７章 正規分布

正規分布（normal distribution）は、偶発的なデータのゆらぎによって生じる統計学で最も基本的な確率分布です。この章では正規分布についてその性質を詳しく見て行きましょう。

7.1 一般の正規分布

正規分布は、平均と分散の 2 つの量によって完全に特徴付けられています。平均μ、分散σ^2の正規分布は、$N(\mu,\sigma^2)$とも書かれます。ここにNは normal の頭文字を表わしています。確率変数Xがこの分布に従うとき、

$$X \sim N(\mu,\sigma^2)$$

のように表わされます。

平均μ、分散σ^2の正規分布の確率密度関数$f(x)$は、以下の式で与えられることが知られています。

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

この章で説明する正規分布の性質は、上の式からすべて導かれますが、この本ではあまりこの式にこだわらないように話を進めます。この関数のグラフを描くと、図 7-1 のようになります。

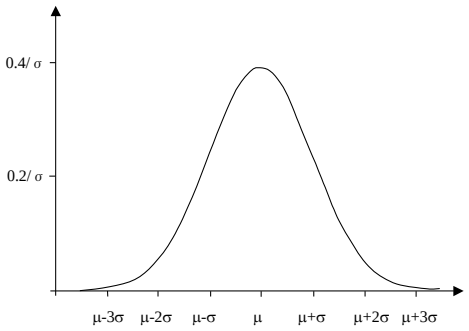

図 7-1　$N(\mu, \sigma^2)$分布の密度関数

ここに左右対称の山の中央が平均値μとなり、中間値も最頻値も平均値に一致します。山の高さは、確率密度関数の重要な性質、全面積が 1 であるというところから求まります。その値$f(\mu)$は、標準偏差のσを用いて、以下のように表わされます。

$$f(\mu)=\frac{1}{\sqrt{2\pi}\sigma}=\frac{0.3989\cdots}{\sigma}$$

確率密度関数の全面積の値は 1 に決まっていますので、分布の拡がりを表わす標準偏差が大きくなると、確率密度関数の山の高さは当然低くなります。

さてこのグラフから、確率変数Xがaの値以下となる確率を考えてみます。これは図 7-2 のグラフでは、$x=a$の位置から左側の面積に相当します。

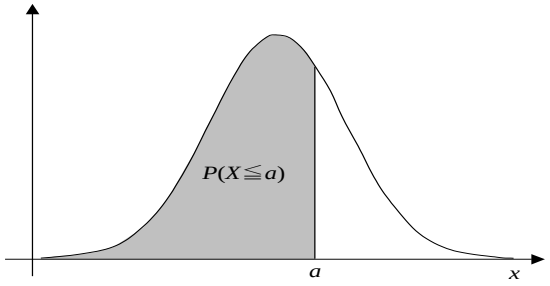

図 7-2 正規分布の確率

面積は積分で表わされる話はしましたので、確率は以下のように表わされます。

確率　　$P(X \le a) = \int_{-\infty}^{a} f(x)dx$

この値は一般に数式による積分では求められず、コンピュータ等による数値計算で値が求められます。平均と分散が与えられた場合のこの確率の計算は、Excel の関数を用いて求めることができますが、次に学ぶ標準正規分布に従う場合の計算の方がより覚え易いので、ここでは説明しないことにします。

全確率はグラフの全範囲の積分ですから、以下のようになります。

全確率　$P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$

平均と分散がそれぞれ μ，σ^2 で表わされるということは、式で表現すると、以下のようになります。

確率変数の平均　$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \mu$

確率変数の分散　$V(X) = E((X-\mu)^2) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \sigma^2$

これらの全確率や平均・分散の計算は、$f(x)$ として上で表わした式を用いると、計算に慣れた人なら簡単に示すことができますが、この本では省略します。

7.2 標準正規分布

ここでは、正規分布の中で特によく利用される特別なものを紹介しましょう。これは、平均が 0 で分散が 1 の正規分布です。平均と分散の記号を使うと、$\mu = 0$, $\sigma^2 = 1$ となります。これは、$N(0,1)$ 分布とも表示され、特別に標準正規分布（standard normal distribution）と呼ばれています。一般的な正規分布の確率密度関数を表わす式の中で $\mu = 0$，$\sigma = 1$ とおくと、標準正規分布に対する以下のような確率密度関数が得られます。

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

この関数をグラフで表わすと、図 7-3 のようになります。

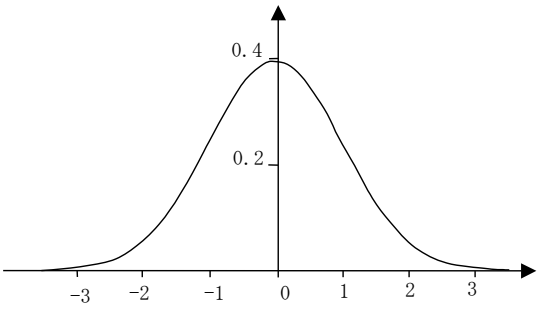

図 7-3 標準正規分布のグラフ

すぐ分かるように、この関数の最大値は、 $f(0)=1/\sqrt{2\pi}=0.3989\cdots$ です。

一般の正規分布では確率の具体的な計算を省略しましたが、ここでは確率変数 X の値 x と図 7-4 で与えられる確率 $p=P(X\le x)$ との関係を Excel によって求めてみます。

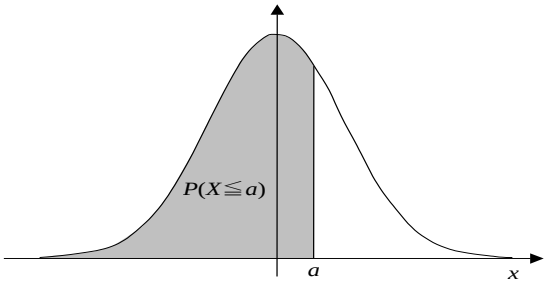

図 7-4 標準正規分布の確率

これらの関係は、以下の 2 つの関数で与えられます。

$$p=normsdist(x) \quad \Leftrightarrow \quad x=norminv(p)$$

この関数は、正規 normal、標準 standard、分布 distribution、逆 inverse という言葉の合成で名前が付けられています。具体的な計算は次の問題でやってみて下さい。

問題

標準正規分布に対して以下の確率を求めよ。

1) $P(X\le 2)$　　2) $P(X\ge 2)$

3) $P(X\ge 1)$　　4) $P(X\le -1,\ X\ge 1)$

5) $P(-1\le X\le 1)$

解答

1) $P(X\le 2)=normsdist(2)=0.97725$

2) $P(X\ge 2)=1-normsdist(2)=0.02275$

3) $P(X\ge 1)=1-normsdist(1)=0.158655$

4) $P(X\le -1,\ X\ge 1)=2\times normsdist(-1)=0.317311$

5) $P(-1\le X\le 1)=normsdist(1)-normsdist(-1)=0.682689$

7.3 正規分布の性質

7.3.1 確率の概数

正規分布は平均と分散によって分布が完全に決まる確率分布です。例えば平均 μ から標準偏差 σ 以内に含まれる確率 $P(\mu-\sigma \le X \le \mu+\sigma)$ は、μ や σ の大きさに関係なくすべて同じ大きさになります。この性質を利用して、平均から標準偏差で測って区切りの良い距離までの確率の概数を覚えておくと、おおよその確率を推測するのに便利です。区切りの良い距離としては、標準偏差の 1 倍、2 倍、3 倍がとられています。その様子を図 7-5 に表わしてみました。

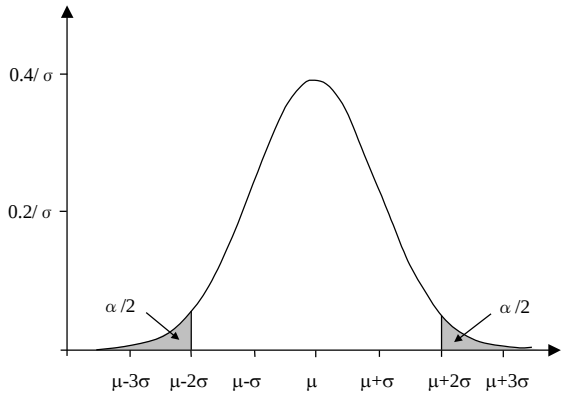

図 7-5 正規分布と確率

これらの範囲に含まれる確率及び両端の確率の合計 α の概数は以下で与えられます。

$P(\mu-\sigma \le X \le \mu+\sigma) = 0.683$ α = 0.317

$P(\mu-2\sigma \le X \le \mu+2\sigma) = 0.954$ α = 0.046

$P(\mu-3\sigma \le X \le \mu+3\sigma) = 0.997$ α = 0.003

細かいところは大変でしょうから、指定された範囲の両端の確率として以下のように覚えておきましょう。

σ までなら 32%、2σ までなら 5%、3σ までなら 0.3%

この数値はいろいろな場面で役に立つはずです。

問題

ある集団の身長の分布は、平均 170cm、標準偏差 10cm の正規分布であった。以下の確率の概数を求めよ。

1) $P(160 \le X \le 180)$ 2) $P(150 \le X \le 190)$ 3) $P(X \ge 190)$

解答

1) 100 - 32 = 68% 2) 100 – 5 = 95% 3) 5 / 2 = 2.5%

7.3.2 偏差値について [Skip OK]

ここでは、試験などでよく利用される偏差値について説明します。データの平均と分散が

$\bar{x}, s^2$ のとき、x の偏差値を以下で定義します。

$$\text{偏差値}=50+10\times\frac{x-\bar{x}}{s}$$

これは暗黙の前提として正規分布に近い分布を想定しています。x の値が平均点 $\bar{x}$ に等しいなら、試験の点数に似た得点として偏差値 50 点とします。そして、3σ 離れたら外側には 0.3%であるということから、計算式が簡単で、試験の点数風に見えるように、標準偏差の幅を 10 点となるように決めています。そうすると、偏差値の範囲は、ほぼ 20 点と 80 点の間に収まるはずです。もちろん、得点の分布は正規分布から外れることもありますので、以下で述べる順位等を考える際には、1 つの目安として偏差値を利用すべきでしょう。

この偏差値を利用すると、正規分布の場合、受験生の中での自分の位置が比較的容易に分かります。例えば、1000 人中偏差値 70 の人の場合、上位に 2σ ずれているわけですから、上側には約 2.5%の人がいます。即ち、上には 25 人程度の人がいることが分かります。具体的に以下の問題をやってみて下さい。

問題

1000 人が受験した試験の成績の分布は、平均 60 点、標準偏差 15 点の正規分布であった。A, B, C 君の点数がそれぞれ 75 点，90 点，45 点であるとき以下の問いに答えよ。

1) A 君の偏差値を求めよ。

2) B 君の偏差値を求めよ。

3) C 君の偏差値を求めよ。

4) B 君の順位はおよそ何番か。

5) C 君の順位はおよそ何番か。

解答

1) 60　　2) 70　　3) 40　　4) およそ 25 番　　5) およそ 840 番

7.3.3 標準正規分布への変換

以前 6.3 節で、確率変数の平均と分散の性質について述べましたが、ここでは確率変数が正規分布に従うときの性質について見てみましょう。

確率変数 X の平均が μ、分散が σ^2 のとき、新しい確率変数 $X'=cX+d$ の平均は $c\mu+d$、分散は $c^2\sigma^2$ で与えられることは、分布の形によらない性質でしたから、X が正規分布でももちろん成り立ちます。では、X の分布を正規分布に限るとどこが違うのでしょうか。それは、X' が平均 $c\mu+d$、分散 $c^2\sigma^2$ の正規分布になるというところです。一般の分布では、1 次式によって新しい確率変数を作った場合、新しい確率変数がどのような分布に従うか、簡単な公式はありません。しかし、正規分布の場合、変換後もやはり正規分布になるところが特徴的です。このことを記号を使って表現すると以下のようになります。

$$X\sim N(\mu,\sigma^2)\quad \text{ならば、}\quad X'=cX+d\sim N(c\mu+d,c^2\sigma^2)$$

この関係を利用すると、一般の正規分布から簡単に標準正規分布に従う確率変数を作り出すことができます。

$$X \sim N(\mu, \sigma^2) \quad \text{ならば、} \quad X' = \frac{X-\mu}{\sigma} \sim N(0,1)$$

この表式は 6.3 節の問題にもなっていました。

ここで述べた性質は、数式を使って比較的簡単に証明することができますが、積分を用いますので省略することにします。

このように正規分布する確率変数は、どんなものでも標準正規分布する確率変数に変えられることは、確率の計算の際に非常に便利です。例えば身長のデータで、平均 172cm，標準偏差 6cm の集団から 1 人選び出したとき、その人が 180cm 以上である確率を求める場合、$X' = \frac{X-\mu}{\sigma}$ の変換から、$x' = \frac{180-172}{6} = 1.333333$ とすると、この値は標準正規分布する確率変数の値に変わっています。そこで Excel の標準正規分布の確率を求める関数を利用して、以下のようになります。

$$\begin{aligned} P(X \geq 180) &= P(X' \geq 1.333333) \\ &= 1 - normsdist(1.333333) \\ &= 0.091211 \cong 0.091 \end{aligned}$$

コンピュータを利用できないとき、正規分布の確率を求めるには数表を用います。そのため殆どの統計学の教科書の巻末には正規分布の数表が付いています。しかし、表は平均や分散の大きさごとに用意することはできませんので、標準正規分布の場合の値が掲載されています。ここで述べた確率変数の性質は、すべての正規分布でこの数表が利用できることを保証しています。

問題

$X \sim N(67.2, 46.35)$ 分布のとき、以下の確率を求めよ。

1) $P(X \leq 60)$　　2) $P(X \geq 80)$　　3) $P(60 \leq X \leq 70)$

解答

1) $0.145127 \cong 0.145$　　2) $0.030046 \cong 0.030$

3) $0.514438 \cong 0.514$

7.3.4 正規分布の合成

ここでは正規分布する確率変数の和について考えます。一般の確率変数 X_1，X_2 について、平均と分散がそれぞれ、μ_1，σ_1^2 及び、μ_2，σ_2^2 で与えられるとき、新しい確率変数 $X = X_1 + X_2$ の平均と分散は、それぞれ $\mu_1 + \mu_2$，$\sigma_1^2 + \sigma_2^2$ で与えられます。これは、6.3 節で述べた一般的性質です。正規分布の場合、和を取った確率変数もやはり正規分布になるというところが重要です。これを記号を用いて表わしてみましょう。

$X_1 \sim N(\mu_1, \sigma_1^2)$，$X_2 \sim N(\mu_2, \sigma_2^2)$ のとき、

$$X = X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

正規分布する確率変数はいくつ足してもやはり正規分布します。

問題

互いに独立な確率変数 X_1，X_2 が、$X_1 \sim N(10,9)$ 分布，$X_2 \sim N(7,16)$ 分布であるとするとき、以下の確率変数 X の分布を求めよ。

1) $X = X_1 + X_2$　　2) $X = 2X_1 + X_2$　　3) $X = X_1 - X_2$

解答

1) $X \sim N(17,25)$　　2) $X \sim N(27,52)$　　3) $X \sim N(3,25)$

問題

互いに独立な確率変数 X_i（$i = 1,2,\cdots,n$）が、それぞれ $N(\mu, \sigma^2)$ 分布に従うとき、以下の変数の分布を求めよ。

$$\bar{x} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

解答

$$\bar{x} \sim N(\mu, \sigma^2/n)$$

問題

ある商品の製造は３つの工程からなり、各工程に要する日数は、以下のような正規分布に従うとする。

	平均	標準偏差
第１工程	3	1
第２工程	10	3
第３工程	5	2

1) 完成までに要する時間の平均と標準偏差を求めよ。

2) 納期を 20 日とするとき、納期に遅れる確率を求めよ。

解答

1) 平均 18 日，標準偏差 3.742 日（分散 14 日 ^{2}）

2) $x = \dfrac{20-18}{3.742} = 0.534522$　より、

$p = 1 - normsdist(0.534522) = 1 - 0.296449 = 0.29$

7.3.5 中心極限定理 [Skip OK]

正規分布に関する性質として、最後に最も重要で利用範囲の広い中心極限定理と呼ばれるものについて説明します。これは簡単に言うと、どんな分布の確率変数でも十分多くの平均

を取ると、その平均の分布は正規分布になるという驚くべき定理です。このことが、これまで正規分布を統計の基本と言ってきた理由であり、正規分布の重要性を示す性質です。以下にこの定理を書いておきましょう。

中心極限定理

独立な確率変数、$X_i\ (i=1,2,\cdots,n)$ が、平均μ_i、分散σ_i^2の一般的な確率分布に従うとき、容易に満たされるある条件のもとで以下となる。

$$\lim_{n\to\infty}\sum_{i=1}^{n}(X_i-\mu_i)\Big/\sqrt{\sum_{i=1}^{n}\sigma_i^2}\sim N(0,1)$$

まず、確率変数X_iについて、$X_i-\mu_i$にすると平均は0、分散はもとのとおりのσ_i^2になります。それを合計した$\sum_{i=1}^{n}(X_i-\mu_i)$については、平均が$0$、分散が$\sum_{i=1}^{n}\sigma_i^2$になります。さらに、$\sum_{i=1}^{n}(X_i-\mu_i)\Big/\sqrt{\sum_{i=1}^{n}\sigma_i^2}$とすると、平均が$0$、分散が$1$になります。これは一般的性質です。中心極限定理はここからが重要で、このnを十分大きくすると、これが正規分布になるというところです。

もう少し実用的な表示法を考えてみましょう。独立な確率変数$X_1, X_2, \cdots, X_n$から、新しい確率変数として$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$を作ります。一般的性質として、確率変数$\overline{X}$の平均は$\frac{1}{n}\sum_{i=1}^{n}\mu_i$、分散は$\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2$となることは容易に分かると思います。中心極限定理はnを十分大きくすると、$\overline{X}$が正規分布になるというところです。

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i=\frac{1}{n}(X_1+X_2+\cdots+X_n)\underset{n\to\infty}{\sim}N\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i,\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2\right)$$

直感的に理解し易い特別な場合として､各確率変数の平均と分散が等しい場合を考えてみましょう。標本のn個のデータの平均を求めるときが、これに相当します。

独立な確率変数$X_1, X_2, \cdots, X_n$が、平均μ、分散σ^2の確率分布に従うとき、確率変数$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$の平均は $n\cdot\frac{\mu}{n}=\mu$、分散は $n\cdot\frac{\sigma^2}{n^2}=\frac{\sigma^2}{n}$、標準偏差は $\frac{\sigma}{\sqrt{n}}$ となります。

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i=\frac{1}{n}(X_1+X_2+\cdots+X_n)\underset{n\to\infty}{\sim}N(\mu,\sigma^2/n)$$

この確率変数$\overline{X}$は標本平均を表わしています。実験データで、n個の測定データの平均を取って1つのデータとするとき、このばらつきの統計量（標準誤差）としてここで与えた標

本平均の標準偏差が用いられます。

以上のことから、たくさんのデータの平均を取るという操作には 2 つの意味があることが分かります。1 つは一般的な性質として、分散の値がデータの個数に反比例して小さくなり測定の精度が上がるということ、もう 1 つは分布の形の分からないデータでも平均化したものは性質が完全に分かっている正規分布に従うということです。後者こそが中心極限定理の本質です。最後に、ここで述べたたくさんのデータというのはどの程度でしょうか。データの分布にもよりますが、6 個程度の平均でもかなり正規分布に近付くようなものもあります。

問題

資料の重さ(mg)を 10 回測定したところ、測定誤差があり以下の結果を得た。平均と標準偏差を求め、それから平均の標準偏差（標準誤差）を求めよ。

71.5, 71.3, 70.8, 71.1, 70.9, 71.2, 71.4, 71.5, 70.9, 71.3

解答

平均　71.19,　　標準偏差　$0.255821 \cong 0.2558$,

平均の標準偏差　$0.080898 \cong 0.08090$

問題

1 つの処理に平均 3.54 分、標準偏差 0.47 分かかるとする。同じ処理を 10 回繰り返すとき、38 分以上かかる確率を求めよ。

解答

10 回の処理で、平均 35.4 分、分散 2.209 、標準偏差 1.486271

$$x = \frac{38 - 35.4}{1.486271} = 1.749345$$

$$p = 1 - normsdist(1.749345) = 0.040116 \cong 0.040$$