正規分布の概要

2019年（作成）
2022年8月（改定）
倉谷　隆博

1．確率密度関数(probability density function)

正規分布の確率密度関数は（式1）、（図1）で表される。

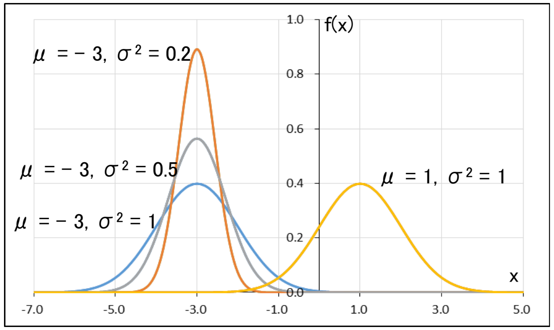
$f(x) = \{ 1 / (2\pi\sigma^2)^{1/2} \} e^{-(x-\mu)^2/(2\sigma^2)}$ ・・・ （式1）

x：確率変数，μ：期待値，σ^2：分散

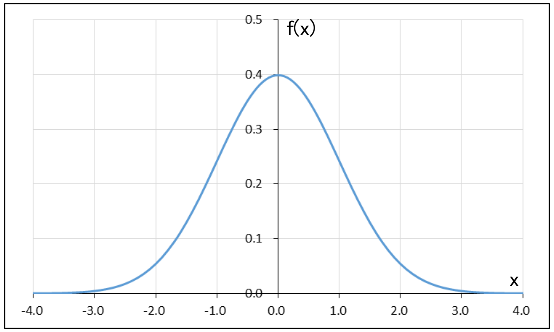
また、正規分布を、期待値が 0, 分散が 1 になるように標準化（変数変換）した標準正規分布の確率密度関数は（式2）、（図2）で表される。

$f(x) = \{ 1/(2\pi)^{1/2} \} e^{-x^2/2}$ ・・・ （式2）

なお、詳細は、資料「いろいろな確率分布」を参照のこと。



（図1）正規分布



（図2）標準正規分布

■確率の計算

正規分布は連続型の確率分布であるので、確率 P は確率密度関数の積分（面積）によって求まる。

$P(a \leqq x \leqq b) = \int_a^b f(x)\,dx = \int_{-\infty}^b f(x)\,dx - \int_{-\infty}^a f(x)\,dx = F(b) - F(a)$ ・・・ （式3）

ここで、P (a ≦ x ≦ b) は、確率変数 X の実現値 x が a 以上、b 以下のときの確率を表す。

また、F(x) は、f(x) の累積分布関数を表す。

（式3）の確率を求める手順を、標準正規分布を使用して（図3）で示す。

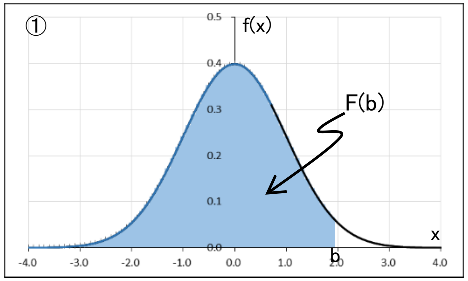
また、（図4）、（図5）、（図6）に 1 σ区間、2 σ区間、3 σ区間の確率を標準正規分布で示す。
いずれのグラフ作成も、Excel の NORM.DIST関数を使用した。

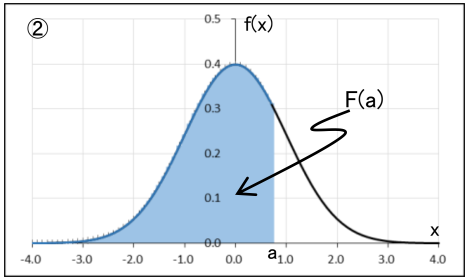
（図4） 1 σ 区間の確率（μ－σ ～ μ＋σ，標準正規分布で－1 ～ +1 の範囲に入る確率）は、68．3 %

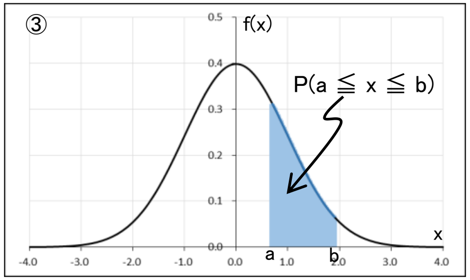
（図5） 2 σ 区間の確率（μ－2σ～ μ＋2σ，標準正規分布で－2 ～ +2 の範囲に入る確率は、95．4%

（図6） 3 σ 区間の確率（μ－3σ～ μ＋3σ，標準正規分布で－3 ～ +3 の範囲に入る確率は、99．7%

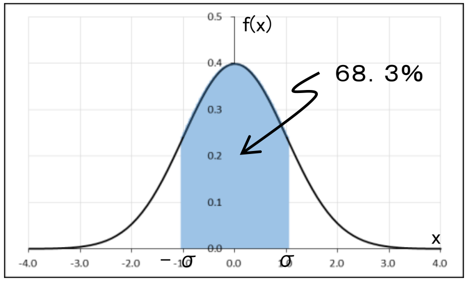
μ ± 3σの範囲から外れる確率は 0．3% になるので、極めて稀で、この場合を俗に、千三つという。



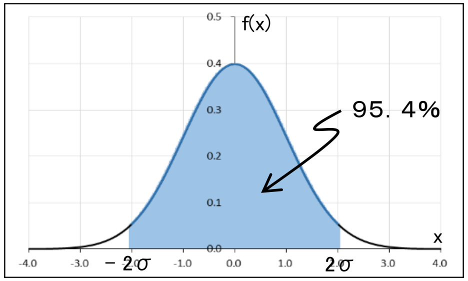




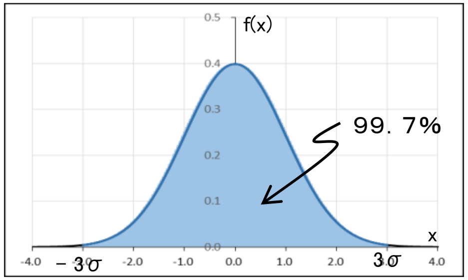
③の面積 = ①の面積 − ②の面積
（図3）確率の計算



（図4）1σ 区間の確率



（図5）2σ 区間の確率



（図6）3σ 区間の確率

2．再生性(reproductive property) 1/2

同じ確率分布に属する互いに独立な確率変数があり、これらの確率変数の定数倍や和がまた同じ確率分布になるとき、この確率分布は再生性を持つという。正規分布、二項分布、負の二項分布、ポアソン分布、ガンマ分布、カイ二乗分布、コーシー分布などは再生性を持つ確率分布である。

ここでは、正規分布の再生性を定数倍と和に分けて確認する。

（1）定数倍の再生性

確率変数 X が期待値 μ , 分散 σ^2 の正規分布に従い、その確率密度関数を f(x) とする。このとき、a を定数として、定数倍の確率変数 a X が、期待値 a μ , 分散 $a^2\sigma^2$ の正規分布 f(x) に従うことを定数倍の再生性という。

（2）和の再生性

確率変数 X_1 が期待値 μ_1 , 分散 $\sigma_1{}^2$ の正規分布に従い、その確率密度関数を $f_1(x)$ とする。確率変数 X_2 が期待値 μ_2 , 分散 $\sigma_2{}^2$ の正規分布に従い、その確率密度関数を $f_2(x)$ とする。このとき、$X_1 + X_2$ が期待値 $\mu_1 + \mu_2$, 分散 $\sigma_1{}^2 + \sigma_2{}^2$ の正規分布 f(x) に従うことを和の再生性という。

■定数倍の再生性の確認

期待値 μ , 分散 σ^2 の正規分布の確率密度関数は、$f(x) = (1 / (2\pi\sigma^2)^{1/2}\, e^{-(x-\mu)^2/(2\sigma^2)}$ で表すことができる。

また、正規分布に従う確率変数 X の定数倍 a X に関する閉区間 [s, t] における確率は、

$$P(s \leqq aX \leqq t) = P(s / a \leqq X \leqq t / a) = \int_{s/a}^{t/a} f(x)\,dx = (1 / (2\pi\sigma^2)^{1/2} \int_{s/a}^{t/a} e^{-(x-\mu)^2/(2\sigma^2)}\,dx$$

ここで、y = a x として置換積分を行う。dy = a dx　積分範囲は x : s / a → t / a に対して、y : s → t である。

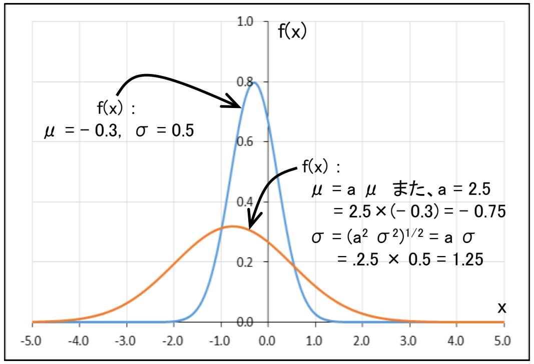
$$P(s \leqq y \leqq t) = (1 / (2\pi\sigma^2)^{1/2} \int_{s}^{t} e^{-(y/a-\mu)^2/(2\sigma^2)} (1 / a)\,dy = (1 / (2\pi a^2 \sigma^2)^{1/2} \int_{s}^{t} e^{-(y-a\mu)^2/(2a^2\sigma^2)}\,dy$$

従って、Y = a X の確率密度関数 f(y) は、改めて、y を x で置き換えて表すと、

$$f(x) = (1 / (2\pi a^2 \sigma^2)^{1/2}\, e^{-(x-a\mu)^2/(2a^2\sigma^2)} \quad \cdots \quad (式4)$$

この確率密度関数は、期待値 a μ , 分散 $a^2\sigma^2$ の正規分布である。

正規分布の定数倍の再生性の一例を、（図7）に示す。



（図7）正規分布の定数倍の再生性

2. 再生性(reproductive property) 2/2

■和の再生性の確認

$X = X_1 + X_2$ の確率密度関数 f(x) は畳み込み積分で表される（別途資料「いろいろな確率分布」の付録17）。

$$f(x) = \int_{-\infty}^{\infty} f_1(x-y)\, f_2(y)\, dy = \int_{-\infty}^{\infty} \{ (1/(2\pi\sigma_1^2)^{1/2})\, e^{-((x-y)-\mu_1)^2/(2\sigma_1^2)} \} \{ (1/(2\pi\sigma_2^2)^{1/2})\, e^{-(y-\mu_2)^2/(2\sigma_2^2)} \}\, dy$$

$$= (1/(2\pi\sigma_1\sigma_2)) \int_{-\infty}^{\infty} e^{-((x-y)-\mu_1)^2/(2\sigma_1^2) - (y-\mu_2)^2/(2\sigma_2^2)}\, dy$$

ここで、非積分関数の指数項を A とおいて、

$$A = -((x-y)-\mu_1)^2/(2\sigma_1^2) + (y-\mu_2)^2/(2\sigma_2^2)) = -1/(2\sigma_1^2\sigma_2^2)\{ \sigma_2^2(-y+(x-\mu_1))^2 + \sigma_1^2(y-\mu_2)^2 \}$$

$$= -1/(2\sigma_1^2\sigma_2^2)\{ \sigma_2^2 y^2 - 2\sigma_2^2(x-\mu_1)y + \sigma_2^2(x-\mu_1)^2 + \sigma_1^2 y^2 - 2\sigma_1^2\mu_2 y + \sigma_1^2\mu_2^2 \}$$

$$= -1/(2\sigma_1^2\sigma_2^2)\{ (\sigma_1^2+\sigma_2^2)y^2 - 2(\sigma_2^2 x - \sigma_2^2\mu_1 + \sigma_1^2\mu_2)y + \sigma_2^2(x-\mu_1)^2 + \sigma_1^2\mu_2^2 \}$$

$$= -1/(2\sigma_1^2\sigma_2^2)\{(\sigma_1^2+\sigma_2^2)(y-(\sigma_2^2(x-\mu_1)+\sigma_1^2\mu_2)/(\sigma_1^2+\sigma_2^2))^2 - (\sigma_2^2(x-\mu_1)+\sigma_1^2\mu_2)^2/(\sigma_1^2+\sigma_2^2) + \sigma_2^2(x-\mu_1)^2 + \sigma_1^2\mu_2^2\}$$

$B = -(\sigma_2^2(x-\mu_1)+\sigma_1^2\mu_2)^2/(\sigma_1^2+\sigma_2^2) + \sigma_2^2(x-\mu_1)^2 + \sigma_1^2\mu_2^2$ とおいて、

$$= (1/(\sigma_1^2+\sigma_2^2))\{ -\sigma_2^4(x-\mu_1)^2 - 2\sigma_2^2(x-\mu_1)\sigma_1^2\mu_2 - \sigma_1^4\mu_2^2 + (\sigma_1^2+\sigma_2^2)(\sigma_2^2(x-\mu_1)^2 + \sigma_1^2\mu_2^2) \}$$

$$= (1/(\sigma_1^2+\sigma_2^2))\{ -\sigma_2^4 x^2 + 2\sigma_2^4 x\mu_1 - \sigma_2^4\mu_1^2 - 2\sigma_2^2 x\sigma_1^2\mu_2 + 2\sigma_2^2\mu_1\sigma_1^2\mu_2 - \sigma_1^4\mu_2^2$$
$$+ (\sigma_1^2+\sigma_2^2)(\sigma_2^2 x^2 - 2\sigma_2^2 x\mu_1 + \sigma_2^2\mu_1^2 + \sigma_1^2\mu_2^2) \}$$

$$= (1/(\sigma_1^2+\sigma_2^2))(-\sigma_2^4 x^2 + 2\sigma_2^4 x\mu_1 - \sigma_2^4\mu_1^2 - 2\sigma_2^2 x\sigma_1^2\mu_2 + 2\sigma_2^2\mu_1\sigma_1^2\mu_2 - \sigma_1^4\mu_2^2$$
$$+ \sigma_1^2\sigma_2^2 x^2 - 2\sigma_1^2\sigma_2^2 x\mu_1 + \sigma_1^2\sigma_2^2\mu_1^2 + \sigma_1^4\mu_2^2 + \sigma_2^4 x^2 - 2\sigma_2^4 x\mu_1 + \sigma_2^4\mu_1^2 + \sigma_1^2\sigma_2^2\mu_2^2)$$

$$= (1/(\sigma_1^2+\sigma_2^2))(-2\sigma_2^2 x\sigma_1^2\mu_2 + 2\sigma_2^2\mu_1\sigma_1^2\mu_2 + \sigma_1^2\sigma_2^2 x^2 - 2\sigma_1^2\sigma_2^2 x\mu_1 + \sigma_1^2\sigma_2^2\mu_1^2 + \sigma_1^2\sigma_2^2\mu_2^2)$$

$$= (\sigma_1^2\sigma_2^2/(\sigma_1^2+\sigma_2^2))(-2x\mu_2 + 2\mu_1\mu_2 + x^2 - 2x\mu_1 + \mu_1^2 + \mu_2^2)$$

$$= (\sigma_1^2\sigma_2^2/(\sigma_1^2+\sigma_2^2))\{ x^2 - 2(\mu_1+\mu_2)x + (\mu_1+\mu_2)^2 \} = (\sigma_1^2\sigma_2^2/(\sigma_1^2+\sigma_2^2))\{ x - (\mu_1+\mu_2) \}^2$$

よって、$A = -1/(2\sigma_1^2\sigma_2^2)\{(\sigma_1^2+\sigma_2^2)(y-(\sigma_2^2(x-\mu_1)+\sigma_1^2\mu_2)/(\sigma_1^2+\sigma_2^2))^2 + (\sigma_1^2\sigma_2^2/(\sigma_1^2+\sigma_2^2))(x-(\mu_1+\mu_2))^2\}$

$$= -((\sigma_1^2+\sigma_2^2)/(2\sigma_1^2\sigma_2^2))\{ y - (\sigma_2^2(x-\mu_1)+\sigma_1^2\mu_2)/(\sigma_1^2+\sigma_2^2) \}^2 - (x-(\mu_1+\mu_2))^2/(2(\sigma_1^2+\sigma_2^2))$$

$$f(x) = (1/(2\pi\sigma_1\sigma_2))\int_{-\infty}^{\infty} \exp\{-((\sigma_1^2+\sigma_2^2)/(2\sigma_1^2\sigma_2^2))\{y-(\sigma_2^2(x-\mu_1)+\sigma_1^2\mu_2)/(\sigma_1^2+\sigma_2^2)\}^2 - (x-(\mu_1+\mu_2))^2/(2(\sigma_1^2+\sigma_2^2))\}\, dy$$

$$= (1/(2\pi\sigma_1\sigma_2))\exp\{-(x-(\mu_1+\mu_2))^2/(2(\sigma_1^2+\sigma_2^2))\} \int_{-\infty}^{\infty} \exp\{-((\sigma_1^2+\sigma_2^2)/(2\sigma_1^2\sigma_2^2))\{y-(\sigma_2^2(x-\mu_1)+\sigma_1^2\mu_2)/(\sigma_1^2+\sigma_2^2)\}^2\}\, dy$$

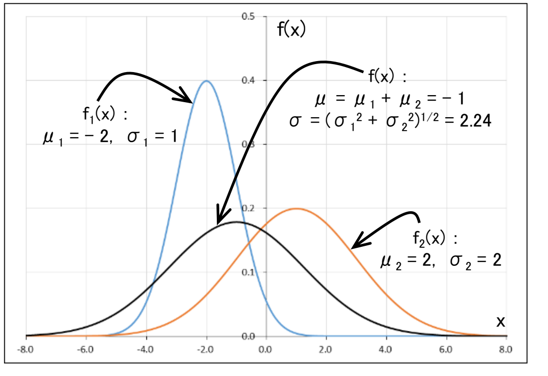
ここで、ガウス積分を使って（別途資料「いろいろな確率分布」の付録9を参照）、積分を計算し、

$$f(x) = (1/(2\pi\sigma_1\sigma_2))\exp\{-(x-(\mu_1+\mu_2))^2/(2(\sigma_1^2+\sigma_2^2))\}\{(2\sigma_1^2\sigma_2^2\pi)/(\sigma_1^2+\sigma_2^2)\}^{1/2}$$

$$f(x) = (1/(2\pi(\sigma_1^2+\sigma_2^2))^{1/2})\, e^{-(x-(\mu_1+\mu_2))^2/(2(\sigma_1^2+\sigma_2^2))}$$ ・・・（式5）

従って、正規分布に従う確率変数の和の確率分布は、期待値も分散も元の確率分布の期待値と分散の和で表されることが分かる。

正規分布の和の再生性の一例を、（図8）に示す。

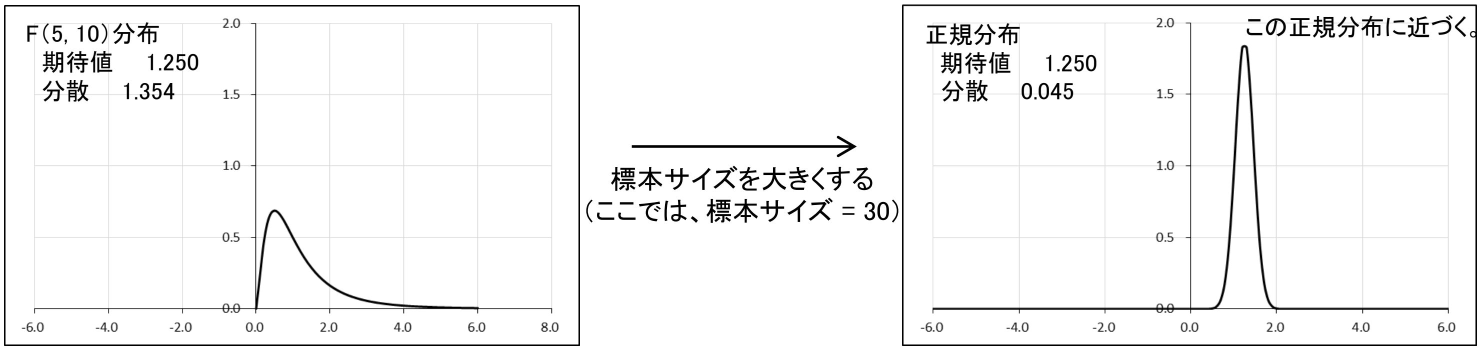


（図8）正規分布の和の再生性

3. 中心極限定理(Central Limit Theorem)

中心極限定理は、母集団からの無作為抽出の試行を十分に多く繰り返せば、母集団がどのような確率分布であっても、母平均 μ と母分散 σ^2 が存在すれば、標本平均の分布が期待値 μ，分散 σ^2/n の正規分布に近づくという定理である(別途資料「いろいろな確率分布」を参照)。
標本サイズを増やせば正規分布に近似できる性質は、推定とか正規分布から派生する確率分布(カイ二乗分布など)を使用した検定に利用される。

例えば(図9)に示すように、期待値 μ = 1.250, 分散 σ^2 = 1.354 の母集団(F 分布を挙げているが、期待値と分散が定まる確率分布であればどんな確率分布であってもいい)を対象に、標本サイズを大きくしたときに(例えば、標本サイズ n = 30)、標本平均の確率分布は、期待値 = 1.250, 分散 = σ^2/n = 1.354/30 = 0.045 の正規分布に近づく。



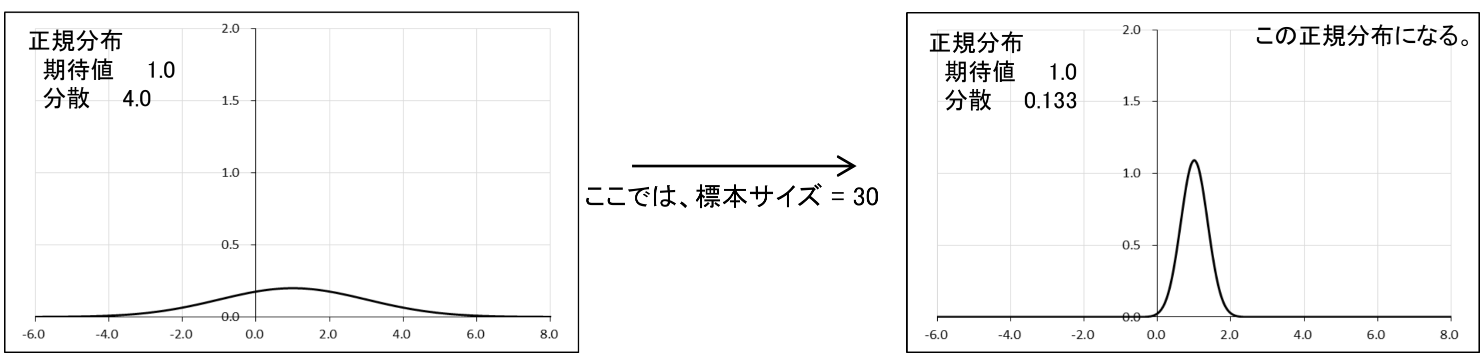
(図9)中心極限定理

なお参考までに、正規分布(期待値 μ，分散 σ^2)から n 個抽出したデータ $X_1, X_2, \cdots, X_n$ の標本平均 $\bar{X}$ の確率分布は、下記に示すとおり、期待値 μ，分散 $\sigma 2 / n$ の正規分布になる(別途資料「いろいろな確率分布」の付録46. 中心極限定理の導出の(1)を参照)。

標本平均の期待値 $E[\bar{X}] = E[(1/n)\sum_{i=1}^{n} Xi] = (1/n)E[\sum_{i=1}^{n} Xi] = (1/n)n\mu = \mu$ ・・・ (式6)

標本平均の分散 $V[\bar{X}] = V[(1/n)\sum_{i=1}^{n} Xi] = (1/n^2)E[\sum_{i=1}^{n} Xi] = (1/n^2)n\sigma^2 = \sigma^2/n$ ・・・ (式7)

例えば、(図10)に示すように、期待値 μ = 1.0，分散 σ^2 = 4.0 の母集団を対象に、標本サイズを大きくしたときに(例えば、標本サイズ n = 30)、標本平均の確率分布は、期待値 = 1.0，分散 = σ^2/n = 0.133 の正規分布になる(正規分布は正規分布のままである)。



(図10)正規分布の標本平均の確率分布

4. 混合正規分布（Contaminated Normal distribution）

複数の正規分布の和が作り出す確率分布を混合正規分布（混合ガウス分布）と呼ぶ。任意の確率密度関数を、複数の正規分布の和によって近似することができる。ちょうど三角関数の和で任意の波形を近似するフーリエ級数に似た考え方になる。クラスタリングなどに使用される。

混合正規分布（式９）は、多変量正規分布（式８）の線型結合で表される。（多変量正規分布は、別途資料「いろいろな確率分布」を参照）。

$g(\mathbf{x}) = \{1/((2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2})\}\, e^{-(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})/2}$ ・・・（式８）

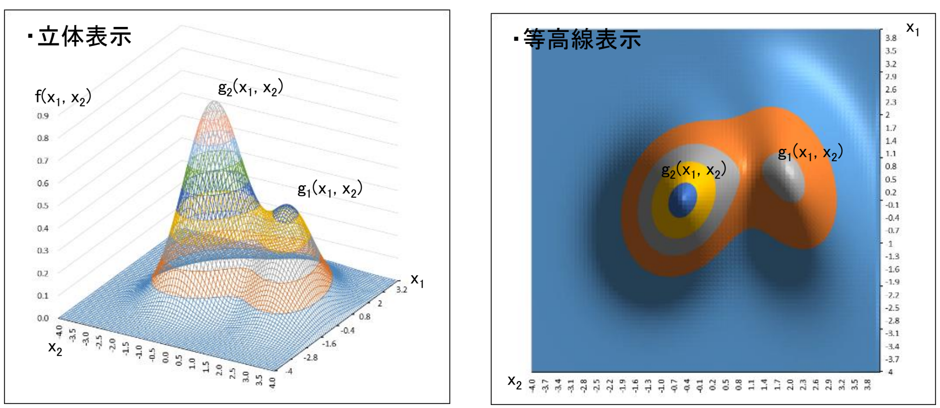
$\mathbf{x}$：確率変数ベクトル　$\boldsymbol{\mu}$：期待値ベクトル　$\boldsymbol{\Sigma}$：分散共分散行列　n：次元数

$f(\mathbf{x}) = \sum_{i=1}^{k} \pi_i\, g_i(\mathbf{x})$ ・・・（式９）

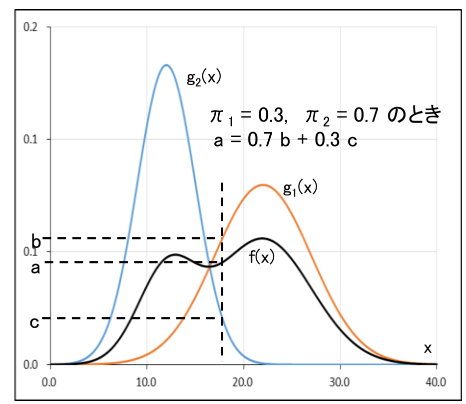
$\mathbf{x}$：確率変数ベクトル，π_i：$g_i(\mathbf{x})$ の混合係数 $\sum_{i=1}^{k} \pi_i = 1$，k：線型結合（混合）する多変量正規分布の数

混合正規分布の（図１１）に示す。
期待値 =（0.5, 2.0）　分散 =（1.2, 0.8）　相関係数 = − 0.3 の多変量正規分布と
期待値 =（0.0, − 0.5）　分散 =（1.0, 0.8）　相関係数 = 0.2 の多変量正規分布とを、混合係数 $\pi_1 = 0.3$, $\pi_2 = 0.7$ で重ね合わせた混合正規分布である。

なお、1 次元の正規分布の混合正規分布の模式図を（図１２）に示す。



（図１１）混合正規分布



（図１２）混合正規分布の模式図

5. 偏差値(deviation score)

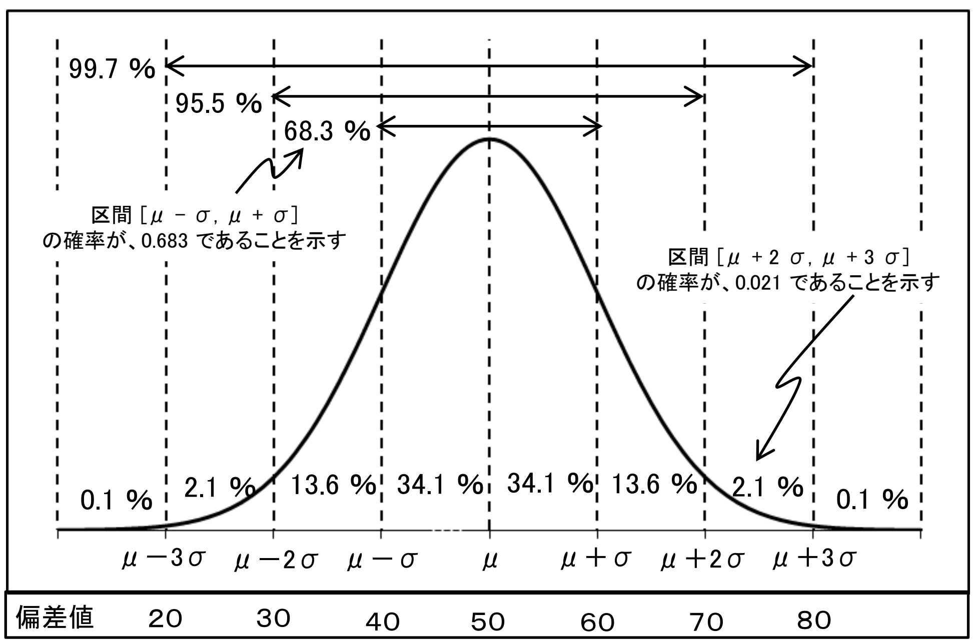
正規分布(平均 = μ, 分散 = σ)を標準正規分布(平均 = 0, 分散 = 1)に標準化した後、使い勝手を考慮して 平均 = 50, 標準偏差 = 10 の正規分布に作り直す。この新たな正規分布に従う確率変数が偏差値である。偏差値の計算式は(式10)である。
偏差値の計算式の肝となるのは、$(x_i - \mu) / \sigma$ (平均 = μ, 標準偏差 = σ)の項である。この項は標準化の式そのものである(資料「いろいろな確率分布」の付録19を参照)。また、この項は x_i が平均 μ からどのくらい離れているかを、その距離を標準偏差 σ の何倍であるかによって示しているという見方ができる。

$$z_i = 10\,(x_i - \mu) / \sigma + 50 \quad \cdots \quad (式10)$$

z_i : x_i の偏差値, x_i : データ, μ : 平均(期待値), σ : 標準偏差(σ^2 : 分散)

(式10)で、$(x_i - \mu) / \sigma$ の値を 10 倍しているが、これは例えば、$x_i = \mu + \sigma$ のとき、$10\,(x_i - \mu)/\sigma = 10\,(\mu + \sigma - \mu)/\sigma = 10$ となることから分かるように、計算結果 10 は x_i が平均 μ から σ 離れていることを示している。つまり、計算結果 10 は標準偏差を示している。
また、(式10)で $(x_i - \mu)/\sigma$ に 50 をプラスしているが、これは $x_i = \mu$ のとき、$z_i = 10\,(x_i - \mu)/\sigma + 50 = 10\,(\mu - \mu)/\sigma + 50 = 50$ となることから、平均を 50 にするための工夫である。よって、(式10)は、平均 = 50, 標準偏差 = 10 の正規分布にするための式であると言える。
なお、(式10)から分かるように、偏差値の取る値の範囲は、－∞(マイナス無限大) から ＋∞(プラス無限大)である。

(図13)に偏差値と正規分布との関係を示す。



(図13)正規分布と偏差値

偏差値を使って、テストの成績(順位)を計算する。
計算には、Excel の NORM.DIST 関数を使う。

例えば、
・テストの点数が偏差値 = 64 であったとすると、
　上位から、(1 − NORM.DIST(64, 50, 10, true)) × 100 = 8.1 %
　のところに位置する。
　よって、200人が受けたテストであれば、
　200 × 8.1 % = 16.2　上から、17 番目になる。

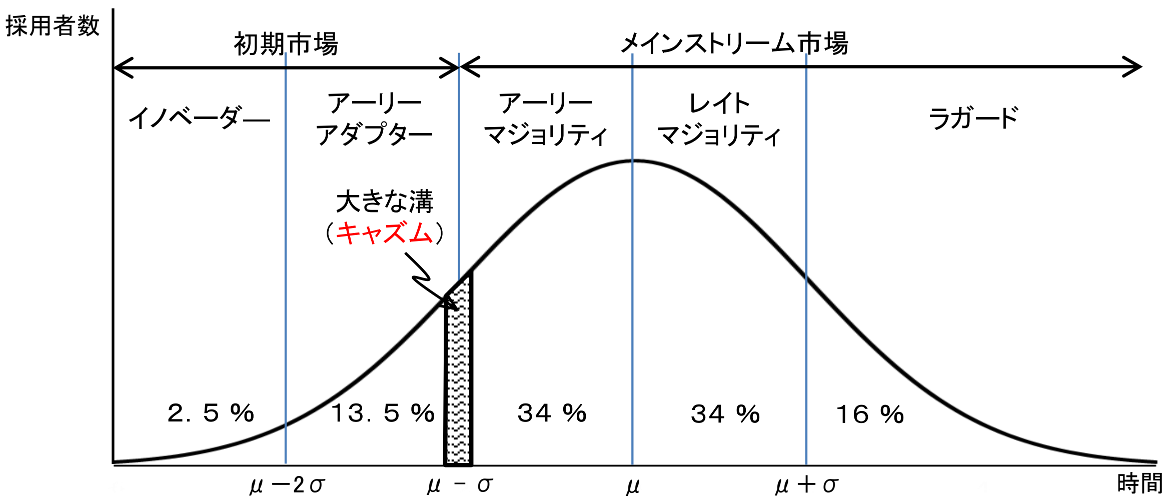
・テストの点数が偏差値 = 39 であったとすると、
　上位から、(1 − NORM.DIST(39, 50, 10, true)) × 100 = 86.4 %
　のところに位置する。
　よって、200人が受けたテストであれば、
　200 × 86.4 % = 172.9　上から、173 番目になる。

7. イノベーター理論(Diffusion of innovations)

イノベーター理論は、商品やサービスの普及段階を表すマーケティング理論である(米国 E.M.ロジャース)。商品やサービスの普及段階を正規分布を使った5段階(表1)に分類して、マーケティング戦略の検討を行う必要があると提案した。
この理論では、商品が普及する鍵は、アーリーマジョリティに大きな影響を与えるアーリーアダプターの攻略にあると、主張している。
また、このイノベーター理論をもとに、初期市場とメインストリーム市場の間には大きな溝(Chasm キャズム)があり、この溝を乗り越えない限り、商品やサービスは初期市場の中でやがて消えていくというキャズム理論が、J.ムーア(米国)によって提唱された。
事例として、ネスレ日本(株)の「ネスカフェ・コーヒーアンバサダー」の取り組みは、アーリーアダプターへのアプローチであった。
また、フリーマーケットの(株)メルカリはUI(User Interface)やUX(User Experience)の改良を積み重ねてキャズムを乗り越えた。

(表1)イノベータ理論の段階

お客様のタイプ	割合(%)	特徴
革新者 (Innovators)	2．5	最も早く製品を使い始める「新しい」ことに価値を置いている層で、訴求ポイントは、「革新的」、「新技術」などである。
初期採用者 (Early Adapters)	13．5	イノベーターほどではないが、トレンドに敏感な層である。オピニオンリーダー、インフルエンサーがこれにあたる。アーリーマジョリティに対して口コミなどで大きな影響を与える。この層の攻略が商品普及の鍵になる。訴求ポイントは、「新しい」ことに加え、具体的なメリットなどである。
前期追随者 (Early Majority)	34	先行する2つの層に比べて慎重な採用姿勢を取る層であるが、情報感度は高くアーリーアダプターから大きな影響を受ける。訴求ポイントは、新製品のメリット、流行に乗り遅れることのデメリットなどである。
後期追随者 (Late Majority)	34	新しいものの採用に消極的な層である。新製品の普及が広がってから採用を検討する。訴求ポイントは、普及率の高さ、採用しても失敗がないことなどである。
遅滞者 (Laggards)	16	保守的で、新しいものに対して全く関心がない層である。訴求ポイントは、高い安心度、製品の定番化などである。



(図14)正規分布とイノベーター理論

参考にした書籍

BellCurve 統計EB：「正規分布のグラフ作成」、 Excelによる正規分布曲線のグラフの作り方 | ブログ | 統計WEB (bellcurve.jp
数学の景色：「正規分布の再生性」、正規分布の再生性とその詳しい証明 | 数学の景色 (mathlandscape.com)
嵯峨山茂樹（東京大学）：「混合正規分布とEMアルゴリズム」、 D4-GaussianMixture.pdf (u-tokyo.ac.jp)
東大 IPC：「イノベーター理論を分かりやすく解説」、 イノベーター理論をわかりやすく解説！【事例あり】(utokyo-ipc.co.jp)
東大 IPC：「キャズム理論とは？」、 キャズム理論とは？キャズムが発生する理由、越えるための7つのポイント (utokyo-ipc.co.jp)
ONE MARKETING Co., LTD：「キャズム理論とは？」、 キャズム理論とは？キャズムを超える方法とその事例 | BtoBマーケティングラボ (onemarketing.jp)