2008 年度前期　統計データ解析Ａ　第５回

20 本くじをひいて，**6** 本しか当たらない確率 **(2)** ー 連続型確率分布と正規分布の計算

連続型確率分布

前回の例では，くじを５回ひいたときの当たり回数の分布を使って，ヒストグラムを説明しました．このときの例での確率変数は「当たる回数」で，それは０，１，２，３，４，５の６通りだけでした．しかし，前回の最後に述べたように，正規分布モデルで表される例としてあげた「測定誤差」などでは，確率変数の値はさまざまな数値になり，６通りに限られる，などということはありません．では「とびとびではなく連続的な値をとる」確率変数の場合，ヒストグラムはどのように描けばよいのでしょうか．

ここで，ヒストグラムの横軸を考えてみてください．前回の例では，横軸は６通りのとびとびの数字を表していました．これを，横軸が例えば「測定誤差」の数値を表していると考えてください．そうすると，ヒストグラムの柱の幅は，横軸の数値の「刻み」(図１の例では「0.1 刻み」）になります．この刻みを階級といいます．

そこで，ヒストグラムの階級の区切りかたをものすごく細かくしたとしましょう．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，２項分布のような，確率変数がとびとびの値（例えば，当たり回数）になる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．

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．それなのにわざわざ「連続型」というものを考えるのは，確率分布モデルは数式で表されるからです．数学では，とびとびの値をとる数式よりも，

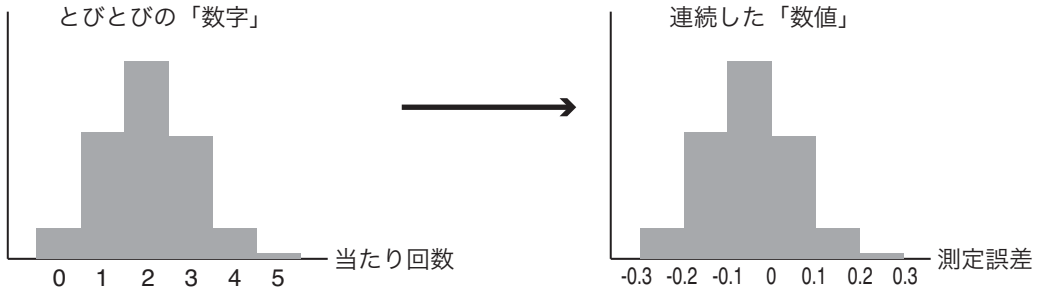

図 1:「数字」から「数値」へ

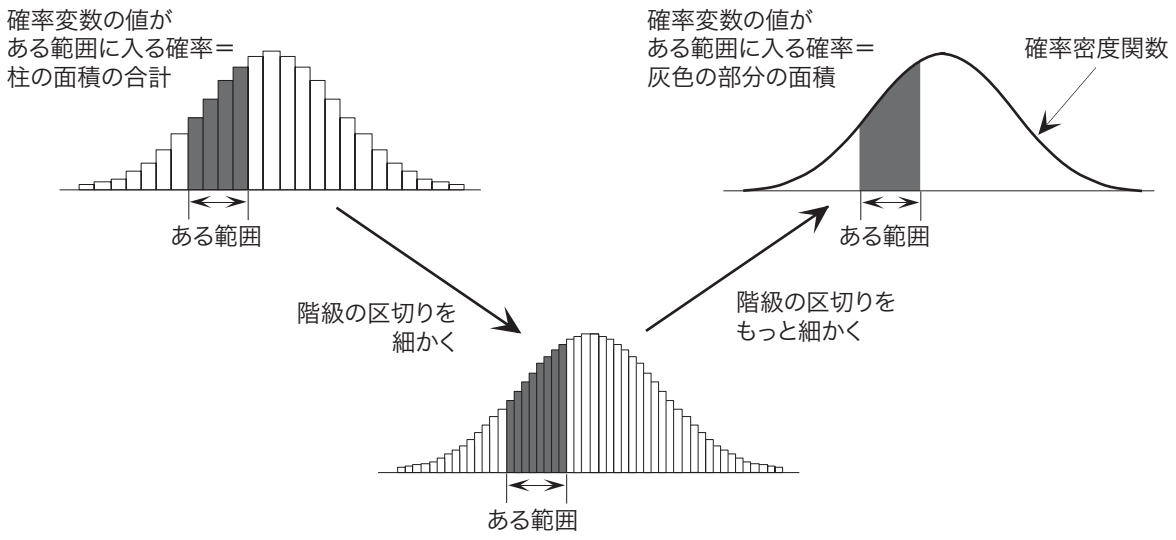

図 2: 連続型確率分布

連続なグラフになるような数式のほうがずっと簡単なのです．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．

正規分布モデルの計算

いろいろ準備をしましたが，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 3 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[1]．

[1]証明は，2006 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．

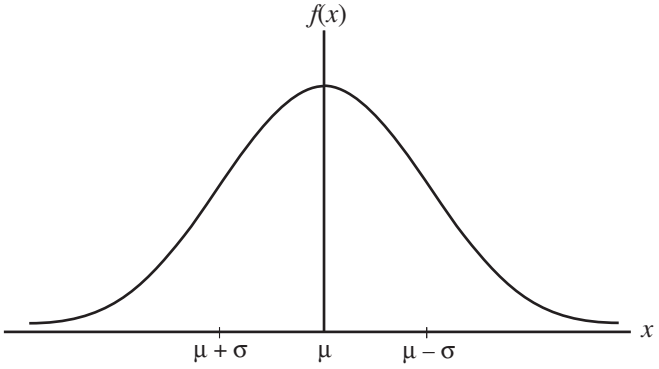

図 3: 正規分布の確率密度関数

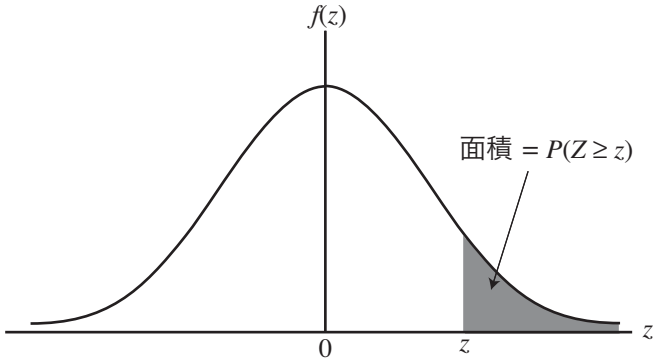

図 4: 標準正規分布の確率密度関数のグラフ上で「確率変数 Z が値 z 以上である確率」$P(Z \geq z)$

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．

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．

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」$P(Z \geq z)$ を計算したもので，確率密度関数のグラフにおいては図 4 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．

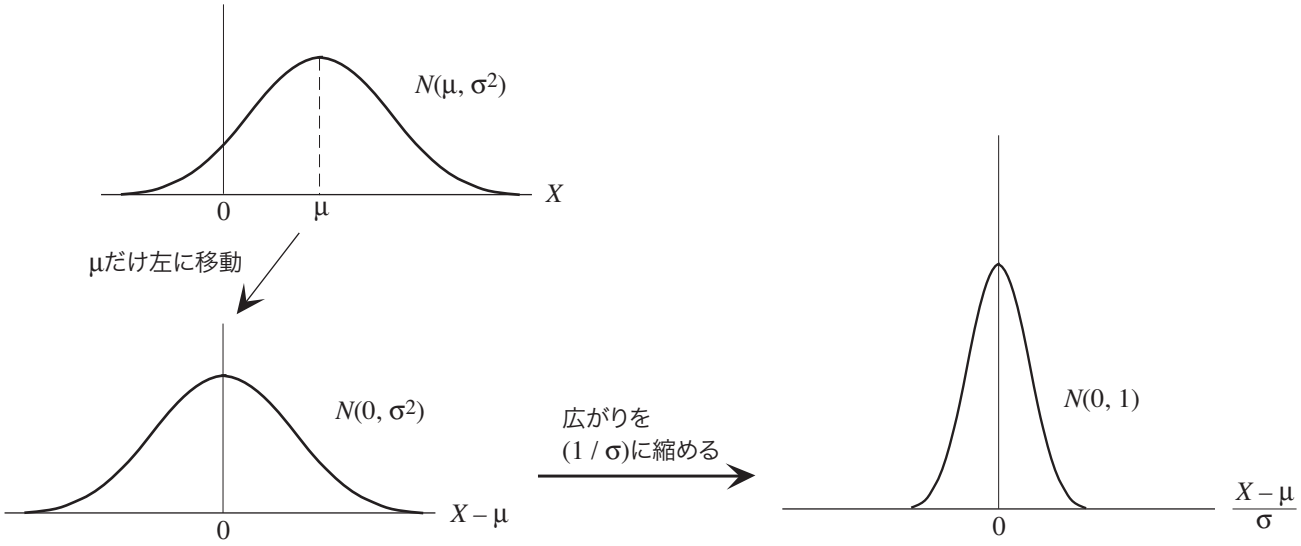

図 5: 正規分布の性質 1

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，$P(X \geq 55)$ と $P(45 \leq X \leq 60)$ を求めてください．

付録：「離散」と「連続」の間

次の問題を考えてみます．

1 秒毎にステップ式に動くのではなく，連続的に動く秒針があるとします．あなたは，好きなときにボタンを押して秒針を止めることができます．針を見ずにボタンを押したとき，

(a) 針が 0 時の位置から 3 時の位置の間に止まる確率はいくらですか．
(b) 針が 0 時ちょうどの位置に止まる確率はいくらですか．

≡ ∧∧ (a) は，時計の周囲のうち「0 時の位置から 3 時の位置の間」の
≡ ・ ・ ≡ 幅が円周全体の 1/4，ということから求めればいいけど，(b) は
( ) ～ どう考えればいいんでしょう？

(a) と同じように考えるんやけど，「『0 時ちょうどの位置』の幅」は，いくらかいな？

＝ ∧◆∧
＝ o- o ＝
( ) ～

針が止まる位置は連続型確率変数と考えることができます．針は一定の速度で動きますから，特定の場所に止まりやすい，止まりにくい，といった偏りはありません．ですから，確率密度関数のグラフは図 A1 のような平坦な形になります．

(a) については，０時の位置から３時の位置に止まる確率は図 A1 の灰色の部分の面積で，０時から３時の間の幅は文字盤１周の 1/4 ですから，この面積もグラフの下の部分全体の面積の 1/4 となり，求める確率は 1/4(0.25) となります．

これに対して，(b) では，文字盤上で「０時ちょうど」の部分の幅は０ですから，そこに止まる確率も０です．この答えについて，こんな疑問を持つ人がいるのではないでしょうか．

> 「１２時ちょうどに針が止まる確率は，『１２時ちょうど』の幅が０だから，０だという．それならば，１２時０分０秒にも１２時０分 0.1 秒にも１２時０分 0.01 秒にも，文字盤の周上のどこに止まる確率もみな０のはずだ．それなのに，『０時から３時までの間』のどこかに止まる確率は 1/4 だという．これはどういうことか」

この疑問に答えるポイントは，文字盤の周をいくら細かく刻んでも，その刻みで文字盤の周全体を埋め尽くすことはできないということです．つまり，１２時０分０秒にも止まる確率も，１２時０分 0.1 秒に止まる確率も，１２時０分 0.01 秒に止まる確率も皆ゼロですが，だからといって「文字盤の周上のどこに止まる確率もみな０」ではないのです．

文字盤の周を，１秒刻み，0.1 秒刻み，0.01 秒刻み，といくらでも細かく刻むことはできます．したがって，文字盤の周に無限個の刻みを並べることができます．このように「びっしり」と並んだ無限個の刻みは，数学の言葉では稠密であるといいます．このような無限個の刻みには，１２時ちょうどの位置から数えて，１番から順に番号をつけることができます．「無限個だが，番号をつけて数えることができる」ことを，「数えられる無限」という意味で可算無限といいます．

一方，文字盤の周上の位置は，例えば１２時ちょうどの位置を０度として，実数値の角度で表現できます．もし，文字盤の周上の角度を表す全ての実数値に１番から番号をつけることができるなら，それは「可算無限個の刻みで，文字盤の周上の全ての点を埋め尽くすことができる」つまり「無限に刻みを細かくすれば，文字盤の周上のどんな位置でも表せる」ことになります．それならば，刻みの各点に止まる確率は０ですから，文字盤の周上のどこに止まる確率も０ということになります．

しかし，実は「全ての実数値に１番から番号をつけることはできない」のです．つまり，「無限個」にも「大小」があり，文字盤の周上の実数値の数は，可算無限個よりもずっと多い，別種の無限個なのです．直観的にいえば，可算無限個の刻みは「びっしり」並んでいるのに対して，実数は「べったり」と並んでいる，ということです[2]．

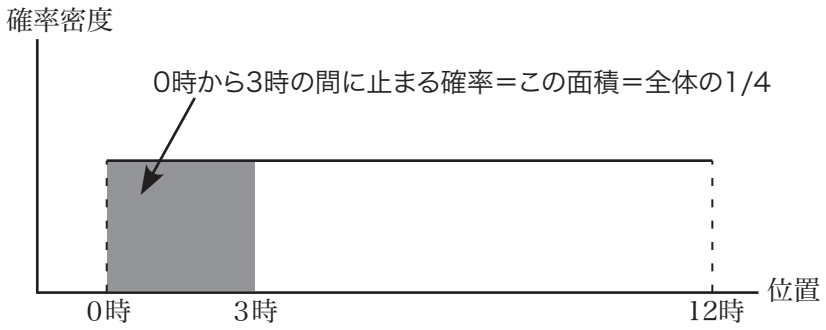

図 A1: 時計の針が止まる位置の確率密度関数

[2]実数は，さきほど述べた「稠密性」だけではなく，「連続性」を持っています．

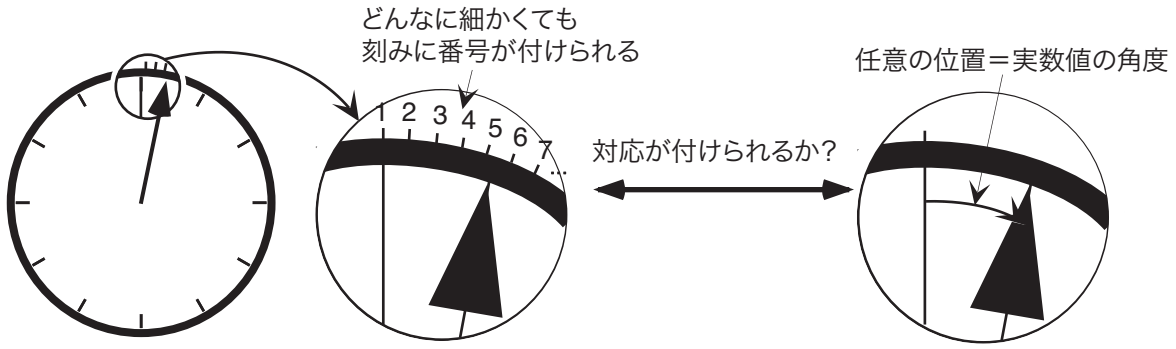

図 A2: 周上の無限個の刻みと，実数値の角度

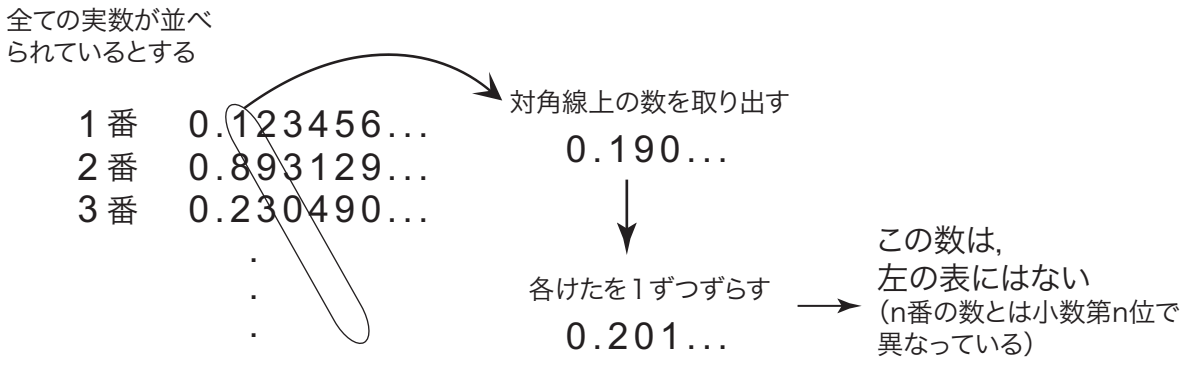

図 A3: 対角線論法

実数が可算でないことは，次に示すカントールの対角線論法で簡単に説明できます．説明を簡単にするため，0 以上 1 未満の実数を考えることにします．この区間のすべての実数は，$0.xxxx\ldots$ の形の，有限小数あるいは循環する無限小数（すなわち有理数），あるいは循環しない無限小数（すなわち無理数）で表されます[3]．さて，すべての実数に 1 番から番号をつけることができるとしましょう．そこで，図 A3 のようにすべての実数を 1 番から順に上から並べた表を作ります．そこで，この表から，「1 番の実数の小数第 1 位，2 番の数の小数第 2 位,...，n 番の数の小数第 n 位...」のように，対角線上の各数字をつなぎあわせた数をつくり，さらにその数の各けたを「$0 \to 1$，$1 \to 2$,...，$9 \to 0$」のように置き換えた数を考えます．この数は，さっきの表の 1 番の数とは小数第 1 位で，2 番の数とは小数第 2 位で,...，n 番の数とは小数第 n 位で... 異なっています．つまり，表のどの数とも異なった数が存在することになり，「すべての実数を並べた表」であるということに矛盾します．つまり，「すべての実数を 1 番から順番に並べることはできない」ということが証明されます．

「無限にも大小がある」という事実が数学に与えた衝撃は，その後の数学をそれ以前のものとは根本的に違ったものにしてしまったほど，大きなものでした．そのあたりを平易に解説した本としては，瀬山士郎「はじめての現代数学」（講談社現代新書 909，ISBN4-06-148909-7）をおすすめします．

[3]正確には，例えば $0.1 = 0.0999\ldots$ のように，有限小数は無限小数の形に統一して表すことにします．