2009 年度前期　データ解析序説　第１１回

正規分布とは

中心極限定理と正規分布モデル

世の中には，さまざまなランダム現象があります．それらが「どういう理由で，どのようにランダムか」を数式で表しているのが，前回説明した確率分布モデルです．しかし，「どういう理由で，どのようにランダムか」が明確にわかる場合というのは，そう多くはありません．ところが，世の中の広い範囲のランダム現象を，ある１種類の確率分布モデルで表すことができる，という定理があります．これが**中心極限定理**で，その確率分布モデルを**正規分布モデル**といいます．

中心極限定理とは，簡単に言うと「あるランダム現象のランダムさの原因が，無数の独立なランダム現象の合計になっているときは，そのランダム現象は概ね正規分布モデルであらわせる」ということです．この場合，「無数の独立なランダム現象」は，どんなものであってもかまいません[1]．

例えば，物の長さを測定するときの測定誤差は，測定するごとに異なるランダム現象となります．しかし，定規の熱による伸び縮み，人間の目の限界，定規を見るときの空気の乱れ，などなど，無数の独立な原因による誤差の合計になっていますから，測定誤差の確率分布は正規分布になります．あるいは，電気回路の雑音は，回路中の金属の原子が熱によって独立に振動し，その合計として現れます．ですから，雑音の瞬間瞬間の強さは正規分布にしたがいます．このように，独立な無数の原因の合計になって現れるランダム現象はたくさんありますから，正規分布モデルで表せる分布は，自然科学，社会科学の分野を問わず，世の中に無数に見つけることができます．

正規分布モデルのパラメータは，期待値と分散です．つまり，期待値と分散がわかれば，「確率変数がある値からある値の範囲にある確率が，いくらになるか」という計算をすることができます．しかも，実際にはいちいち計算する必要すらなく，後半で説明するように，すでに計算結果が書いてある数表を使うことで，簡単に確率を知ることができます．なお，「期待値がμ，分散がσ^2」である正規分布を$N(\mu,\sigma^2)$と書きます．

正規分布モデルの計算

では，正規分布モデルにしたがう確率変数がある範囲の値になる確率を，数表を使って求める方法を説明します．前回も述べましたが，正規分布モデルのパラメータは期待値と分散で，確率変数Xの確率分布が期待値μ，分散σ^2の正規分布であることを，「確率変数Xは正規分布$N(\mu,\sigma^2)$にしたがう」あるいはさらに短く「$X \sim N(\mu,\sigma^2)$」と書きます．正規分布の確率密度関数のグラフは図１のようになります．期待値μをとる確率密度がいちばん高く，左右対称に広がっています．

正規分布には，次の大変重要な性質があります[2]．

１．確率変数Xが期待値μ，分散σ^2の正規分布$N(\mu,\sigma^2)$にしたがうとき，確率変数$(X-\mu)/\sigma$は正規分布$N(0,1)$にしたがう．

---

[1]本当は，まったくどんなものでもよいというわけではなく，いろいろ制約がありますが，省略します．くわしくは，2008 年度後期「情報統計学」の第 12 回の講義録を参照してください．

[2]証明は，2008 年度後期の浅野の講義「情報統計学」の第６回の講義録をネットで参照してください．

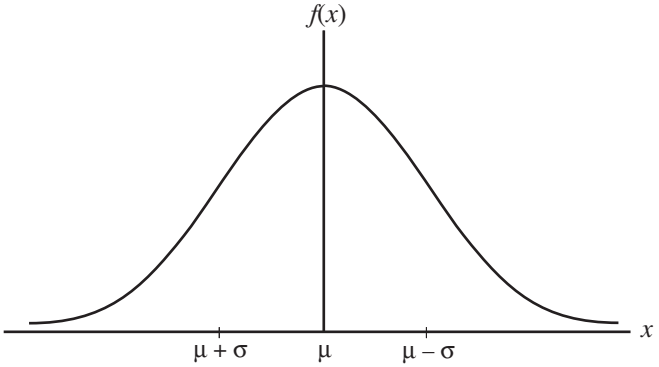

図 1: 正規分布の確率密度関数

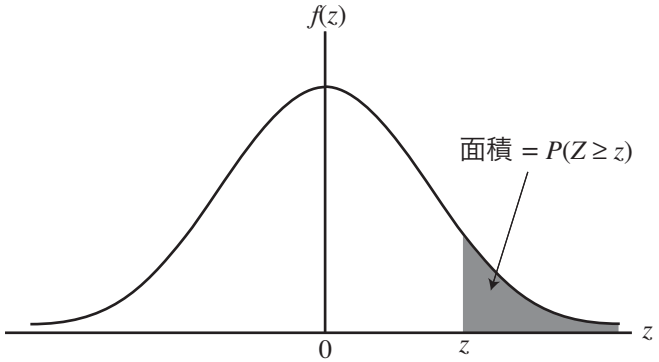

図 2: 標準正規分布の確率密度関数のグラフ上で「確率変数 Z が値 z 以上である確率」$P(Z \geq z)$

この $N(0,1)$ を**標準正規分布**といいます．「確率変数 $(X-\mu)/\sigma$」とは，確率変数 X の「すべての可能な値」について，いずれも μ をひいて σ で割るという操作を行なって，新しい確率変数を作ったものです．この計算は，第 3 回の講義で説明した，度数分布についての「標準得点」の計算と同じものです．この性質を，この講義では以後「正規分布の性質 1 」とよぶことにします．

2．$X_1, ..., X_n$ が独立で，いずれも正規分布 $N(\mu, \sigma^2)$ にしたがうならば，それらの平均 $\overline{X} = (X_1 + ... + X_n)/n$ は $N(\mu, \sigma^2/n)$ にしたがう

この性質は，次のように考えることができます．前回の講義で，母集団と標本について説明しました．そこで，母集団分布が正規分布 $N(\mu, \sigma^2)$ であるとしましょう．そのとき，無作為抽出される標本は，ランダムな数値である確率変数であり，母集団分布と同じ正規分布 $N(\mu, \sigma^2)$ という確率分布にしたがいます．

では，標本としてデータひとつを取り出すのではなく，n 個のデータを取り出すことにしましょう．「標本となるデータの数」を，統計学では標本の**サイズ**といい，ここではサイズが n となります．また，データの数が多いことを「標本サイズが大きい」といいます．

標本として取り出される n 個のデータを $X_1, ..., X_n$ とし，各々が無関係に無作為抽出で選ばれるとします．このとき，$X_1, ..., X_n$ は**独立**となります[3]．

[3]統計学でいう「独立」の正確な意味は，この講義では扱っていません．くわしくは，2008 年度後期「情報統計学」第 2 回

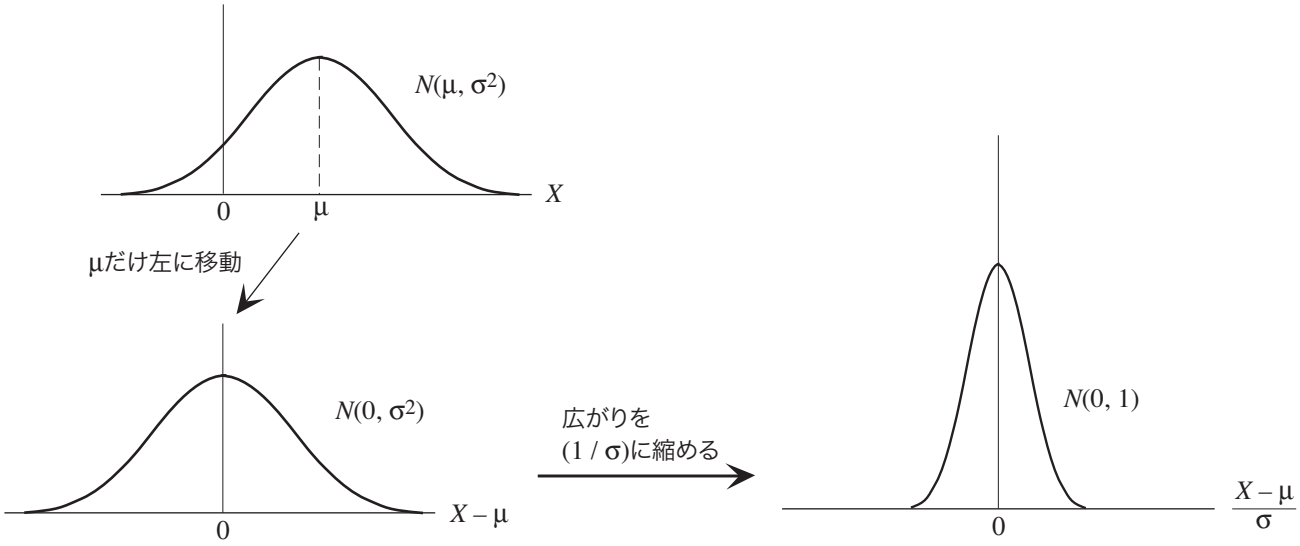

図 3: 正規分布の性質 1

このとき，n 個のデータの平均 $(X_1 + ... + X_n)/n$ を**標本平均**とよび，$\bar{X}$ で表します．標本として取り出される各々のデータは，ランダムな数値ですから，それらをいくつか集めて計算した標本平均もまたランダムです．そこで，標本の期待値・分散ではなく，標本平均自体の期待値・分散を考えてみましょう．

図 4 のように，標本の各データにたとえ極端に母平均からかけはなれた値があっても，いくつか集めて平均すると，その極端な値は相殺されてしまいます．ですから，標本平均は，極端な値になる可能性が小さくなります．したがって，標本平均の分散は母分散より小さくなります[4]．

この性質は，それだけではなく，標本平均の期待値が母平均と同じであること，標本平均の確率分布もやはり正規分布になることを述べています[5]．この性質を，この講義では「正規分布の性質２」とよぶことにします．

ここでいう n，すなわち標本サイズが大きくなると，標本平均の分散が小さくなります．ということは，母集団の分散が大きくても，データをたくさん集めれば，標本平均はそうばらついた値になる可能性は少ないので，いま 1 回だけ集めた標本から計算した標本平均も，おそらく母平均に近い，ということになります．このことは，「くじを 1 本だけひいても当たる確率は全くわからないが，くじをたくさんひけば当たる確率はだいたいわかってくる」という，先に述べたごく当たり前の事実に対応しています．

正規分布のこれらの性質は，次回説明する「区間推定」をはじめとする統計的推測の技法において，大変重要な意味を持ちます．

正規分布の数表の見方

「標準正規分布にしたがう確率変数が，ある範囲の値をとる」確率は，数表から簡単に知ることができます．配布した数表は「標準正規分布にしたがう確率変数 Z がある値 z 以上である確率」$P(Z \geq z)$ を

を参照してください．

[4]それがなぜ $1/n$ になるのかは，2008 年度後期「情報統計学」第 4 回を参照してください．

[5]標本平均の確率分布がやはり正規分布になることを**正規分布の再生性**といいます．これらの証明は，2008 年度後期「情報統計学」第 6 回の講義録（付録）を参照してください．

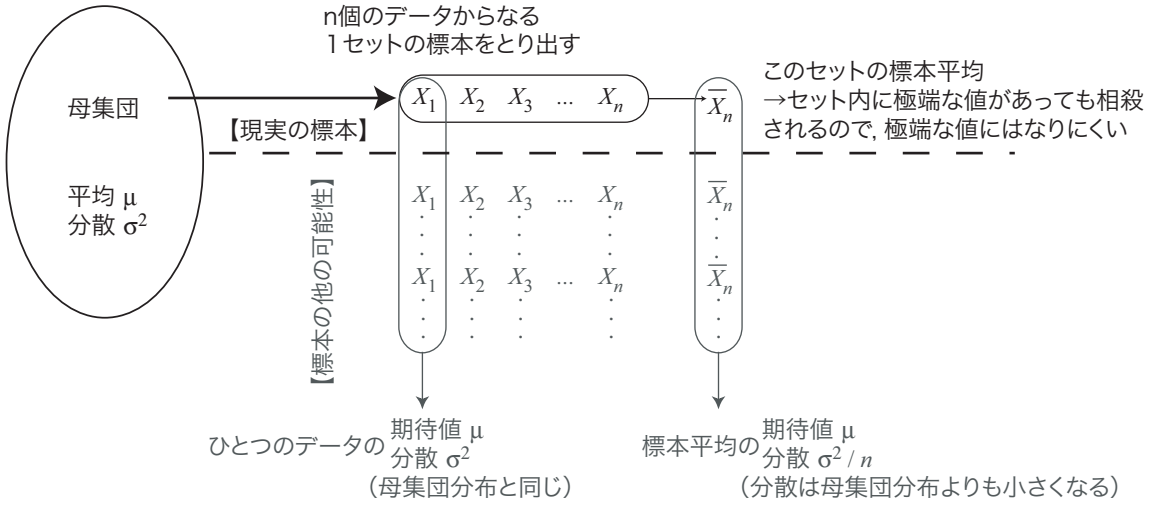

図 4: 標本平均のしたがう確率分布

計算したもので，確率密度関数のグラフにおいては図 2 のグレーの部分の面積になります．標準正規分布の確率密度関数は $z = 0$ に対して左右対称なので，数表は $z \geq 0$ についてのみ掲載されています．

さきほどの「正規分布の性質 1 」を使うと，期待値・分散がどんな値の正規分布でも，それにしたがう確率変数 X がある値 x 以上である確率を，この数表だけで求めることができます．例えば，期待値 50，分散 10^2 である正規分布 $N(50, 10^2)$ にしたがう確率変数 X が 60 以上である確率，すなわち $P(X \geq 60)$ を求めてみましょう．$Z = (X - 50)/10$ のように変換すると，性質 1 から確率変数 Z は標準正規分布 $N(0, 1)$ にしたがいます．また，$X = 60$ のとき $Z = (60 - 50)/10 = 1$ ですから，求める確率は $P(Z \geq 1)$ です．数表から，$P(Z \geq 1) = 0.15866$ であることがわかります．