Bernoulli 分布

Bernoulli (ベルヌーイ) 分布は、「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試行 (Bernoulli 試行) の結果を 0 と 1 で表した分布です。式で表すと、次のようになります。

$$
\begin{aligned}
P(X=1) &= \phi \\
P(X=0) &= 1-\phi \\
P(X=x) &= \phi^x(1-\phi)^{1-x}
\end{aligned}
$$

以下では、確率分布の基本的なことから上の式を説明していきます。また、Bernoulli 分布の期待値や分散などの計算も行なっていきます。

♣ 確率分布とは

例として、「サイコロ 2 個振ったときの出た目の和 X」を考えましょう。この X は、どのような値となるかが確率法則によって決まるような変数で、一般に**確率変数**と呼ばれます。この確率変数 X に対する確率分布 (確率変数がとる値とその値をとる確率の対応の様子) は次の表のようになります。

X	2	3	4	5	6	7	8	9	10	11	12
確率	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

サイコロ 2 個の目の出かたの場合の数は、全部で $6 \times 6 = 36$ 通りです。これより、

$$
\begin{aligned}
X=2:\ (1,1) &\quad \to \frac{1}{36} \\
X=3:\ (1,2),(2,1) &\quad \to \frac{2}{36} \\
X=4:\ (1,3),(2,2),(3,1) &\quad \to \frac{3}{36} \\
&\ \vdots
\end{aligned}
$$

といった要領で、X の取りうる値、それぞれに対する確率を求めることができます。「確率変数 X が $X=2$ となる確率は $\frac{1}{36}$ である」ということを、数式では次のように表現します。

$$
P(X=2) = \frac{1}{36}
$$

P は、確率 (probability) の頭文字となっています。確率分布で最も重要なことは、確率分布では全体の合計が 1 となるということです。

$$P(X=2)+P(X=3)+\cdots+P(X=12)=\sum_{k=2}^{12}P(X=k)=1$$

$P(X=2)$ から $P(X=12)$ までの和を、総和記号 $\sum$ を用いて表現しました[26]。

♣ Bernoulli 分布とは

Bernoulli 分布は、上で説明した確率分布の 1 つの例です。先ほどは「サイコロ 2 個振ったときの出た目の和」という結果が複数あるような試行 (起こりうる結果がいくつかあり、どれが起こるかは偶然で決まる手順のこと) を考えました。ここでは、起こりうる結果が 2 つしかない、Bernoulli 試行と呼ばれる試行を考えます。例えば、コインを投げるという試行は、結果が表が出るか裏が出るかの 2 つなので、Bernoulli 試行となっています。コインを投げた結果、表が出ることを $X=1$、裏が出ることを $X=0$ と表しましょう。表が出る確率を ϕ (ファイと読みます) とすると[27]、数式では

$$P(X=1)=\phi$$

となります。ϕ は確率を表すので、必ず 0 以上 1 以下の数です。このことを数学の記号で

$$\phi\in[0,1]$$

と表します。ϕ が 0 以上 1 以下の領域 ($[0,1]$) に含まれている ($\in$) という意味です。確率分布では全体の合計 (今の場合は $X=0$ と $X=1$ の 2 つの合計) は 1 となるので、

$$P(X=1)+P(X=0)=1$$

です。これより、

$$P(X=0)=1-\phi$$

と分かります。この $P(X=1)$ と $P(X=0)$ の式を 1 つにまとめると、次のように表すことができます[28]。

$$P(X=x)=\phi^x(1-\phi)^{1-x}$$

実際に $x=1$、$x=0$ を入れてみれば、上の 2 つの式が再現できることが確認できます。

$$P(X=1)=\phi^1(1-\phi)^{1-1}=\phi$$
$$P(X=0)=\phi^0(1-\phi)^{1-0}=1-\phi$$

[26] 総和記号については、softmax 関数のところで詳しく説明しています。
[27] 完璧なコインなら $\phi=0.5$ ですが、コインが歪んでいたりすると 0.5 からズレるでしょう。
[28] この $P(X=x)$ のことを「確率質量関数 (probability mass function)」と言います。

0 乗は 1 となることに注意して下さい[29]。以上で Bernoulli 分布の説明はおしまいです。最初式を見た時は難しそうに見えたかもしれませんが、実は非常にシンプルな分布なのです。

♣ 期待値と分散

確率分布を特徴づける量として、**「期待値 (expected value)」** と **「分散 (variance)」** があります。それぞれがどのような量か、確認しておきましょう。

期待値とは「確率変数が取る値を、確率によって重み付けした平均値」です。例えば、300 円の宝くじ 1 枚の期待値が 100 円であった場合、その宝くじには 100 円の価値が期待できるということです。当たったくじ、外れたくじの総合で見て、平均すると 1 枚あたり 100 円の価値であったということです。期待値という名前は、確率変数が取ると「期待」される値であることから名付けられています。数式では、期待値は次のように定義されます。

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

確率変数 X の取り得る全ての値 k について、その値とその値が出る確率を掛け合わせたものを足し合わせる、ということです[30]。例えば、最初に考えた「サイコロ 2 個振ったときの出た目の和 X」の期待値を計算しましょう。上の式をこの例の場合に適用すると、

$$\begin{aligned} E[X] &= \sum_{k=2}^{12} kP(X = k) \\ &= 2 \times P(X = 2) \ + \ 3 \times P(X = 3) \ + \ \cdots \ + \ 12 \times P(X = 12) \\ &= 2 \times \frac{1}{36} + 3 \times \frac{2}{36} + \cdots + 12 \times \frac{1}{36} = 7 \end{aligned}$$

となります。つまり、サイコロ 2 個振ったときの出た目の和の期待値は 7 と分かります。

次に分散を確認します。分散とは「確率変数のばらつき具合を表すための指標」です。分散が大きいほどばらつきが大きく、逆に分散が小さいほど確率変数の散らばりが少なく平均の近くにあることになります。確率変数 X の分散は、数式では次のように表現されます。

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

[29] softmax 関数のところで説明しています。
[30] 上の式では添え字 k の動く範囲が一般には異なるので、特に範囲は明記していません。

つまり、確率変数 X の期待値 $E[X]$ からのズレの 2 乗の期待値です。この式では分かりにくいので、総和記号を使って表すと、次のようになります。

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

これは次のように計算を進めることができます。

$$\begin{aligned}\mathrm{Var}[X] &= \sum_k \left(k^2 - 2E[X]k + E[X]^2\right) P(X = k) \\ &= \sum_k k^2 P(X = k) - 2E[X] \sum_k kP(X = k) + E[x]^2 \sum_k P(X = k)\end{aligned}$$

ここで、1 つ目の項は期待値の定義より、

$$\sum_k k^2 P(X = k) = E[X^2]$$

つまり、確率変数 X の 2 乗の期待値となります。2 つ目の項も期待値の定義より、

$$2E[X] \sum_k kP(X = k) = 2E[X] \times E[X] = 2E[X]^2$$

最後の項は、確率の合計が 1 であることを使うと、

$$E[X]^2 \sum_k P(X = k) = E[X]^2 \times 1 = E[X]^2$$

これより、分散は次のようになることが分かります。

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

つまり、2 乗の期待値から期待値の 2 乗を引いた量となっています。分散を計算する際には、最初の定義の式で計算しても良いですし、この 2 乗の期待値から期待値の 2 乗を引く方法で計算しても同じ結果となります。状況に応じて、計算が簡単な方で計算を行えば良いでしょう。

♣ Bernoulli 分布の期待値と分散

それでは、最後に Bernoulli 分布の期待値と分散を計算してみましょう。Bernoulli 分布は

$$P(X = 1) = \phi$$
$$P(X = 0) = 1 - \phi$$

でした。1 か 0 しかないので、計算はとても簡単です。先ず期待値は、前のセクションで説明した定義に従って計算すると、

$$E[X] = \sum_{k=0}^{1} kP(X = k) = 0 \times P(X = 0) + 1 \times P(X = 1) = \phi$$

となり、Bernoulli 分布の期待値は ϕ であることが分かります。次に分散を計算します。ここでは、2 乗の期待値から期待値の 2 乗を引く方法で分散を計算しましょう。2 乗の期待値は、

$$E[X^2] = \sum_{k=0}^{1} k^2 P(X = k) = 0^2 \times P(X = 0) + 1^2 \times P(X = 1) = \phi$$

となり、2 乗の期待値も ϕ であることが分かります。これより、

$$\mathrm{Var}[X] = E[X^2] - E[X]^2 = \phi - \phi^2 = \phi(1 - \phi)$$

となり、Bernoulli 分布の分散は $\phi(1 - \phi)$ であることが分かります。

Gauss 分布

一般的な Gauss (ガウス) 分布は、次の式で与えられます。

$$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\beta}^{-1}) = \sqrt{\frac{\det(\boldsymbol{\beta})}{(2\pi)^n}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\beta}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

この分布は**「多変量 Gauss 分布」**や**「多変量正規分布」**とも呼ばれます。複雑な上記の式の意味を理解することが、この解説プリントの目標です。

♣ 1 変数の Gauss 分布

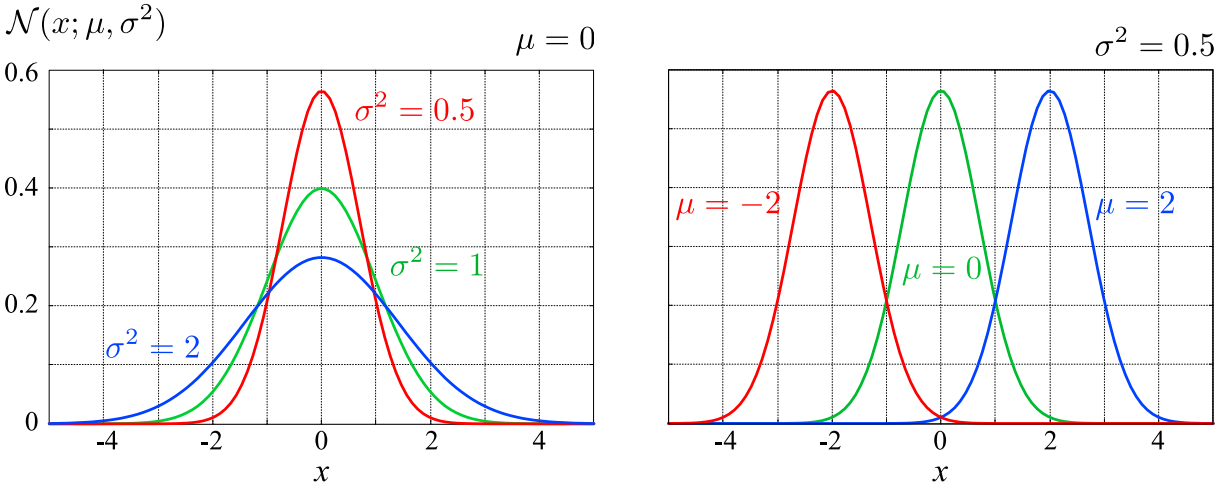
最初に書いた一般的な Gauss 分布は、複数の変数 (多変数) に対する確率分布を表すものです。ここでは先ず、変数が 1 つの場合を考えましょう。これを基本に、多変数の場合に拡張していくことで、最初の式の意味を理解することができます。Gauss 分布 (正規分布ともいいます) は、次の式で与えられる確率分布です。

$$\mathcal{N}(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

exp というのは、底がネイピア数の指数関数のことです。つまり、

$$\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) = e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

です。特に指数が複雑な場合には、e^A よりも $\exp(A)$ という書き方の方が見やすいため、この書き方がよく使われます。x は確率変数です。Bernoulli 分布の場合には、確率変数は 0 か 1 のみを取る「離散型」の変数でした[31]。この Gauss 分布では連続的にあらゆる値を取る「連続型」の変数を考えるため、Gauss 分布も次のように連続的なグラフとなります。



[31] 離散というのは、とびとび、という意味です。

σ^2 と μ (ミュー) の値を変化させると分布の様子が変化することが分かります。このように σ^2 と μ は分布を特徴づける量となっており[32]、σ^2 は「分散」、μ は「平均」と呼ばれています。分散 σ^2 を大きくすると、曲線の幅が広がり、確率変数の取り得る値のばらつきが大きくなることが分かります。つまり、σ^2 は確率変数の取り得る値のばらつきをコントロールするパラメータです。一方で平均 μ を変化させると、曲線の中心の位置がずれることが分かります。Gauss 分布は、この平均 μ を中心に左右対称な曲線となります。

以上で 1 変数の Gauss 分布の基本的な事項はおしまいですが、最後にいくつか細かなコメントをしておきます。先ず、Gauss 分布の表し方

$$\mathcal{N}(x; \mu, \sigma^2)$$

の表記法についてです。$\mathcal{N}$ は正規分布 (normal distribution) の頭文字です。そして連続変数 x に対する確率分布なので、$\mathcal{N}(x)$ と表します。ただ、Gauss 分布は 2 つのパラメータ μ と σ^2 によって分布が変化します。そのため、これらのパラメータの値が与えられた上での、x に対する確率分布という意味で $\mathcal{N}(x; \mu, \sigma^2)$ と表しています。もう 1 つのコメントは、

$$\mathcal{N}(x; \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

についている $\sqrt{\dfrac{1}{2\pi\sigma^2}}$ の部分についてです。この部分は定数なので、単に分布全体の大きさを変化させるだけで、分布の分散や平均を変化させることはない、いわばどうでもよい部分です。なのに何やら複雑な係数となっているのは、$\mathcal{N}(x; \mu, \sigma^2)$ が確率分布の基本ルール「確率分布では全体の合計が 1 となる」を守るようにするためです。少しややこしいのは、Gauss 分布の場合には確率変数が連続なので、単純な足し算で「全体の合計」を計算することはできない点です。このような連続的な変数の場合の全体の合計の計算には、積分と呼ばれる特殊な “足し算” を使います。分からない場合には無視してもらって大丈夫ですが、数式で表すと、

$$\int_{-\infty}^{\infty} \mathcal{N}(x; \mu, \sigma^2)\, dx = 1$$

となるようになっています。とにかく、$\mathcal{N}(x; \mu, \sigma^2)$ が確率分布の基本ルールを守るようにするために、なにやら複雑な係数がついているのだと思っておけば十分です[33]。

[32] このような変数をパラメータ (parameter) と呼びます。
[33] このような係数のことを、一般に「規格化係数」と言います。

♣ 1 変数から多変数へ

一般に何か測定したとすると、その測定結果は様々な要素に依存しているでしょう。こういった様々な要素をまとめて扱うためには、1 変数の確率分布のみでは不十分で、多変数の確率分布を考える必要が出てきます。ここでは、前のセクションで説明した 1 変数の Gauss 分布を多変数の場合に拡張するための準備を行なっていきます。

ベクトル

n 個の確率変数 $x_i\ (i = 1, \cdots, n)$ があったとします。このように多くの変数を扱う際には、**「ベクトル (vector)」**と呼ばれる量を導入すると便利です[34]。

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

n 個の変数 x_1 から x_n を縦に並べて、それをひとつの「グループ」とみなしています。このベクトルは、プログラミング言語でいうところの「配列」です。n 個の変数を全てバラバラの変数名で扱うと大変なので、配列を導入して n 個の変数を 1 つのグループとして扱えるようにしたのと同じことです。このベクトル $\boldsymbol{x}$ に対応する量で、次のような量を導入しておきましょう[35]。

$$\boldsymbol{x}^{\mathrm{T}} = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}$$

$\boldsymbol{x}$ は縦に変数を並んでいるため「縦ベクトル」、$\boldsymbol{x}^{\mathrm{T}}$ では変数が横に並んでいるため「横ベクトル」と呼ぶことにしましょう[36]。

ベクトルは、複数の数を 1 つのグループにまとめた量なので、通常の数とは計算のルールが異なります。例として、2 つのベクトルを考えます。

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

この 2 つのベクトルの足し算と引き算は、次のようになります。

---

[34] ここではベクトルは太字で表すことにしますが、$\vec{x}$ と書く場合もあります。
[35] T は転置 (transpose) という操作を表す記号ですが、あまり気にしなくて大丈夫です
[36] 本当は、列ベクトル、行ベクトルと呼ばれますが、行と列がややこしいので横と縦にしています。

$$\boldsymbol{x}+\boldsymbol{y}=\begin{pmatrix}x_1\\x_2\\\vdots\\x_n\end{pmatrix}+\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix}=\begin{pmatrix}x_1+y_1\\x_2+y_2\\\vdots\\x_n+y_n\end{pmatrix}$$

$$\boldsymbol{x}-\boldsymbol{y}=\begin{pmatrix}x_1\\x_2\\\vdots\\x_n\end{pmatrix}-\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix}=\begin{pmatrix}x_1-y_1\\x_2-y_2\\\vdots\\x_n-y_n\end{pmatrix}$$

これがベクトルの足し算と引き算のルールです。同じ位置の変数を足せば (引けば) よいだけです。次に掛け算についてです。ベクトルの場合、実は普通の数とは違って掛け算といっても色々な種類の掛け算があるのですが[37]、ここでは多変量 Gauss 分布を理解するのに必要な掛け算のみを確認しましょう (単に掛け算といった場合には、通常この計算を指します)。

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}=\begin{pmatrix}x_1&x_2&\cdots&x_n\end{pmatrix}\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix}=x_1y_1+x_2y_2+\cdots+x_ny_n$$

いくつか注意点を述べます。先ず、このベクトルの計算は、横ベクトルと縦ベクトルの積のみに適用できる計算です。ですから、

$$\boldsymbol{x}\boldsymbol{y}=\begin{pmatrix}x_1\\x_2\\\vdots\\x_n\end{pmatrix}\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix},\qquad\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}^{\mathrm{T}}=\begin{pmatrix}x_1&x_2&\cdots&x_n\end{pmatrix}\begin{pmatrix}y_1&y_2&\cdots&y_n\end{pmatrix}$$

といった掛け算はありません[38]。また、横ベクトルと縦ベクトルの順番にも気をつけて下さい。

$$\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}=\begin{pmatrix}x_1\\x_2\\\vdots\\x_n\end{pmatrix}\begin{pmatrix}y_1&y_2&\cdots&y_n\end{pmatrix}$$

みたいな計算も今回は扱いません[39]。普通の数と違って、掛け算の順番も大事なのがベクトル

[37]内積、外積、テンソル積など色々あります。
[38]内積 $\boldsymbol{x}\cdot\boldsymbol{y}$ や外積 $\boldsymbol{x}\times\boldsymbol{y}$ みたいな計算はありますが、ここでは気にしなくて大丈夫です。
[39]テンソル積と呼ばれる掛け算なのですが、気にしなくてよいです。

の計算の大きな特徴です。もう1つ重要な注意点は、ベクトルの掛け算 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}$ の結果は、ベクトルではないという点です。

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = x_1y_1 + x_2y_2 + \cdots + x_ny_n$$

において、$x_1y_1 + x_2y_2 + \cdots + x_ny_n$ は n 個の数を足した単なる数です。このような数を、ベクトルと区別するために**「スカラー (scalar)」**と呼びます。別にスカラーは何も新しい数ではなく、$2+3=5$ のような普通の計算はスカラーの足し算に他なりません。ですから、ベクトル同士の掛け算の結果は、普通の数になるのだと思っておけば十分です。

行列

ベクトルをさらに発展させたものとして**「行列 (matrix)」**と呼ばれる量があります。これは例えば次のような量です[40]。

$$\boldsymbol{M} = \begin{pmatrix} M_{11} & M_{12} & \cdots & M_{1n} \\ M_{21} & M_{22} & \cdots & M_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ M_{n1} & M_{n2} & \cdots & M_{nn} \end{pmatrix}$$

上のように複数の数を四角状に並べたものです。なぜこんな量を考えるのでしょうか？ 行列のありがたみを感じるために、例として、3人の英語、数学、国語の点数が下の表のようになっている場合を考えましょう。

	英語	数学	国語
A さん	80	90	70
B さん	55	100	65
C さん	85	60	95

行列を使うと、このデータを次のように1つの文字 $\boldsymbol{M}$ で表すことができるのです。

$$\boldsymbol{M} = \begin{pmatrix} M_{\text{A, 英語}} & M_{\text{A, 数学}} & M_{\text{A, 国語}} \\ M_{\text{B, 英語}} & M_{\text{B, 数学}} & M_{\text{B, 国語}} \\ M_{\text{C, 英語}} & M_{\text{C, 数学}} & M_{\text{C, 国語}} \end{pmatrix} = \begin{pmatrix} 80 & 90 & 70 \\ 55 & 100 & 65 \\ 85 & 60 & 95 \end{pmatrix}$$

このように、複数のデータを1つにまとめにして扱えるため、多変数のデータを扱う場合にこの行列という量が大活躍するのです。

[40] 下の行列は特に、$n \times n$ 行列と呼ばれます。大きさ (サイズ) は $n \times n$ のように表されます。

行列に関しても様々な計算のルールがあります。ここでは多変量 Gauss 分布を理解するのに必要となる、ベクトルと行列の掛け算を確認しましょう。大きな行列、ベクトルを考えると大変なので、例として次の 2×2 の行列とベクトルを考えましょう。

$$\boldsymbol{M} = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}, \qquad \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

ベクトルと行列の掛け算としては、次の 2 パターンがあります。

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M} = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} = \begin{pmatrix} x_1 M_{11} + x_2 M_{21} & x_1 M_{12} + x_2 M_{22} \end{pmatrix}$$

$$\boldsymbol{M}\boldsymbol{x} = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 M_{11} + x_2 M_{12} \\ x_1 M_{21} + x_2 M_{22} \end{pmatrix}$$

ややこしいですが、これが計算のルールです。よく見ると、どのような規則で計算が行われているか分かると思います。大事なのは、$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}$ の結果は横ベクトルに、$\boldsymbol{M}\boldsymbol{x}$ の結果は縦ベクトルになる点です。

上の 2 つのパターンのベクトルと行列の掛け算を組み合わせた、次のような式を考えましょう。

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x} = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

このような量は、$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}$、$\boldsymbol{M}\boldsymbol{x}$、どちらの部分から計算しても構いません。

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x} = \big(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}\big)\boldsymbol{x} = \begin{pmatrix} x_1 M_{11} + x_2 M_{21} & x_1 M_{12} + x_2 M_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}}\big(\boldsymbol{M}\boldsymbol{x}\big) = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} x_1 M_{11} + x_2 M_{12} \\ x_1 M_{21} + x_2 M_{22} \end{pmatrix}$$

いずれにしても、先ほど確認した、横ベクトルと縦ベクトルの積となります。そのため、どちらで計算しても、

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x} = M_{11}x_1^2 + \big(M_{12} + M_{21}\big)x_1 x_2 + M_{22}x_2^2$$

となります。大事なことは、$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x}$ の計算結果は、行列でもベクトルでもなく、普通の数 (スカラー) となるという点です。ここでは、2×2 の行列を例に説明しましたが、この計算はそのまま $n \times n$ の行列での計算に応用できます。

♣ 多変量 Gauss 分布の式を理解していく

以上で基本的な準備は完了しました。もう 1 度、最初の式を見てみましょう。

$$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\beta}^{-1}) = \sqrt{\frac{\det(\boldsymbol{\beta})}{(2\pi)^n}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\beta}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

以下では 1 変数の場合の式

$$\mathcal{N}(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

と比較しながら、最初の式を理解していきます。先ず、確率変数が 1 つの場合には単に x と書いていましたが、複数 (n 個) の確率変数が含まれる場合には、n 個の確率変数 x_i $(i = 1, \cdots, n)$ をまとめたベクトル $\boldsymbol{x}$ を導入すると便利です。

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

確率変数が n 個あると、それぞれのデータに対して平均値 μ が存在するため、μ も次のようなベクトルとなっています。

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

ですから、2 つのベクトルの引き算は、

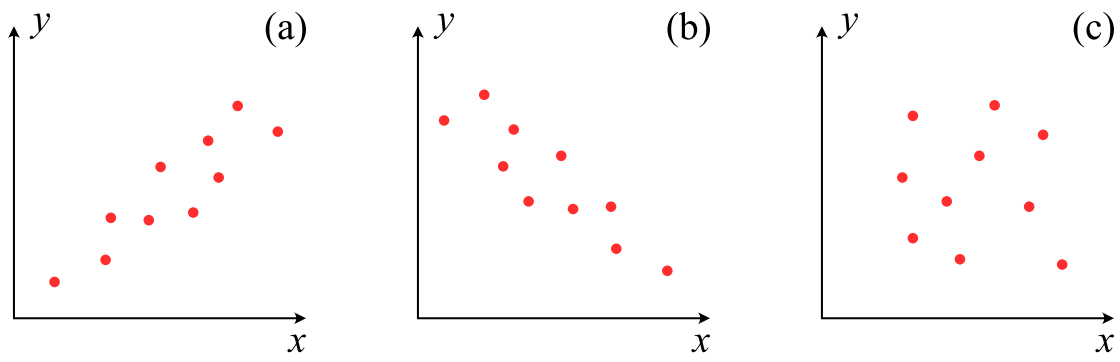
$$\boldsymbol{x}-\boldsymbol{\mu} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \\ \vdots \\ x_n-\mu_n \end{pmatrix}$$

となっています。多変量 Gauss 分布に含まれる $\boldsymbol{\beta}$ は**「精度行列 (precision matrix)」**と呼ばれる $n \times n$ の行列です[41]。この行列は、確率変数の散らばり具合を表す量で、共分散と呼ばれる量を一般化したものだと考えられます。2 つの確率変数 x, y に対して、

$$\mathrm{Cov}(x,y) = E\big[\big(x-E[x]\big)\big(y-E[y]\big)\big] = E[xy]-E[x]E[y]$$

[41] 逆共分散行列 (inverse covariance matrix) とも呼ばれます

を x と y の共分散といいます。$E[x]$ は Bernoulli 分布のところで説明した確率変数 x の期待値です。なにやらややこしい量ですが、この量は 2 つの変数 x, y の「相関」を表しています。相関というのは、2 つの変数の関係性のことです。例えば、確率変数 x, y のデータ点を 10 点ほどプロットしたら、次のようになったとしましょう。



(a) の場合には、x が増加すると y も増加する傾向にあることが分かります。この場合には、x と y の間には「正の相関がある」と言います。一方、(b) の場合には、x が増加すると y が減少する傾向にあることが分かります。この場合には、x と y の間には「負の相関がある」と言います。そして (c) の場合には、特に x と y の間には特に関係がなくバラバラにデータが広がっています。このような場合を「無相関」といいます。共分散 $\mathrm{Cov}(x, y)$ の値は、正の相関がある場合には正、負の相関がある場合には負の数となります。そして無相関の場合には、共分散の値は 0 となります。このように共分散 $\mathrm{Cov}(x, y)$ は、x と y の相関を反映した量となっています。

共分散について最後にいくつかコメントしておきます。共分散に関しては、定義から次の関係式が成り立ちます。

$$\mathrm{Cov}(x_1, x_2) = \mathrm{Cov}(x_2, x_1)$$

x_1 と x_2 を入れ替えても値は変わりません。また、x と x に関する共分散は、

$$\mathrm{Cov}(x, x) = E[x^2] - E[x]^2$$

となります。2 乗の期待値から期待値の 2 乗を引いた量は、Bernoulli 分布のところで出てきた「分散」に他なりません。つまり、

$$\mathrm{Cov}(x, x) = \mathrm{Var}(x)$$

ということになります。

この共分散を用いて、精度行列 $\boldsymbol{\beta}$ (の逆行列) は次のように与えられます。

$$\boldsymbol{\beta}^{-1} = \begin{pmatrix} \mathrm{Cov}(x_1, x_1) & \mathrm{Cov}(x_1, x_2) & \cdots & \mathrm{Cov}(x_1, x_n) \\ \mathrm{Cov}(x_2, x_1) & \mathrm{Cov}(x_2, x_2) & \cdots & \mathrm{Cov}(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(x_n, x_1) & \mathrm{Cov}(x_n, x_2) & \cdots & \mathrm{Cov}(x_n, x_n) \end{pmatrix}$$

つまり n 個の確率変数 x_i の関係を調べるためには、2 つの変数を選び出してそれらの共分散を考えることになりますが、2 つの変数の選び方にはたくさんパターンがあります。それらの情報を 1 つにまとめてしまったのが精度行列 $\boldsymbol{\beta}$ なのです。尚、上の式では $\boldsymbol{\beta}$ の表式ではなく、$\boldsymbol{\beta}$ の「逆行列」と呼ばれる行列 $\boldsymbol{\beta}^{-1}$ の方が綺麗な形となるので、逆行列の形で書いています[42]。一般のサイズの行列の逆行列の計算の仕方は、大学レベルの数学となり、かなりややこしいのでここでは説明を省略します。ただ実際には、逆行列の計算はコンピュータで簡単にできてしまうので、使う分には計算アルゴリズムの詳細を知らなくても全然問題ありません。

以上で必要な行列 $\boldsymbol{\beta}$ とベクトル $\boldsymbol{x}$, $\boldsymbol{\mu}$ が揃いました。これらの行列、ベクトルを用いて多変量 Gauss 分布は構成されています。

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\beta}^{-1}) = \sqrt{\frac{\det(\boldsymbol{\beta})}{(2\pi)^n}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\beta} (\boldsymbol{x} - \boldsymbol{\mu}) \right)$$

$(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\beta} (\boldsymbol{x} - \boldsymbol{\mu})$ の部分は、行列とベクトルの積となっています。先ほどのベクトルと行列のところで説明しましたが、このような横ベクトル、行列、縦ベクトルの積は、普通の数 (スカラー) となります。つまり、$(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\beta} (\boldsymbol{x} - \boldsymbol{\mu})$ もルールに従って計算すれば、スカラーの量となります。そして exp の前についている

$$\sqrt{\frac{\det(\boldsymbol{\beta})}{(2\pi)^n}}$$

は、1 変数の場合の $\sqrt{\frac{1}{2\pi\sigma^2}}$ に対応する係数です。$\det(\boldsymbol{\beta})$ というのは、$\boldsymbol{\beta}$ の**「行列式 (determinant)」**と呼ばれる量です。一般のサイズの行列の det を求めるアルゴリズムも逆行列と同様、かなり複雑なので (大学レベルの数学となります)、ここでは説明を省略しますが、行列に対してスカラーを返すような計算です。つまり、$\det(\boldsymbol{\beta})$ は普通の数であり、パッケージやライブラリを利用すればどのプログラミング言語でも簡単に計算できる量となっています。この係数も 1 変数の場合と同様に、分布 $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\beta}^{-1})$ が確率分布の基本ルール「全て合計したら 1 になる」を満たすようにするためにくっついています。

[42] 精度行列 $\boldsymbol{\beta}$ の逆行列 $\boldsymbol{\beta}^{-1}$ は、分散共分散行列と呼ばれ、一般には $\boldsymbol{\Sigma}$ と表します。

以上で、多変量 Gauss 分布に関する説明はおしまいです。一般の n 変数の場合だとこのようにややこしくて分かりにくいので、最後に 2 変数の場合の多変量 Gauss 分布を具体的に考えます。具体例をみれば、より理解が深まるかと思います。

♣ 2 変数の多変量 Gauss 分布

では、具体的に $n = 2$ の場合、つまり 2 変数に対する Gauss 分布を考えてみましょう。確率変数のベクトル $\boldsymbol{x}$、平均値のベクトル $\boldsymbol{\mu}$ は、この場合、

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

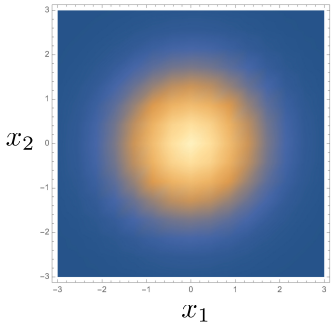
となります。平均は 1 変数の場合に説明したように、全体の位置をずらすだけなので、ここでは簡単のために、$\mu_1 = \mu_2 = 0$ としておきましょう。精度行列 $\boldsymbol{\beta}$ の逆行列 (つまり分散共分散行列) は、この場合

$$\boldsymbol{\beta}^{-1} = \begin{pmatrix} \mathrm{Cov}(x_1, x_1) & \mathrm{Cov}(x_1, x_2) \\ \mathrm{Cov}(x_2, x_1) & \mathrm{Cov}(x_2, x_2) \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(x_1) & \mathrm{Cov}(x_1, x_2) \\ \mathrm{Cov}(x_1, x_2) & \mathrm{Var}(x_2) \end{pmatrix}$$

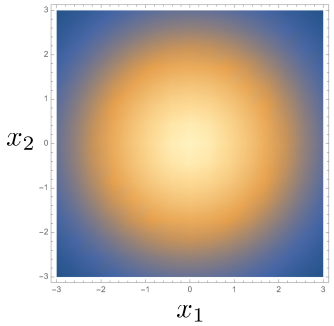
となっています。前のセクションで説明したように、$\mathrm{Cov}(x_1, x_2) = \mathrm{Cov}(x_2, x_1)$ であること、$\mathrm{Cov}(x_1, x_1) = \mathrm{Var}(x_1)$ であることを用いています。以下では先ず、x_1 と x_2 が無相関である場合を考えましょう。この場合には、共分散 $\mathrm{Cov}(x_1, x_2)$ は 0 となるので、

$$\boldsymbol{\beta}^{-1} = \begin{pmatrix} \mathrm{Var}(x_1) & 0 \\ 0 & \mathrm{Var}(x_2) \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

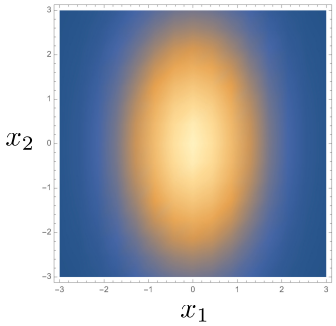
となります。ここでは、1 変数の場合の Gauss 分布で分散を σ^2 と表現していたため、それを真似して $\mathrm{Var}(x_1) = \sigma_1^2$、$\mathrm{Var}(x_2) = \sigma_2^2$ としました。σ_1^2、σ_2^2 の値をいくつか変えて多変量 Gauss 分布のグラフを描いてみると、次のようになります。



$$\boldsymbol{\beta}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



$$\boldsymbol{\beta}^{-1} = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$$



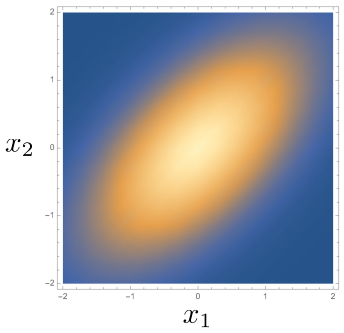
$$\boldsymbol{\beta}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$

上の図では、色が薄い (白っぽい) ほど確率分布の値が大きく、濃い (青っぽい) ほど小さくなっています。1 変数の場合と同様に、σ_1^2 と σ_2^2 は、それぞれの変数方向の分布の広がり具合を変化させるパラメータとなっていることが分かります。σ_1^2 を大きくすると、x_1 方向に分布が広がる様子が見て取れます。

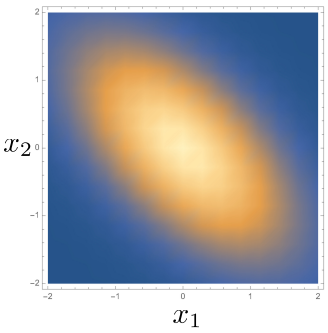
次に、x_1 と x_2 が相関を持っており、$\mathrm{Cov}(x_1, x_2)$ が 0 でない場合を考えましょう。例として、分散共分散行列が

$$\boldsymbol{\beta}^{-1} = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}, \qquad \boldsymbol{\beta}^{-1} = \begin{pmatrix} 1 & -0.6 \\ -0.6 & 1 \end{pmatrix}$$

の 2 つの場合の多変量 Gauss 分布のグラフを描いてみると、次のようになります。



$$\boldsymbol{\beta}^{-1} = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$$



$$\boldsymbol{\beta}^{-1} = \begin{pmatrix} 1 & -0.6 \\ -0.6 & 1 \end{pmatrix}$$

$\mathrm{Cov}(x_1, x_2) = 0$ の場合とは異なり、分布が歪んで $\mathrm{Cov}(x_1, x_2) > 0$ の場合には正の相関、$\mathrm{Cov}(x_1, x_2) < 0$ の場合には負の相関を反映した分布となっていることが分かります。このように、多変量 Gauss 分布では精度行列 (分散共分散行列) の値を調整することで、分布の広がり具合や歪み具合を調整できるようになっています。