2016 年度秋学期　統計学　第１１回

分布の「型」を考える ー 確率分布モデルと正規分布

確率分布モデル

前回の講義の最後に，標本平均は母平均の推測結果としてもそうおかしくはない，ただし間違った推測をしてしまう可能性はある，という説明をしました。

では，間違った推測をしてしまう可能性，つまり推測の不確かさを，確率の形で計算するには，どうすればよいでしょうか？ 計算をするには，問題が式と数で表されていなければなりません。しかし，この計算のてがかりとなる母集団分布（＝標本がしたがう確率分布）は，数字の集まりにすぎず，式ではあらわされていません。これでは，計算はできません。

そこで用いるのが，第 7 回の回帰分析のときにも出てきた「モデル」の考え方です。ここでは，母集団分布（＝標本がしたがう確率分布）が，ある数式で表されるものだと仮定してしまいます。いわば，母集団分布のヒストグラムが，ある式であらわされる関数のグラフになっていると考えるのです。この数式を**確率分布モデル**といい，母集団のなりたちに合わせていろいろなものが考えられています。

関数を式で表す場合，そのグラフの大きさや位置を調整する数字があります。この数字は，回帰分析のときにもでてきた**パラメータ**です。回帰分析のとき，回帰直線を $y = a + bx$ という式で表したとき，直線は直線でもどういう直線であるかをパラメータ a や b で表し，これらを最小２乗法で求めました。確率分布モデルにも，同じようにパラメータがあります。これから，確率分布モデルのパラメータを標本から推測する方法を説明していきます。

連続型確率分布

ここまでの講義で，度数分布→確率分布→確率変数という順に進めてきた説明では，連続した数値の区間である階級を，ひとつの数値で代表する「階級値」というものをわざわざ導入することで，確率変数は 172.5cm のつぎは 177.5cm というように「とびとび」の値をとる，と考えてきました。

しかし，数学では，とびとびの値をとる数式よりも，連続なグラフになるような数式のほうがずっと簡単なはずです。ですから，ヒストグラムを数式で表すために，「とびとびではなく連続的な値をとる」確率変数というのは考えられないのでしょうか。

そこで，ヒストグラムの説明のところで述べた，「ヒストグラムの柱は，分割することができる」という性質を使います。ヒストグラムの各柱をどんどん細かく分割することで，階級の区切りかたを「十分に」細かくしたとします。このような確率分布は，値がとびとびにならない，「ある範囲内のどんな値にでもなることができる」確率分布と考えることができます（図 2)。このような確率分布を**連続型確率分布**といい，これに対し，確率変数が（階級に区切ったのではなく）本来とびとびの値（例えば，くじびきの当たり回数）になるような確率分布を**離散型確率分布**といいます。

連続型確率分布では，確率変数が「ある 1 つの値」をとる確率ではなく，「ある範囲の値」をとる確率を考えます。離散型確率分布で確率変数が「ある範囲の値」をとる確率は，確率変数のある範囲内の値に対応する確率を合計したものです。ヒストグラム上でこれを見ると，ある範囲内にある「柱」の面積を合計したものになります（図 2 の左)。「ヒストグラムで度数を表しているのは柱の高さではなく柱の面積」であるからです。

これを，階級の区切りが見えないほど細かくなったヒストグラムで考えると，柱の境目は見えなくなっ

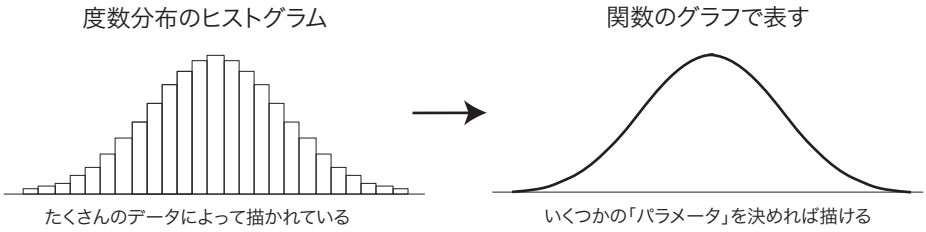

図 1: 確率分布モデル

ているので，灰色の部分の面積がそれに相当します（図 2 の右)。この面積は，数学では「『ヒストグラムの上端をつないだグラフで表される関数』の『ある範囲』での積分」といいます。この「ヒストグラムの上端をつないだグラフで表される関数」を**確率密度関数**といいます。

この講義では積分の計算をすることはありませんが，特定の確率分布モデルでの積分の値を計算してまとめた数表はよく用います。

「確率変数がある範囲の値に入る確率」
＝「確率密度関数のグラフの下の部分のうち，この範囲にあたる部分の面積」
（＝「確率密度関数のこの範囲での積分」）

という関係は，今後の講義でよく出てきますので，よく理解してください。

確率密度関数は確率変数がとりうる各値の「現れやすさ」を表してはいますが，確率そのものではないことに注意してください。「連続型確率変数がある 1 つの値をとる確率」は，確率密度関数の値ではありません。「連続型確率変数がある 1 つの値をとる確率」は，範囲の幅が 0 ですからその範囲に対応するグラフの下の部分の面積も 0 で，すなわち 0 であることに注意しましょう。また，グラフの下の部分全体の面積は，「確率変数の値が，とりうる値の範囲全体のどこかにある確率」ですから 1 (100%) となります。

ところで，現実のデータは，必ず何桁かの数字で表されるわけですから，どんなに細かく表現してても必ず「デジタル」，すなわち「とびとび（離散的)」です。連続型確率分布というのは，あくまで数式で表しやすくするための手段だと考えてください。

正規分布モデル

代表的な連続型確率分布モデルである**正規分布モデル**は一番応用範囲の広い確率分布モデルで，世の中には正規分布モデルであらわせるような母集団分布がたくさんあります。これは，**中心極限定理**という定理があるからです。中心極限定理とは，ひとことでいうと「母集団のデータが分布している（ばらついている）原因が，無数の独立な原因によるデータの分布の合計になっているときは，母集団分布は概ね正規分布になる」ということです。

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 3 のようになります。期待値 μ をとる確率密度がい

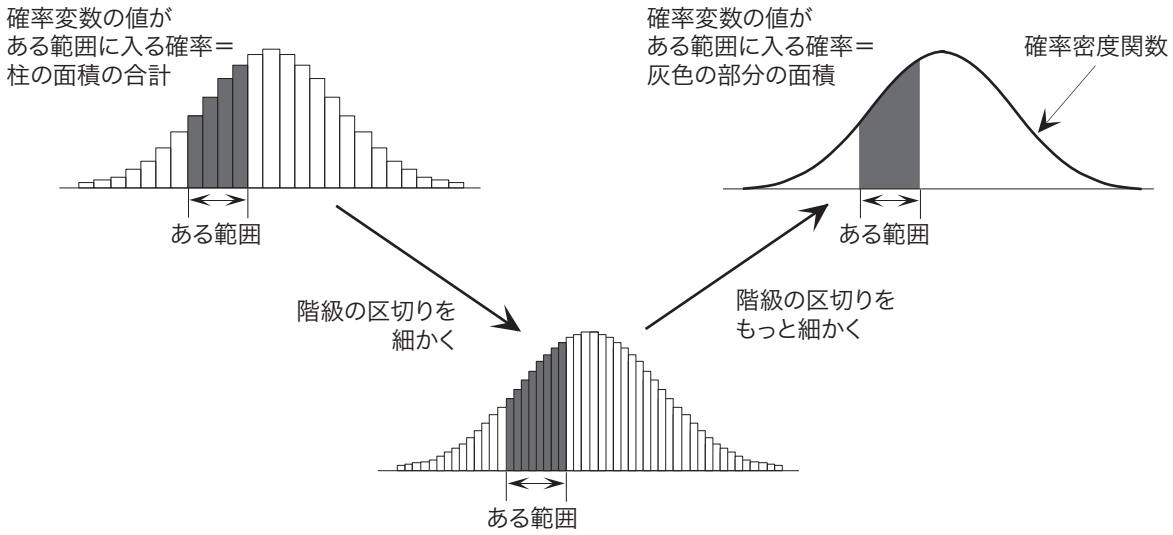

図 2: 連続型確率分布

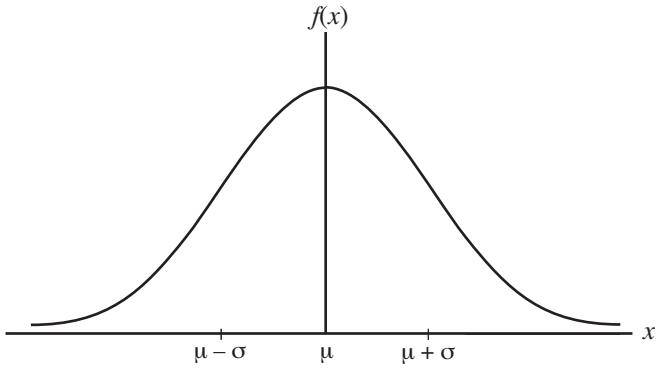

図 3: 正規分布の確率密度関数

ちばん高く，左右対称に広がっています。また，グラフの中央部の広がりが，標準偏差 σ に対応しています[1]。なお，期待値から離れると確率密度は小さくなりますが，0 にはなりませんので，グラフは左右とも無限に広がっています。

図 4 は，期待値 0 で標準偏差が 0.5, 1.0, 1.5 の場合に，正規分布の確率密度関数を描いたものです。標準偏差が大きくなるほど，グラフの中央部の広がりが大きくなり，そのぶん山の高さが低くなります。

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。元の確率変数 X の期待値が μ，分散が σ^2

[1]正確にいうと，横軸の $\mu+\sigma, \mu-\sigma$ の位置が，グラフの変曲点にあたります。
[2]くわしくは，私の講義「解析応用」第 3 部を参照してください。

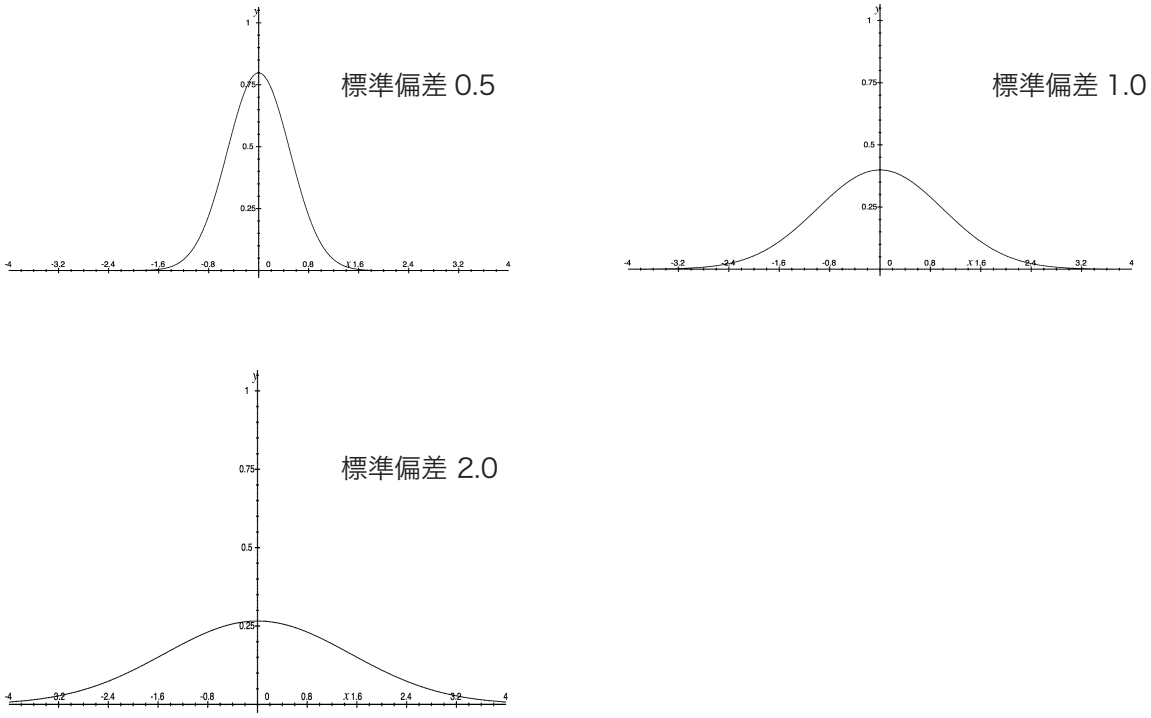

図 4: いろいろな標準偏差の場合の正規分布

のとき，確率変数 $(X-\mu)/\sigma$ の期待値が 0，分散が 1 になることは，第 5 回の「標準得点」のところで説明した通りです。ここで述べていることは，それだけでなく，**元の確率変数 X が正規分布にしたがうならば，変換後の確率変数もやはり正規分布にしたがう**，ということです。この $N(0,1)$ を，**標準正規分布**といいます。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。図 5 で，この操作を確認してください。

> **2．$X_1,\ldots,X_n$ 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 $(X_1+\cdots+X_n)/n$ は $N(\mu,\sigma^2/n)$ にしたがいます[3]。**

「$X_1,\ldots,X_n$ が独立で，いずれも正規分布 $N(\mu,\sigma^2)$ にしたがう」という条件にあてはまるのは，$X_1,\ldots,X_n$ が，正規分布 $N(\mu,\sigma^2)$ にしたがう母集団分布から無作為抽出された標本であるときです。このとき，$(X_1+\cdots+X_n)/n$ は標本平均です。第１０回の講義で，標本平均の期待値は μ，分散は σ^2/n であると述べました。ここで述べているのは，それだけでなく，**母集団分布が正規分布であるならば，標本平均もやはり正規分布にしたがう**，ということです。この性質を，この講義では以後「正規分布の性質 2」とよぶことにします。

正規分布の数表の見方

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求めるには，積分の計算をしなければなりません。しかし，実際にはその必要はなく，数表を使って求めることができます。

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」$P(Z \geqq z)$ を計算したもの

[3] これを**正規分布の再生性**といいます。

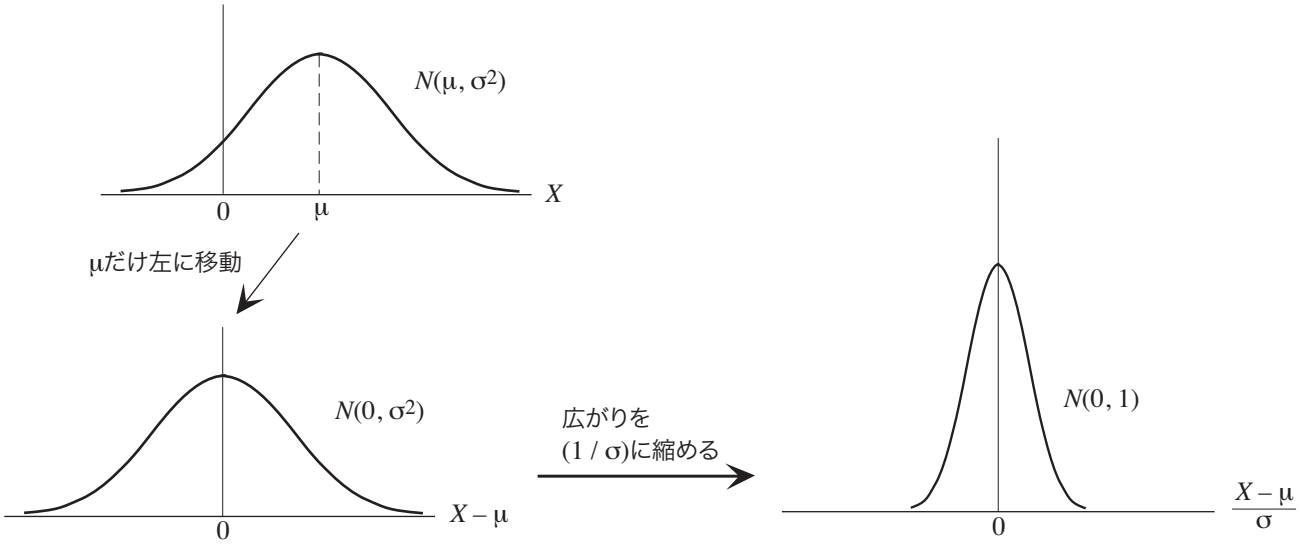

図 5: 正規分布の性質 1

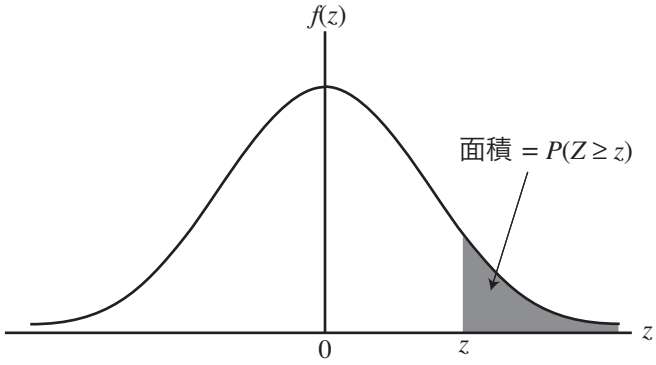

図 6: 標準正規分布の確率密度関数のグラフ上で「確率変数 Z が値 z 以上である確率」$P(Z \geqq z)$

で，確率密度関数のグラフにおいては図 6 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 $z \geqq 0$ 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 $P(X \geqq 60)$ 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 $P(Z \geqq 1)$ です。数表から，$P(Z \geqq 1) = 0.15866$ であることがわかります。

今日の演習

確率変数 X が正規分布 $N(50, 10^2)$ にしたがうとき，(1) $P(X \geqq 55)$，(2) $P(45 \leqq X \leqq 60)$，をそれぞれ求めてください。