を考えると当然と思われる式である。例えば式 (1.24) を言葉で表現するなら以下のような意味になる。

複合事象（A, B）の同時エントロピー（不確定度）は、事象 A のエントロピー（不確定度）と事象 A が決定した後の事象 B のエントロピー（不確定度）との和である。

このように表現すれば当然であると思われるが、あえてこの式を変形しながら確認していこう。まず同時エントロピーの定義から

$$H(A,B) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\Big(p(A_i, B_j) \log p(A_i, B_j) \Big)$$

ここで、$p(A_i, B_j) = p(A_i)\, p(B_j \mid A_i)$ なので

$$H(A,B) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\Big(p(A_i, B_j) \log \big[p(A_i)\, p(B_j \mid A_i)\big] \Big)$$

対数の法則から $\log ab = \log a + \log b$ なので $\log\big[p(A_i)\, p(B_j \mid A_i)\big]$ を分離し、シグマ記号を中に入れて、

$$\begin{aligned} H(A,B) &= -\sum_{i=1}^{n}\sum_{j=1}^{m}\Big(p(A_i, B_j) \log p(A_i) + p(A_i, B_j) \log p(B_j \mid A_i) \Big) \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(A_i, B_j) \log p(A_i) - \sum_{i=1}^{n}\sum_{j=1}^{m} p(A_i, B_j) \log p(B_j \mid A_i) \end{aligned}$$

ここから、右辺の第一項とと第二項を分けて、それぞれ $H(A)$ と $H(B \mid A)$ になる事をしめす。まず、右辺の第一項に $p(A_i, B_j) = p(B_j \mid A_i)p(A_i)$ を代入して j に関係のない要素でシグマをくくりだすと第一項は

$$\begin{aligned} -\sum_{i=1}^{n}\sum_{j=1}^{m} p(A_i, B_j)\log p(A_i) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(B_j \mid A_i)p(A_i)\log p(A_i) \\ &= -\sum_{i=1}^{n} p(A_i)\log p(A_i)\sum_{j=1}^{m} p(B_j \mid A_i) \end{aligned}$$

ここで式（1.12）で示したように $\sum_j p(B_j \mid A_i) = 1$ なので、第一項は以下のようになり、事象 A のエントロピー $H(A)$ に他ならない。

$$-\sum_{i=1}^{n} p(A_i)\log p(A_i) \quad = H(A)$$

ついで、第二項に $p(A_i, B_j) = p(B_j \mid A_i)p(A_i)$ を代入してやって変形すると、式（1.22）の条件付きエントロピーの式[*7]に他ならない。

$$-\sum_{i=1}^{n}\sum_{j=1}^{m} p(A_i, B_j)\log p(B_j \mid A_i) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(A_i)\, p(B_j \mid A_i)\log p(B_j \mid A_i) = H(B \mid A)$$

以上より

$$H(A, B) = H(A) + H(B \mid A)$$

同様に、３つの事象 $A, B, , C$ について以下が成立する。$H_{AB}(C)$ は A と B が決定した後の C についての不確定度を表すエントロピーである。

性質 1.10. 【３つの複合事象のエントロピー】

$$H(A, B, C) = H(A) + H(B \mid A) + H(C \mid AB) \tag{1.26}$$

さらに、以下の性質がなりたつ。

性質 1.11.

$$H(B \mid A) \geqq 0 \tag{1.27}$$

$H(B \mid A)$ はエントロピー $H(B \mid A_i)$ の期待値（平均）であり、エントロピーは非負であるから上記が成立する。なお

$$H(B \mid A) = 0$$

が成立するのは、すべての i について $H(B \mid A_i) = 0$ となる場合[*8]で、どの A_i が起きても、A_i が決定すれば

[*7] 条件付きエントロピーの式（1.22）を再掲載すると、

$$H_A(B) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(A_i)\, p_{A_i}(B_j)\ \log\ p_{Ai}(B_j)$$

[*8] $p(A_i)$ はすべてが 0 でないとする。

完全に B_j が定まる場合である。つまり事象系 B が事象系 A に完全に従属している場合である。

性質 1.12. 【種々のエントロピーの相互関係】

$$H(A)+H(B) \geqq H(A,B) \tag{1.28}$$
$$H(A) \geqq H(A \mid B)$$
$$H(B) \geqq H(B \mid A) \tag{1.29}$$
$$H(A,B) = H(A)+H(B \mid A)$$
$$= H(B)+H(A \mid B) \tag{1.30}$$

本来式（1.28）を数式で証明すればそこから各種関係が導かれるのだが、ここでは図 7 のようにベン図によるイメージで把握しておく。上記の式の不等号がついた式の等号が成立するのは事象 A と事象 B が独立の時である。

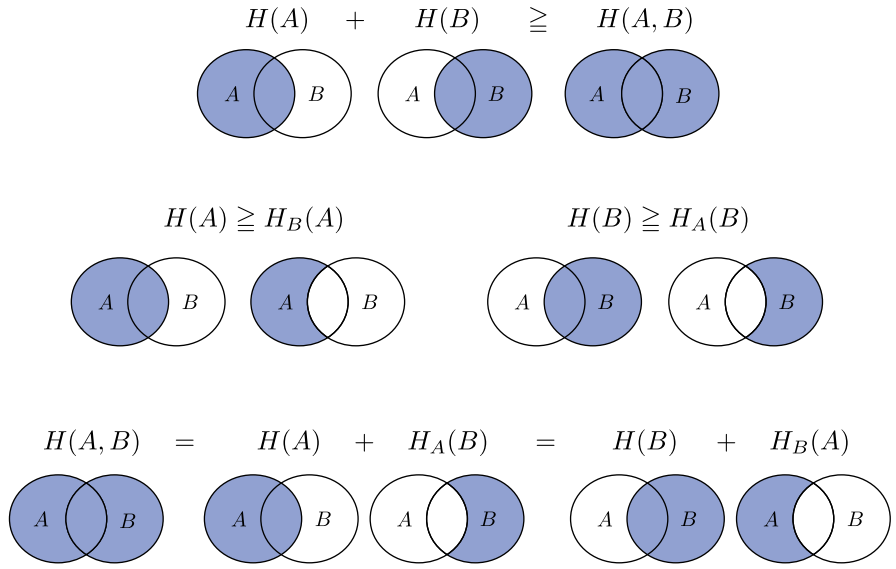


図 7　種々のエントロピーの相互関係

図 7 の最初の関係式 1.28 について補足しておく。この $H(A)+H(B) \geqq H(A,B)$ を言葉で表現しておくと、「事象 A と B に関するそれぞれのエントロピーを足した量は、A と B を合わせたときの同時エントロピーよりも大きくなるか、または等しくなる。」という表現になる。これは、**エントロピーの劣加法性（subadditivity）**と呼ばれる。

$H(A)+H(B)$ は、何も知らないで両方バラバラに予測する場合の不確かさで、$H(A,B)$ は、両方を一緒に観測するときの不確かさである。つまり「一緒に扱うほうが不確かさが減る」＝冗長性（重なり）が省かれるということになる。

なぜこのようなことが起こるのか？

- 2 つの事象が独立の場合
  事象 A を「サイコロ（1〜6）を振った時の目」、事象 B を「コイン（表 or 裏）」とすると、それぞれのエント

ロピーは $H(A) = \log_2 6$ であり、$H(B) = \log_2 2$ であり、同時エントロピーは $H(A,B) = H(A)+H(B)$ となる。

- 2 つの事象に関連が強い場合
  事象 A を「気温 (高 or 低)」で事象 B を「アイスを買ったかどうか (買う or 買わない)」とする。この場合は、A =「高」なら、B =「買う」、A =「低」なら、B =「買わない」という傾向があり、もし A が判るれば B は必ず判るとすると B に関する不確かさは 0 となる。つまり、$H(B \mid A) = 0 \Rightarrow H(A,B) = H(A)$。ただし、このように関連が強い場合でも別々に見てしまうと $H(A) + H(B) > H(A)$ となる。これは図の交わった部分を 2 回数えてることに起因する。つまり、「同時に扱ったほうが、必要な情報量は減る」ことになる。

また式 (1.29) は、シャノンの基本不等式 (Shannon's fundamental inequality) と呼ばれるもので、$H(A) \geqq H_B(A)$ の式の意味している事は「事象 B の何が起きたかを知った時は、B を知らない時よりも、A に関するエントロピー（不確定度）が減少する事」を意味している。これは実感にも近い。例えば

- 事象 A：明日雨が降るかどうか
- 事象 B：明日台風が接近するかどうか

とうような事象を想定した場合、台風が接近するのであれば、明日雨が降る確率は高くなる。つまり事象 B がわかった時点で事象 A のエントロピー（不確実性）が減少する事になる。

1.3 関数方程式

関数方程式とは、未知の関数の満たす性質が方程式で表されている場合に，その方程式を満たす関数を求める問題である。以下のようなものが代表的なものである。

関数方程式

$$f(x+y) = f(x) + f(y) \quad \rightarrow \quad f(x) = ax \tag{1.31}$$
$$f(xy) = f(x) + f(y) \quad \rightarrow \quad f(x) = a\log x \tag{1.32}$$

基本的な解法は以下

- 変数に具体的な値を入れて、特殊な x についての値を求める
- $f(x)$ の導関数 $f'(x)$ を求める
- $f'(x)$ を積分して $f(x)$ を求める

■$f(x+y) = f(x) + f(y)$ の性質を持った関数を求める

これは線形性の定義そのものであり、求める関数が $f(x) = ax$ となるのは至極当然のように思える。

- この関数は原点を通る
  $x = y = 0$ を代入すると、$f(0) = f(0) + f(0)$。この両辺から $f(0)$ を引いて $f(0) = 0$。つまり、この関数は原点を通る。
- この関数は奇関数である
  次に、$y = -x$ とすると、$f(x-x) = f(x) + f(-x)$ となり、左辺は $f(x-x) = f(0) = 0$ なので、$0 = f(x) + f(-x)$ となり、$f(x) = f(-x)$ となる。つまり、この関数は奇関数である。
- 導関数を求める
  以下のように、微分 $f'(x)$ 結果は定数になり、その値は微分関数に $x = 0$ の値を入れた値 $f'(0)$ になる。
  まず、導関数の定義から

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(h)}{h}$$

  $f(x+y) = f(x) + f(y)$ が成り立つなら、それを分母の $f(x+h)$ に適応して

$$f'(x) = \lim_{h\to 0} \frac{f(x) + f(h) - f(h)}{h} = \lim_{h\to 0} \frac{f(h)}{h}$$

  この式は、$f'(x)$ を求めているにもかかわらず x が全く現れていない。ということはどんな x に対しても同じ値になることを意味している。また、この関数は原点を通り、$f'(x)$ の値はどんな x でも同じなので、$f'(0)$ で求めることができる。定数を a とすると

$$a = f'(0) = \lim_{h\to 0} \frac{f(0+h) - f(0)}{h}$$

- 積分して関数 $f(x)$ を求める
  導関数が定数になるという事から $f'(x) = a$ とすると、それを積分する事で元の関数は $f(x) = ax + c$

と表すことができる。ここで $f(0) = 0$ である事から、$c = 0$ となるので、結局求める関数は $f(x) = ax$ となる

■$f(xy) = f(x) + f(y)$ の性質を持った関数を求める

掛け算が足し算になるという性質をもった関数として対数関数 $f(x) = \log x$ がすぐに思い浮かぶと思う。実際に $a \log x$ ならば、対数関数の法則 $\log(xy) = \log x + \log y$ からこの式が成立する事がわかる。

- この関数は $x = 1$ の時 0 である
  $x = 1$ を代入すると、$f(y) = f(1) + f(y)$。この両辺から $f(y)$ を引いくと $f(1) = 0$。
- 導関数を求める
  普通の微分定義は：
  $$f'(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon}$$
  でも今回は「掛け算」が出てくるので、加法ではなく乗法的に増やす形にしてみる。
  $$f'(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon x) - f(x)}{\epsilon x}$$
  この分母の $f(x + \epsilon x)$ を以下のように、x をかっこの外に取り出し、与えられた関数の性質 $f(xy) = f(x) + f(y)$ を使って変形すると
  $$f(x + \epsilon x) = f\{(1 + \epsilon)x\} = f(1 + \epsilon) + f(x)$$
  この式の右辺と左辺から $f(x)$ を引いて、さらに ϵx で割ると
  $$\frac{f(x + \epsilon x) - f(x)}{\epsilon x} = \frac{1}{x} \frac{f(1 + \epsilon)}{\epsilon}$$
  ここで、$\epsilon \to 0$ の極限をとると、左辺は明らかに $f(x)$ の導関数 $f'(x)$ となるので、
  $$f'(x) = \lim_{\epsilon \to 0} \frac{1}{x} \frac{f(1 + \epsilon)}{\epsilon}$$
  ここで、先に $f(1) = 0$ であることを示しているので、$x = 1$ を代入すると以下のように x が消えて、x とは関係のないある定数に近づくことがわかる。
  $$f'(1) = \lim_{\epsilon \to 0} \frac{f(1 + \epsilon)}{\epsilon}$$
  この定数を以下のように c とおく
  $$\lim_{\epsilon \to 0} \frac{f(1 + \epsilon)}{\epsilon} = c$$
  結果、$f'(x)$ は以下のように書くことができる。
  $$f'(x) = c\frac{1}{x}$$
- 積分して関数 $f(x)$ を求める
  上記を積分すると
  $$f(x) = c \log_\epsilon x + d$$
  ここで $f(1) = 0$ である事から、$d = 0$ となるので、結局求める関数は $f(x) = c \log_\epsilon x$ となる。

【別の解法】微分する方法

与えられた関数の性質についての以下の式を y で微分する

$$f(xy) = f(x) + f(y)$$

左辺 $f(xy)$ は合成関数なので、連鎖律（chain rule）を使って

$$\frac{d}{dy}f(xy) = f'(xy) \cdot \frac{d}{dy}(xy) = f'(xy) \cdot x$$

右辺の $f(x)$ は、y で微分する場合は定数であり 0 なので $f'(y)$ のみが残る。つまり、

$$xf'(xy) = f'(y)$$

ここで $y = 1$ のときを考えると

$$xf'(x) = f'(1)$$

$f'(a)$ は定数なので、$f'(1) = a$ とおくと上の式は

$$xf'(x) = a$$
$$f'(x) = \frac{a}{x}$$

これを積分すると

$$f(x) = a\log x + d$$

元の式に $x = 1$ を代入すると $f(y) = f(1) + f(y)$ であり $f(1) = 0$。この $f(1) = 0$ を上の式に代入すると $d = 0$ となるので、

$$f(x) = a\log x$$

1.4　ラグランジェの未定乗数法

ラグランジュの未定乗数法とは、制約条件のもとで関数の極値（最大・最少値）を求めるための方法。例えば二変数の場合は以下のようになる。

制約条件 $g(x,y)=0$ のもとで目的関数 $f(x,y)$ を最大・最小を求める問題は、ラグランジュ (Lagrange) 乗数を λ とし、

$$L(x,y,\lambda)=f(x,y)-\lambda g(x,y) \tag{1.33}$$

と置き、関数 L をそれぞれ 3 つの変数 x,y,λ で偏微分した値をゼロとした連立方程式を解く事で求める事ができる。

$$\frac{\partial L}{\partial x}=\frac{\partial L}{\partial y}=\frac{\partial L}{\partial \lambda}=0 \tag{1.34}$$

■ラグランジュの未定乗数法の解き方

この解き方を考えるにあたって、以下のような例題を考えてみる[*9]。

例題 1.1. x、y が $x+y=1$ という制約条件を満たす場合、つまり $g(x,y)=x+y-1=0$ の元で、目的関数 $f(x,y)=x^2+y^2$ の最小値を求める。

- **図形的に解く**

  最初に理解しやすいように、図形的に解を求めてみる。

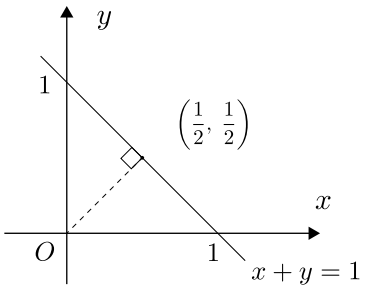


図 8　例題の図形的解法

まず、x、y が $x+y=1$ を満たす点は、$y=-x+1$ であり図 8 のような直線上の点であり、求めたい x、y はこの直線上に存在する事が制約条件になる。次に、$f(x,y)=x^2+y^2$ は、原点とその直線上の点 (x,y) との距離 $\sqrt{x^2+y^2}$ の二乗であり、原点からその直線上の点で最も近いのは図のように直線に垂線を下ろした足となる。

[*9] YouTube チャンネル AIcia Solid Project のラグランジュの未定乗数法の説明 https://www.youtube.com/watch?v=2-E4XiHQEcM&t=1333s に沿って記載している。

この点の座標は図 8 からわかるように、$\left(\frac{1}{2}, \frac{1}{2}\right)$ である。つまりこの点で最小値を取り、その最小値は以下のように $\frac{1}{2}$ となる。

$$x^2 + y^2 = \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

- **ラグランジュの未定乗数法で解く**

  今度はこれをラグランジュの未定乗数法で解いてみる。問題を式 (1.33) のように関数 L の形に書くと

$$\begin{aligned} L(x, y, \lambda) &= f(x, y) - \lambda g(x, y) \\ &= x^2 + y^2 - \lambda(x + y - 1) \end{aligned}$$

  これを x、y、λ で偏微分して、それをゼロと置いて 3 つの方程式を解く。

$$\begin{cases} \dfrac{\partial L}{\partial x} = 2x - \lambda = 0 \\ \dfrac{\partial L}{\partial y} = 2y - \lambda = 0 \\ \dfrac{\partial L}{\partial \lambda} = -(x + y - 1) = 0 \end{cases}$$

  第一式と第二式から

$$x = \frac{1}{2}\lambda, \quad x = \frac{1}{2}\lambda$$

  これを第三式に代入して解くと $\lambda = 1$ となるので、

$$x = \frac{1}{2}, \quad x = \frac{1}{2}, \quad \lambda = 1$$

  なので $f(x, y)$ の最小値は以下のように $\frac{1}{2}$ となる。

$$x^2 + y^2 = \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

■ラグランジェの未定乗数法の意味

この式（1.34）についてその意味を考えていこう。まず最後の項の $\dfrac{\partial L}{\partial \lambda} = 0$ であるが、

$$\frac{\partial L}{\partial \lambda} = \frac{\partial \left(f(x,y) - \lambda g(x,y) \right)}{\partial \lambda} = -g(x,y) = 0$$

であり、制約条件 $g(x,y) = 0$ を意味しているに過ぎない。本質は第一項と第二項で、この二つを並記すると

$$\frac{\partial L}{\partial x} = \frac{\partial \left(f(x,y) - \lambda g(x,y) \right)}{\partial x} = 0$$
$$\frac{\partial L}{\partial y} = \frac{\partial \left(f(x,y) - \lambda g(x,y) \right)}{\partial y} = 0$$

この式を変形し、さらに $f(x,y)$ を f、$g(x,y)$ を g と書くと

$$\frac{\partial f}{\partial x} = \lambda \frac{\partial g}{\partial x}$$
$$\frac{\partial f}{\partial y} = \lambda \frac{\partial g}{\partial y}$$

これを行列表記すると

$$\begin{pmatrix} \dfrac{\partial f}{\partial x} \\ \dfrac{\partial f}{\partial y} \end{pmatrix} = \lambda \begin{pmatrix} \dfrac{\partial g}{\partial x} \\ \dfrac{\partial g}{\partial y} \end{pmatrix}$$

以上のように、式（1.34）は以下の条件と同じである。

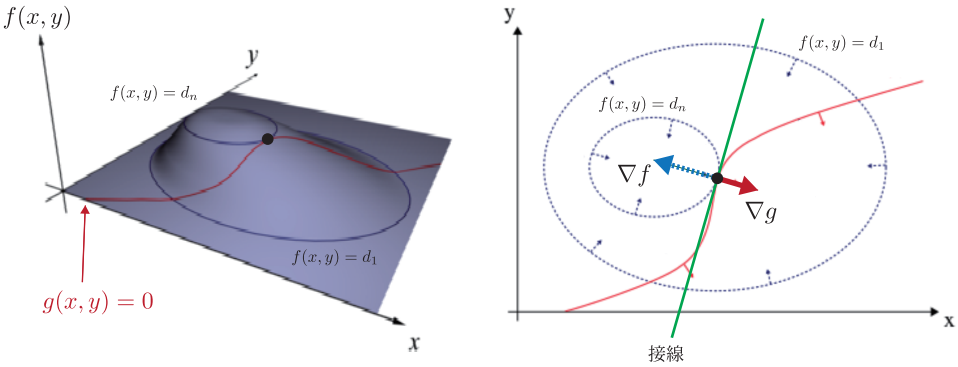
$$g(x,y) = 0 \qquad \text{かつ} \qquad \nabla f = \lambda \nabla g \tag{1.35}$$

ここで、∇f、∇g は

$$\nabla f = \begin{pmatrix} \dfrac{\partial f}{\partial x} & \dfrac{\partial f}{\partial y} \end{pmatrix}^t \qquad \nabla g = \begin{pmatrix} \dfrac{\partial g}{\partial x} & \dfrac{\partial g}{\partial y} \end{pmatrix}^t$$

この ∇f および ∇g は、点 (x,y) における関数 f と g の勾配 (gradient) と呼ばれるものである。この勾配こそが、何故ラグランジュの未定乗数法が成立するかのイメージを導く。

先の 2 変数の事例を図にしたのが図 9 である。(a) の鳥観図をみるとちょうど黒丸あたりに $g(x,y) = 0$ を制約条件とした場合の最大値があると推測される。この鳥観図の中に青線のように同じ高さ d の等高線を描くことを考えよう。そしてその高さを d_1 から徐々に大きくしながら頂点に向かってひとつづつ等高線を描くイメージをしよう。そして、途中でちょうど d_n の時に等高線と $g(x,y) = 1$ が一点で交わったとする。それを xy 平面上に描いたのが図 9 の (b) の等高線図である。



(a) 鳥観図　　(b) 等高線図

図 9　ラグランジュの未定乗数法のイメージ

この時、$g(x,y)=0$ と等高線 $f(x,y)=d_n$ は XY 平面上でも 1 点で交わり接線を共有する。この接線に対してちょうど垂直に交わるベクトルが勾配ベクトル ∇f と ∇g である。

つまり、式（1.35）の $\nabla f=\lambda\nabla g$ は、∇f と ∇g が平行である事を意味しており

XY 平面上の関数 $g(x,y)=0$ に沿って少しづつ移動しながら、∇f と ∇g を計算し、この二つが平行になる点 (x,y) を探すようなイメージでとらえればよい。

例題 1.2. 関数 $g(x,y)=2x+10y-15$ についての制約条件 $g(x,y)=0$ のもとで、目的関数 $f(x,y)=x^2+y^2+3$ の極値を求めよう。

$$L=f(x,y)-\lambda\quad g(x,y)$$
$$L=x^2+y^2+3-\lambda(2x+10y-15)$$

とおいて、L を x、y、λ で偏微分して 0 とおくと以下の連立方程式が得られる。

$$\begin{cases}\dfrac{dL}{dx}=2x-2\lambda=0 & (1.36)\\ \dfrac{dL}{dy}=2y-10\lambda=0 & (1.37)\\ \dfrac{dL}{d\lambda}=2x+10y-15=0 & (1.38)\end{cases}$$

式（1.36）より $x=\lambda$、式（1.37）より $y=5x$、これらを式（1.38）に代入して x だけの式にして x を求めて、あとは逆算すると

$$p=(x,y,\lambda)=\left(\frac{15}{52},\frac{75}{52},\frac{15}{52}\right)$$

となる。

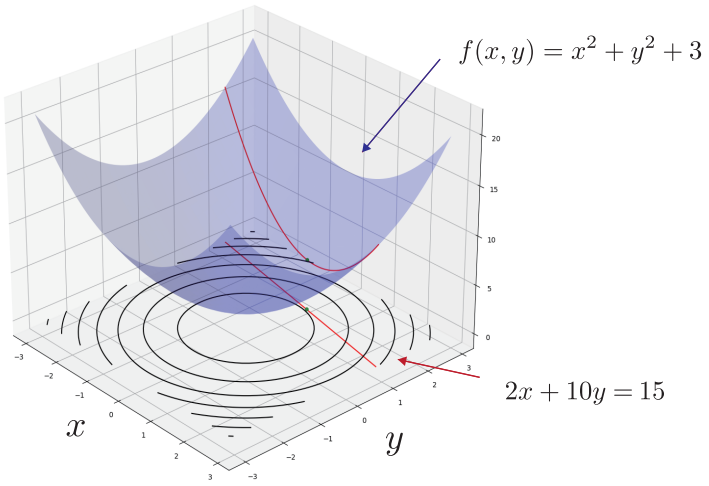


図 10　例題 1.2 の答えを示す図

以下のソースコードは、sympy という代数計算ライブラリを使って上記の例題を解いて、さらに図 10 のグラフを描く Python プログラム。

ソースコード 2　ラグランジュ未定乗数法の例題を解くプログラム

```
# -*- coding: utf-8 -*-
"""
ラグランジュの未定乗数法
Created␣on␣Sun␣May␣␣9␣07:17:43␣2021
@author:␣hiros
"""
import numpy as np
import sympy as sy  #数式処理用ライブラリ

import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
#sy.init_printing()

# ３つの変数を宣言する
x, y, l = sy.symbols('x␣y␣λ')
# 目的関数と制約関数を準備する
def purpose(x, y):      #目的関数
    return x**2 + y**2 + 3
def restrict(x, y):     #制約関数
    return 2*x + 10*y - 15
def gradient(x):        #図を作るときに使う制約関数の軌跡をとる関数
    return (-2*x + 15) / 10

# 目的関数f(x,y) 制約関数g(x,y) ラグランジュ関数L(x,y)
fx = purpose(x, y)
```

```
gx = restrict(x, y)
L = fx - l * gx

# L を x,y,l で偏微分して連立方程式を解く
dx = sy.diff(L, x)
dy = sy.diff(L, y)
dl = sy.diff(L, l)
p = sy.solve([dx, dy, dl])

# 格子状のXY 平面メッシュ（最少ー３，最大３，間隔０．１）を作る
xs = np.arange(-3, 3, 0.1)
ys = np.arange(-3, 3, 0.1)
X, Y = np.meshgrid(xs, ys)

# figure インスタンスを生成し、figure インスタンスの axes を生成
fig = plt.figure()
ax = Axes3D(fig)
#バージョン 3.4から,明示的にAxes3D を生成し Figure に追加が必要
fig.add_axes(ax)

# 目的関数f(x,y)のサーフェイスを描く
ax.plot_surface(X, Y, purpose(X, Y), color='blue', alpha=0.2)
# 制約関数g(x,y)のサーフェイス上のどこを通るかの軌跡を描く
ax.plot(xs, gradient(xs), [purpose(x, gradient(x)) for x in xs], color="red")
# 連立方程式を解いて求めた極値の点を示す
ax.scatter([p[x]], [p[y]], [purpose(p[x], p[y])], color='green')
# 底面に的関数f(x,y)の等高線を描画
ax.contour(X, Y, purpose(X, Y), colors = "black", offset = 0)
# 底面に制約関数g(x,y)を描く
ax.plot(xs, gradient(xs),color="red")
# 底面に連立方程式を解いて求めた極値の点を示す
ax.scatter([p[x]], [p[y]],color='green')
plt.show()
```

2　ダイバージェンス

統計学や情報理論をはじめとした広い分野で、ダイバージェンス (divergence) という言葉が出てくる。ダイバージェンス（divergence）は、2 つの確率分布の「近さ」や「ずれの大きさ」を表す指標である。一般的な距離関数とは異なり、対称性（交換則）を満たさないため、厳密には距離とは呼べないが、距離的な概念と捉えることができる。そうした二つの確率分布の距離に近い概念が以下の 3 つである。

2 つの確率分布の近さの指標

- 【相互情報量】　クロス集計表における同時分布と周辺分布の差異から、2 つの変数間の統計的関連性を定量化する指標。

$$I(A;B) = \sum_i \sum_j p(A_i, B_j) \log \frac{p(A_i, B_j)}{p(A_i)\, p(B_j)} \tag{2.1}$$

- 【相対エントロピー】　KL ダイバージェンスと呼ばれ、2 つの確率分布の近さの代表的指標。特に「真の分布」と「仮定した分布」の違いを測る際に使われ、機械学習やベイズ推論などで広く用いられる。

$$D(P||Q) = \sum_{x \in A} P(x) \log \frac{P(x)}{Q(x)} \tag{2.2}$$

- 【交差エントロピー】　クロスエントロピーと呼ばれ、機械学習の誤差を評価する関数としてよく使用される

$$H(P,Q) = -\sum_{x \in A} P(x) \log Q(x) \tag{2.3}$$

以下、これらについて詳細をまとめていく。

なお、ここでは触れないが、このように確率分布の隔たり具合を論じる分野として情報幾何学があるようである。確率分布を定めるパラメータ空間に距離（ダイバージェンス）を定義して上げる事で、確率分布の隔たり具合を論じるもののようである。情報幾何学では上記以外の各種の近さの指標が出てくるようである。

2.1　相互情報量

相互情報量はクロス表の周辺確率と同時確率の関係から定義される。事象 A と事象 B とに関係があるとすると、A が何であるかを知った事によって B が何であるかについての情報を得たことにもなる。これは周辺確率と同時確率との関係に現れる。

例えば

- 事象 A：体温が平熱より高いかどうか
- 事象 B：風邪を引いているかどうか

とする。このとき、体温を測ることで熱があるかどうかがわかると、風邪を引いているかどうかについての情報を得ることができる。この「体温を計ることによる風邪の状態に関する不確かさの減少量」を相互情報量 $I(A;B)$ と呼ぶ。

相互情報量は、次のようにエントロピー[*10]を使って表される[*11]。

$$I(A;B) = H(B) - H(B \mid A) \tag{2.4}$$

これは、事象 A を知ることで、事象 B に関する不確かさがどれだけ減るかを示す。一方、以下のように表現することもできる。

$$I(A;B) = H(A) - H(A \mid B)$$

これは、相互情報量が対称的であること（$I(A;B) = I(B;A)$）を示している。

このことをベン図で表すと図 11 のように、エントロピー同士の重なり部分として視覚的に捉えることができる。

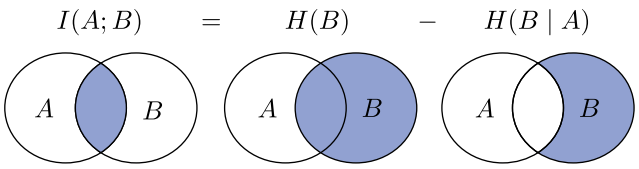


図 11　相互情報量

相互情報量を確率による計算式で表していってみよう。まず式（2.4）は

$$I(A;B) = H(A) - H(A \mid B)$$

と書いても同じなので、この式を確率を用いて書き下すと

$$I(A;B) = -\sum_i p(A_i) \log p(A_i) + \sum_i \sum_j p(A_i, B_j) \log p(A_i \mid B_j)$$

[*10] $H(B)$ は、風邪をひいているかどうか（事象 B）のエントロピー（不確かさ）、$H_A(B)$ は、体温が分かっているという条件のもとでの風邪を引いているかどうかのエントロピー（条件付きエントロピー）

[*11] 相互情報量は $I(A,B)$ と表す事もある

ここで、右辺を変形するために以下の式を代入すると

$$p(A_i) = \sum_j p(A_i, B_j) \qquad p(A_i \mid B_j) = \frac{p(A_i, B_j)}{p(B_j)}$$

$$\begin{aligned} I(A;B) &= -\sum_i \sum_j p(A_i, B_j) \log p(A_i) + \sum_i \sum_j p(A_i, B_j) \log \frac{p(A_i, B_j)}{p(B_j)} \\ &= \sum_i \sum_j p(A_i, B_j) \log \frac{1}{p(A_i)} + \sum_i \sum_j p(A_i, B_j) \log \frac{p(A_i, B_j)}{p(B_j)} \\ &= \sum_i \sum_j p(A_i, B_j) \log \frac{p(A_i, B_j)}{p(A_i)p(B_j)} \end{aligned}$$

この式が相互情報量の定義となる

定義 2.1. 【相互情報量】同時確率 $p(A, B)$ を有する 2 つの確率変数 A と B がある場合、相互情報量 $I(A;B)$ を以下のように定義する。

$$I(A;B) = \sum_i \sum_j p(A_i, B_j) \log \frac{p(A_i, B_j)}{p(A_i)\, p(B_j)} \tag{2.5}$$

12 ページの計算事例で確認してみよう。表を再掲すると以下の表。

	B_1 風邪ひいている	B_2 風邪ひいてない	$p(A_i)$
A_1 熱がある	0.55	0.05	0.60
A_2 熱がない	0.10	0.30	0.40
$p(B_j)$	0.65	0.35	

$$\begin{aligned} I(A;B) &= \sum_i \sum_j p(A_i, B_j) \log \frac{p(A_i, B_j)}{p(A_i)\, p(B_j)} \\ &= 0.55 \times \log \frac{0.55}{0.65 \times 0.60} + 0.05 \times \log \frac{0.05}{0.35 \times 0.60} + 0.10 \times \log \frac{0.10}{0.65 \times 0.40} + 0.30 \times \log \frac{0.30}{0.35 \times 0.40} \\ &= 0.36 \end{aligned}$$

ここで、熱を測る前の B についてのエントロピー $H(B)$ は

$$H(B) = -(0.65 \times \log_2 0.65 + 0.35 \times \log_2 0.35) = 0.93$$

さらに熱を測ったあとの B についてのエントロピー $H(B \mid A)$ を求めよう。まず熱を測ったあとの条件付き確率は

$$p(B_1 \mid A_1) = \frac{0.55}{0.6} = 0.92 \qquad p(B_2 \mid A_1) = \frac{0.05}{0.6} = 0.08$$
$$p(B_1 \mid A_2) = \frac{0.10}{0.4} = 0.25 \qquad p(B_2 \mid A_2) = \frac{0.30}{0.4} = 0.75$$

なので「熱がありなし」の条件付きエントロピーは

$$H(B \mid A_1) = -\left(0.92 \times \log 0.92 + 0.08 \times \log 0.08\right) = 0.41$$
$$H(B \mid A_2) = -\left(0.25 \times \log 0.25 + 0.75 \times \log 0.75\right) = 0.81$$

上記より、熱を測ったあとの B についてのエントロピーは

$$\begin{aligned} H(B \mid A) &= p(A_1)\ H_{A_1}(B) + p(A_2)\ H_{A_2}(B) \\ &= 0.60 \times 0.41 + 0.4 \times 0.81 = 0.57 \end{aligned}$$

以上のように検温前のエントロピー $H(B)$ は 0.93。検温を済ませた後[*12]のエントロピー $H_A(B)$ は 0.57 となる。この差分が検温によって得られた情報量 $I(A;B)$ を意味している。

別の解釈をすると、風邪かどうかを診断するのに必要な情報量は 0.93 ビットであり、その必要な情報量に対して熱を測る事によって 0.57 ビットが得られ、あと残り 0.36 ビットの不確実性が残っているという解釈もできる。

相互情報量 $I(A;B)$ は、事象 B を特定するために、事象 A がわかった事によってどの程度の「不確かさ」が減少したかを示す。なので二つの事象が完全に独立ならば得られる情報量は「全くない」ので $I(A;B) = 0$、２つが完全に従属な時は「事象 B の何が起こったかを知る」のと同じなので最大となり $I(A;B) = H(B)$ となる。このように相互情報量 $I(A;B)$ は、２つの確率変数 A と B がどのくらい関連があるかを示す尺度と考える事ができる。

上記のことを計算式でも確認しよう。相互情報量 $I(A;B)$ は以下のように計算さえる。

$$I(A;B) = \sum_{i,j} p(A_i, B_j) \log \frac{p(A_i, B_j)}{p(A_i) \cdot p(B_j)}$$

- 分子： $p(A_i, B_j)$ 実際に観測された「一緒に起きる確率」
- 分母： $p(A_i)p(B_j)$ A と B が独立だったら期待される同時確率

つまり、相互情報量は、**実際の共起確率と「独立だったら」の期待値との比**の対数（差）である。もし事象 A と B が独立なら、以下のように計算上も相互情報量 $I(A;B) = 0$ となる。

$$p(A_i, B_j) = p(A_i) \cdot p(B_j) \Rightarrow \log \frac{p(A_i, B_j)}{p(A_i)p(B_j)} = \log 1 = 0$$

[*12] 検温の結果で熱があったかどうかは問わない。検温した後の風邪かどうかの不確実性である。

2.2 相対エントロピー (KL ダイバージェンス)

次に相対エントロピーについて示す。「相対エントロピー」は「ダイバージェンス」または「カルバック・ライブラー情報量」(Kullback–Leibler divergence) と呼ばれる。

定義 2.2. 【相対エントロピーの定義 (K L情報量)】
アルファベット空間 A に値をとる 2 つの確率分布 $P(x)$、$Q(x)$ $(x \in A)$ に対して、相対エントロピー $D(P||Q)$ を以下のように定義する。

$$D(P||Q) = \sum_{x \in A} P(x) \log \frac{P(x)}{Q(x)} \tag{2.6}$$

ただし、$P(x) > 0$ のときは必ず $Q(x) > 0$ であるものとする。

上記定義中の「アルファベット空間」というのは、取りうる記号や値の集合のことで、確率変数がとりうるすべての値（記号）を集めた集合のことである。また、また上記定義中の「ただし」以下について補足すると、関数 $\log 0$ は定義されないため、以下のような拡張的な定義を用いる：

- $P(x) = 0$ のとき：
$$P(x) \log \frac{P(x)}{Q(x)} = 0 \qquad (Q(x) \text{ が } 0 \text{ でも } > 0 \text{ でも})$$
- $Q(x) = 0$, $P(x) > 0$ のとき：
$$P(x) \log \frac{P(x)}{Q(x)} = \infty$$

■相対エントロピーの意味をベイズ的に解釈する この相対エントロピーをベイズ的に解釈してみよう。$X = \{x_1, x_2, \cdots, x_n\}$ をアルファベット空間とし、各 x に対しての確率 $Q(x)$ が定まっているする。つまり、ベイズ確率でいう事前分布が定まっているとする。今 X に関する新たなデータ I を知ったとし、 その結果に従う (条件付き) 確率が $P(x)$ になったとする。つまりベイズ確率でいう事後分布が求まったとする時、得られる情報量は

$$\begin{aligned}(-\log Q(x)) - (-\log P(x)) &= \log P(x) - \log Q(x) \\ &= \log \frac{P(x)}{Q(x)}\end{aligned}$$

となる。この情報量に各 x_i の事後確率分布をかけて、得られた情報の期待値を算出したものが以下であり、これが相対エントロピーとなる。

$$\sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}$$

例えば、ある病気の有病率（事前確率）を $Q(x)$ とし、ある検査結果（情報）を得たあとでの確率（事後確率）を $P(x)$ とする。このとき、「事後でどれだけ確信を深められたか？」は

$$\log \frac{P(x)}{Q(x)}$$

で表される。それを「すべての可能性について、どれだけ信念が更新されたか」の平均を取ると、それが相対エントロピーとなる。

$$D(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

相対エントロピーは、2 つの確率変数の近さを意味すると同時に、事前確率 $Q(x)$ が事後確率 $P(x)$ に変わった場合に、どの程度エントロピー（不確かさ）が減ったのかという事を意味している事になる。つまりなんらかの学習によって予測がより正確になった程度として解釈可能である。これが相対エントロピー（KL ダイバージェンス）がよく機会学習に現れる理由である。

性質 2.1. 【相対エントロピーの性質】

$$D(P||Q) \geqq 0 \tag{2.7}$$

$D(P||Q) = 0$ となるのは $P = Q$ の場合

$$D(P||Q) \neq D(Q||P) \tag{2.8}$$

この相対エントロピー (KL ダイバージェンス) は、二つの分布の「近さ」を意味しており距離っぽい性質を持っているが、上記のように交換則が成立しない $D(P||Q) \neq D(Q||P)$ ので、一般的な距離とは異なる。つまり「どちら」から「どちら」に値を測るかで結果が変わってくる。

一般的には、この相対エントロピー (KL ダイバージェンス) はパラメータ推定の当てはまりの良さを測る指標として使われる。つまり以下のように、真の分布が $p(x)$ だとし $q(x|\theta)$ という確率分布を準備して、θ を色々変えて KL ダイバージェンスを計算してみて、値が小さくなるように $q(x|\theta)$ を選べば、$p(x)$ を近似できたことになる。

$$D(P||p(x|\theta)) = \sum_{x \in A} P(x) \log \frac{P(x)}{p(x|\theta)}$$

2.3 交差エントロピー

続いて、交差エントロピー (クロスエントロピー cross entropy) についてみていく。交差エントロピーは、正解値と推定値の比較を行うときによく使用される。実際に機械学習などの損失関数（誤差関数）としてもよく用いられる。

定義 2.3. 【交差エントロピーの定義 *(Cross Entropy)*】
正解の確率分布を $P(x)$、推定した確率分布を $Q(x)$ としたとき、交差エントロピーは、以下のようになる。

$$H(P,Q) = -\sum_{x \in A} P(x) \log Q(x) \tag{2.9}$$

$Q(x) = P(x)$ で交差エントロピーは最小となり、$P(x)$ または $Q(x)$ のエントロピーと同じになる。例えば真の分布 $P(x)$ と推定した分布 $Q(x)$ が一致している場合、交差エントロピーは「元々の確率分布が持つ予測のしにくさ（エントロピー)」と同じになる。しかし、真の分布と観測データの分布が一致していない場合、エントロピーに加えて、「分布がズレている分」だけエントロピーが増加する事になる。

■具体例　以下のような教師データ T を学習させた結果、Z_1 が得られたとした場合、この学習結果の誤差をクロスエントロピー関数で評価してみる。

$$T = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \qquad Z_1 = \begin{bmatrix} 0.5 \\ 0.2 \\ 0.3 \end{bmatrix}$$

単純に交差エントロピーの定義に当てはめて

$$\begin{aligned} H(T, Z_1) &= -\{0 \times \log 0.5 + 1 \times \log 0.2 + 0 \times \log 0.3\} \\ &= 0.70 \end{aligned}$$

もし学習によって得られた結果が

$$Z_2 = \begin{bmatrix} 0.06 \\ 0.90 \\ 0.04 \end{bmatrix}$$

であったとすると

$$\begin{aligned} H(T, Z_2) &= -\{0 \times \log 0.06 + 1 \times \log 0.90 + 0 \times \log 0.04\} \\ &= 0.05 \end{aligned}$$

■何故 2 乗和誤差より良いのか　誤差を見積もる関数として、クロスエントロピーは 2 乗和誤差よりも使われる事が多い。それはクロスエントロピーの方が収束が早いからである。

それを考えるにあたって 2 クラス分類の交差エントロピーを考える。例えば、コインを投げて表がでるか裏が出るかの事象を考えてみる。この時に表を 1、裏を 0 と符号化する。いまこのコインの表がでる真の確率を $P(1) = p$ とし、表が出る推定値を $Q(1) = q$ とする。当然 $P(0) = 1 - p$ であり、$Q(0) = 1 - q$ である。ここ

で、交差エントロピーは式（2.9）から

$$
\begin{aligned}
H(P,Q) &= -\sum_{x=0}^{1} P(x)\log Q(x) \\
&= -p\log q - (1-p)\log(1-q)
\end{aligned}
$$

また二乗和誤差は以下のようになる。

$$
\begin{aligned}
L(P,Q) &= \frac{1}{2}\sum_{x=0}^{1}\{P(x)-Q(x)\}^2 \\
&= \frac{(p-q)^2+(q-p)^2}{2}
\end{aligned}
$$

いま表が出る真の確率を 0.5 つまり、$p=0.5$ としたとき、q を $0-1$ まで変化させた時の 2 つの式の値は図 12 のようになる。このように交差エントロピーの方が正しい値から離れるとより大きな値をとる。それによって真値からの誤差が大きい程乖離も大きくなり、より真値に素早く近づきやすいという性質が生まれる。

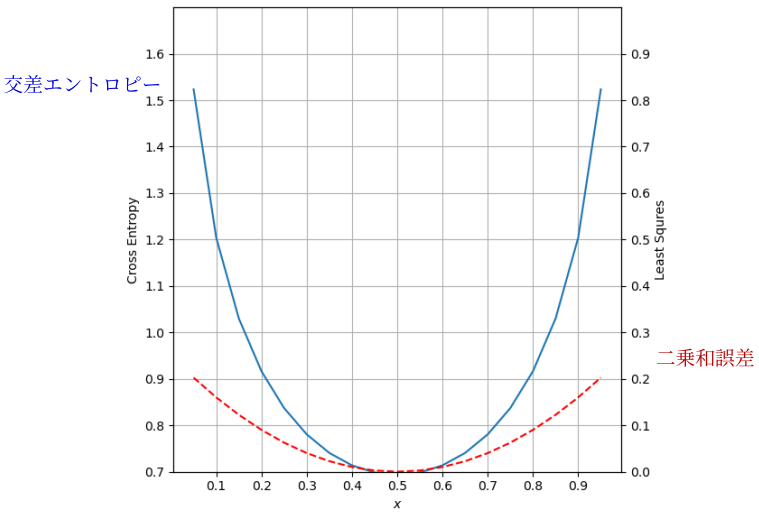


図 12　交差エントロピーと二乗和誤差のグラフ

この図を描写するプログラムは以下になる。

ソースコード 3　交差エントロピーと二乗和誤差のグラフを描くプログラム

```
import numpy as np
import matplotlib.pyplot as plt

def Cross_Entropy(p,q):
    # log(0)を避けるために微小な値を設定して加算する
    delta = 1e-7
    return -p * np.log(q+delta) - (1-p) * np.log(1-q+delta)
```

```
def Least_squares_error(p,q):
    return ((p−q)**2 + (q−p)**2) /2

p = 0.5
x = np.arange(0.05,1.0,0.05)
y = Cross_Entropy(p,x)
y2 = Least_squares_error(p, x)


# figure インスタンスを生成し、figure インスタンスの axes を生成
fig = plt.figure()
ax1 = fig.add_subplot(1,1,1)
ax2 = ax1.twinx()

ax1.grid(True)
ax1.yaxis.set_ticks(np.arange(0.6, 1.7, 0.1))
ax1.set_ylim(0.7, 1.7)
ax1.xaxis.set_ticks(np.arange(0, 1, 0.1))
ax2.yaxis.set_ticks(np.arange(0, 1, 0.1))
ax2.set_ylim(0, 1)
ax1.plot(x, y)
ax2.plot(x, y2, linestyle='dashed', color='red')
ax1.set_xlabel(r"$x$")
ax1.set_ylabel(r"Cross␣Entropy")
ax2.set_ylabel(r"Least␣Squres")
plt.show()
```

2.4　3つのエントロピーの関係

■ダイバージェンスと相互情報量　相互情報量と相対エントロピー（つまりダイバージェンス）とは密接な関係がある。相互情報量の定義式（2.5）を再掲すると以下。

$$I(A;B) = \sum_i \sum_j p(A_i, B_j) \log \frac{p(A_i, B_j)}{p(A_i)\, p(B_j)}$$

ここで、アルファベット空間を $A = \{00, 01, 10, 11\}$ とし、$P(x) = p(A_i, B_j)$、$Q(x) = p(A_i)\, p(B_j)$ とおくと、この式は

$$\sum_{x \in A} P(x) \log \frac{P(x)}{Q(x)}$$

となり相対エントロピーの定義式（2.6）と同じである。つまり図 13 のように相互情報量は、同時確率 $p(A_i, B_j)$ と周辺確率の積 $p(A_i)p(B_j)$ との相対エントロピーに等しい。

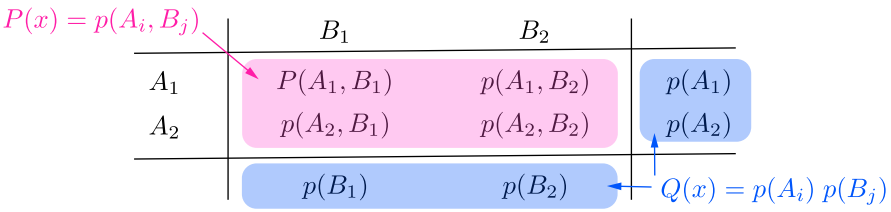


図 13　相互情報量と相対エントロピー

当然ながら事象 A と B が独立している時は、同時確率と周辺確率の積は等しい。つまり $p(A_i, B_j) = p(A_i)p(B_j)$ なので $\dfrac{P(x)}{Q(x)} = 1$ であり、相対エントロピー（ダイバージェンス）もゼロである。

■ダイバージェンスと交差エントロピー　ダイバージェンスの式 (2.6) は以下のように変形でき

$$\begin{aligned} D(P||Q) &= \sum_{x \in A} P(x) \log \frac{P(x)}{Q(x)} \\ &= \sum_{x \in A} P(x) \log P(x) - \sum_{x \in A} P(x) \log Q(x) \\ &= H(P) - H(P, Q) \end{aligned}$$

つまり、ダイバージェンスとはエントロピー - 交差エントロピーであるといえる。

■ダイバージェンスと最尤推定　n 個の確率変数 $x_1, x_2, \cdots, x_n$ は互いに独立で，同じパラメータ θ をもった確率分布から得られたとする。ここで、パラメータが θ であるという条件の元で x_i というデータが得られる確率を $q(x|\theta)$ と表すとすると、尤度関数 $L(\theta)$ は式（4.2）でしめしたように以下で表される。

$$L(\theta) = q(x_1|\theta) \cdot q(x_2|\theta) \cdot \cdots \cdot q(x_n|\theta) = \prod_{i=1}^{n} q(x_i|\theta)$$

また、尤度関数の積を和に変えるために、対数をとった対数尤度関数は式（4.3）であり以下

$$\begin{aligned} l(\theta) &= \log L(\theta) \\ &= \log q(x_1|\theta) + \log q(x_2|\theta) + \cdots + \log q(x_n|\theta) \\ &= \sum_{i=1}^{n} \log q(x_i|\theta) \end{aligned}$$

一方でダイバージェンスは、真の確率分布を $p(x_i)$ としパラメータ θ による推定確率分布を $q(x_i|\theta)$ とすると

$$\begin{aligned} D(P||Q) &= \sum_{i=1}^{n} p(x_i) \log \frac{p(x_i)}{q(x_i|\theta)} \\ &= \sum_{i=1}^{n} p(x_i) \log p(x_i) - \sum_{i=1}^{n} p(x_i) \log q(x_i|\theta) \end{aligned}$$

3　情報源

連続した事象 A を考える。つまり、$A_1, A_6, A_1, \cdots, A_n, \cdots$ のように事象が系列となって次から次へと出てくる場合を考える。こうした系列を生み出している元を**情報源**と呼ぶ。例えば英語の文章は、アルファベット 26 文字と句読点が次々と出てきた系列であると考える事ができ、その背景にアルファベットを発生させる情報源が想定できる。

3.1　用語の整理

いま $A_1, A_6, A_1, \cdots, A_n, \cdots$ のようなデータが情報源から得られるとする時、事象 A がとり得る要素の母集合

$$A = \{A_1, A_2, \cdots, A_n\}$$

を**情報源アルファベット**と呼ぶ。ただし、このアルファベットという呼び方は英字の $A - Z$ という 26 文字と句読点に限定した意味ではなく、比喩的な呼び方であって、「A_1 偶数」、「A_2 奇数」のようなものを $\{1, 0\}$ というように記号化した場合についてもこの母集合 $A = \{1, 0\}$ を情報源アルファベットと呼ぶ。また、それぞれの要素 A_i を**情報源記号**と呼ぶ。また、アルファベットの要素数が p 個の時、これを p **元アルファベット**とも呼ぶ[*13]。

情報源には、以下のように「記憶のない情報源」と「記憶のある情報源」と呼ばれるものに分類される。

記憶のない情報源　時刻 t に要素 A_i がでる確率が、それ以前の要素系列 $x_{t-s}, \cdots, x_{t-2}, x_{t-1}$ とは関係しない場合。例えば、サイコロを連続して振った場合の目の出方などで、前に何の目が出たから次の目がでやすいという事はない。

記憶のある情報源　時刻 t に要素 A_i がでる確率が、それ以前に得られた要素系列 $x_{t-s}, \cdots, x_{t-2}, x_{t-1}$ が何であったかに影響される。例えば、今日の天気が「晴れ」である確率は、昨日や一昨日の天気に影響されるような場合である。

■マルコフ情報源 (Markov source)　「記憶のある情報源」の事をマルコフ情報源という。いいかえれば、マルコフ情報源とは「過去の有限個の記号の生起が次の記号の生起に影響する情報源」であると言える。また、過去の s 個の記号に依存して x_i が決定するような情報源を図 14 のように s 重マルコフ情報源と呼ぶ。

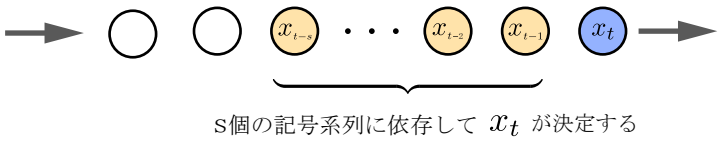


図 14　s 重マルコフ情報源

次にこの情報源の確率を考える。いま時刻 $t(t = 1, 2, \cdots)$ に発生した記号の系列を $x_1, x_2, \cdots$ と表すとする。この $x_1, x_2, \cdots$ のそれぞれの x_i に情報源アルファベットの $A = \{A_1, A_2, \cdots, A_n\}$ の中からひとつずつ

[*13] 例えば「A_1 偶数」、「A_2 奇数」の母集合 $A = \{1, 0\}$ ならば 2 元アルファベットとなる

要素が抽出されて、実際の時系列データ（例えば $A_4, A_{10}, A_3, \cdots, A_6, \cdots$）が決定されていくものとする。その時に、今までに出た $s(s = 1, 2, \cdots)$ 個のデータ系列

$$x_{t-s}, \cdots, x_{t-2}, x_{t-1}$$

によって、次にどの要素が出るかが確率論的に決まるとすると、時刻 t に要素 x_t の出る確率を

$$p(x_t | x_{t-s} \ \cdots \ x_{t-2} \ x_{t-1})$$

という条件付き確率として記述する事が出来、時刻 $t-s$ から時刻 t までの系列が得られる確率は

$$p(x_{t-s} \ \cdots \ x_{t-1} \ x_t) = p(x_t | x_{t-s} \ \cdots \ x_{t-2} \ x_{t-1}) p(x_{t-s} \ \cdots \ x_{t-2} \ x_{t-1})$$

と表す事ができる。

■遷移確率行列と状態遷移図　$m = 2$ の場合、すなわち 2 重マルコフ情報源を考えてみる。A の値は 2 元アルファベット

$$A = \{0, 1\}$$

とする。時刻 t の記号 x_t の生起確率は以下のようになる。

$$p(x_t | x_{t-2}, x_{t-1})$$

ここで要素 x_{t-2}, x_{t-1} 及び x_t は、$(1, 0)$ のどちらかをとるので以下の 8 種類の組み合わせである。

$$\begin{array}{cccc} p(0|00), & p(0|01), & p(0|10), & p(0|11) \\ p(1|00), & p(1|01), & p(1|10), & p(1|11) \end{array}$$

さらに、2 つの要素 x_{t-2}, x_{t-1} の組み合わせを一つの状態と捉えなおすと、以下の 4 つの状態が考えられる。

$$q_1 = (0, 0), \quad q_2 = (0, 1), \quad q_3 = (1, 0) \quad q_4 = (1, 1)$$

このように記憶している s 個の要素で状態を定義すると、データが得られる毎にある状態から次の状態に変化していると捉える事ができ、その推移を示したものが図 15 である。これを**状態遷移図**と呼ぶ。

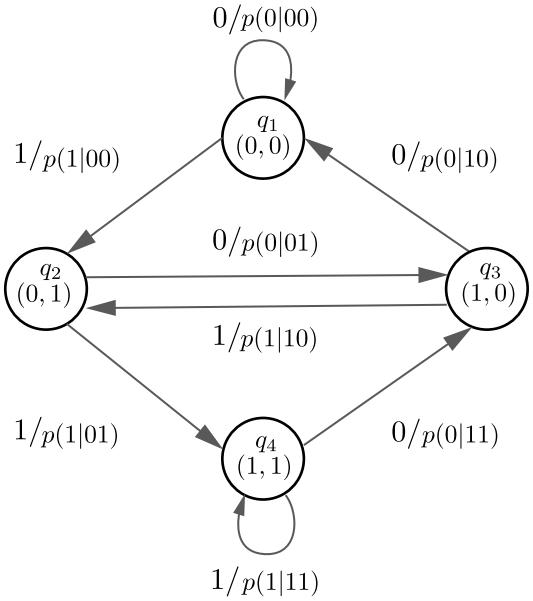


図 15　状態遷移図

状態遷移図の丸は状態を示し、遷移する可能性のある状態と状態を矢印で結び、そこに確率を書き入れている。例えば、図 15 の q_1 から q_2 に向かう矢印についている $1/p(1|00)$ は、「状態が q_1 の時に 1 が生起すると、状態 q_2 に変わり、その生起確率は $p(1|00)$ である。」事を意味する。

また、この状態遷移図は行列であらわす事が可能である。まず、ある状態 qi から qj の状態に変化する確率（遷移確率と呼ぶ）を

$$p_{ij} = p(q_j|q_i)$$

とあらわす。それによって、ある状態からある状態への変化を以下のように行列であらわす事ができる。これを**遷移確率行列**と呼ぶ。この場合は状態が 4 つ存在するので 4×4 行列となる。

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{pmatrix}$$

3.2　マルコフ情報源のエントロピー

先の遷移確率行列は、それぞれの状態毎のそれぞれの出力（4×2 個）の確率である。マルコフ情報源のエントロピーを計算するにあたって、まず 4 つの状態 q_1, q_2, q_3, q_4 がどのような確率で出現するかを求める事を考えよう。

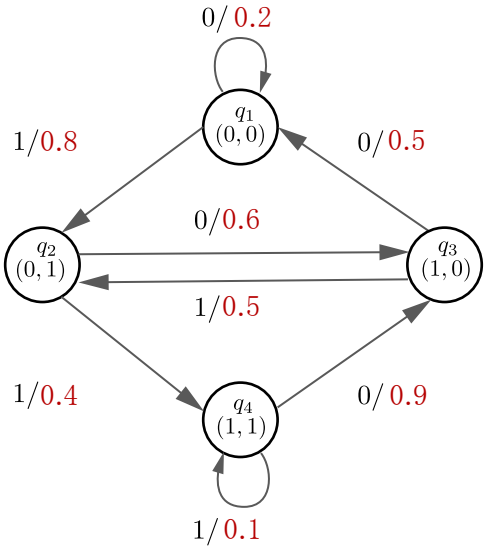
このマルコフ情報源は、最初の状態（初期状態）から遷移を開始し、時間が十分経過した後に定常状態になる、つまり初期状態に依存しなくなると考える。この定常状態に到達した時の q_1, q_2, q_3, q_4 の 4 つ状態が発生する確率を w_1, w_2, w_3, w_4 であるとする。当然この 4 つ以外の状態は存在しないので、その確率の合計は 1 となる。つまり、

$$w_1 + w_2 + w_3 + w_4 = 1 \tag{3.1}$$

ここから先は、以下のような具体的な値を当てはめて考えてみる。

$$p(0|00) = 0.2, \quad p(0|01) = 0.6, \quad p(0|10) = 0.5, \quad p(0|11) = 0.9$$
$$p(1|00) = 0.8, \quad p(1|01) = 0.4, \quad p(1|10) = 0.5, \quad p(1|11) = 0.1$$

この場合の状態遷移図と遷移確率行列は図 16 となる。



$$P = \begin{pmatrix} 0.2 & 0.8 & 0 & 0 \\ 0 & 0 & 0.6 & 0.4 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.9 & 0.1 \end{pmatrix}$$

図 16　状態遷移図と遷移確率行列の事例

この時、状態 q_1 に注目すると、状態 q_1 になるのは以下の 2 つである。

- 直前の状態が q_1 で、確率 0.2 で再び q_1 に遷移した場合
- 直前の状態が q_3 で、確率 0.5 で q_1 に遷移した場合

なので w_1 は以下のように表すことができる。

$$w_1 = 0.2w_1 + 0.5w_3$$

この式を含めて、同様に 4 つの状態について書いた以下の連立方程式（3.2）と先の方程式（3.1）との 5 つの連立方程式を解けばよい

$$\begin{cases} w_1 = 0.2w_1 + 0.5w_3 \\ w_2 = 0.8w_1 + 0.5w_3 \\ w_3 = 0.6w_2 + 0.9w_4 \\ w_4 = 0.4w_2 + 0.1w_4 \end{cases} \tag{3.2}$$

この方程式を解くと

$$w_1 = 0.2036, \quad w_2 = 0.3258, \quad w_3 = 0.3258, \quad w_4 = 0.1448$$

以上の計算を行列であらわしておこう。定常状態のそれぞれの状態になる確率を

$$\omega = (w_1, w_2, w_3, w_4)$$

とすると式 (3.2) は

$$\begin{pmatrix} w_1 & w_2 & w_3 & w_4 \end{pmatrix} = \begin{pmatrix} w_1 & w_2 & w_3 & w_4 \end{pmatrix} \begin{pmatrix} 0.2 & 0.8 & 0 & 0 \\ 0 & 0 & 0.6 & 0.4 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.9 & 0.1 \end{pmatrix}$$

$$\omega = \omega P$$

というようにあらわす事ができる。

【参考】この５つの連立方程式を解く Python プログラムが以下

```
import sympy as sy
w1,w2,w3,w4 = sy.symbols("w1 w2 w3 w4")
f1 = 0.8*w1 - 0.5*w3
f2 = 0.8*w1 - w2 + 0.5*w3
f3 = 0.6*w2 - w3 + 0.9*w4
f4 = 0.4*w2 - 0.9*w4
f5 = 1 - w1 - w2 - w3 - w4
sy.solve([f1,f2,f3,f4,f5])
```

ここからこのマルコフ情報源のエントロピーを計算しよう。エントロピーは式（1.5）のようにあらわす事ができる。状態 q_i になる確率 p_i が n 個あった場合は以下のようになる。

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

秋の事例の場合、状態が q_1 のときに出現する値は 0 と 1 であり、それぞれの出現確率が 0.2 と 0.8 なので

$$H_1 = -0.2 \log 0.2 - 0.8 \log 0.8 = 0.7219$$

同様にして

$$H_2 = -0.6 \log 0.6 - 0.4 \log 0.4 = 0.9710$$
$$H_3 = -0.5 \log 0.5 - 0.5 \log 0.5 = 1$$
$$H_4 = -0.9 \log 0.9 - 0.1 \log 0.1 = 0.4690$$

これはそれぞれの状態のエントロピーである。情報源全体のエントロピーは、すべての状態についての期待値をとればよいので

$$\begin{aligned} H &= w_1 H_1 + w_2 H_2 + w_3 H_3 + w_4 H_4 \\ &= 0.2036 \times 0.7219 + 0.3258 \times 0.9710 + 0.3258 \times 1 + 0.1448 \times 0.4690 = 0.8570 \end{aligned}$$

つまり、マルコフ情報源のエントロピーは以下で定義できる。

定義 3.1. 定常分布 $\omega = (w_1, w_2, \cdots, w_k)$ が存在するマルコフ情報源のエントロピー H は、生起する要素であるアルファベット空間を $A = \{a_1, a_2, \cdots, a_l\}$ とし、起こりえる状態を $(q_1, q_2, \cdots, q_k)$ とすると、条件付き発生確率は $p(a_j|q_i)$ とあらわす事ができるので、以下のように定義される。

$$\begin{aligned} H &= \sum_{i=1}^{k} w_i H_1 \\ &= -\sum_{i=1}^{k} w_i \sum_{j=1}^{l} p(a_j|q_i) \log p(a_j|q_i) \end{aligned} \tag{3.3}$$

4　最尤推定

最尤推定法 (Maximum Likelihood Estimation: MLE) とは、最も『尤もらしい』＝最尤なパラメータを推定する方法。統計的仮設検定（Statistical Hypothesis Testing）との違いは、統計的仮説検定が「事前に立てた仮説の真偽を評価する」のに対して、最尤推定は「確率分布を前提としその確率分布のパラメータを推定する」という考え方である。以下の事例を元に考えてみる。

4.1　コイン投げの事例

> コインが 1 枚ある。このコインはどうもイカサマなコインらしく、表の出る確率が 1/2 でないらしい。このコインの表の出る確率を調べるために、コインを 3 回投げたところ、2 回表が出た。さて、このコインの表が出る確率はいくつだろうか？

普通に考えれば、3 回投げて 2 回表が出たのだから $2/3 = 66.7\%$ だろうと推定される。実際に最尤推定法でも同じ答えがでる。ここではあえて最尤推定の考え方で導いてみる。

このコインが持っている「表が出る確率」を未知のパラメータ θ とする。今、「3 回投げたら 2 回表が出る」という現象が起こった場合、この θ が幾つであると考えるのが「最も尤もらしいか？」という事を考えてみる。この確率は θ の関数であると考えられる。なので、未知のパラメータを θ として、「表、表、裏」というデータが得られる確率を $L(\theta)$ と書くと、表が出る確率が θ なら、裏が出る確率は $(1-\theta)$ なので、$L(\theta)$ は、

$$L(\theta) = \theta \cdot \theta \cdot (1-\theta) = \theta^2(1-\theta)$$

と表す事ができる。この $L(\theta)$ を尤度関数と呼ぶ。試しに、$\theta = 0.1$ であったり、$\theta = 0.9$ の場合の $L(\theta)$ を計算してみると

$$L(0.1) = 0.1^2 \cdot 0.9 = 0.009 = 0.9\%$$
$$L(0.9) = 0.9^2 \cdot 0.1 = 0.081 = 9.1\%$$

この θ と $L(\theta)$ との関係をグラフに表すと図 17 のようになる。

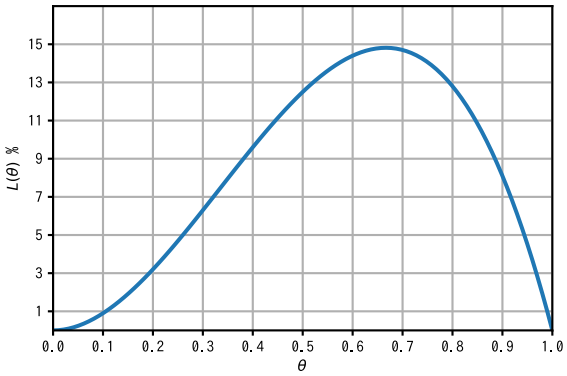


図 17　事例の最尤推定グラフ

この $L(\theta)$ を微分して 0 となるところを探せば、それが尤度 $L(\theta)$ を最大化する値である。 ここで、$L(\theta) = \theta^2(1-\theta) = \theta^2 - \theta^3$ なのでこれを微分すると

$$\frac{d}{d\theta}L(\theta) = 2\theta - 3\theta^2 = \theta(2-3\theta)$$

微分した結果がゼロになるのは $\theta = 0$ または $\theta = 2/3$。$\theta = 0$ の場合は尤度 $L(\theta)$ もゼロとなるので、尤度 $L(\theta)$ を最大にする値は $\theta = 2/3$。これは最初の「3 回投げて 2 回表が出たのだから、このコインの表の出る確率は $2/3 = 66.7\%$ だろう」という推定と同じ結果である。

4.2　二項分布の最尤推定値

先の事例を一般化する。先の事例は二項分布の最尤推定値を求めている事と同じ。二項分布（binomial distribution）は、結果が成功か失敗のいずれかである試行（ベルヌーイ試行と呼ばれる）を独立に n 回行ったときの成功回数を確率変数とする離散確率分布で、$B(n, \theta)$ と表される。

このベルヌーイ試行において、ある事象（試行回数 n と成功回数 k）が特定された時、「この現象が起こる尤もらしさ」を求める。この「尤もらしさ」（最尤推定値）は、未知のパラメータである成功確率 θ をによって変化するので、θ の関数である。これを尤度関数と呼び、以下の $L(\theta)$ のように表す事ができる

$$L(\theta) = {}_n\mathrm{C}_k \theta^k (1-\theta)^{n-k} \tag{4.1}$$

$$\begin{aligned} n: &\quad \text{試行回数} \\ k: &\quad \text{成功回数} \end{aligned}$$

$${}_n\mathrm{C}_k = \frac{n!}{k!(n-k)!} = \binom{n}{k}$$

これを微分してゼロとなる値を調べれば、最尤推定値が得られるのだが、この微分が複雑なので対数をとった対数尤度関数と呼ばれる関数

$$l(\theta) = \log L(\theta)$$

を微分する[*14]。対数をとる事によって積を和に、商を差にして演算を簡易化できる。

まずは $l(\theta)$ を展開すると

$$\begin{aligned} l(\theta) &= \log\left\{ {}_nC_k\theta^k(1-\theta)^{n-k} \right\} \\ &= \log\left\{ \frac{k!}{n!(n-k)!}\theta^k(1-\theta)^{n-k} \right\} \\ &= \log(k!) - \log\{n!(n-k)!\} + \log\theta^k + \log(1-\theta)^{n-k} \\ &= \log(k!) - \log(n!) - \log(n-k)! + \log\left(\theta^k\right) + \log(1-\theta)^{n-k} \end{aligned}$$

この $\log(k!) - \log(n!) - \log(n-k)!$ は θ と関係しない定数で、θ について微分するとゼロとなるので、

$$\begin{aligned} \frac{d}{d\theta}l(\theta) &= \frac{d}{d\theta}\left\{\log\left(\theta^k\right) + \log(1-\theta)^{n-k}\right\} \\ &= k\cdot\frac{d}{d\theta}\log\theta + (n-k)\cdot\frac{d}{d\theta}\log(1-\theta) \end{aligned}$$

ここで

$$\frac{d}{d\theta}\log\theta = \frac{1}{\theta} \quad , \quad \frac{d}{d\theta}\log(1-\theta) = -\frac{1}{(1-\theta)} \qquad \text{合成関数の微分を用いる}[*15]$$

である事から

$$\begin{aligned} \frac{d}{d\theta}l(\theta) &= \frac{k}{\theta} - \frac{(n-k)}{(1-\theta)} \\ &= \frac{k(1-\theta) - (n-k)\theta}{\theta(1-\theta)} \\ &= \frac{k-n\theta}{\theta(1-\theta)} \end{aligned}$$

これがゼロになるという事から、分子＝0 とおいて

$$\begin{aligned} k - n\theta &= 0 \\ \theta &= \frac{k}{n} \end{aligned}$$

確かに、「n 回の試行で k 回成功した場合、この成功確率はいくつか？」という問題において、最も「尤もらしい」という推定値は k/n となる。

[*14] 自然対数を底とする対数関数は単調増加の 1 対 1 対応している関数なので、対数を最大化する θ は、そのまま元の関数 $L(\theta)$ を最大化する

[*15] $(1-\theta)$ の微分をするために、$(1-\theta) = t$ とおいて合成関数の微分公式から

$$\begin{aligned} \frac{d}{d\theta}\log(1-\theta) &= \frac{d}{dt}\log t\cdot\frac{d}{d\theta}(1-\theta) \\ &= \frac{1}{t}\cdot(-1) = -\frac{1}{(1-\theta)} \end{aligned}$$

4.3 尤度関数の一般化

さらに二項分布によって得られる事象の確率に対する尤度関数を一般化する事を考えてみる。さきの事例では、「本来は未知のパラメータ（表が出る確率）」が θ であるという条件の元で、「表が出る確率」θ と「裏が出る確率」$(1-\theta)$ を利用して、「3 回投げて 2 回表が出た」という現象の最尤推定値を

$$L(\theta) = \theta \cdot \theta \cdot (1-\theta)$$

とした。この式は「表」「表」「裏」という三回の試行での確率を掛け合わせたモノだが、n 回の試行での確率を掛け合わせたものとして一般化していく。

【尤度関数の定義】

n 個の確率変数 $x_1, x_2, \cdots, x_n$ は互いに独立で，同じパラメータ θ をもった確率分布から得られたとする。ここで、パラメータが θ であるという条件の元で x_i というデータが得られる確率を $f(x|\theta)$ と表すとすると、尤度関数 $L(\theta)$ は以下のように表す事ができる。

$$L(\theta) = f(x_1|\theta) \cdot f(x_2|\theta) \cdot \cdots \cdot f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{4.2}$$

【対数尤度関数の定義】

この尤度関数は積和 $(\prod)$ で定義されており微分計算が複雑。そのため、尤度関数の対数をとった以下の対数尤度関数 $l(\theta)$ で考えるのが最尤法による推定を考える定石。

$$\begin{aligned} l(\theta) &= \log L(\theta) \\ &= \log f(x_1|\theta) + \log f(x_2|\theta) + \cdots + \log f(x_n|\theta) \\ &= \sum_{i=1}^{n} \log f(x_i|\theta) \end{aligned} \tag{4.3}$$

【対数尤度関数の微分】

対数尤度関数を微分したものが下式。

$$\frac{d}{d\theta} l(\theta) = \sum_{i=1}^{n} \frac{1}{f(x_i|\theta)} \cdot \frac{d}{d\theta} f(x_i|\theta) \tag{4.4}$$

■対数尤度関数の微分の確認　対数尤度関数の定義式 (4.3) の微分が式 (4.4) となる事を確認しよう。合成関数の微分公式[*16] を使う。まずは、$t = f(x_i|\theta)$ とおいた時の $\log f(x_i|\theta)$ の微分は

$$\begin{aligned}\frac{d}{d\theta}\log f(x_i|\theta) &= \frac{d}{dt}\log t \cdot \frac{dt}{d\theta} \\ &= \frac{1}{t}\cdot\frac{d}{d\theta}t\end{aligned}$$

ここで、t を戻せば

$$\frac{d}{d\theta}\log f(x_i|\theta) = \frac{1}{f(x_i|\theta)}\cdot\frac{d}{d\theta}f(x_i|\theta)$$

対数尤度関数は、$\log f(x_i|\theta)$ を $i = 1$ から n まで加算する。加算において微分の線形性（加えたものの微分は微分したものを加えるのと同じ）が成り立つので

$$\frac{d}{d\theta}l(\theta) = \sum_{i=1}^{n}\frac{1}{f(x_i|\theta)}\cdot\frac{d}{d\theta}f(x_i|\theta)$$

■二項分布の事例の確認　次に、先の事例「3 回投げて 2 回表が出た」をこの式に当てはめて計算してみよう。3 回の試行で「表」「表」「裏」という結果が得られたとすると

- 1 番目は「表」：$f(x_1|\theta) = \theta$ であり、$\dfrac{d}{d\theta}f(x_1|\theta) = 1$
- 2 番目は「表」：$f(x_2|\theta) = \theta$ であり、$\dfrac{d}{d\theta}f(x_2|\theta) = 1$
- 3 番目は「裏」：$f(x_3|\theta) = (1-\theta)$ であり、$\dfrac{d}{d\theta}f(x_2|\theta) = -1$

これを対数尤度関数の微分式 (4.4) に代入して

$$\begin{aligned}\frac{d}{d\theta}l(\theta) &= \sum_{i=1}^{n}\frac{1}{f(x_i|\theta)}\cdot\frac{d}{d\theta}f(x_i|\theta) \\ &= \frac{1}{f(x_1|\theta)}\cdot\frac{d}{d\theta}f(x_1|\theta) + \frac{1}{f(x_2|\theta)}\cdot\frac{d}{d\theta}f(x_2|\theta) + \frac{1}{f(x_3|\theta)}\cdot\frac{d}{d\theta}f(x_3|\theta) \\ &= \frac{1}{\theta}\cdot 1 + \frac{1}{\theta}\cdot 1 + \frac{1}{(1-\theta)}\cdot -1 \\ &= \frac{2}{\theta} - \frac{1}{(1-\theta)} \\ &= \frac{2-3\theta}{\theta(1-\theta)}\end{aligned}$$

この微分がゼロとなる値が最尤推定値となるので、分子=ゼロとなる θ の値を求めると、$2-3\theta = 0$ より $\theta = 2/3$ となる。

[*16] 合成関数の微分公式 (**??**) は以下。

$$\frac{df}{dx} = \frac{df}{dg}\cdot\frac{dg}{dx}$$

4.4 正規分布の最尤推定値

平均 μ で分散 σ の正規分布の式は

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\dfrac{(x-\mu)^2}{2\sigma^2}} \tag{4.5}$$

いまこの正規分布から $x_1, x_2, \cdots, x_n$ のような n 個のデータが独立に生成された場合を考える。この時、尤度関数は以下の式のように表すことができる。

$$\begin{aligned} L(\mu,\sigma;x_1,x_2,\cdots,x_n) &= L(\mu,\sigma;x_1)\times L(\mu,\sigma;x_2)\times\cdots\times L(\mu,\sigma;x_n) \\ &= \prod_{k=1}^{n} L(\mu,\sigma;x_k) \\ &= \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\dfrac{(x_k-\mu)^2}{2\sigma^2}} \end{aligned}$$

また対数尤度関数は、積が和に変わるので

$$\begin{aligned} l(\mu,\sigma;x_1,x_2,\cdots,x_n) &= \log\left(L(\mu,\sigma;x_1,x_2,\cdots,x_n)\right) \\ &= n\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum_{k=1}^{n}\log\left(e^{-\frac{(x_k-\mu)^2}{2\sigma^2}}\right) \\ &= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2 \end{aligned} \tag{4.6}$$

■平均の最尤推定値　正規分布の平均の最尤推定値を求めよう。上記の対数尤度関数（式 4.6）を μ で偏微分すると

$$\begin{aligned} \frac{\partial}{\partial\mu}l(\mu) &= \frac{\partial}{\partial\mu}\left(-\frac{n}{2}\log\left(2\pi\sigma^2\right)\right) - \frac{\partial}{\partial\mu}\left[\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2\right] \\ &= -\frac{\partial}{\partial\mu}\left[\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2\right] \\ &= -\frac{1}{2\sigma^2}\left[\frac{\partial}{\partial\mu}(x_1-\mu)^2 + \frac{\partial}{\partial\mu}(x_2-\mu)^2 + \cdots + \frac{\partial}{\partial\mu}(x_n-\mu)^2\right] \\ &= -\frac{1}{2\sigma^2}\left[-2(x_1-\mu) - 2(x_2-\mu) - \cdots - 2(x_n-\mu)\right] \\ &= \frac{1}{\sigma^2}\sum_{k=1}^{n}(x_k-\mu) \end{aligned}$$

この値を 0 とおいて平均 μ について解くと

$$\frac{1}{\sigma^2}\sum_{k=1}^{n}(x_k-\mu)=0$$
$$\sum_{k=1}^{n}(x_k-\mu)=0$$
$$\sum_{k=1}^{n}x_k-n\mu=0$$

より

$$\mu=\frac{1}{n}\sum_{k=1}^{n}x_k$$

となり、最尤推定値は得られたデータ $x_1, x_2, \cdots, x_n$ の平均に他ならない。

■標準偏差の最尤推定値　次に正規分布の標準偏差の最尤推定値を求めよう。上記の対数尤度関数（式 4.6）を σ で偏微分すると

$$\begin{aligned}
\frac{\partial}{\partial\sigma}l(\sigma) &= \frac{\partial}{\partial\sigma}\left(-\frac{n}{2}\log\left(2\pi\sigma^2\right)\right)-\frac{\partial}{\partial\sigma}\left[\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2\right]\\
&= \frac{\partial}{\partial\sigma}\left(-\frac{n}{2}\log(2\pi)\right)+\frac{\partial}{\partial\sigma}\left(-\frac{n}{2}\log\left(\sigma^2\right)\right)-\frac{\partial}{\partial\sigma}\left[\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2\right]\\
&= \frac{\partial}{\partial\sigma}\left(-\frac{n}{2}\log(2\pi)\right)+\frac{\partial}{\partial\sigma}\left(-n\log(\sigma)\right)-\frac{\partial}{\partial\sigma}\left[\frac{1}{2}\sigma^{-2}\sum_{k=1}^{n}(x_k-\mu)^2\right]\\
&= -\frac{n}{\sigma}-\left[\frac{1}{2}(-2)\sigma^{-3}\sum_{k=1}^{n}(x_k-\mu)^2\right]\\
&= -\frac{n}{\sigma}+\frac{\sum_{k=1}^{n}(x_k-\mu)^2}{\sigma^3}
\end{aligned}$$

この式の値を 0 とおいて σ について解くと

$$\frac{\sum_{k=1}^{n}(x_k-\mu)^2}{\sigma^3}=\frac{n}{\sigma}$$

両辺に $\frac{\sigma^3}{n}$ をかけて

$$\sigma^2=\frac{\sum_{k=1}^{n}(x_k-\mu)^2}{n}$$
$$\sigma=\sqrt{\frac{\sum_{k=1}^{n}(x_k-\mu)^2}{n}}=\sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_n-\mu)^2}{n}}$$

標準偏差 σ の最尤推定量が導出できた。同時に分散 σ^2 の最尤推定量が標本分散であることがわかる[*17]。

[*17] 標本分散をそのまま σ^2 の推定量とするのは不偏性の観点から問題があり、通常は σ^2 の推定量には不偏分散 (標本分散を $\frac{n}{n-1}$ 倍したもの) を用いる

5　ベイズ統計

5.1　用語の準備

■確率　まずは確率とは何かというと、「不確かさの量」である。以下のような定義がある。

定義 5.1. ラプラスの確率の定義
ある試行で起こりうる結果（事象）が有限個で、すべて等しく起こりやすいと考えられるとき、ある事象 A の確率 $P(A)$ は以下で表される。

$$P(A) = \frac{\text{事象 } A \text{ が起こる場合の数}}{\text{全ての起こりうる場合の数}}$$

実際にデータの観測に基づく定義として、頻度論での定義がある。

定義 5.2. 頻度論での確率の定義
頻度論の確率とは「確率とは、同じ試行を無限に繰り返したときに、ある事象が起こる割合の極限」という考え方で、ある事象 A が n 回の試行で k 回起こったときの確率を以下のように定義する。

$$P(A) = \lim_{n\to\infty} \frac{k}{n}$$

しかし、ラプラスの定義でのように有限や数えられる場合の「場合の数で割る」だけでは、連続的な現象を正しく扱えないため、以下のような公理が出来た。

公理 5.1. コルモゴロフの確率の公理

(1)	$0 \le p(x) \le 1$	確率は非負で 1 以下
(2)	$\sum_x p(x) = 1\,,\quad \int_{-\infty}^{\infty} p(x)dx = 1$	全事象の確率を足し合わせると 1
(3)	$p(A \cup B) = p(A) + p(B)$	確率は足し算できる

■同時確率　2 つの事象 A、B を考えた時、図 18 のようにこの二つの事象が同時に起こる事象を $A \cap B$ と表し、この事象 $A \cap B$ が起こる確率を同時確率といい $P(A \cap B)$ と表す。

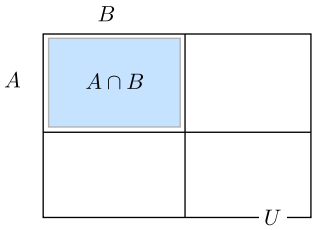


図 18　同時確率

■条件付確率　ある事象 A が起こったという条件のもとで事象 B が起こる確率を条件付確率といい $P(B|A)$ と表す。条件付確率 $P(B|A)$ は、図 19 のように、事象 A を全体と考えたときに、事象 B が起こる確率の事をさし、以下の式（5.1）と式（5.2）によって計算できる。

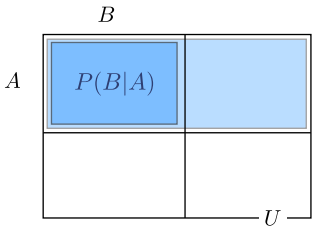


図 19　条件付確率

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{5.1}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{5.2}$$

■乗法定理　先の式（5.1）の両辺に $P(A)$ をかけたものが確率の乗法定理と言われるもので、式（5.1）と式（5.2）から、式（5.3）と式（5.4）が導出される。

$$P(A \cap B) = P(A)\ P(B|A) \tag{5.3}$$

$$P(A \cap B) = P(B)\ P(A|B) \tag{5.4}$$

■事象の独立性　事象 A と事象 B が関連しない場合、つまり事象 A が起こっても起こらなくても、事象 B の起こる確率には何も影響しない場合を**独立**とい、関連する場合は**従属**という。つまり、

$$P(B|A) = P(B|\bar{A}) = P(B) \tag{5.5}$$

事象 A と事象 B が独立の場合は、式（5.1）と式（5.2）は

$$P(B|A) = P(B) \tag{5.6}$$

$$P(A|B) = P(A) \tag{5.7}$$

となるので、式（5.3）と式（5.4）も以下のようになる。一般にこれが事象の独立の判定式で、同時発生確立が周辺確率の掛け算になっている事が独立の条件である。

$$P(A \cap B) = P(A)\ P(B) = P(B)\ P(A) \tag{5.8}$$

例題 5.1. サイコロを振って出た目を X とする

1. 「X が 3 で割り切れる事」と「X が偶数である事」は独立か？
2. 「X が素数である事」と「X が偶数である事」は独立か？

実際に確認するには式 (5.8) に当てはめればよい。

1. の解答：独立

   3 で割り切れる目は $\{3,6\}$。偶数は $\{2,4,6\}$。3 で割り切れる目でかつ偶数は $\{6\}$ のみなので $P(B \cap A) = \frac{1}{6}$、$P(A) = \frac{1}{3}$、$P(B) = \frac{1}{2}$ であり、$P(A \cap B) = P(A)\ P(B)$ が成立するので独立。

   一見すると、$A \cap B = \{6\}$ で、事象 A と事象 B とは同じ要素を共有しており独立でないように勘違いする場合があるが、独立というのは事象と事象の確率の関係であり、分割表の個数を見ると 1 行目と 2 行目が 1 : 2、1 列名と 2 列目が 1 : 1 と同じ個数の比率（＝確率）になっている。

	B	$\bar{B}$	
A	6	3	6,3
$\bar{A}$	2,4	1,5	1,2,4,5
	2,4,6	1,3,5	

2. の解答：独立でない

   素数は $\{2,3,5\}$。偶数は $\{2,4,6\}$。分割表より、$P(A \cap B) = \frac{1}{6}$、$P(A) = \frac{1}{2}$、$P(B) = \frac{1}{2}$ であり、$P(A \cap B) = P(A)\ P(B)$ が成立しないので、独立ではない。

   上の分割表は行の 1 行目と 2 行目の比率が 1 列目でも 2 列目でも 1 ： 2 となっている。一方この分割表は 1 行目と 2 行目の比率が 1 列目は 1 ： 2 で、2 列目は 2 ： 1 となっており交絡している事が判る。

	B	$\bar{B}$	
A	2	3,5	2,3,5
$\bar{A}$	4,6	1	1,4,6
	2,4,6	1,3,5	

5.2　ベイズの定理

■ベイズの定理の式の導出　先の式（5.3）と式（5.4）を再掲する。

$$P(A \cap B) = P(A)\ P(B|A)$$
$$P(A \cap B) = P(B)\ P(A|B)$$

より左辺は同じ $P(A \cap B)$ であり以下の式が成立する。

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

この式を変形した以下の式がベイズの定理 (Bayes) と言われる式である。

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{5.9}$$
$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{5.10}$$

たとえば式（5.9）を言葉で表現すると

$$B\ のもとで\ A\ が起こる確率 = \frac{A\ のもとで\ B\ の起こる確率 \times A\ の起こる確率}{B\ の起こる確率}$$

右辺の分子は $A \cap B$ に過ぎず少し回りくどい表現だが、図 20 のように左上の $P(A \cap B)$ を導出するに当たって、縦に見た場合と横に見た場合の関係から、A と B との役割を変換しているというイメージで捉えればよい。

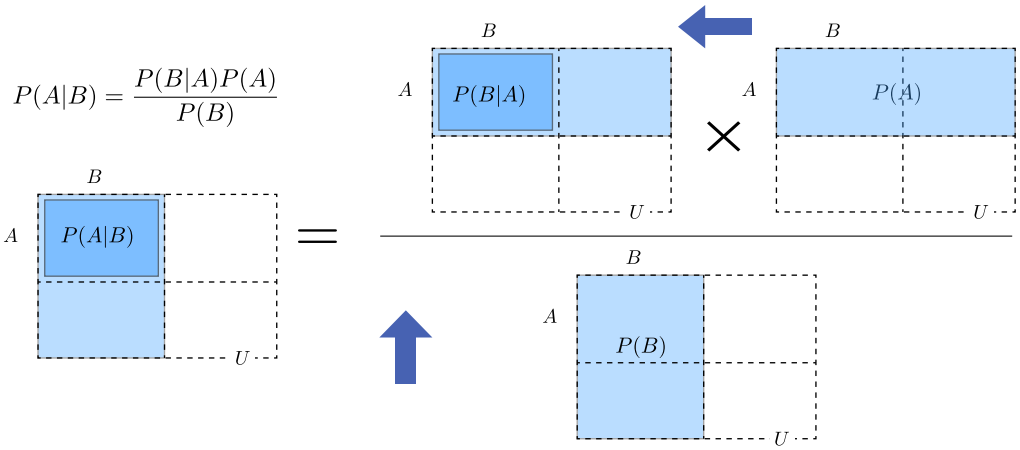


図 20　ベイズの定理のイメージ

■因果関係を調べる式として解釈する

ベイズの定理がとても有効なのは、これを因果関係の式として解釈でき、起こった現象からその原因となった仮説についての確率を計算するのに使える事である。例えば、A を仮説 or 原因 (Hypoethesis)、B をデータ or 結果 (Data) として解釈してみよう。

定義 5.3. 【ベイズの基本公式】
ある結果 D（$Data$）が得られた時、その結果データが原因仮説 H $(Hypoethesis)$ によって起こったと考えられる確率は以下のように定義される。

$$P(H|D) = \frac{P(D|H)\ P(H)}{P(D)} \tag{5.11}$$

これらには以下のような名前がついている。

$$\overbrace{P(H|D)}^{\text{事後分布}} = \frac{\overbrace{P(D|H)}^{\text{尤度関数}}\ \overbrace{P(H)}^{\text{事前分布}}}{\underbrace{P(D)}_{\text{エビデンス}}}$$

また、原因となる仮説 H は、図 21 のように複数存在する方が一般的である。

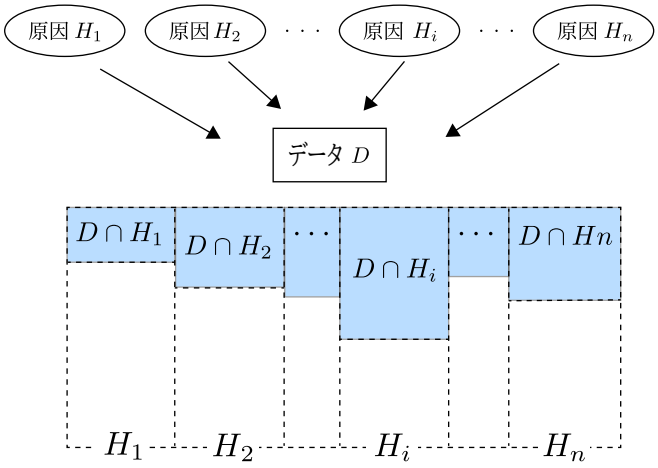


図 21　原因が複数ある場合

その場合、i 番目の原因仮説 H_i によってデータ D が引き起こされる確率を以下のように表現できる。

$$P(H_i|D) = \frac{P(D|H_i)\ P(H_i)}{P(D)} \quad (\text{i=1,2,}\cdots\text{,n}) \tag{5.12}$$

この式の $P(D)$ は原因仮説 $H_1, H_2, \cdots, H_n$ の元でのデータ D の確率の和であり、確率の乗法公式[*18]を用いて変形すると

$$\begin{aligned} P(D) &= P(D \cap H_1) + P(D \cap H_2) + \cdots + P(D \cap H_n) \\ &= P(D|H_1)P(H_1) + P(D|H_2)P(H_2) + \cdots + P(D|H_n)P(H_n) \end{aligned}$$

この結果を上記の式（5.12）に代入する事で以下のベイズの展開公式が得られる。

[*18] 式（5.3）と式（5.4）

定義 5.4. 【ベイズの展開公式】
データ D は原因 $H_1, H_2, \cdots, H_n$ のどれかひとつによって引き起こされると仮定する。いまデータ D が得られた場合、その原因が仮説 H_i である確率は以下のように定義される。

$$P(H_i|D) = \frac{P(D|H_i)\ P(H_i)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2) + \cdots + P(D|H_n)P(H_n)}$$
$$= \frac{P(D|H_i)\ P(H_i)}{\sum_{i=1}^{n} P(D|H_i)P(H_1)} \tag{5.13}$$

この展開公式を使った例題をあげておく。

【例題】　飛行機事故の調査によれば、その原因は「操縦ミス」「整備不良」「管制ミス」などがある。ここでは原因をこの３つであるとし、それぞれの発生確率 $P(H_i)$ と事故の原因になる確率 $P(D|H_i)$ が以下の表のような確率であるとする。

	発生確率 $P(H_i)$	事故につながる確率 $P(D|H_i)$
操縦ミス (H_1)	0.6	0.02
整備不良 (H_2)	0.3	0.03
管制ミス (H_3)	0.1	0.01

いま事故が起こったとして、その事故が上記の３つの原因による可能性がどの程度かを計算する。

この例題を、横軸に「事象の発生確率 $P(H_i)$」、縦軸に「その事象が発生した時に事故につながる確率 $P(D|H_i)$」をとって、図で表すと以下のようになる。

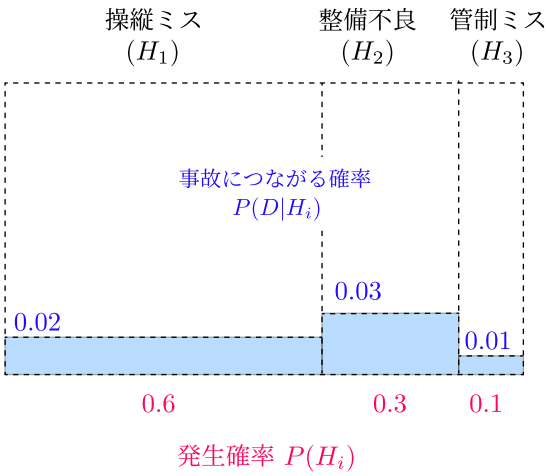


図 22　飛行機事故の原因推定

そして、それぞれの原因仮説 (H_1, H_2, H_3) によってこの事故が起こった可能性がどの程度かを計算すると以下のようになり、「操縦ミス (H_1) である可能性が最も高い事になる。

$$P(H_1|D) = \frac{0.02 \times 0.6}{0.02 \times 0.6 + 0.03 \times 0.3 + 0.01 \times 0.1} = 0.545$$
$$P(H_2|D) = \frac{0.03 \times 0.3}{0.02 \times 0.6 + 0.03 \times 0.3 + 0.01 \times 0.1} = 0.409$$
$$P(H_3|D) = \frac{0.01 \times 0.1}{0.02 \times 0.6 + 0.03 \times 0.3 + 0.01 \times 0.1} = 0.045$$

このようにベイズの基本公式は、ある事象が生じたというデータを入手した時に、その事象の原因が何である可能性が高いかを計算するのに有効である。

■ベイズ理論を理解する３つのキーワード　ベイズの展開公式（式 5.13）には４つの確率 $P()$ が含まれているが、その中の３つの確率には名称がつけられている。図 23 のように「事後確率」「尤度」「事前確率」の３つである。

$$\text{事後確率}\ P(H_i|D) = \frac{\text{尤度}\ P(D|H_i)\ \text{事前確率}\ P(H_i)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2) + \cdots + P(D|H_n)P(H_n)}$$

図 23　ベイズ理論を理解する３つのキーワード

この３つのキーワードの意味は表 4 のようになっている。例えば先の飛行機事故の事例でいえば、**「事後確率」**とは「事故があったという事を知ったときに、その原因を調べる前に原因は何であるかを予測」する事であり、**「尤度」**とは言葉の意味からいって「もっともらしさの程度」の事で「どの原因がもっともらしいか」を意味する。ここでは「操縦ミスや整備不良が生じたら普通どの程度の確率で事故に結び付くかを示す」ものである。そして**「事前確率」**とは、その仮説が生じる確率で、「そもそも原因となった操縦ミスや整備不要などのミスが生じる一般的な確率」を意味している。

表 4　ベイズ理論を理解する３つのキーワード

確率記号	名称	意味
$P(H_i\|D)$	事後確率	データ D が得られた時の原因が H_1 である確率
$P(D\|H_i)$	尤度	原因 H_i が生じた場合に結果 D が生じる確率
$P(H_i)$	事前確率	データ D を得る前に持っていた原因 H_i が生じる確率

5.3　ベイズ更新

ベイズ更新とは、得られたデータをもとに事前確率を更新していく事であり、データによって学習していく過程をモデル化できる。このベイズ更新を説明するにあたって、例題を準備する。

【例題】　図 24 のように赤玉と青玉が混ざって入っている壺が 3 つあるとする。3 つの壺の玉の総数は同じだが、壺 1 には赤玉：青玉の比が 1 ： 2 で、壺 2 には 2 ： 1 で、壺 3 には 3 ： 0 の比率で混ざっているとする。これら 3 つの壺の一つを選び、さらに選んだ壺から一つ玉を取り出した時、それが赤玉であった。赤玉であったと知った時に、どの壺が最初に選ばれた可能性が高いかの確率を求めよう。ただし、3 つの壺が選ばれる確率は順に　壺 1 ：壺 2 ：壺 3 = 3 ： 2 ： 1 であるとする。

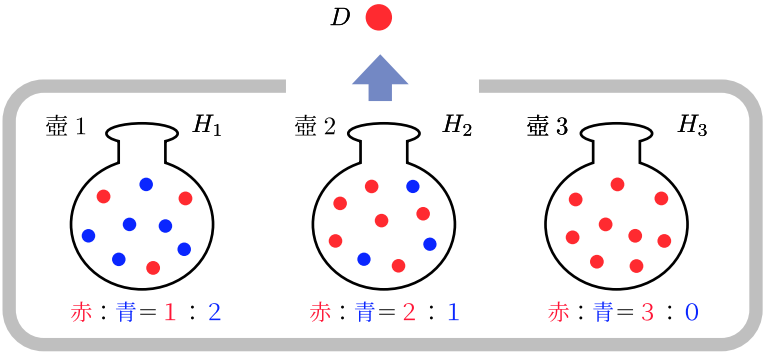


図 24　壺と玉の問題

まずそれぞれの壺が選ばれた（H_1, H_2, H_3）とした時に、赤玉がでる D という確率は、それぞれの壺の赤玉と青玉の比率のから

$$P(D|H_1) = \frac{1}{3}, \quad P(D|H_2) = \frac{2}{3}, \quad P(D|H_3) = \frac{3}{3}$$

また、それぞれの壺が選ばれる確率は、文章中の定義から

$$P(H_1) = \frac{3}{6}, \quad P(H_2) = \frac{2}{6}, \quad P(H_3) = \frac{1}{6}$$

これらを以下のベイズの展開公式式（5.13）に代入すれば

$$P(H_3|D) = \frac{P(D|H_3)\ P(H_3)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2) + P(D|H_3)P(H_3)}$$
$$= \frac{\frac{3}{3} \times \frac{1}{6}}{\frac{1}{3} \times \frac{3}{6} + \frac{2}{3} \times \frac{2}{6} + \frac{3}{3} \times \frac{1}{6}} = \frac{\frac{1}{6}}{\frac{10}{18}} = \frac{3}{10}$$

同様に計算すると

$$P(H_1|D) = \frac{3}{10}, \quad P(H_2|D) = \frac{4}{10}, \quad P(H_3|D) = \frac{3}{10}$$

このように「赤玉が得られた」というデータで可能性が最も高いのは壺２であるというのが結論になる。何故、赤玉の比率が最も多い壺３でないのかというと、それぞれの壺が選ばれる事前確率が「壺１：壺２：壺３＝３：２：１」という条件がある為である。

■理由不十分の原則　つぎに、この事前確率の条件がなかった場合の事を考えてみる。「何も情報がなければ起こる事象の確率は同等」という発想ですすめる。これを**「理由不十分の原則」**と呼ぶ。この場合、それぞれの壺が選ばれる事前確率は、

$$P(H_1) = \frac{1}{3}, \quad P(H_2) = \frac{1}{3}, \quad P(H_3) = \frac{1}{3}$$

この事前確率を使って計算すると結果は

$$P(H_1|D) = \frac{1}{6}, \quad P(H_2|D) = \frac{1}{3}, \quad P(H_3|D) = \frac{1}{2}$$

つまり、もし事前確率がなくすべての事象が棟確率と考えるならば、赤玉が得られたというデータのもとで、もっとも高い確率なのは「壺３が選ばれていると推測する事」である。

■複数のデータで更新する　今までは赤玉が得られたというデータを得た場合について考えてきたが、次に以下のように、複数のデータが得られた場合、得られたデータによって事前確率を更新する事で確率が変化していく様子を捉えよう。

> 先の例題と同様に最初に３つの壺から一つを選び、以降は壺は選ばないとする。その上で、最初に選んだ壺の中から玉を取り出す試行を三回行ったところ、赤玉、赤玉、青玉の順番にデータが得られたとする。この場合に、どの壺が選ばれていた可能性が高いかを計算する。

いままでと同様に、以下のベイズの展開公式に当てはめて計算する

$$P(H_3|D) = \frac{P(D|H_3)\ P(H_3)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2) + P(D|H_3)P(H_3)}$$

まず、得られたデータを以下のように記号化しておく。

D_r：「赤玉が得られた」　　D_b：「青玉が得られた」

今回の例題では、$D_r \to D_r \to D_b$ という順番に３つのデータが得られた場合を考える。

１回目　赤玉 D_r　最初の事前確率は３つの壺のどれが選ばれたかについては、すべてが等確率と考える。最初に得られたのは赤玉 D_r なので、上記と一緒でそれぞれの事後確率は

$$P(H_1|D_r) = \frac{1}{6}, \quad P(H_2|D_r) = \frac{1}{3}, \quad P(H_3|D_r) = \frac{1}{2}$$

２回目　赤玉 D_r　１回目の結果を事前確率とする。つまり $P(H_1) = \frac{1}{6},\ P(H_2) = \frac{1}{3},\ P(H_3) = \frac{1}{2}$ とすると事後確率はそれぞれ

$$P(H_1|D_r) = \frac{1}{14}, \quad P(H_2|D_r) = \frac{4}{14}, \quad P(H_3|D_r) = \frac{9}{14}$$

つまり壺３が選ばれた H_3 確率が高まったことになる。

3回目　青玉 D_b 今度は出たのは青玉である。今までと同様に上記を事前確率とすると

$$P(H_1|D_r) = \frac{1}{3}, \quad P(H_2|D_r) = \frac{2}{3}, \quad P(H_3|D_r) = 0$$

青玉が出た事によって壺 3 が選ばれた H_3 確率はゼロになり、壺 2 が選ばれた H_2 確率が最も高くなった。

■ベイズ更新について定式化しておく

まずベイズの展開公式 (5.13) は以下

$$P(H_i|D) = \frac{P(D|H_i)\ P(H_i)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2) + \cdots + P(D|H_n)P(H_n)}$$

$D_1, D_2, \cdots, D_n$ とデータを得たとして、その結果から仮説 H_i の確率をベイズ更新する事を考える。その時、右辺の分母は 1 回目の更新でも n 回目の更新でも常に定数なので除いて、比例式 $\propto$ で表すと、1 回目のデータに基づく確率は

$$P(H_i \mid D_1) \propto \prod_i P\left(D_1 \mid H_i\right) P(H_i)$$

2 回目のデータからベイス更新した仮説 H_i の確率を $P(H_i \mid D_1, D_2)$ と表すと

$$\begin{aligned} P(H_i \mid D_1, D_2) \ &\propto \prod_i P\left(D_2 \mid H_i\right) \underline{P(H_i \mid D_1)} \quad \leftarrow 1 \text{ 回目から計算された事前確率} \\ &= \prod_i P\left(D_2 \mid H_i\right) \underline{P\left(D_1 \mid H_i\right) P(H_i)} \end{aligned}$$

n 回目のデータからベイス更新した仮説 H_i の確率は

$$P(H_i \mid D_1, D_2, \cdots, D_n) \ \propto \prod_i P\left(D_n \mid H_i\right) \underline{P(H_i \mid D_1, D_2, \cdots, D_{n-1})} \quad \leftarrow n-1 \text{ 回目までの結果}$$

このように、計算結果から事前確率を更新していくのがベイズ更新である。

■逐次合理性　壺を選んだ後の玉のでる順番がもし異なっていたとしたらどうなるのか。例えば「青玉⇒赤玉⇒赤玉」という順番であったとしよう。その場合も三回目終了時点で、どの壺が選ばれていたかの確率は全く同じになる。これを**「逐次合理性」**と呼ぶ。

■例題

【例題】　ベイズ更新についての例題 (1) と (2) のそれぞれの確率を求めよ。

(1) 5% の人がかかっている病気 A があります。実際に病気 A にかかっている人が検診 B を受けると 90% の確率で陽性となり、病気にかかっていない人が検診 B を受けると 70% の確率で陰性となります。さて、検診 B を受けた結果が陽性だった場合、実際に病気 A にかかっている確率はどのくらいでしょうか？

(2) この人に対して、新たに検査 C が行われることになりました。実際に病気 A にかかっている人が検診 C を受けると 80% の確率で陽性となり、病気にかかっていない人が検診 C を受けると 90% の確率で陰性となります。検診 B の結果が陽性だったあと、さらに検診 C を受けてその結果も陽性だった場合、病気 A にかかっている確率はどのくらいでしょうか。

以下のように事象を記号化する

病気仮説　H_1：病気 A にかかっている　H_2：病気 A にかかっていない
検診結果　B_1：検診 B で陽性となった　B_2：検診 B で陰性となった
検診結果　C_1：検診 B で陽性となった　C_2：検診 B で陰性となった

$$P(H_1|B_1) = \frac{P(B_1|H_1)P(H_1)}{P(B_1|H_1)P(H_1) + P(B_1|H_2)P(H_2)}$$

$$P(H_1|C_1) = \frac{P(C_1|H_1)P(H_1)}{P(C_1|H_1)P(H_1) + P(C_1|H_2)P(H_2)}$$

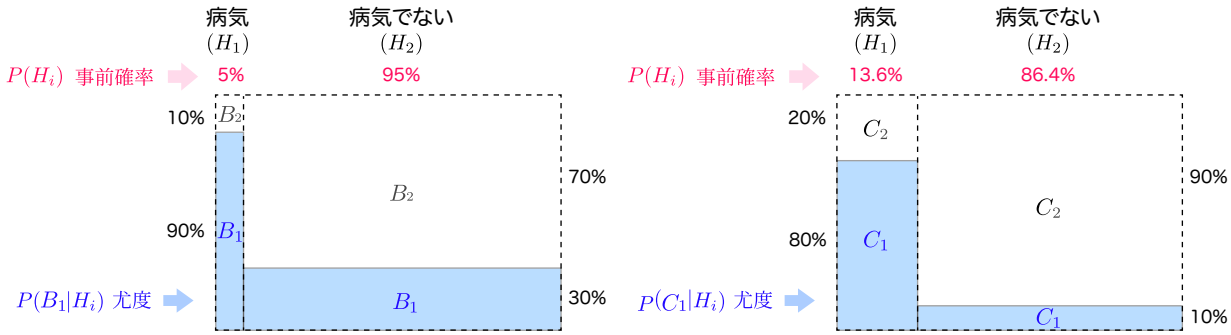


図 25　ベイズ更新の例題の図

(1) 解答

$$P(H_1|B_1) = \frac{P(B_1|H_1)P(H_1)}{P(B_1|H_1)P(H_1) + P(B_1|H_2)P(H_2)}$$

$$= \frac{0.9 \times 0.05}{0.9 \times 0.05 + 0.3 \times 0.95} = 0.136$$

(2) 解答

$$P(H_1|C_1) = \frac{P(C_1|H_1)P(H_1)}{P(C_1|H_1)P(H_1) + P(C_1|H_2)P(H_2)}$$
$$= \frac{08 \times 0.136}{0.8 \times 0.136 + 0.1 \times 0.864} = 0.58$$

このように、最初の検査 B で陽性と出た段階では、病気の確率は 13% であった。それが、二回目の検査 C でも陽性だった事で 58% まで高まった事になる。

■事前確率の重要性　計算した結果の確率が、一般的な常識からみて意外な結果になる場合がある。そういった場合の多くは事前確率が影響している。

【例題】　ある病気を発見する検査 T に関して、次の事が知られている。

- 病気にかかっている人を検査すると、98 %の確率で「病気である」と正しく判定される。
- 病気にかかっていない人を検査すると、5 %の確率で誤って「病気である」と判定される。
- 統計をとると母集団の中で、この病気にかかっている人は 3 %、かかっていない人は 97 %である。

この母集団から無作為に抽出された一人に検査 T を適用し、「病気である」と判定されたとき、この人が本当に病気にかかっている確率を求めよ。

以下のように記号化すれば、求めたい確率は $P(H_1|D)$ である。

H_1 :「この病気にかかっている」
H_2 :「この病気にかかっていない」
D :「この病気にかかっている（陽性）と判断された」

この確率を求めるには以下のベイズの展開公式に当てはめればよい

$$P(H_1|D) = \frac{P(D|H_1) \cdot P(H_1)}{P(D|H_1) \cdot P(H_1) + P(D|H_2) \cdot P(H_2)}$$

尤度の算出　与えられた文章から、検査 T の精度については以下の確率になる。

$P(D|H_1) =$ 病気の人が陽性と判断される確率 $= 0.98$
$P(D|H_2) =$ 病気でない人が陽性と判断される確率 $= 0.05$

事前確率の設定　次に事前確率も、文章の定義から以下のようになっている

$P(H_1) =$ 病気の人である確率 $= 0.03$
$P(H_2) =$ 病気でない人である確率 $= 0.97$

事後確率の計算 尤度と事前確率から計算すると

$$P(H_1|D) = \frac{0.98 \times 0.03}{0.98 \times 0.03 + 0.05 \times 0.97} = 0.377$$

「病気の人の９８％」を正しく診断する検査で、「病気である」（陽性）と判定されたとしても、本当に「病気である」確率は３８％程度という事になる。これは思ったより低い確率である。そうなった理由は事前確率の低さにある。つまり「この病気にかかる人が３％程度」で非常に低いから、たとえ判定が陽性でも病気ではないと判断される確率が高くなる。

逆に検査結果が「病気でない（陰性）」と判断された場合はどうだろうか？ 陰性であったという結果を D_2 とすると

$$P(D_2|H_1) = \text{病気の人が陰性と判断される確率} = 0.02$$
$$P(D_2|H_2) = \text{病気でない人が陰性と判断される確率} = 0.95$$

より、この人が本当は病気である（陽性）である確率は

$$\begin{aligned} P(H_1|D_2) &= \frac{P(D_2|H_1)\cdot P(H_1)}{P(D_2|H_1)\cdot P(H_1) + P(D_2|H_2)\cdot P(H_2)} \\ &= \frac{0.02\times 0.03}{0.02\times 0.03 + 0.95\times 0.97} = 0.00065 \end{aligned}$$

つまりほとんど０％であるという事であり、「病気でない（陰性）」と判定されたら確実に「病気ではない」という事になる。

5.4 ナイーブベイズフィルター

ベイズ理論がシンプルな形で適応されている事例。メール本文に出現する単語をもとに迷惑メールかどうかを判断する。単純ベイズフィルター（Naive Bayes Filter）と呼ばれる。

【例題】 迷惑メールと通常メールを調べた結果、以下の４つの単語が、迷惑メールと通常メールに下の表のような確率で含まれている事がわかった。また、迷惑メールと通常メールの比率は７：３の割合であった。

検出された単語	H_1 迷惑メール	H_2 通常メール
プレゼント	0.6	0.1
無料	0.5	0.3
統計	0.01	0.4
経済	0.05	0.5

いま受け取ったメールに、「プレゼント、無料、経済」という順番でこれらの単語が検出されたとき、このメールが迷惑メールかどうかを判定する。ただし、それぞれの単語の出現は独立である（つまりある単語が出たら他のある単語が出やすいという事はない）とする。

モデルの要素を整理して記号化していこう。まず原因仮説については以下の表のように記号化する。

原因	意味
H_1	受信メールが迷惑メールである
H_2	受信メールが通常メールである

得られたデータについては以下の表のように記号化する。

データ	意味
D_1	受信メールに「プレゼント」という単語が検出された
D_2	受信メールに「無料」という単語が検出された
D_3	受信メールに「統計」という単語が検出された
D_4	受信メールに「経済」という単語が検出された

得られたデータは順番も考慮すると、受け取ったメールの本文に、「プレゼント」→「無料」→「経済」という単語がこの順番に出現しているという事であり、それを一文字で

$$D = (D_1, D_2, D_4)$$

とあらわすとする。ここでそれぞれの単語の出現率は独立であるとしているので

$$P(D|H_1) = P(D_1, D_2, D_4|H_1) = P(D_1|H_1)P(D_2|H_1)P(D_4|H_1)$$

つまり、迷惑メールに３つの単語が順番に出てくる確率は、迷惑メールにそれぞれの単語が出現する確率の積になっている事になる。

求めたいのはデータ D が得られたときの「迷惑メールである確率 (H_1)」と「通常メールである確率 (H_2)」であるので、以下の２つの式を比較すればよい。

$$P(H_1|D) = \frac{P(D|H_1)P(H_1)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)}$$

$$P(H_2|D) = \frac{P(D|H_2)P(H_2)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)}$$

この２つの確率の分母は同じなので分子だけを比較して、迷惑メールである確率が高いか、通常メールである確率が高いかを判断すればよい。という事は結局、図 26 のように縦にかけ合わせた値の大小を比較すればよい事になる。

検出された単語　$P(H_i|D)$

$P(H_i\|D)$	H_1 迷惑メール	H_2 通常メール
D_1　プレゼント	0.6	0.1
D_2　無料	0.5	0.3
D_3　統計	0.01	0.4
D_4　経済	0.05	0.5

事前確率　$P(H_i)$

$P(H_i)$	H_1 迷惑メール	H_2 通常メール
事前確率	0.7	0.3

$P(D|H_1)P(H_1)$　　$P(D|H_2)P(H_2)$

縦にかけ合わせた値の大小を比較すると判別ができる。

図 26　迷惑メールフィルター

図 26 のように縦に掛け算すると

$$P(D|H_1)P(H_1) = 0.0105$$
$$P(D|H_2)P(H_2) = 0.0045$$

ここから、迷惑メールである確率と通常メールである確率を計算すると

$$P(D|H_1) = \frac{0.0105}{0.0105 + 0.0045} = 0.7$$
$$P(D|H_2) = \frac{0.0045}{0.0105 + 0.0045} = 0.3$$

７割の確率で「迷惑メール」として判断すべきという事になる。

5.5　ベイズ更新とシグモイド関数

事前確率をもとに条件付き確率を計算し、次にその計算結果を事前確率に代入して計算を続ける・・・というベイズ更新を続けていった時、図 27 のように横軸に試行回数、縦軸に各回の計算された確率をとると、そのグラフがシグモイド関数と同様の曲線を描くという話。

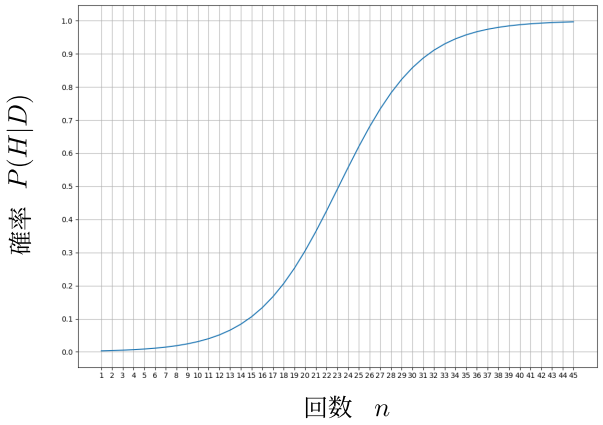


図 27　試行回数を横軸にとるとシグモイド曲線を描く

まずは事例について

【例題】　バレンタインディーにチョコをもらったとする。本命の場合にチョコをあげる確率は 0.65、義理の場合にチョコを上げる割合は 0.5 とする。事前確率は本命か義理か半々で確率 0.5 とする。この事例について、毎年チョコをもらったとし、毎年毎に事前確率を更新していった場合の「本命確率」の変化を調べよう。

	H_1 本命	H_2 義理
チョコあげた D_1	0.65	0.5
チョコあげない D_2	0.35	0.5

まず最初にチョコをもらった場合の本命確率は

$$P(H_1|D_1) = \frac{P(D_1|H_1)P(H_1)}{P(D_1|H_1)P(H_1) + P(D_1|H_2)P(H_2)}$$
$$= \frac{0.65 \times 0.5}{0.65 \times 0.5 + 0.5 \times 0.5} = 0.56522$$

毎年チョコをもらっていくとして、前の年の事後確率を新たに事前確率 $P(H_1),\quad P(H_2)$ として更新して計算

していくと

$$
\begin{aligned}
2\text{年目} &= \frac{0.65 \times 0.56522}{0.65 \times 0.56522 + 0.5 \times 0.43478} = 0.65000 \\
3\text{年目} &= \frac{0.65 \times 0.65000}{0.65 \times 0.65000 + 0.5 \times 0.35000} = 0.70712 \\
4\text{年目} &= \frac{0.65 \times 0.70712}{0.65 \times 0.70712 + 0.5 \times 0.29288} = 0.75837
\end{aligned}
$$

このように「本命である確率」はチョコをもらうたびに上昇している事になる。ちなみに、次の５年目ではチョコをもらえなかったとすると確率は以下のように下がる事になる。

$$
\begin{aligned}
P(H_1|D_2) &= \frac{P(D_2|H_1)P(H_1)}{P(D_2|H_1)P(H_1) + P(D_2|H_2)P(H_2)} \\
&= \frac{0.35 \times 0.75837}{0.35 \times 0.75837 + 0.5 \times 0.24163} = 0.68721
\end{aligned}
$$

以上のベイズ更新をシミュレーションしたのが図 27 で、最初の事前確率として本命である確率を $P(H_1) = 0.003$ とし４４回チョコをもらい続けたとして、本命確率がどのように上昇していくかをシミュレーションしたものである。

何故、このような図になるのかを調べる前に条件付き確率の式を以下のように変形しておく、右辺の分母と分子を $P(D_1|H_1)P(H_1)$ で割って

$$
\begin{aligned}
P(H_1|D_1) &= \frac{P(D_1|H_1)P(H_1)}{P(D_1|H_1)P(H_1) + P(D_1|H_2)P(H_2)} \\
&= \frac{1}{1 + \dfrac{P(D_1|H_2)P(H_2)}{P(D_1|H_1)P(H_1)}}
\end{aligned}
$$

ここで尤度の比は常に一定なので α とおいてしまう。

$$
\alpha = \frac{P(D_1|H_2)}{P(D_1|H_1)}
$$

また、$P(H_2) = 1 - P(H_1)$ なので $P(H_2) = p$ とおくと

$$
\frac{P(H_2)}{P(H_1)} = \frac{p}{1-p}
$$

以上のように、事前の本命確率を p とすると、チョコをもらったという事後の本命確率 p_{next} は

$$
p_{next} = \frac{1}{1 + \alpha \cdot \dfrac{p}{1-p}}
$$

この α は定数なので、結局次の確率 p_{next} は今の「本命確率 p」と「義理確率 $(1-p)$」との比によって決まっている事になる。

■オッズ Odds の意味　「起こる確率」÷「起こらない確率」のことをオッズと言います。つまり、確率 p で起こる事象に対して、

$$\text{オッズ} = \frac{\text{起こる確率}}{\text{起こらない確率}} = \frac{p}{1-p}$$

のことをオッズと言います。そしてオッズ比とは、オッズを用いて、ある 2 つの事象の起こりやすさを比較するものです。例えば、20 %の確率で起こることと 10 %の確率で起こることがあるとしたら、オッズはそれぞれ

$$\frac{0.2}{1-0.2} = 0.250$$
$$\frac{0.1}{1-0.1} = 0.111$$

となるので、オッズ比は、

$$\frac{0.250}{0.111} = 2.25$$

となります。オッズ比が 1 であれば、2 つの事象の間に差はないと判断できます。

■ロジット logit 変換　オッズの対数を取る事をロジット変換といいます。実は、オッズに対数を取ることで比率を等間隔にする事が可能になります。例えば、10 %の確率と 90 %の確率のオッズは

$$\frac{0.1}{1-0.1} = 0.1111$$
$$\frac{0.9}{1-0.9} = 9$$

それぞれの対数をとると

$$\log \frac{0.1}{1-0.1} = -2.197225$$
$$\log \frac{0.9}{1-0.9} = +2.197225$$

図 28 にオッズとロジットのグラフを示す。図 28 のように、オッズは確率 p が 1 に近づくにつれて大きな数になっていくのに対して、ロジット変換をすると、0〜1 という範囲で動く確率のようなデータを説明変数とする値は $(0.5, 0)$ を中心に「$-\infty$」から「$+\infty$」に広がった対称な形を描く。これは対数なので

$$y = \log \frac{p}{1-p} = \log p - \log(1-p)$$

となり $y = 0$ なら $p = (1-p)$ で $p = 0.5$ となり、これを原点に対象になる。

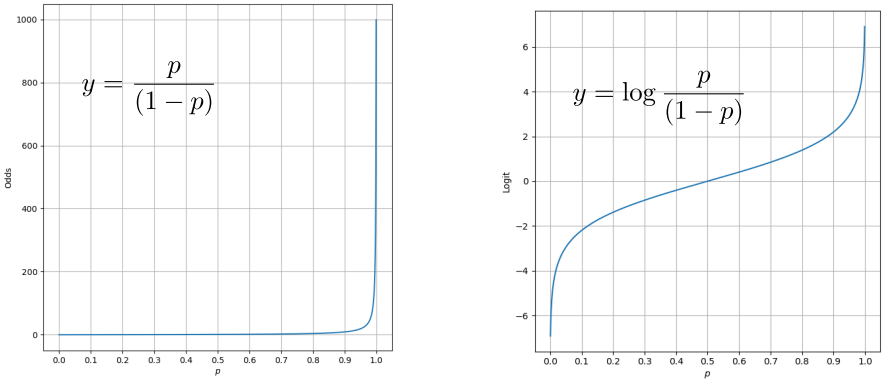


図 28　オッズとロジットのグラフ

■ロジスティック関数　　ロジット関数の逆関数がロジスティック関数。シグモイド関数とも呼ばれます。

ロジット関数は

$$y = \log \frac{p}{1-p}$$

これの逆関数を求めると

$$\begin{aligned} e^y &= \frac{p}{1-p} \\ e^y - e^y p &= p \\ p &= \frac{e^y}{e^y + 1} = \frac{1}{1 + e^{-y}} \end{aligned}$$

つまり

$$g(x) = \frac{1}{1 + e^{-x}} \tag{5.14}$$

標準ロジスティック関数は任意の数を確率に変換する関数」だと考えることができます。

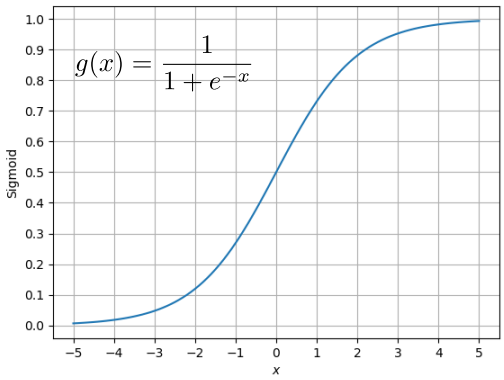


図 29　ロジスティック関数のグラフ

5.6 自然共役事前分布

共役事前分布とは、ベイズ統計を扱う際に、複雑な計算を回避するために考えられた事前分布で、共役事前分布を用いて事後分布を求めると、**事後分布** $P(\theta \mid x)$ が**事前分布** $P(\theta)$ と同じ分布になるという特性がある。

ベイズの定理は、以下のように定式化される。

$$\text{事後分布} = \frac{\text{尤度分布} \times \text{事前分布}}{\text{周辺尤度}} \qquad P(\theta \mid x) = \frac{P(x \mid \theta)P(\theta)}{P(x)}$$

このとき、「事前分布と事後分布が同じ族」になるように、**尤度と相性の良い分布**を選ぶと、計算しても 事後分布の形が事前分布と同じ族になるので計算が楽になる。こうして、尤度と組み合わせる事で、事後分布が事前分布と同じ族になるように選ばれた事前分布を共役事前分布という。

以下のようなものが共役事前分布として知られている。この共役事前分布に尤度をかけて事後分布を求めると、事後分布の形が事前分布と同じになる。

表 5　共役事前分布一覧

母数が規定する確率分布	共役事前分布	事後分布
ベルヌーイ分布	ベータ分布	ベータ分布
二項分布	ベータ分布	ベータ分布
正規分布（ σ^2 既知）	正規分布	正規分布
正規分布（ σ^2 未知）	逆ガンマ分布	逆ガンマ分布
ポアソン分布	ガンマ分布	ガンマ分布
多項分布	ディリクレ分布	ディリクレ分布

共役事前分布がどのような形の分布になるかは、データを取ってくる母集団の確率分布（これを以下『母数が規定する確率分布』とする）によって決定される。例えば、母数の規定する確率分布が二項分布の場合、事前分布をベータ分布に設定すれば、事後分布もベータ分布になる。

このように、母数が規定する確率分布に対して、適切な事前分布を持ってくれれば、事後分布は事前分布と同じ形の分布になる。すると、事前分布と事後分布が同じ形になりベイズの更新が容易になる。

■共役事前分布の事例

表が出る確率の母数を θ とした時、表が x 回出る確率は二項分布に従う。これが母数が基底する確率分布であり尤度関数となる。

$$P(x \mid \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

この事前分布として以下のベータ関数をもってくる。

$$P(\theta) = \mathrm{Beta}(\theta \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

ベイズの定理の分母の周辺尤度 $P(x)$ は定数なので、以下のように比例式 $\propto$ で表す事ができる。

$$P(\theta \mid x) \propto P(x \mid \theta)P(\theta)$$

これに二項分布とベータ分布を代入すると、結果は以下のようにベータ分布と同じ形になる。

$$P(\theta \mid x) \propto \theta^x (1-\theta)^{n-x} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}$$
$$\propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

ただし通常は、共役事前分布は表 5 のようなものから選び、そのパラメータをデータや理論に基づいて設定する事になる。

■事後分布の推定方法

事後分布を求める方法として以下がある。

(1) **自然共役事前分布**を使って、解析的に解を求める
(2) **MAP 推定**を使って、確率分布全体を求める事を諦め、代わりに事後分布が最大の点だけを求める
(3) **MCMC 法** (Markov Chain Monte Carlo) を使って、事後分布をサンプリングする

5.7 MAP 推定

MAP 推定 (MAP:Maximum A Posteriori Estimation) は前の節で述べたように確率分布全体を求める事を諦め、代わりに事後分布が最大の点だけを求める方法である。なので、この MAP 推定と最尤推定 (MLE:Maximum Likelihood Estimation) は似ていることになる。実際に、MAP 推定と最尤推定（MLE）は「考え方の構造」はほぼ同じで、事前分布があるかないかだけが異なる。なので、後で述べるように予め何も情報がない場合に事前分布を無情報事前分布として設定する場合があり、その場合は MAP 推定値が最尤推定値に一致 (ベイズ論と頻度論とが整合) する。

まず最尤推定とはなにかを復習しておこう。通常の確率分布（確率密度関数や確率質量関数）は、「あるパラメータ（例：正規分布の平均や分散、コインの表が出る確率など）が与えられたときに、データが生成される確率（または確率密度)」を表す。例えば、コインの表が出る確率が 0.5 のとき、10 回投げて 7 回表が出る確率を調べるという目的で使用する。

これに対して、最尤推定は先に「観測されたデータ」が手元にあるときに、そのデータが最も『もっとも』生成されるような、確率分布のパラメータは何かを考える。つまり以下の表のように変数と固定値が逆 j になっていることになる。この「もっともらしい度合い」を尤度（ゆうど、likelihood）と呼び $L(\theta \mid X)$ と表記する。

確率分布 $P(X \mid \theta)$	変数：	X（データ）
	固定値：	θ（パラメータ）
	意味：	パラメータ θ が既知のときに、データ X が得られる確率
尤度関数 $L(\theta \mid X)$	変数：	θ（パラメータ）
	固定値：	X（観測されたデータ）
	意味：	観測データ X が得られたときに、どのパラメータ θ が最ももっともらしい か

このように、通常母数（因）は決定していて、データが観測される（果）という因果関係を想定するが、最尤推定はこの因果関係を逆にし、データを観測した場合に母数を推定しようという考え方で、その点はベイス推定と同じである。では、何がちがうかというと、MAP 推定と最尤推定の違いは、以下の定義のように事前分布 $P(\theta)$ を用いているかどうかの違いである。

定義 5.5. *MAP* 推定

観測データ x を得た後の、パラメータ θ の「最も確からしい値」を求める方法で、式で書くと以下のようになる。

$$\theta_{MAP} = \arg\max_{\theta} P(\theta|x) = \arg\max_{\theta} \left[P(x|\theta)P(\theta)\right]$$

ここで、$P(\theta|x)$ は事後分布、$P(x|\theta)$ は尤度（データの確率)、$P(\theta)$ は事前分布（パラメータの先入観）

■尤度分布と事前分布は推定する

ベイズ推定では、尤度分布と事前分布は、以下のようにあらかじめ与えている事が前提になる。

$$\text{事後分布} = \frac{\text{尤度分布} \times \text{事前分布}}{\text{周辺尤度}} \qquad P(\theta \mid x) = \frac{P(x \mid \theta)P(\theta)}{P(x)}$$

- 尤度分布は、パラメータ θ が与えられているときの データの確率で、既知のモデルと考える。
- 事前分布は、原因 θ が起こる確率で、データを観測する前の「信念」として自分で仮定して与える。

以下の表でいえば、後ろの 2 つの列の確率をあらかじめ与えた上で、事実 x が観測できたときの原因が θ である確率 $P(\theta \mid x)$ を求める問題である。

事例	解決したい課題 $P(\theta \mid x)$		事前に設定する情報	
	事実 x	原因 θ	尤度分布 $P(x \mid \theta)$	事前分布 $P(\theta)$
赤・青が入った袋	青 or 赤	どの袋が選ばれた？	袋別の赤玉・青玉の出現率	どの袋を選ぶ事が多いか
飛行機事故の原因	事故が発生	どの原因が発生した？	原因別の事故発生率	どの原因が統計的に多いか
病気の検査方法	陽性 or 陰性	病気だといいきれる？	病気ありなしでの陽性率	病気の統計的出現率

■MAP 推定の手順

Step.1 データを眺めて事後分布を推定する

図 30 のように、データを観察して分布を想定する。そして、分布全体を求めるのは諦めて、代わりに事後分布が最大のとなる平均の点だけを求める（σ=1 は既知とする）。

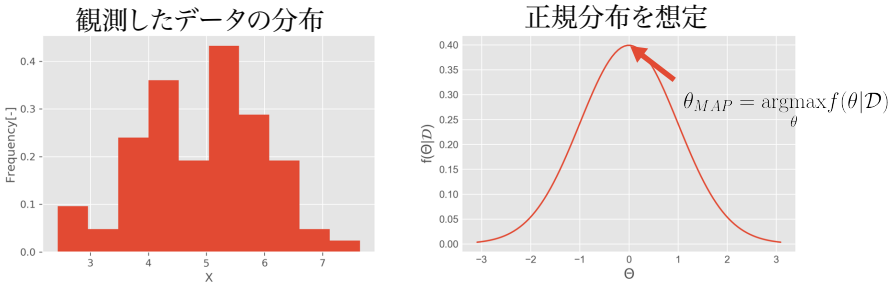


図 30　観察データから事後分布を推定する

Step.2 事前分布を設定する

次に必要なのは、以下の式の事前分布 $f(\theta)$ を設定しておくことである。

$$f(\theta \mid \mathcal{D}) \propto f(\mathcal{D} \mid \theta) f(\theta)$$

ここで「まだ何も知らないし、特に偏った仮定をしたくない」というときに使うのが「無情報事前分布（uninformative prior）」で、事後分布に影響を与えない一様分布か Normal$(\theta \mid 0, 100$ のような広い範囲で一定値になるような正規分布を設定する。$f(\theta) =$ const なので

$$\begin{aligned} \theta_{MAP} = \underset{\theta}{\text{argmax}} f(\theta \mid \mathcal{D}) &= \underset{\theta}{\text{argmax}} f(\mathcal{D} \mid \theta) f(\theta) \\ &= \text{const.} \ \underset{\theta}{\text{argmax}} f(\mathcal{D} \mid \theta) \end{aligned}$$

となって、これは最尤推定値に他ならない。

■無情報事前分布の変形

無情報事前分布をプログラムで実装するために、変形しておく。一般に観察データ D の確率は 0 以下でそれを複数書けると小さな数になる。そうした小さな数を正確に計算するために以下のように対数を取り、マイナスをかけて最大を求める代わりに最小を計算する方法に変えておく。これを NLL(Negative Log Likelihood) と呼ぶ。

$$\begin{aligned}\theta_{MAP} &= \underset{\theta}{\operatorname{argmax}} f(\mathcal{D} \mid \theta) \\ &= \underset{\theta}{\operatorname{argmax}} \left(\log_{10} f(\mathcal{D} \mid \theta)\right) \\ &= \underset{\theta}{\operatorname{argmin}} \left(-\log_{10} f(\mathcal{D} \mid \theta)\right) \\ &= \underset{\theta}{\operatorname{argmin}} (NLL)\end{aligned}$$

また観測データが正規分布 $\mathrm{Normal}(\mathcal{D} \mid \theta)$ から独立にサンプリングされているとすると、観測したデータ系列を $x_1, x_2, \cdots, x_n$ だとすると以下の式のように積和 $\prod$ で表される。

$$\begin{aligned}f(\mathcal{D} \mid \theta) &= \mathrm{Normal}(x_1 \mid \theta) \times \mathrm{Normal}(x_2 \mid \theta) \times \cdots \times \mathrm{Normal}(x_n \mid \theta) \\ &= \prod_i \mathrm{Normal}\left(x_i \mid \theta\right)\end{aligned}$$

これを $NLL = -\log_{10} f(\mathcal{D} \mid \theta)$ に代入すると、対数計算は積を和に変える（$\log ab = \log a + \log b$）ので以下のように変形出来る。

$$\begin{aligned}NLL &= -\log_{10} \prod_i \mathrm{Normal}\left(x_i \mid \theta\right) \\ &= -\sum_i \log_{10} \mathrm{Normal}\left(x_i \mid \theta\right)\end{aligned}$$

■MAP 推定の実装

このプログラムは平均値の判らない正規分布から、８０個のサンプルデータを取ってきたデータから、平均値を推定するプログラムである。

ソースコード 4　MAP 推定の実装

```
import numpy as np
from scipy import optimize
import matplotlib.pyplot as plt
from scipy import stats
import pandas as pd

plt.style.use("ggplot")

df = pd.read_excel("../data/MAP_sample.xlsx", index_col="id")

plt.hist(df["value"])

def likelihood(mu, *args):
    li = -np.log10(stats.norm.pdf(mu, loc=args))
    return(np.sum(li))

optimize.minimize(likelihood, 1, args=df["value"])
```

6 離散型確率分布

最初に「確率変数」や「確率関数」や「確率分布」という用語を説明しておく。

用語 6.1. 【確率空間】
以下の 3 つの組 $(\Omega, \mathcal{F}, P)$ を確率空間という。

- 標本空間 Ω
- Ω の部分集合で、確率 T の定義されたものの集まり $\mathcal{F}$
- $\mathcal{F}$ に属する部分集合に対して定義された確率 P

用語 6.2. 【標本空間 Ω】
与えらえた試行において起こり得る個々の事象（*event*）を標本点（*sample point*）と呼び、試行において起こり得るすべての標本点からなる集合を標本空間（*sample space*）と呼ぶ。標本空間を Ω で表し、標本空間に属する個々の標本点を ω で表す。$\omega \in \Omega$ となる。

用語 6.3. 【確率変数】
確率変数 X とは、全事象 $\Omega = \omega_1, \omega_2, \omega_3, \cdots$ の一つ一つの事象 $\omega_1, \omega_2, \omega_3, \cdots$ に数値 $x_1, x_2, x_3, \cdots$ が割り当てられており、その数値のいずれかを取る場合で、それぞれの数値が出る確率が決まっている場合に、この $X = \{x_1, x_2, x_3, \cdots, x_n\}$ を確率変数と呼ぶ。

この $x_1, x_2, x_3, \cdots, x_n$ を確率変数 X の実現値と呼び、実現値が、$0, 1, 2, \cdots$ のようにとびとびの値をとる時、X を離散型の確率変数と呼ぶ。

用語 6.4. 【確率関数】
X の各実現値 $x_1, x_2, \cdots, x_n$ に対する確率 $p_1, p_2, \cdots, p_n$ が定まっている時、各実現値から確率への関数 $f(x_i) = p_i$ を確率関数と呼ぶ。

用語 6.5. 【確率分布】
確率変数 X のすべての実現値 $\chi = x_1, x_2, \cdot, x_n$ に対して確率 $p_1, p_2, \cdots, p_n$ が定まっている時「X の確率分布 *(Probability Distribution)* が与えられているという。

一般的に確率分布は確率関数によって与えられる。この散型の確率関数 $f(x)$ が持つべき性質は以下の性質をみたす。

性質 6.1. X がとり得る離散値の集合を χ とした時、確率関数 $f(x)$ は以下の性質を満たす

$$f(x) > 0 \quad (x \in \chi) \qquad \text{かつ} \qquad \sum_{x \in \chi} f(x) = 1 \tag{6.1}$$

図 31 に主な確率分布を示す

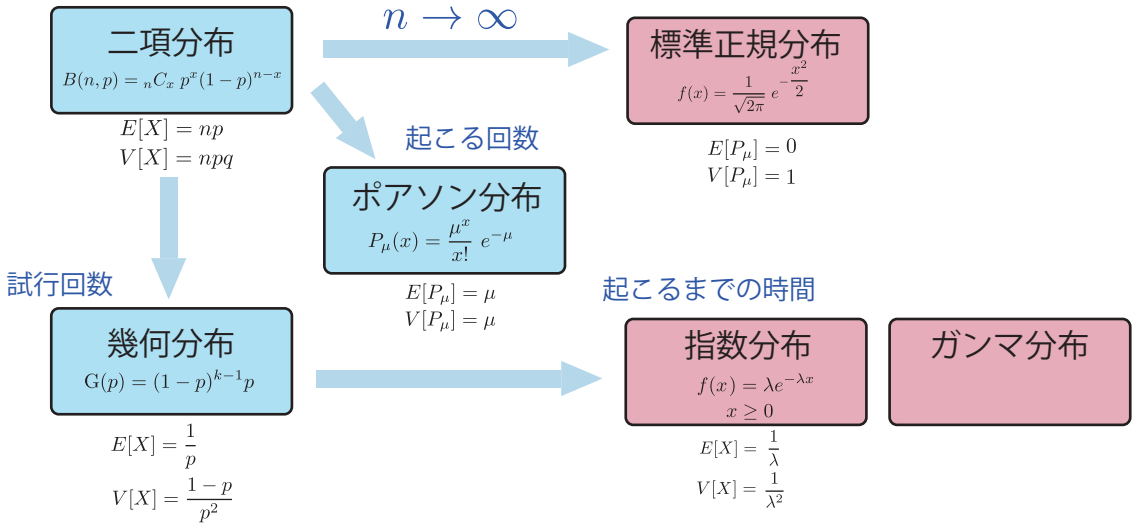


図 31　代表的な確率分布

6.1 ベルヌーイ分布

「成功か失敗か」「表か裏か」「勝ちか負けか」のように 2 種類のみの結果しか得られないような実験、試行をベルヌーイ試行 (Bernoulli trial) または二項試行 (binomial trial) と呼ぶ。ベルヌーイ試行は以下の 3 つの条件を満たすような試行である。

1. 試行の結果は 2 種類のみの事象になる。例えば成功 (1) または失敗 (0) のいずれかになる。
2. 各試行は独立で、前に出た事象と次の事象には関連がない。
3. 確率は試行を通じて一定、例えば成功確率 p 、失敗確率 $(1 - p)$ はどの試行でも同じ確率。

ベルヌーイ分布とは、そうしたベルヌーイ試行の結果を 0 と 1 で表した場合の確率関数の分布を指す。

定義 6.1. 【ベルヌーイ分布】
ベルヌーイ試行において、確率変数を X とし、事象が生じた場合を $x = 1$、そうでない場合を $x = 0$ という値をとるものとし、$x = 1$ が生じる確率を p とすると、確率関数は以下のように書くことができる。これをベルヌーイ分布 *(Bernoulli distribution)* と呼び、$Ber(x)$ という記号で表す。

$$Ber(x) = p^x(1-p)^{1-x} \qquad \text{ここで} \quad x = \{1, 0\} \tag{6.2}$$

ちなみに当たり前だが、どちらかが生じる確率は以下のように 1 となる。

$$Ber(1) + Ber(0) = p + (1-p) = 1$$

ベルヌーイ分布の平均は

$$\begin{aligned} E(X) &= \sum_{x=0}^{1} xp^x(1-p)^{1-x} \\ &= 0 \times p^0 \times (1-p)^1 + 1 \times p^1 \times (1-p)^0 = p \end{aligned} \tag{6.3}$$

分散は $V(X) = E[X^2] - (E[X])^2$ を使って求める[*19] 。まず

$$\begin{aligned} E[X^2] &= \sum_{x=0}^{1} x^2 p^{x^2}(1-p)^{1-x^2} \\ &= 0^2 \times p^{0^2} \times (1-p)^{1-0^2} + 1^2 \times p^{1^2} \times (1-p)^{1-1^2} = p \end{aligned}$$

なので

$$\begin{aligned} V(X) &= E[X^2] - E[X]^2 \\ &= p - p^2 = p(1-p) \end{aligned} \tag{6.4}$$

[*19] この式は以下のように展開できる。

$$\begin{aligned} V(X) &= \sum_i^n p_i(x_i - \mu)^2 = \sum_i^n p_i x_i^2 - 2\mu \sum_i^n p_i x_i + \mu^2 \sum_i^n p_i \\ &= \sum_i^n p_i x_i^2 - 2\mu \times \mu + \mu^2 \times 1 == \sum_i^n p_i x_i^2 - \mu^2 = E[X^2] - (E[X])^2 \end{aligned}$$

この式を微分してゼロをおくと、$p = \frac{1}{2}$ の時に分散が最大となる事がわかる。

6.2　二項分布と幾何分布

ベルヌーイ試行 は 1 回の試行で成功 or 失敗を記録する最も基本的な試行で、二項分布も幾何分布もベルヌーイ試行に基づく確率分布である。その違いは、以下のように、二項分布は**成功回数に着目**した分布で、幾何分布 は**試行回数に着目**した分布である。

- **二項分布（Binomial Distribution）**
  n 回のベルヌーイ試行を行い、そのうち 成功する回数 X を確率変数とする分布。
- **幾何分布（Geometric Distribution）**
  ベルヌーイ試行を繰り返し、最初の成功が出るまでの試行回数 X を確率変数とする分布。

■二項分布と幾何分布の確率関数

独立なベルヌーイ試行を n 回繰り返した時の和の分布が二項分布である。例えば、コインを n 回投げて表の出た回数を X とすると、X は確率変数でその実現値である x は、$x = 0, 1, 2, \cdots, n$ という n 個の値をとることになる。

定義 6.2.　【二項分布】
n 回のベルヌーイ試行を行うものとする。確率変数を X として成功事象が生じた回数を x とする。この成功事象が生じる確率を p、生じない確率を $q = 1 - p$ とすると、二項分布の確率関数は以下のように書くことができる。

$$\begin{aligned} X \sim Bin(n,p) &= {}_nC_x\, p^x q^{n-x} \\ &= {}_nC_x\, p^x (1-p)^{n-x} \end{aligned} \tag{6.5}$$

$X \sim Bin(n,p)$ は確率変数 X が二項分布 $Bin(n,p)$ に従う事を意味する。$Bin(n,p)$ という記号（または $B(n,p)$ とも書くことがある）からわかるように二項分布のパラメータは試行回数 n と成功確率 p の 2 つであり、この 2 つで分布が決まる。この式（6.5）の ${}_nC_x$ は、n 回の試行の内のどこかで成功事象が x 回生じた場合の組み合わせの数であり以下のように計算される。

$${}_nC_x = \frac{n!}{x!(n-x)!}$$

そして、その時の確率は成功事象が x 回で、$n-x$ 回は成功していないので、確率は 2 つの積であり $p^x(1-p)^{n-x}$ である。その結果、二項分布の確率関数は式（6.5）のように表す事ができる。

幾何分布とは、独立なベルヌーイ試行を繰り返し「最初の成功が k 回目に出る」 という確率で、何回目で成功したかを確率変数 X とする確率分布である。例えば、「 コインを投げ続けて、初めて表（成功）が出るまでの回数」であり、前の $k-1$ 回はすべて失敗し、k 回目で初めて成功する 確率。

定義 6.3. 【幾何分布】

成功確率 p のベルヌーイ試行を繰り返し、最初の成功が出るまでの試行回数を確率変数 X とすると、$X = k$ となる確率関数は以下のように表す事ができる。

$$X \sim Geo(p) = (1-p)^{k-1}p, \quad k = 1, 2, 3, \ldots \tag{6.6}$$

$X \sim \mathrm{Geo}(p)$ は確率変数 X が幾何分布 $\mathrm{Geo}(p)$ に従う事を意味する。幾何分布のパラメータは成功確率 p のみである。X は最初の成功が出るまでの試行回数（成功を含む）なので、「$k-1$ 回の失敗の後に、k 回目で初めて成功する確率」は、$(1-p)^{k-1}p$ となる。

■二項分布の平均と分散の導出

性質 6.2. 【二項分布の平均と分散】

$$\begin{aligned} &\text{平均} \quad E[X] = np &(6.7)\\ &\text{分散} \quad V[X] = npq \\ &\qquad\quad = np(1-p) &(6.8)\end{aligned}$$

二項分布の平均の導出

まず平均を求めよう。二項分布の定義式（6.5）より平均値は（値 x と確率 ${}_nC_x\, p^x q^{n-x}$ の積和なので、

$$E[X] = \sum_{x=0}^{n} x \cdot \frac{n!}{x!(n-x)!}\, p^x (1-p)^{n-x} \tag{6.9}$$

この式を変形していって $E[X] = np$ を導出するのだが、その時に、「n 個から x 個を選んでさらに x から 1 つを選ぶ」組み合わせと「先に n 個から 1 個を選び残りの $n-1$ 個から $x-1$ 個をとる」組み合わせが同じになるという性質[*20]を利用する。式で書くと以下のようになる。

$$x \cdot {}_nC_x = x \cdot \frac{n!}{x!(n-x)!} = \frac{n(n-1)!}{(x-1)!(n-x)!} = n \cdot {}_{n-1}C_{x-1}$$

この性質を使って $y = x - 1$ とおいて y の式に変形する事で平均を求める。

式 6.9 について、$x = 0$ の場合は $\sum$ の中全体も 0 なので $\sum$ を $x = 1$ から加算しても同じである。さらに

[*20] わかりやすいように事例で説明すると、「先に n 人の中から x 人の委員を選び、さらにその中から一人の委員長を選ぶ」場合の起こり得る場合の数は $x \cdot {}_nC_x$ で表す事ができる。そして今度は「先に n 人から一人の委員長を選び、さらに残った $n-1$ 人から $k-1$ 人の委員を選ぶ」場合の場合の数は $n \cdot {}_{n-1}C_{k-1}$ となる。この二つの事象は同じなので、$x \cdot {}_nC_x = n \cdot {}_{n-1}C_{x-1}$

この式の $\sum$ の中の x と $x! = x \cdot (x-1), \cdots, 2 \cdot 1$ の x とを通分して消去すると

$$\begin{aligned} E[X] &= \sum_{x=1}^{n} x \cdot \frac{n!}{x!(n-x)!}\, p^x (1-p)^{n-x} \\ &= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!}\, p^x (1-p)^{n-x} \end{aligned}$$

ここで n と p とを $\sum$ の外に出す

$$E[X] = np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!}\, p^{x-1} (1-p)^{n-x}$$

ここで $y = x-1$、$m = n-1$ とおくと、範囲 $1 \leqq x \leqq n$ は $0 \leqq y \leqq m$ になる。また

$$n - x = m - y$$

となる。これらを先の式の $\sum$ 記号の中の式に代入していく

$$\begin{aligned} E[X] &= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!}\, p^{x-1} (1-p)^{n-x} \quad \text{この式に代入すると} \\ &= np \sum_{y=0}^{m} \frac{m!}{y!(m-y)!} p^y (1-p)^{(m-y)} \\ &= np \sum_{y=0}^{m} {}_mC_y p^y (1-p)^{(m-y)} \end{aligned}$$

この $\sum_{y=0}^{m} {}_mC_y p^y (1-p)^{(m-y)}$ は、確率 p の事象が m 回の試行で y 回起こる確率を $y=0$ から $y=m$ まで足した値を意味するので、当然 1 となる[*21] 。なので

$$E[X] = np$$

二項分布の分散の導出

分散もほぼ同じような計算手順をとるが、分散は $V(X) = E[X^2] - (E[X])^2$ を使って求める。平均 $E[X]$ はすでに求めたので、$E[X^2]$ を計算する。

この $E[X^2]$ を計算するにあたって、最初に $x^2 = x(x-1) + x$ という関係と「和の期待値は期待値の和」である事を利用して、$E[X^2]$ を以下のように展開する

$$E[X^2] = E[X(X-1) + X] = E[X(X-1)] + E[X] \tag{6.10}$$

この式より、$E[X]$ はすでに求めているので、$E[X(X-1)]$ を求めればよい事になる。

[*21] これは二項定理を使って確認する事もできる。そもそも二項定理は

$$(a+b)^n = {}_nC_0 a^n b^0 + {}_nC_1 a^{n-1} b^1 + \cdots + {}_nC_{n-1} a^1 b^{n-1} + {}_nC_n a^0 b^n = \sum_{i=0}^{n} {}_nC_i \;\; a^{n-i} b^i$$

元の式の $\sum_{y=0}^{m} {}_mC_y p^y (1-p)^{(m-y)}$ をみると、$(p + (1-p))^m$ を二項定理によって展開したものになっている。さらに、ここで $p + (1-p) = 1$ なのでこれは必ず 1 となる。

では $E[X(X-1)]$ を求めよう。まず、二項分布の定義式（6.5）より $x(x-1)$ の期待値は

$$\begin{aligned} E[X(X-1)] &= \sum_{x=0}^{n} x(x-1)P(x) \\ &= \sum_{x=0}^{n} x(x-1)\frac{n!}{x!(n-x)!}\, p^x(1-p)^{n-x} \end{aligned}$$

これを変形していく。まず、この $\sum$ 記号の中の計算において $x=0$ および $x=1$ の時は $x(x-1)$ が 0 になるので和は $x=2$ からとる。次に $x(x-1)$ と $x!$ とを通分する。つぎに分数の分子を $n! = n(n-1)(n-2)!$ とし $p^x = p^2p^{x-2}$ を代入する。そして最後に $n(n-1)p^2$ を $\sum$ の外に出すと以下のようになる。

$$\begin{aligned} E[X(X-1)] &= \sum_{x=2}^{n} x(x-1)\frac{n!}{x!(n-x)!}\, p^x(1-p)^{n-x} \\ &= \sum_{x=2}^{n} \frac{n!}{(x-2)!(n-x)!}\, p^x(1-p)^{n-x} \\ &= \sum_{x=2}^{n} \frac{n(n-1)(n-2)!}{(x-2)!(n-x)!}\, p^2p^{x-2}(1-p)^{n-x} \\ &= n(n-1)p^2 \sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!(n-x)!}\, p^{x-2}(1-p)^{n-x} \end{aligned}$$

ここで $z = x-2$、$l = n-2$ とおくと、範囲 $2 \leqq x \leqq n$ は $0 \leqq z \leqq l$ になる。また

$$n - x = l - z$$

となる。これらを先の式の $\sum$ 記号の中の式に代入していく

$$\begin{aligned} E[X(X-1)] &= n(n-1)p^2 \sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!(n-x)!}\, p^{x-2}(1-p)^{n-x} \\ &= n(n-1)p^2 \sum_{z=0}^{l} \frac{l!}{z!(l-z)!} p^z(1-p)^{l-z} \\ &= n(n-1)p^2 \sum_{z=0}^{l} {}_lC_z\, p^z(1-p)^{l-z} \end{aligned}$$

この $\sum_{z=0}^{l} {}_lC_z p^z(1-p)^{(l-z)}$ は、確率 p の事象が l 回の試行で z 回起こる確率を $z=0$ から $z=l$ まで足した値を意味するので、当然 1 となる。

$$E[X(X-1)] = n(n-1)p^2$$

以上により

$$\begin{aligned} V(X) &= E[X^2] - (E[X])^2 \\ &= E[X(X-1)] + E[X] - (E[X])^2 \\ &= n(n-1)p^2 + np - (np)^2 \\ &= np(1-p) \end{aligned}$$

■幾何分布の平均と分散の導出　　幾何分布は最初に成功するまでの確率分布である。その平均と分散は以下のように表される。例えば、サイコロを振って「6の目が出る」という事象を成功事象とすると「6の目が出る」確率は $p = \frac{1}{6}$ であり、期待値はその逆数の6、つまり6回降れば1回は「6の目が出る」という事が期待できる事になる。

性質 6.3.　【幾何分布の平均と分散】

$$\text{平均} \qquad E[X] = \frac{1}{p} \tag{6.11}$$

$$\text{分散} \qquad V[X] = \frac{1-p}{p^2} \tag{6.12}$$

幾何分布の平均の導出

期待値 $E[X]$ は、確率変数 X の重み付き平均なので、以下の式で定義される。

$$E[X] = \sum_{k=1}^{\infty} kP(k)$$

これに式 6.6 の幾何平均の確率関数 $P(k) = (1-p)^{k-1}p$ を代入すると、

$$E[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1}p \tag{6.13}$$

以下の等比級数の公式の無限版**??**を利用する。

$$\sum_{k=1}^{\infty} kr^{k-1} = \frac{1}{(1-r)^2}, \quad (|r| < 1)$$

この公式に r = 1 - p を代入すると、

$$\sum_{k=1}^{\infty} k(1-p)^{k-1} = \frac{1}{p^2}$$

この左辺は. 式 6.13 のように両辺に p をかけて

$$\sum_{k=1}^{\infty} p \cdot k(1-p)^{k-1} = p \cdot \frac{1}{p^2}$$

この左辺が $E[X]$ にほかならないので

$$E[X] = \frac{1}{p}$$

幾何分布の分散の導出

先に期待値 $E[X]$ を求めてあるので、分散は $E[X^2]$ を求めて以下の式に当てはめて計算する。

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2$$

$E[X^2]$ は、 2 乗の期待値で $P(k) = (1-p)^{k-1}p$ なので以下の式で定義される。

$$E[X^2] = \sum_{k=1}^{\infty} k^2 P(k) = \sum_{k=1}^{\infty} k^2 (1-p)^{k-1}$$

ここで、以下の等比無限級数の公式から導出された公式 6.15 を活用する（公式の導出は 88 ページ参照)。

$$\sum_{k=1}^{\infty} k^2 r^{k-1} = \frac{1+r}{(1-r)^3}, \quad (|r| < 1)$$

この式に $r = 1-p$ を代入すると、

$$\sum_{k=1}^{\infty} k^2 (1-p)^{k-1} = \frac{1+(1-p)}{p^3} = \frac{2-p}{p^3}$$

したがって、

$$E[X^2] = p \cdot \frac{2-p}{p^3} = \frac{2-p}{p^2}$$

最後に分散の定義式に当てはめて

$$\begin{aligned} \mathrm{Var}(X) &= E[X^2] - (E[X])^2 \\ &= \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{2-p}{p^2} - \frac{1}{p^2} \\ &= \frac{2-p-1}{p^2} = \frac{1-p}{p^2} \end{aligned}$$

■モーメント（moment）

「モーメント（moment)」は、元々は物理学（力学）由来の概念で「分布の形を記述する量」である。例えば、力のモーメント（トルク）は、$M = rF$ で表される。これは力 F を作用点までの距離 r で重み付けしたもので、中心点からの力の分布度合いを示している。

統計学でモーメントを数学的に明確に導入したのは、数学者の ピアソン（Karl Pearson） で、データの分布を分析する上で、平均や分散だけでなく、歪度や尖度などが考案され、それらを統一的・定量的に評価する方法として発展してきた。

例えば、正規分布 (ガウス分布)、指数分布、一様分布の分布はそれぞれ形状が異なるが、以下の表のように、モーメントを用いると共通の枠組みで特徴づけられ、分布の形状を数値で客観的に比較でき、3 次モーメントが 0 の分布は対称的、正の値は右に歪んでいることがわかる。

表 6　確率分布のモーメント比較

モーメント	**正規分布**	**指数分布**	**一様分布** [0,a]
1 次モーメント (平均)	μ	$1/\lambda$	$a/2$
2 次中心モーメント (分散)	σ^2	$1/\lambda^2$	$a^2/12$
3 次モーメント (歪度)	0	2	0
4 次モーメント (尖度)	3	9	1.8

この n 次モーメント $E[X^n]$ は、確率変数 X の n 乗の期待値 として定義される。

$$E[X^n] = \sum_k k^n P(X = k) \quad \text{（離散分布）}$$

$$E[X^n] = \int_{-\infty}^{\infty} x^n f(x) dx \quad \text{（連続分布）}$$

離散分布の場合、1 次モーメントと 2 次モーメントは以下のようになり、1 次モーメント は「重心」を意味しており平均に関連し、2 次モーメントは「広がり」を意味しており分散に関連する。

- 1 次モーメント

$$E[X] = \sum_{k=1}^{\infty} kP(X = k)$$

- 2 次モーメント

$$E[X^2] = \sum_{k=1}^{\infty} k^2 P(X = k)$$

ただし、分散は 2 次モーメントそのものではなく、2 次モーメントから 1 次モーメントの影響を取り除いたものである。

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2$$

無限等比級数公式の変形 1

以下の無限等比級数の公式を導出する。

$$\sum_{k=0}^{\infty} r^k = \frac{1}{(1-r)^2} \qquad (|r| < 1) \tag{6.14}$$

最初に部分和 S_n を以下のように定義する。

$$S_n = \sum_{k=0}^{n} r^k = 1 + r + r^2 + \cdots + r^n$$

S_n の両辺に r を掛ると、元の式と比較すると右辺は 1 項ずれた形になる。

$$rS_n = r + r^2 + r^3 + \cdots + r^{n+1}$$

S_n から rS_n を引く。

$$S_n - rS_n = (1 + r + r^2 + \cdots + r^n) - (r + r^2 + \cdots + r^{n+1})$$

右辺の多くの項は相殺され $1 - r^{n+1}$ となり、以下のように整理できる。

$$S_n(1-r) = 1 - r^{n+1}$$

よって、部分和 S_n は

$$S_n = \frac{1 - r^{n+1}}{1-r}, \quad (r \neq 1)$$

ここで極限を取る。つまり $n \to \infty$ とすると、$|r| < 1$ の場合、 $r^{n+1} \to 0$ となるので、

$$\sum_{k=0}^{\infty} r^k = \lim_{n\to\infty} \frac{1 - r^{n+1}}{1-r} = \frac{1}{1-r}.$$

さらに、以下のように両辺を r で微分する。

$$\frac{d}{dr}\left(\sum_{k=0}^{\infty} r^k\right) = \frac{d}{dr}\left(\frac{1}{1-r}\right)$$

左辺の微分は、

$$\frac{d}{dr}\left(\sum_{k=0}^{\infty} r^k\right) = \sum_{k=1}^{\infty} k r^{k-1}$$

右辺の微分は、

$$\frac{d}{dr}\left(\frac{1}{1-r}\right) = \frac{1}{(1-r)^2}$$

よって、以下の式 6.14 が導出できた。

$$\sum_{k=1}^{\infty} k r^{k-1} = \frac{1}{(1-r)^2}, \quad (|r| < 1)$$

無限等比級数公式の変形 2

先に求めた無限級数の公式 6.14 を使って、以下の無限等比級数の公式を導出する。

$$\sum_{k=1}^{\infty} k^2 r^{k-1} = \frac{1+r}{(1-r)^3}, \qquad (|r|<1) \tag{6.15}$$

先に求めた無限等比級数の公式が以下 。

$$\sum_{k=1}^{\infty} k r^{k-1} = \frac{1}{(1-r)^2}$$

さらに、これを r について微分する。

$$\frac{d}{dr}\left(\sum_{k=1}^{\infty} k r^{k-1}\right) = \frac{d}{dr}\left(\frac{1}{(1-r)^2}\right)$$

右辺を微分すると、

$$\frac{d}{dr}\left(\frac{1}{(1-r)^2}\right) = \frac{2}{(1-r)^3}$$

一方、左辺の微分は、

$$\frac{d}{dr}\left(\sum_{k=1}^{\infty} k r^{k-1}\right) = \sum_{k=1}^{\infty} k^2 r^{k-1}$$

よって、

$$\sum_{k=1}^{\infty} k^2 r^{k-1} = \frac{2}{(1-r)^3}$$

6.3 Python で二項分布を描く

■Python で二項分布を描く scipy.stats というパッケージを使う。二項分布は binom 関数によって作る。以下は 生起確率 $p = 0.4$ のベルヌーイ試行を $n = 20$ 回行った時の成功した数を 100 個抽出した場合であり、以下のような二項分布に従う変数 X を 100 個抽出し、それをヒストグラム表示したものである。

$$X \sim Bin(20, 0.4) = {}_{20}C_x \ 0.4^x 0.6^{20-x}$$

binom 関数によって生成されるのはランダムな試行のシミュレーション結果で試行する毎に異なる。一方 scipy.stats パッケージの binom.pmf 関数は理論値を計算する関数で、折れ線はその理論値を表示したものである。

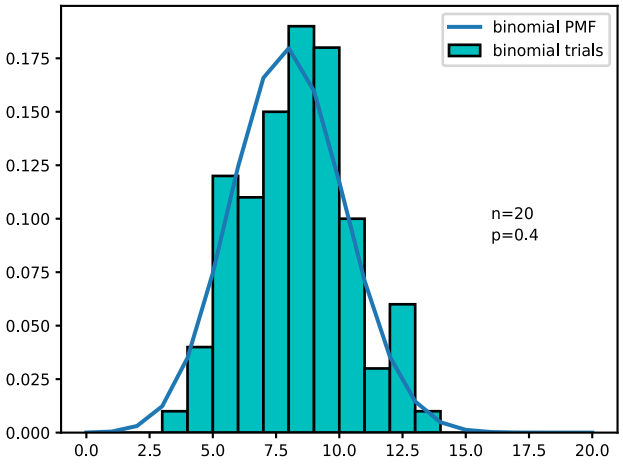


図 32　二項分布

ソースコード 5　二項分布を描くプログラム

```
# -*- coding: utf-8 -*-
"""
二項分布を計算する
Created on Sun Jul 25 10:20:11 2021

@author: hiros
"""

import numpy as np
from scipy.stats import binom
import matplotlib.pyplot as plt

np.random.seed(0)

n = 20
p = 0.4
```

```
min_k = 0
max_k = n

# 生起確率p の試行を n 回行った場合の成功の数を１００個抽出
data = np.random.binomial(n, p, size=100)
# 試行n 回のうち、確率 p の事象が起こる回数(上のnumpy 版と同じ)
#data = binom.rvs(n, p, size=100, random_state=0)

#min_k から max_k+1までの等差数列を作る
k = np.arange(min_k, max_k + 1)
#scipy.stats.binom.pmf(k, n, p) でP（x＝k）の時の二項分布の理論値を計算
binom_pmf = binom.pmf(k, n, p)

fig = plt.figure()
ax = fig.add_subplot(111)

ax.plot(k, binom_pmf, label="binomial␣PMF")
ax.hist(data, bins=n, range=(min_k, max_k), density=True,
    color='c', edgecolor='k', label="binomial␣trials")
ax.legend(loc="upper␣right")
ax.text(4*n/5, max(binom_pmf)/2+0.01, "n={}".format(n))
ax.text(4*n/5, max(binom_pmf)/2, "p={}".format(p))

plt.show()
```

■確率を変化させた場合の二項分布のグラフを描く　試行回数を３０回とし成功確率を $p = 0.1$ から $p = 0.9$ まで変化させた二項分布のグラフを描いてみると、以下のようになる。

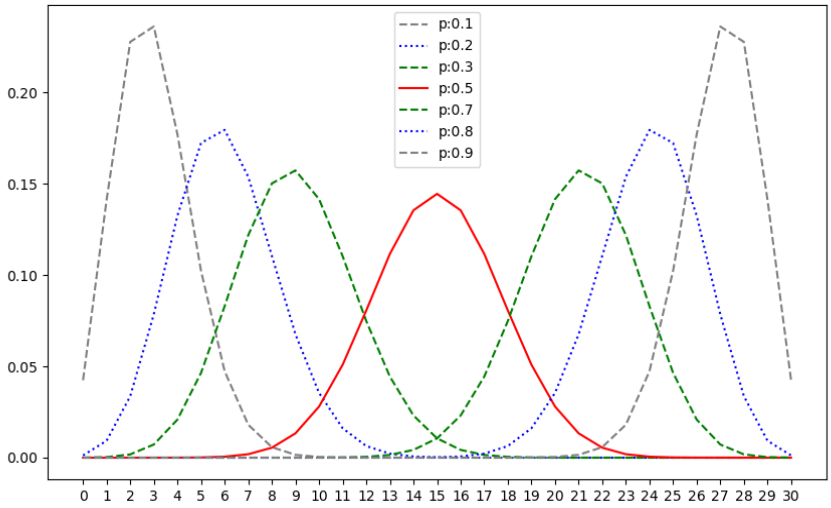


図 33　確率を変化した場合の二項分布のグラフ

ソースコード 6　確率を変化させた場合の二項分布を描くプログラム

```
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt

n = 30
linestyles = ['--', ':', '--', '-', '--', ':','--']
colorstyles = ['gray', 'blue', 'green', 'red', 'green', 'blue','gray']

fig = plt.figure(figsize=(10,6))
ax = fig.add_subplot(111)

x_set = np.arange(n+1)
for p, ls, cs in zip([0.1, 0.2, 0.3, 0.5, 0.7, 0.8, 0.9], linestyles,colorstyles):
    rv = stat s.binom(n, p)
    ax.plot(x_set, rv.pmf(x_set), label=f'p:{p}', ls=ls, color=cs)

ax.set_xticks(x_set)
ax.legend()
plt.show()
```

6.4　ポアソン分布

ポアソン分布は主に「ランダムに起きる事故・病気の発症」など頻繁に起こらない事象について「特定の期間中に何回起こる確率が何%あるのか？」を計算するのに用いられる。このポアソン分布は二項分布の回数 n を無限大に大きくした場合として導かれる。

何故、二項分布の極限がポアソン分布になるかをみていこう。事例として、ある地域の自動車交通事故における５日間の死亡事故の発生が図 34 のようであったとする。この５日間の事故は６件で、赤×印が死亡事故の発生タイミングを表している。

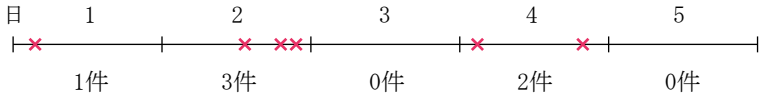


図 34　死亡事故の発生

ここで１日を非常に細かいタームに分割したとする。例えば１日を１分タームの 1440 個 (24 時間 ×60 分) に分割したと考えると、一つ一つのタームに１件の事故が起きる確率は非常に小さいので、ひとつのタームに２件の事故が同時に起こる可能性はほとんど無視できると考えられる。そうすると各タームにおいては、１件の事故が起こるか起こらないかのどちらかである。

いま１日を n 個のタームに分割し、それぞれのタームにおいて死亡事故の起こる確率を p とすると、１日に x 件の死亡事故が起こるという現象は、n 個のタームの内の x 個のタームにおいて死亡事故が起こるという事を意味する。なので x の確率分布 $P(x)$ は二項分布にしたがい以下のように表す事ができる。

$$P(x) = {}_nC_x \; p^x (1-p)^{n-x} \tag{6.16}$$

この式 (6.16) の分割数 n を無限に大きくしていったものがポアソン分布である。

■ポアソン分布の導出～ポアソンの極限定理　あとで導出するが、式 (6.16) の分割数 n を無限に大きくしていった極限が式 (6.17) であり、逆にこの式 (6.17) をポアソン分布の定義とする。

定義 6.4.　【ポアソン分布 *(Poisson Distribution)*】
確率変数 X の確率密度関数が、以下の式で与えられる確率分布をパラメータ μ のポアソン分布という。ここで、$\mu > 0$ であるとする。

$$P_\mu(x) = \frac{\mu^x}{x!} \; e^{-\mu} \tag{6.17}$$

この μ は生起確率であり、後で述べるように平均値でもある。このようにポアソン分布は、平均値である μ というひとつのパラメータのみによって特徴つけられる分布である。

では、実際にこの式 (6.16) の分割数 n を無限に大きくすることで式 (6.17) を導いてみよう。まずは n と p を二項分布の平均値を μ で置き換えよう。二項分布の平均値 μ は 81 ページの式 (6.7) のように

$$\mu = np$$

と表す事ができる。

では、分割数 n を大きくするというのはどのような意味があるのかを考えてみる。例えば、いま 1 日を 1 分タームで n 個に分割していたとすると、それを 30 秒タームにすると考えると、分割数は $2n$ になるが死亡事故が起こる確率は逆に半分 $p/2$ となる。つまり、分割数を大きくするというのは、$n \times p$ を一定にしたままで n を大きくすることであり、当たり前だが平均値 $\mu = np$ は常に一定である。

では、実際に $\mu = np$ が」が一定という条件のもとで n を無限に大きくした時の極限を求めていこう。まずは式 (6.16) の組み合わせ部分の ${}_nC_x$ を展開すると以下のようになる[*22]。

$$\lim_{n\to\infty} {}_nC_x\ p^x(1-p)^{n-x} = \lim_{n\to\infty} \left\{ \frac{n\times(n-1)\times\cdots\times(n-x+1)}{x!}\ p^x(1-p)^{n-x} \right\}$$

つぎに p を置き換える。$\mu = np$ より $p = \dfrac{\mu}{n}$ を代入すると

$$\lim_{n\to\infty} {}_nC_x\ p^x(1-p)^{n-x} = \lim_{n\to\infty} \left\{ \frac{n\times(n-1)\times\cdots\times(n-x+1)}{x!}\ \left(\frac{\mu}{n}\right)^x \left(1-\frac{\mu}{n}\right)^{n-x} \right\}$$

この右辺の極限記号内の第一項の分子 $n\times(n-1)\times\cdots\times(n-x+1)$ は x 個の掛け算であり、さらに第二項の $\left(\dfrac{\mu}{n}\right)^x = \dfrac{\mu^x}{n^x}$ の分母 n^x も x 個の掛け算である。この事を利用して第一項と第二項の分母を入れ替えると以下のようになる[*23]。

$$\lim_{n\to\infty} {}_nC_x\ p^x(1-p)^{n-x} = \lim_{n\to\infty} \left\{ \left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-x+1}{n}\right)\left(\frac{\mu^x}{x!}\right)\left(1-\frac{\mu}{n}\right)^{n-x} \right\}$$

そして右辺の極限記号内の最後の項 $\left(1-\dfrac{\mu}{n}\right)^{n-x}$ を n 乗と x 乗とに分けて展開しておくと以下のようになる。

$$\lim_{n\to\infty} {}_nC_x\ p^x(1-p)^{n-x} = \lim_{n\to\infty} \left\{ \left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-x+1}{n}\right)\left(\frac{\mu^x}{x!}\right)\left(1-\frac{\mu}{n}\right)^{n}\left(1-\frac{\mu}{n}\right)^{-x} \right\}$$

ここから多くの項が 1 になる事を利用して変形を進める。つまり、

$$\lim_{n\to\infty}\left(\frac{n}{n}\right)=1,\quad \lim_{n\to\infty}\left(\frac{n-1}{n}\right)=1,\quad \lim_{n\to\infty}\left(\frac{n-x+1}{n}\right)=1,\quad \lim_{n\to\infty}\left(1-\frac{\mu}{n}\right)^{-x}=1$$

[*22] 順列と組み合わせの計算式は以下で、ここでは単純に ${}_nC_x$ を展開しただけ。

$${}_nP_x = n\times(n-1)\times\cdots\times(n-x+1)$$
$${}_nC_x = \frac{{}_nP_x}{x!} = \frac{n\times(n-1)\times\cdots\times(n-x+1)}{1\times2\times\cdots\times x}$$

[*23] 以下のように第二項の分子 n^x を第一項に分配して、逆に第一項の分母を第二項に出した。

$$\begin{aligned}\frac{n\times(n-1)\times\cdots\times(n-x+1)}{x!}\ \left(\frac{\mu}{n}\right)^x &= \frac{n\times(n-1)\times\cdots\times(n-x+1)}{n\times n\times\cdots\times n}\ \left(\frac{\mu^x}{x!}\right)\\ &= \left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-x+1}{n}\right)\ \left(\frac{\mu^x}{x!}\right)\end{aligned}$$

を利用すると以下のように多くの項が消去できる。さらに以下では、n と関係のない $\dfrac{\mu^x}{x!}$ を極限記号の前に出している。

$$\begin{aligned}\lim_{n\to\infty} {}_nC_x\ p^x(1-p)^{n-x} &= \lim_{n\to\infty}\left\{ \cancel{\left(\frac{n}{n}\right)}\left(\cancel{\frac{n-1}{n}}\right)\cdots\left(\cancel{\frac{n-x+1}{n}}\right)\left(\frac{\mu^x}{x!}\right)\left(1-\frac{\mu}{n}\right)^n\cancel{\left(1-\frac{\mu}{n}\right)^{-x}}\right\} \\ &= \frac{\mu^x}{x!}\lim_{n\to\infty}\left(1-\frac{\mu}{n}\right)^n \end{aligned} \tag{6.18}$$

最後にこの式 (6.18) の極限部分が $e^{-\mu}$ になっている事を示す。ネイピア数 e の定義は式 (付録 A.1) のように

$$e = \lim_{h\to 0}(1+h)^{\frac{1}{h}}$$

である。いま求めたいものは以下の式である。

$$\lim_{n\to\infty}\left(1-\frac{\mu}{n}\right)^n$$

この式を e で表すために、ちょっと強引だが、以下のように

$$h = -\frac{\mu}{n}$$

と置き換えると、$n\to\infty$ という事は h で考えると $h\to 0$ であり、n を h で表すと

$$n = -\frac{\mu}{h}$$

である。これを求めたい式に代入すると、以下の右辺のように h の式として表現できる。

$$\begin{aligned}\lim_{n\to\infty}\left(1-\frac{\mu}{n}\right)^n &= \lim_{h\to 0}(1+h)^{-\frac{\mu}{h}} \\ &= \left\{\lim_{h\to 0}(1+h)^{\frac{1}{h}}\right\}^{-\mu} = e^{-\mu}\end{aligned}$$

この結果を式 (6.18) に代入すると

$$\lim_{n\to\infty} {}_nC_x\ p^x(1-p)^{n-x} = \frac{\mu^x}{x!}\ e^{-\mu}$$

となりポアソン分布の式 (6.17) を導く事ができた。

■合計が１になる事の証明　確率分布なので当たり前であるが、まずはポアソン分布の合計が 1 になる事を確認しよう。この性質は後にポアソン分布の平均や分散を求める際に使う。

まず求めたい合計を式で表そう。ポアソン分布は無限大までの分布だが、それぞれの x の値は離散値なのでシグマ記号を用いて以下のように表す事ができる[*24]

$$\sum_{x=0}^{\infty} P_\mu(x) = \sum_{x=0}^{\infty}\frac{\mu^x}{x!}\ e^{-\mu} = e^{-\mu}\sum_{x=0}^{\infty}\frac{\mu^x}{x!} \tag{6.19}$$

[*24] 二番目の式はシグマに関係のない $e^{-\mu}$ をシグマ記号の前にだした。このシグマ記号の項が

$$\sum_{x=0}^{\infty}\frac{\mu^x}{x!} = e^{\mu}$$

と変形できる事を示し、前にだした $e^{-\mu}$ との積が $e^{-\mu}\cdot e^{\mu} = 1$ である事を示せばよい。

この式をマクローリン展開を使って変形していく。まず関数 $f(\mu)$ のマクローリン展開は、式 (付録 B.1) のように以下となる。

$$f(\mu) = f(0) + f'(0)\mu + \frac{f''(0)}{2!}\mu^2 + \frac{f'''(0)}{3!}\mu^3 + \cdots + \frac{f^{(n)}(0)}{n!}\mu^n + \cdots$$

ここで $f(\mu) = e^{\mu}$ とおくと、自然数 e^{μ} の微分は e^{μ} であり[*25]、以下のように関数 $f(x)$ 及び n 階の導関数 $f^n(x)$ の x に値 0 を入れたものはすべて 1 となる。

$$f(0) = 1, \quad f'(0) = 1, \quad \cdots, \quad f^{(n)}(0) = 1$$

これにより、e^{μ} のマクローリン展開は

$$\begin{aligned} e^{\mu} &= 1 + \mu + \frac{1}{2!}\mu^2 + \frac{1}{3!}\mu^3 + \cdots + \frac{1}{n!}\mu^n + \cdots \\ &= \sum_{x=0}^{\infty} \frac{\mu^x}{x!} \end{aligned}$$

となる。

この式を先の式 (6.19) に代入することで、以下のようにポアソン分布の合計値は 1 となる事がわかる。

$$\begin{aligned} \sum_{x=0}^{\infty} P_{\mu}(x) &= \sum_{x=0}^{\infty} \frac{\mu^x}{x!}\, e^{-\mu} \\ &= e^{-\mu} \cdot \left\{ \sum_{x=0}^{\infty} \frac{\mu^x}{x!} \right\} = e^{-\mu} \cdot e^{\mu} = 1 \end{aligned}$$

■ポアソン分布の平均と分散　ポアソン分布の平均と分散を求める。結論からいうと以下のように平均も分散も「事象が生起する確率（パラメータ μ）」と同じになる。

性質 6.4. ポアソン分布の平均値を $E[P_{\mu}]$、分散を $V[P_{\mu}]$ と表すと、いずれも平均値 μ と同じになる。

$$E[P_{\mu}] = \mu \tag{6.20}$$
$$V[P_{\mu}] = \mu \tag{6.21}$$

まずは平均値を求めよう。そもそもポアソン分布は式 (6.17) のように以下の式で定義される。

$$P_{\mu}(x) = \frac{\mu^x}{x!}\, e^{-\mu}$$

なので、その平均は以下である。

$$E[P_{\mu}] = \sum_{x=0}^{\infty} x\, \frac{\mu^x}{x!}\, e^{-\mu}$$

[*25] 自然数 e とは、**??**で示すように、a^x を微分しても a^x となるような特別な底 a の値を e と定めているので、e^x の微分は e^x となる。

この式を次のように変形する。

まず $x=0$ の時の値は 0 なのでシグマ記号は $x=1$ から開始させても同じである。また、シグマ記号の中の x と分母の $x!$ とを通分し分母を $(x-1)!$ に変形。さらに μ をシグマ記号の前に出し分子を μ^{x-1} に変形する。以上により平均を以下のように表す事ができる。

$$\begin{aligned} E[P_\mu] &= \sum_{x=0}^{\infty} x \ \frac{\mu^x}{x!} \ e^{-\mu} \\ &= \mu \ \sum_{x=1}^{\infty} \ \frac{\mu^{x-1}}{(x-1)!} \ e^{-\mu} \end{aligned}$$

ここで、$y=x-1$ とおくと、第二項は y を確率変数とするポアソン分布の合計になり、その値は 1 となる。

$$\sum_{x=1}^{\infty} \ \frac{\mu^{x-1}}{(x-1)!} \ e^{-\mu} = \sum_{y=0}^{\infty} \ \frac{\mu^y}{y!} \ e^{-\mu} = 1$$

よって

$$E[P_\mu] = \mu \cdot \sum_{x=1}^{\infty} \ \frac{\mu^{x-1}}{(x-1)!} \ e^{-\mu} = \mu \cdot 1 = \mu$$

となりポアソン分布の平均値が μ である事が確認できた。

次に分散をもとめよう。分散 $(V[X])$ はそれぞれの確率変数の値から期待値（平均）を引いた偏差の二乗の期待値として定義される。つまり

$$V[X] = E[\ (X - E[X])^2\ ]$$

である。この式を変形すると以下のようになる[*26]。

$$V[X] = E[X^2] - E[X]^2$$

ここにポアソン分布の式

$$P_\mu(x) = \frac{\mu^x}{x!} \ e^{-\mu}$$

を代入すると

$$V[X] = \sum_{x=0}^{\infty} \ x^2 \cdot \frac{\mu^x}{x!} \cdot e^{-\mu} - \mu^2$$

これが求めようとする式である。ここでいきなりであるが、$x^2 = x(x-1) + x$ という関係を使って上の式に当てはめると、

$$V[X] = \sum_{x=0}^{\infty} \ x(x-1) \cdot \frac{\mu^x}{x!} \cdot e^{-\mu} + \sum_{x=0}^{\infty} \ x \cdot \frac{\mu^x}{x!} \cdot e^{-\mu} - \mu^2$$

[*26] 期待値は、定数 a に対して $E[a]=a$。定数 a と確率変数 X との積に対して $E[aX]=aE[X]$ となる性質を持つので以下のように変形できる。

$$\begin{aligned} V[X] &= E[(X-E[X])^2] = E[(X-\mu)^2] = E[X^2 - 2\mu X - \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - \mu^2 = E[X^2] - \{E[X]\}^2 \end{aligned}$$

この第二項はポアソン分布の平均の式であり μ に他ならないので以下のようになる。

$$V[X] = \sum_{x=0}^{\infty} x(x-1) \cdot \frac{\mu^x}{x!} \cdot e^{-\mu} + \mu - \mu^2 \tag{6.22}$$

この第一項を変形しよう。まず $x(x-1)$ と分母の $x!$ を相殺し、mu^x から μ^2 をくくりだしてシグマ記号の外におく。さらにこの式は $x=0$ と $x=1$ の時の値はゼロとなるのでシグマは $x=2$ からでも同じである。以上から第一項は以下のように変形できる。

$$\sum_{x=0}^{\infty} x(x-1) \cdot \frac{\mu^x}{x!} \cdot e^{-\mu} = \mu^2 \sum_{x=2}^{\infty} \frac{\mu^{x-2}}{(x-2)!} \cdot e^{-\mu}$$

ここで、$z = x-2$ とおくと、この式は以下のようになる。ここで、z に関するシグマ内の式はポアソン分布の合計なので 1 となっている。以上から式 (6.22) の第一項は

$$\mu^2 \sum_{x=2}^{\infty} \frac{\mu^{x-2}}{(x-2)!} \cdot e^{-\mu} = \mu^2 \sum_{z=0}^{\infty} \frac{\mu^z}{z!} \cdot e^{-\mu} = \mu^2$$

となる。これを式 (6.22) に代入すると、以下のようになる。

$$V[X] = \mu^2 + \mu - \mu^2 = \mu$$

このように、ポアソン分布の分散は μ となる。

6.5 Python でポアソン分布を描く

■Python でポアソン分布を描く　二項分布と同様に Scipy の統計モジュール scipy.stats というパッケージを使う。

パッケージの読み込みは以下のように指定する[*27]。

ソースコード 7　二項分布、ポアソン分布を描く scipy.stats モジュールの読み込み

```
from scipy.stats import binom
from scipy.stats import poisson
```

これらのパッケージに含まれる関数を表 7 にしめす。二項分布のパラメータは、試行数 n、生起確率 p、ポアソン分布のパラメータは生起確率 mu である。

表 7　二項分布とポアソン分布を計算する関数

	ランダム値 (size 個の値)	確率密度 ($x = k$ の時の値)	累積確率 ($x = k$ までの累積)
二項分布	binom.rvs(n,p,size)	binom.pmf(k,n,p)	binom.cdf(k,n,p)
ポアソン分布	poisson.rvs(mu,size)	poisson.pmf(k,mu)	poisson.cdf(k,mu)

ポアソン分布は、試行回数 n が大きく、確率 p が小さい時には二項分布の近似式になる事がわかっている。ほぼ $n \geq 50$、$p \leq 0.1$ くらいが目途である。以下の図 35 は、試行回数を固定値 $n = 50$ とし、確率を $p = 0.05$ ～$p = 0.5$ まで変化させた場合のポアソン分布（実線）と二項分布（点線）のグラフである。

[*27] from scipy import stats というように stats 全体を読み込んでもよい。その場合は使用する時に、stats.binom.rvs(n=50, p=20, size=100)　というように stats から指定する。

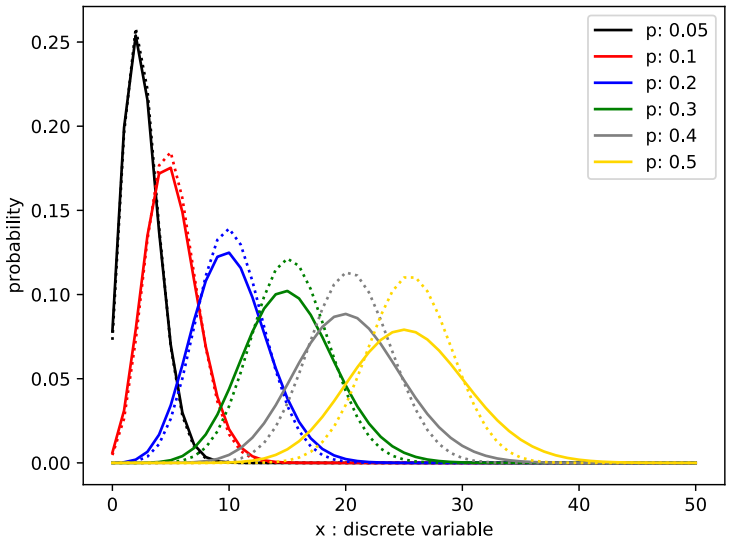


図 35　確率を変化させた場合のポアソン分布（実線）と二項分布（点線）のグラフ

ソースコード 8　確率を変化させた場合のポアソン分布と二項分布のグラフを描くプログラム

```
# -*- coding: utf-8 -*-
"""
␣二項分布とポアソン分布の違い　実線がポアソン分布、点線が二項分布
␣Created␣on␣Mon␣Oct␣11␣21:14:59␣2021
␣@author:␣hiros
␣"""

from scipy.stats import poisson
from scipy.stats import binom
import numpy as np
import matplotlib.pyplot as plt
#%matplotlib inline

fig = plt.figure()
ax = fig.add_subplot(111)

n = 51
p_lst = np.array([0.05,0.1,0.2,0.3,0.4,0.5])
mu_lst = n * p_lst
x = np.arange(n)
colorstyles = ['black','red','blue','green','gray','gold']

for p, m,cs in zip(p_lst,mu_lst,colorstyles):
    ax.plot(x, poisson.pmf(x, m),label=f'p:␣{p}',ls='-',color=cs)
    ax.plot(x, binom.pmf(x, n, p),ls=':',color=cs)
```

```
ax.legend()
plt.ylabel("probability")
plt.xlabel("x␣:␣discrete␣variable")
fig.savefig("Z:\latex_document\InformationTheory\graphics\img.pdf")
```

7　連続値型確率分布

7.1　連続型の場合は面積が確率になる

定義 7.1.　【確率密度関数】
連続型確率変数 X に対して以下の式を満たす関数 $f(x)$ が存在する時、$f(x)$ を確率変数 X の確率密度関数といい、確率変数 X は確率密度関数 $f(x)$ の確率分布に従うという。

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx \tag{7.1}$$

離散型分布では確率密度と確率は同じだが、連続型分布では確率密度と確率は異なっており、確率密度関数の一定の幅の面積が確率となる。また、確率密度関数は確率の定義から、下式のように全体の面積は 1 となるように規格化されている。

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

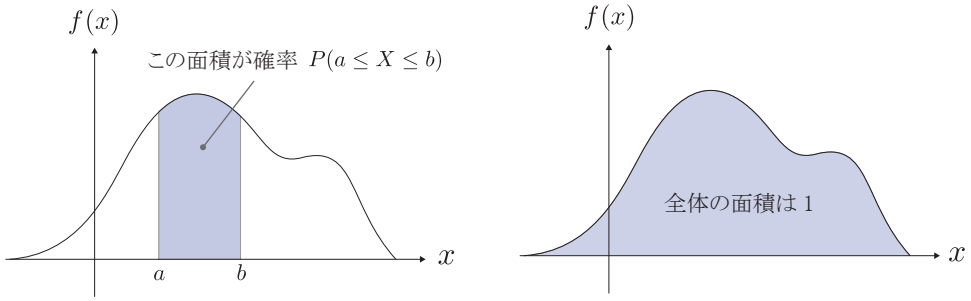


図 36　確率密度関数の面積が確率になる

性質 7.1.　【平均、分散、標準偏差】
連続型確率変数 X が確率密度関数 $f(x)$ の確率分布に従う時、平均、分散、標準偏差は以下のように表す事ができる。

$$\text{平均 } \mu = E[X] = \int_{-\infty}^{\infty} x\,f(x)\,dx \tag{7.2}$$

$$\text{分散 } \sigma^2 = V[X] = \int_{-\infty}^{\infty} (x-\mu)^2\,f(x)\,dx = E[X^2] - E[X]^2 \tag{7.3}$$

$$\text{標準偏差 } \sigma = \sqrt{V[X]} \tag{7.4}$$

式 (7.3) の分散を確認しよう。

$$\begin{aligned}\int_{-\infty}^{\infty}(x-\mu)^2 f(x)\,dx &= \int_{-\infty}^{\infty}(x^2-2\mu x+x^2) f(x)\,dx \\ &= \int_{-\infty}^{\infty} x^2 f(x)\,dx - 2\mu \underbrace{\int_{-\infty}^{\infty} x f(x)\,dx}_{\text{つまり } E[X]} + \mu^2 \underbrace{\int_{-\infty}^{\infty} f(x)\,dx}_{\text{規格化されているので } 1} \\ &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - E[X]^2\end{aligned}$$

7.2　１変数の変数変換と特徴量の変化

確率密度関数によって確率密度を表現する事ができたので、次に確率密度関数の操作をしてみる。具体的には一次関数による変数変換をしながら、その表現がどのように変化するのかを見てみる。また同様に、一次変数によって平均、分散という特徴量がどのように変化するかをみてみる。

■確率密度関数の変数変換

図 37 のように、X の確率密度関数 $f(x)$ が与えられているとし、$Y = 2X + 3$ という一次変換をした Y の確率密度関数 $g(y)$ がどのようになっているかを調べよう。

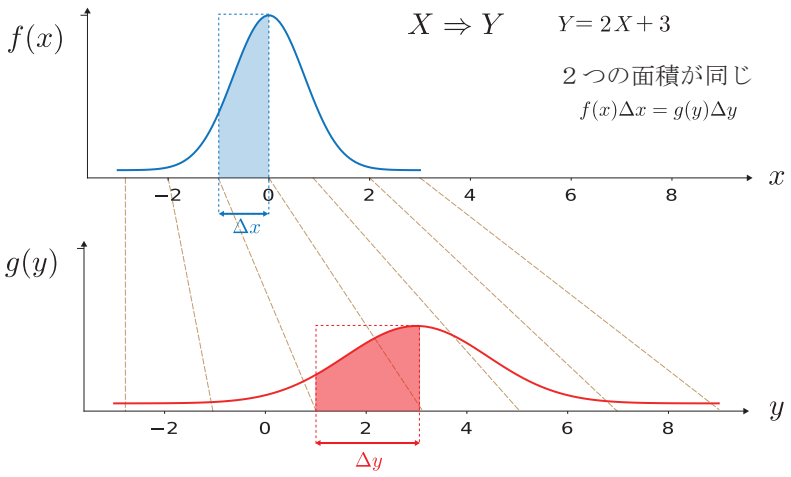


図 37　確率密度関数の変数変換のイメージ

まず、$Y = 2X + 3$ なので y 軸は x 軸の 2 倍に拡大されている事が想定される。一方で縦軸はというと、$g(y)$ は確率密度関数なので全面積が 1 という条件があり、横軸が拡大された分縦軸は縮小されているはずである。つまり、横方向に引き伸ばしたような形になっているはずである。

確認していこう。まず微小区間 Δx を取ったとする。その時対応する微小区間 Δy が存在する。その時、図 37 の影の部分のように、Δx と高さ $f(x)$、Δy と高さ $g(y)$ をかけた面積は同じであるはずである。つまり

$$f(x)\Delta x = g(y)\Delta y$$

この Δx と Δy を 0 に近づけた極限を dx, dy とすると、以下のように書ける。

$$f(x)dx = g(y)dy$$

これを変形すると、以下のようになる[*28]。

$$g(y) = f(x)\left|\frac{dx}{dy}\right| \tag{7.5}$$

具体的に $y = 2x + 3$ 事例でみてみる。x について解くと $x = \frac{1}{2}(y-3)$ なので $\frac{dx}{dy} = \frac{1}{2}$ となる。つまりこの事例の場合は、式 (7.5) は、下式となる[*29]。

$$g(y) = \frac{1}{2}f(x)$$

このように、元の関数 $f(x)$ を $Y = 2X + 3$ によって変換する事で、横に 2 倍で縦に半分になるように引き伸ばされた事を意味しており、確率密度 $g(y)$ は、$f(x)$ に対して $\frac{dx}{dy}$（この場合だと $\frac{1}{2}$）をかける必要がある。以下のようなイメージでとらえておけば良い。

「t の世界に変数変換する」ためには、元の関数 $f(x)$ を t で表すだけでなく、
Δx を新しい変数 t の変化に変換するために、変化率 $\frac{\Delta x}{\Delta t}$ をかける必要がある。

■一次変換による確率変数変換の性質

一次変換 $Y = aX + b$ によって確率変数を変換したときの平均と分散がどのように変化するかを調べよう。そのために、確率変数 X についての平均を μ_x、分散を σ_x^2 として確率変数 X に $Y = aX + b$ という一次変換を施すとする。そして、その変換後の確率変数 Y についての平均を μ_y、分散を σ_y^2 とする。この時に、変換後の平均 μ_y、分散 σ_y^2 を変換前の平均 μ_x、分散 σ_x^2 で表してみる。

【平均 μ_y】

一次変換式より $y = ax + b$ で、式 (7.5) から $g(y) = f(x)\left|\frac{dx}{dy}\right|$ なので、新しい確率変数 Y の平均 μ_y は

$$\mu_y = \int_{-\infty}^{\infty} y\, g(y)\, dy = \int_{-\infty}^{\infty} (ax+b) f(x) \frac{dx}{\cancel{dy}} \cancel{dy} = \int_{-\infty}^{\infty} (ax+b) f(x) dx$$

[*28] この $\left|\frac{dx}{dy}\right|$ は拡大縮小率を意味しており、多変数の場合のヤコビアンに相当する。ここでは、あえてヤコビアンと同様に絶対値をつけて表現した。

[*29] しかしながら、この右辺はまだ x の関数なので、それを y の関数に変換するために $y = \frac{1}{2}(y-3)$ を代入して

$$g(y) = \frac{1}{2} \cdot f\left(\frac{1}{2}(y-3)\right)$$

この $(ax+b)$ を展開してやると

$$\begin{aligned}\mu_y &= a\underbrace{\int_{-\infty}^{\infty} x\ f(x)\ dx}_{\mu_x\text{に他ならない}} + b\underbrace{\int_{-\infty}^{\infty} f(x)dx}_{\text{全ての確率なので }1}\\ &= a\mu_x + b\end{aligned}$$

【分散 σ_y^2】

上記の計算と同じく $y=ax+b$ と $g(y)=f(x)\left|\dfrac{dx}{dy}\right|$ と使う。さらに上記の結果 $\mu_y = a\mu_x + b$ を使う。

新しい確率変数 Y の分散 σ_y^2 は

$$\begin{aligned}\sigma_y^2 &= \int_{-\infty}^{\infty}(y-\mu_y)^2\ g(y)\ dy = \int_{-\infty}^{\infty}\{(ax+b)-(a\mu_x+b)\}^2 f(x)\frac{dx}{\cancel{dy}}\cancel{dy}\\ &= \int_{-\infty}^{\infty}(ax-a\mu_x)^2 f(x)dx = \int_{-\infty}^{\infty} a^2(x-\mu_x)^2 f(x)dx\\ &= a^2\underbrace{\int_{-\infty}^{\infty}(x-\mu_x)^2 f(x)dx}_{\sigma_x^2\text{に他ならない}}\\ &= a^2\sigma_x^2\end{aligned}$$

7.3　多変数の場合の基本

統計計算では２つ以上の確率変数の和や、確率変数相互の関係を調べる方法が重要。ここでは多変数の場合の同時分布や周辺分布、条件付き分布について調べる。その前に、まず２変数に拡張した重積分について説明する。

■重積分について

１変数の定積分が面積を表しているのに対して、２変数の重積分は体積を意味する。図 64(b) のような関数 $z = f(x, y)$ と xy 平面上の長方形 K があるとし、図 64(a) のように平面 K の範囲 $a \leqq x \leqq b$、$c \leqq y \leqq d$ を x 軸を n 個、y 軸を m 個に区分してあるものとする。その時、各区分の代表点 $P_{ij} = (x_i, y_j)$ の関数値 $f(P_{ij})$ を高さとするひとつひとつの直方体の体積を集めた V を求めよう。

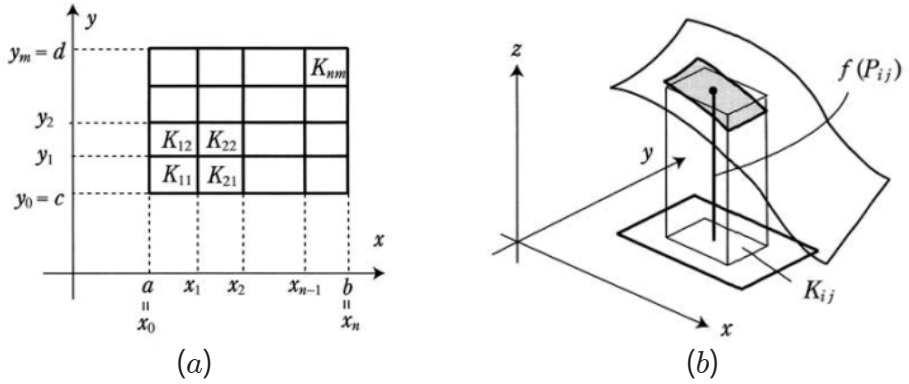


図 38　重積分と体積

ひとつひとつの区画を $\Delta x_i = x_i - x_{i-1}$、$\Delta y_i = y_i - y_{i-1}$ と表示すると

$$\begin{aligned} V = & f(P_{11})\Delta x_1\Delta y_1 + f(P_{21})\Delta x_2\Delta y_1 + \cdots + f(P_{n1})\Delta x_n\Delta y_1 + \\ & f(P_{12})\Delta x_1\Delta y_2 + f(P_{22})\Delta x_2\Delta y_2 + \cdots + f(P_{n2})\Delta x_n\Delta y_2 + \\ & \vdots \\ & f(P_{1m})\Delta x_1\Delta y_m \quad + \quad f(P_{2m})\Delta x_2\Delta y_m \quad + \cdots + f(P_{nm})\Delta x_n\Delta y_m \end{aligned}$$

シグマ記号であらわすと、

$$V = \sum_{i}^{n}\sum_{j}^{m} f(P_{ij})\Delta x_i\Delta y_j$$

定義 7.2. 重積分の定義　　$n \to \infty$、$m \to \infty$ の時に、この体積の和の極限が存在するならば、それを $f(x, y)$ の領域 D における重積分と呼び、以下のように表す。

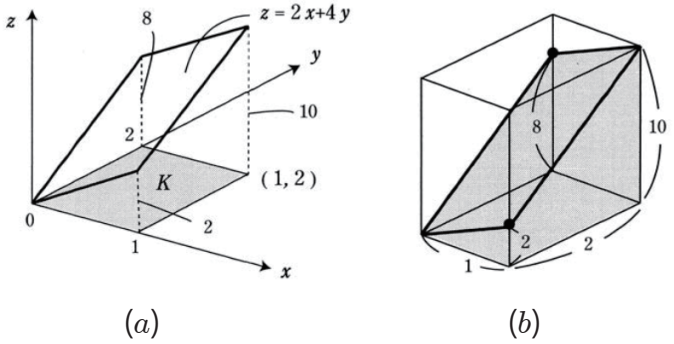
$$\int\int_D f(x, y)\, dxdy = \lim_{n,m\to\infty} \sum_{i}^{n}\sum_{j}^{m} f(P_{ij})\Delta x_i\Delta y_j \tag{7.6}$$

簡単な事例を元に計算過程を追いかけてみる。

例題 7.1. 双一次関数である $z = 2x + 4y$　の長方形領域 $K(0 \leqq x \leqq 1, \quad 0 \leqq y \leqq 2)$ 上での重積分

$$I = \int\int_K (2x + 4y)dxdy$$

を求める



(a)　　(b)

図 39　重積分の事例

図形的に求める

この関数 $z = 2x + 4y$ は平面をつくり、(x, y) が $f(1, 0) = 2$、$f(1, 2) = 10$、$f(0, 2) = 8$ なので、図 65 の (b) の斜線部分が求める体積。これは、$1 \times 2 \times 10$ の直方体の半分になるので

$$1 \times 2 \times 10\nabla \cdot 2 = 10$$

累次積分で求める

累次積分とは、重積分

$$\int\int f(x, y)dxdy$$

を解くときに

$$\int \left\{ \int f(x, y)dx \right\} dy$$

というように、先に x で積分して、その結果を次に y で積分をするという 2 段構成にする積分方法で、

「逐次積分」とも呼ばれる。実際に事例でやってみる。

$$
\begin{aligned}
V &= \int_0^2 \int_0^1 (2x+4y)\, dxdy \\
&= \int_0^2 \left\{ \int_0^1 (2x+4y)dx \right\} dy \\
&= \int_0^2 \left\{ [x^2+4yx]_1 - [x^2+4yx]_0 \right\} dy \\
&= \int_0^2 (1+4y)dy \\
&= [y+2y^2]_2 - [y+2y^2]_0 \\
&= 10
\end{aligned}
$$

というように、先の 2 つの結果と同じである。

■同時分布の確率分布

まずは連続型 2 変数の場合の確率密度関数の定義。

定義 7.3. 【同時分布の確率密度関数】
連続型の 2 つの確率変数 X、Y について、$a \leqq X \leqq b$ かつ $c \leqq y \leqq c$ となる確率 $P(a \leqq x \leqq b, c \leqq y \leqq d)$ が、

$$P(a \leqq x \leqq b, c \leqq y \leqq d) = \int_c^d \int_a^b f(x,y)dx\, dy \tag{7.7}$$

で表される時、$f(x,y)$ を変数 X、Y の同時分布の確率密度関数という。

図 40 の $Z = f(x,y)$ は確率密度を表していて、領域 A は $A = \{(x,y) \mid a \leqq X \leqq b, c \leqq Y \leqq d\}$ という XY 平面上の集合で、その上に立っている赤の柱の体積が確率 $P(a \leqq x \leqq b, c \leqq y \leqq d)$ を表す。

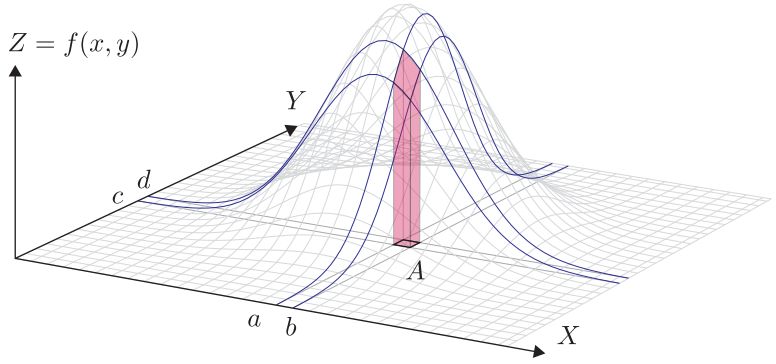


図 40　連続型 2 変数の確率密度関数と確率

当然 $f(x,y)$ は確率密度関数なので、全て 0 以上であり、全ての確率の合計は 1 である。つまり、

$$f(x,y) \geqq 0 \qquad \text{であり、} \qquad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$$

定義 7.4. 【周辺確率密度】

確率変数 X の周辺確率密度 $f(x)$ を以下のように定義する。

$$f(x) = \int_{-\infty}^{\infty} f(x,y)\ dy \tag{7.8}$$

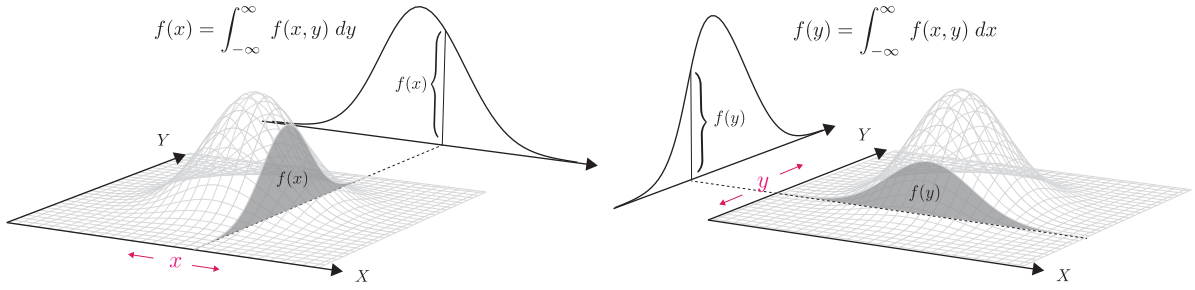
確率変数 Y の周辺確率密度 $f(y)$ を以下のように定義する。

$$f(y) = \int_{-\infty}^{\infty} f(x,y)\ dx \tag{7.9}$$

周辺確率密度関数のイメージは図 41。この図の (a) のように、X の周辺確率密度ならば $X = x$ に固定して y について積分する。つまり

$$f(x) = \int_{-\infty}^{\infty} f(x,y)\ dy$$

を計算する。この $f(x)$ は図の網掛けのような断面積を意味する。この断面積を x の関数として X 軸にそって動かして投影したグラフが $f(x)$ となり、これが X の周辺確率密度関数となる。



(a) Xの周辺確率密度　　(b) Yの周辺確率密度

図 41　周辺確率密度

定義 7.5. 【条件付き確率関数】

$f(x) \neq 0$ なる x に対して、$X = x$ を与えた時の $Y = y$ の条件付き確率関数を以下のように定義する。

$$f(y|x) = \frac{f(x,y)}{f(x)} \tag{7.10}$$

$X = x$ を与えた時なので、図 42 のように、$X = x$ の時の断面の形状が確率を意味すると考えれば良い。この断面のグラフが $X = x$ の時の確率密度 $f(y|x)$ を表しているが、確率の合計が 1 であるという条件を満たしていない。

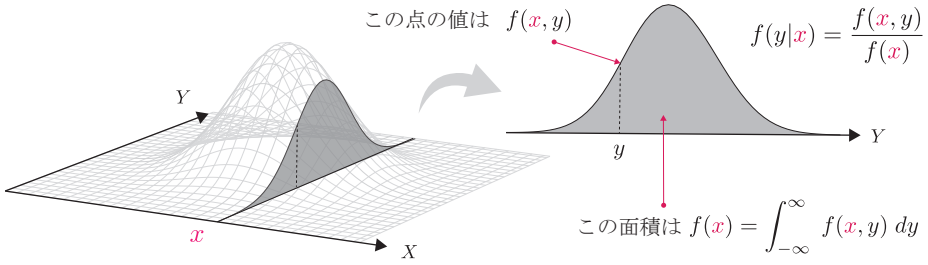


図 42　条件付き確率関数

確率の合計を 1 にするには、グラフのそれぞれの値 $f(x,y)$ をこの断面全体の面積で割って規格化すればよい。実は、この断面の面積は式 (7.8) で表される X の周辺確率にほかならないので、条件付き確率は式 (7.10) で表す事ができる。

7.4　多変数の変数変換とヤコビアン

多変数の連続値型確率分布の変数変換も、１変数の時と同様に変数変換の拡大率が重要になる。多変数の場合は拡大率がヤコビアンと呼ばれる行列式で表される。

■置換積分と変化率

まずは１変数の置換積分を変化率という視点で見直してみる。$y = f(x)$ という関数に対して、$x = g(t)$ と置いた時の置換積分の式は、

$$\int f(x)dx = \int f(g(t))\frac{dx}{dt}dt$$

この $\dfrac{dx}{dt}$ は $x = g(t)$ の接線であり t に対する x の変化率、つまり t が少し動いた時にどの程度 x が動くかを意味している。

つまり「x の世界」から「t の世界」に変数変換する為には以下の２つが必要。

- 元の関数 $f(x)$ を t で表す
- Δx を Δt の変化に変換するために変化率 $\dfrac{\Delta x}{\Delta t}$ をかける

■ヤコビアンとその意味

多変数の重積分において、１変数の変化率を意味するものがヤコビアン (Jacobian) と呼ばれる行列式である。まずは２変数の場合の重積分で確認してみる。変数 x、y の空間を変数 u、v の空間に変換する事を考える。その対応を示す関数を $x = \varphi(u, v)$、$y = \psi(u, v)$ としたとき（φ はファイ、ψ はプサイと読む）、x、y の全微分は以下の式（∂ はラウンドと読む）のようになる。

$$dx = \frac{\partial \varphi}{\partial u}\,du + \frac{\partial \varphi}{\partial v}\,dv$$
$$dy = \frac{\partial \psi}{\partial u}\,du + \frac{\partial \psi}{\partial v}\,dv$$

これを行列を用いて表すと以下のようになる。

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = \begin{pmatrix} \dfrac{\partial \varphi}{\partial u} & \dfrac{\partial \varphi}{\partial v} \\ \dfrac{\partial \psi}{\partial u} & \dfrac{\partial \psi}{\partial v} \end{pmatrix} \begin{pmatrix} du \\ dv \end{pmatrix}$$

この行列をヤコビ行列とよび、以下のように慣習的に行列 J で表す事が多い。

$$J = \begin{pmatrix} \dfrac{\partial \varphi}{\partial u} & \dfrac{\partial \varphi}{\partial v} \\ \dfrac{\partial \psi}{\partial u} & \dfrac{\partial \psi}{\partial v} \end{pmatrix}$$

この行列 J の行列式をヤコビアン (Jacobian) または関数行列式とよび、慣習的に以下のように表す。

$$|J| = \frac{\partial(\varphi, \psi)}{\partial(u.v)} = \begin{vmatrix} \dfrac{\partial\varphi}{\partial u} & \dfrac{\partial\varphi}{\partial v} \\ \dfrac{\partial\psi}{\partial u} & \dfrac{\partial\psi}{\partial v} \end{vmatrix}$$

このヤコビ行列 J が何を意味しているかを考えよう。まず下式のように、ヤコビ行列 J は新しい座標空間 (u, v) のちょっとの変化 (du, dv) を、元の座標空間 (x, y) の変化量 (dx, dy) に変換する一次変換行列であると考えられる。

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = \begin{pmatrix} \dfrac{\partial\varphi}{\partial u} & \dfrac{\partial\varphi}{\partial v} \\ \dfrac{\partial\psi}{\partial u} & \dfrac{\partial\psi}{\partial v} \end{pmatrix} \begin{pmatrix} du \\ dv \end{pmatrix} \qquad J = \begin{pmatrix} \dfrac{\partial\varphi}{\partial u} & \dfrac{\partial\varphi}{\partial v} \\ \dfrac{\partial\psi}{\partial u} & \dfrac{\partial\psi}{\partial v} \end{pmatrix}$$

つまり、ヤコビ行列 J は新しい座標での変化量を元の座標での変化量に対応させる一次変換であり、その行列式 $|J|$ はこの一次変換の拡大率を意味している。以上のように、このヤコビ行列をつかった変数変換は以下のようになる。

定義 7.6. 重積分の変数変換とヤコビアン

2 変数関数 $f(x, y)$ の重積分

$$I = \int\int_D f(x, y)\, dxdy$$

において、変数 (x, y) を $x = \varphi(u, v)$ と $y = \psi(u, v)$ という関数によって変数 (u, v) に変換したとき、被積分関数 $f(x, y)$ が、$g(s, t) = f(\varphi(u, v), \psi(u, v))$ に変換され、領域 D が領域 E に変換されたとすると以下のように表す事ができる。

$$I = \int\int_E |J|\, g(u, v)\, du\, dv \tag{7.11}$$

ここで

$$|J| = \frac{\partial(\varphi, \psi)}{\partial(u, v)} = \begin{vmatrix} \dfrac{\partial\varphi}{\partial u} & \dfrac{\partial\varphi}{\partial v} \\ \dfrac{\partial\psi}{\partial u} & \dfrac{\partial\psi}{\partial v} \end{vmatrix}$$

上記の事は、1 変数の場合の置換積分と同様に、以下のような操作をするイメージで理解すれば良いと思う。

「(s, t) の世界に変数変換する」ためには、元の関数 $f(x, y)$ を $k(s, t)$ に変換するだけでなく、
dx、dy を新しい変数 ds、dt に変換するために、変化率 $|J|$ をかける。

例題 7.2. 確率変数 X、Y による同時分布の確率密度関数 $f(x,y)$ があり、それを以下のように変数変換したとする。その時の Z、W の確率密度関数 $g(z,w)$ はどのように変換されるか？

$$\begin{cases} Z &= 3X+Y \\ W &= X+2Y \end{cases}$$

$f(x,y)$ の x と y を z と w で表した式 $g(z,w)$ を求める事なので、上記の連立方程式をといて

$$\begin{cases} X = \dfrac{2Z-W}{5} \\ Y = \dfrac{3W-Z}{5} \end{cases}$$

変換後の点 (z,w) に対応する変換前の点 (x,y) は以下のようになる。

$$(x,y) = \left(\frac{2z-w}{5}, \frac{3w-z}{5}\right)$$

今求めたいのは $g(z,w)$ の値であり、$g(z,w)$ は拡大縮小率を J とすると以下のように表すことができる。

$$g(z,w) = J \cdot f\left(\frac{2z-w}{5}, \frac{3w-z}{5}\right)$$

次に、この時の拡大縮小率を考えてみる。この変数変換は、図 67 のように元々の基底ベクトル $e_x=(1,0)$ と $e_y=(0,1)$ をそれぞれ $e_z=(3,1)$ と $e_w=(1,2)$ に移す。元の基底ベクトルがつくる四角形の面積は 1 である。変換後の基底ベクトル $e_z=(3,1)$ と $e_w=(1,2)$ がつくる平行四辺形の面積を求めれば拡大率がわかる。

図 67 のように図形的に解いてみる。青の四角形と赤の三角形と黄色の三角形の面積を $4\times3=12$ から引いて 5 となる。つまり面積が 5 倍になっているので、確率密度は $1/5$ となる。

$$g(z,w) = \frac{1}{5} \cdot f\left(\frac{2z-w}{5}, \frac{3w-z}{5}\right)$$

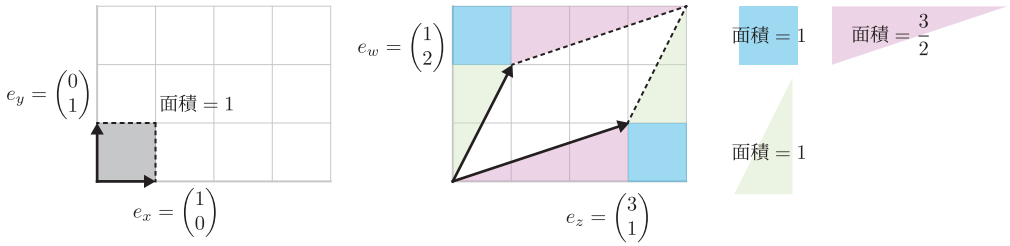


図 43　変数変換による面積の変化

この面積の拡大率を求める過程を行列を用いながら解いていこう。まず与えられた変数変換を行列表現すると以下。

$$\begin{pmatrix} Z \\ W \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$$

この一次変換行列を以下のように表現すると、この行列 A の行列式 $|A|$ が面積の拡大縮小率を表している。

$$A = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$$

実際に求めると、２次元の行列式は以下で計算する事ができる。

$$X = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{ならば、} \qquad |X| = (ad - bc)$$

なので、以下のように面積は５倍となる。

$$|A| = (3 \times 2 - 1 \times 1) = 5$$

例題 7.3. 一対一対応しているが線形変換でない変数変換について考える。確率変数 X、Y による同時分布の確率密度関数 $f(x, y)$ があり、それを以下のように変数変換したとする。その時の Z、W の確率密度関数 $g(z, w)$ はどのように変換されるか？

$$\begin{cases} Z &= Xe^Y \\ W &= Y \end{cases}$$

この変換は図 68 のように場所によって拡大率が異なる変換である。

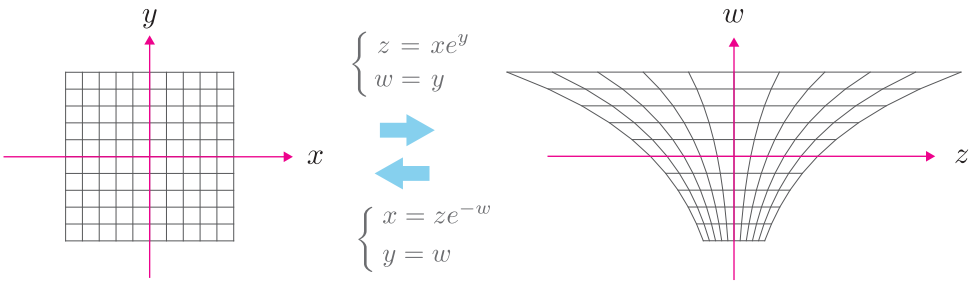


図 44　面積拡大率が場所によって異なる場合

まず与えられた変換式を X と Y について解くと

$$\begin{cases} X = Ze^{-W} \\ Y = W \end{cases}$$

つまり、変換後の座標が (z, w) であったとすると、その場合に対応する変換前の座標 (x, y) は以下のように表す事ができる。

$$(x, y) = (ze^{-w}, w)$$

また、変換による拡大率を $|J|$ とするとその確率密度関数 $g(z, w)$ は、以下のように表す事ができる。

$$g(z, w) = |J| \, f(ze^{-w}, w)$$

この $|J|$ を求めるのであるが、面積がどのように拡大されるかは場所によって異なるので、各座標点 (x, y) における面積拡大率を調べる。簡単に想定すると、y 軸方向は $w = y$ なので拡大率はゼロで、x 軸方向は $z = xe^y$ なので e^y 倍されている事になる。ここで求めたいのは (z, w) の式なので z で表すと $x = ze^{-w}$ より e^{-w} 倍となる。以上より、確率密度関数 $g(z, w)$ は下式のようになると想定される。

$$g(z, w) = \frac{1}{e^w}\, f(ze^{-w}, w)$$

次に、先の事例と同様に面積の拡大率を求める過程を行列を用いながら解いていこう。まず与えられた変数変換を行列表現すると以下。

$$\begin{pmatrix} z \\ w \end{pmatrix} = \begin{pmatrix} e^y & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

したがってその行列式は $|e^y|$ となる。これを新しい座標系である z と w の座標で表すと $|e^w|$ なので、新しい座標系での関数は以下のように表現できる。

$$g(z, w) = \frac{1}{|e^w|}\, f(ze^{-w}, w)$$

ソースコード 9　面積拡大率が場所によって異なる場合のグラフを作成するプログラム

```
import numpy as np
import matplotlib.pyplot as plt

#変換関数
def fx(x, y):
        z = x * np.exp(y)
        w = y
        return z, w

x_min = −2.8 ; x_max = 2.8
y_min = −2.8 ; y_max = 2.8

#変換前のグラフを描く
fig, ax = plt.subplots()
ax.set_xlim(x_min, x_max) ; ax.set_ylim(y_min, y_max)
X, Y = np.meshgrid(np.arange(−1, 1.2, 0.2),np.arange(−1, 1.2, 0.2))
plt.plot(X,Y) ; plt.plot(X.T,Y.T)

#変換後のグラフを描く
fig, ax2 = plt.subplots()
ax2.set_xlim(x_min, x_max) ; ax2.set_ylim(y_min, y_max)
W, Z = fx(X, Y)
plt.plot(W,Z) ; plt.plot(W.T,Z.T)

plt.show()
```

7.5 ベータ分布

ベータ分布はベイズ統計において特に重要な役割を担っている。その理由の一つが、形状が非常に柔軟であるため、事前確率分布として扱いやすいという点が挙げられる。

ベータ分布は連続型の確率分布の 1 つで、成功数 a と失敗数 b が分かっている試行に関して、成功率 p の分布を表す。

定義 7.7. ベータ分布

ベータ分布とは，確率密度関数が以下であるような確率分布。

$$f(x \mid a,b) = C\, x^{a-1}(1-x)^{b-1} \qquad (0 \le x \le 1) \tag{7.12}$$

ただし，a,b はパラメータ（正の実数）であり，C は規格化定数。

$$C = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} = \frac{1}{B(a,b)} \qquad a,b \text{ が整数なら } C = \frac{(a+b-1)!}{(a-1)!(b-1)!}$$

■ベータ分布とベイズの定理

ベータ分布はベイズの定理を使って「コイン投げにおいて表が出た時に、そのコインの表がでる確率」を表している分布である。つまり、事前分布を一様分布とし，尤度が二項分布（コイン投げ）であるときの事後分布となっている。その事を確認しよう。

- まず各変数を以下のように定義する。

  x　観測された成功回数（例：表が出た回数）
  n　試行回数（コインを投げた回数）
  θ　コインの表が出る確率（未知のパラメータ）

- 観測された成功回数が x であった時、コインの表の出る確率 θ を求めるベイズの定理は

$$P(\theta \mid x) = \frac{P(x \mid \theta)P(\theta)}{P(x)}$$

- 周辺尤度 $P(x)$ は θ に依存しない定数なので比例式で書けば以下のように書ける。

$$P(\theta \mid x) \propto P(x \mid \theta)P(\theta)$$

- いっぽう、表が出る確率が θ で x 回表が出る確率は二項分布に従うので、以下のようになる。このように二項分布が尤度関数になる。

$$P(x \mid \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

- ここで事前分布が一様分布とするので、

$$P(\theta) = 1 \quad (\text{ ただし } 0 \le \theta \le 1)$$

- さらに、組み合わせ $\binom{n}{x}$ は定数である。なので $P(\theta)$ と定数の 2 つは比例式で書くなら無視できる。つまり、以下のように表す事ができる。

$$P(\theta \mid x) \propto \theta^x (1-\theta)^{n-x}$$

- ここで、$a = x+1, \quad b = n-x+1$ を代入してやれば、以下のようにベータ分布の式 (7.12) と同じになる。

$$P(\theta \mid x) \propto \theta^{a-1} (1-\theta)^{b-1}$$

■パラメータを変えた時のベータ分布のグラフの変化

表が出る確率 θ が不明であるコインを何回か投げて，表が m 回，裏が n 回出たとしま す。このとき「表が出る確率の予測値」は，パラメータが $(a,b) = (m+1, n+1)$ であ るベータ分布に従うと考えることができる。パラメータを変えた場合のベータ分布の違いを図 45 に示した。

$(a,b) = (1,1)$ のときは、ベータ分布は図 45 の青い直線のような一様分布になる。 $(a,b) = (1,1)$ なので $m = n = 0$ のときであり、そもそもコインを投げていないときに該当する。この時は「情報が全く無いので，θ は一様分布に従う」と解釈できます。

$(a,b) = (2,3)$ のときは，ベータ分布は図 45 の赤い曲線のようになる。つまり，$m = 1, n = 2$ のときは、表が出る確率は $\frac{1}{3}$ を中心とした緩いカーブを描いている。

$(a,b) = (4,7)$ のときは，ベータ分布は図 45 の緑の曲線のようになります。つまり，$m = 3, n = 6$ のときは、表が出る確率は $\frac{1}{3}$ に近く、さきほどより試行回数が多いので平均値の確率が高まって，歪度の狭い分布になっている。図 45 を描く Python コードがソースコード 10 である。

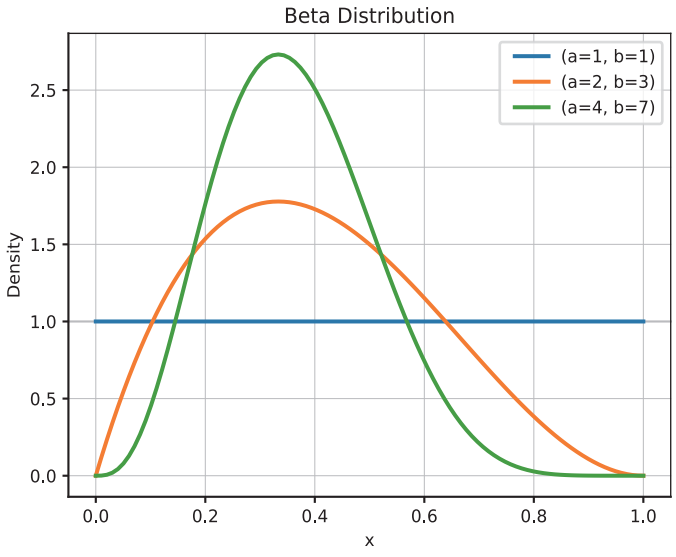


図 45　パラメータを変えた場合のベータ分布

また、a と b の比率が変化すると以下のように変化する。

「まだデータ $(a,b) = (1,1)$ が少ない」 → ベータ分布は平ら（全ての p が等しくありそう）

「たくさん裏が出た $(a, b) = 2, 8)$」 → ベータ分布は 右に山ができる（高い p がもっともらしい）
「たくさん表が出た $(a, b) = 8, 2)$」 → ベータ分布は 左に山ができる（低い p がもっともらしい）

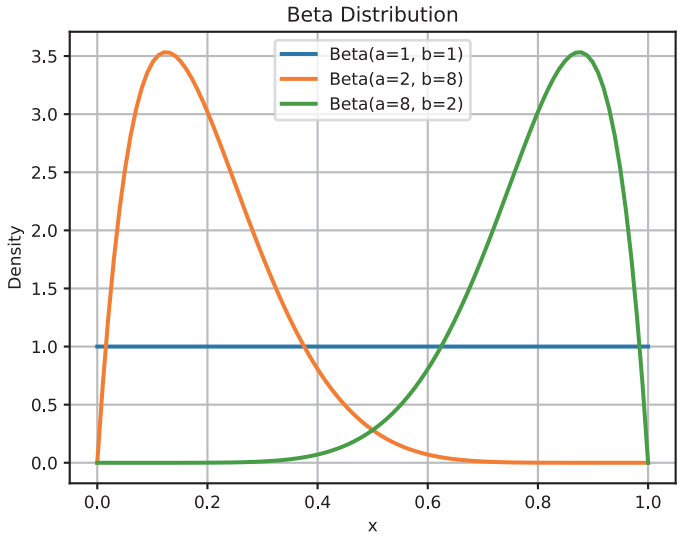


図 46　パラメータを変えた場合のベータ分布 02

通常の確率分布は「データの確率」であるのに対して、ベータ分布は「パラメータ（確率そのもの）の確率」を表していると言える。このように、「確率自体に分布を与える」という考え方は、ベイズ統計の中心概念の一つで、その他にベータ分布の多変量版であるディリクレ分布（Dirichlet distribution）や確率の代わりにそのロジット変換（$\log \frac{p}{1-p}$）に対して正規分布を仮定したロジスティック正規分布（Logistic Normal distribution）などがある。

ソースコード 10　パラメータを変えた場合のベータ分布

```
mport numpy as np
import matplotlib.pyplot as plt
from scipy.stats import beta

# パラメータ a, b を設定
params = [(1, 1), (2, 3), (4, 7)]

# x の範囲を設定
x = np.linspace(0, 1, 100)

# ベータ分布の確率密度関数(PDF) を計算し描画
for a, b in params:
        y = beta.pdf(x, a, b)
        plt.plot(x, y, label=f'Beta(a={a}, b={b})')
        plt.xlabel('x')
        plt.ylabel('Density')
        plt.title('Beta Distribution')
        plt.legend()
        plt.grid(True)

plt.savefig('beta_distribution.pdf')
plt.show()
```

8　正規分布

平均が $\mu = 0$、分散 $\rho = 1$ の正規分布を標準正規分布 (standard normal distribution) と呼び、$N(0,1)$ というように表す。標準正規分布は以下のような確率密度関数で定義される[*30]。

定義 8.1. 標準正規分布の確率密度関数は以下

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}} \tag{8.1}$$

$1/\sqrt{2\pi}$ は定数なので除くと、この標準正規分布曲線の式の基本部分は自然対数の $e^{-\frac{x^2}{2}}$ の部分。さらに e にかかっている $\frac{1}{2}$ も係数と見なして省けば、この関数の基本骨格は

$$f(x) = e^{-x^2}$$

であると考えてよい。この e^{-x^2} のグラフを表したのが図 47。骨格部分だけだが、既にグラフの形状は正規分布の形になっている。またこのグラフをよく見ると、以下の性質を持つ事がわかる。

- 常に非負である
- 左右対称である
- $x \to +\infty$ や $x \to -\infty$ では 0 に収束する
- $x = 0$ の時に最大値となる

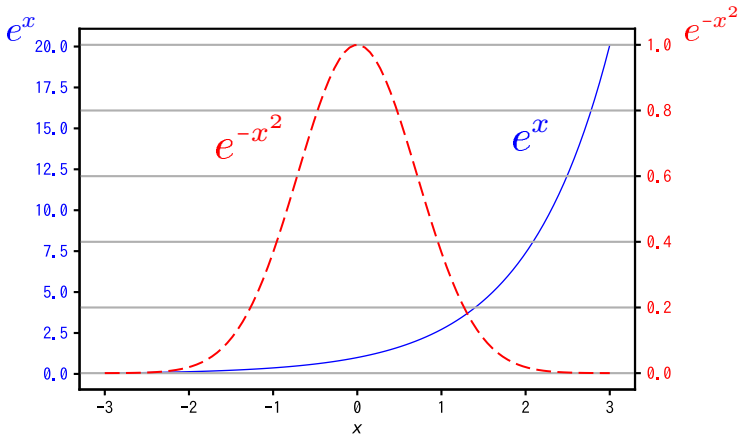


図 47　e^{-x^2} のグラフ

[*30] 同じ式であるが、指数関数部分の表現を簡素化するために以下のように表す事もある

$$f(z) = \frac{1}{\sqrt{2\pi}}\, \exp\left(-\frac{z^2}{2}\right)$$

しかしながらこの関数の積分値は 1 となっていない。確率密度関数として機能させるには、取り得る範囲全体を積分した時に値が 1 になっていなければならない。そこで以下では、この基本骨格部分の関数 $f(x) = e^{-x^2}$ の範囲全体を積分した時に 1 になるように変換する。その後、さらに分散も 1 になるように変換する。そして、その結果が式 (8.1) になる事を示していく。

8.1 積分値を１にする

$f(x) = e^{-x^2}$ の積分を求めるのだが、結論からいうと以下のように $\sqrt{\pi}$ となる。

公式 8.1. e^{-x^2} の積分

$$\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi} \tag{8.2}$$

以下でこの関数の積分値を導いていく。ただし、この解法はかなりテクニカルである。どこがテクニカルかというと、求めようとしているのは以下の N なのだが、

$$N = \int_{-\infty}^{+\infty} e^{-x^2} dx$$

N を直接解くのではなく、以下の二重積分 I の値を求めて、I と N の関係から N を求めようという、回りくどい導き方であるという点である。

$$I = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-(x^2+y^2)} dxdy$$

以下に、順をおって確認していく。

■I と N の関係　まず、I と N の関係が $I = N^2$ となっており、I の積分値の平方根が N の値である事を示そう。最初に、

$$e^{-(x^2+y^2)} = e^{-x^2} \cdot e^{-y^2}$$

より I を変形する。そして x で積分してから、y で積分する。その時に、x に関与しないものは定数として扱うと

$$I = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} e^{-x^2} \cdot e^{-y^2} \, dx \right] dy$$

$$= \int_{-\infty}^{+\infty} e^{-y^2} \left[\int_{-\infty}^{+\infty} e^{-x^2} \, dx \right] dy$$

ここで、

$$\int_{-\infty}^{+\infty} e^{-x^2} \, dx = N$$

であり、N は y での積分では定数とみなせるので上の式 I は

$$I = N \times \int_{-\infty}^{+\infty} e^{-y^2} \, dy$$

となる。さらにこの e^{-y^2} の積分値も x と y とが異なるだけで、その同じ N である。つまり

$$\int_{-\infty}^{+\infty} e^{-y^2}\, dy = N$$

であり、結果的に $I = N \times N = N^2$ となる。

■I を積分する　I を求めれば、その平方根が N の値である事がわかったので、以下の式 I を積分しよう。

$$I = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-(x^2+y^2)} dxdy$$

この計算過程も極座標にしてから、合成関数の微分をつかって・・とややこしいので (1)〜(3) までの順番に示す。

(1) **極座標変換**　最初に、被積分関数を極座標変換する。

$$x = r\cos\theta \quad , \quad y = r\sin\theta$$

とおくと、

$$-(x^2+y^2) = -(r^2\cos^2\theta + r^2\sin^2\theta) = -r^2$$

なので、

$$e^{-(x^2+y^2)} = e^{-r^2}$$

さらに、公式付録 C.4 を用いて極座標に変換した式 I は以下となる。この積分区間は平面すべてを対象区間とするので $0 \leq r \leq \infty$ となり、θ は $0 \leq \theta \leq 2\pi$ となる。

$$I = \int_0^{\pi} \int_0^{+\infty} r \cdot e^{-r^2} drd\theta \tag{8.3}$$

ちなみに、被積分関数が $r \cdot e^{-r^2}$ となっている。ここで r をかけているのは、極座標変換によって生じたものである。詳細は 190 ページ参照。

(2) **微分して原始関数を求める**　次は、この e^{-r^2} についての積分を考えるのだが、積分の前に、$z = e^{-r^2}$ とおいてこの関数を微分してみる。合成関数の微分の式 (**??**) を利用するために、$u = -r^2$ とおくと

$$\frac{du}{dr} = -2r$$

また、$z = e^{-r^2}$ は $z = e^u$ と表す事ができ（e^u の微分は e^u のままなので)、

$$\frac{dz}{du} = e^u$$

この 2 つを、合成関数の公式 (**??**) をつかって合成して

$$\frac{dz}{dr} = \frac{dz}{du} \cdot \frac{du}{dr} = e^u \cdot (-2r) = -2re^{-r^2}$$

微分した結果が求めたい re^{-r^2} となるためには、$-1/2$ をかけておけば良いので

$$\text{原始関数} = -\frac{1}{2}e^{-r^2}$$

であれば良い事が判る。

(3) **積分する**　最後に、この原始関数を使って式 (8.3) を積分するのだが、この式 (8.3) をよく見ると、被積分関数に θ は関与していないので、原始関数を θ（つまり微分すると 1 となる関数）として、先に θ で累次積分してしまう。

$$I = \int_0^{+\infty} re^{-r^2} \left[\int_0^{2\pi} \theta d\theta \right] dr$$

θ の積分は

$$\int_0^{2\pi} \theta d\theta = [\theta]_0^{2\pi} = 2\pi$$

なので、θ に関する積分は定数になるので前に出してしまう。後は以下のようにスムーズに積分できる。

$$\begin{aligned} I &= 2\pi \int_0^{+\infty} re^{-r^2} dr \\ &= 2\pi \left[-\frac{1}{2}e^{-r^2} \right]_0^{+\infty} \\ &= 2\pi \left\{ 0 - \left(-\frac{1}{2} \right) \right\} = \pi \end{aligned}$$

以上より、$I = \pi$ となる事が判った。一方で、N は先に述べたように $N = I^2$ という関係があるので、最後の結果としては $N = \sqrt{\pi}$ となる。以上のようにして e^{-x^2} の積分は下式。

$$\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}$$

この積分が $\sqrt{\pi}$ となる事がわかったので、積分値が 1 になるようにするには、$\sqrt{\pi}$ で割ってやればよい。つまり新たな関数は

$$f(x) = \frac{1}{\sqrt{\pi}} e^{-x^2} \tag{8.4}$$

8.2 分散を１にする

次に、ここまでで求めた関数

$$f(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}$$

の分散を求めて、分散を 1 にするように元の関数 $f(x)$ を変換する。ここでもいくつかの手順を踏む。まずは「分散を期待値で表し」、「平均値を求め」、最後に「分散をもとめて」、分散が 1 になるように元の関数 $f(x)$ を変換しよう。

■分散を期待値で表す　まず分散を期待値で表す事を考える。期待値はその確率分布から得られるであろうと期待できる値を意味し、Expectation（期待）の頭文字の E を用いて表す。確率変数 x の期待値とは「確率変数がとる値とその値をとる確率の積を全て足し合わせたもの」であり、連続型確率変数の場合は積分記号を用いて表現する。つまり

$$E(x) = \int_{-\infty}^{+\infty} x \cdot f(x)\, dx$$

となり、この $E(x)$ は平均値そのものである。

定義 8.2. 分散を期待値で表すと以下のようになる

$$V(x) = E(x^2) - \{E(x)\}^2 \tag{8.5}$$

まず、この式 (8.5) を確認してみる。分散の計算手順は、「それぞれのデータ x_i から平均 μ を引いた値の二乗 $(x_i - \mu)^2$ とその値が得られる確率 $f(x)$ との積和」なので

$$\begin{aligned} V(x) &= \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx \\ &= \int_{-\infty}^{+\infty} (x^2 - 2\mu x + \mu^2) f(x) dx \\ &= \int_{-\infty}^{+\infty} x^2 \cdot f(x)\, dx - 2\mu \int_{-\infty}^{+\infty} x \cdot f(x)\, dx + \mu^2 \int_{-\infty}^{+\infty} f(x)\, dx \end{aligned}$$

ここで、

$$\int_{-\infty}^{+\infty} x \cdot f(x)\, dx = \mu \quad , \quad \int_{-\infty}^{+\infty} f(x)\, dx = 1$$

なので

$$\begin{aligned} V(x) &= \int_{-\infty}^{+\infty} x^2 \cdot f(x)\, dx - 2\mu^2 + \mu^2 \\ &= \int_{-\infty}^{+\infty} x^2 \cdot f(x)\, dx - \mu^2 \end{aligned}$$

ここで、

$$\int_{-\infty}^{+\infty} x^2 \cdot f(x) \quad = E(x^2)$$

つまり、この式は x^2 の期待値の定義 $E(x^2)$ に他ならない。また $\mu^2 = \{E(x)\}^2$ なので、最終的に

$$V(x) = E(x^2) - \{E(x)\}^2$$

■平均値を求める　つぎに、確認した $V(x) = E(x^2) - \{E(x)\}^2$ を用いて分散を求めるのだが、その前に平均値を求めよう。求めたい確率分布の関数は

$$f(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}$$

なので

$$E(x) = \int_{-\infty}^{+\infty} x \cdot f(x)\, dx = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} x e^{-x^2} dx \tag{8.6}$$

先に求めたように、

$$\text{原始関数} \qquad f(x) = -\frac{1}{2} e^{-x^2}$$

を微分する。$u = -x^2$ とおいて合成関数の微分式 (**??**)　を適用すると

$$\frac{df}{du} = \frac{d}{du} \left(-\frac{1}{2} \cdot e^u \right) = -\frac{1}{2} \cdot e^u = -\frac{1}{2} \cdot e^{-x^2} \qquad , \qquad \frac{du}{dx} = -2 \cdot x$$

となるので

$$\frac{df}{dx} = \frac{df}{du} \cdot \frac{du}{dx} = -\frac{1}{2} \cdot e^{-x^2} \cdot -2x = x e^{-x^2}$$

となる。まさに求める非積分関数であるので式 (8.6) は

$$\begin{aligned} E(x) &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} x e^{-x^2} dx \\ &= \frac{1}{\sqrt{\pi}} \left[-\frac{1}{2} e^{-x^2} \right]_{-\infty}^{+\infty} \\ &= 0 \end{aligned}$$

つまり平均値は 0 となる。これは、最初に $\exp\left(-x^2\right)$ のグラフを描いた図 47 でみたように、この関数が $x \to +\infty$ や $x \to -\infty$ では 0 に収束することからも当たり前。

■分散を求める　　$E(X) = 0$ なので、分散は $V(x) = E(x^2) - \{E(x)\}^2 = E(x^2)$ となる。つまり、$E(x^2)$ を求めればよい。求めたい確率分布関数は

$$f(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}$$

なので、

$$E(x^2) = \int_{-\infty}^{+\infty} x^2 \cdot f(x)\, dx = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} x^2 e^{-x^2} dx \tag{8.7}$$

この式を

$$E(x^2) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} x \left(x e^{-x^2} \right) dx$$

ととらえて以下の部分積分を利用する。定積分の部分積分の公式 (**??**) は $F' = f$ とすると、

$$\int_a^b f \cdot g \, dx = [F \cdot g \,]_a^b - \int_a^b F \cdot g' \, dx$$

$$\begin{aligned} f = xe^{-x^2} \qquad &\text{とおくと} \qquad F = -\frac{1}{2}e^{-x^2} \\ g = x \qquad &\text{とおくと} \qquad g' = 1 \end{aligned}$$

なので式 (8.7) は

$$\begin{aligned} E(x^2) &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} x \left(xe^{-x^2} \right) dx \\ &= \frac{1}{\sqrt{\pi}} \left\{ \left[\left(-\frac{1}{2}e^{-x^2} \right) \cdot x \right]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} \left(-\frac{1}{2}e^{-x^2} \right) \cdot 1 \, dx \right\} \end{aligned}$$

ここで e^{-x^2} は $-\infty$ 及び $+\infty$ の時はゼロに収束するので、第一項は以下のようにゼロになる。

$$\left[\left(-\frac{1}{2}e^{-x^2} \right) \cdot x \right]_{-\infty}^{+\infty} = 0$$

また公式 (8.2) より、

$$\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}$$

なので、

$$\begin{aligned} E(x^2) &= \frac{1}{\sqrt{\pi}} \left\{ \int_{-\infty}^{+\infty} \left(-\frac{1}{2}e^{-x^2} \right) \cdot 1 \, dx \right\} \\ &= \frac{1}{\sqrt{\pi}} \cdot -\frac{1}{2} \int_{-\infty}^{+\infty} e^{-x^2} \, dx \\ &= \frac{1}{\sqrt{\pi}} \cdot -\frac{1}{2} \cdot \sqrt{\pi} \\ &= \frac{1}{2} \end{aligned}$$

ここまでで以下の関数の分散が 1/2 となることが分かった。

$$f(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}$$

この分散を 1 にする為に、x 方向の縮尺を広げよう。x 方向の変化は標準偏差分の $1/\sqrt{2}$ が必要

$$x : \eta = \frac{1}{\sqrt{2}} : 1 \qquad \Rightarrow \qquad x = \frac{\eta}{\sqrt{2}}$$

このように x 方向の縮尺を広げた新たな変数 η(イータ) という変数に変換すると

$$f(\eta) = \frac{1}{\sqrt{\pi}} e^{-\frac{1}{2}\eta^2}$$

ただし、この関数は変数変換をしたので、合計積分値が 1 になるようにするために修正が必要。上記の関数を積分しよう。

$$\begin{aligned}\int_{-\infty}^{+\infty} f(\eta)d\eta &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{\pi}} e^{-\frac{1}{2}\eta^2} d\eta \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}\eta^2} d\eta\end{aligned}$$

ここで、

$$z^2 = \frac{1}{2}\eta^2 \qquad \text{となるような変換、つまり} \qquad z = \frac{\eta}{\sqrt{2}}$$

とうように変数 z に変換する関数を考えて、置換積分をする（式 (**??**) 参照)。

$$\frac{dz}{d\eta} = \frac{1}{\sqrt{2}} \qquad \text{なので、} \qquad d\eta = \sqrt{2}dz \qquad \text{より}$$

元の積分は

$$\begin{aligned}\int_{-\infty}^{+\infty} f(\eta)d\eta &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{-z}\sqrt{2}\, dz \\ &= \frac{\sqrt{2}}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{-z}\, dz \\ &= \frac{\sqrt{2}}{\sqrt{\pi}} \cdot \sqrt{\pi} = \sqrt{2}\end{aligned}$$

領域すべての積分値を 1 にするには、関数を $\sqrt{2}$ で割っておけば良いので、求める確率密度関数は

$$g(\eta) = \frac{1}{\sqrt{\pi}} e^{-\frac{\eta^2}{2}} \qquad \text{を } \sqrt{2} \text{ で割って、} \qquad f(\eta) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{\eta^2}{2}}$$

以上によって、標準正規分布が導出できた。

9　共分散行列

9.1　期待値と平均・分散・共分散

■期待値

最初に期待値の概念を拡張しておく。

定義 9.1.　【関数の期待値】

$\varphi(X)$ を確率変数 X の関数とすると $\varphi(x)$ の期待値を以下のように定義し、$\varphi(X)$ の期待値と呼ぶ。

$$\text{離散値}\quad E[\varphi(X)] = \sum_{i=1}^{n} \varphi(x_i) f(x_i) \tag{9.1}$$

$$\text{連続値}\quad E[\varphi(X)] = \int_{-\infty}^{\infty} \varphi(x) f(x) dx \tag{9.2}$$

期待値をこのように関数に関する期待値として一般化しておくと、その特殊な場合として $\varphi(X) = X$ の場合の期待値が以下のように平均となる。

定義 9.2.　【期待値と平均】

$$\text{離散値}\quad E[X] = \sum_{i=1}^{n} x_i f(x_i) \tag{9.3}$$

$$\text{連続値}\quad E[X] = \int_{-\infty}^{\infty} x f(x) dx \tag{9.4}$$

同様に、２変数関数 $f(x, y)$ の場合の期待値も、X、Y に関する関数を $\varphi(x, y)$ とすると以下になる。

$$E[X, Y] = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \varphi(x, y) f(x, y) dx \right) dy = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \varphi(x, y) f(x, y) dy \right) dx$$

期待値は以下のような線形性を持っている。

定義 9.3.　【期待値の線形性】

確率変数 X と Y と任意の定数 a、b について以下の線形性が成り立つ。

$$E[aX + bY] = aE[X] + bE[Y] \tag{9.5}$$

これによって以下のように、確率変数 X に関する新たな関数 $\varphi(X) = 3x + 3x^3$ の期待値を計算する事が簡

易になる。

$$E[2X+3X^2] = \int_{-\infty}^{\infty}(2x+3x^2)f(x)dx$$
$$= 2\int_{-\infty}^{\infty} xf(x)dx + 3\int_{-\infty}^{\infty} x^3 f(x)dx = 2E[X] + 3E[X^3]$$

つまり、ある確率変数 X があった時、その変数を使って様々な変数変換した場合の期待値を計算できるようになる。

例題 9.1. 確率変数 X と Y の同時分布の確率密度関数 $f(x,y)$ が以下の式で与えられている時、関数 $\varphi(x,y) = xy$ の期待値 $E[XY]$ を求めよ。

$$f(x,y) = \begin{cases} x+y & (0 \leq x \leq 1 \text{ かつ } 0 \leq y \leq 1) \\ 0 & (\text{その他}) \end{cases}$$

$$E[XY] = E[\varphi(x,y)] = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\varphi(x,y)f(x,y)dx\right)dy$$
$$= \int_0^1\left(\int_0^1 xy(x+y)dx\right)dy = \int_0^1\left(\int_0^1 x^2y + xy^2\,dx\right)dy = \int_0^1\left[\frac{y}{3}x^3 + \frac{y^2}{2}x^2\right]_{x=0}^{x=1}dy$$
$$= \int_0^1\left(\frac{y}{3}+\frac{y^2}{2}\right)dy = \left[\frac{1}{6}y^2 + \frac{1}{6}y^3\right]_{x=0}^{x=1} = \frac{1}{6}+\frac{1}{6} = \frac{1}{3}$$

■分散

ついで期待値と分散の定義。

定義 9.4. 【期待値と分散】

ある確率変数 X の平均を μ とすると分散は以下のように表すことができる。

$$V[X] = E[(X-\mu)^2] \tag{9.6}$$
$$V[X] = E[X^2] - \{E[X]\}^2 \tag{9.7}$$

確率関数を $f(x)$ として期待値を展開すると、以下のようなる。

$$\text{離散値} \quad V[X] = \sum_{i=1}^{n}(x_i-\mu)^2 f(x_i) \tag{9.8}$$
$$\text{連続値} \quad V[X] = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx \tag{9.9}$$

式 (9.7) は、以下のように式 (9.6) から導く事ができる。

$$\begin{aligned} V[X] = E[(X-\mu)^2] &= E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - 2E[X]E[X] + \{E[X]\}^2 \\ &= E[X^2] - \{E[X]\}^2 \end{aligned}$$

■共分散

期待値と共分散の定義。

定義 9.5. 【共分散】

2 つの確率変数 X、Y の期待値がそれぞれ μ、ν だったとする。この時 X と Y との共分散 *(covariance)* を以下のように定義する

$$Cov[X,Y] = E[(X-\mu)(Y-\nu)] \tag{9.10}$$

つまり、離散値及び連続値の共分散は以下のように計算する事ができる

$$\text{離散値} \quad Cov[X,Y] = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-\mu)(y_i-\nu)f(x_i,x_j)$$

$$\text{連続値} \quad Cov[X,Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu)(y-\nu)f(x,y)\,dx\,dy$$

その他以下のような表記をする事がある

$$Cov(X,Y)\,, \quad \sigma(X,Y)\,, \quad \sigma_{xy}\,, \quad s_{xy}$$

特に、標本に関する統計量を意味する場合は σ_{xy} や s_{xy} を使う場合が多い。

定義 9.6. 【共分散の別定義】

共分散 *(covariance)* は以下のように定義する場合もある

$$Cov[X,Y] = E[XY] - E[X][Y] \tag{9.11}$$

これを共分散の定義とする場合もあるが、定義式（9.22）から導く事ができる。まずは定義より

$$Cov[X,Y] = E[(X-\mu)(Y-\nu)] = E[XY - \nu X - \mu Y + \mu\nu]$$

期待値の線形性の式（9.5）より

$$\begin{aligned} E[XY - \nu X - \mu Y + \mu\nu] &= E[XY] - \nu E[X] - \mu E[Y] + \mu\nu \\ &= E[XY] - E[Y]\ E[X] - E[X]\ E[Y] + E[X]\ E[Y] \\ &= E[XY] - E[Y]\ E[X] \end{aligned}$$

例題 9.2. $(X,Y) = (-6,-7),(8,-5),(-4,7),(10,9)$ がいずれも確率 $\frac{1}{4}$ で出る。このときの共分散 $Cov[X,Y]$ を求めよ。

$$\mu = E[X] = -6 \times \frac{1}{4} + 8 \times \frac{1}{4} - 4 \times \frac{1}{4} + 10 \times \frac{1}{4} = 2$$
$$\nu = E[Y] = -7 \times \frac{1}{4} - 5 \times \frac{1}{4} + 7 \times \frac{1}{4} + 9 \times \frac{1}{4} = 1$$

なので

$$\begin{aligned} Cov[X,Y] &= E[(X-\mu)(Y-\nu)] \\ &= (-6-2)(-7-1) \times \frac{1}{4} + (8-2)(-5-1) \times \frac{1}{4} + (-4-2)(7-1) \times \frac{1}{4} + (10-2)(9-1) \times \frac{1}{4} \\ &= 14 \end{aligned}$$

例題 9.3. 確率変数 X、Y の確率密度関数 (X,Y) が以下の式で与えられている時の共分散 $Cov[X,Y]$ を求めよ。

$$f(x,y) = \begin{cases} x+y & (0 \le x \le 1 \text{ かつ } 0 \le y \le 1) \\ 0 & (\text{その他}) \end{cases}$$

$$\begin{aligned} \mu = E[X] &= \int_0^1 \int_0^1 x(x+y)\,dx\,dy = \int_0^1 \int_0^1 x^2 + xy\,dx\,dy = \int_0^1 \left[\frac{1}{3}x^3 + \frac{1}{2}x^2 y\right]_{x=0}^{x=1} dy \\ &= \int_0^1 \left(\frac{1}{3} + \frac{1}{2}y\right) dy = \left[\frac{1}{3}y + \frac{1}{4}y^2\right]_{y=0}^{y=1} = \frac{7}{12} \end{aligned}$$

$$\begin{aligned} \nu = E[Y] &= \int_0^1 \int_0^1 y(x+y)\,dx\,dy = \int_0^1 \int_0^1 xy + y^2\,dx\,dy = \int_0^1 \left[\frac{1}{2}x^2 y + xy^2\right]_{x=0}^{x=1} dy \\ &= \int_0^1 \left(\frac{1}{2}y + y^2\right) dy = \left[\frac{1}{4}y^2 + \frac{1}{3}y^3\right]_{y=0}^{y=1} = \frac{7}{12} \end{aligned}$$

以下の積分と式の展開はかなり面倒なので、ソースコード (11) のような python の sympy を使った。

$$
\begin{aligned}
Cov[X,Y] = E[(X-\mu)(Y-\nu)] &= \int_0^1 \int_0^1 (x-\mu)(y-\nu)(x+y)\, dx\, dy \\
&= \int_0^1 \int_0^1 \left(x-\frac{7}{12}\right)\left(y-\frac{7}{12}\right)(x+y)\ dx\, dy \\
&= \int_0^1 \left[x^3\left(\frac{y}{3}-\frac{7}{36}\right)+x^2\left(\frac{y^2}{2}-\frac{7y}{12}+\frac{49}{288}\right)+x\left(-\frac{7y^2}{12}+\frac{49y}{144}\right)\right]_{x=0}^{x=1} dy \\
&= \int_0^1 \left(-\frac{y^2}{12}+\frac{13y}{144}-\frac{7}{288}\right)\ dy \\
&= \left[-\frac{y^3}{36}+\frac{13y^2}{288}-\frac{7y}{288}\right]_{x=0}^{x=1} \\
&= -\frac{1}{144}
\end{aligned}
$$

ソースコード 11　sympy を使った計算過程

```
import sympy as sp
from fractions import Fraction  # 分須Fraction クラスをインポート

x,y = sp.symbols("x␣y") #シンボルを定義するには SymPy のシンボルクラスを使う

#最初の式
eq1 = (x−Fraction(7,12))∗(y−Fraction(7,12))∗(x+y) ; print(sp.latex(eq1))
#x で積分
eq2 = sp.integrate(eq1,x) ; print(sp.latex(eq2))
#その不定積分に対して x=1を代入
eq3 = eq2.subs(x,1) ; print(sp.latex(eq3))
#その結果を y で積分
eq4 = sp.integrate(eq3,y) ; print(sp.latex(eq4))
#その不定積分に対して y=1を代入
eq5 = eq4.subs(y,1) ; print(sp.latex(eq5))
```

【共分散の性質】

X、Y、Z を確率変数とし a、b を定数とする時、共分散に関して以下が成立する

交換則	$Cov[X,Y] = Cov[Y,X]$	(9.12)
分配則	$Cov[X,Y+Z] = Cov[X,Y] + Cov[X,Z]$	(9.13)
定数加算	$Cov[X+a,Y+b] = Cov[X,Y]$	(9.14)
定数倍	$Cov[aX,bY] = ab\,Cov[X,Y]$	(9.15)

■交換則

定義より $Cov[Y,X] = E[(Y-\nu)(X-\mu)]$ は、$Cov[X,Y] = E[(X-\mu)(Y-\nu)]$ と同じ。

■分配則

期待値の別定義 (式 9.11) である $Cov[X,Y] = E[XY] - E[X][Y]$ と期待値の線形性（式 9.5）である $E[aX+bY] = aE[X] + bE[Y]$ を使う。

$$\begin{aligned} Cov[X,Y+Z] &= E[X(Y+Z)] - E[X]\,E[Y+Z] \\ &= E[XY+XZ] - E[X]\,(E[Y]+E[Z]) \\ &= E[XY] + E[XZ] - E[X]\,E[Y] - E[X]\,E[Y] \\ &= \{E[[XY] - E[X]\,E[Y]\}\,\{E[XZ] - E[X]\,E[Z]\} \\ &= Cov[X,Y] + Cov[X,Z] \end{aligned}$$

■定数加算

以下より $Cov[X',Y'] = E[(X'-E[X'])(Y'-E[Y'])] = E[(X-\mu)(Y-\nu)] = Cov[X,Y]$ は明らか。
つまり、$Cov[X+a,Y+b] = Cov[X,Y]$

$$\begin{aligned} X' = X+a \quad \text{とすると、} \quad E[X'] = \mu + a \text{ なので、} \quad X' - E[X'] = (X-\mu) \\ Y' = Y+b \quad \text{とすると、} \quad E[Y'] = \nu + b \text{ なので、} \quad Y' - E[Y'] = (Y-\mu) \end{aligned}$$

■定数倍

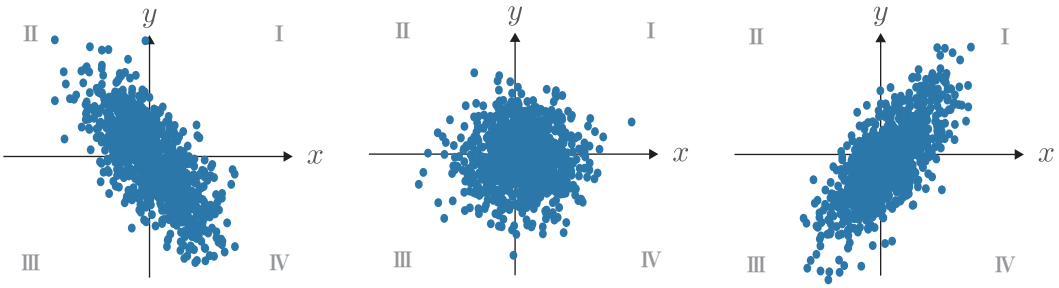
$Cov[aX,bY] = E[a(X-\mu)b(Y-\nu)] = abE[(X-\mu)(Y-\nu)] = ab\,Cov[X,Y]$ より明らか。

9.2　相関係数

■相関係数の定義

共分散 $Cov[X,Y]$ は X と Y との関係を表す。図 48 のように、X と Y との共分散は $Cov[X,Y] = E[(X-\mu)(Y-\nu)]$ なので、相関係数の符号は $(X-\mu)$ と $(Y-\nu)$ の符号によって以下のようになる。

- (a)$(X-\mu)$ と $(Y-\nu)$ の符号が反対が多い（Ⅱ 象限と Ⅳ 象限が優位）場合はマイナス
- (b)$(X-\mu)$ と $(Y-\nu)$ の符号の反対と同じが均等（I 象限〜Ⅳ 象限が均等）の場合はゼロ
- (c)$(X-\mu)$ と $(Y-\nu)$ の符号が同じが多い（I 象限と Ⅲ 象限が優位）場合はプラス



(a) マイナス：ⅡとⅣが優勢　　(b) ゼロ：I ～Ⅳが均等　　(c) プラス：I とⅢが優勢

図 48　分布の傾向と共分散・相関

ただし、2 つの変数の関係の違いを共分散の大きさだけで比較する事はできない。何故なら式 (9.15) のように定数倍をした場合は共分散の大きさも定数倍される。なので、例えば $Cov[aX,aY] = a^2\,Cov[X,Y]$ のように 2 つの変数を a 倍したとすると、共分散の大きさは a^2 倍されるが、これはスケールが変化しただけで 2 つの変数の関係は同じである。

こうしたスケールの問題を除いて 2 つの変数の関係だけを考えるためには 2 つの変数の縮尺を揃える必要がある。この縮尺を揃えるために用いられる方法が標準化と言われる以下のような方法である。

定義 9.7.　【標準化】
次の式のように、データを平均 0 標準偏差が 1 になるように変換する方法を標準化と呼ぶ。

$$X' = \frac{X - E[X]}{\sqrt{V[X]}}\,, \quad Y' = \frac{Y - E[Y]}{\sqrt{V[Y]}} \tag{9.16}$$

標準化すると平均 0 で分散 1 になる事を示そう。まずは上記の式の $E[X]\,, E[Y]\,, \sqrt{V(X)}\,, \sqrt{V(Y)}$ は定数なので、簡略化のために $E[X] = \mu\,,\quad E[Y] = \nu\,,\quad \sqrt{V(X)} = \sigma_X\,,\quad \sqrt{V(Y)} = \sigma_Y$ のように書くと以下のように書き換えられる。

$$X' = \frac{X - \mu}{\sigma_X}\,, \quad Y' = \frac{Y - \nu}{\sigma_Y}$$

- **平均が 0 になる事を示す**

  定数 μ の期待値は μ に他ならない。つまり $E[\mu] = \mu$ なので、

$$E[X'] = E\left[\frac{X-\mu}{\sigma_X}\right] = \frac{1}{\sigma_X}\{E[X] - E[\mu]\} = \frac{1}{\sigma_X}\{\mu - \mu\} = 0$$

  同様にして $E[Y'] = 0$ となる。

- **分散が 1 になる事を示す**

  $\sigma_X^2 = V[X] = E[(X-\mu)(X-\mu)]$ を利用する。

$$V[X'] = E\left[\frac{X-\mu}{\sigma_X}\cdot\frac{X-\mu}{\sigma_X}\right] = \frac{E[(X-\mu)(X-\mu)]}{\sigma_X^2} = \frac{E[(X-\mu)(X-\mu)]}{E[(X-\mu)(X-\mu)]} = 1$$

  同様にして $V[Y'] = 1$ となる。

つまり、標準化によって平均 0 で標準偏差 1 になるように縮尺を揃える事によって、2 つの変数の関係を比較できるようにしたのが相関係数である。

定義 9.8. 【相関係数】

相関係数は以下のように定義される。

$$\rho_{XY} = Cov[X', Y'] = \frac{Cov[X,Y]}{\sqrt{V[X]}\sqrt{V[Y]}} = \frac{Cov[X,Y]}{\sigma_X\sigma_Y} \tag{9.17}$$

この変数 X'、Y' は変数 X、Y を標準化した変数。つまり

$$X' = \frac{X-\mu}{\sigma_X}, \quad Y' = \frac{Y-\nu}{\sigma_Y}$$

相関係数は r_{xy} とあらわす場合がある。とくに標本に関する統計量を扱う場合は、共分散を s_{xy}、それぞれの標準偏差を s_x、s_y、相関係数は r_{xy} または r とあらわす事が多い。

式 (9.17) の証明には、共分散の別定義式 (9.11)　$Cov[X,Y] = E[XY] - E[X][Y]$　及び、上記で示した $E[X'] = 0$、$E[Y'] = 0$ を使う。まず X、Y を標準化した変数を X'、Y' とする。

$$\begin{aligned}\rho_{XY} &= Cov[X', Y'] \\ &= E[X'Y'] - E[X'][Y'] = E[X'Y'] \\ &= E\left[\frac{X-\mu}{\sigma_X}\cdot\frac{Y-\nu}{\sigma_Y}\right] = \frac{E[(X-\mu)(Y-\mu)]}{\sigma_X\sigma_Y} = \frac{Cov[X,Y]}{\sigma_X\sigma_Y}\end{aligned}$$

このように、相関係数とは共分散 $Cov[X,Y]$ を 2 つの変数の標準偏差 σ_X と σ_Y とで割ったものになる。

例題 9.4. $(X,Y)=(-6,-7),(8,-5),(-4,7),(10,9)$ がいずれも確率 $\frac{1}{4}$ で出る。このときの相関係数 ρ_{XY} を求めよ。

式 9.1 のように、各期待値は確率関数が $f(x_i)$ なので、$E[\varphi(X)]=\sum_{i=1}^{n}\varphi(x_i)f(x_i)$ となる。

$$\mu=E[X]=-6\times\frac{1}{4}+8\times\frac{1}{4}-4\times\frac{1}{4}+10\times\frac{1}{4}=2$$
$$\nu=E[Y]=-7\times\frac{1}{4}-5\times\frac{1}{4}+7\times\frac{1}{4}+9\times\frac{1}{4}=1$$

$$\sigma_X^2=E[(X-\mu)^2]=64\times\frac{1}{4}+36\times\frac{1}{4}+36\times\frac{1}{4}+64\times\frac{1}{4}=50$$
$$\sigma_Y^2=E[(Y-\nu)^2]=64\times\frac{1}{4}+36\times\frac{1}{4}+36\times\frac{1}{4}+64\times\frac{1}{4}=50$$

$$\begin{aligned}Cov[X,Y]&=E[(X-\mu)(Y-\nu)]\\&=(-6-2)(-7-1)\frac{1}{4}+(8-2)(-5-1)\frac{1}{4}+(-4-2)(7-1)\frac{1}{4}+(10-2)(9-1)\frac{1}{4}=14\end{aligned}$$

$$\rho_{XY}=\frac{Cov[X,Y]}{\sigma_X\cdot\sigma_Y}=\frac{14}{\sqrt{50}\cdot\sqrt{50}}=\frac{14}{50}=0.28$$

例題 9.5. 確率変数 X、Y の確率密度関数 (X,Y) が以下の式で与えられている時の相関係数 ρ_{XY} を求めよ。

$$f(x,y)=\begin{cases}x+y & (0\le x\le 1 \text{ かつ } 0\le y\le 1)\\ 0 & (\text{その他})\end{cases}$$

求めたい相関係数は

$$\rho_{XY}=\frac{Cov[X,Y]}{\sigma_X\sigma_Y}$$

まず、X と Y との平均を求めると

$$\begin{aligned}\mu=E[X]&=\int_0^1\int_0^1 x(x+y)\,dx\,dy=\int_0^1\int_0^1 x^2+xy\,dx\,dy=\int_0^1\left[\frac{1}{3}x^3+\frac{1}{2}x^2y\right]_{x=0}^{x=1}dy\\&=\int_0^1\left(\frac{1}{3}+\frac{1}{2}y\right)dy=\left[\frac{1}{3}y+\frac{1}{4}y^2\right]_{y=0}^{y=1}=\frac{7}{12}\end{aligned}$$

$$\begin{aligned}\nu=E[Y]&=\int_0^1\int_0^1 y(x+y)\,dx\,dy=\int_0^1\int_0^1 xy+y^2\,dx\,dy=\int_0^1\left[\frac{1}{2}x^2y+xy^2\right]_{x=0}^{x=1}dy\\&=\int_0^1\left(\frac{1}{2}y+y^2\right)dy=\left[\frac{1}{4}y^2+\frac{1}{3}y^3\right]_{y=0}^{y=1}=\frac{7}{12}\end{aligned}$$

次に、分散 σ_X^2 は式 (9.7)、つまり $V[X]=E[X^2]-\{E[X]\}^2$ を用いて算出する。

$$\rho_X^2 = E[X^2] - \mu^2 = \int_0^1 \int_0^1 x^2(x+y)\,dx\,dy - \mu^2 = \int_0^1 \int_0^1 (x^3 + x^2 y)\,dx\,dy - \mu^2$$
$$= \int_0^1 \left[\frac{1}{4}x^4 + \frac{1}{3}yx^3\right]_{x=0}^{x=1} dy - \mu^2 = \int_0^1 \left(\frac{1}{4} + \frac{1}{3}y\right) dy - \mu^2 = \left[\frac{1}{4}y + \frac{1}{6}y^2\right]_{y=0}^{y=1} - \mu^2$$
$$= \frac{10}{24} - \left(\frac{7}{12}\right)^2 = \frac{11}{144}$$

$$\rho_Y^2 = E[Y^2] - \nu^2 = \int_0^1 \int_0^1 y^2(x+y)\,dx\,dy - \nu^2 = \int_0^1 \int_0^1 (y^2 x + y^3)\,dx\,dy - \nu^2$$
$$= \int_0^1 \left[\frac{1}{2}y^2x^2 + y^3 x\right]_{x=0}^{x=1} dy - \nu^2 = \int_0^1 \left(\frac{1}{2}y^2 + y^3\right) dy - \nu^2 = \left[\frac{1}{6}y^3 + \frac{1}{4}y^4\right]_{y=0}^{y=1} - \nu^2$$
$$= \frac{10}{24} - \left(\frac{7}{12}\right)^2 = \frac{11}{144}$$

$$Cov[X,Y] = E[(X-\mu)(Y-\nu)] = \int_0^1 \int_0^1 (x-\mu)(y-\nu)(x+y)\,dx\,dy$$
$$= \int_0^1 \int_0^1 \left(x - \frac{7}{12}\right)\left(y - \frac{7}{12}\right)(x+y)\;\,dx\,dy$$
$$= \int_0^1 \left[x^3\left(\frac{y}{3} - \frac{7}{36}\right) + x^2\left(\frac{y^2}{2} - \frac{7y}{12} + \frac{49}{288}\right) + x\left(-\frac{7y^2}{12} + \frac{49y}{144}\right)\right]_{x=0}^{x=1} dy$$
$$= \int_0^1 \left(-\frac{y^2}{12} + \frac{13y}{144} - \frac{7}{288}\right) dy$$
$$= \left[-\frac{y^3}{36} + \frac{13y^2}{288} - \frac{7y}{288}\right]_{x=0}^{x=1}$$
$$= -\frac{1}{144}$$

以上より

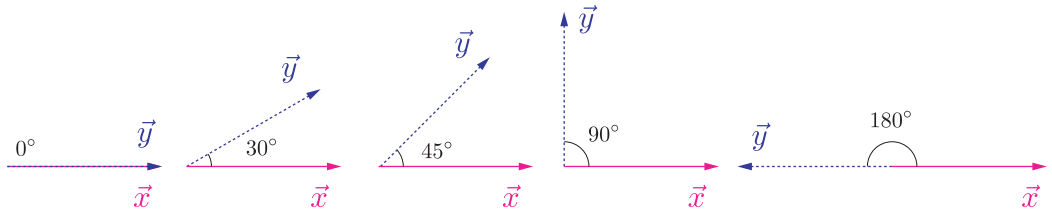
$$\rho_{XY} = \frac{Cov[X,Y]}{\sigma_X \sigma_Y} = \frac{-\frac{1}{144}}{\sqrt{\frac{11}{144}} \cdot \sqrt{\frac{11}{144}}} = -\frac{1}{144} \cdot \frac{144}{11} = -\frac{1}{11}$$

■相関係数と内積

相関係数は、平均偏差ベクトル $\mathbf{x}$ と $\mathbf{y}$ のなす角度 θ の余弦 $\cos\theta$ に等しい

$$\cos\theta = \frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{x}||\mathbf{y}|} = \frac{Cov[X,Y]}{\sqrt{V[X]}\sqrt{V[Y]}} = \frac{\sigma_{XY}}{\sigma_X\cdot\sigma_Y} = \rho_{XY} \tag{9.18}$$

標準偏差で割っているので 2 つのベクトルの長さは 1 で、そのなす角度と相関係数の関係は以下。



$\rho_{XY} = 1$　　$\rho_{XY} = \frac{\sqrt{3}}{2} \fallingdotseq 0.866$　　$\rho_{XY} = \frac{\sqrt{2}}{2} \fallingdotseq 0.707$　　$\rho_{XY} = 0$　　$\rho_{XY} = -1$

相関係数のとり得る範囲は $-1 \leq \cos\theta \leq 1$ なので、相関係数 ρ_{XY} も

$$-1 \leq \rho_{XY} \leq 1 \tag{9.19}$$

式 (9.18) を導出していく。まず、確率変数 X と Y から、それぞれの平均 $E[X] = \mu$、$E[Y] = \nu$ を引いて並べた平均偏差ベクトルをつくり、以下のように $\mathbf{x}$、$\mathbf{y}$ とする。

$$\mathbf{x} = \begin{pmatrix} X_1 - \mu \\ X_2 - \mu \\ \vdots \\ X_n - \mu \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} Y_1 - \nu \\ Y_2 - \nu \\ \vdots \\ Y_n - \nu \end{pmatrix}$$

このベクトル $\mathbf{x}$ と $\mathbf{y}$ との内積は、そのなす角度を θ とすると以下のように表す事ができる。内積の定義式については式 (付録 D.4) 参照。

$$\mathbf{x}\cdot\mathbf{y} = <\mathbf{x},\mathbf{y}> = |\mathbf{x}||\mathbf{y}|\cos\theta$$

上記の式を変形して

$$\cos\theta = \frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{x}||\mathbf{y}|} \tag{9.20}$$

この右辺を共分散 σ_{XY} や標準偏差 σ_x, σ_Y で表していく。そのために内積が成分の積和である事を利用

する。

$$
\begin{aligned}
\mathbf{x}\cdot\mathbf{y} &= (X_1-\mu)(Y_a-\nu)+(X_2-\mu)(Y_2-\nu)+\cdots+(X_n-\mu)(Y_n-\nu)\\
&= \sum_{i=1}^{i=n}(X_i-\mu)(Y_i-\nu) = Cov[X,Y] = \sigma_{XY}\\
\mathbf{x}\cdot\mathbf{x} &= (X_1-\mu)(X_a-\nu)+(X_2-\mu)(X_2-\nu)+\cdots+(X_n-\mu)(X_n-\nu)\\
&= \sum_{i=1}^{i=n}(X_i-\mu)^2 = V[X] = \sigma_X^2 \qquad (\mathbf{x}\cdot\mathbf{x}=|\mathbf{x}|^2 \text{ なので } |\mathbf{x}|=\sigma_X)\\
\mathbf{y}\cdot\mathbf{y} &= (Y_1-\mu)(Y_a-\nu)+(Y_2-\mu)(Y_2-\nu)+\cdots+(Y_n-\mu)(Y_n-\nu)\\
&= \sum_{i=1}^{i=n}(Y_i-\mu)^2 = V[Y] = \sigma_Y^2 \qquad (\mathbf{y}\cdot\mathbf{y}=|\mathbf{y}|^2 \text{ なので } |\mathbf{y}|=\sigma_Y)
\end{aligned}
$$

式 (9.20) にこれらの値を代入すると

$$
\cos\theta = \frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{x}||\mathbf{y}|} = \frac{Cov[X,Y]}{\sqrt{V[X]}\sqrt{V[Y]}} = \frac{\sigma_{XY}}{\sigma_X\cdot\sigma_Y} = \rho_{XY} \tag{9.21}
$$

9.3 共分散行列

2 変数での相関係数を多変数の相関行列に拡張していく。

■共分散行列を求める

統計的な標本データの観点から説明する。n 個のサンプルのそれぞれについて m 項目の測定をし、結果として $(n \times m)$ 個の測定値が得られたとする。図のように表形式で示す場合は、縦にサンプル、横に測定項目をとる。一般にサンプル数の方が項目数より多いのでこの方が利便性が良い。

サンプル (サンプル数:n)	項目 (項目数:m) 1	2	⋯	m
1	X_{11}	X_{12}	⋯	X_{1m}
2	X_{21}	X_{22}	⋯	X_{2m}
⋮	⋮	⋮	⋱	⋮
n	X_{n1}	X_{n2}	⋯	X_{nm}

図 49　計測によって得られた生データ

このデータをそのまま行列表示したものを素得点行列 X_0 とする。

$$X_0 = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1m} \\ X_{21} & X_{22} & \cdots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nm} \end{pmatrix}$$

定義 9.9.　【基準化得点行列】

このデータ行列 X_0 の各要素を $x_{ij} = X_{ij} - \mu_j$ のように変換した行列を平均偏差得点行列と呼ぶ。

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} = \begin{pmatrix} X_{11} - \mu_1 & X_{12} - \mu_2 & \cdots & X_{1m} - \mu_m \\ X_{21} - \mu_1 & X_{22} - \mu_2 & \cdots & X_{2m} - \mu_m \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} - \mu_1 & X_{n2} - \mu_2 & \cdots & X_{nm} - \mu_m \end{pmatrix} \tag{9.22}$$

定義 9.10.　【共分散行列】

平均偏差得点行列 X に対して以下のような演算をしたものを共分散行列、または分散共分散行列と呼ぶ。

$$C = \frac{1}{n} X^t X = \frac{1}{n} \begin{pmatrix} \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} & \cdots & \sum_i x_{i1}x_{in} \\ \sum_i x_{i2}x_{i1} & \sum_i x_{i2}^2 & \cdots & \sum_i x_{i2}x_{in} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_i x_{in}x_{i1} & \sum_i x_{in}x_{i2} & \cdots & \sum_i x_{in}^2 \end{pmatrix} = \begin{pmatrix} s_1^2 & s_{12}^2 & \cdots & s_{1n}^2 \\ s_{21}^2 & s_2^2 & \cdots & s_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1}^2 & s_{n2}^2 & \cdots & s_n^2 \end{pmatrix} \tag{9.23}$$

分散行列 C の対角要素は、項目 j（列）の分散となる。

$$s_j^2 = \frac{1}{n}\sum_i (X_{ij} - m_j)^2 = \frac{1}{n}\sum_i x_{ij}^2$$

また j 行 k 列の要素は、項目 j と項目 k（列）の共分散となる。

$$s_{jk}^2 = \frac{1}{n}\sum_i (X_{ij} - m_j)(X_{ik} - m_k) = \frac{1}{n}\sum_i x_{ij}x_{ik}$$

ついで、確率変数ベクトルという視点から考えてみる。つまり、複数の変数の確率分布があって、その分布に従う複数の確率変数をひとまとめにしてベクトルとして扱っていく。まず、n 個の確率変数 $X_1, X_2, \cdots, X_n$ を並べた列ベクトルを確率変数ベクトルと定義する。この確率変数ベクトルは、確率変数を並べたもので、確率の値を並べたものではないので注意。

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

この確率変数ベクトル X の期待値は以下。$E[X]$ は平均値を縦に並べたベクトルとなる。

【確率変数ベクトルの期待値】

$$E[X] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{pmatrix} \tag{9.24}$$

この確率変数ベクトル X の共分散行列は以下。単に行列の分散ともいい $V[X]$ と表記する。また、$\sum$ と表記する場合もある。

【確率変数ベクトルの共分散行列】

$$V[X] = E[(X-\mu)(X-\mu)^T] \tag{9.25}$$

$$= \begin{pmatrix} V[X_1] & Cov[X_2, X_1] & \cdots & Cov[X_n, X_1] \\ Cov[X_1, X_2] & V[X_2] & \cdots & Cov[X_n, X_2] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[X_1, X_n] & Cov[X_2, X_n] & \cdots & V[X_n] \end{pmatrix}$$

ただし μ は平均ベクトルで $\mu = E[X]$

公式 9.1. 【共分散行列の別の求め方】

縦ベクトルで表記した確率変数 X に対して以下が成立する。

$$V[X] = E[XX^T] - E[X]E[X]^T \tag{9.26}$$

分散の別の計算式 $V[X] = E[X^2] - \{E[X]\}^2$ に該当するベクトル版と考えればよい。これを確認していくまえに、行列についての演算法則と確率行列の期待値の演算法則について確認しよう。

- **行列の演算法則の確認**

  A、B、C を同じ次元の行列、x、y を縦ベクトル、k をスカラーの定数とした場合、以下が成立する

結合則	$(AB)C = A(BC)$
分配則	$A(B+C) = AB + AC$
交換則は成立しない	$AB \neq BA$
線形性 (1)	$A(x+y) = Ax + Ay$
線形性 (2)	$A(kx) = k(Ax)$

- **確率行列の期待値の法則の確認**

  X と Y を確率変数を成分とする確率行列、A を成分が定数の定数行列、y を縦ベクトル、k をスカラーの定数とした場合、以下が成立する

定数倍	$E[kX] = k\,E[X]$
確率行列同士の和	$E[X+Y] = E[X] + E[Y]$
定数行列との和	$E[X+A] = E[X] + A$
定数行列の期待値	$E[A] = A$
転置行列の期待値	$E[X^T] = E[X]^T$
内積の期待値	$E[a^T X] = a^T E[X]$

- **式 9.26 の確認**

  $E[X] = \mu$ と置く、μ は定数の縦ベクトルになる。

$$\begin{aligned} V[X] &= E[(X-\mu)(X-\mu)^T] = E[(X-\mu)(X^T-\mu^T)] = E[XX^T - X\mu^T - \mu X^T + \mu\mu^T] \\ &= E[XX^T] - E[X]\mu^T - \mu E[X^T] + \mu\mu^T = E[XX^T] - \mu\mu^T - \mu\mu^T + \mu\mu^T \\ &= E[XX^T] - \mu\mu^T = E[XX^T] - E[X]E[X]^T \end{aligned}$$

この確率変数ベクトルの確率密度関数と期待値について考えよう。2 変数の同時分布の確率密度関数については式 (7.7) でしめした。これを多変数に拡大しよう。

- **2 変数の場合の復習**

  まず 2 変数の場合は、107 ページで示したように、連続型の 2 つの確率変数 X、Y について、$a \leqq X \leqq b$ かつ $c \leqq y \leqq c$ となる確率 $P(a \leqq x \leqq b, c \leqq y \leqq d)$ が以下の式で示される時、$f(x,y)$ を変数 X、Y

の同時分布の確率密度関数と定義した。

$$P(a \leqq x \leqq b, c \leqq y \leqq d) = \int_c^d \int_a^b f(x,y)dx\, dy$$

これを図示したのが下図。図の赤の柱の体積が確率 $P(a \leqq x \leqq b, c \leqq y \leqq d)$ を表している。

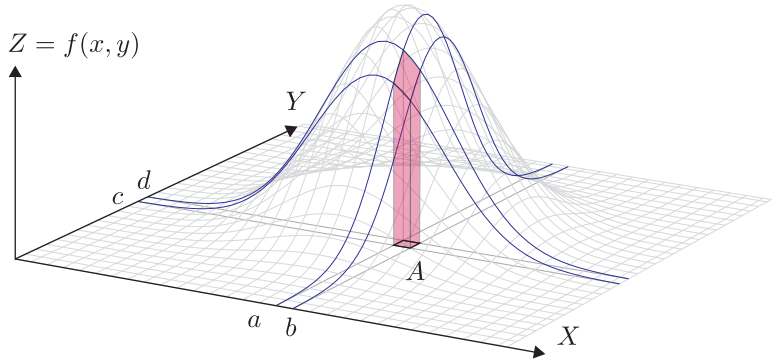


図 50　連続型 2 変数の確率密度関数と確率

- **多変数の場合の確率密度の表記**

  以下のように、n 個の確率変数を成分に持つ確率変数ベクトルを $\mathbf{x}$ と表記し、ベクトル $\mathbf{x}$ を入力として一つの値を返す確率密度関数を $f_X(\mathbf{x})$ と表記する。

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad , \qquad f_X(\boldsymbol{x}) = f_{X_1,X_2,\cdots,X_n} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

  実際に確率を求めるには以下のように重積分をする

$$P(X \text{ がある範囲 } D \text{ にはいる}) = \int \cdots \int F_X(\boldsymbol{x})dx_1 \cdots dx_n$$

  これを簡略化して以下のように表記する。この R^n は n 次元のベクトル空間全体にわたって積分する事を意味する。

$$P(X \text{ がある範囲 } D \text{ にはいる}) = \int_{R^n} f_X(\boldsymbol{x})\, d\boldsymbol{x}$$

公式 9.2.　【期待値】ベクトル確率変数 $\boldsymbol{x}$ の期待値 $E[\boldsymbol{x}]$ は以下。また、$f_X(\boldsymbol{x})$ の結果のベクトル値に関数 $g(\boldsymbol{x})$ を施した結果の期待値 $E[g(\boldsymbol{x})]$ は

$$E[\boldsymbol{x}] = \int_{R^n} \boldsymbol{x} f_X(\boldsymbol{x}) d\boldsymbol{x} \tag{9.27}$$

$$E[g(\boldsymbol{x})] = \int_{R^n} g(\boldsymbol{x}) f_X(\boldsymbol{x}) d\boldsymbol{x} \tag{9.28}$$

■共分散行列の一次変換

公式 9.3. 【共分散行列の一次変換】

$\boldsymbol{x}$ を確率変数ベクトルとすると、定数行列 A によって $\boldsymbol{x}$ を一次変換した時の分散行列は以下となる。

$$V[A\boldsymbol{x}] = A\,V[\boldsymbol{x}]\,A^T \tag{9.29}$$

$A\boldsymbol{x}$ は以下のような $m \times n$ の行列 A による一次変換である。

$$A\boldsymbol{x} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$A\boldsymbol{x}$ の期待値は、確率変数でない定数の行列 A は外に出す事ができて

$$E[A\boldsymbol{x}] = AE[\boldsymbol{x}] = A\mu$$

これを用いて、共分散行列を求める[*31]と

$$\begin{aligned} V[A\boldsymbol{x}] &= E[(A\boldsymbol{x} - A\mu)(A\boldsymbol{x} - A\mu)^T] \\ &= E[A(\boldsymbol{x} - \mu)\{A(\boldsymbol{x} - \mu)\}^T] = E[A(\boldsymbol{x} - \mu)\left\{(\boldsymbol{x} - \mu)^T A^T\right\}] \\ &= E[A(\boldsymbol{x} - \mu)(\boldsymbol{x} - \mu)^T A^T] = AE[(\boldsymbol{x} - \mu)(\boldsymbol{x} - \mu)^T]A^T \end{aligned}$$

この $E[(\boldsymbol{x} - \mu)(\boldsymbol{x} - \mu)^T]$ は $V[\boldsymbol{x}]$ なので、

$$V[A\boldsymbol{x}] = A\,V[\boldsymbol{x}]\,A^T$$

ここで

$$V[\boldsymbol{x}] = \begin{pmatrix} V[x_1] & Cov[x_2, x_1] & \cdots & Cov[x_n, x_1] \\ Cov[x_1, x_2] & V[x_2] & \cdots & Cov[x_n, x_2] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[x_1, x_n] & Cov[x_2, x_n] & \cdots & V[x_n] \end{pmatrix}$$

[*31] 式の展開途中で、転置行列の演算 $(AB)^T = B^T A^T$ を使っている

■共分散行列から任意の方向のばらつきを調べる

分散行列は変数のばらつきに関する完全な情報を持っている。例えば以下のような 2 変数の確率変数ベクトル $\boldsymbol{x}$ の分散行列の対角成分は、「座標軸方向のばらつき」を意味している。

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad , \quad V[\boldsymbol{x}] = \begin{pmatrix} V[x_1] & Cov[x_2, x_1] \\ Cov[x_1, x_2] & V[x_2] \end{pmatrix}$$

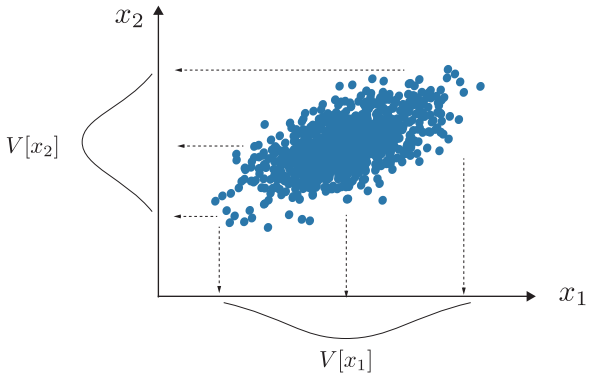


図 51　座標軸方向のばらつき

実は共分散行列があれば、座標軸の方向だけでなく、あらゆる方向でのばらつきを計算する事ができる。例えば図 52 のようなベクトル $\boldsymbol{u}$ 方向のばらつきは以下となる。

$$V[\boldsymbol{u}^t \boldsymbol{x}] = \boldsymbol{u}^t V[\boldsymbol{x}] \boldsymbol{u} \tag{9.30}$$

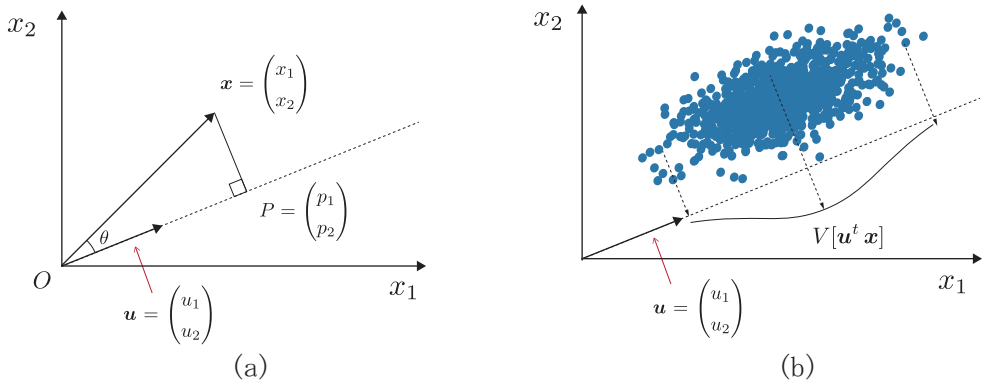


図 52　特定方向でのばらつき

式 (9.30) を求めていこう。まずベクトル $\boldsymbol{u}$ を $\boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$ とし、さらに基準化されているとする。つまり $|\boldsymbol{u}| = 1$ であるとする。ここで図 (52) の (a) のように任意のベクトル $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ をベクトル $\boldsymbol{u}$ に下ろした垂

線の点を $P = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$ としたときの $|OP|$ を内積を用いながら求めよう。

まずは、内積の定義式 (付録 D.1) と $|\boldsymbol{u}| = 1$ より以下が成立。

$$\boldsymbol{u}^T\boldsymbol{x} = |\boldsymbol{u}|\,|\boldsymbol{x}|\cos\theta = |\boldsymbol{x}|\cos\theta$$

この $|\boldsymbol{x}|\cos\theta$ は長さ OP に他ならないので下式で求める事ができる。

$$|OP| = \boldsymbol{u}^T\boldsymbol{x} = \begin{pmatrix} u_1 & u_2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1u_1 + x_2u_2$$

この値を z とする[*32]。つまり

$$z = \boldsymbol{u}^T\boldsymbol{x}$$

この時、得られた値はスカラーであるが、元々の $\boldsymbol{x}$ は確率変数ベクトルなので、得られた z も新たな確率変数である。この新しい確率変数 z のばらつき $V[z]$ を求めてみよう。

- $E[z]$ を求める

$$E[z] = E[\boldsymbol{u}^T\boldsymbol{x}] = \boldsymbol{u}^T E[X] = \boldsymbol{u}^T\mu \qquad (E[X] = \mu \quad \text{とおく})$$

- $V[z]$ を求める

$$\begin{aligned}
V[z] &= E[(z - E[z])(z - E[z])^T] \\
&= E[(\boldsymbol{u}^T\boldsymbol{x} - \boldsymbol{u}^T\mu)(\boldsymbol{u}^T\boldsymbol{x} - \boldsymbol{u}^T\mu)^T] = E[(\boldsymbol{u}^T\boldsymbol{x} - \boldsymbol{u}^T\mu)(\boldsymbol{x}^T\boldsymbol{u} - \mu^T\boldsymbol{u})] \\
&= E[\boldsymbol{u}^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{u} - \boldsymbol{u}^T\boldsymbol{x}\mu^T\boldsymbol{u} - \boldsymbol{u}^T\mu\boldsymbol{x}^T\boldsymbol{u} - \boldsymbol{u}^T\mu\mu^T\boldsymbol{u}] \\
&= E[\boldsymbol{u}^T(\boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{x}\mu^T - \mu\boldsymbol{x}^T - \mu\mu^T)\boldsymbol{u}] \\
&= \boldsymbol{u}^T E[(\boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{x}\mu^T - \mu\boldsymbol{x}^T - \mu\mu^T)]\boldsymbol{u} \qquad \because \boldsymbol{u} \text{ は定数} \\
&= \boldsymbol{u}^T E[((\boldsymbol{x} - \mu)(\boldsymbol{x} - \mu)^T)]\boldsymbol{u} = \boldsymbol{u}^T V[X]\boldsymbol{u}
\end{aligned}$$

[*32] 点 P は長さ 1 のベクトル $\boldsymbol{u}$ の延長線上なので $\vec{OP}$ を求めるなら以下のようにベクトル $\boldsymbol{u}$ を $|OP|$ 倍しなければならない。

$$\vec{OP} = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = |OP|\boldsymbol{u} = (x_1u_1 + x_2u_2)\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

10　多次元正規分布

多次元正規分布は、正規分布の多次元版であり、多変量正規分布ともいわれる。

10.1　多次元標準正規分布

最初に、平均 0 で分散 1 の標準正規分布に従う互いに独立な確率変数ベクトルについて考えてみる。ちなみに「互いに独立な確率変数」であれば、以下のように同時密度関数が周辺密度の積に分解できる。

定義 10.1.　【確率変数ベクトルの独立】
n 個の確率変数 $z_1, z_2, \cdots, z_n$ が独立であるとは、任意の実数 $a_1, a_2, \cdots, a_n$ に対して以下の式が成立する事である。

$$P(z_1 < a_1, z_2 < a_2, \cdots, z_n < a_n) = P(z_1 < a_1)P(z_2 < a_2)\cdots P(z_n < a_n)$$

確率密度関数でいえば、$\boldsymbol{z}$ を確率変数ベクトルとした時、同時密度関数 $f_{\boldsymbol{Z}}(\boldsymbol{z})$ が、周辺密度関数の積 $g(z_1)g(z_2)\cdots g(z_n)$ に分解されるとき独立であるという。

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = g(z_1)g(z_2)\cdots g(z_n) \tag{10.1}$$

この多次元標準正規分布の確率密度関数は以下のようになる。

定義 10.2.　【多次元標準正規分布の確率密度関数】
$\boldsymbol{Z}$ を n 次元標準正規分布 $N(\boldsymbol{o}, \boldsymbol{I})$ に従う確率変数とするとき、以下のように確率変数ベクトルを $\boldsymbol{z}$、ゼロベクトルを $\boldsymbol{o}$、そして単位行列を I としたとき、

$$\boldsymbol{z} = \begin{pmatrix} z_1 \\ z_1 \\ \vdots \\ z_n \end{pmatrix}, \qquad \boldsymbol{o} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \boldsymbol{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

確率密度関数は以下で表す事ができる。

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{1}{\sqrt{2\pi}^n}\ e^{-\frac{1}{2}||\boldsymbol{z}||^2} \tag{10.2}$$

この $||\boldsymbol{z}||^2$ をベクトルの内積で表示すると

$$||\boldsymbol{z}||^2 = \left(\sqrt{z_1^2 + z_2^2 + \cdots + z_n^2}\right)^2 = \boldsymbol{z}^T\boldsymbol{z}$$

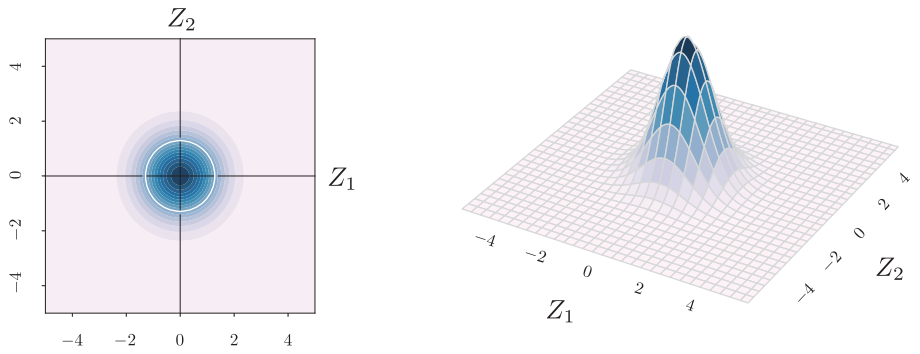


図 53　2 次元標準正規分布の確率密度関数

図 53 は二次元の標準正規分布の確率密度関数を表す。なお図 53 の左図の白丸は分散を表す。またこの図を描くプログラムをリスト 12 に示す。

この多次元標準正規分布 $N(\boldsymbol{o}, \boldsymbol{I})$ の期待値ベクトルと共分散行列は以下のようになる。これは各変数 $z_1, z_2, \cdots, z_n$ が平均 0 でばらつき 1 の標準正規分布に従う事から当然。また各変数は独立なので $Cov[Z_i, Z_j] = 0\ (i \neq j)$ となる。

$$E[\boldsymbol{Z}] = \begin{pmatrix} E[Z_1] \\ E[Z_2] \\ \vdots \\ E[Z_n] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \boldsymbol{o}$$

$$V[\boldsymbol{Z}] = \begin{pmatrix} V[Z_1] & Cov[Z_1, Z_2] & \cdots & Cov[Z_1, Z_n] \\ Cov[Z_2, Z_1] & V[Z_2] & \cdots & Cov[Z_2, Z_n] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[Z_n, Z_1] & Cov[Z_n, Z_2] & \cdots & V[Z_n] \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 1 \end{pmatrix} = \boldsymbol{I}$$

ソースコード 12　2 次元標準正規分布の確率密度関数を描くプログラム

```
import numpy as np
import matplotlib.pyplot as plt

# データ列を作成する
a = 6;    sls = 0.1
x = np.arange(−a, a , sls)
y = np.arange(−a, a , sls)
X, Y = np.meshgrid(x, y)
Z = X.astype(np.float64)

# 一次変換した標準正規分布を計算する
mx = 2;    my = 1            # 平衡移動 x-mx y-my
Wt = np.deg2rad(0)           # 回転行列
Dr = np.array([[np.cos(Wt), −np.sin(Wt)], [np.sin(Wt), np.cos(Wt)]])
Dt = np.array([[1.2, 0], [0, 1.2]])          # 拡大縮小行列
D = np.linalg.inv(Dr @ Dt)
for i in range(Y[0].size):
    for j in range(X[0].size):
        Wk = D @ np.array([X[i, j] − mx, Y[i, j] − my])
        Z[i, j] = np.exp(−(Wk[0] ** 2 + Wk[1] ** 2) / 2) / (2 * np.pi)

# グラフ表示
fig = plt.figure()
ax = fig.add_subplot(122, projection='3d')
bx = fig.add_subplot(121)
x_min = −a ; x_max = a
y_min = −a ; y_max = a
ax.set_xlim(x_min, x_max) ; ax.set_ylim(y_min, y_max) ;    ax.set_zlim(0, 0.2)
bx.set_xlim(x_min, x_max) ; bx.set_ylim(y_min, y_max)
# 右の三次元サーフェス
ax.grid(False) ; ax.set_zticks([])
ax.plot_surface(X, Y, Z, cmap="PuBu" , rstride=4, cstride=4, edgecolor="lightgray" ,linewidth=0.5)
ax.set_xlabel("X", fontsize=9) ; ax.set_ylabel("Y", fontsize=9)
# 左の濃度マップ
bx.contourf(X, Y, Z,levels=15, cmap='PuBu' )
bx.plot([x_min,x_max],[0,0],color="black", linewidth=0.6)
bx.plot([0,0], [y_min,y_max],color="black", linewidth=0.6)
bx.set_xlabel("X", fontsize=20) ; bx.set_ylabel("Y", fontsize=20)
# 信頼楕円の描画
theta = np.linspace(0, 2*np.pi, 20)
EX = np.cos(theta) ;    EY = np.sin(theta)
EX, EY = Dr @ Dt @ np.array([EX, EY])
bx.plot(EX + mx, EY + my, color="w")

plt.show()
```

10.2　多次元の標準正規分布を一次変換して様々な正規分布をつくる

一般の多次元の正規分布は、標準正規分布を一次変換して、スケールを変えたり、原点を変えたり、回転させたりして得る事ができる。その様子を見ていこう。

■スケーリングとシフト

先の多次元標準正規分布に従う n 個の確率変数ベクトル $\boldsymbol{Z}$ に対して以下のような一次変換をする事を考えてみる。ここで、σ は正のスカラー定数、μ は n 次元の定数の縦ベクトルとする。

$$\boldsymbol{X} = \sigma\boldsymbol{Z} + \boldsymbol{\mu}$$

この期待値と分散行列は

$$E[\boldsymbol{X}] = E[\sigma\boldsymbol{Z} + \boldsymbol{\mu}] = \sigma E[\boldsymbol{Z}] + \boldsymbol{\mu} = \sigma \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \boldsymbol{\mu}$$

$$\begin{aligned} V[\boldsymbol{X}] &= E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T] \\ &= E[(\sigma\boldsymbol{Z} + \boldsymbol{\mu} - \boldsymbol{\mu})(\sigma\boldsymbol{Z} + \boldsymbol{\mu} - \boldsymbol{\mu})^T] = \sigma^2 E[\boldsymbol{Z}\boldsymbol{Z}^T] = \sigma^2 V[\boldsymbol{X}] = \begin{pmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \ddots & \\ & & & \sigma^2 \end{pmatrix} = \sigma^2\boldsymbol{I} \end{aligned}$$

この一次変換によって平均 $\boldsymbol{\mu}$、分散 $\sigma^2\boldsymbol{I}$ の正規分布になる。この多次元正規分布を $N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ と表す。図 54 は具体的に標準正規分布 $\boldsymbol{Z}$ に対して以下の式で一次変換した $N(\boldsymbol{\mu}, 1.1^2\boldsymbol{I})$ の 2 次元正規分布のグラフである。

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = 1.1^2 \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

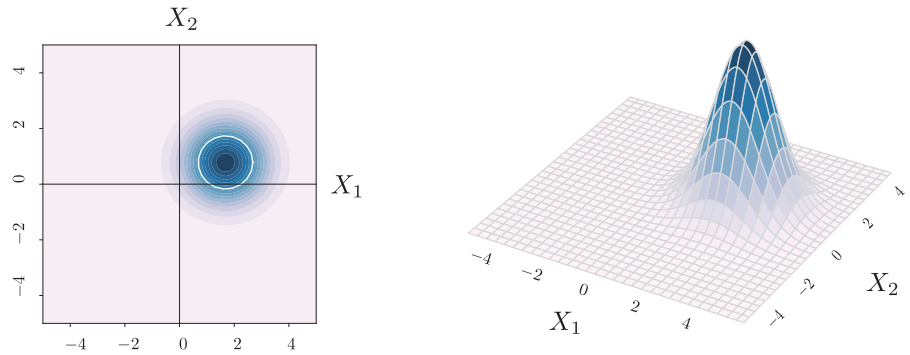


図 54　多次元正規分布 $N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$

■縦横の伸縮

スケーリングでは全方向に均等に σ 倍したが、軸によって伸縮の倍率を変えると図 55 のように楕円状の分布になる。これは n 個の多次元標準正規分布に従う確率変数ベクトル $\boldsymbol{Z}$ に対しての以下の式のように各成分を別々の定数で伸縮した事になる。

$$\boldsymbol{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix}, \qquad X = \begin{pmatrix} \sigma_1 Z_1 \\ \sigma_2 Z_2 \\ \vdots \\ \sigma_n Z_n \end{pmatrix}$$

これを行列で表すと以下のようになる。

$$\boldsymbol{X} = D\boldsymbol{Z}\,, \qquad D = \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{pmatrix}$$

この期待値と分散行列は

$$E[\boldsymbol{X}] = E[D\boldsymbol{Z}] = DE[\boldsymbol{Z}] = D\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \boldsymbol{o}$$

$$V[\boldsymbol{X}] = E[(D\boldsymbol{Z})(D\boldsymbol{Z})^T] = DE[\boldsymbol{Z}\boldsymbol{Z}^T]D^T = DV[\boldsymbol{Z}]D^T = D\boldsymbol{I}D^T = D^2$$

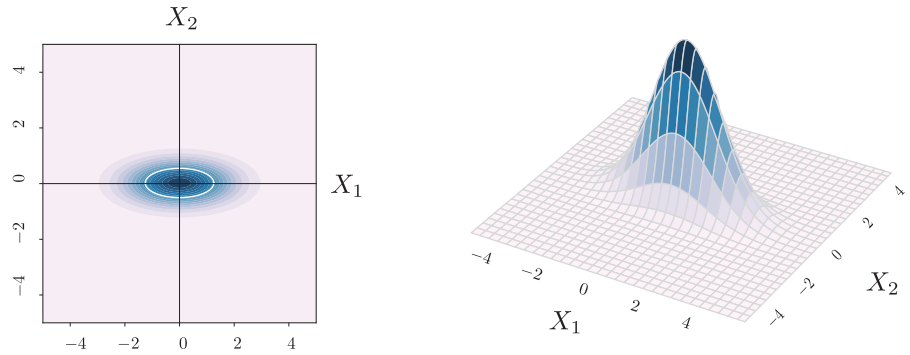


図 55　多次元正規分布 $N(\boldsymbol{o}, D^2)$

■回転

図 55 の多次元正規分布をさらに時計周りに 45° 回転させたものを考えると図 56 のようになる。一般的に、原点を中心とする期待値が $\boldsymbol{o}$ の多次元正規分布 $N(\boldsymbol{o}, \boldsymbol{V})$ はこのような形をしている。

回転という操作を行列で表すと、直交行列をかけるという操作になる。直交行列とは以下の式を満たす n 次正方行列 Q の事 (202 ページの節 D.4 を参照) であり、その列ベクトルはお互いに正規直交基底で出来て

いる。。

$$Q^tQ = I$$

この直交行列による写像は、２つのベクトルのなす角度と長さの両方を変えない写像であるという特徴がある。つまり図形を合同な図形に移す写像であり、そういった写像は幾何学的にいうと回転または鏡映（裏表を変える）となる。

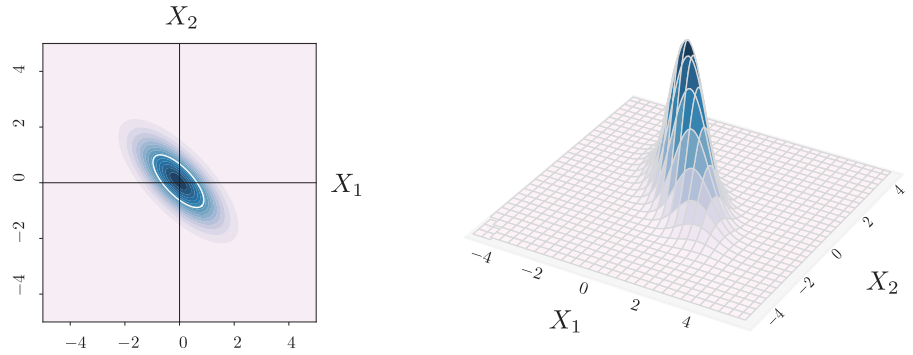


図 56　多次元正規分布 $N(\boldsymbol{o}, \boldsymbol{V})$

この一般的な原点を中心とする期待値が $\boldsymbol{o}$ の多次元正規分布 $N(\boldsymbol{0}, \boldsymbol{V})$ をつくり出す過程を振り返ってみる。

1. $N(\boldsymbol{o}, I)$ の確率変数ベクトルをつくる
   多次元標準正規分布 $N(\boldsymbol{o}, I)$ に従う確率変数ベクトル $\boldsymbol{Z}$ を持ってくる

$$\boldsymbol{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix}$$

   この期待値と分散行列は

$$E[\boldsymbol{Z}] = \boldsymbol{o}\,, \quad V[\boldsymbol{Z}] = I$$

2. 各変数毎に拡大縮小して $N(\boldsymbol{o}, D^2)$ の確率変数ベクトルにする
   その $\boldsymbol{Z}$ に下式のように対角行列 D をかける事で、各成分毎に拡大縮小率を変えた新しい確率変数ベクトル $\boldsymbol{X}$ をつくる

$$\boldsymbol{X} = D\boldsymbol{Z} = \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} = \begin{pmatrix} \sigma_1 Z_1 \\ \sigma_2 Z_2 \\ \vdots \\ \sigma_n Z_n \end{pmatrix}$$

   この期待値と分散行列は

$$E[\boldsymbol{X}] = \boldsymbol{o}\,, \quad V[\boldsymbol{X}] = D^2$$

3. 全体を回転して $N(\boldsymbol{o}, \boldsymbol{V})$ の確率変数ベクトルにする
   その $\boldsymbol{X}$ に直交行列 Q をかけて $\boldsymbol{Y} = Q\boldsymbol{X}$ をつくる。この期待値と分散行列を求めよう。

期待値 $E[\boldsymbol{Y}]$ は、$E[\boldsymbol{X}] = \boldsymbol{o}$ より

$$E[\boldsymbol{Y}] = E[Q\boldsymbol{X}] = QE[\boldsymbol{X}] = \boldsymbol{o}$$

共分散行列 $V[\boldsymbol{Y}]$ は、$V[\boldsymbol{X}] = D^2$ より

$$V[\boldsymbol{Y}] = E[(Q\boldsymbol{X})(Q\boldsymbol{X})^T] = E[Q(XX^T)Q^T] = QV[\boldsymbol{X}]Q^T = QD^2Q^t$$

このように原点を中心とした多次元の標準正規分布に従う確率ベクトル $\boldsymbol{Z}$ を変数毎に伸縮し、さらに回転する事で、一般の多次元正規分布をつくりだす事ができ、その共分散行列は以下のように表す事ができる。

$$\boldsymbol{V} = Q\,D^2\,Q^T \tag{10.3}$$

逆に、今 $\boldsymbol{V}$ という共分散行列が与えられたとする。分散行列は $Cov[Z_1, Z_2] = Cov[Z_2, Z_1]$ なのでお互いの対角要素が同じ値の対称行列[*33]になる。この対称行列である $\boldsymbol{V}$ という共分散行列を対角行列 D と直交行列 Q によって上記のように分解する事ができれば、分布状況の見通しがわかりやすくなる。

いま上記の式 (10.3) を満たすような直交行列 Q が見つかったと過程すると $Q^TQ = QQ^T = I$ なので、式 (10.3) の両辺に左から Q^T をかけて右から Q をかけると以下のような対角行列が抽出できる。

$$Q^T\boldsymbol{V}Q = Q^T(Q\,D^2\,Q^T)Q = D^2$$

この D^2 の要素は各確率変数の分散になる。

$$D^2 = \begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_n^2 \end{pmatrix} = \begin{pmatrix} V[Z_1] & & & \\ & V[Z_2] & & \\ & & \ddots & \\ & & & V[Z_n] \end{pmatrix}$$

このように対称行列 V を直交行列 Q を用いて対角行列 D に変換する事を対角化と呼ぶ。分散行列だけをみていると n 個の確率変数が複雑に絡み合っているように見える場合でも、対角化することによって複雑に見えた関係が実は独立な n 個の変数が合成された結果であると捉え直す事が可能になる。

[*33] 対称行列とは、任意の行列 A の行と列を入れ替えたとき、元の行列 A と等しくなるもの

10.3　分散行列の対角化

分散行列のように要素が実数の対称行列は必ず対角化できる。いったんこれを前提にする。いま分散行列 V に対して、ある適切な直交行列 Q を持ってきて、Q^TVQ が対角行列になるようにできたとし、できた対角行列を以下のように Λ と表すものとする。

$$Q^TVQ = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} = \Lambda$$

この両辺に Q をかける。Q は直交行列で $QQ^T = I$ なので、左辺は $QQ^TVQ = VQ$ となり左辺は $Q\Lambda$。つまり以下のように表す事ができる。

$$VQ = Q\Lambda$$

これを成分表示すると

$$\begin{pmatrix} - & \boldsymbol{v}_1 & - \\ & \vdots & \\ - & \boldsymbol{v}_n & - \end{pmatrix} \begin{pmatrix} | & & | \\ \boldsymbol{q}_1 & \cdots & \boldsymbol{q}_n \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ \boldsymbol{q}_1 & \cdots & \boldsymbol{q}_n \\ | & & | \end{pmatrix} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

この両辺の 1 列目を取り出すと

$$\begin{pmatrix} - & \boldsymbol{v}_1 & - \\ & \vdots & \\ - & \boldsymbol{v}_n & - \end{pmatrix} \begin{pmatrix} | \\ \boldsymbol{q}_1 \\ | \end{pmatrix} = \lambda_1 \begin{pmatrix} | \\ \boldsymbol{q}_1 \\ | \end{pmatrix}$$

列ごとにみると以下のような構造。

$$\begin{pmatrix} - & \boldsymbol{v}_1 & - \\ & \vdots & \\ - & \boldsymbol{v}_n & - \end{pmatrix} \begin{pmatrix} | \\ \boldsymbol{q}_1 \\ | \end{pmatrix} = \lambda_1 \begin{pmatrix} | \\ \boldsymbol{q}_1 \\ | \end{pmatrix} \quad , \cdots , \quad \begin{pmatrix} - & \boldsymbol{v}_1 & - \\ & \vdots & \\ - & \boldsymbol{v}_n & - \end{pmatrix} \begin{pmatrix} | \\ \boldsymbol{q}_n \\ | \end{pmatrix} = \lambda_n \begin{pmatrix} | \\ \boldsymbol{q}_n \\ | \end{pmatrix}$$

つまり、求めたい直交行列 Q の各列は以下のような行列 V の固有ベクトルを求める式で構成されている[*34]。

$$V\boldsymbol{q}_n = \lambda\boldsymbol{q}_n$$

以上の事から、直交行列 Q は分散行列 V の固有ベクトル $\boldsymbol{q}_n$ を求めて、その固有ベクトルを列ベクトルとして並べる事によってつくる事ができる事がわかる。具体的な手順は以下。

1. 与えられた分散行列 V の固有値 $\lambda_1, \cdots, \lambda_n$ を求める
2. 各固有値 λ_n に対応する固有ベクトル $\boldsymbol{p}_n$ を求める
3. 各固有ベクトルの長さを 1 にそろえる。つまり $\boldsymbol{q}_n = \boldsymbol{p}_n / ||\boldsymbol{p}_n||$

[*34] 固有値及び固有ベクトルについては、209 ページの固有値と固有ベクトルの定義を参照

4. 固有ベクトルを以下のように列方向に並べて Q をつくる

$$Q = \begin{pmatrix} | & & | \\ \boldsymbol{q}_1 & \cdots & \boldsymbol{q}_n \\ | & & | \end{pmatrix}$$

5. 対応する固有値を対角成分に並べた行列 Λ をつくる

$$\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

この直交行列 Q と対角行列 Λ を用いれば分散行列は $VQ = Q\Lambda$ と表せ、この両辺に左から Q^T をかける事で

$$Q^t V Q = \Lambda$$

というように対角化できる。

11　MCMC の原理

MCMC 法（エムシーエムシー法）は、マルコフ連鎖モンテカルロ法（Markov Chain Monte Carlo）の略で、ベイズ統計などで使われる「複雑な確率分布からのサンプリング方法」である。ベイズの事後分布 $P(\theta \mid x)$ を知りたいけど、分布が複雑すぎて、直接は描けないし、積分もできない場合に、MCMC 法でたくさんの「サンプル（標本）」を作って、その分布を近似する方法で、以下のように 2 つのアイデアを組み合わせた方法である。

モンテカルロ法（Monte Carlo）　ランダムにサンプルを作って、分布を推定する方法
マルコフ連鎖（Markov Chain）　現在の状態だけに依存して、次の状態が決まるルールで動く

11.1　モンテカルロ法

特定の個人が考案して命名されたわけではなく、ランダム性や確率を利用して問題を解決する一連の計算アルゴリズムを指す総称である。「モンテカルロ」という名前は、カジノで有名なモナコの都市モンテカルロに由来しており、この手法が乱数（サイコロを振るようなランダム性）を用いることから、第二次世界大戦中にロスアラモス国立研究所の科学者たちによって名付けられた。

π の値を求める事を考えよう。図 57 のように乱数を発生させて、円の中に入る $x^2 + y^2 \leq 1$ 確率を求める。発生させる乱数の数を多くすれば、赤の円の中に入る確率は $\frac{1}{4}\pi$ に近づくはずである。

具体的には以下のような手順で計算をする。

1. $0 \sim 1$ の一様乱数を 2 個発生させそれを x, y 座標とし
2. $[0,1] \times [0,1]$ の正方形の中にランダムに点を打つ
3. その点が $x^2 + y^2 \leq 1$ なら、その点は円の中にあるとする
4. 全体の点の数に対して、円の中に入った点の割合を数える
5. 無数に点を発生させれば、円に入る割合は $\fallingdotseq \frac{\pi}{4}$ に漸近する

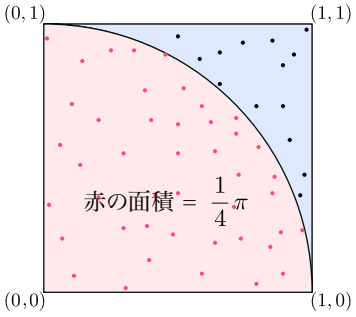


図 57　モンテカルロ法で円周率を求める

■モンテカルロ法の実装

ソースコード 13　モンテカルロ法の実装

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from ipywidgets import interact

plt.style.use("ggplot")
np.random.seed(123)

NMC = 10000
xmc = np.random.rand(NMC)
ymc = np.random.rand(NMC)
#振る舞いを与えるデコレータ 0からNMC までスライダーを動かせるようにする
@interact(mcs=(0,NMC,1))
def animation(mcs=0):
    plt.figure(figsize=(6,6))
    plt.xlim([0,1])
    plt.ylim([0,1])
    #円を描く
    x = np.arange(0,1,0.001)
    y = (1 - x ** 2) ** 0.5
    y2 = np.ones(x.shape[0])    #円の塗りつぶしに使うすべてが 1 の配列
    plt.plot(x,y)
    plt.fill_between(x, y, alpha=0.3)
    plt.fill_between(x, y, y2,alpha=0.3)
    r = (xmc[:mcs] ** 2 + ymc[:mcs] ** 2) ** 0.5
    accept = np.where(r<=1, 1, 0)
    accept_ratio = np.sum(accept) / mcs
    plt.scatter(xmc[:mcs], ymc[:mcs], color="black", marker=".")
    plt.show()
    print("Monte␣Carlo:␣",accept_ratio)
    print("Analytical␣Solution:␣", np.pi / 4.0)
```

■モンテカルロ法の適用場面

このように、モンテカルロ法とは「乱数を使って近似計算をする手法」の総称である。「分布の形がまったくわからない」場合には、モンテカルロ法は使えないが、「確率密度の形はわかるけど、サンプリングが難しい」という状況ではモンテカルロ法が大活躍する。以下にどんな場合に使われているかを示す。

目的	名前	必要なもの	成功条件
分布から平均や確率を計算	単純モンテカルロ	乱数生成	分布からサンプルが取 れること
複雑な関数の面積を計算	面積モンテカルロ	境界がわかる	内外の判定ができるこ と
複雑な分布からサン プルしたい	棄却サンプリング	比例関数 $\propto p(x)$ がわかる	上からおおう関数 $q(x)$ がある事
一般的な事後分布サンプリング	MCMC	$p(x)$ の比がわかる	正規化定数が不要でも成り立つ

特にモンテカルロ法はベイズ推論でよく使われる。ベイズ推定の事後分布では、以下のような形で事後分布

を考える。

$$p(\theta \mid D) \propto p(D \mid \theta) \cdot p(\theta)$$

このとき、右辺の積は計算できても、左辺を正規化する「定数 Z 」はわからないことが多い。 こういった場合に、モンテカルロ法（特に MCMC）ではこの比だけわかれば十分なので、多くの現実的な応用が可能である。

11.2　棄却サンプリング

事後分布がベータ分布に従う時にモンテカルロ法を使って、ベータ分布に従ったデータをサンプリングしたい。この再に棄却サンプリングという方法を用いる。

まずベータ分布は以下で定義される。

$$\mathrm{Beta}(x \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

このとき、$B(\alpha, \beta)$ はベータ関数で、以下の積分で定義されている。

$$B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta$$

ベータ分布の関数を数値解析的に求める事は難しくないが、複数のデータをサンプリングして得るのは容易ではない[*35]。ここでは「ベータ分布に従う複数のデータをサンプリングする」という事を棄却サンプリングという手法で行ってみる。

■具体的な手順

目標分布である Beta 分布 $f(x)$ に従う乱数の発生は諦める。その代わり、サンプリングが簡単な一様分布 $g(x)$ を使う。これを提案分布という（ただし $f(x) \leq Mg(x)$ となるように定数 M を調整している）。この時に、乱数を発生させて提案分布内部の座標が得て、それが目標分布内にあれば採用し、なければ棄却するという考え方でサンプル集団を作り出す。

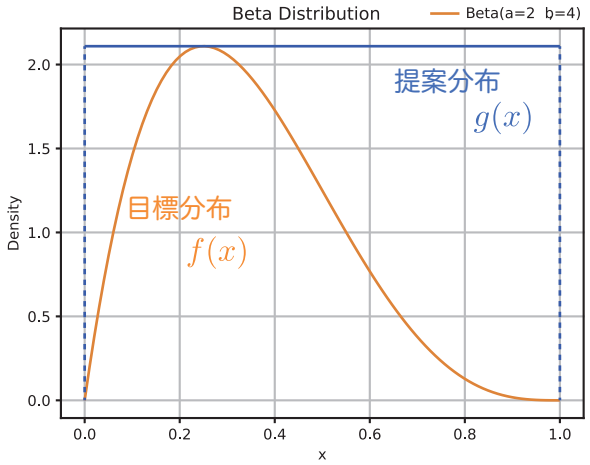


図 58　棄却サンプリング

具体的には以下のような手順になる。

(1) 0〜1 の乱数を発生させ候補の x 座標をサンプリング
(2) 0〜$Mg(x)$ までの乱数 r を引き、次式を判定 $r \leq f(x)$

[*35] Python にはベータ分布に従う複数のデータをサンプリングする関数 scipy.stats.beta.rvs がある

(3) 真ならばその x のカウントを 1 増やす。偽ならばそのサンプルは棄却する
(4) (2) と (3) を N 回繰り返す
(5)　各 x でカウント数のヒストグラムを作成

■棄却サンプリングの実装

Python のプログラムをコード 14 に示した。この中で、`f = beta(a=a, b=b).pdf` という命令が出てくるが、これはベータ分布をオブジェクトとして使うという宣言であって、以下のような関数表現と等価である。

```
def f(x):
        return beta.pdf(x, a=1.5, b=2.0)
```

ただし、オブジェクトとして扱うと、以下のように分布の様々な性質をもったオブジェクトとして扱える。

```
dist = beta(a=1.5, b=2.0)
y = dist.pdf(x)
z = dist.rvs(1000)
```

それによって、多様な操作が、パラメータを再指定せず一貫して扱えるようになる。使えるのは以下のような分布の特性である。

- pdf()（確率密度関数）
- cdf()（累積分布関数）
- rvs()（乱数サンプリング）
- mean(), var() など

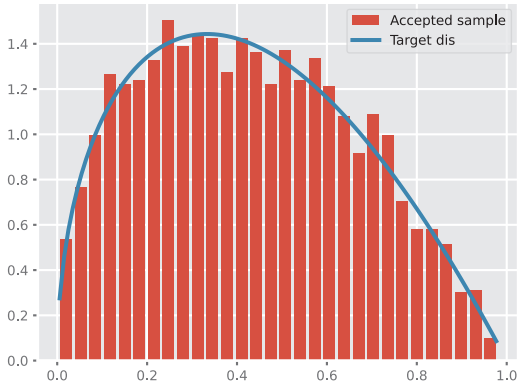


図 59　棄却サンプリングの実装プログラムの出力

ソースコード 14　棄却サンプリングの実装

```
import numpy as np
from scipy import stats
from scipy import optimize as opt
from scipy.stats import beta, uniform
import matplotlib.pyplot as plt
%matplotlib inline

plt.style.use("ggplot")
np.random.seed(123)

# ベータ分布オブジェクトのパラメータ設定
f = beta(a=a, b=b).pdf
res = opt.fmin(lambda x: −f(x), 0.3)
y_max = f(res)

NMCS = 5000
x_mcs = uniform.rvs(size=NMCS)                #rvs は一様分布からの乱数を生成(X 軸 )
r = uniform.rvs(size=NMCS) * y_max            #y 軸の最大値までの乱数を生成
accept = x_mcs[r <= f(x_mcs)]                 #ベータ分布より小さかったらアクセプトする
plt.hist(accept, density=True, bins=30, rwidth=0.8, label="Accepted␣sample")
x = np.linspace(beta.ppf(0.001, a, b), beta.ppf(0.999, a, b), 100)
plt.plot(x, beta.pdf(x, a, b), label = "Target␣dis")
plt.legend()
plt.savefig('rejection_sampling.pdf␣')
plt.show()
```

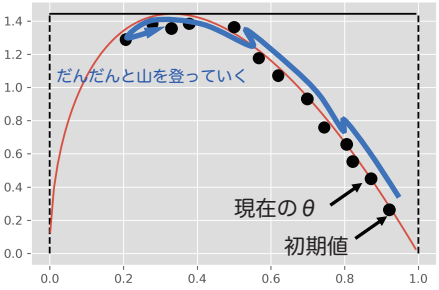
11.3　MCMC と定常分布

先ほどの棄却サンプリングは、必要な関数を取り囲む一様分布を設定した。いわば関数を取り囲む四角形を設定した事になる。その中に点をランダムにうって領域内か、領域外かを判断していく方法であった。1 次元ならそれほど難しくないが、次元（変数の数）が増えると、計算や解析が急激に難しくなる現象があり、それを「次元の呪い（Curse of Dimensionality）」と呼ぶ。

例えば、1 次元（線）で囲めば、長さ 10 の区間に、10 点おけばだいたい均等にカバーできるが、二次元（正方形）なら $10 \times 10 = 100$ 点 必要である。それでも 1 辺に 10 点置けば、全体をカバー可能だが、これが 10 次元（超立方体）となると、1 辺に 10 点置こうとすると必要な点の数は $10^{10} = 10,000,000,000$ 点にもなる。このように急激に必要な点が巨大化し、すべての点を調べることが現実的でないだけでなく、高次元空間ではデータがスカスカで、平均や分布が安定しない。そこで「前の試行結果を使って、次の乱数を発生させえるという」マルコフチェーンを組み合わせた MCMC が役に立つことになる。

具体的には右図のように、以下のような流れで計算をする。

1. 初期値 θ_0 を適当に決める
2. 乱数で今の θ から新しい θ_{new} を探してくる
3. 次の条件式を判定する　　$f(\theta_{\text{new}} \mid D) > f(\theta \mid D)$
4. 真であれば状態を更新。偽の場合はある程度の確率で受容
5. 2 − 4 を繰り返す

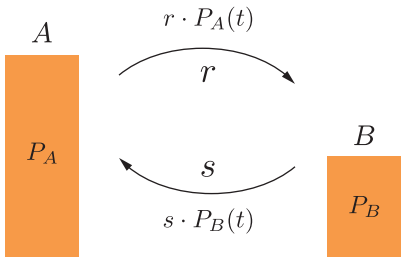


このような流れで計算を進めた時に、ある一定の定常状態に落ち着くことが必要であり、次にマルコフ過程が定常状態になるための条件について説明する。

■マルコフ過程の定常状態

いま右図のように、状態 A と B があり、A → B の遷移率が r で、B → A の遷移率が s であるとする。その時、ある時刻 t における A の確率 $P_A(t)$ は次のように変化する。

$$\frac{dP_A(t)}{dt} = s \cdot P_B(t) - r \cdot P_A(t)$$



ここで、状態 B にいる確率は補数なので以下となる。

$$P_B(t) = 1 - P_A(t)$$

$$\frac{dP_A(t)}{dt} = s \cdot P_B(t) - r \cdot P_A(t)$$

これを代入すると、ある時刻 t における A の確率 $P_A(t)$ は以下のように表す事ができる。

$$\frac{dP_A(t)}{dt} = s(1 - P_A(t)) - rP_A(t) = s - (r + s)P_A(t)$$

ここで、定常状態（steady state）とは？「時間が十分経って、確率がもう変化しなくなった状態」であり、

$$\frac{dP_A(t)}{dt} = 0$$

になるときである。このときの $P_A(\infty)$ を 定常確率という。いま定常状態になったとすると以下が成立する。

$$0 = s - (r+s)P_A$$

これを解くと以下のようになる。

$$P_A = \frac{s}{r+s}$$

$$P_B = 1 - P_A = \frac{r}{r+s}$$

このように定常状態とは、$\boxed{\text{B から A への流入確率}} = \boxed{\text{A から B への流出確率}}$ の状態であり、$s \cdot P_B = r \cdot P_A$ の状態である。この状態を**詳細釣り合い（detailed balance）**という。

■詳細釣合い（detailed balance）

事後分布を推定する例について説明していこう。図 61 のように事後分布があった時、 θ、θ' の起こりやすさは $f(\theta)$、$f(\theta')$ のように表す事ができる[*36]。また、θ から θ' への移りやすさを $f(\theta' \mid \theta)$ 、θ' から θ への移りやすさを $f(\theta \mid \theta')$ とすれば、それぞれの流入量と流出量が釣り合っている状態は以下のように表す事ができる。

$$f(\theta'|\theta)\, f(\theta) = f(\theta|\theta')\, f(\theta') \tag{11.1}$$

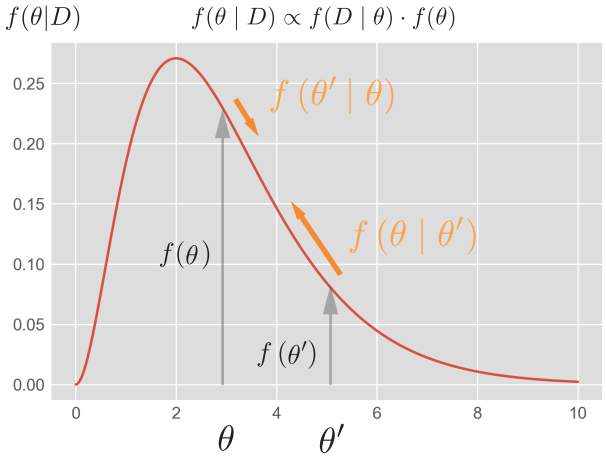


図 60　詳細つり合いのイメージ

詳細つり合いとは、このように任意の 2 つの θ、θ' での確率の流れが等しいことを意味しており、定義範囲内でのどの 2 点をとっても詳細まで釣り合っているので**詳細つり合い**と呼ばれる。その場合には、確率分布は定常分布になる。

[*36] 本来は $f(\theta \mid D)$、$f(\theta' \mid D)$ と表記すべきだが、観測値である D は同じ値なので省略し、それぞれ $f(\theta)$、$f(\theta')$ と表記した。

これを微分方程式で書くと以下のように表すことができる。この式を **Master 方程式**と呼ぶ。

$$\frac{\mathrm{d}f(\theta)}{\mathrm{d}t} = \sum_{\theta'} \left(-f\left(\theta'|\theta\right) f(\theta) + f\left(\theta|\theta'\right) f\left(\theta'\right)\right) \tag{11.2}$$

この右辺の第一項 $-f\left(\theta'|\theta\right) f(\theta)$ は θ のポイントから θ' への流出を示し、第二項 $f\left(\theta|\theta'\right) f\left(\theta'\right)$ は θ' から θ への流入を示す。この流出と流入のつり合いが様々な θ' で成立しているので θ' での和 $\sum_{\theta'}$ を取っている。

式 (11.1) を式 (11.2) に代入すると、以下のように時刻 t によらず一定の定数となり $f(\theta)$ は定常分布になる。

$$\frac{\mathrm{d}f(\theta)}{\mathrm{d}t} = 0 \qquad \Leftrightarrow \qquad f(\theta) = \text{ const }.$$

式 (11.1) の $f(\theta'|\theta)$、$f(\theta|\theta')$ を**遷移核（transition kernel）**と呼ぶ。遷移核とは「ある状態から次の状態へ移る確率を定めるルール（関数）」のことである。この遷移核が $f\left(\theta'|\theta\right) f(\theta) = f\left(\theta|\theta'\right) f\left(\theta'\right)$ を満たすものを「詳細つり合い」といい、それを満たす遷移核は 1 つではなく、以下のようなものがある

1. Metropolis-Hastings（M-H アルゴリズム）
2. Gibbs サンプラー（熱浴法）
3. ハミルトニアンモンテカルロ

11.4　M-H アルゴリズム

メトロポリスヘイスティング法（Metropolis-Hastings algorithm）は、メトロポリス（Nicholas Metropolis）とヘイスティング（W.K. Hastings）という二人の研究者の名前に由来している。

■MH 法の考え方　MH 法の目的は、今までの MCMC の目的と同じで、次のようなベイズの定理に基づく事後分布からのサンプリングをすることである。

$$f(\theta|D) = \frac{p(D|\theta) \cdot p(\theta)}{p(D)}$$

サンプルがたくさんあれば、横軸 θ に対してその出現頻度が分布の高さ（密度）を反映するので、確率密度関数を直接描くことはできなくても、ヒストグラムなどで「形」を近似できる。つまり、サンプリングすることによって分布の形を“経験的に”知ることが出来る。しかし、この分布はふつうは正規化定数 $p(D)$ を計算するのがとても難しいので、直接サンプリングすることができない。そのため、正規化定数を必要としないように工夫された方法がいくつかある。たとえば、比率を用いて受容判定を行うメトロポリス・ヘイスティングス法や、条件付き分布だけでサンプリングするギブスサンプラー、さらに勾配情報を使って効率的に候補を提案するハミルトニアンモンテカルロ法などがある。

MH 法のポイントは比率を用いることで、$p(D)$ を使わないという点である。「今いる場所と、そこから“移動先”として提案された 1 点」を比率で比較して、「どっちの方が事後分布 $p(\theta \mid D)$ で起こりやすいか」を判断する仕組みである。

表 8　MH 法での各確率の扱い

項目	MH 法での扱い
$p(\theta)$	事前分布として 与えられる（前提）
$p(D\|\theta)$	尤度として 計算される
$p(D)$	使わない（$p(D)$ は比で打ち消されるから）
$f(\theta \mid D)$	比例形 $p(\theta\|D) \propto p(D\|\theta) \cdot p(\theta)$ だけで OK。定数倍は不要

MH 法で各確率をどのように捉えているかを表 8 に示した。まず $p(\theta)$ は前提として与えて、そこから尤度 $p(D|\theta)$ を計算する。しかし $p(D)$ は使わない。今の候補 θ と提案された新しい候補 θ' を、下式の第一項のような事後分布の高さ（= 尤度 × 事前分布）の比で比較する。つまり比にすることで絶対値を使わなくても、どちらがより事後分布で“起こりやすいかを判断することができる。

$$\frac{p(D|\theta') \cdot p(\theta')}{p(D|\theta) \cdot p(\theta)} \cdot \frac{q(\theta|\theta')}{q(\theta'|\theta)}$$

この「確率が高い方を選ぶ」という判断基準は MH 法の設計思想である。なぜなら目標は「事後分布と同等な分布になるサンプル列を得ること」だから、そのためにはより事後分布の密度が高い場所には頻繁に訪れる（受け入れる）ようにし、密度が低い場所にはあまり行かない or 行ってもすぐ戻るというようなロジックを組み立てて、サンプリングしたいからである。

また、さらにこの式の第二項の $q()$ は、確率が θ から θ' に変化した時の流入量と流出量との比になっている。 $q(\theta|\theta')$ と $q(\theta'|\theta)$ の比は、「方向性のゆがみ」を調整するための補正項で、この補正があることで、非対称な提案でも事後分布が定常分布になる

また、マルコフ連鎖が時間とともに特定の分布（定常分布）に収束するためには、以下の詳細つり合い条件を満たす必要がある。

$$f\left(\theta'|\theta\right)f(\theta) = f\left(\theta|\theta'\right)f\left(\theta'\right)$$

ところが、この $f(\theta'|\theta)$ はまだ判っていない。そこで、$f(\theta'|\theta)$ を導くために $q(\theta'|\theta)$ という設計しやすい分布を仮設定する。その上で、その分布を受容する確率 $\alpha(\theta,\theta')$ をかけて、以下のように全体の遷移確率を構成しなおおす。

$$f(\theta'|\theta) = q(\theta'|\theta)\cdot\alpha(\theta,\theta')$$

この仮においた分布 $q(\theta' \mid \theta)$ は設計者が設定する分布なので「提案分布」と呼ばれる。この提案分布には、正規分布や非対称なドリフトつき分布などが用いられる。

■MH 法のアルゴリズム　ついで、詳細つり合いを満たすように修正していくやりかたについて述べる。

提案分布を $q(\theta'|\theta)$ として計算した結果が以下のようになっており、詳細つり合いが満たされていない場合

$$q\left(\theta'|\theta\right)f(\theta) > q\left(\theta|\theta'\right)f\left(\theta'\right)$$

以下のように両辺が等号になるように補整係数 r を導入する。

$$rq\left(\theta'|\theta\right)f(\theta) = q\left(\theta|\theta'\right)f\left(\theta'\right)$$

これを r について解くと、以下のような r を導入することになる。

$$r = \frac{q\left(\theta|\theta'\right)f\left(\theta'\right)}{q\left(\theta'|\theta\right)f(\theta)}$$

(1) 初期値 θ を適当に決める

(2) 提案分布 $q\left(\theta'|\theta\right)$ から乱数を引いて、新しい θ' を探してくる

(3) 次の条件式を判定する

$$q\left(\theta'|\theta\right)f(\theta) > q\left(\theta|\theta'\right)f\left(\theta'\right)$$

(4) **真の場合** θ から θ' への流れが強い事を意味するので以下の r を使って補正する

$$r = \frac{q\left(\theta \mid \theta'\right)f\left(\theta'\right)}{q\left(\theta' \mid \theta\right)f(\theta)}$$

偽の場合 θ から θ' への流れが弱いことを意味する。低いところに行きやすくすると分布が崩れるので、θ' を受け入れる

(5) (2) − (4) を繰り返す

■ランダムウォーク HM 法

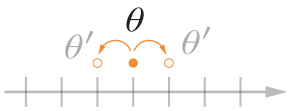
ランダムウォーク・メトロポリス・ヘイスティングス法（Random Walk Metropolis-Hastings, 略して RW-MH 法）は、M-H 法の中でも非常によく使われる方法である。これは、現在の位置 θ の周りに少しラン

ダムに動いて、新しい候補 θ' を探す方法である。提案分布（proposal distribution）$q(\theta'|\theta)$ は、対称な分布にする。ここでは以下のように正規分布 (平均 0、標準偏差 1) をとり、θ の変化幅を ϵ で指定する。

$$\theta' = \theta + \epsilon\,\text{Normal}(0,1)$$

ランダムウォークは左右対称な分布なので、$q(\theta'|\theta) = q(\theta|\theta')$ であり、以下のように補正係数 r は事後分布の比となる。

$$r = \frac{q\left(\theta \mid \theta'\right) f\left(\theta'\right)}{q\left(\theta' \mid \theta\right) f(\theta)} = \frac{f\left(\theta'\right)}{f(\theta)}$$



また、この受容確率 r をつかった受け入れるかどうかの判断は以下のようにする。

(1) 受容確率 r に基づいて、
新しい候補 θ' の密度が 現在の状態 θ より高い場合（$r \geq 1$ の場合）は受け入れる。
逆に低い場合（$r < 1$ の場合）は、次の乱数に従って受け入れるかどうかを決める

(2) 乱数 $u \sim \text{Uniform}(0,1)$ を使って、受け入れるか否かを「確率的に」決める
r が低いが、 乱数 u よりも大きい（か同じ）ときは、受け入れる
r が低く、 乱数 u よりも小さいときは、 拒否し状態は変えない

このように、２段階にしているのは、「らしいところに滞在しやすく、あまりらしくないところにも確率的に時々飛ぶ」という挙動を実現するためである。

■M-H アルゴリズムの python による実装

以下がベータ関数を事後確率として設定した場合の MH アルゴリズムで、アルゴリズム自体は非常に短い。

ソースコード 15　MH アルゴリズムの本体

```
#M-H アルゴリズムの本体
theta = 0.8 # 初期値
NMCS = 20000    # MCS の数
epsilon = 0.5   # 探索幅
theta_mcs = [theta] # MCS の結果を保存するリスト
for i in range(NMCS):  # MCS の数だけ繰り返す
    # 探索幅の範囲内にランダムに新しいtheta を生成する
    theta_new = theta + epsilon * np.random.randn()
    # 新しいtheta のベータ関数の密度が前の theta よりも高いかをチェック
    if beta.pdf(theta_new, a, b) > beta.pdf(theta, a, b):
        # 新しいtheta の密度が今の theta よりも高い場合、theta を受け入れる
        theta = theta_new
    # 新しいtheta が前の theta よりも低い場合、確率的な判断をする
    else:
        # 受容確率r を計算する
        r = beta.pdf(theta_new, a, b) / beta.pdf(theta, a, b)
        # 乱数を発生させ、r より小さいなら新しい theta を受け入れる
        if np.random.rand() < r:
            theta = theta_new
    # theta の値を保存
    theta_mcs.append(theta)
# 結果をデータフレームに変換
df = pd.DataFrame(theta_mcs)
```

結果は以下のように、横軸をサンプリングしていった結果、元のベータ関数に近いヒストグラムが描けている。

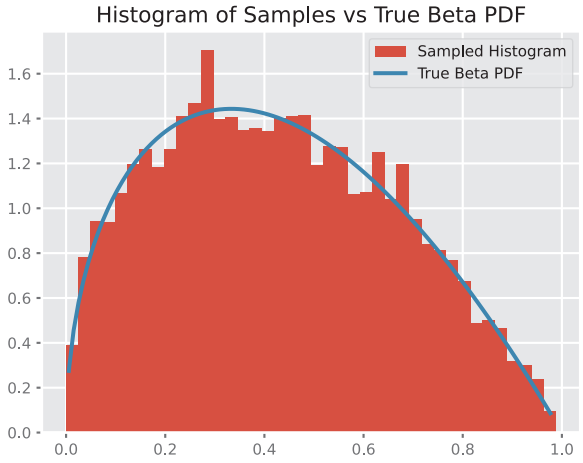


図 61　詳細つり合いのイメージ

以下はソースコード全体

ソースコード 16　n = 2 の場合のエントロピーを描くプログラム

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import beta
import pandas as pd

plt.style.use("ggplot")
np.random.seed(123)

a, b = 1.5, 2.0

#最初の図（ベータ分布を描く）
plt.figure()
x = np.linspace(beta.ppf(0.001,a,b), beta.ppf(0.999, a,b), 100)
plt.plot(x, beta.pdf(x, a, b))
plt.title("True␣Beta␣Distribution")

#M-H アルゴリズムの本体
theta = 0.8 # 初期値
NMCS = 20000    # MCS の数
epsilon = 0.5   # 探索幅
theta_mcs = [theta] # MCS の結果を保存するリスト
for i in range(NMCS):  # MCS の数だけ繰り返す
# 探索幅の範囲内にランダムに新しいtheta を生成する
theta_new = theta + epsilon * np.random.randn()
# 新しいtheta のベータ関数の密度が前の theta よりも高いかをチェック
if beta.pdf(theta_new, a, b) > beta.pdf(theta, a, b):
# 新しいtheta の密度が今の theta よりも高い場合、theta を受け入れる
theta = theta_new
# 新しいtheta が前の theta よりも低い場合、確率的な判断をする
else:
# 受容確率r を計算する
r = beta.pdf(theta_new, a, b) / beta.pdf(theta, a, b)
# 乱数を発生させ、r より小さいなら新しい theta を受け入れる
if np.random.rand() < r:
theta = theta_new
# theta の値を保存
theta_mcs.append(theta)
# 結果をデータフレームに変換
df = pd.DataFrame(theta_mcs)

# theta のトレースプロットとヒストグラムを描画する
plt.figure()
plt.plot(df[0])
plt.xlabel("MCS")
plt.ylabel("$\Theta$")
plt.title("Trace␣Plot␣of␣Theta")
```

```
# ヒストグラムとベータ分布の確率密度関数を比較する
plt.figure()
plt.hist(df[0][1000:], density=True, bins=40, label="Sampled Histogram")
x = np.linspace(beta.ppf(0.001,a,b), beta.ppf(0.999, a,b), 100)
plt.plot(x, beta.pdf(x, a, b), label="True Beta PDF")
plt.legend()
plt.title("Histogram of Samples vs True Beta PDF")

plt.show()
```

plt.figure() の役割

`plt.figure()` を呼ぶと、新しい描画キャンバス（Figure）が作られる。これを使わないと、すべてが同じキャンバスに上書きされてしまう。

ベータ分布を描くためのデータを作っている部分

以下がベータ分布用の x 軸を作っている部分。

```
x = np.linspace(beta.ppf(0.001,a,b), beta.ppf(0.999, a,b), 100)
```

・この `beta.ppf(q, a, b)` の `ppf` は Percent Point Function（＝累積分布関数（CDF）の逆関数）で、`beta.ppf(0.001, a, b)` は、「ベータ分布 Beta(a, b) において、下位 0.1% の値」を意味し、`beta.ppf(0.999, a, b)` は、「上位 99.9% の点（＝ほぼ最大値）」を意味する。

・`np.linspace(`**最小値, 最大値**`, 100)` は、`linspace()` は等間隔に並んだ点を作る関数で、ベータ分布の 0.1%〜99.9% の範囲を 100 分割 して、$x = [x_1, x_2, \ldots, x_{100}]$ という形で「描画のための x 軸」を作っている

付録 A　自然対数の底 (Napier 数）e について

次いで、自然対数の底となる e を準備する。a^x を微分しても a^x となるような特別な底 a の値を e と定めて、実際にその e の値を求めるという都合の良い事をする。実際、この自然対数は色んな所で活躍する数である。

まず、指数関数 $y = a^x$ の a の値が変化した場合どのようなグラフになるかを示したのが図 62 の (a) である。$a^0 = 1$ なので、全ての場合に $(0, 1)$ を通るが、a の値が大きくなるほど急激に立ち上がっている。

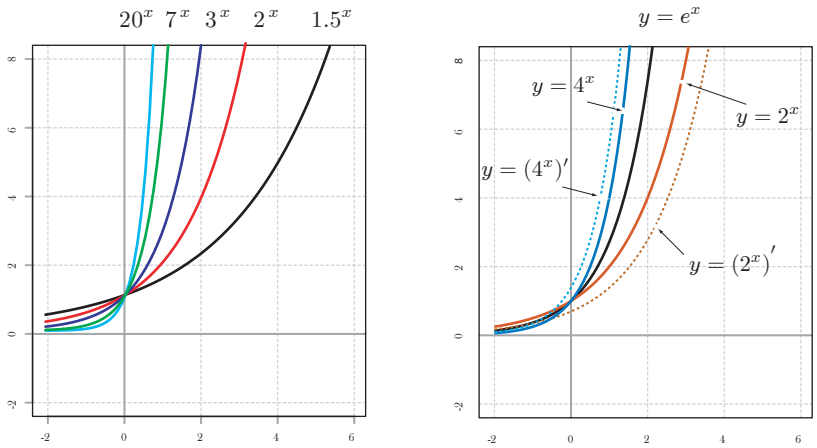


図 62　指数関数の変化と指数関数の微分

また、図 62 の (b) は、$y = 2^x$ と $y = 4^x$ のグラフを描き、さらにその微分 $y = (2^x)'$ と $y = (4^x)'$ を描いたものである。底が 2 と 4 の場合、それぞれ微分したものが一致していない。つまり、2^x と $(2^x)'$、4^x と $(4^x)'$ のグラフは一致していない。このグラフが一致しているような a の値を e と定義するのである。

つまり、指数関数が微分しても変わらないように指数関数の底 e を定めるのである

では、その数はどういう数なのか、微分の定義式に当てはめてみてゆこう

$$\begin{aligned}(a^x)' &= \lim_{h\to 0}\frac{a^{x+h}-a^x}{h}\\ &= \lim_{h\to 0}\frac{a^x\cdot a^h-a^x}{h}\\ &= a^x\cdot\lim_{h\to 0}\frac{a^h-1}{h}\end{aligned}$$

つまり

$$\lim_{h\to 0}\frac{a^h-1}{h} = 1$$

となるような a を求めれば良いわけである。今、h が限りなくゼロに近いけどゼロではないと考えた時、上記の式が成立したものとして変形してゆくと

$$\begin{aligned} a^h - 1 &= h \\ a^h &= 1 + h \qquad \text{両辺を } \frac{1}{h} \text{ 乗して} \\ (a^h)^{\frac{1}{h}} &= (1+h)^{\frac{1}{h}} \\ a &= (1+h)^{\frac{1}{h}} \end{aligned}$$

つまり、h を限りなくゼロに近づけた時の $(1+h)^{\frac{1}{h}}$ を e と定義してあげれば良い。

定義 付録 A.1. 自然対数の底（*Napier* 数）e を以下のように定義する。

$$e = \lim_{h \to 0} (1+h)^{\frac{1}{h}} \tag{付録 A.1}$$

$$e = \lim_{n \to \infty} (1 + \frac{1}{n})^n \tag{付録 A.2}$$

式 (付録 A.2) の n は、$n = \dfrac{1}{h}$ と逆数に変換したものである。それによって n をゼロではなく無限大 ∞ に近づける事になるが式としては同じ事である。

■自然対数の底の値を求める　式 (付録 A.2) を利用して、e の値を求めておこう。表 9 のように、順次 n を増加させていってその値がどんな値に近づくかを調べると、$e = 2.718281828459\cdots$ となる

表 9　自然対数の底を求める

n	$(1+\frac{1}{n})^n$	$=$ 値
1	$(1+1)^1$	$= 2$
10	$(1+0.1)^{10}$	$= 2.59374\cdots$
100	$(1+0.01)^{100}$	$= 2.70481\cdots$
1000	$(1+0.001)^{1000}$	$= 2.71692\cdots$
10000	$(1+0.0001)^{10000}$	$= 2.71815\cdots$

自然対数の底の値は

$$e = \lim_{n \to \infty} (1 + \frac{1}{n})^n = 2.718281828459\cdots \tag{付録 A.3}$$

■本当に指数関数の微分が変わらないかの確認　先ほどの式に当てはめて、本当に指数関数の微分が変わらないかの確認してみよう。まず先ほどと同様に、微分の定義式に当てはめる。この時式の指数関数の底 a を e に

変えると

$$\begin{aligned}(e^x)' &= \lim_{h\to 0}\frac{e^{x+h}-e^x}{h}\\ &= \lim_{h\to 0}\frac{e^x\cdot e^h-e^x}{h}\\ &= e^x\cdot\lim_{h\to 0}\frac{e^h-1}{h}\end{aligned}$$

ここで、定義式 (付録 A.1) を利用して e^h を求めると

$$e^h = \lim_{h\to 0}(1+h)$$

これは、h がゼロに近づけば近づくほど、e^h は 1 に近づくという事を意味している。よって (e^h-1) はゼロに近づく、つまり

$$\lim_{h\to 0}\frac{e^h-1}{h} \Rightarrow \frac{0}{0}=1 \qquad \text{(付録 A.4)}$$

ここで h は、ゼロに近づくが決してゼロにはならない、つまり $h\neq 0$ なので・・割り算が出来て 1 になっている。したがって、

$$(e^x)' = e^x\cdot\lim_{h\to 0}\frac{e^h-1}{h} = e^x$$

■点 $(0,1)$ における接線の傾きがちょうど 1 である事の確認

微分しても変わらないような数として e を定義した。この e を底とした指数関数について、もう少しその性質を見ておこう。図 63 のように、指数関数 $y=e^x$ と対数関数 $y=\log x$ は、$y=x$ に関して対象であり、$y=e^x$ は点 $(0,1)$ を通り、$y=\log x$ は点 $(1,0)$ を通り、両方とも 1 で座標軸と交わる。

この時、底を e にすると、$y=e^x$ 上の点 $(0,1)$ における接線の傾きがちょうど 1 になる。

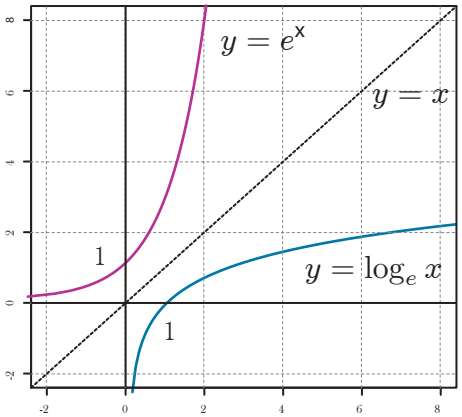


図 63　指数関数と対数関数のグラフ

この事を確認しよう。まず、$y = e^x$ 上の点 $(0, 1)$ における接線の傾きは

$$\begin{aligned} f'(0) &= \lim_{h \to 0} \frac{f(0+h) - f(0)}{h} \\ &= \lim_{h \to 0} \frac{e^{0+h} - e^0}{h} \\ &= \lim_{h \to 0} \frac{e^h - 1}{h} \end{aligned}$$

この式は、式 (付録 A.4) と同じで 1 であり

$$f'(0) = 1$$

付録 B　マクローリン展開とオイラーの公式

B.1　マクローリン展開

マクローリン展開を用いると，一般の関数 $f(x)$ を多項式で近似することができ、三角関数、指数関数、対数関数を多項式のように扱うことができる。x の関数 $f(x)$ のマクローリン展開は以下[*37]。

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots + \frac{f^{(n)}(0)}{n!}x^n + \cdots \qquad \text{(付録 B.1)}$$

■マクローリン展開の確認　　まず、x の関数 $f(x)$ が以下のような形の無限級数で表されると仮定する。一旦、無前提に仮定するのである。

$$f(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n + \cdots$$

いま、この両辺を繰り返し微分すると

$$\begin{aligned} f'(x) &= a_1 + 2a_2x + 3a_3x^2 + \cdots + na_nx^{n-1} + \cdots \\ f''(x) &= \quad 2a_2 + 3\cdot 2a_3x + \cdots + n(n-1)a_nx^{n-2} + \cdots \\ f'''(x) &= \quad\quad 3\cdot 2a_3 + \cdots + n(n-1)(n-2)a_nx^{n-3} + \cdots \end{aligned}$$

これらの式で $x = 0$ とおくと定数項だけが残るので

$$\begin{aligned} f(0) &= a_0 \\ f'(0) &= a_1 \\ f''(0) &= 2!a_2 \\ f'''(0) &= 3!a_3 \\ &\cdots \\ f^{(n)}(0) &= n!a_n \end{aligned}$$

したがって、元の関数 $f(x)$ の係数 a_n は

$$a_n = \frac{f^{(n)}(0)}{n!}$$

つまり

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots + \frac{f^{(n)}(0)}{n!}x^n + \cdots$$

と表す事ができる。

[*37] $f(x)$ の $x = 0$ を中心としたテイラー展開のことを特にマクローリン展開と呼ぶ

B.2　三角関数・指数関数のマクローリン展開

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \qquad \text{(付録 B.2)}$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \qquad \text{(付録 B.3)}$$

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \cdots \qquad \text{(付録 B.4)}$$

■$\sin x$ のマクローリン展開の確認　sin の微分式と cos の微分式式が以下である事を利用する

$$(\sin x)' = \cos x \quad 、 \quad (\cos x)' = -\sin x$$

$f(x) = \sin x$ とおいて微分を重ねると

$$f^{(1)}(x) = \cos x, \quad f^{(2)}(x) = -\sin x, \quad f^{(3)}(x) = -\cos x, \quad f^{(4)}(x) = \sin x$$
$$f^{(5)}(x) = \cos x, \quad f^{(6)}(x) = -\sin x, \quad f^{(7)}(x) = -\cos x, \quad f^{(8)}(x) = \sin x$$

$x = 0$ の時のこれらの値は

$$f^{(1)}(0) = 1, \quad f^{(2)}(0) = 0, \quad f^{(3)}(0) = -1, \quad f^{(4)}(0) = 0$$
$$f^{(5)}(0) = 1, \quad f^{(6)}(0) = 0, \quad f^{(7)}(0) = -1, \quad f^{(8)}(0) = 0$$

先にみたように、マクローリンの展開の式付録 B.1 は

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots + \frac{f^{(n)}(0)}{n!}x^n + \cdots$$

この式に代入して

$$\begin{aligned}\sin x &= 0 + x + \frac{-1}{3!}x^3 + \frac{1}{5!}x^5 + \frac{-1}{7!}x^7 + \cdots \\ &= x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \frac{1}{7!}x^7 + \cdots\end{aligned}$$

■$\cos x$ のマクローリン展開の確認　これは、今の結果と sin の微分の式の $(\sin x)' = \cos x$ を利用すれば簡単。

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

なので、この両辺を微分して

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots$$

■e^x のマクローリン展開の確認　指数関数の微分の式から、e^x は微分しても e^x のままなので $f(x) = e^x$ とすると

$$f^{(n)} = e^x$$

つまり、

$$e^x = 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots + \frac{f^{(n)}(0)}{n!}x^n + \cdots$$

B.3　オイラーの公式

$$e^{i\theta} = \cos\theta + i\sin\theta \qquad \text{(付録 B.5)}$$

■オイラーの公式を確認する　先の指数関数の展開式 (付録 B.4) は

$$e^x = 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \frac{1}{5!}x^5 + \frac{1}{6!}x^6 + \frac{1}{7!}x^7 + \cdots$$

ここで、$x = i\theta$ とおくと

$$\begin{aligned} e^{i\theta} &= 1 + \frac{1}{1!}i\theta + \frac{1}{2!}i\theta^2 + \frac{1}{3!}i\theta^3 + \frac{1}{4!}i\theta^4 + \frac{1}{5!}i\theta^5 + \frac{1}{6!}i\theta^6 + \frac{1}{7!}i\theta^7 + \cdots \\ &= 1 + \frac{1}{1!}i\theta - \frac{1}{2!}\theta^2 - \frac{1}{3!}i\theta^3 + \frac{1}{4!}\theta^4 + \frac{1}{5!}i\theta^5 - \frac{1}{6!}\theta^6 - \frac{1}{7!}i\theta^7 + \cdots \\ &= \left(1 - \frac{1}{2!}\theta^2 + \frac{1}{4!}\theta^4 - \frac{1}{6!}\theta^6 - \cdots\right) + i\left(\frac{1}{1!}\theta - \frac{1}{3!}\theta^3 + \frac{1}{5!}\theta^5 - \frac{1}{7!}\theta^7 + \cdots\right) \end{aligned}$$

ここで

$$\begin{aligned} \cos\theta &= 1 - \frac{1}{2!}\theta^2 + \frac{1}{4!}\theta^4 - \frac{1}{6!}\theta^6 + \cdots \\ \sin\theta &= \theta - \frac{1}{3!}\theta^3 + \frac{1}{5!}\theta^5 - \frac{1}{7!}\theta^7 + \cdots \end{aligned}$$

より

$$e^{i\theta} = \cos\theta + i\sin\theta$$

付録 C　重積分

1 変数の定積分を 2 変数に拡張したものを重積分と呼ぶ。定積分が面積を表しているのに対して、重積分は体積を意味する。重積分の定義にはリーマン積分の考え方（ページ**??**参照）を使う。変数 x, y が定義する xy 平面を細かく区分して、それぞれの区分毎に「区分面積と関数値 $z = f(x, y)$」との積和をとる。そして、この区分面積を極限まで小さくしていく事で重積分を定義する。

C.1　重積分の定義

図 64(b) のような関数 $z = f(x, y)$ と xy 平面上の長方形 K があるとし、図 64(a) のように平面 K の範囲 $a \leqq x \leqq b$、$c \leqq y \leqq d$ を x 軸を n 個、y 軸を m 個に区分してあるものとする。その時、各区分の代表点 $P_{ij} = (x_i, y_j)$ の関数値 $f(P_{ij})$ を高さとするひとつひとつの直方体の体積を集めた V を求めよう。

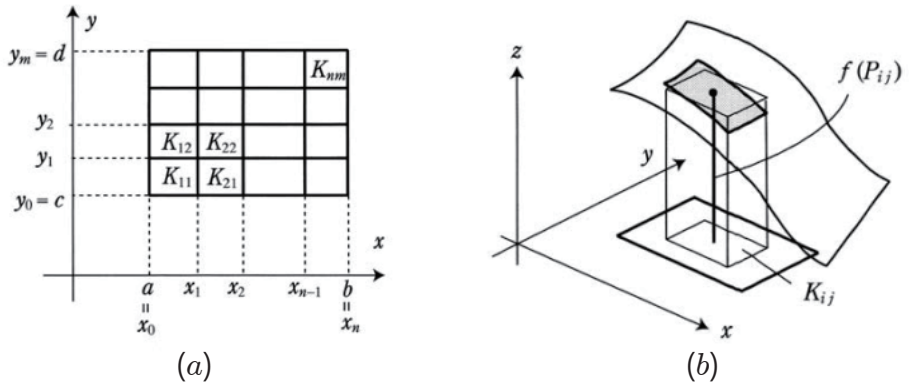


図 64　重積分と体積

ひとつひとつの区画を $\Delta x_i = x_i - x_{i-1}$、$\Delta y_i = y_i - y_{i-1}$ と表示すると

$$\begin{aligned} V = & f(P_{11})\Delta x_1\Delta y_1 + f(P_{21})\Delta x_2\Delta y_1 + \cdots + f(P_{n1})\Delta x_n\Delta y_1 + \\ & f(P_{12})\Delta x_1\Delta y_2 + f(P_{22})\Delta x_2\Delta y_2 + \cdots + f(P_{n2})\Delta x_n\Delta y_2 + \\ & \vdots \\ & f(P_{1m})\Delta x_1\Delta y_m \quad + \quad f(P_{2m})\Delta x_2\Delta y_m \quad + \cdots + f(P_{nm})\Delta x_n\Delta y_m \end{aligned}$$

シグマ記号であらわすと、

$$V = \sum_i^n \sum_j^m f(P_{ij})\Delta x_i\Delta y_j$$

定義 付録 C.1. 重積分の定義　　$n \to \infty$、$m \to \infty$ の時に、この体積の和の極限が存在するならば、それを $f(x, y)$ の領域 D における重積分と呼び、以下のように表す。

$$\int\int_D f(x, y)\, dxdy = \lim_{n,m\to\infty} \sum_i^n \sum_j^m f(P_{ij})\Delta x_i\Delta y_j \qquad \text{(付録 C.1)}$$

表 10 のように $\int\int$ が $\sum\sum$ に、dx, xy が $\Delta x\Delta y$ に対応している事に注目。

表 10　極限と有限

極限の世界		有限の世界
$\int\int$	$\Leftrightarrow$	$\sum\sum$
dx、dy	$\Leftrightarrow$	Δx、Δy

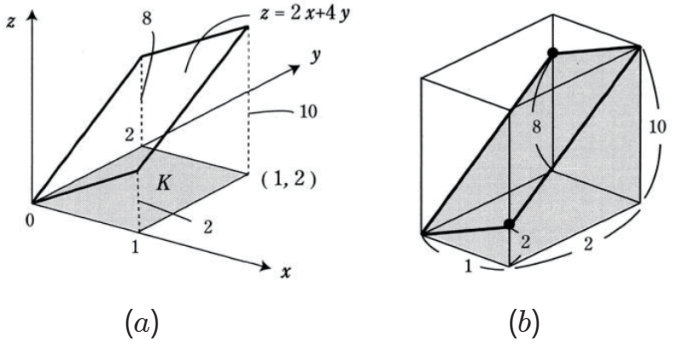
極限の世界の $f(x,y)\times dx\times dy$ を有限の世界でみると、「高さ $f(P_{ij})$」×「底面積 $(\Delta x\times\Delta y)$」を示していると考えてよい。

■計算事例　簡単な事例を元に計算過程を追いかけてみる。

例題 付録 C.1. 双一次関数である $z=2x+4y$　の長方形領域 $K(0\leqq x\leqq 1,\quad 0\leqq y\leqq 2)$ 上での重積分

$$I=\int\int_K(2x+4y)dxdy$$

を求める



(a)　(b)

図 65　重積分の事例

図形的に求める

この関数 $z=2x+4y$ は平面をつくり、(x,y) が $f(1,0)=2$、$f(1,2)=10$、$f(0,2)=8$ なので、図 65 の (b) の斜線部分が求める体積。これは、$1\times 2\times 10$ の直方体の半分になるので

$$1\times 2\times 10\nabla\cdot 2=10$$

リーマン和で求める

まず、以下のように $n\times m$ 個の小区間に区分する

区間 $0\leqq x\leqq 1$ を n 個に分解して

$$x_0=0,\ x_1=\frac{1}{n},\ x_2=\frac{2}{n},\ \cdots,\ x_i=\frac{i}{n},\ \cdots,\ x_n=\frac{n}{n}$$

区間 $0 \leqq y \leqq 2$ を m 個に分解して

$$y_0 = 0,\ y_1 = \frac{2}{m},\ y_2 = \frac{4}{m},\ \cdots,\ \ y_i = \frac{2j}{m},\ \cdots,\ y_m = \frac{2m}{m}$$

そして、各区間の代表点 $P_{ij} = (p_i, p_j)$ を各小区間の右上の頂点とすると、それぞれの頂点は以下のように表す事ができる。

$$P_{ij} = 2x_i + 4y_j = 2\frac{i}{n} + 4\frac{2j}{m}$$

リーマン和 V は、各小区間の面積が $1/n \times 2/m$ なので

$$V = \sum_{1 \leqq i \leqq n} \sum_{1 \leqq j \leqq m} \left(2\frac{i}{n} + 4\frac{2j}{m}\right) \frac{1}{n}\frac{2}{m}$$

まず j を固定して i を動かすと[*38]以下のような i の消えた式になる。

$$\begin{aligned}
&\left(2\frac{1}{n} + 4\frac{2j}{m}\right)\frac{2}{nm} + \left(2\frac{2}{n} + 4\frac{2j}{m}\right)\frac{2}{nm} + \cdots + \left(2\frac{n}{n} + 4\frac{2j}{m}\right)\frac{2}{nm} \\
&= \frac{2}{nm}\left\{2\left(\frac{1+2+\cdots+n}{n}\right) + 4\frac{2nj}{m}\right\} \\
&= \frac{2}{nm}\left\{(n+1) + \frac{8n}{m}j\right\}
\end{aligned}$$

つぎに j を動かすと

$$\begin{aligned}
&\frac{2}{nm}\left\{(n+1) + \frac{8n}{m}1\right\} + \frac{2}{nm}\left\{(n+1) + \frac{8n}{m}2\right\} + \cdots + \frac{2}{nm}\left\{(n+1) + \frac{8n}{m}m\right\} \\
&= \frac{2}{nm}\left\{m(n+1) + \frac{8n}{m}(1+2+\cdots+m)\right\} \\
&= \frac{2}{nm}\left\{m(n+1) + 4n(m+1)\right\}
\end{aligned}$$

このように i を先に、次に j を動かして合計したものが V。この V を以下のように変形。

$$\begin{aligned}
V &= \frac{2}{nm}\left\{m(n+1) + 4n(m+1)\right\} \\
&= 2\left(\frac{n+1}{n}\right) + 8\left(\frac{m+1}{m}\right) \\
&= 2\left(1 + \frac{1}{n}\right) + 8\left(1 + \frac{1}{m}\right)
\end{aligned}$$

この時、区分をどんどんと小さくする、つまり n と m を無限大に近づけていくと

$$n \to \infty \text{ ならば} \quad \frac{1}{n} = 0 \quad , \quad m \to \infty \text{ ならば} \quad \frac{1}{m} = 0$$

なので、$V = 2\,(1+0) + 8\,(1+0) = 10$ となり、先に図形的に求めた結果と同じ。

[*38] 途中の $\dfrac{1+2+\cdots+n}{n}$ の変形では $(1+2+\cdots+n) = \dfrac{n(n+1)}{2}$ を利用

累次積分で求める

累次積分とは、重積分

$$\int\int f(x,y)dxdy$$

を解くときに

$$\int\left\{\int f(x,y)dx\right\}dy$$

というように、先に x で積分して、その結果を次に y で積分をするという 2 段構成にする積分方法で、「逐次積分」とも呼ばれる（参照フビニの定理付録 C.1）。実際に事例でやってみる。

$$\begin{aligned}
V &= \int_0^2\int_0^1 (2x+4y)\,dxdy \\
&= \int_0^2\left\{\int_0^1 (2x+4y)dx\right\}dy \\
&= \int_0^2\left\{[x^2+4yx]_1-[x^2+4yx]_0\right\}dy \\
&= \int_0^2 (1+4y)dy \\
&= [y+2y^2]_2-[y+2y^2]_0 \\
&= 10
\end{aligned}$$

というように、先の 2 つの結果と同じである。

定理 付録 C.1. フビニの定理　積分区間で連続な関数であれば、重積分は累次積分に変形することが出来る。

$z=f(x,u)$ の長方形 $K(a\leqq x\leqq b, c\leqq y\leqq d)$ における重積分

$$\int\int_K f(x,u)dxdy$$

は、次の累次積分で計算できる。

$$\int_c^d\left\{\int_a^b f(x,y)dx\right\}dy$$

C.2　重積分の変数変換とヤコビアン

重積分においても、１変数の置換積分と同様に変数変換を用いてより簡単に計算する事ができる。特に、正規分布の積分を求める時など計算を簡単にする為には変数変換が必要。その際に置換積分の変化率のように変数変換の拡大率が重要になる。その拡大率を表す行列式のことをヤコビアンと言う。

■置換積分と変化率　　まずは１変数の置換積分における変化率を再考する。以下が置換積分の式 (**??**)

$$\int f(x)dx = \int f(g(t)) \cdot g'(t)dt$$

この式はまた以下のようにも書くことができる。

$$\int f(x)dx = \int f(g(t))\frac{dx}{dt}dt$$

この $\dfrac{dx}{dt}$ は、変数変換に用いる関数 $x = g(t)$ の接線であり t に対する x の変化率、つまり t が少し動いた時にどの程度 x が動くかを意味している。表 10 で説明した有限の世界でいえば、$\dfrac{\Delta x}{\Delta t}$ を意味している事になる。

「t の世界に変数変換する」ためには、元の関数 $f(x)$ を t で表すだけでなく、

Δx を新しい変数 t の変化に変換するために、変化率 $\dfrac{\Delta x}{\Delta t}$ をかける必要がある。

多変数の重積分において、１変数の変化率を意味するものがヤコビアン (Jacobian) と呼ばれる行列式である。

■ヤコビアンとその意味　　まずは２変数の場合の重積分で確認してみる。変数 x、y の空間を変数 u、v の空間に変換する事を考える。その対応を示す関数を $x = \varphi(u, v)$、$y = \psi(u, v)$ としたとき（φ はファイ、ψ はプサイと読む)、x、y の全微分は式 (**??**) のように

$$dx = \frac{\partial \varphi}{\partial u}\,du + \frac{\partial \varphi}{\partial v}\,dv$$
$$dy = \frac{\partial \psi}{\partial u}\,du + \frac{\partial \psi}{\partial v}\,dv$$

と表す事ができる。これを行列を用いて表すと

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = \begin{pmatrix} \dfrac{\partial \varphi}{\partial u} & \dfrac{\partial \varphi}{\partial v} \\ \dfrac{\partial \psi}{\partial u} & \dfrac{\partial \psi}{\partial v} \end{pmatrix} \begin{pmatrix} du \\ dv \end{pmatrix}$$

となる。この行列をヤコビ行列とよび、慣習的に行列 J で表す事が多い。つまり、

$$J = \begin{pmatrix} \dfrac{\partial \varphi}{\partial u} & \dfrac{\partial \varphi}{\partial v} \\ \dfrac{\partial \psi}{\partial u} & \dfrac{\partial \psi}{\partial v} \end{pmatrix}$$

また、この行列 J の行列式をヤコビアン (Jacobian) または関数行列式とよび、慣習的に以下のように表す。

$$|J| = \frac{\partial(\varphi, \psi)}{\partial(u.v)} = \begin{vmatrix} \dfrac{\partial\varphi}{\partial u} & \dfrac{\partial\varphi}{\partial v} \\ \dfrac{\partial\psi}{\partial u} & \dfrac{\partial\psi}{\partial v} \end{vmatrix}$$

このヤコビ行列 J が何を意味しているかというと、図 66 のように、元の座標空間 (u, v) を新しい座標空間 (x, y) に対応させる一次変換行列であると考えられる。このように行列 J が元の座標から新しい座標への一次変換だとすると、J の行列式 $|J|$ は、この一次変換の拡大率を意味している。

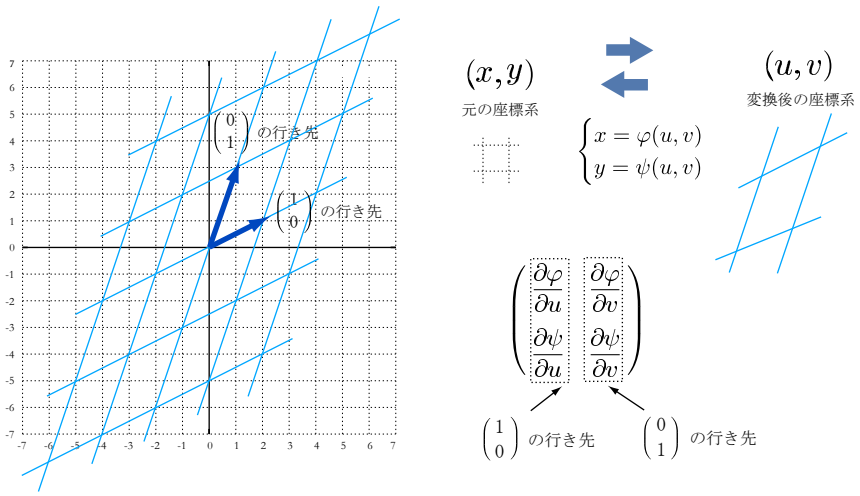


図 66　ヤコビ行列の意味

■重積分の変数変換　このヤコビ行列をつかった重積分の変数変換公式について述べる。

公式 付録 C.1. 重積分の変数変換

2 変数関数 $f(x, u)$ の重積分

$$I = \int\int_D f(x, y)\, dxdy$$

において、変数を $x = g(s, t)$ から $y = h(s, t)$ に変換したとき、被積分関数 $f(x, y)$ が、$k(s, t) = f(g(s, t), h(s, t))$ に変換され、領域 D が領域 E に変換されたとすると以下のように表す事ができる。

$$I = \int\int_E k(s, t)\, |J|\, ds\, dt \qquad \text{(付録 C.2)}$$

ここで

$$|J| = \frac{\partial(\varphi, \psi)}{\partial(u.v)} = \begin{vmatrix} \dfrac{\partial\varphi}{\partial u} & \dfrac{\partial\varphi}{\partial v} \\ \dfrac{\partial\psi}{\partial u} & \dfrac{\partial\psi}{\partial v} \end{vmatrix}$$

上記の事は、１変数の場合の置換積分の式 (**??**) と同様に、以下のような操作をするイメージで理解すれば良いと思う。

「(s,t) の世界に変数変換する」ためには、元の関数 $f(x,y)$ を $k(s,t)$ に変換するだけでなく、
dx、dy を新しい変数 ds、dt に変換するために、変化率 $|J|$ をかける必要がある。

■ヤコビアンの多変数への拡張　ヤコビアンを多変数に拡張しておこう。関数 $f_1, f_2, \cdots, f_n$ として行列で表すと

$$\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix} = \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \cdots & \dfrac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \cdots & \vdots \\ \dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{pmatrix} \begin{pmatrix} dx_1 \\ dx_2 \\ \vdots \\ dx_n \end{pmatrix}$$

以下のように、この行列を J とした場合の行列式 $|J|$ が、多変数のヤコビアンであり、$|J| = \dfrac{\partial(f_1, f_2, \cdots, f_n)}{\partial(x_1, x_2, \cdots, x_n)}$ と表す。

$$|J| = \begin{vmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \cdots & \dfrac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \cdots & \vdots \\ \dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{vmatrix}$$

以上のように、このヤコビ行列をつかった変数変換は以下のようになる。

２変数関数 $f(x,y)$ の重積分

$$I = \int\int_D f(x,y)\, dxdy$$

において、変数 (x,y) を $x = \varphi(u,v)$ と $y = \psi(u,v)$ という関数によって変数 (u,v) に変換したとき、被積分関数 $f(x,y)$ が、$g(s,t) = f(\varphi(u,v), \psi(u,v))$ に変換され、領域 D が領域 E に変換されたとすると以下のように表す事ができる。

$$I = \int\int_E |J|\, g(u,v)\, du\, dv \qquad \text{(付録 C.3)}$$

ここで

$$|J| = \frac{\partial(\varphi,\psi)}{\partial(u,v)} = \begin{vmatrix} \dfrac{\partial\varphi}{\partial u} & \dfrac{\partial\varphi}{\partial v} \\ \dfrac{\partial\psi}{\partial u} & \dfrac{\partial\psi}{\partial v} \end{vmatrix}$$

上記の事は、1 変数の場合の置換積分と同様に、以下のような操作をするイメージで理解すれば良いと思う。

「(s,t) の世界に変数変換する」ためには、元の関数 $f(x,y)$ を $k(s,t)$ に変換するだけでなく、
dx、dy を新しい変数 ds、dt に変換するために、変化率 $|J|$ をかける必要がある。

例題 付録 C.2. 関数 $f(x,y)$ があり、それを以下のように変数変換したとする。その時の関数 $g(z,w)$ はどのように変換されるか？

$$\begin{cases} z &= 3x+y \\ w &= x+2y \end{cases}$$

$f(x,y)$ の x と y を z と w で表した式 $g(z,w)$ を求める事なので、上記の連立方程式をといて

$$\begin{cases} x = \dfrac{2z-w}{5} \\ y = \dfrac{3w-z}{5} \end{cases}$$

変換後の点 (z,w) に対応する変換前の点 (x,y) は以下のようになる。

$$(x,y) = \left(\frac{2z-w}{5}, \frac{3w-z}{5}\right)$$

今求めたいのは $g(z,w)$ の値であり、$g(z,w)$ は拡大縮小率を J とすると以下のように表すことができる。

$$g(z,w) = J \cdot f\left(\frac{2z-w}{5}, \frac{3w-z}{5}\right)$$

次に、この時の拡大縮小率を考えてみる。この変数変換は、図 67 のように元々の基底ベクトル $e_x = (1,0)$ と $e_y = (0,1)$ をそれぞれ $e_z = (3,1)$ と $e_w = (1,2)$ に移す。元の基底ベクトルがつくる四角形の面積は 1 である。変換後の基底ベクトル $e_z = (3,1)$ と $e_w = (1,2)$ がつくる平行四辺形の面積を求めれば拡大率がわかる。

図 67 のように図形的に解いてみる。青の四角形と赤の三角形と黄色の三角形の面積を $4 \times 3 = 12$ から引いて 5 となる。つまり面積が 5 倍になっているので、確率密度は $1/5$ となる。

$$g(z,w) = \frac{1}{5} \cdot f\left(\frac{2z-w}{5}, \frac{3w-z}{5}\right)$$

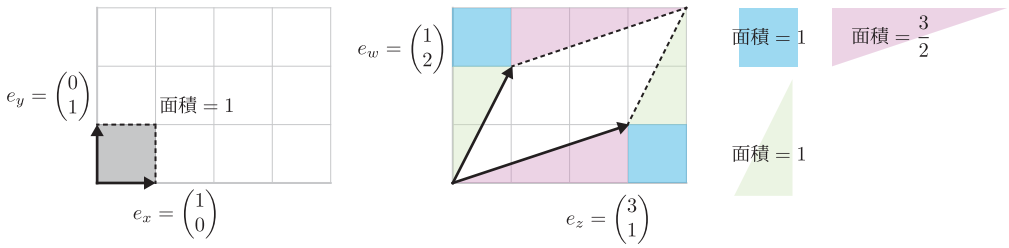


図 67　変数変換による面積の変化

この面積の拡大率を求める過程を行列を用いながら解いていこう。まず与えられた変数変換を行列表現すると以下。

$$\begin{pmatrix} z \\ w \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

この一次変換行列を以下のように表現すると、この行列 A の行列式 $|A|$ が面積の拡大縮小率を表している。

$$A = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$$

２次元の行列式は以下。

$$X = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{ならば、} \qquad |X| = (ad - bc)$$

なので、以下のように面積は５倍となる。

$$|A| = (3 \times 2 - 1 \times 1) = 5$$

例題 付録 C.3. 一対一対応しているが線形変換でない変数変換について考える。関数 $f(x, y)$ があり、それを以下のように変数変換したとする。その時の z、w の確率密度関数 $g(z, w)$ はどのように変換されるか？

$$\begin{cases} z &= xe^y \\ w &= y \end{cases}$$

この変換は図 68 のように場所によって拡大率が異なる変換である。

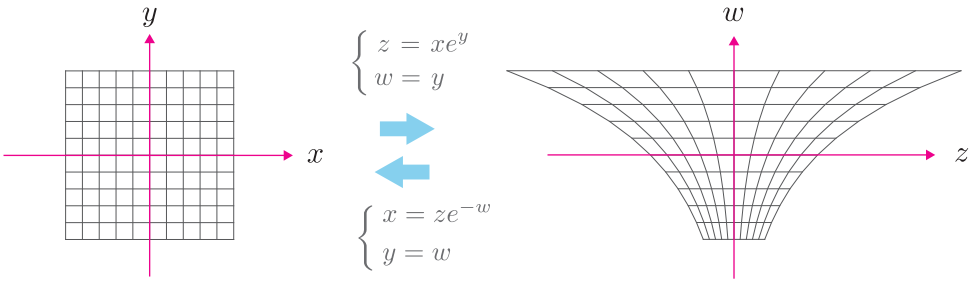


図 68　面積拡大率が場所によって異なる場合

まず与えられた変換式を x と y について解くと

$$\begin{cases} x = ze^{-w} \\ y = w \end{cases}$$

つまり、変換後の座標が (z, w) であったとすると、その場合に対応する変換前の座標 (x, y) は以下のように表す事ができる。

$$(x, y) = (ze^{-w}, w)$$

また、変換による拡大率を $|J|$ とするとその確率密度関数 $g(z,w)$ は、以下のように表す事ができる。

$$g(z,w) = |J|\ f(ze^{-w}, w)$$

この $|J|$ を求めるのであるが、面積がどのように拡大されるかは場所によって異なるので、各座標点 (x,y) における面積拡大率を調べる。簡単に想定すると、y 軸方向は $w = y$ なので拡大率はゼロで、x 軸方向は $z = xe^y$ なので e^y 倍されている事になる。ここで求めたいのは (z,w) の式なので z で表すと $x = ze^{-w}$ より e^{-w} 倍となる。以上より、確率密度関数 $g(z,w)$ は下式のようになると想定される。

$$g(z,w) = \frac{1}{e^w}\ f(ze^{-w}, w)$$

次に、先の事例と同様に面積の拡大率を求める過程を行列を用いながら解いていこう。まず与えられた変数変換を行列表現すると以下。

$$\begin{pmatrix} z \\ w \end{pmatrix} = \begin{pmatrix} e^y & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

したがってその行列式は $|e^y|$ となる。これを新しい座標系である z と w の座標で表すと $|e^w|$ なので、新しい座標系での関数は以下のように表現できる。

$$g(z,w) = \frac{1}{|e^w|}\ f(ze^{-w}, w)$$

ソースコード 17　面積拡大率が場所によって異なる場合のグラフを作成するプログラム

```
import numpy as np
import matplotlib.pyplot as plt

#変換関数
def fx(x, y):
z = x * np.exp(y)
w = y
return z, w

x_min = −2.8 ; x_max = 2.8
y_min = −2.8 ; y_max = 2.8

#変換前のグラフを描く
fig, ax = plt.subplots()
ax.set_xlim(x_min, x_max) ; ax.set_ylim(y_min, y_max)
X, Y = np.meshgrid(np.arange(−1, 1.2, 0.2),np.arange(−1, 1.2, 0.2))
plt.plot(X,Y) ; plt.plot(X.T,Y.T)

#変換後のグラフを描く
fig, ax2 = plt.subplots()
ax2.set_xlim(x_min, x_max) ; ax2.set_ylim(y_min, y_max)
W, Z = fx(X, Y)
plt.plot(W,Z) ; plt.plot(W.T,Z.T)

plt.show()
```

C.3　重積分の極座標への変数変換

■極座標のヤコビアンについて

直交座標と極座標は互いに変換可能で、図 69 のような関係がある。

$$x = r\cos\theta \qquad y = r\sin\theta$$

なので二次元平面上の同じ点を $(x, y) = (r\cos\theta, r\sin\theta)$ とあらわす事ができる。具体的に r と θ を x と y から求めると以下のようになる。

$$r = \sqrt{x^2 + y^2} \qquad \theta = \tan^{-1}\frac{y}{x}$$

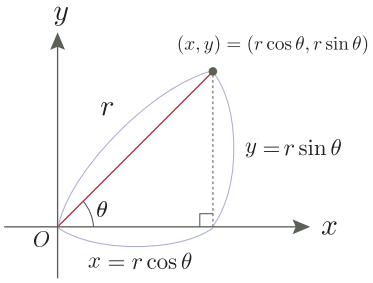


図 69　極座標変換

ここで関数 $f(x, y)$ を極座標系で表した関数 $f(r\cos\theta, r\sin\theta)$ の編微分を考える。まずは、極座標系での偏微分は

$$r \text{ で偏微分：} \qquad \frac{\partial x}{\partial r} = \cos\theta\,, \quad \frac{\partial y}{\partial r} = \sin\theta$$

$$\theta \text{ で偏微分：} \qquad \frac{\partial x}{\partial \theta} = -r\sin\theta\,, \quad \frac{\partial y}{\partial \theta} = r\cos\theta$$

求めたいのは r や θ に関する偏微分 $\dfrac{\partial f}{\partial r}$、$\dfrac{\partial f}{\partial \theta}$ を元の座標系の x や y の偏微分 $\dfrac{\partial f}{\partial x}$、$\dfrac{\partial f}{\partial y}$ で表す事である。上記の偏微分を合成関数の微分公式に当てはめていくと

$$\frac{\partial f}{\partial r} = \frac{\partial f}{\partial x}\cdot\frac{\partial x}{\partial r} + \frac{\partial f}{\partial y}\cdot\frac{\partial y}{\partial r} = \cos\theta\cdot\frac{\partial f}{\partial x} + \sin\theta\cdot\frac{\partial f}{\partial y}$$

$$\frac{\partial f}{\partial \theta} = \frac{\partial f}{\partial x}\cdot\frac{\partial x}{\partial \theta} + \frac{\partial f}{\partial y}\cdot\frac{\partial y}{\partial \theta} = -r\sin\theta\cdot\frac{\partial f}{\partial x} + r\cos\theta\cdot\frac{\partial f}{\partial y}$$

これを行列で表すと

$$\begin{pmatrix} \dfrac{\partial f}{\partial r} & \dfrac{\partial f}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial f}{\partial x} & \dfrac{\partial f}{\partial y} \end{pmatrix} \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$$

この式の以下の部分をヤコビ行列 J と呼ぶ。これは元も座標系の偏微分を新しい座標系の偏微分に変換する行列となる。

$$J = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$$

この時、ヤコビアン $|J|$ は以下。ヤコビアンは変換の拡大率を意味しており、極座標系への拡大率は r となる。

$$|J| = \begin{vmatrix} \dfrac{\partial x}{\partial r} & \dfrac{\partial x}{\partial \theta} \\ \dfrac{\partial y}{\partial r} & \dfrac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix}$$
$$= \cos\theta \cdot r\cos\theta - (-r)\sin\theta \cdot \sin\theta = r(cos^2\theta + \sin^2\theta) = r$$

■重積分の極座標変換

公式 付録 C.2. 重積分の極座標変換

重積分の領域 D が極座標で以下の範囲のとき

$$\varphi(\theta) \leqq r \leqq \mu(\theta) \quad , \qquad \alpha \leqq \theta \leqq \beta$$

直交座標系の重積分は以下のように極座標系の重積分に変換できる

$$\int\int_D f(x,y)\, dxdy = \int_\alpha^\beta \int_{\varphi(\theta)}^{\mu(\theta)} f(r\cos\theta, r\sin\theta)\, r\, dr\, d\theta \qquad \text{(付録 C.4)}$$

1 変数の $y = f(x)$ の積分を考えるときに x を細かい区分にして、それぞれの y を求めてその総和を考えた。同様に図 70 の (a) の網掛け領域 D を細かい区分に分割して総和を求める事にする。

図 70 の (a) の D を、微小な範囲の半径 Δr と角度 $\Delta\theta$ をもった n 個の「2 つの扇形の差分から作られる疑似四角形（缶詰のパイン形状）」で埋め尽くすとする。ひとつひとつの疑似四角形は図 70 の (b) のようになる。

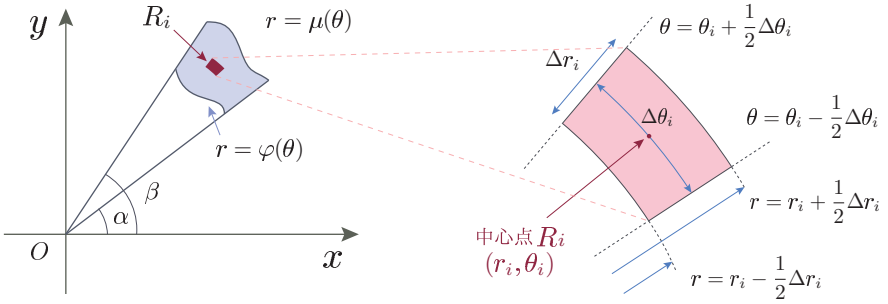


図 70　極座標での微分イメージ

この図 70 の (b) ような「バームクーヘン状の形状」は二つの扇形の差分にすればよい。扇型の面積 S は[*39]

$$S = r^2 \times \frac{\theta}{2}$$

なので、i 番目の小区間の面積を R_i とすると

$$\begin{aligned} R_i &= \left(r_i + \frac{1}{2}\Delta r_i\right)^2 \cdot \frac{\Delta\theta_i}{2} - \left(r_i - \frac{1}{2}\Delta r_i\right)^2 \cdot \frac{\Delta\theta_i}{2} \\ &= \left\{\left(r_i + \frac{1}{2}\Delta r_i\right)^2 - \left(r_i - \frac{1}{2}\Delta r_i\right)^2\right\} \cdot \frac{\Delta\theta_i}{2} \\ &= 2 \cdot r_i \cdot \Delta r_i \cdot \frac{\Delta\theta_i}{2} \\ &= r_i \cdot \Delta r_i \cdot \Delta\theta_i \end{aligned}$$

これで各区間の面積が算出できた。あとはそれぞれの高さをかけて体積にすればよい。図 71 のように、この小区間 i の体積 S_i は

$$S_i = f(r_i \cos\theta_i,\ r_i \sin\theta_i)\ r_i \cdot \Delta r_i \cdot \Delta\theta_i$$

となる。この S_i を n 個分足し合わせれば体積の近似値となる。さらに分割数 n をどんどんと増やして無限大にすれば求める体積 V となる。

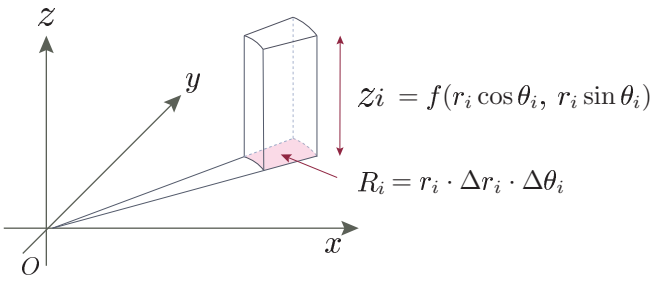


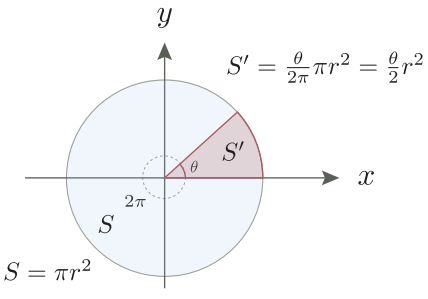
図 71　極座標小区間の体積計算イメージ

つまり、体積 V は

$$V = \lim_{n\to\infty} \sum_{i=1}^{n} f(r_i \cos\theta_i,\ r_i \sin\theta_i) r_i \cdot \Delta r_i \cdot \Delta\theta_i$$

[*39] 扇型の面積は右図のように求める。円全体の面積は $S = \pi r^2$。角度 θ がラジアンで表されているとすると一周 $360°$ は 2π ラジアンなので、円全体の面積の $\dfrac{\theta}{2\pi}$ が求めたい扇形の面積。つまり、

$$S' = \frac{\theta}{2\pi} \times \pi r^2 = \frac{\theta}{2} \times r^2$$



この式は、式 (付録 C.1) と同様に、lim をとることで、Δr、$\Delta\theta$ が dr、$d\theta$ となり、以下の式の右辺になる。またこの時の積分範囲は、領域 D の範囲（$\varphi(\theta) \leqq r \leqq \mu(\theta)$、$\alpha \leqq \theta \leqq \beta$）から設定される。

$$\int\int_D f(x,y)\,dxdy = \int_\alpha^\beta \int_{\varphi(\theta)}^{\mu(\theta)} f(r\cos\theta, r\sin\theta)\; r\; dr\; d\theta$$

付録 D　内積と直交

D.1　内積の定義とそのイメージ

ここでは内積を定義し、そのイメージを物理的現象から掴んでおこう。

内積の定義と性質

ゼロでない 2 つのベクトルを a、b とし、そのなす角度を θ とするとき、内積は以下のように定義される。

$$a \cdot b =< a, b >= |a||b| \cos\theta \tag{付録 D.1}$$

また成分で表現すれば以下のように、成分同士の積和になる。

$$< x, y >= x_1 y_1 + x_2 y_2 + \cdots + x_n y_n \tag{付録 D.2}$$

これをベクトル表示すると

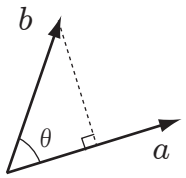
$$< x, y >= \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = x^t y \tag{付録 D.3}$$

また 2 つのベクトルのなす角度が 90° ならば、$\cos\theta = 0$ なので、$< x, y >= 0$ の時は 2 つのベクトルは直交すると言える。

定理 付録 D.1.　内積の定義

以下の図のように、ゼロでない 2 つのベクトル $\boldsymbol{a}, \boldsymbol{b}$ のなす角度を θ とするとき、以下のように定義されるものを $\boldsymbol{a}$,$\boldsymbol{b}$ の内積 *(inner product)* またはスカラー積 *(scalar product)* と呼ぶ。一般に、内積は $\boldsymbol{a} \cdot \boldsymbol{b}$ または、$< \boldsymbol{a}, \boldsymbol{b} >$ と表記される。

$$\boldsymbol{a} \cdot \boldsymbol{b} =< \boldsymbol{a}, \boldsymbol{b} >= |\boldsymbol{a}||\boldsymbol{b}| \cos\theta \tag{付録 D.4}$$



■内積と仕事量　内積は物理的な仕事の定義を考えると意味が理解しやすい。例えば、図 72 のように、物体に斜めの力 F を加えて、水平方向右に距離 s だけ動かしたときの仕事量を考えよう。物理的には、力 F が物体にした仕事量 W は、どれだけの力をどれだけの距離加えたか、つまり「力×距離」で定義される。

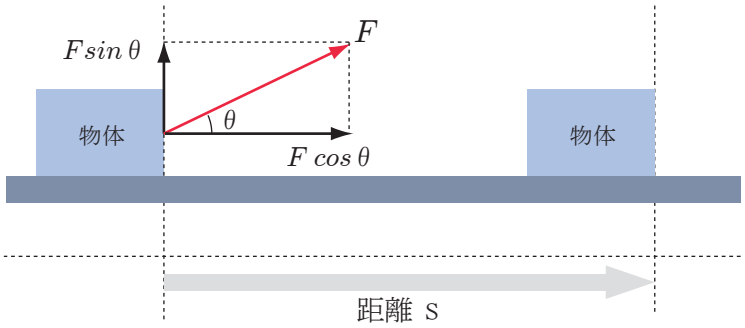


図 72　物理的な仕事の定義

今、図 72 のように力を斜め右上に加えているので、図のようにその力を水平方向と垂直方向に分解すると、物体を移動させるために役立った力は $F\cos\theta$ のみであり、$F\sin\theta$ は移動に関しては実質貢献していない。なので仕事 W は、

$$\begin{aligned} W &= s \times F\cos\theta \\ &= |s||F|\cos\theta \end{aligned}$$

であり、まさに仕事 W という物理的概念を内積で表現できる事になる。

■内積とベクトルの直交　内積は、ベクトルの直交性と関連深い。内積の定義から

$$\cos\theta = \frac{< x, y >}{|x||y|}$$

なので、２つのベクトルのなす角度が 90° ならば、$\cos\theta = 0$ なので、

$< x, y > = 0$	なら２つのベクトルは直交
$< x, y > = |x||y|$	なら２つのベクトルは平行

という事がいえる。

D.1.1　内積を成分表示する

ついで、内積が成分の積で表す事ができることを示そう。

■内積の線形性　まずは準備として、内積演算が線形性を持っている事を示す。演算が線形性を持っているという事は、以下の２つの式が成立する事である。

$$\begin{aligned} < x_1 + x_2, y > &= < x_1, y > + < x_2, y > \\ < \alpha x_1, y > &= \alpha < x_1, y > \end{aligned}$$

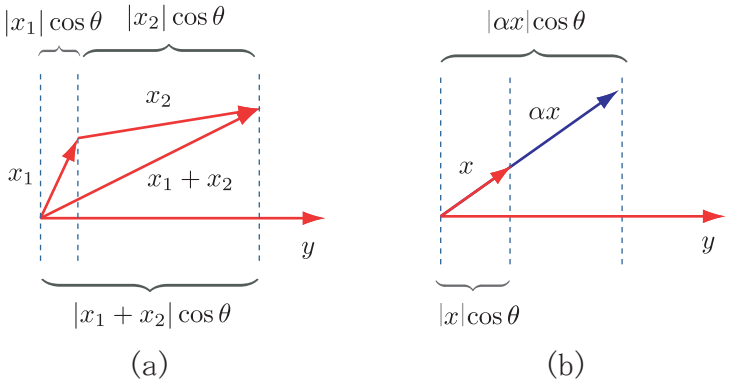


図 73　内積演算は線形演算である

$< x_1 + x_2, y >=< x_1, y > + < x_2, y >$ について確認しよう。図 73 の (a) のように、

$$|x_1 + x_2|\cos\theta = |x_1|\cos\theta + |x_2|\cos\theta$$

が成立する。この両辺に $|y|$ をかけると

$$|x_1 + x_2||y|\cos\theta = |x_1||y|\cos\theta + |x_2||y|\cos\theta$$

つまり

$$< x_1 + x_2, y >=< x_1, y > + < x_2, y >$$

$< \alpha x_1, y >= \alpha < x_1, y >$ についても、図 73 の (b) からすぐに導ける。

■内積の成分表示　内積演算が線形演算である事が確認でき、準備が出来たので内積を成分表示してみよう。いま、以下のような 2 つのベクトル x と y があったとしよう。

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

ここで、同じベクトルの内積は角度 0、つまり $\cos\theta = 1$ なので、$< x, x >= |x||x| = |x|^2$ である。なので

$$|x + y|^2 =< x + y, x + y >$$

さらに内積の線形性を用いて右辺を展開すると

$$\begin{aligned} |x + y|^2 &=< x + y, x + y > \\ &=< x, x > +2 < x, y > + < y, y > \\ &= |x|^2 + 2 < x, y > +|y|^2 \end{aligned}$$

この式を変形しよう。$< x, y >$ を左辺に移項して

$$< x, y >= \frac{1}{2}\left\{|x + y|^2 - |x|^2 - |y|^2\right\}$$

ここで、$|x+y|^2$ を成分表示してやろう。

$$\begin{aligned}|x+y|^2 &= (x_1+y_1)^2+(x_2+y_2)^2+\cdots+(x_n+y_n)^2\\ &= (x_1^2+x_2^2+\cdots+c_n^2)+2(x_1y_1+x_2y_2+\cdots+x_ny_n)+(y_1^2+y_2^2+\cdots+y_n^2)\\ &= |x|^2+2(x_1y_1+x_2y_2+\cdots+x_ny_n)+|y|^2\end{aligned}$$

なので、上記の $< x,y >= \dfrac{1}{2}\left\{|x+y|^2-|x|^2-|y|^2\right\}$ に当てはめると

$$< x,y >= x_1y_1+x_2y_2+\cdots+x_ny_n$$

というように成分同士の積和で表される事になる。さらに、これをベクトルで表せば

$$< x,y >= \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}\begin{pmatrix} y_1\\ y_2\\ \vdots\\ y_n \end{pmatrix} = x^ty$$

D.2 正規直交系

定義 付録 D.1. 正規直交系の定義
$S = \{a_1, a_2, \cdots, a_n\}$ がベクトル空間 R の部分集合で以下の 2 条件を満たすとき、S を R における正規直交系という。

1. S のどのベクトルも長さが *1* である
$$a \in S \Rightarrow |a| = 1$$

2. S の異なるどの 2 つのベクトルも直交する
$$a_1, a_2 \in S \Rightarrow < a_1, a_2 >= 0$$

これをクロネッカーのデルタ[*40]を用いて表すと

$$< a_i, a_j >= \delta_{ij} \qquad (i, j = 1, 2, \cdots, n)$$

と書ける。こうした正規直交系のベクトルはお互いに線形独立でもある。

定理 付録 D.2. 正規直交系のベクトルは互いに線形独立である
0 でない k 個のベクトル $a_1, a_2, \cdots, a_k$ のどの 2 つも直交するならば、$a_1, a_2, \cdots, a_k$ は線形独立である。

この事を確認しよう。線形独立である事を確認するには、**??**ページの定義**??**のように、もしあるスカラー $c_1, c_2, \cdots, c_k$ によって

$$c_1 a_1 + c_2 a_2 + \cdots + c_k a_k = 0$$

と表した時、$c_1 = c_2 = \cdots = c_k = 0$ になる事が示せれば良い。そこで、この両辺と a_i との内積をとると

$$c_1 < a_1, a_i > + \cdots + c_i < a_i, a_i > + \cdots + c_k < a_k, a_i >= 0$$

左辺のうち $< a_i, a_i >$ 以外の項は、これらのベクトルが直交しているので 0 となる。したがってこの式は

$$c_i < a_i, a_i >= 0 \quad (i = 1, \cdots, k)$$

となる。ところが、正規直交系のベクトルは $< a_i, a_i >= |a_i|^2 \neq 0$ なので、$c_i = 0$ でなければならない。この事が全ての i について成り立つので、$c_1 = c_2 = \cdots = c_k = 0$ でなければならない。つまり、これらは線形独立である。

また特に、n 個のベクトルの組 $\{a_1, a_2, \cdots, a_n\}$ が n 次元ベクトル空間 R の基底で、しかも正規直交系をなすならば、それらを**正規直交基底**と呼ぶ。

[*40] クロネッカーのデルタ (Kronecker delta) とは、以下のような関係を表す記号で、いろいろな場面で有用である。例えば、単位行列は $(I = \delta_{ij})$ と書けたり、n次元直交座標の基底ベクトルの内積は、$< e_i, e_i >= \delta_{ij}$ と書ける。

$$\delta_{ij} = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases}$$

定理 付録 D.3. 座標値は各基底ベクトルとの内積で求まる

n 個のベクトルの組 $\{e_1, e_2, \cdots, e_n\}$ が n 次元ベクトル空間 R の正規直交基底であるとすると、ベクトル空間 R の任意のベクトル x は

$$x = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n$$

と表すことができ、その座標値 $\{x_1, x_2, \cdots, x_n\}$ は、ベクトル x と各基底ベクトルとの内積 $x_i =< x, e_i >$ で求める事が出来る。

座標値が、$x_i =< x, e_i >$ で求める事が出来る事を確認しよう。実際に x と e_i の内積を求めると、$< e_i, e_i >= \delta_{ij}$（$i = j$ なら 0、$i \neq j$ なら 1）なので、以下のように、$< e_i, e_i >$ 以外の項はゼロになり、e_i の成分が求められる。

$$\begin{aligned} < x, e_i > &=< x_1 e_1 + x_2 e_2 + \cdots + x_n e_n, e_i > \\ &= x_1 \underbrace{< e_1, e_i >}_{0} + \cdots + x_i \underbrace{< e_i, e_i >}_{1} + \cdots + x_n \underbrace{< e_n, e_i >}_{0} = x_i \end{aligned}$$

つまり、$\{e_1, e_2, \cdots, e_n\}$ が正規直交基底なら、任意のベクトル x の座標値を求めるには、ベクトル x と各基底ベクトルの内積を取ればよい。ちなみに、ベクトル x と各基底ベクトルの内積は $|e_i| = 1$ なので

$$< x, e_i >= |x||e_1| \cos\theta = |x| \cos\theta$$

となり、図 74 のように、各座標系へ下ろした垂線の足の長さを意味している。これをベクトル x を基底ベクトルへ射影した長さといい、基底ベクトルをその長さ倍したもの $< x, e_i > e_i$ を**射影ベクトル**という。

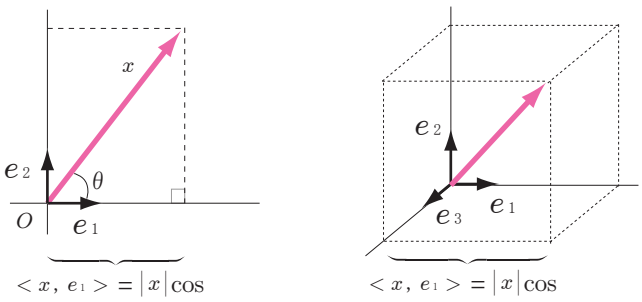


図 74　正規直交基底ベクトルへの射影が座標値

D.3　シュミットの直交化法

n 次元計量ベクトル空間 V は必ず正規直交基底を持つことが出来る。次に、n 個の線形独立なベクトルから正規直交基底を作る方法を示そう。

D.3.1　シュミットの直交化

シュミットの直交化

n 次元計量ベクトル空間 V の n 個の基底 $\{a_1, a_2, \cdots, a_n\}$ に対して、次の式で定まる $\{e_1, e_2, \cdots, e_n\}$ は正規直交基底となる。この方法を**グラム・シュミットの直交化法** (Gram-schmidt orthonormalization) と呼ぶ。

$$
\begin{aligned}
e_1 &= \frac{a_1}{|a_1|} \\
e_2 &= \frac{a'_2}{|a'_2|} \qquad \text{ただし、} \quad a'_2 = a_2 - < e_1, a_2 > e_1 \\
e_3 &= \frac{a'_3}{|a'_3|} \qquad \text{ただし、} \quad a'_3 = a_3 - < e_1, a_3 > e_1 - < e_2, a_3 > e_2 \\
&\cdots \\
e_n &= \frac{a'_n}{|a'_n|} \qquad \text{ただし、} \quad a'_n = a_2 - \sum_{k=1}^{n-1} < e_k, a_n > e_k
\end{aligned}
$$

シュミットの直交化の手順の原理は先に述べたベクトルへの射影である。つまり、図 75 のように、線形独立は 2 つのベクトル a_1 と a_2 をとってきて、a_2 ベクトルを a_1 ベクトルに射影したベクトルを a'_1 とし、次に $a_2 - a'_1$ を求めれば、新たに直交する 2 つのベクトル a'_1 と a'_2 を作ることができる。これを繰り返すのである。

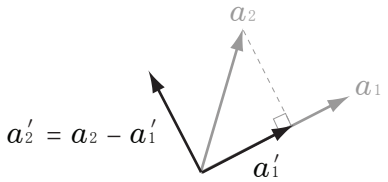
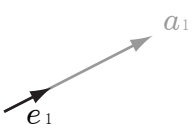


図 75　2 つのベクトル a_1 と a_2 を直交させる

■シュミットの直交化の手順　もう少し詳しく手順を説明しよう。

1. まず a_1 をもってきて、これを長さを 1 に正規化して e_1 とする

$$e_1 = \frac{a_1}{|a_1|}$$



2. 次に、a_2 をもってきて、$a'_2 = a_2 - \alpha e_1$ とおき、この a'_2 が e_1 と直交するように α を定める。

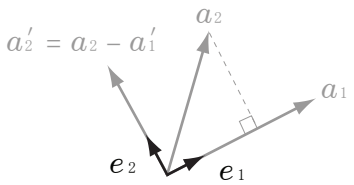
a_2' と e_1 の内積をとると直交するので

$$< e_1, a_2' >=< e_1, a_2 > -\alpha < e_1, e_1 >= 0$$

ここで $< e_1, e_1 >= 1$ なので

$$\alpha =< e_1, a_2 >$$

のように α を定めると、a_2' と e_1 は直交する。これを正規化して e_2 とする。つまり

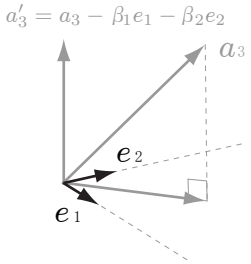


$$e_2 = \frac{a_2'}{|a_2'|} \qquad \text{ただし、} \quad a_2' = a_2 - < e_1, a_2 > e_1$$

3. さらに、a_3 をもってきて、$a_3' = a_3 - \beta_1 e_1 - \beta_2 e_2$ とおいて、e_1 と e_2 とに直交するように a_3' を定める。

   a_3' と e_1 および e_2 との内積をとると

$$< e_1, a_3' >=< e_1, a_3 > -\beta_1 < e_1, e_1 > -\beta_2 < e_1, e_2 >= 0$$
$$< e_2, a_3' >=< e_2, a_3 > -\beta_1 < e_2, e_1 > -\beta_2 < e_2, e_2 >= 0$$



   ここで、$< e_1, e_1 >= 1$、$< e_1, e_2 >=< e_2, e_1 >= 0$ なので

$$\beta_1 =< e_1, a_3 >$$
$$\beta_2 =< e_2, a_3 >$$

   のように β_1、β_2 を定めると、a_3' は e_1 と e_2 とに直交する。

   これを正規化して

$$e_3 = \frac{a_3'}{|a_3'|} \qquad \text{ただし、} \quad a_3' = a_3 - < e_1, a_3 > e_1 - < e_2, a_3 > e_2$$

4. 以下、同様にして、$e_4, \cdots, e_n$ を求めるれば、空間 V の n 個の基底 $\{a_1, a_2, \cdots, a_n\}$ を元に、正規化直交基底 $\{e_1, e_2, \cdots, e_n\}$ を作り出す事ができる。

■具体例　具体的な事例でシュミットの直交化法を確認しよう。

例題 付録 D.1. シュミットの直交化法を用いて、次の線形独立なベクトル a_1、a_2、a_3 から正規直交基底を作れ。

$$a_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad a_2 = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \qquad a_3 = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$$

まずは e_1 を求めよう。$|a_1| = \sqrt{3}$ より、長さを 1 に正規化すると

$$e_1 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

ついで、e_2 を求めよう。

$$a_2' = a_2 - < e_1, a_2 > e_1 = \begin{pmatrix}1\\2\\3\end{pmatrix} - \frac{6}{\sqrt{3}} \cdot \frac{1}{\sqrt{3}} \begin{pmatrix}1\\1\\1\end{pmatrix} = \begin{pmatrix}1-2\\2-2\\3-2\end{pmatrix} = \begin{pmatrix}-1\\0\\1\end{pmatrix}$$

$|a_2'| = \sqrt{2}$ なので長さ 1 に正規化すると

$$e_2 = \frac{1}{\sqrt{2}} \begin{pmatrix}-1\\0\\1\end{pmatrix}$$

最後に、e_3 を求めよう。

$$\begin{aligned} a_3' &= a_3 - < e_1, a_3 > e_1 - < e_2, a_3 > e_2 \\ &= \begin{pmatrix}1\\3\\2\end{pmatrix} - \frac{6}{\sqrt{3}} \cdot \frac{1}{\sqrt{3}} \begin{pmatrix}1\\1\\1\end{pmatrix} - \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} \begin{pmatrix}-1\\0\\1\end{pmatrix} = \begin{pmatrix}1\\3\\2\end{pmatrix} - \begin{pmatrix}2\\2\\2\end{pmatrix} - \frac{1}{2} \begin{pmatrix}-1\\0\\1\end{pmatrix} = \frac{1}{2} \begin{pmatrix}-1\\2\\-1\end{pmatrix} \end{aligned}$$

長さを 1 に正規化するために、まず長さを求めると、

$$|a_3'| = \sqrt{\frac{1}{4}(1+4+1)} = \sqrt{\frac{6}{4}} = \frac{\sqrt{6}}{2}$$

なので、

$$e_3 = \frac{2}{\sqrt{6}} \cdot \frac{1}{2} \begin{pmatrix}-1\\2\\-1\end{pmatrix} = \frac{1}{\sqrt{6}} \begin{pmatrix}-1\\2\\-1\end{pmatrix}$$

以上により、a_1、a_2、a_3 から作った正規直交基底をなすベクトルは

$$e_1 = \frac{1}{\sqrt{3}} \begin{pmatrix}1\\1\\1\end{pmatrix}, \quad e_2 = \frac{1}{\sqrt{2}} \begin{pmatrix}-1\\0\\1\end{pmatrix}, \quad e_3 = \frac{1}{\sqrt{6}} \begin{pmatrix}-1\\2\\-1\end{pmatrix}$$

ちなみに、上では a_1 からシュミットの直交化を施したが、a_2 から行う事もできる。その場合は

$$e_1 = \frac{1}{\sqrt{14}} \begin{pmatrix}1\\2\\3\end{pmatrix}, \quad e_2 = \frac{1}{\sqrt{21}} \begin{pmatrix}4\\1\\-2\end{pmatrix}, \quad e_3 = \frac{1}{\sqrt{6}} \begin{pmatrix}-1\\2\\-1\end{pmatrix}$$

となり、別の正規直交基底になる。つまり、n 次元計量空間には必ず正規直交基底を作る事ができるが、その正規直交基底は 1 つではなく、任意に設定する事ができる。

D.4　直交行列について

列ベクトルが正規直交基底で出来ている n 次元の正方行列 A は直交行列と呼ばれる行列である。この行列の n 個の列ベクトルは全て長さが 1 で、互いに直交するので、

$$\begin{pmatrix} - & a_1 & - \\ - & a_2 & - \\ & \vdots & \\ - & a_1 & - \end{pmatrix} \begin{pmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

つまり、$A^tA = I$ という性質をもっている。この行列による写像が何を意味しているかを見ていこう。

直交変換

正方行列 A が**直交行列**であれば、行列 A による写像には、以下のような特徴がある。

- 2 つのベクトルのなす角度を変えない写像である。
- ベクトルの長さを変えない写像である。
- 行列 A による写像は図形を合同な図形に写像する。

このような直交行列による写像を**直交変換**という。

■直交行列の定義と性質　　まずは直交行列を定義してその性質を整理しておこう。

定義 付録 D.2. 直交行列の定義

n 次正方行列 A が

$$A^tA = I \tag{付録 D.5}$$

を満たすとき、行列 A を**直交行列**という。

この定義より以下の事がいえる。

定義 付録 D.3. 直交行列の性質

A が直交行列なら

$$A^t = A^{-1} \tag{付録 D.6}$$
$$|A| = \pm 1 \tag{付録 D.7}$$
$$AA^t = I \tag{付録 D.8}$$

式付録 D.6 は $A^tA = I$ より明白である。では、式付録 D.7 を確認しよう。まず、定義から $|A^tA| = |I| = 1$。ここで、**??**ページの節**??**で述べたように $|A| = |A^t|$ なので、$|A^tA| = |A|^2 = 1$ となる。なので $|A| = \pm 1$ である。

また、$A^tA = I$ が成立するならば、$AA^t = I$ も成立する。何故ならば、A は逆行列を持つので正則であり $A^tA = I$ の左から A をかけた $AA^tA = A$ に右から A^{-1} をかけると、$AA^t = I$ となるからである。

■直交行列による写像は長さも角度も保存する　次に直交行列による写像を考えよう。

定義 付録 D.4. 直交変換の性質

直交行列 A による写像は、ベクトルの長さを変えない。つまり

$$|Ax| = |x| \tag{付録 D.9}$$

また、任意の 2 つのベクトル x、y のなす角度を変えない。つまり

$$< Ax, Ay >=< x, y > \tag{付録 D.10}$$

まず直交行列 A によってベクトルを写像すると Ax となる。これが x と長さが変わらない事をしめそう。

$$|Ax|^2 =< Ax, Ax >= (Ax)^t(Ax) = x^tA^tAx$$

ここで $A^tA = I$ より

$$|Ax|^2 = x^tx = |x|^2$$

なのでベクトルの長さを変えない。また、逆にベクトルの長さを変えない行列を直交行列と定義する事もできる。つまり、$< Ax, Ax >= x^tA^tAx = x^tx$ が成り立つとして、$x^t(A^tA - I)x = 0$ が任意の x について成り立つ事から、$A^tA = I$ を導いてもよい。

次に、2 つのベクトル x、y を直交行列 A で写像しても、その角度が変わらない事を示そう。ベクトルのなす角度は

$$\cos\theta = \frac{< x, y >}{|x||y|}$$

である。ここで、A による写像はベクトルの長さを変えないので、$|x||y| = |Ax||Ay|$ であり、$< Ax, Ay >=< x, y >$ が示せればよい。これも $A^tA = I$ である事を用いれば

$$< Ax, Ay >= (Ax)^tAy = x^tA^tAy = x^ty =< x, y >$$

となるので、直交行列による写像は、ベクトルの長さも角度も変えないという事が判る。

■直交行列による写像は合同変換である　直交行列による写像は長さも角度も変えない。

また当然ながら $A0 = 0$ なので、原点を動かさない写像である。このようにベクトルの長さ、ベクトル同士のなす角度を変えず、原点も移動しない変換を**合同変換**と呼ぶ。このように図形を合同なまま変換するものには図 76 のように**回転変換**と**鏡映変換**がある。

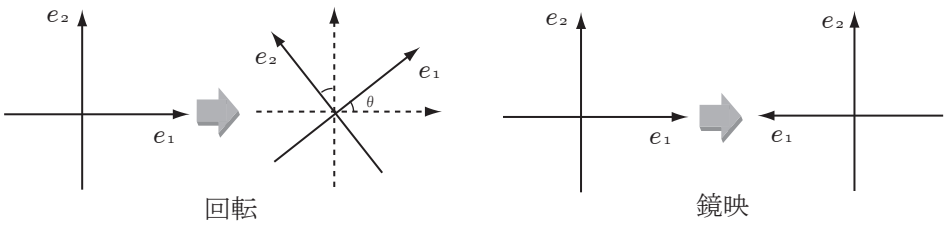


図 76　合同変換には回転と鏡映がある

鏡映変換は表と裏をひっくり返す変換であり、回転のみでは実現できない事がわかるであろう。また回転と鏡映の違いは A の行列式の違いで判る。

$$\begin{array}{ll} \text{回転} & |A| = 1 \\ \text{鏡映} & |A| = -1 \end{array}$$

D.5 シュミットの直交化と QR 分解

$A = (a_1, a_2, \cdots, a_n)$ とするとき、$a_1, a_2, \cdots, a_n$ から、 Gram-Schmidt の直交化を行って正規直交基底 $q_1, q_2, \cdots, q_n$ を作る計算は、行列 A の QR 分解を求めていることになる。

QR 分解

行列 A を正則行列とするとき、直交行列 Q と、上三角行列 R で

$$A = QR$$

満たすものが存在する。特に R の対角成分は正であるように取ることができ、そういうものに限ると分解は一意的である。これを A の QR **分解**と呼ぶ。

$$\left(\begin{array}{c|c|c|c|c} a_1' & a_2' & \cdots & a_n' \end{array}\right) = \left(\begin{array}{c|c|c|c|c} a_1 & a_2 & \cdots & a_n \end{array}\right) \left(\begin{array}{c|c|c|c|c} 1 & \alpha_{12} & \alpha_{13} & \cdots & \alpha_{1n} \\ 0 & 1 & \alpha_{23} & \cdots & \alpha_{2n} \\ 0 & 0 & 1 & \cdots & \alpha_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

$$a_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad a_2 = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \qquad a_3 = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$$

a_1'、a_2'、a_3' を a_1、a_2、a_3 で表してみよう。まず

$$a_1' = a_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \tag{付録 D.11}$$

とおく、ついで a_2' を求めると、$|a_1'| = \sqrt{3}$ であり、$< a_1', a_2 >= 6$ なので

$$a_2' = a_1 - \frac{< a_1', a_2 >}{|a_1'|} \cdot \frac{a_1'}{|a_1'|} = a_2 - 2a_1'$$

$a_1' = a_1$ なので、

$$a_2' = a_2 - 2a_1 = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} - 2 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \tag{付録 D.12}$$

a_3' は $< a_1', a_3 >= 6$、$< a_2', a_3 >= 1$、$|a_2'| = \sqrt{2}$

$$\begin{aligned} a_3' &= a_3 - \frac{< a_1', a_3 >}{|a_1'|} \cdot \frac{a_1'}{|a_1'|} - \frac{< a_2', a_3 >}{|a_2'|} \cdot \frac{a_2'}{|a_2'|} \\ &= a_3 - \frac{6}{\sqrt{3}} \cdot \frac{1}{\sqrt{3}} a_1 - \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} a_2' = a_3 - 2a_1 - \frac{1}{2} a_2' \end{aligned}$$

ここで、$a'_2 = a_2 - 2a_1$ なので

$$\begin{aligned} a'_3 &= a_3 - 2a_1 - \frac{1}{2}(a_2 - 2a_1) = a_3 - \frac{1}{2}a_2 - a_1 \\ &= \begin{pmatrix}1\\3\\2\end{pmatrix} - \frac{1}{2}\begin{pmatrix}1\\2\\3\end{pmatrix} - \begin{pmatrix}1\\1\\1\end{pmatrix} = \begin{pmatrix}-1/2\\1\\-1/2\end{pmatrix} \end{aligned} \qquad \text{(付録 D.13)}$$

この 3 つの式（式付録 D.11〜式付録 D.13）をまとめると、a'_1、a'_2、a'_3 を以下のように a_1、a_2、a_3 で表す事ができる。

$$\begin{aligned} a'_1 &= a_1 \\ a'_2 &= a_2 - 2a_1 \\ a'_3 &= a_3 - \frac{1}{2}a_2 - a_1 \end{aligned}$$

これを行列で表現すると以下のように、元のベクトルを列ベクトルとする行列 A と上三角行列（これを N と表記する）の積になる。

$$\left(\begin{array}{c|c|c} a'_1 & a'_2 & a'_3 \end{array}\right) = \left(\begin{array}{c|c|c} a_1 & a_2 & a_3 \end{array}\right) \begin{pmatrix} 1 & -2 & -1 \\ 0 & 1 & -1/2 \\ 0 & 0 & 1 \end{pmatrix} \qquad \text{これを、} A' = AN \text{ とおく}$$

さらに、q_1、q_2、q_3 を求める為には、a'_1、a'_2、a'_3 をそれぞれの長さで割れば良い。それぞれの長さは、$|a'_1| = \sqrt{3}$、$|a'_2| = \sqrt{2}$、$|a'_3| = \sqrt{6}/2$ なので、

$$\left(\begin{array}{c|c|c} q_1 & q_2 & q_3 \end{array}\right) = \left(\begin{array}{c|c|c} a'_1 & a'_2 & a'_3 \end{array}\right) \begin{pmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{2}{\sqrt{6}} \end{pmatrix} \qquad \text{これを、} Q = A'D \text{ とおく}$$

この 2 つの式、$A' = AN$ と $E = A'D$ を変形して行こう。まず、上三角行列の行列式は対角成分のかけ算であり[*41]、N の対角成分は必ず 1 になるので逆行列は $|A| \neq 0$[*42]。なので逆行列を持つ。その逆行列を N^{-1} とすると $A' = AN$ の両辺に逆行列をかけて

$$A = A'N^{-1}$$

また、対角行列 D の逆行列はそれぞれの対角成分の逆数であり、これを D^{-1} とすると、$E = A'D$ より

$$A' = QD^{-1}$$

この 2 つの式から、

$$A = QD^{-1}N^{-1}$$

となる。この例の値を求めてみよう。まずは Q を求めると、$Q = A'D$ なので

$$Q = \begin{pmatrix} 1 & -1 & -1/2 \\ 1 & 0 & 1 \\ 1 & 1 & -1/2 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{2}{\sqrt{6}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \end{pmatrix}$$

*41 ??ページの式??
*42 ??ページ参照

ついで、

$$D^{-1}N^{-1} = \begin{pmatrix} \sqrt{3} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & \frac{\sqrt{6}}{2} \end{pmatrix} \begin{pmatrix} 1 & 2 & 2 \\ 0 & 1 & \frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \sqrt{3} & 2\sqrt{3} & 2\sqrt{3} \\ 0 & \sqrt{2} & \frac{\sqrt{2}}{2} \\ 0 & 0 & \frac{\sqrt{6}}{2} \end{pmatrix}$$

この $D^{-1}N^{-1}$ を R とおけば、

$$A = QR = \begin{pmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \end{pmatrix} \begin{pmatrix} \sqrt{3} & 2\sqrt{3} & 2\sqrt{3} \\ 0 & \sqrt{2} & \frac{\sqrt{2}}{2} \\ 0 & 0 & \frac{\sqrt{6}}{2} \end{pmatrix}$$

というように、行列 A を直交ベクトル Q と上三角行列 R の積に分解できる。

付録 E　固有値と固有ベクトル

行列 A による一次変換 $y = Ax$ をベクトル x からベクトル y への写像として捉えた場合、行列 A の固有ベクトルを考える事で以下のような事がわかり、行列 A による写像の性質を考察する上で見通しが良くなる。

固有値・固有ベクトルの働き

固有ベクトルとは方向の変わらないベクトルである　式 $y = Ax$ を、行列 A によってベクトル x がベクトル y に写像されたと考える。その時、固有ベクトルとは写像 A によって方向が変わらないベクトルの事である

固有ベクトルを座標軸にとれば作用が簡単にイメージできる　あるベクトルに行列 A を作用させた結果は、その点を各固有ベクトルの方向に分解し、それぞれの成分を固有値倍して、合成した点に移される事になる。

E.0.1　固有ベクトルとは方向の変わらないベクトルである

$y = Ax$ という一次変換を考えるた時、行列 A が xy 平面上の点をどこに移すかを考えよう。

行列 A を $A = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}$ とすると、図 77 のように、黒点が赤点に移動する。

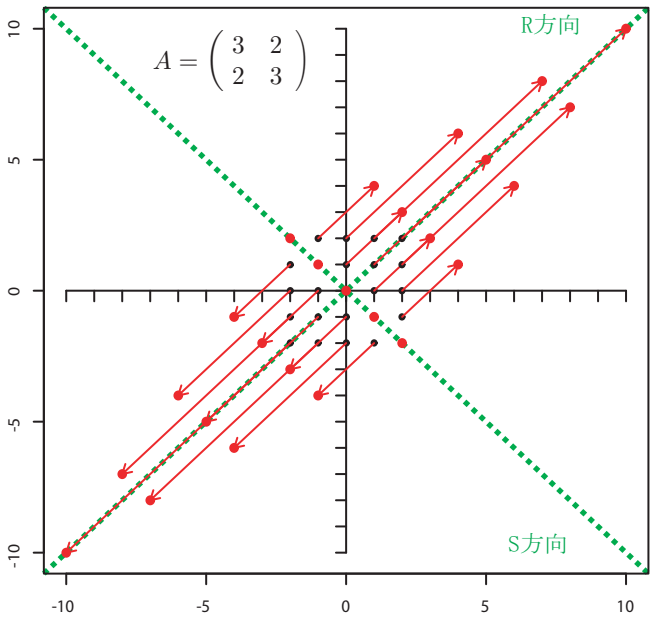


図 77　行列 A によってそれぞれの点が何処に移るか

図 77 の± 45° に引いた二つの緑の点線（R 方向、S 方向）上の黒点に注目して欲しい。この線上の点は原点から± 45° の方向の線上に伸び縮しているだけである。つまり、R 方向（ベクトル $(1, 1)$ の整数倍）と S 方

向（ベクトル $(-1,1)$ の整数倍）は、その方向を変えずに、それぞれ５倍と１倍されている。

このように行列 A による変換によって、

1. 方向が変わらないベクトルを**固有ベクトル**という
2. その伸び縮の倍率を**固有値**という

つまり、ある x をベクトルとし、行列 A でベクトル x を変換した結果がベクトル x の定数倍で表せるという事であり、以下のように定義される。

定義 付録 E.1.　【固有値と固有ベクトル】
n 次の正方行列 A に対して、以下の式が成立するような定数 λ とベクトル x が存在するとき，λ を行列 A の固有値、x を λ に対する固有ベクトルと言う。

$$Ax = \lambda x \tag{付録 E.1}$$

E.0.2　固有空間で表すと行列の作用が簡単にイメージできる

このような固有ベクトルを座標軸にする[*43]と、一見複雑に見えてる行列 A による写像の作用を、非常に簡単に表す事ができる。実際に固有ベクトルを座標軸にとって新しい座標軸で表してみよう。図 78 は、さきの行列 A によって、点 $x=(1,2)$ が点 $y=(7,8)$ に移っている様子である。

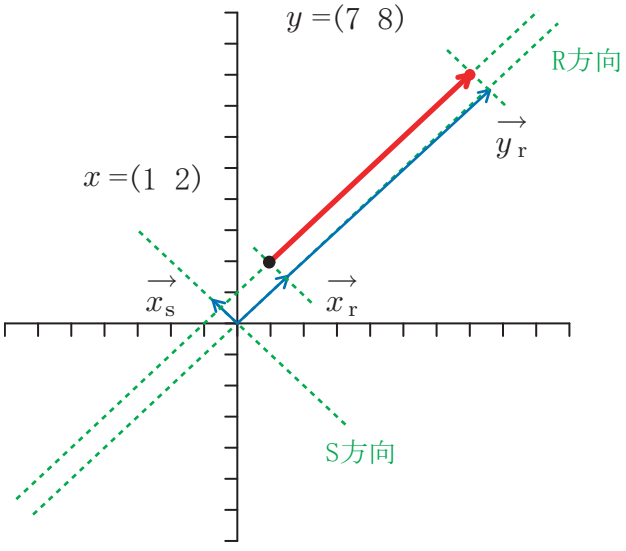


図 78　固有ベクトルの方向に分解して合成する

図 78 のように、点 x を R 方向、S 方向という固有ベクトルの方向に分解する。つまり $\overrightarrow{x} = \overrightarrow{x_r} + \overrightarrow{x_s}$ とす

[*43] ある一時独立なベクトル群 $\boldsymbol{B}$ を座標軸とする事を「$\boldsymbol{B}$ を基底とする」と表現する場合がある

る。そうすると、x の像 y は、$\overrightarrow{x_r}$ の 5 倍と $\overrightarrow{x_s}$ の 1 倍を加えたものになる。つまり、$\overrightarrow{y} = 5\overrightarrow{x_r} + \overrightarrow{x_s}$ というように簡単になるのである。

あるベクトルに行列 A を作用させた結果は、その点を各固有ベクトルの方向に分解し、それぞれの成分を固有値倍して、合成した点に移される事になる。

と言うことは、この固有ベクトルを座標軸にとってあげれば、一見複雑に見える行列による作用を簡単に表現することが出来るはずである。次の節では、その事を調べてみよう。

E.1 固有ベクトルを基底にした世界でベクトル・行列を表現する

固有ベクトルを基底にしてベクトル・行列を表すと簡単に表現できる

1. 固有ベクトルを座標軸とする表現をすれば、ベクトル x は、$x' = P^{-1}x$ と表せる。
2. 固有ベクトルを座標軸とする表現をすれば、A という写像は、$P^{-1}AP$ と表現できる。
3. さらに $P^{-1}AP$ は以下のように簡略化して表現できる。

$$P^{-1}AP = \Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

4. このように行列を簡単な対角行列に変換する事を対角化と呼ぶ。

E.1.1 固有ベクトルを基底にしてベクトルを表現する

では、具体的にある点 x を行列 A の固有ベクトルを基底とした新しい座標軸で表すとするとどのように表現されるかを調べよう。まず、行列 A を n 次の正方行列とする。そして、任意のベクトル x がある基底の元で

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

と表されているとする。さらに、行列 A の固有ベクトルを $p_1, \cdots, p_n$ とし、それらをまとめて

$$P = \begin{pmatrix} | & & | \\ p_1 & \cdots & p_n \\ | & & | \end{pmatrix}$$

と書くことにする[*44]。

さて、固有ベクトルを基底としたときにベクトル $x = (x_1, \cdots, x_n)'$ がどのように表現されるかを考えよう。当然、同じベクトルでも基底を変えると表現が変わる。固有ベクトルを基底とするとベクトル x が

[*44] この固有ベクトルを横に並べた行列 P をモードマトリクスと呼ぶ事がある。

$x' = (x'_1, \cdots, x'_n)'$ というように表されたとしよう。これを先の行列 P を用いて書くと以下のようになる。

$$Px' = x'_1 \begin{pmatrix} p_{11} \\ \vdots \\ p'_{1n} \end{pmatrix} + x'_2 \begin{pmatrix} p_{21} \\ \vdots \\ p'_{2n} \end{pmatrix} + \cdots + x_n \begin{pmatrix} p_{n1} \\ \vdots \\ p'_{n2} \end{pmatrix}$$

$$= \begin{pmatrix} | & & | \\ p_1 & \cdots & p_n \\ | & & | \end{pmatrix} \begin{pmatrix} x'_1 \\ \vdots \\ x'_n \end{pmatrix}$$

この Px' が x と同じものなので

$$x = Px'$$

後で述べるが、相異なる固有値に対応する固有ベクトルはお互いに線形独立である事がわかっている。なので、この行列 P は逆行列を持つので

$$x' = P^{-1}x$$

今度は、$y = Ax$ の y を考えよう。これもまったく同様に

$$Py' = \begin{pmatrix} | & & | \\ p_1 & \cdots & p_n \\ | & & | \end{pmatrix} \begin{pmatrix} y'_1 \\ \vdots \\ y'_n \end{pmatrix}$$

というように表現でき、

$$y = Py'$$
$$y' = P^{-1}y$$

となる。

E.1.2 固有ベクトルを基底にして行列を表現する

では次に、行列 A そのものを新しく A の固有ベクトルを基底軸として表すとどのような表現になるかを調べよう。元の基底軸上での $y = Ax$ という変換があるとして、これを新しい基底軸上で表現すればよいのである。

まず、x と y を新しい基底軸での表現でかくと、$x = Px'$ であり、$y = Py'$ である。なので、$y = Ax$ に、それぞれ $x = Px'$ と $y = Py'$ を代入すると

$$Py' = APx'$$

この両辺に左から P^{-1} をかけると、

$$y' = P^{-1}APx \qquad \text{(付録 E.2)}$$

となる。

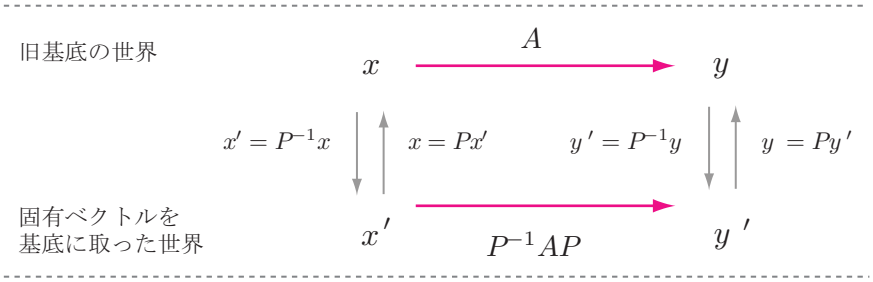


図 79　固有ベクトルを基底に取った世界でのベクトルと行列の表現

以上をまとめると、図 79 のように

1. 固有ベクトルを座標軸とする表現をすれば、ベクトル x は、$x' = P^{-1}x$ と表せる。
2. 固有ベクトルを座標軸とする表現をすれば、A という写像は、$P^{-1}AP$ と表現できる。

実は、この $P^{-1}AP$ という写像はもっと簡単に表す事ができる。そのことを示してみよう。行列 A の n 個の相異なる固有値を $\lambda_1, \cdots, \lambda_n$ とし、それらに対応する固有ベクトルを $p_1, \cdots, p_n$ としよう。ここで、固有値と固有ベクトルは、

$$Ap_i = \lambda_i p_i \ (i = 1, \cdots, n)$$

と表すことができる。これらをまとめて行列表現するために、固有ベクトルをまとめた行列を P、対応する固有値を対角行列に持つ行列 Δ とする。これら用いると、$i = 1, \cdots, n$ の n 個の関係をまとめて

$$A \begin{pmatrix} | & & | \\ p_1 & \cdots & p_n \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ p_1 & \cdots & p_n \\ | & & | \end{pmatrix} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

と表す事ができる[*45]。つまり

$$AP = P\Delta$$

のように表す事ができる。ここで、行列 P は逆行列をもつので、両辺に P^{-1} をかけると

$$P^{-1}AP = \Delta$$

となる。つまり、

固有ベクトルを基底軸にとれば行列 A は以下のように簡略化して表現できる。

$$P^{-1}AP = \Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \qquad \text{(付録 E.3)}$$

[*45] 列を定数倍するときは対角行列を右から、逆に行を定数倍する時は対角行列を左からかければよい。

このように行列を簡単な対角行列に変換する事を**対角化**と呼んでいる。

E.2 さて、一体なにがうれしいの？

これは一体何がうれしいのか？

色んな計算が見通し良くなる

1. 行列 A の階乗の計算が以下のように見通しよくなる。
2. 相関行列の固有ベクトルを座標軸にすれば相関行列が Λ と簡単にかける。

■計算が見通しよくなる　まず、左から P、右から P^{-1} をかけてやって

$$A = P\Lambda P^{-1} \tag{付録 E.4}$$

たとえば、A^n を考えてみると以下のように、隣り合う $P^{-1}P$ が相殺し合って単位行列 I となり、簡単な式 $P\Lambda^n P^{-1}$ で表せる事がわかる。

$$A^n = (P\Lambda \underbrace{P^{-1})(P}_{I}\Lambda \underbrace{P^{-1})\cdots(P}_{I\cdots}\Lambda \underbrace{P^{-1})(P}_{I}\Lambda P^{-1}) = P\Lambda^n P^{-1} \tag{付録 E.5}$$

■相関行列自体が簡単にかける　また、よくデータ解析でててくる相関行列 R_X も、固有ベクトルを座標軸にすると Λ と簡単にかける。

まず相関行列の固有ベクトルを並べたモードマトリクス P を作ってやり、それを新しい座標軸にする。そうすると、X というデータ行列は、あたらしい座標軸でのデータ行列 $X' = XP$ というように計算できる。

なので、

$$R_{\acute{X}} = \acute{X}^t\acute{X} = (XP)^t XP = P^t X^t XP = P^t R_X P$$

ここで、もともと P が相関行列の固有値であるから、$R_X P = P\Lambda$ なので、

$$P^t R_X P = P^t P\Lambda$$
$$R_{\acute{X}} = \Lambda$$

相関行列を簡単にかけると何が嬉しいのか？

例えば、点 x と点 y の距離は、元の座標軸では $(x-y)^t R(x-y)$ であるが、あたらしい座標軸では $(x'-y')^t\Lambda(x'-y')$ と簡単になる。

E.3　固有値と固有ベクトルの求め方

ここでは、正方行列 A を考える事にしよう。固有値と固有ベクトルの関係式は、式**??**より

$$Ax = \lambda x$$

であった。この式は

$$(A - \lambda I)x = 0$$

と表す事ができる。この式から λ を求めるには、この式が $x \neq 0$ の解を持つ必要がある。つまり、

$$|A - \lambda I| = 0 \tag{付録 E.6}$$

が成立しなければならない。何故ならば、もし $|A - \lambda I| \neq 0$ ならば、$(A - \lambda I)$ には逆行列が存在し、$x = 0$ が解として一意に定まってしまい、$Ax = \lambda x$ の固有値 λ が定まらないからである。この式付録 E.6 を**固有方程式**という。この固有方程式を解くことで、固有値と固有ベクトルを計算する事ができる。

■固有値と固有ベクトルを求める　　では、実際に先の行列 A について、固有方程式を解いて固有値と固有ベクトルを求めてみよう。

$$A = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}$$

なので

$$\begin{aligned} |A - \lambda I| &= \begin{vmatrix} 3-\lambda & 2 \\ 2 & 3-\lambda \end{vmatrix} = (3-\lambda)^2 - 4 \\ &= \lambda^2 - 6\lambda + 5 = (\lambda - 5)(\lambda - 1) = 0 \end{aligned}$$

となり $\lambda = 5, 1$ が求まる。そして、$Ax = \lambda x$ に代入してそれぞれの固有ベクトルを求めればよい。

$\lambda = 5$ の固有ベクトルを求める　$Ax = \lambda x$ に代入して $Ax = 5x$。つまり

$$\begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 5 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

具体的に展開すると

$$\begin{cases} x_1 - x_2 = 0 \\ x_1 - x_2 = 0 \end{cases}$$　なので、固有ベクトルは　$p_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

$\lambda = 1$ の固有ベクトルを求める　$Ax = \lambda x$ に代入して $Ax = x$。つまり

$$\begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

具体的に展開すると

$$\begin{cases} x_1 + x_2 = 0 \\ x_1 + x_2 = 0 \end{cases}$$　なので、固有ベクトルは　$p_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$