２章　確率分布と検定

2.1 確率密度関数

量的データの集計方法としてヒストグラムを描くということを前章で学びましたが、ここではデータの数を十分大きくしていった場合のヒストグラムの形を考えます。図のようにデータ数を多く取って行くと、ヒストグラムはきめが細かくなり、ヒストグラムの上端を繋いだグラフは次第に滑らかになって行き、ある形に近づいて行きます。この形を変数の確率密度関数または単に密度関数と呼びます。但し、縦軸の度数はこの関数が囲む面積が 1 になるような数値に書き換えています。

図 2.1.1　ヒストグラムの極限と密度関数

例えばテストの点数を考える場合、以下の図のような位置に 40 点と 60 点があるとしましょう。確率密度関数のその間にある領域の面積を S とすると、S は 40 点から 60 点の間にいる人の割合になります。

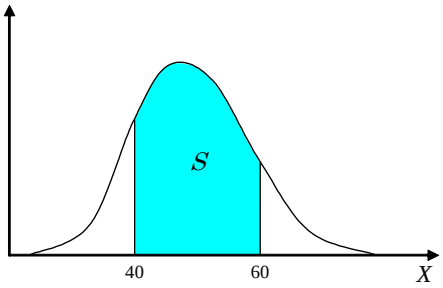

図 2.1.2　面積と確率

もう少し言い換えるとこの集団の中から一人選ぶとすると、40 点から 60 点の間の人を選ぶ確率は S になります。もちろん割合または確率ですので、この面積は 0 以上 1 以下です。またこれを 0 点から 100 点までとしますと、全員その中に入りますので、割合または確率は 1 になります。確率密度関数が囲む全面積が 1 になるように縦軸を書き換えていると言った理由は全確率が 1 になるように設定するためです。

よく「ある変数の分布は？」という言い方をしますが、これはこの確率密度関数がどんな形かということです。以後代表的な分布について見て行きましょう。

2.2 正規分布と標準正規分布

統計学の基本的な分布は正規分布（normal distribution）と呼ばれる分布です。これはデータの測定誤差などランダムな現象から生じる分布です。また、正規分布以外の分布からで

も、ある程度大きな数（値ではない）のデータの平均値は正規分布することが知られています。これは中心極限定理と言って統計学では大変重要な定理です。この節ではこの正規分布について見て行きましょう。

正規分布は平均と分散で確率密度関数の形が完全に決まる分布です。ある量 X が平均 μ（ミュー）、分散 σ^2（シグマ²）の正規分布をしていることを $X \sim N(\mu,\sigma^2)$ と表します。正規分布していることを正規分布に従う、とも言います。この正規分布の確率密度関数は以下のようにきれいな富士山形をしています。

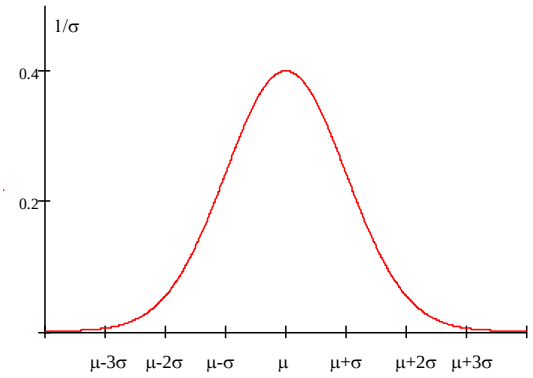

図 2.2.1　$N(\mu,\sigma^2)$ の確率密度関数

正規分布には平均と標準偏差の値がどのように変わってもこの形を維持するという面白い性質があります。すなわち平均と分散の値に関わらず以下の性質があることが示されています。

$P(\mu-\sigma \le X \le \mu+\sigma) = 0.683$　　両側　32%

$P(\mu-2\sigma \le X \le \mu+2\sigma) = 0.954$　　両側　5%

$P(\mu-3\sigma \le X \le \mu+3\sigma) = 0.997$　　両側　0.3%

ここに $P(a \le X \le b)$ は変数 X が a から b の値を取る確率を表します。例えば、ある試験の結果が、平均 65 点、標準偏差 8 点の正規分布をする場合、一番上の式を利用すると、57（=65-8）点から 73（=65+8）点までの間に約 68%の人がいることになります。この値は非常に重要で、特に両側確率の 32%, 5%, 0.3%という概数は統計を学ぶ際にはぜひ覚えておくべきでしょう。

正規分布の中で特に重要なものがあります。それは平均 0、分散 1 の正規分布です。この分布の確率密度関数は以下の図のような形をしており、計算機が発達する前から細かい範囲で確率を与える表が作られていました。

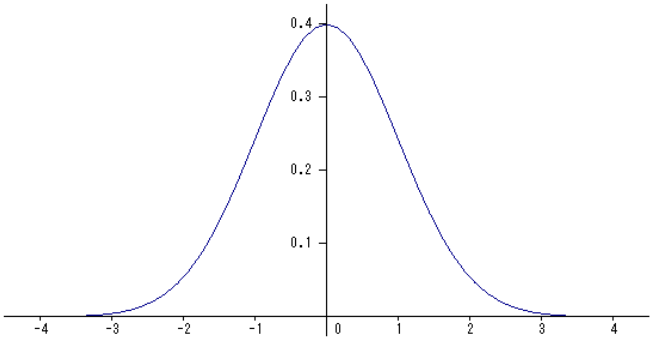

図 2.2.2　標準正規分布の確率密度関数

しかし、実際の分布が都合よく標準正規分布になるということはまずないのに、なぜ詳しく調べられたのでしょうか。それは正規分布に以下の性質があるからです。

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

これだけでは何のことか分からないと思いますので、前に使った試験の点数の分布 $N(65, 8^2)$ を例に上の式の意味を考えてみます。例えばこの試験で 70 点以上の人の確率（割合） $P(X \geq 70)$ を計算しようとしても、当然一般の正規分布では細かな確率は与えられていませんのでそのままでは不可能です。そこで上の式を利用して、変数 X を X' に変換します。この場合 $\mu = 65$、$\sigma = 8$ ですから、変換の式は以下のようになります。

$$X' = \frac{X-65}{8} \sim N(0,1)$$

この式を使って $X \geq 70$ を X' の関係に直すと

$$X' = \frac{X-65}{8} \geq \frac{70-65}{8} = 0.625$$

$X' \geq 0.625$ となります。この X' は標準正規分布に従うので、細かい確率も計算できます。すなわち結果は以下のようになります。

$$P(X \geq 70) = P(X' \geq 0.625) = 0.2660$$

このようにして標準正規分布の確率を詳しく求めるということは、一般の正規分布の確率を詳しく求めると同じであることが分かりました。

一般の分布の場合、上の変換は単に平均を 0 にして分散を 1 にする変換で、正規分布のときだけ変換後も正規分布であり続けるところが大きな特徴です。このようにして変数 X を変数 X' に変換することを標準化と呼び、後で学ぶ多変量解析ではよく利用されます。

正規分布の性質は他にもいろいろありますが、興味のある人は「基礎からの統計学」を参照して下さい。

問題（１個のデータについて）

体重の平均 20kg、標準偏差 2kg（分散 4kg^2）の子供 1000 人の集団がある。データは正規分布するとして以下の問いに概数（大体の値）で答えよ。

１）22kg の子供は重い方から大体何%か　　［　　］%
２）24kg の子供は重い方から大体何%か　　［　　］%
３）24kg の子供は重い方から大体何番目か　［　　］番目
４）18kg の子供は重い方から大体何%か　　［　　］%
５）18kg の子供は重い方から大体何番目か　［　　］番目

問題（１個のデータについて）

前の問題で、子供の体重から平均の 20kg を引き、その結果を標準偏差の 2kg で割るとする。以下の問いに答えよ。

１）20kg の子供の値はいくらになるか　　［　　］
２）22kg の子供の値はいくらになるか　　［　　］
３）17kg の子供の値はいくらになるか　　［　　］
４）この計算結果は平均［　　］、分散［　　］の正規分布になる。
５）３）について、<u>17kg 以下</u>となる正確な確率を求めよ。
　Excel の関数 normsdist(x)を用いると　　［　　］
６）２）について、<u>22kg 以上</u>となる正確な確率を求めよ。
　Excel の関数 normsdist(x)を用いると　　［　　］
７）17kg 以上、22kg 以下となる正確な確率を工夫して求めよ。
　Excel の関数 normsdist(x)を用いると　　［　　］

解答（１個のデータについて）

１）22kg の子供は重い方から大体何%か　　［16］%
２）24kg の子供は重い方から大体何%か　　［2.5］%
３）24kg の子供は重い方から大体何番目か　［25］番目
４）18kg の子供は重い方から大体何%か　　［84］%
５）18kg の子供は重い方から大体何番目か　［840］番目

解答（１個のデータについて）

１）20kg の子供の値はいくらになるか　　［0］
２）22kg の子供の値はいくらになるか　　［1］
３）17kg の子供の値はいくらになるか　　［-1.5］
４）この計算結果は平均［0］、分散［1］の正規分布になる。
５）３）について、<u>17kg 以下</u>となる正確な確率を求めよ。
　=normsdist(-1.5)　　［0.066807］
６）２）について、<u>22kg 以上</u>となる正確な確率を求めよ。
　=1-normsdist(1)　　［0.158655］
７）17kg 以上、22kg 以下となる正確な確率を工夫して求めよ。
　=normsdist(1)-normsdist(-1.5)　　［0.774538］