@IT
atmarkIT

社会人１年生から学ぶ、やさしい確率分布【Excel／エクセルで学べる】

羽山博［著］

この連載は、データをさまざまな角度から分析し、その背後にある有益な情報を取り出す方法を学ぶ『社会人 1 年生から学ぶ、やさしいデータ分析』連載（記述統計と回帰分析編）の続編で、確率分布に焦点を当てています。

この確率分布編では、推測統計の基礎となるさまざまな確率分布の特徴や応用例を説明します。身近に使える表計算ソフト（Microsoft Excel や Google スプレッドシート）を使いながら具体的に事例を見ていきます。

必要に応じて、Python のプログラムでの作成例にも触れることにします。

数学などの前提知識は特に問いません。中学・高校の教科書レベルの数式が登場するかもしれませんが、必要に応じて説明を付け加えるのでご心配なく。肩の力を抜いてぜひとも気楽に読み進めてください。

やさしいデータ分析【確率分布編】　新連載開始！

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の確率分布編です。第 1 回は出発点として、推測統計の基礎となる確率分布の意味や種類、特徴を解説します。離散型分布と連続型分布の違いや種類、確率分布を表す確率質量関数／確率密度関数と累積分布関数の意味や特徴などを見ながら連載の全体像を紹介します。

羽山博（2024 年 05 月 09 日）

確率分布は難しくない ～ 推測統計の基礎を固めよう

確率分布がどう役に立つのか？

2023 年度（令和 5 年度）に実施された文部科学省の全国学力・学習状況調査（いわゆる学力テスト）では、公立小学校 6 年生の国語の平均正答数は **9.4 問**（14 問中）、標準偏差は **2.9 問**でした（国立教育政策研究所の報告書による）。

さて、ある学習塾で公立小学校に通う 6 年生 10 人の生徒に同じテストを受けさせたところ、平均正答数が **9.7 問**だったとしましょう。皆さんは、9.7 − 9.4 = 0.3 の差をどのように考えるでしょうか（図 1）。学習塾の経営者としては全国の平均よりも正答数が多かったと喜びたいところでしょう。しかし、この差は「たまたま」なのかもしれません。

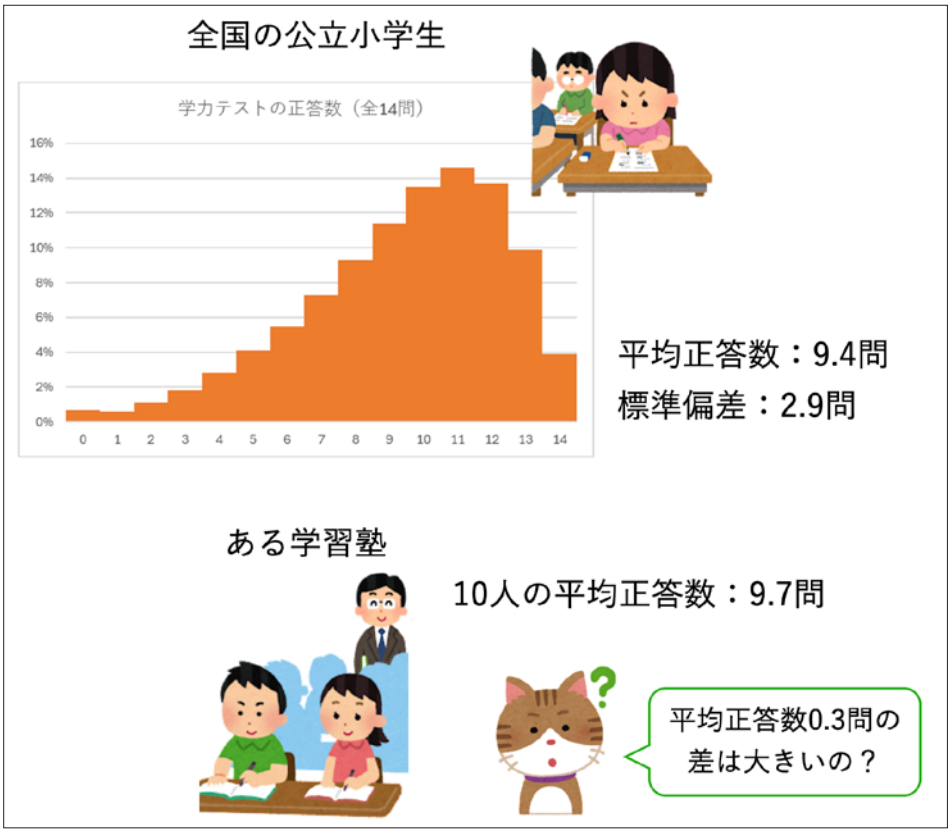


図 1　学力テストの結果をどう捉えるか？

図中のグラフは上の出典資料から作成したもの。**0.3 問**の差をどう捉えればいいだろうか。学習塾の生徒の方が正答数が多いのか、たいした差ではないのか？

→目次に戻る

上のような問いに対して、何らかの根拠を持って判断を下すためには推測統計の知識が必要になります。結論だけ言うと「正答数が多いとは言えない」ということになるのですが、なぜそうなるのかは今後の連載の中でお話しします（後のお楽しみということにしましょう）。

そして、その推測統計の基礎となっているのが**確率分布**です。この例は、公立小学校 6 年生全体を**母集団**として、その正答数が**正規分布**に従っている（正規分布になっている）という前提でのお話です。確率分布を知ることにより、母集団の性質を推測したり、平均値に差があるかどうかを確かめたりすることができます。

2023 年度（令和 5 年度）は公立の小学 6 年生のうち **94.6%**の生徒が参加しているので、母集団を公立小学校の 6 年生全体と考えています。なお、学力テストは、機会均等／学習状況の改善などを目的とする一方で、自治体ごとに成績を公表するのは、過度に競争をあおる可能性があるなど、さまざまな問題点も指摘されています。

確率分布とは何か、ということに関してはちょっと後回しにして、まず、確率分布がデータ分析／データサイエンスにおいて、どのような位置にあるのかを紹介しておきましょう。

データ分析と確率分布の関係とは？

『やさしいデータ分析』の記述統計と回帰分析編で、最初にデータ分析とデータサイエンスの全体像を紹介しました。図 2 は、その際に掲載した図の右側に、確率分布などがどう関係してくるのかを描き加えたものです。それぞれの領域／分野はお互いに関係していますが、これからの連載でお話しする内容に関しては、確率分布は推測統計の基礎として密接に関係しています。

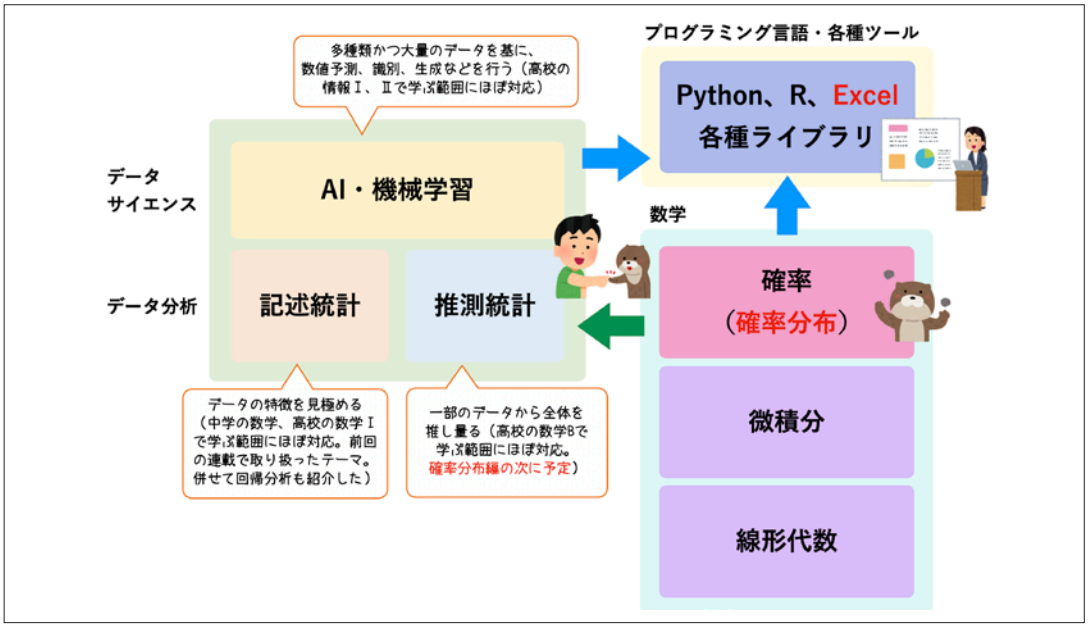


図 2　データ分析／データサイエンスと数学、各種ツールの関係

数学はデータ分析やデータサイエンスの基礎となっている。とりわけ、確率分布は推測統計の基礎として極めて重要。なお、プログラミング言語や各種ツールはデータ分析やデータサイエンスを進める上で必須となる道具で、数式や関数などのシミュレーションや可視化にも使われる。

→目次に戻る

推測統計とは、一部のデータ（標本）を基に、その元となるデータ（母集団）の性質を「推し測る」ためのさまざまな方法のことです。そのためには、確率分布の知識が必要になります。つまり、「このようなデータであれば確率的にこうなるはずだ」→「実際に得られたデータはこうだ」→「母集団はこのような分布だろう」といった推測を行うわけです。仮説検定（以下の注釈を参照）を含む推測統計は、データ分析やデータサイエンスの一つの柱となります。

図 1 の中の記述統計（と回帰分析）については、すでに述べたように前回の連載で解説しました。数学とプログラミングについては、数学 ×Python プログラミング入門で解説しています。また、AI・機械学習で使われる数学については AI・機械学習の数学入門で解説しています。その中で、確率分布についても代表的なものについて考え方や計算の方法を数学的に解説しています。

この連載では、実感を持って理解できるように、Excel を使って手を動かしながらさまざまな確率分布の形や応用例を見ていきます。なお、仮説検定とは、サンプルとして取り出したデータを基に 2 つのグループの平均には差があるかどうかといったことを、一定の根拠に基づいて判断する方法です。確率分布は仮説検定を行うための基礎となります。

今回は、連載の開始に当たって、確率分布を理解する上で重要となる以下のキーワードについて解説し、その後、連載の内容を紹介します。

- **離散型分布と連続型分布** …… 確率変数の取り得る値が幾つかに限られる（飛び飛びの値を取る）ものが離散型分布。確率変数が範囲内のどの値でも取れるものが**連続型分布**。
- **確率質量関数／確率密度関数** …… 確率分布を関数 $y = f(x)$ として表したもの。離散型分布の場合は**確率質量関数**と呼ばれ、連続分布の場合は**確率密度関数**と呼ばれる。
- **累積分布関数** ……確率質量関数の累計や確率密度関数の積分値を関数 $Y = F(x)$ として表したもの。x 以下となる確率を表すもの。

これらのキーワードを初めて聞く方にとっては、いったい何を言っているのか分からないと思われるかもしれませんが、ここから少しずつ具体的に解説していきます。

そもそも確率分布って何？

では、始めます。「そもそも」のところからスタートです。そもそも確率分布とは、いったい何なのでしょうか。

「**確率**」に関しては、中学や高校の数学でも登場したので、基本的な意味についてはご存じだと思います。全ての事象（出来事）の中で、ある事象が起こる割合のことですね。例えば、どの目も同じように出る 6 面体のサイコロ（以下、単に「サイコロ」と呼びます）であれば、1 の目が出る確率は **1/6**（6 分の 1）です。全体の事象が **6 通り**あって、1 の目が出るという事象は **1 通り**ですから、**1/6** だというわけですね。、

一方の「**分布**」とは、どの値がどの位置にどれぐらいあるかということです。

ということは、**確率分布**とは、どの確率がどの位置にどれぐらいあるかということですね。サイコロの目であれば、どの目が出る確率も **1/6** なので、それぞれの目が出る確率分布は図 3 のように表されます。

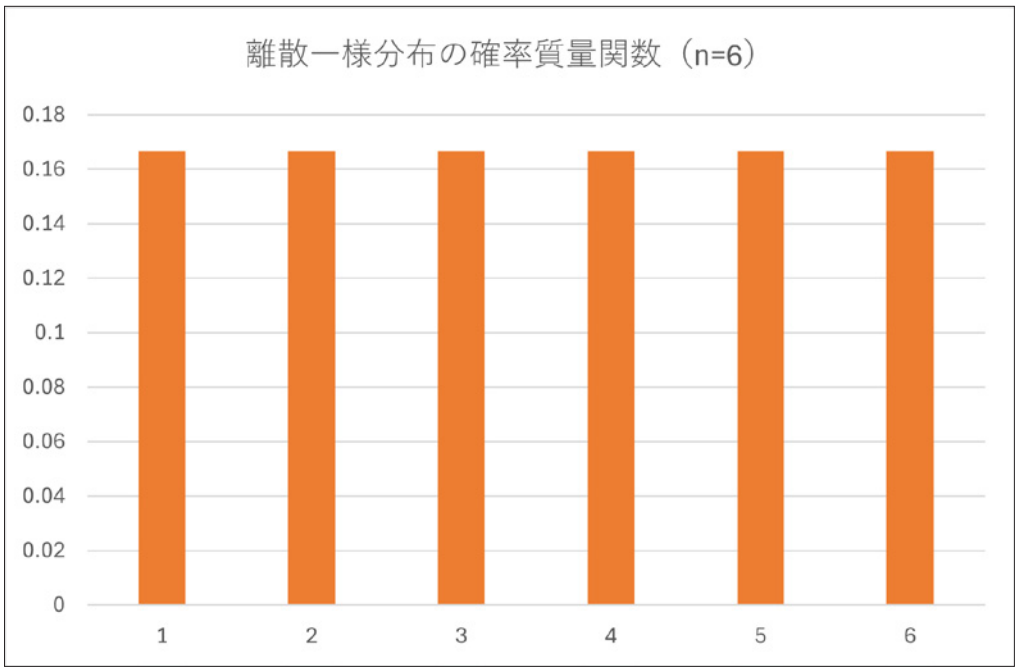


図 3　サイコロの目の確率分布（離散一様分布の例）

サイコロのそれぞれの目は 1/6 = 0.166... の確率で出るので、グラフにするとこのようになる。横軸がそれぞれの目、縦軸が確率となる。横軸の値を表す変数は**確率変数**と呼ばれる。グラフのタイトルにある**確率質量関数**の意味は後述する。

ここで重要なことは、図 3 が、実際にサイコロを振って出た目の数を基に求めたものではなく、理論的にはこうなるはずだ、ということです。確率分布とは、そのような理論的な分布のことです。記述統計で扱った度数分布表などの「分布」は実際に得られたデータの分布でした。この違いを意識しておいてください。

もう 1 つ、気が付くことがあるのではないかと思います。**確率分布の全ての値の合計は 1 となっている**ということです。全事象の確率が **1（= 100%）**となるので当然のことですが、このことも重要です。

図 3 の分布は、どの事象の確率も同じですね。このような確率分布を**一様分布**と呼びます。また、横軸（確率変数）の値が 1、2、3、4、5、6 と飛び飛びになっています。そのような確率分布を**離散型確率分布**または**離散確率分布**と呼びます。従って、図 2 の分布は**離散一様分布**と呼ばれます。

離散型確率分布と連続型確率分布

もう 1 つ、離散型確率分布の例を見ておきましょう。先ほどのサイコロを 5 回振ったときに 1 の目が何回か出る確率分布は図 4 のようになります。ここでは、理屈は抜きにして結果だけを紹介します。1 の目が 0 回しか出ない、1 回出る、2 回出る、3 回出る、4 回出る、5 回出るという 6 つの場合があります。横軸（確率変数）が図 2 とは異なることに注意してください。

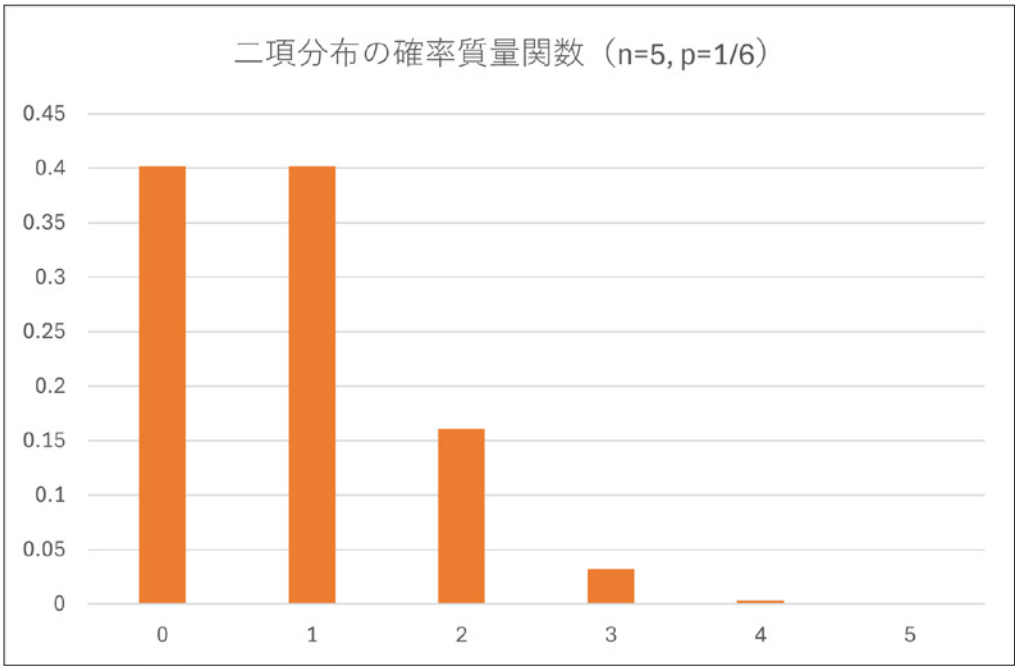


図 4　サイコロを 5 回振って、1 の目が何回か出る確率（二項分布の例）
1 の目が出ない（0 回出る）場合と 1 回出る確率がいずれも **0.4 程度**。一方、1 の目が 5 回とも出る確率は **0.0001 程度**（グラフではほとんど見えない）。

図 4 のような分布は、**二項分布**と呼ばれるものです。二項分布（やその他の分布）がどのようなものかは、この連載で少しずつ解説していきます。二項分布は離散分布なので、いちいち**離散二項分布**と呼ばずに、単に**二項分布**と呼びます。

ところで、横軸の値が飛び飛びでない（範囲内のどの値でも取れるような）場合もありそうですね。そのような分布を**連続型確率分布**または**連続確率分布**と呼びます。図 3 で見た一様分布には、**連続一様分布**もあります。連続型確率分布として最も有名なものは**正規分布**です。今のところはやはり理屈抜きで、分布の形だけを確認しておきましょう（図 5）。

→目次に戻る

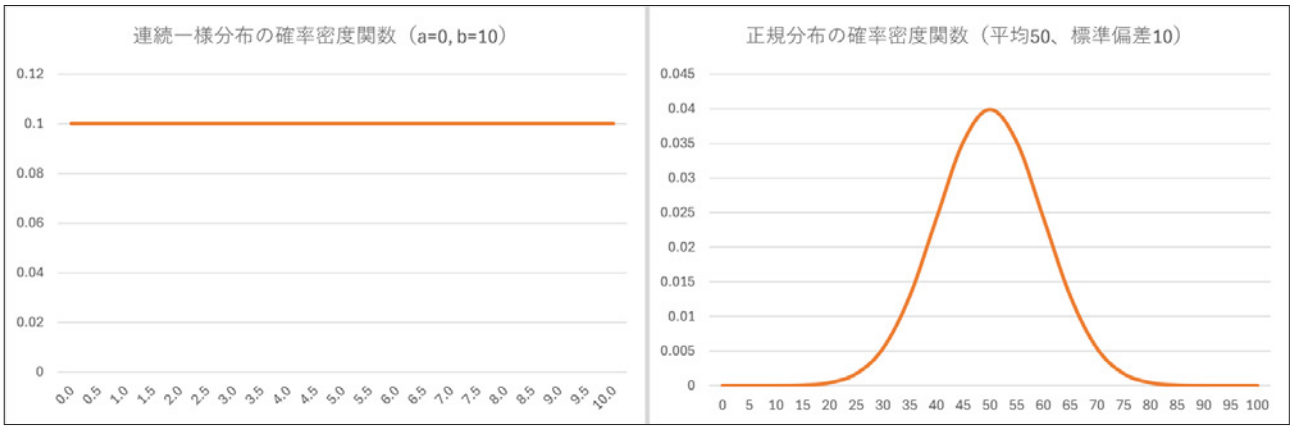


図 5　連続一様分布と正規分布

連続型確率分布の場合、値が飛び飛びではないので棒グラフではなくスムーズな曲線で表される。左の例は、**0** ～ **10** の範囲の連続一様分布。目盛は **0.5** 刻みで表示してあるが、確率変数の値 x は **0** ～ **10** の範囲のどの値を取ることもでき、どの値も同じ確率で現れる（理想的な一様乱数のようなもの）。右の例は、平均が **50**、標準偏差が **10** の正規分布の例。正規分布はさまざまな場面で登場する。

離散型確率分布の場合、確率変数の値 x に対する y の値はその事象が起こる確率です。しかし、連続型確率分布の場合、確率変数の値 x に対する y の値はその事象が起こる確率では**ない**ことに注意が必要です（後述する累積分布関数の微分係数であると考えられます）。例えば、図 5 の正規分布のグラフでは x = 50 のとき、y = 0.3989 ですが、x = 50 である確率が 0.3989 であるというわけではありません。

確率質量関数／確率密度関数と累積分布関数

図 4 や図 5 を見ると、横軸の値を x、縦軸の値を y としたとき、y は x の関数になっていることが分かります。つまり、x に値（例えば **1** など）を入れると、それに対応する y の値が自動的に決まるわけです。このような関数を、離散型分布の場合は**確率質量関数**と呼び、連続型分布の場合は**確率密度関数**と呼びます。

関数である……ということは、数式で表せますね。一応、以下に記しておきますが、今のところはあまり気にしなくてもけっこうです。今後の連載の中で分かりやすく解説します（二項分布では、x の代わりに k という文字を使っています）。

二項分布の確率質量関数

$$f(k) =_n C_k \cdot p^k (1-p)^{n-k}$$

n：試行の回数、p：事象が起こる確率、k：事象が起こる回数

正規分布の確率密度関数

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

π：円周率、σ：標準偏差、σ^2：分散、e：自然対数の底、μ：平均

→目次に戻る

また、x **以下の確率**を表す関数のことを**累積分布関数**と呼びます。離散型確率分布の場合、累積分布関数は、x に対する確率質量関数の累計値を関数として表したものです。サイコロの例であれば、k = 2 以下の累積分布関数の値は、k = 0 ～ 2 の確率質量関数の値を全て累積した値になります（図 6）。

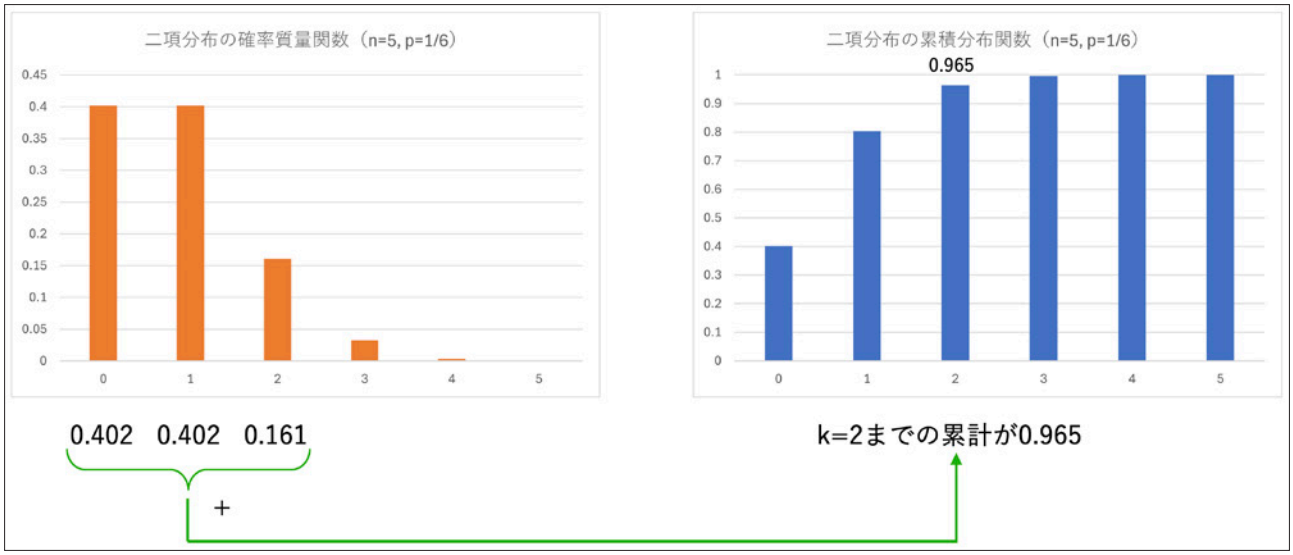


図 6　二項分布の確率質量関数と累積分布関数

離散型確率分布の場合、累積分布関数の値はそれまでの値の累計となる。この例の二項分布であれば、例えば、k = 2 に対する累積分布関数の値はそれまでの累計、つまり 0.402 + 0.402 + 0.161 = 0.965 となる。当然のことながら、一番右の棒の高さは全ての確率を足した値なので **1** になる。

連続型確率分布の場合、累積分布関数は確率密度関数を積分した値を関数として表したものになります。例えば、平均 **50**、標準偏差 **10** の正規分布であれば、x = 60 以下の累積分布関数の値は、確率密度関数を−∞～60 まで積分した値になります（図 7 の左側、グレーの部分の面積）。x =−∞～∞について、積分値（面積）を求め、プロットしていくと累積分布関数のグラフ（図 7 の右側）になります。ただし、図 7 では x = 0 ～ 100 の部分だけを表示してあります。

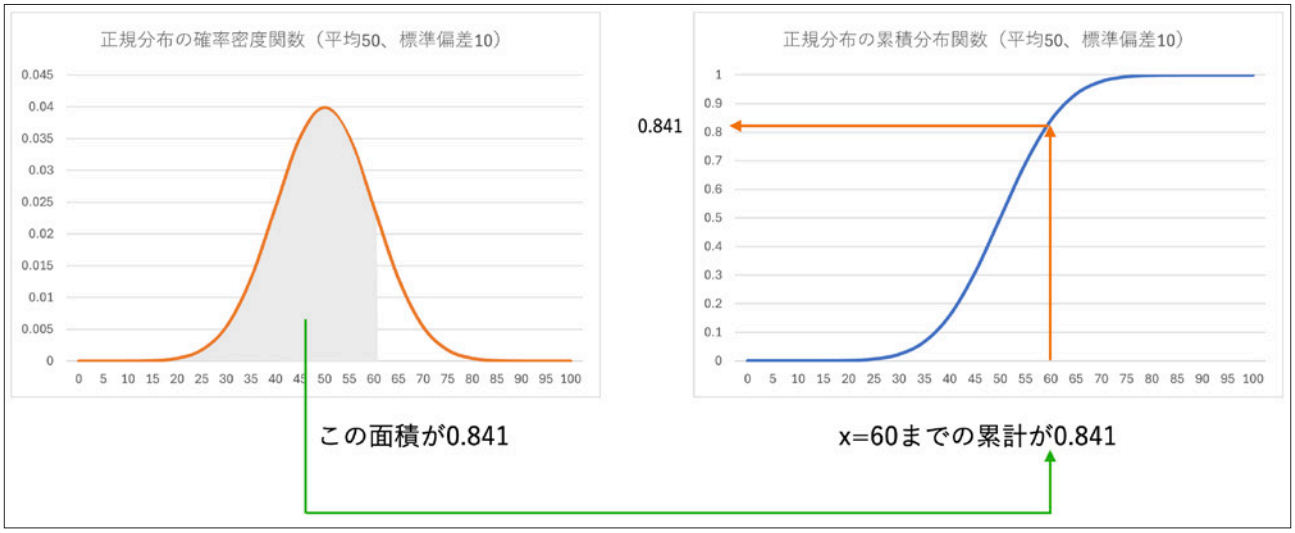


図 7　正規分布の確率密度関数と累積分布関数

連続型確率分布の場合、累積分布関数の値はそれまでの積分値となる。正規分布では定義域（x の取り得る値の範囲＝「台」とも呼ばれる）が−∞～∞までとなるので、この例の場合、x = 60 に対する累積分布関数の値は確率密度関数を−∞～ 60 まで積分した値、つまり図 7 のグレーの部分の面積となる。累積分布関数の右端の y の値は限りなく **1** に近づく。

→目次に戻る

ところで、二項分布では、n の値と p の値が決まれば、確率質量関数や累積分布関数が一意に決まります。同様に、正規分布では平均μと標準偏差σが決まれば確率密度関数や累積分布関数が一意に決まります。このように、関数を一意に決める値のことを**母数**または**パラメーター**と呼びます。

推測統計では、累積分布関数の値（累積確率）やその逆関数の値が重要となります。**累積分布関数の逆関数**とは、累積確率から、それに対する x の値を求める関数ということです。例えば、図 7 に示した正規分布の例であれば、累積確率が **0.9** となる x の値は **62.816** です。縦軸の **0.9** の値から逆に横軸の x の値を求めればいいというわけです。

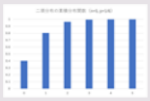
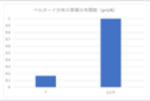
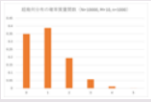
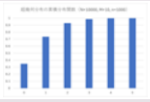
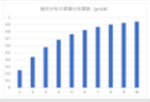
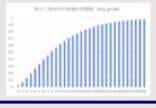
確率分布早わかり ～ この連載で取り扱う内容

というわけで、この連載では、さまざまな確率分布について、その意味や利用例などを、Excel を使って手を動かしながら学んでいきます。表 1 は次回以降少しずつ見ていく内容なので、現時点では右側の説明は気にせず、この連載でどのような分布を取り扱うのかをざっと眺めておいていただくだけで構いません。

◇ ◇ ◇ ◇ ◇ ◇ ◇

繰り返しになりますが、確率分布は区間推定や検定などの推測統計の基礎となる考え方です。一歩ずつ着実に進めていけるように、やさしく説明するつもりです。次回からの連載にぜひご期待ください。

◆ 離散確率分布

回	分　布	確率質量関数	累積分布関数	説明
1（今回）	離散一様分布			全ての事象が同じ確率で起こる場合の分布。グラフはサイコロのそれぞれの目が出る確率の分布。
2	ベルヌーイ分布			結果が2つの場合に分かれるような試行（ベルヌーイ試行）の分布。グラフはサイコロを1回投げたときに1の目とそれ以外の目が出る確率の分布。
2	二項分布			ベルヌーイ試行を何回か繰り返したときに、それぞれの事象が起こる確率の分布。グラフはサイコロを5回投げたときに1の目が何回か出る確率の分布。
3	超幾何分布			製品の抜き取り検査のように、取り出したものを元に戻さない場合に、ある事象が何回か起こる確率の分布。グラフは10000本中10本の当たりくじがある福引きで、くじを1000本引いたときに何本かの当たりがある確率の分布。
4	ポアソン分布			まれにしか起こらない事象が起こる確率の分布。グラフは100人中平均0.5人がかかる疾病に何人かがかかる確率。
5	幾何分布			ある事象がちょうど何回目から起こる確率の分布。グラフは、1/4で当選するくじにちょうどk回目で当選する確率の分布と、k回目までに当選する確率の分布。
5	負の二項分布			ある事象が何回か起こるまでに、それ以外の事象が起こる確率の分布。グラフはサイコロを投げたときに1の目が2回出るまでに他の目が何回か出る確率の分布。

◆ 連続確率分布

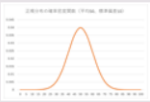
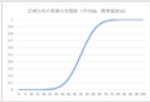
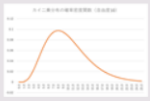
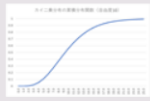
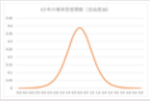
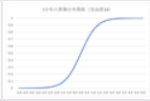
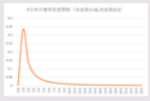
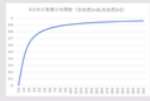
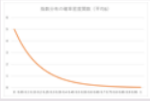
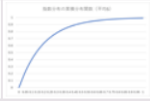
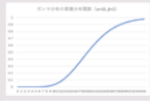
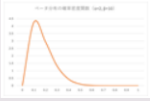
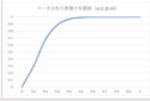
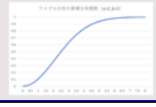
回	分　布	確率密度関数のグラフ	累積分布関数のグラフ	説明
1（今回）	連続一様分布			全ての事象が同じ確率で起こる場合の分布。グラフは0以上10以下の（理想的な）一様乱数の確率分布。
6	正規分布			誤差の確率分布など、さまざまな場面で使われる分布。グラフは平均50、標準偏差10の正規分布。平均0、標準偏差1の正規分布を特に標準正規分布と呼ぶ。
7	カイ二乗分布			母分散の区間推定や検定、適合性の検定、独立性の検定に使われる分布。グラフは自由度10のカイ二乗分布。
8	t分布			母平均の差の検定、相関係数の検定、回帰式の係数の検定などに使われる確率分布。グラフは自由度10のt分布。
9	F分布			母分散の比の検定、分散分析、回帰式の当てはまりの検定などに使わる分布。グラフは自由度（10, 2）のF分布。
10	指数分布			eの$-x$乗で表されるような分布（正確には$\lambda e^{-\lambda x}$）。グラフは1時間に平均5人の顧客が来る店で、何時間で次の客が来るかの分布。
11	ガンマ分布			グラフはα=10、β=2のガンマ分布。例えば、2分に1台の車が通過する地点で、10台の車が通過するまでの時間。αが正の整数の場合はアーラン分布と呼ばれる。
12	ベータ分布			ベルヌーイ試行で、ある事象が起こる確率の分布。グラフはα=2、β=10のベータ分布。ベータ分布は作業時間の見積りなどにも使われる。
13	ワイブル分布			xを時間として、故障率が$x^{\alpha-1}$の関数で表される場合に、故障率や寿命を求めるのに使われる分布。グラフはα=2、β=3のワイブル分布。

表 1　連載の内容

おおむね、この表の流れに沿って、さまざまな分布について詳しく説明していく。ただし、離散一様分布と連続一様分布については簡単なので、今回の説明のみとする。表には確率質量関数／確率密度関数、累積分布関数のグラフも併せて掲載したが、パラメーターが変わると異なる形になることがある。今の段階では、だいたいの雰囲気を捉えるだけで十分

→目次に戻る

［データ分析］二項分布とベルヌーイ分布 ～ 離散型確率分布の基本

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載（確率分布編）の第 2 回。推測統計の基礎となる確率分布のうち、離散型確率分布で代表的なベルヌーイ分布と二項分布の意味や特徴などを解説します。

羽山博（2024 年 06 月 06 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の確率分布編、第 2 回です。前回は出発点として、推測統計の基礎となる確率分布の意味や種類、特徴を解説しました。今回はベルヌーイ分布と二項分布を取り上げ、それらの意味や特徴などを見ていきます。

3 割打者が 3 人続くと確実に点を取れるのか？

野球に興味のない方にとっては、ちょっとなじみのないお話になるかもしれませんが、最近のプロ野球では 3 割打者がほとんどいなくなっています。2013 年は 15 人（セ・リーグ 5 人、パ・リーグ 10 人）だったのが、2023 年はたったの 5 人（セ・リーグ 3 人、パ・リーグ 2 人）でした。これは、投手の技術が向上したことなどが原因だと言われています。

3 割打者の人数は、規定打席以上の打者について筆者が個人的に集計したものです。公式記録とは食い違っている可能性もあります。2024 年の記録については、日本野球機構のページから見ることができます。

当然のことながら、ヒットが連続すれば、得点が得られる可能性が高くなります。話を簡単にするために、3 人の打者がヒットを 2 本以上打つと得点できるものとして、3 割打者が 3 人続くとどれぐらいの確率で得点できるかを考えてみましょう。ただし、相手の投手によって打率が変わったり、打者の調子に波があったりすることはなく、3 人とも平均して 3 割の打率をキープしているものと考えます。

図 1　3 割打者が 3 人続いたときに、ヒットが 2 本以上出る確率は？

1 打席でヒットを打つ確率は 0.3、凡退する確率は 0.7。では、3 打席で 2 本以上ヒットを打つ確率はいくらになるだろうか。3 割打者が 3 人続くと本当に怖いのか？

上のような問いに答えるには、ベルヌーイ分布と二項分布の知識が利用できます。まず、ベルヌーイ分布から見ていきましょう。

ベルヌーイ分布：2 種類の結果が得られる場合の確率分布

サイコロを投げたり、くじを引いたりする行為のことを**試行**と呼びます。厳密には、同じ条件で繰り返し行うことのできる実験や観測などのことを意味します。また、試行の結果を事象と呼びます。平たく言えば、試行とは「何かをやること」、**事象**とはその結果として起こる「出来事」です。

さまざまな試行のうち、1 回の試行で結果が 2 つに分かれるものを**ベルヌーイ試行**と呼びます。図 1 の例も、1 回の試行が「ヒットであるか、そうでないか」の 2 つに分かれるので、ベルヌーイ試行です。

ベルヌーイ試行で、成功を k = 1、失敗を k = 0 と表し、成功する確率を p と表すと、その確率分布、つまり**ベルヌーイ分布**の確率質量関数 f(k) は以下のように表されます。

$$f(k) = \begin{cases} p & k = 1 \\ 1 - p & k = 0 \end{cases} \qquad (1)$$

→目次に戻る

3 割打者の例で言えば、成功の確率は **0.3**、失敗の確率は 1 − 0.3 = **0.7** です。ちなみに、成功とは「うまくいったかどうか」という意味ではなく、「目的の事象が起こったかどうか」ということを表します。

ベルヌーイ分布は極めてシンプルな分布なので、わざわざ可視化しなくても理解できるとは思いますが、一応、Excel でグラフを描いておきます（図 2）。

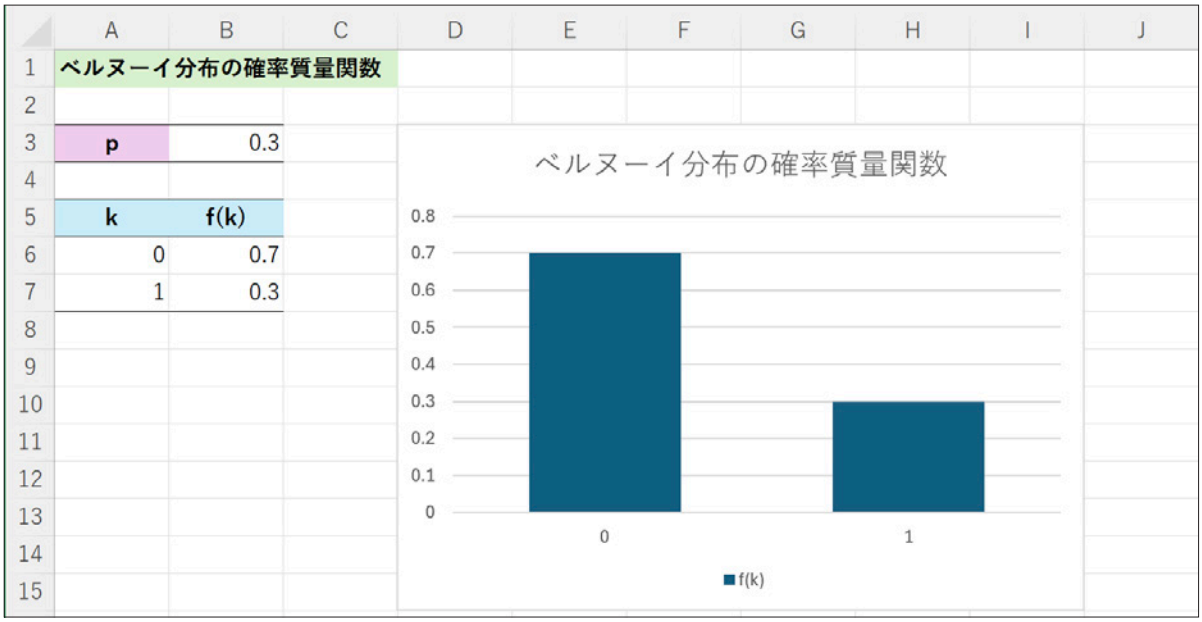


図 2　ベルヌーイ分布の確率質量関数

確率変数の値 k は **0** または **1**。ここでは、確率変数の値 k が昇順に並ぶようにするため、k の値としてセル **A6** に **0**、セル **A7** に **1** と入力してある。従って、セル **B6** には =1-B3、セル **B7** には =B3 と入力してある（上で示した（1）式とは順序が逆になっている）。

グラフ作成の手順は以下の通りです。サンプルファイルをこちらからダウンロードし、表計算ソフト Microsoft Excel で［ベルヌーイ分布］ワークシートを開いて試してみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。なお、Google スプレッドシートでの操作方法は、サンプルファイル内に記載しています。

◆ Excel での操作方法

- セル **B5 ～ B7** を選択する
- ［挿入］タブを開き、［縦棒／横棒グラフの挿入］ボタンをクリックして［集合縦棒］を選択する
- ［グラフのデザイン］タブを開き、［データの選択］ボタンをクリックする
- ［データソースの選択］ダイアログボックスで［横（項目）軸ラベル］の下の［編集］ボタンをクリックする
- ［軸ラベル］ダイアログボックスで［軸ラベルの範囲］ボックスをクリックし、セル **A6 ～ A7** を選択する
- ［OK］をクリックして［軸ラベル］ダイアログボックスを閉じる
- ［OK］をクリックして［データソースの選択］ダイアログボックスを閉じる

あとは、グラフのサイズを変更したり、タイトルを設定したりして、見やすくすれば完成です（以降、これらの見た目を整えるための操作については省略します）。

二項分布：ベルヌーイ試行を何回か繰り返したときの確率分布

ベルヌーイ分布は、1 回の試行の確率分布でした。野球の例で言えば、ある 1 人の打者について、3 割の確率でヒットを打つことが分かっている場合に、ある打席でヒットを打つか凡退するかの確率分布です。今回は 3 人の打者についての確率を知りたいので、ベルヌーイ試行を繰り返す必要があります。そのような場合の確率分布が**二項分布**（Binomial distribution）です。

実は、この連載の第 1 回で二項分布の確率質量関数をすでに紹介しています。以下のような式でした。数式が苦手な方は目眩（めまい）を起こしそうになるかもしれませんが、今のところはあまり気にせず先に進んでください。後で、具体的に解説します。また、BINOM.DIST 関数を使えば簡単に答えが求められます。

$$f(k) =_n C_k \cdot p^k (1-p)^{n-k} \qquad (2)$$

n：試行の回数、p：事象が起こる確率、k：事象が起こる回数

${}_nC_k$ は、n 個の中から k 個のものを選ぶ組み合わせ数です。高校までの数学では、${}_nC_k$ と表しますが、一般には、

$$\binom{n}{k}$$

と表します。従って、（2）式は、以下のように表されます。

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad (3)$$

では、野球の例で具体的に考えてみましょう。あらかじめイメージが湧くように、最初に確率質量関数と累積分布関数のグラフを掲載しておきます（図 3）。グラフの作成方法については、後で解説します。この連載の第 1 回でも解説しましたが、**確率質量関数**とは、ある確率変数に対する事象が起こる確率です。この例では、ヒットを打つ打者の人数が確率変数なので、例えば、2 人の打者がヒットを打つ確率（k = 2 に対する確率）などが確率質量関数の値に当たります。一方の**累積分布関数**は、ある確率変数の値以下の事象が起こる確率の累計です。例えば、2 人**以下**の打者がヒットを打つ確率（k = 0 の場合と k = 1 の場合と k = 2 の場合の累計）が累積分布関数の値に当たります。全ての場合を表とグラフにしたのが図 3 です。

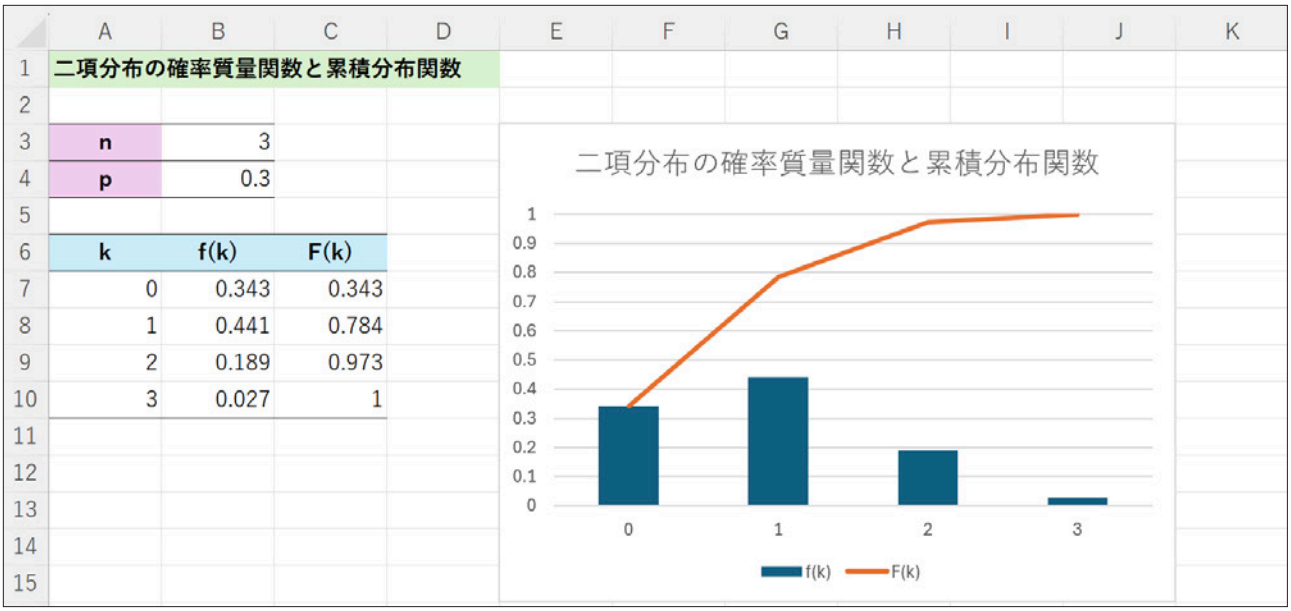


図 3　3 人の 3 割打者のうち何人かがヒットを打つ確率（確率質量関数と累積分布関数）

確率変数の値 k は「ヒットを打つ人数」。f(k) が確率質量関数の値。例えば、n = 3 人のうち、k = 0 人がヒットを打つ確率 f(0) は **0.343**。F(k) は累積分布関数の値。例えば、n = 3 人のうち、ヒットを打つのが k = 1 人以下である確率 F(1) は **0.784**。棒グラフは確率質量関数を、折れ線グラフは累積分布関数を可視化したもの。

ここから、図 3 に示した f(k) の具体的な値が、(3) 式でどのようにして求められるかを見ていきます。3 人の打者のうち何人かがヒットを打つ確率について、全ての場合を洗い出してみましょう。

3 人の 3 割打者が 3 人ともヒットを打つ確率

図 3 の確率変数の値 k は昇順に並んでいますが、逆に、k の値の大きい方から見ていきます。つまり、3 人ともヒットを打つ確率からです。これは簡単です。図 4 をご覧ください。

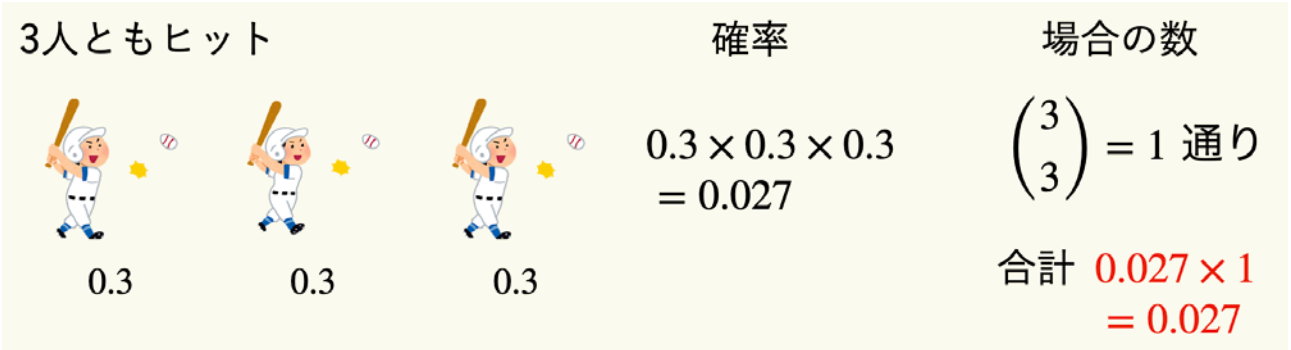


図 4　3 割打者が 3 人ともヒットを打つ確率

3 人ともヒットを打つ確率は 0.3×0.3×0.3 = **0.027**。つまり、**2.7%**となる。3 連打が出るのはかなり確率の低い事象。実際のところ、投手の調子によってはメッタ打ちにされることもあるが、あくまでも常に 3 割の確率でヒットが出るとした場合の話。

3 人ともヒットを打つ確率は 0.3×0.3×0.3 = **0.027** です。このような場合は 1 通りしかありません。その、1 通りというのは、3 人の中から 3 人を選ぶ場合の数です。他に選びようがないですね。

→目次に戻る

ここで、注意すべき点は（本来は、もっと早く言っておくべきことなのですが）、それぞれの試行が**独立**であることが大前提になっているということです。前の打者の結果によって、次の打者の行動（ヒットの確率）が変わるとすれば、独立ではありません（そのような関係を**従属**と呼びます）。実際の試合では、前の打者が 3 塁打を打ったので次の打者は外野フライでもいいとか、強打者であっても 1 点を競っている場合はスクイズを試みるといったことがありますが（……といったところが野球の面白さではあるのですが）、ここでは、前の打者の結果とは関係なく、次の打者は常に 3 割の確率でヒットを打つものとしています。

3 人の 3 割打者のうち 2 人がヒットを打つ確率

次に、2 人がヒットを打つ確率を考えてみましょう。2 人がヒットを打つ確率は 0.3×0.3×0.7 で求められます。ただし、1 人目と 2 人目がヒット、1 人目と 3 人目がヒット、2 人目と 3 人目がヒットという 3 通りの場合があります（図 5）。

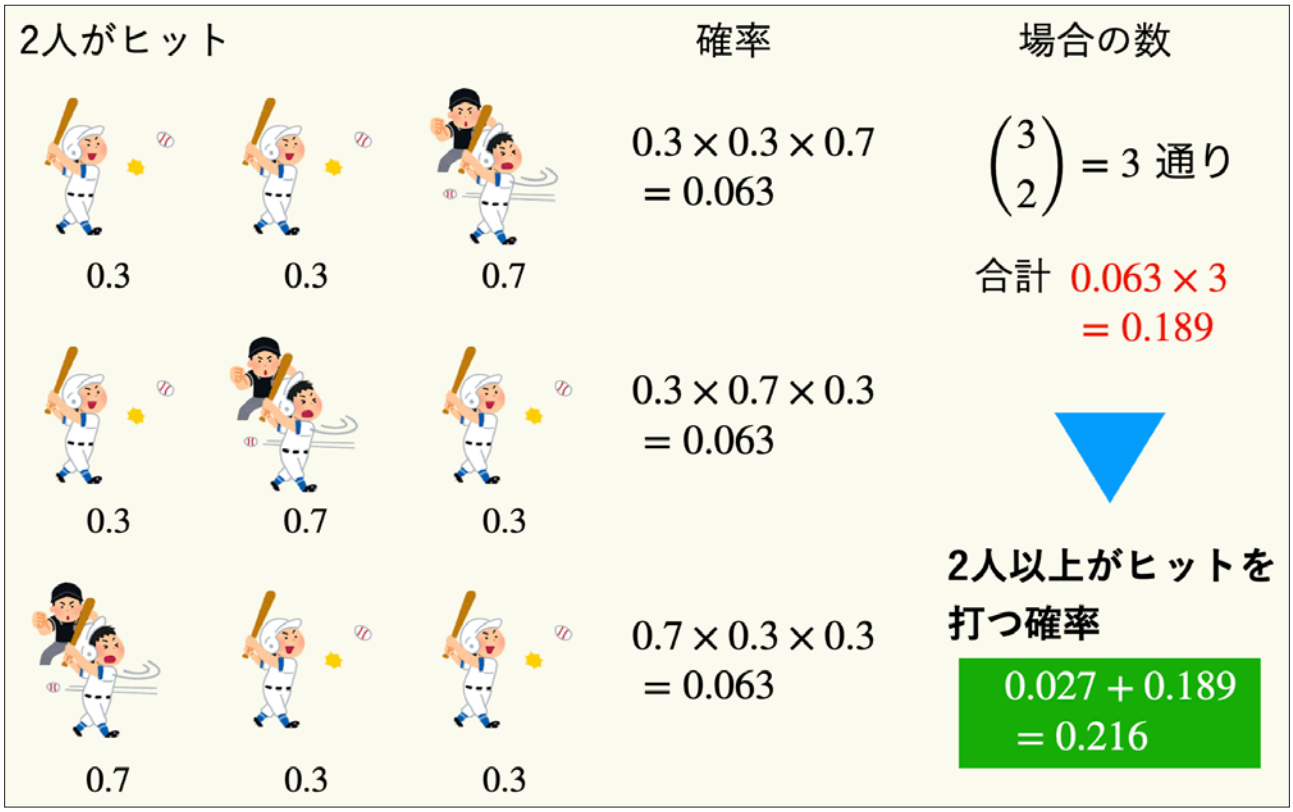


図 5　3 人の 3 割打者のうち 2 人がヒットを打つ確率

2 人がヒットを打つのは 3 通りの場合があり、それぞれの確率は 0.3×0.3×0.7 = **0.063** となる（掛け算の順序が異なっているだけでいずれも **0.063** となる）。従って、合計すると 0.063×3 = **0.189** となる。

図 4 と図 5 から、2 人以上がヒットを打つ確率は 0.027 + 0.189 = **0.216** となることが分かります。2 割ちょっとですね。上でも述べたように実際には一筋縄にはいきませんが、3 割打者が 3 人続いたとしても 8 割方は得点されないと考えれば、投手も気が楽になるのではないでしょうか。

さて、ここで、少し計算を一般化してみましょう。少しゆっくりと考えながら読んでみてください。

→目次に戻る

図 5 の場合。成功確率を p とすると、それぞれのヒットが出る確率は $p^2 \times (1-p)^1$ となっていることが分かります。p の指数の **2** は成功数、つまり k の値ですね。一方の (1 − p) の指数の **1** は、試行数を n とすると、n − k の値です。n は試行数の **3**、 k は成功数の **2** なので n − k = 1 ですね。ということで、それぞれの確率は、以下のように表されます。

$$p^k(1-p)^{n-k} \qquad (4)$$

これが 3 通りあったわけです。3 通りというのは、3 人の中からヒットを打つ 2 人を選ぶ場合の数です。つまり、

$$\binom{3}{2}$$

です。**3** というのは n の値、**2** というのは k の値なので、一般に、

$$\binom{n}{k} \qquad (5)$$

と表されます。これらを掛けると、以下のようになりますね。

$$\binom{n}{k} p^k(1-p)^{n-k} \qquad (6)$$

これは（3）式の右辺です。以下に再掲します。

$$f(k) = \binom{n}{k} p^k(1-p)^{n-k} \qquad (3)$$

組み合わせ数の求め方

（5）式の組み合わせ数の求め方も確認しておきましょう。以下の公式を使います。

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad (7)$$

! は**階乗**を表します。n! は、n × (n − 1) × ⋯ × 1 です。ただし、この公式を正直に使うよりも、以下の方法で計算するのが楽です。

$$\binom{n}{k} = \frac{\overbrace{n \times (n-1) \times \cdots}^{k個}}{k!}$$

n = 3、k = 2 でやってみましょう。以下のようになりますね。穴埋め問題にしています。

$$\binom{3}{2} = \frac{3 \times \boxed{ア}}{\boxed{イ} \times 1} = 3$$

→目次に戻る

答え：ア＝２　、イ＝２

なお、以下の等式が成り立ちます。

$$\binom{n}{k} = \binom{n}{n-k} \qquad (8)$$

例えば、n = 12 個のうちから k = 9 個選ぶ組み合わせ数は、

$$\binom{12}{9} = \frac{12 \times 11 \times 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4}{9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} = 220$$

で計算するよりも、

$$\binom{12}{12-9} = \binom{12}{3} = \frac{12 \times 11 \times 10}{3 \times 2 \times 1} = 220$$

の方が簡単です。

続いて、１人しかヒットを打たない場合と、１人もヒットを打たない場合についても、確率を求めてみましょう。

3 人の 3 割打者のうち 1 人がヒットを打つ確率

１人がヒットを打つ確率は 0.3×0.7×0.7 で求められます。ただし、１人目がヒット、２人目がヒット、３人目がヒットという３通りの場合があります（図 6）。

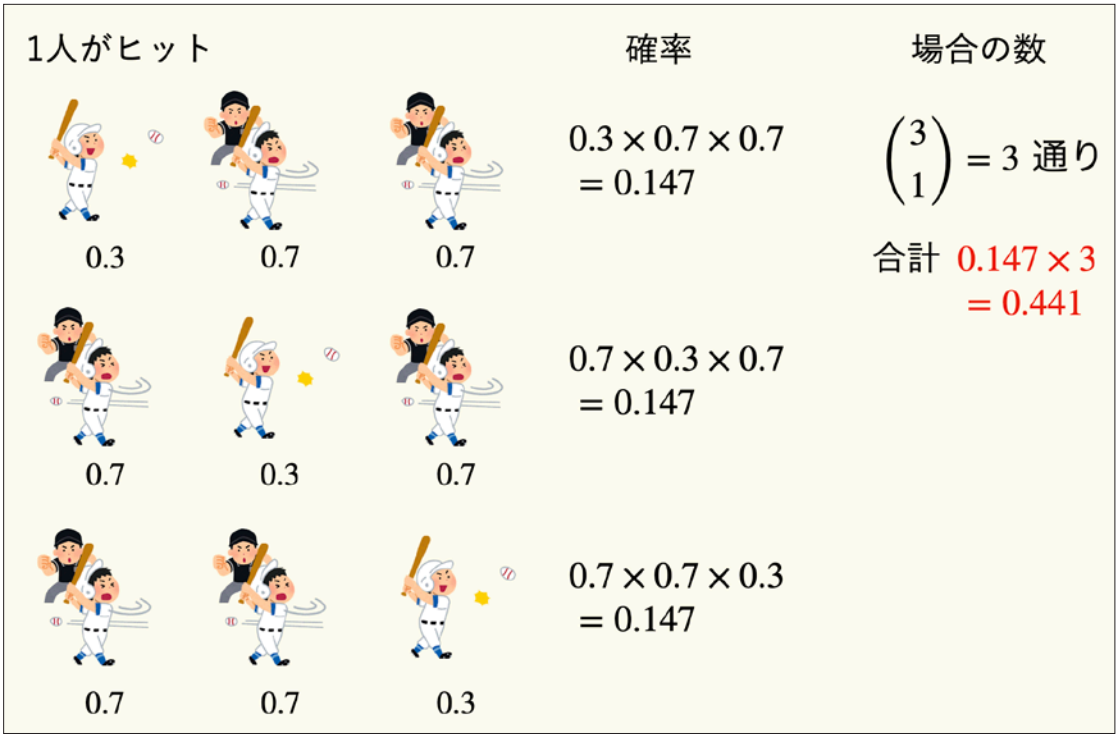


図 6　3 人の 3 割打者のうち 1 人がヒットを打つ確率

１人がヒットを打つのは３通りの場合があり、それぞれの確率は 0.3×0.7×0.7 = **0.147** となる（掛け算の順序が異なっているだけでいずれも **0.147** となる）。従って、合計すると 0.147×3 = **0.441** となる。

→目次に戻る

この例については、公式に従って計算してみましょう。

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad (3)$$

に n = 3、k = 1 を代入すればいいですね。まず、組み合わせ数を計算しておきましょう。k = 1 のときは、計算するまでもなく、

$$\binom{n}{1} = n$$

なので、組み合わせ数は **3** です。p = 0.3 なので、以下のようになります。これも穴埋め問題にしています。

$$\begin{aligned} f(1) &= \binom{3}{1} 0.3^1 (1-\boxed{ア})^{3-\boxed{イ}} \\ &= 3 \times 0.3 \times 0.7^2 \\ &= 0.441 \end{aligned}$$

答え：ア= 0.3　、イ= 1

3 人の 3 割打者のうち 0 人がヒットを打つ確率

3 人のうち 0 人がヒットを打つということは、誰もヒットを打たない（全て確率は **0.7**）ということですね。これは簡単です（図 7）。

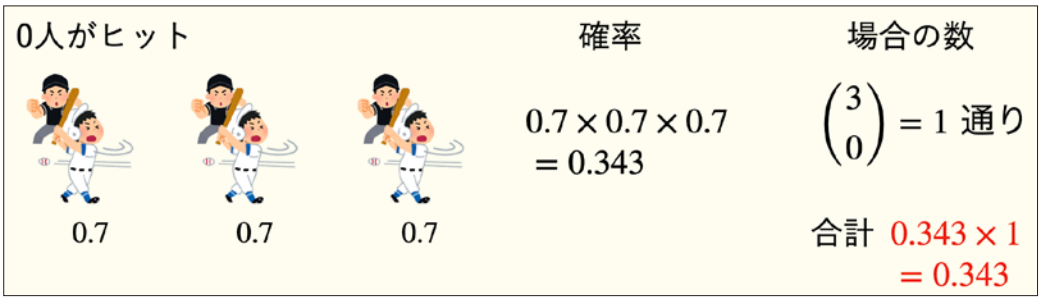


図 7　3 人の 3 割打者のうち 0 人がヒットを打つ確率

3 人ともヒットを打たない確率は 0.7×0.7×0.7 = **0.343**。つまり、**34.3%**となる。3 人の中から 0 人を選ぶ組み合わせ数は **1** なので（後述）、合計も **0.343** となる。

3 人の中から 0 人を選ぶ組み合わせ数というのはちょっと想像しづらいですが、（8）式に当てはめると、**1** であることが分かります。

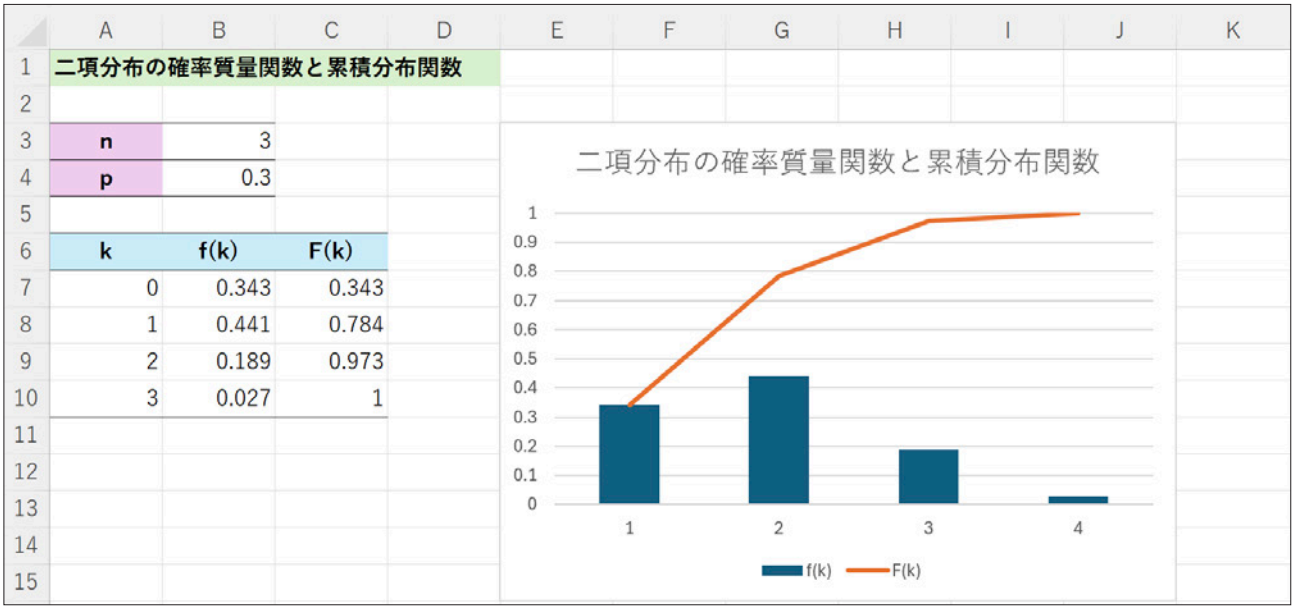
$$\begin{aligned} \binom{3}{0} &= \binom{3}{3-0} \\ &= \binom{3}{3} \\ &= 1 \end{aligned}$$

図 4 ～図 7 の確率を合計すると、0.027 + 0.189 + 0.441 + 0.343 = **1** となることも確認しておいてください。

→目次に戻る

二項分布の確率質量関数と累積分布関数を可視化してみよう

ここまでは、具体的な例を基に、(3) の公式を理解することに重点を置いてきました。ここからは、確率質量関数や累積分布関数の値を求めたり、可視化したりする方法を見ていきます。図 3 のような表やグラフを作成する方法を見ていこうというわけですね。図 3 を再掲して、その後に手順を示すことにします（図 8）。



二項分布の確率質量関数と累積分布関数		
n	3	
p	0.3	
k	f(k)	F(k)
0	0.343	0.343
1	0.441	0.784
2	0.189	0.973
3	0.027	1

図 8　3 人の 3 割打者のうち何人がヒットを打つか（確率質量関数と累積分布関数）

確率質量関数と累積分布関数の値は（3）式に従って計算してもよいが、いずれも BINOM.DIST 関数で求められる。

グラフ作成の手順は以下の通りです。サンプルファイルをこちらからダウンロードし、Excel で［二項分布］ワークシートを開いて試してみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。なお、Google スプレッドシートでの操作方法は、サンプルファイル内に記載しています。

◆ **Excel での操作方法**

- セル **B7** に =BINOM.DIST(A7:A10,B3,B4,FALSE) と入力する
  - ・古いバージョンの Excel でスピル機能が使えない場合は、あらかじめ結果が求められるセル範囲（セル **B7** ～ **B10**）を選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
- セル **C7** に =BINOM.DIST(A7:A10,B3,B4,TRUE) と入力する
  - ・スピル機能が使えない場合は、セル範囲（セル **C7** ～ **C10**）を選択して関数の入力終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
- セル **B6** ～ **C10** を選択する
- ［挿入］タブを開き、［複合グラフの挿入］ボタンをクリックして［集合縦棒 - 折れ線］を選択する
- ［グラフのデザイン］タブを開き、［データの選択］ボタンをクリックする
- ［データソースの選択］ダイアログボックスで［横（項目）軸ラベル］の下の［編集］ボタンをクリックする
- ［軸ラベル］ダイアログボックスで［軸ラベルの範囲］ボックスをクリックし、セル **A7** ～ **A10** を選択する
- ［OK］をクリックして［軸ラベル］ダイアログボックスを閉じる
- ［OK］をクリックして［データソースの選択］ダイアログボックスを閉じる

BINOM.DIST 関数の引数は以下のように指定します。

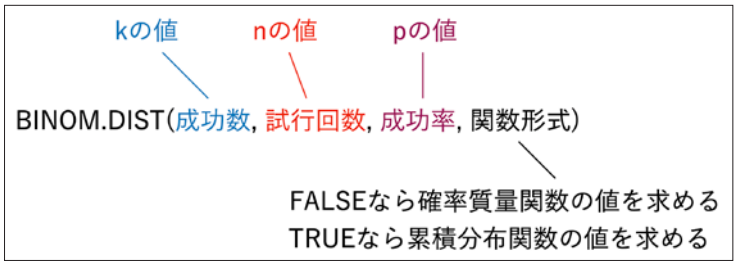


図 9　BINOM.DIST 関数に指定する引数

図 8 の例では、成功数（k に当たる値＝目的の事象が起こる回数）として A7:A10 というセル範囲を指定しているので、スピル機能によりセル **A7** ～ **A10** の成功数に対する確率質量関数の値や累積分布関数の値が一度に求められる。

なお、サンプルファイルには、（3）式に従って確率質量関数と累積分布関数の値を求めた例も含めてあります。興味のある方はご参照ください。

◇ ◇ ◇ ◇ ◇ ◇ ◇

今回は、離散型確率分布の基本的な例として、ベルヌーイ分布と二項分布を取り上げました。実は、二項分布は、連続型確率分布として最もよく使われる正規分布の基礎ともなっています。また、ある打者が例えば 10 打数のうち 4 回ヒットを打ったとき、（たまたま調子が良かっただけかもしれないので）その打者は 3 割打者であると自信を持って言えるか、といった検定（二項検定）のためにも使われます。これらのことについても、いずれ紹介したいと思います。

さて、次回は、離散型確率分布の別の例として、超幾何分布を取り上げます。超幾何分布は、非復元抽出（引いたくじを元に戻さないような場合）の確率を求めるのに使われます。次回もお楽しみに！

この記事で取り上げた関数の形式

関数の利用例については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

二項分布の確率質量関数や累積分布関数の値を求めるための関数

BINOM.DIST 関数：二項分布の確率質量関数や累積分布関数の値を求める

形式

BINOM.DIST(成功数 , 試行回数 , 成功率 , 関数形式)

引数

- **成功数：**目的の事象が起こる回数を指定する。
- **試行回数：**試行の総数を指定する。
- **成功率：**目的の事象が起こる確率を指定する。
- **関数形式：**以下の値を指定する。
  - **FALSE** …… 成功数に対する確率質量関数の値を求める
  - **TRUE** …… 成功数までの累積分布関数の値を求める

［データ分析］超幾何分布
～ くじ引き（非復元抽出）の確率を求める！

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載（確率分布編）の第 3 回。まず「非復元抽出（例：くじ引き）とは何か」を説明。その確率分布である超幾何分布を取り上げ、その意味や特徴などを解説します。

羽山博（2024 年 06 月 27 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の確率分布編、第 3 回です。前回は離散分布の基礎としてベルヌーイ分布と二項分布を取り上げました。今回は「非復元抽出とは何か」をお話しした後、その確率分布である超幾何分布を取り上げ、その意味や特徴などを見ていきます。

温泉旅行が何本か当選する確率を求めるには？

商店街の福引（ふくびき）で、1 等の温泉旅行にチャレンジする光景を思い浮かべてみてください。手元にある福引券 3 枚と交換に「ガラガラ」を 3 回、回したとき、1 等を 1 本引く確率を求めてみたいと思います。

ガラガラは「ガラポン」とも呼びます。機種にもよりますが、2000 ～ 3000 個の玉が入るものが多いようです。なお、正式名称は「新井式回転抽出機」だそうです。

この例は、コイン投げやサイコロ、あるいは前回の記事で見たヒットが出る確率と決定的に異なる点があります。それはいったい何でしょうか。図 1 を見ながら、違いを考えてみてください。

ここでは、話を分かりやすくするため、1 等の温泉旅行が 10 本中 3 本あるものとします。つまり、ガラガラには玉が 10 個入っていて、当たりが 3 個あるというわけです。実際にはそんなに当たりやすい福引はありませんが、ヒットが出る確率と比較しやすい値としました。その場合に、3 回のチャレンジのうち、1 等を 1 本引く確率を考えます。

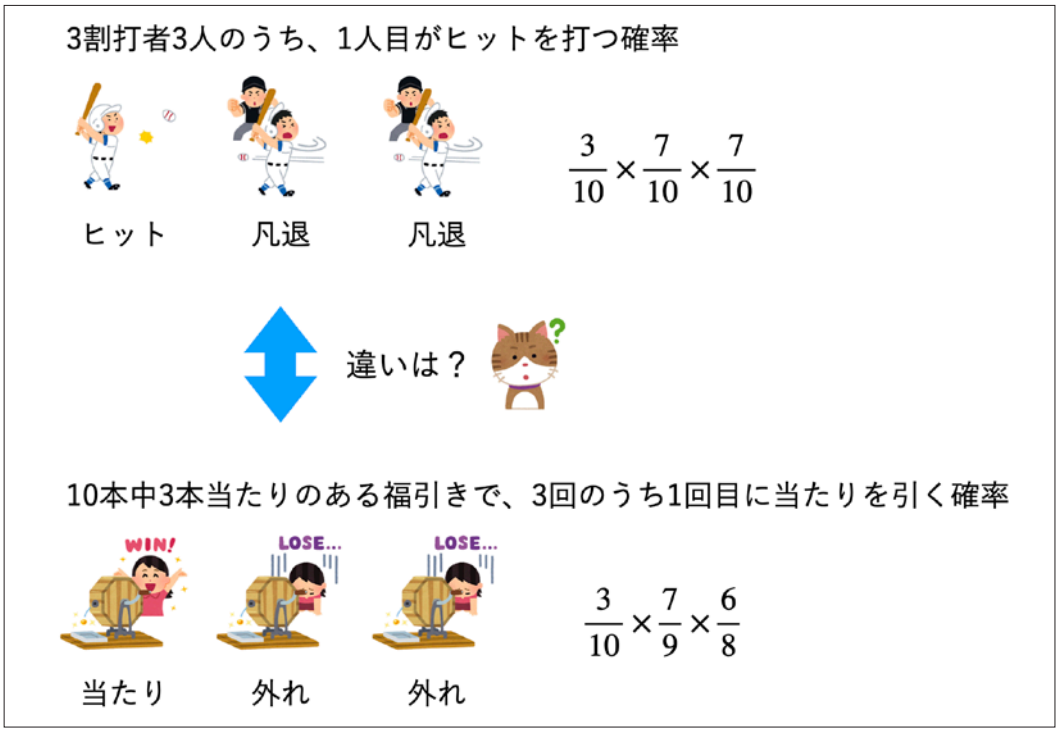


図 1　復元抽出と非復元抽出（3 人の打者がヒットを 1 本打つ確率と、3 回の福引で 1 等を 1 本引く確率の違い）

違いを分かりやすくするため、それぞれの確率を分数で表してある。ヒットが出る確率は常に 3/10（10 分の 3）。しかし、福引の場合は前の試行で当たりが出たかどうかによって次の確率が変わる。図には、1 本目がヒット（当たり）の確率だけを示している。実際には、2 本目がヒット（当たり）の場合と 3 本目がヒット（当たり）の場合もある。それらの場合も含めた確率については後述する。

野球の例では、それぞれの打者がヒットを打つかどうかは**独立**であるという前提でした。つまり、前の打者がヒットであっても、凡退であっても、次の打者がヒットを打つ確率は 3/10 ＝ **0.3**、凡退する確率は 7/10 ＝ **0.7** となります。

しかし、福引の場合は、ガラガラを回すたびに玉が 1 つずつ減っていきます。前回の試行で当たりが出たか、外れが出たかによって、次の試行で当たりが出る確率が変わります。例えば、最初が当たりである確率は 3/10 ＝ **0.3** ですが、次は玉が 1 つ減るので全体は 9 個となり、当たりも 2 個に減ります。外れは相変わらず 7 個のままです。その場合、2 番目の玉が当たりである確率は 2/9 で、外れである確率は 7/9 です。このように、前の試行の結果が次の試行の確率に影響するような場合を**従属**と呼びます。

野球の例は、例えば赤玉が 3 個、白玉が 7 個の合計 10 個の玉が入っている壺（つぼ）から玉を取り出した後、取り出した玉を**元に戻す**場合と同じです。玉の総数も赤玉と白玉の個数も変わりません。このような取り出し方を**復元抽出**と呼びます。

一方、福引の例は、取り出した玉を**元に戻さない**場合と同じです。玉を取り出すたびに玉の個数は 1 つずつ減っていきます。また、赤玉を取り出せば赤玉の個数が減り、白玉を取り出せば白玉の個数が減ります。このような取り出し方を**非復元抽出**と呼びます。

コラム　条件付き確率とは

このコラムは、話の流れから少し外れるので、先に進みたい方は読み飛ばしていただいても構いません。

福引の例では、1 回目で当たりを引いた後に外れを引く確率と、1 回目で外れを引いた後にまた外れを引く確率は異なりますね。「1 回目に当たりを引く」という事象を A と表し、「2 回目に外れを引く」という事象を B と表すと、1 回目に当たりを引いたという条件の下で、2 回目に外れを引く確率は P(B|A) と表わされます。この P(B|A) は**条件付き確率**と呼ばれます。

条件付き確率については、以下の式（条件付き確率の乗法公式）が成り立ちます。

$$P(A \cap B) = P(A)P(B|A)$$

- P(A ∩ B) …… 1 回目に当たりを引き、かつ 2 回目に外れを引く確率（1 回目と 2 回目の両方を満たす確率のこと）
- P(B|A)…… 1 回目に当たりを引いた場合に、2 回目に外れを引く確率（2 回目の確率のこと）

図 1 の 1 回目と 2 回目だけを取り上げると、以下のようになります。

$$P(A) = \frac{3}{10}$$
$$P(B|A) = \frac{7}{9}$$

なので、

$$P(A \cap B) = P(A)P(B|A) = \frac{3}{10} \times \frac{7}{9}$$

となり、図 1 の福引きの例で見た最初の 2 回に当てはまります。

条件付き確率の乗法公式はベイズ統計学の出発点となる公式です。詳細については、［AI・機械学習の数学］機械学習でよく使われる「ベイズの定理」を理解するをご参照ください。

超幾何分布は非復元抽出の確率分布

では、福引の例で、3 回のうち 1 回当たりを引く確率を求めてみます。図 1 では 1 本目が当たりの例を見ましたが、2 本目が当たりの場合と、3 本目が当たりの場合もあります。

非復元抽出で、総数が N 個、総数のうちの成功数が M 回、試行回数が n 回、試行のうちの成功数が k 回の場合の確率分布を**超幾何分布**（ちょうきかぶんぷ、Hypergeometric distribution）と呼びます。最初に超幾何分布の公式を示しておきます。確率変数が取る値を k とします。後で具体的な例を見ながら超幾何分布の意味を解きほぐしていくので、公式が理解できなくても気にせず先に進めてください（計算についても、後述する Excel の HYPGEOM.DIST 関数を使えば簡単にできます）。

$$f(k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \qquad (1)$$

この式にある ( ) は、ベクトルや行列ではなく、組み合わせ数を表します。計算方法については、前回の記事で説明しましたので、忘れた場合はそちらを復習してください。

まずは、N、M、n、k が、福引の例でどの値に当たるかを確認しておきましょう。

値	意味	福引の例での値は？
N	総数	10
M	総数のうちの成功数	3 ← 1等の玉の個数
n	試行回数	3 ← 何回ガラガラを回せるか
k	試行のうちの成功数	1 ← 1等の玉が出た個数

（1）式に値を当てはめると、当たりが 1 本出る確率が求められます（図 1 で見た 1 本目が当たりの場合だけでなく、2 本目が当たりの場合、3 本目が当たりの場合も含めた確率です）。( ) で表す組み合わせ数の計算方法は後で説明します。ここでは計算結果だけを確かめてください。

$$f(1) = \frac{\binom{\boxed{ア}}{1}\binom{\boxed{イ}-3}{3-\boxed{ウ}}}{\binom{10}{\boxed{エ}}} = \frac{3 \cdot 21}{120} = \frac{21}{40}$$

答え：ア＝ 3、イ＝ 10、ウ＝ 1　、エ＝ 3

取りあえず確率は求められました。ここからは福引の例を具体的に見ながら、上の結果と一致することを確認します。また、上の公式の意味についても、あらためて説明します。

すでに述べたように、図 1 で見たのは 3 回の試行のうち、1 回目に当たりを引く確率でした。3 回の試行のうち、当たりを 1 本引くのは、2 回目に当たりを引く場合と、3 回目に当たりを引く場合があります。これらを全て図にしてみましょう（図 2）。

図 2　10 本中 3 本の当たりがある福引に 3 回チャレンジして当たりを 1 本引く確率

1 回チャレンジするたびにガラガラの玉が 1 つずつ減っていくので、分母は 10×9×8。当たりを引くと当たりの玉が減り、外れを引くと外れの玉が減るので、分子は 3（当たりの場合の数）と 7×6（外れの場合の数）を掛けたものになる。それらが 3 通りあるので、後述する式で確率が求められる。

→目次に戻る

いずれの場合も、分母は 10×9×8 で、分子は（掛け算の順序は異なりますが）3×7×6 です。3 通りの場合があるので、求める確率は以下のようになります。

$$\frac{3\cdot 7\cdot 6}{10\cdot 9\cdot 8}\times 3=\frac{\cancel{3}\cdot 7\cdot 6}{10\cdot \cancel{9}\cdot 8}\times \cancel{3}$$

$$=\frac{7\cdot \overset{3}{\cancel{6}}}{10\cdot \underset{4}{\cancel{8}}}$$

$$=\frac{21}{40}$$

公式で求めた結果と一致しましたね。ここであらためて公式を確認しておきましょう。

$$f(k)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \qquad (1)$$

福引の例の N = 10, M = 3, n = 3, k = 1 を当てはめて、意味を見てみます（図 3）。N − M = 10 − 3 = 7, n − k = 3 − 1 = 2 です。

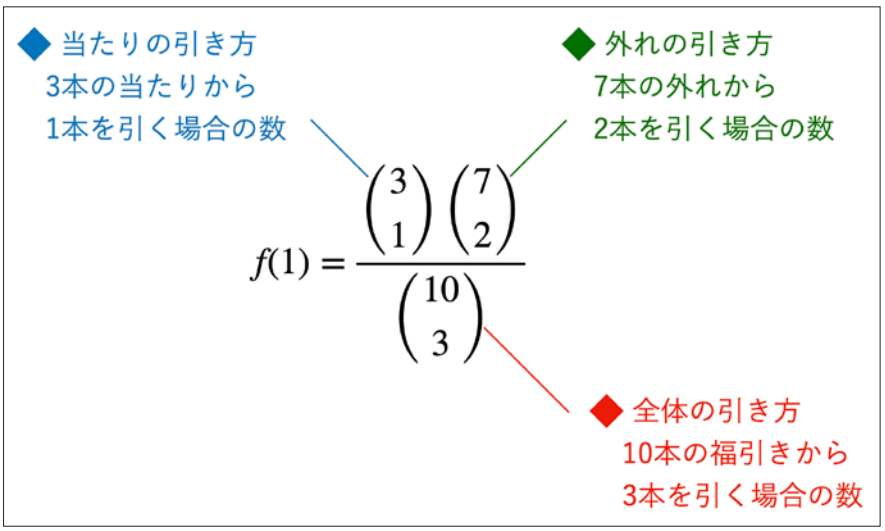


図 3　非復元抽出の確率を具体的に見ていく

公式は複雑そうに見えるが、結局のところ、全体の引き方の中で、当たりを引く引き方と外れを引く引き方が何通りなのかということ。

→目次に戻る

図 3 の式を計算してみましょう。分母は以下のようになります。

$$\binom{10}{3} = \frac{10 \cdot \cancel{9}^{3} \cdot \cancel{8}^{4}}{\cancel{3} \cdot \cancel{2} \cdot 1} = 120$$

分子は以下のようになります。

$$\binom{3}{1}\binom{7}{2} = \frac{3}{1} \cdot \frac{7 \cdot \cancel{6}^{3}}{\cancel{2} \cdot 1} = 63$$

従って、f(1) の値は以下のようになり、図 2 での計算と一致しますね。

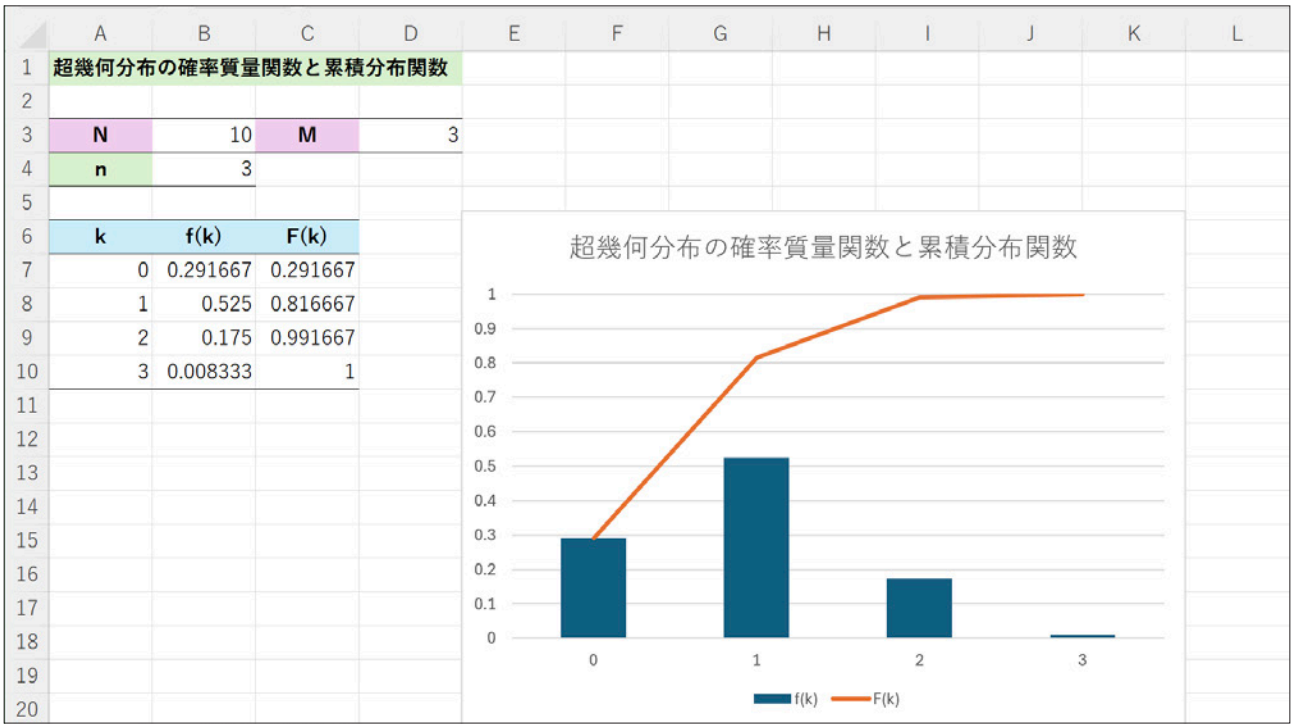
$$f(1) = \frac{\binom{3}{1}\binom{7}{2}}{\binom{10}{3}} = \frac{63}{120} = \frac{21}{40}$$

→目次に戻る

超幾何分布の確率質量関数と累積分布関数を可視化してみよう

ここまでは、当たりを 1 本引く確率しか見てきませんでした。つまり、確率変数の値 k = 1 の場合だけでした。福引の例であれば、全て外れ（k = 0）から、3 本とも当たり（k = 3）の場合までがあるので、それらの確率を全て求め、確率質量関数と累積分布関数を可視化してみましょう。

（1）式を使って計算することもできますが、Excel の HYPGEOM.DIST 関数を使えば簡単です。結果は図 4 のようになります。グラフ作成の手順は図 4 の後に記しておきます。



	A	B	C	D
1	超幾何分布の確率質量関数と累積分布関数			
2				
3	N	10	M	3
4	n	3		
5				
6	k	f(k)	F(k)	
7	0	0.291667	0.291667	
8	1	0.525	0.816667	
9	2	0.175	0.991667	
10	3	0.008333	1	

図 4　超幾何分布の確率質量関数と累積分布関数

確率質量関数と累積分布関数の値は（1）式に従って計算してもよいが、いずれも HYPGEOM.DIST 関数で求められる。

グラフ作成の手順は以下の通りです。サンプルファイルをこちらからダウンロードし、[超幾何分布］ワークシートを開いて試してみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。なお、Google スプレッドシートの HYPEGEOM.DIST 関数には最後の引数がない（確率質量関数の値しか求められない）ことに注意が必要です。具体的な操作方法は、サンプルファイル内に記載しています。

◆ **Excel での操作方法**

- セル **B7** に =HYPGEOM.DIST(A7:A10,B4,D3,B3,FALSE) と入力する
  - ・古いバージョンの Excel でスピル機能が使えない場合は、あらかじめ結果が求められるセル範囲（セル **B7** ～ **B10**）を選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
- セル **C7** に =HYPGEOM.DIST(A7:A10,B4,D3,B3,TRUE) と入力する
  - ・古いバージョンの Excel でスピル機能が使えない場合は、あらかじめ結果が求められるセル範囲（セル **C7** ～ **C10**）を選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
- セル **B6** ～ **C10** を選択する
- ［挿入］タブを開き、［複合グラフの挿入］ボタンをクリックして［集合縦棒 - 折れ線］を選択する
- ［グラフのデザイン］タブを開き、［データの選択］ボタンをクリックする
- ［データソースの選択］ダイアログボックスで［横（項目）軸ラベル］の下の［編集］ボタンをクリックする
- ［軸ラベル］ダイアログボックスで［軸ラベルの範囲］ボックスをクリックし、セル **A7** ～ **A10** を選択する
- ［OK］をクリックして［軸ラベル］ダイアログボックスを閉じる
- ［OK］をクリックして［データソースの選択］ダイアログボックスを閉じる

HYPEGEOM.DIST 関数の引数は図 5 のように指定します。**母集団**とは集団全体のことで、福引の例で言えば、ガラガラの中に入っている玉全体のことです。**標本**とは母集団から取り出したもののことです。図 5 には、超幾何分布の確率質量関数の公式も併せて示してあるので、対応を確認しておいてください。

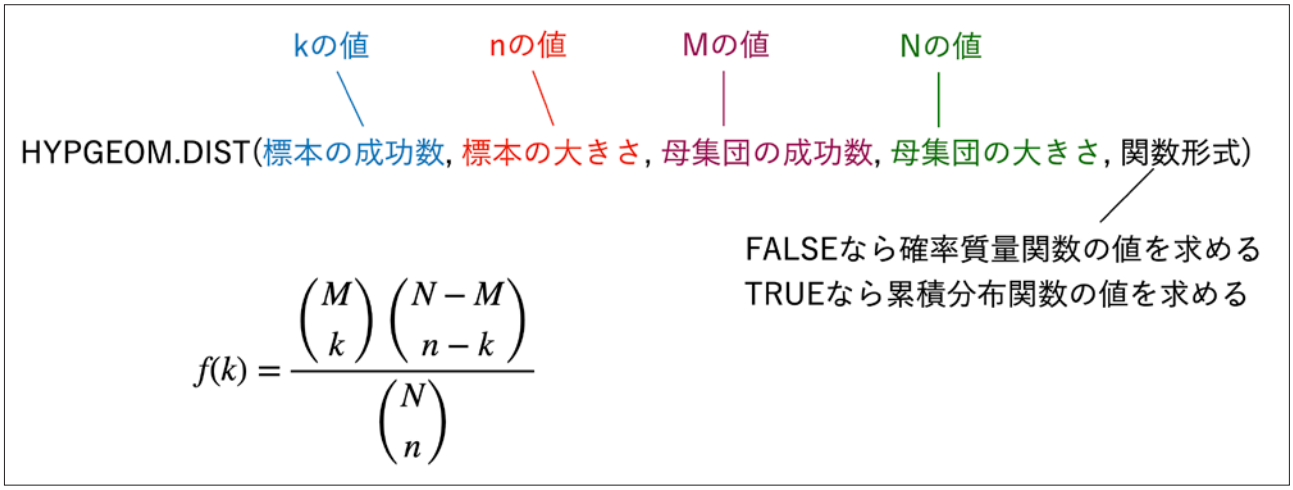


図 5　HYPGEOM.DIST 関数に指定する引数

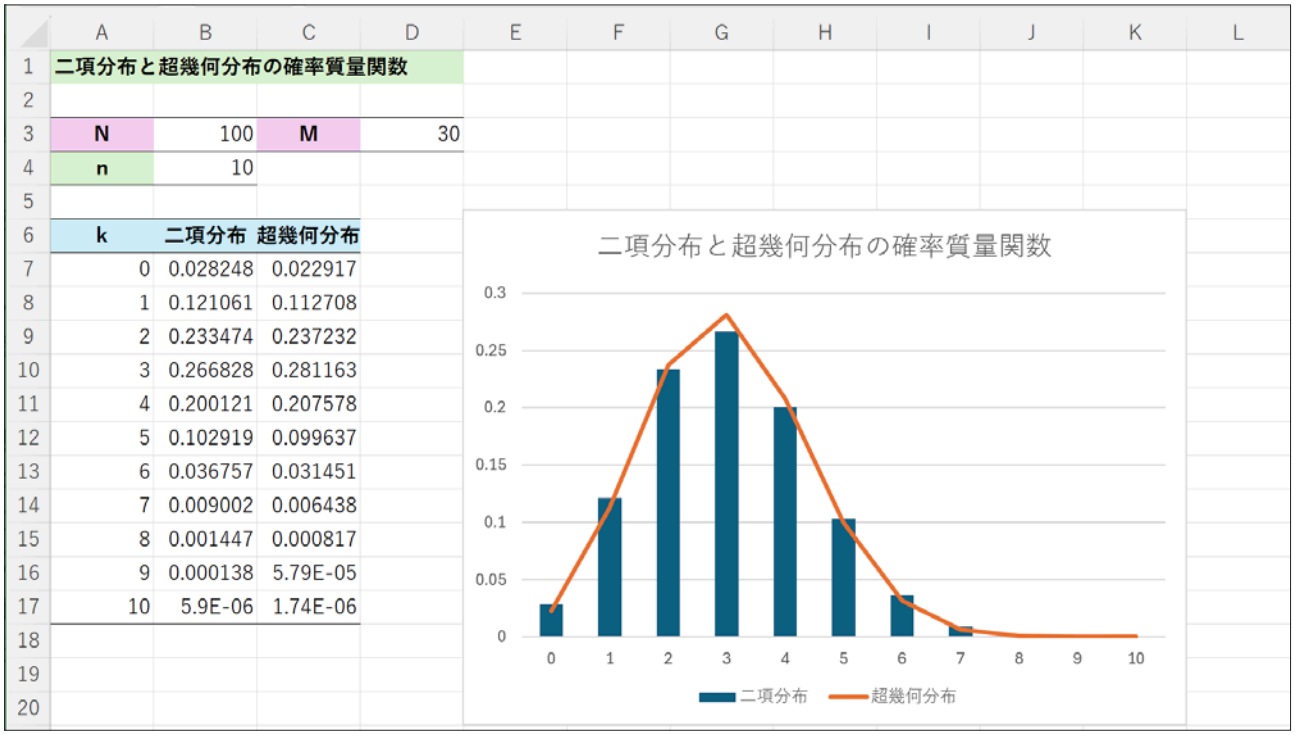
図 4 の例では、標本の成功数（k に当たる値＝目的の事象が起こる回数）として A7:A10 というセル範囲を指定しているので、スピル機能によりセル **A7** ～ **A10** の成功数に対する確率質量関数の値や累積分布関数の値が一度に求められる。Excel のヘルプには、2 番目の引数が「標本数」と記されているが、標本数という用語は、標本の抽出を何回行ったかを表したり、標本が何種類あるか（グループの数）を表したりするのに使われるので、誤解のないように「標本の大きさ」とした。

なお、サンプルファイルには、（1）式に従って確率質量関数と累積分布関数の値を求めた例も含めてあります。興味のある方はご参照ください。

→目次に戻る

二項分布と超幾何分布

二項分布と超幾何分布の違いは、復元抽出であるか非復元抽出であるかということです。母集団が大きくなると二項分布と超幾何分布の値はほぼ等しくなります。例えば、100 万本のくじから 1 本引いたとしても、残りの数はほとんど変わらないからです（残りは 99 万 9999 本ですね）。サンプルファイルには当たりが 30 本ある 100 本のくじを 10 回引いたときの、二項分布と超幾何分布の f(0) ～ f(10) の値を求め、グラフを描いた例を含めてあります（図 6）。N = 100 でもかなり近い値になっていることが分かります。



	A	B	C	D
1	二項分布と超幾何分布の確率質量関数			
2				
3	N	100	M	30
4	n	10		
5				
6	k	二項分布	超幾何分布	
7	0	0.028248	0.022917	
8	1	0.121061	0.112708	
9	2	0.233474	0.237232	
10	3	0.266828	0.281163	
11	4	0.200121	0.207578	
12	5	0.102919	0.099637	
13	6	0.036757	0.031451	
14	7	0.009002	0.006438	
15	8	0.001447	0.000817	
16	9	0.000138	5.79E-05	
17	10	5.9E-06	1.74E-06	

図 6　N の値が大きいときの二項分布と超幾何分布

二項分布の確率質量関数を棒グラフで、超幾何分布の確率質量関数を折れ線グラフで描いてみた。N = 100 の場合でも、二項分布と超幾何分布がほぼ重なることが分かる。

◇ ◇ ◇ ◇ ◇ ◇ ◇

今回は、非復元抽出の確率分布として超幾何分布を取り上げました。超幾何分布はここで見た例の確率を求めるだけでなく、データの件数が少ない場合の独立性の検定などにも使われます。これらのことについても、いずれ紹介したいと思います。

さて、次回は、離散型確率分布の別の例として、まれにしか起こらない事象の確率を求めるために使われるポアソン分布を取り上げます。次回もお楽しみに！

この記事で取り上げた関数の形式

関数の利用例については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

超幾何分布の確率質量関数や累積分布関数の値を求めるための関数

HYPGEOM.DIST 関数：超幾何分布の確率質量関数や累積分布関数の値を求める

形式

HYPGEOM.DIST(標本の成功数 , 標本の大きさ , 母集団の成功数 , 母集団の大きさ , 関数形式)

引数

- **標本の成功数：** 標本の中で目的の事象が起こる数を指定する。
- **標本の大きさ：** 母集団から取り出した標本の数（試行の総数）を指定する。
- **母集団の成功数：** 母集団の中での目的の事象が起こる数を指定する。
- **母集団の大きさ：** 母集団全体の数を指定する。
- **関数形式：** 以下の値を指定する。
  - **FALSE** …… 標本の成功数に対する確率質量関数の値を求める
  - **TRUE** …… 標本の成功数までの累積分布関数の値を求める

［データ分析］ポアソン分布 ～ 100 年に 1 人の天才は何人現れる？

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載（確率分布編）の第 4 回。ポアソン分布とは、出来事（事象）が、まれにしか起こらない場合に、独立な試行を何回も繰り返したときの確率分布です。そのような事例を紹介した後、確率の求め方や可視化の方法などを解説していきます。

羽山博（2024 年 07 月 11 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の確率分布編、第 4 回です。前回は、福引などのように、引いたくじを元に戻さない**非復元抽出**の例と、その確率分布である**超幾何分布**を取り上げました。今回は、まれにしか起こらない事象の例を紹介した後、その確率分布であるポアソン分布を取り上げ、その意味や特徴などを見ていきます。

100 人に 1 人の天才は何人いる？

「最近の日本人で 100 年に 1 人の天才は」というと、皆さんは誰の名前を思い浮かべるでしょうか。ChatGPT に聞いてみたところ、将棋の藤井聡太、野球の大谷翔平、医学者の山中伸弥が例として挙がりました（敬称略、以下同様）。皆さんが思い浮かべた人と一致したでしょうか（実際の ChatGPT では、質問の仕方などによって違う名前が表示されるかもしれません）。

日本人と限定せずに聞いてみたところ、アインシュタイン、レオナルド・ダ・ヴィンチ、モーツァルトの名が挙がりました。歴史上の人物なので、1000 年に 1 人といった方が適切かもしれませんね。今回登場するポアソン分布の式を導出したフランスの数学者シメオン・ドニ・ポアソンも、天才の一人かもしれません。ちなみに、筆者は、数学者ペレルマンの名前を挙げます。

それにしても、さまざまな分野で 100 年に 1 人の天才が登場しますね。その確率分布について考えてみましょう。100 年間で**平均して** 1 人の天才が登場することが分かっているものとして、100 年間のうちに 2 人の天才が現れる確率はどれぐらいでしょうか。

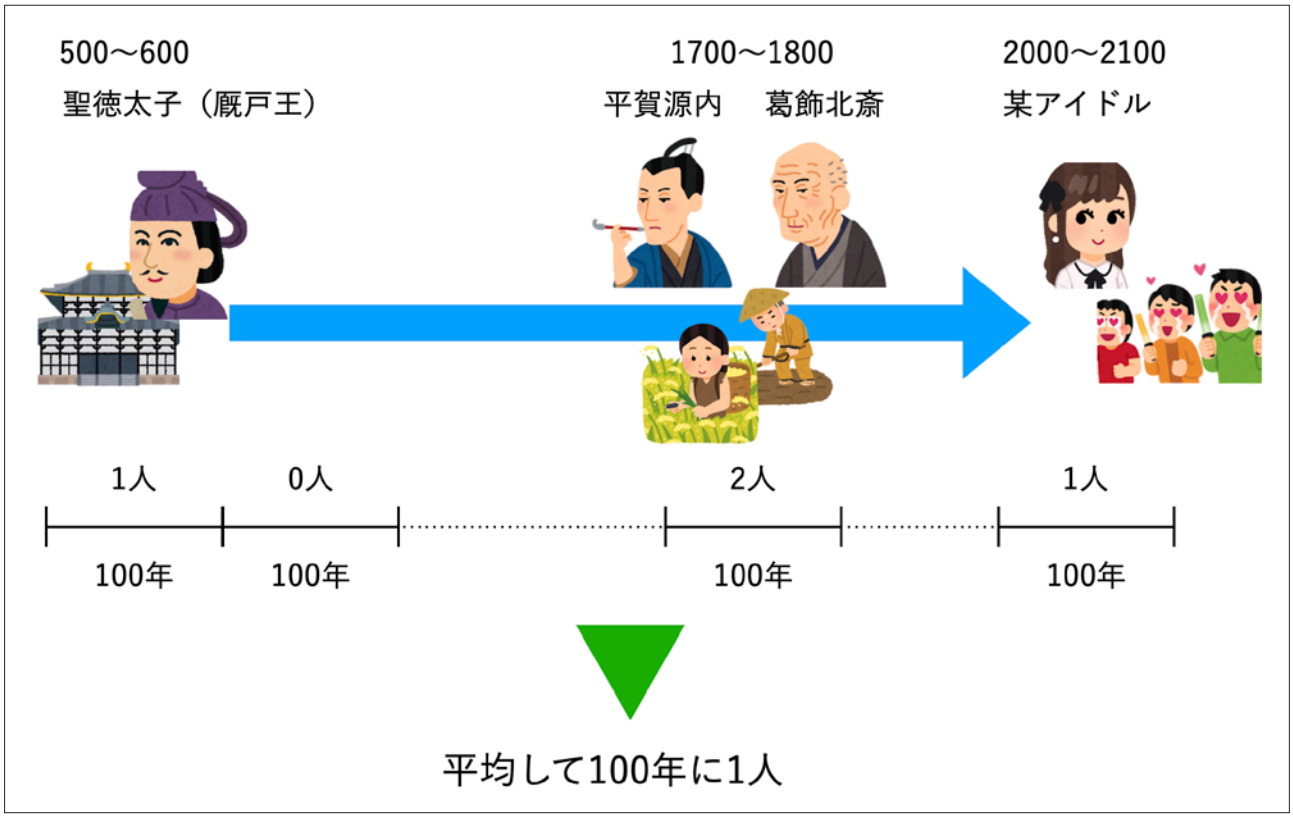


図 1　100 年に平均して 1 人の天才が登場するイメージ

平賀源内の生年は享保 3（1728）年、葛飾北斎の生年は宝暦 10（1760）年。なお、イラストに時代背景の異なるもの（飛鳥時代の聖徳太子（厩戸王：うまやどのおう）と奈良時代の東大寺など）があるがご容赦のほど。

このような、まれにしか起こらない事象の確率は**ポアソン分布**（Poisson distribution）と呼ばれる確率分布に従います。ポアソン分布では各試行は独立であることが前提となっています。まずは理屈抜きで、以下に示すポアソン分布の確率質量関数の式に当てはめて、確率を求めてみましょう。

$$f(k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad (1)$$

（1）式の k が確率変数の値です。つまり、何人天才が現れるかということですね。この場合は、100 年間のうちに 2 人の天才が現れる確率なので **2** です。

λ（ラムダ）は事象が起こる平均の数です。この場合は、平均して 1 人なので **1** ですね。

e は自然対数の底で、e = 2.7182…という値です。

ただ、細かい計算になるので、実際に計算していただかなくても構いません。数式中のλと k にどのような値を指定すればいいのかを確認しておいてください。

$$f(2) = \frac{\boxed{ア}^2 \times 2.7182^{-1}}{\boxed{イ}!} = 0.183945258$$

→目次に戻る

答え：ア＝ 1、イ＝ 2

第 1 回でも簡単に触れましたが、確率分布の形状を決定する定数値（例えば平均や標準偏差など）は**母数（パラメーター）**と呼ばれます。ポアソン分布の母数（＝ポアソン分布の確率質量関数を一意に決める値、つまりその確率分布の形状を完全に決定する値）はλであることも確認しておいてください。なお、k は確率変数の値（ある事象が起こる回数などの具体的な数値）であり、母数（パラメーター）ではありません。

コラム　まれにしか起こらないってどういうこと？ ～ 二項分布とポアソン分布の関係

すでに述べたように、ポアソン分布はまれにしか起こらない事象が起こる場合に、その事象が何回起こるかを求めるのに使われる確率分布です。しかし、「まれにしか起こらない」というのはずいぶんとあいまいな表現ですね。その意味を明確にしておきましょう。

実は、二項分布の試行回数を n、成功確率（目的の事象が起こる確率）を $p = \lambda / n$ とし、λを一定のままとし、試行回数 n を無限大に近づけていくと（その場合、p が小さくなっていきます）、平均λのポアソン分布になります。このことを式で表すと以下のようになります。

$$\lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} = \lim_{n\to\infty} \binom{n}{k} (\lambda/n)^k (1-\lambda/n)^{n-k}$$
$$= \frac{\lambda^k e^{-\lambda}}{k!}$$

この式にある最初の $\binom{n}{k}$ は、組み合わせ数を表します。計算方法については、第 2 回の記事で説明しましたので、忘れた場合はそちらを復習してください。

この式の証明は高校数学の知識でできますが、やや難しいので省略します。ここでは、目的の事象が起こる確率 p が小さい＝試行回数 n に比べ、事象が起こる回数λが小さい、ということが「まれにしか起こらない」ということだと理解していただければよいかと思います。このことについては、次の項で関数とグラフを使って確かめてみます。

ところで、二項分布の特殊な形であるなら、何もポアソン分布なんて使わずに二項分布だけで事足りるのでは、と思われるかもしれません。しかし、それぞれの確率密度関数（上の式の左辺と右辺）を見比べてみると、右辺のポアソン分布の方が簡単で、計算が楽であることが分かります。また、母数も二項分布では n と p の 2 つですが、ポアソン分布ではλだけなので、母数の推定も簡単であるなどのメリットがあります（母数の推定については推測統計編でお話する予定です）。

ポアソン分布の確率質量関数と累積分布関数を可視化してみよう

ここまでは、100 年に 1 人の天才が 2 人現れる確率だけを見てきました。つまり、確率変数の値 k = 2 の場合だけでした。当然のことながら 100 年に 1 人の天才が現れない（k = 0 である）場合もあるわけです。そこで、k = 0 ～ 5 までについてポアソン分布の確率質量関数と累積分布関数を可視化してみましょう。

（1）式を使って計算することもできますが、Excel の POISSON.DIST 関数を使えば簡単です。結果は図 2 のようになります。作成の手順は図 2 の後に記しておきます。

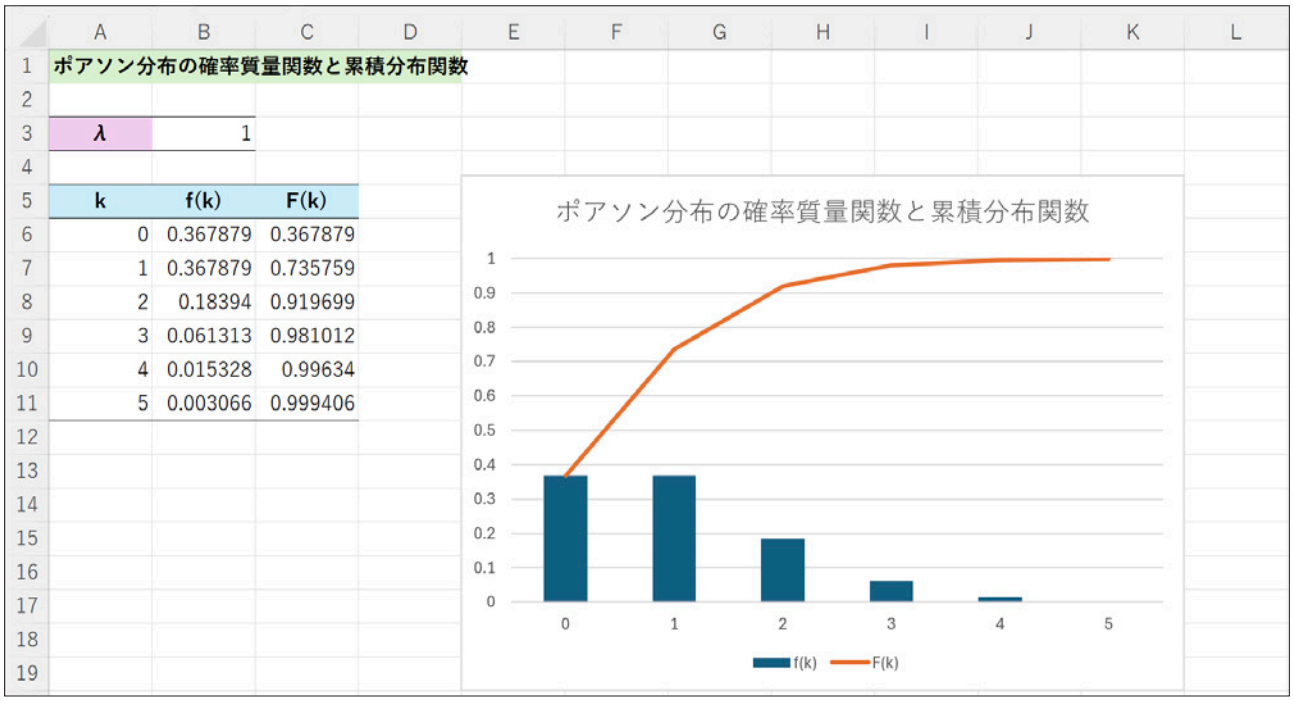


図 2　ポアソンの確率質量関数と累積分布関数

確率質量関数と累積分布関数の値は（1）式に従って計算してもよいが、いずれも POISSON.DIST 関数で求められる。

グラフ作成の手順は以下の通りです。サンプルファイルをこちらからダウンロードし、[ポアソン分布]ワークシートを開いて試してみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。具体的な操作方法は、サンプルファイル内に記載しています。

◆ **Excel での操作方法**

- セル **B6** に =POISSON.DIST(A6:A11,B3,FALSE) と入力する
  - ・古いバージョンの Excel でスピル機能が使えない場合は、あらかじめ結果が求められるセル範囲（セル **B6** ～ **B11**）を選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
- セル **C6** に =POISSON.DIST(A6:A11,B3,TRUE) と入力する
  - ・古いバージョンの Excel でスピル機能が使えない場合は、あらかじめ結果が求められるセル範囲（セル **C6** ～ **C11**）を選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
- セル **B5** ～ **C11** を選択する
- ［挿入］タブを開き、［複合グラフの挿入］ボタンをクリックして［集合縦棒 - 折れ線］を選択する
- ［グラフのデザイン］タブを開き、［データの選択］ボタンをクリックする
- ［データソースの選択］ダイアログボックスで［横（項目）軸ラベル］の下の［編集］ボタンをクリックする
- ［軸ラベル］ダイアログボックスで［軸ラベルの範囲］ボックスをクリックし、セル **A6** ～ **A11** を選択する
- ［OK］をクリックして［軸ラベル］ダイアログボックスを閉じる
- ［OK］をクリックして［データソースの選択］ダイアログボックスを閉じる

POISSON.DIST 関数の引数は図 3 のように指定します。母数がλだけなので簡単ですね。

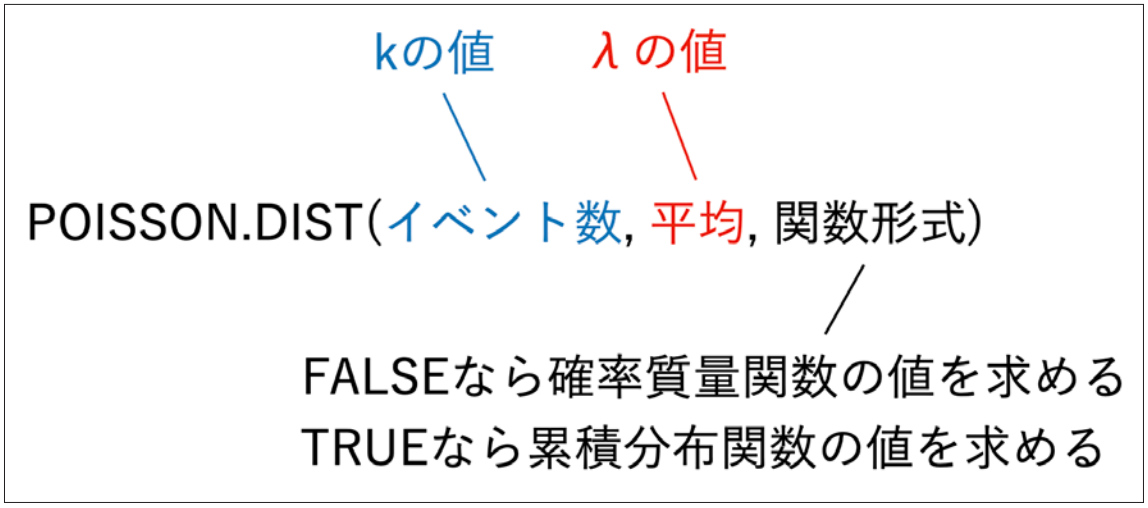


図 3　POISSON.DIST 関数に指定する引数

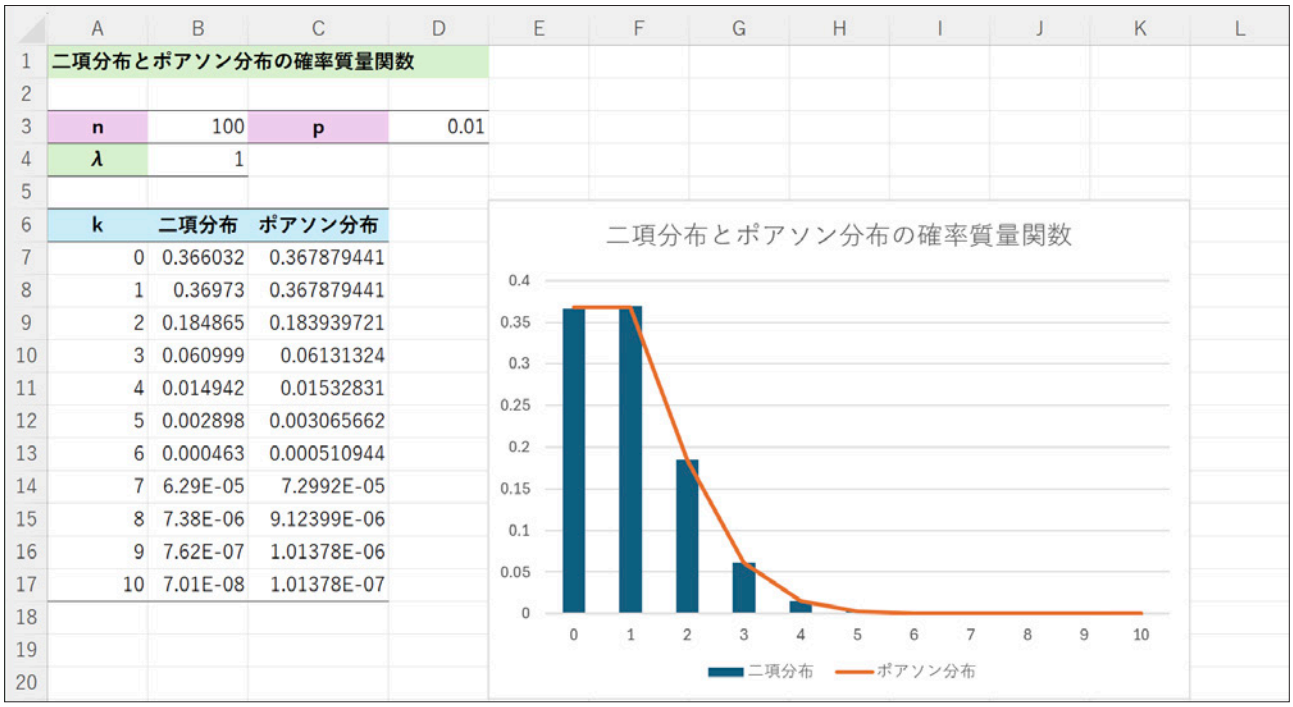
図 2 の例では、イベント数として A6:A11 というセル範囲を指定しているので、スピル機能によりセル A6 ～ A11 のイベント数に対する確率質量関数の値や累積分布関数の値が一度に求められる。イベント数とは、目的の事象が起こる回数、つまり k に当たる値のこと。

なお、サンプルファイルには、（1）式に従って確率質量関数と累積分布関数の値を求めた例も［ポアソン分布 (公式に従って作成)］ワークシートに含めてあります。興味のある方はご参照ください。

上のコラムでも述べたように、二項分布で p ＝λ /n とし、λを一定のままとし、試行回数 n を無限大に近づけていくと、平均λのポアソン分布になります。例えば、100 年に 1 人の天才が現れるなら、n ＝ 100、λ＝ 1 なので、p ＝ 1/100 です。

→目次に戻る

そこで、n = 100、p = 1/100 の二項分布の確率質量関数と、λ = 1 のポアソン分布の確率質量関数について k = 0 ～ 10 までの値を求め、可視化してみましょう（図 4）。サンプルファイルの［二項分布とポアソン分布］ワークシートを開いてください。作成の手順は図 4 の後に記しておきます。



	A	B	C	D
1	二項分布とポアソン分布の確率質量関数			
2				
3	n	100	p	0.01
4	λ	1		
5				
6	k	二項分布	ポアソン分布	
7	0	0.366032	0.367879441	
8	1	0.36973	0.367879441	
9	2	0.184865	0.183939721	
10	3	0.060999	0.06131324	
11	4	0.014942	0.01532831	
12	5	0.002898	0.003065662	
13	6	0.000463	0.000510944	
14	7	6.29E-05	7.2992E-05	
15	8	7.38E-06	9.12399E-06	
16	9	7.62E-07	1.01378E-06	
17	10	7.01E-08	1.01378E-07	
18				
19				
20				

図 4　n の値が大きく p の値が小さいときの二項分布とλ =np のポアソン分布

二項分布の確率質量関数を棒グラフで、ポアソン分布の確率質量関数を折れ線グラフで描いてみた。n = 100 の場合でも、二項分布とポアソン分布がほぼ重なることが分かる。n = 1000 や n = 10000 にすると、さらに値が近づく（セル B3 の値を変えて試してみるとよい）。

◆ Excel での操作方法

- セル **D3** に =B4/B3 と入力する
- セル **B7** に =BINOM.DIST(A7:A17,B3,D3,FALSE) と入力する
  - 古いバージョンの Excel でスピル機能が使えない場合は、あらかじめ結果が求められるセル範囲（セル **B7** ～ **B17**）を選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
- セル **C7** に =POISSON.DIST(A7:A17,B4,FALSE) と入力する
  - 古いバージョンの Excel でスピル機能が使えない場合は、あらかじめ結果が求められるセル範囲（セル **C7** ～ **C17**）を選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
- セル **B7** ～ **C17** を選択する
- ［挿入］タブを開き、［複合グラフの挿入］ボタンをクリックして［集合縦棒 - 折れ線］を選択する
- ［グラフのデザイン］タブを開き、［データの選択］ボタンをクリックする
- ［データソースの選択］ダイアログボックスで［横（項目）軸ラベル］の下の［編集］ボタンをクリックする
- ［軸ラベル］ダイアログボックスで［軸ラベルの範囲］ボックスをクリックし、セル **A7** ～ **A17** を選択する
- ［OK］をクリックして［軸ラベル］ダイアログボックスを閉じる
- ［OK］をクリックして［データソースの選択］ダイアログボックスを閉じる

ポアソン分布の応用例

これまで、一定の期間である事象が起こる確率を見てきました。ポアソン分布は他にも、お昼時の 1 時間に平均 10 人の顧客が訪れる（$\lambda = 10$ の）店舗で、1 時間に k 人の顧客が訪れる確率を求めたり、10000 個の製品の中で平均 1 個の不良品が発生する（$\lambda = 1$ の）場合に、10000 個中 k 個の不良品が発生する確率を求めたりするのに使えます。100 連ガチャを引いたときにレアアイテムが幾つか得られる確率なども求められますね。

店舗の例であれば、お昼時に多くて何人の顧客が訪れるかを見積もることができ、店員の最適な配置などに利用できます。この場合、1 時間に 18 ～ 20 人の顧客が訪れる確率を計算すると以下のようになります（サンプルファイルの［ポアソン分布 (応用例)］ワークシートに k = 0 ～ 20 の場合の計算例とグラフが含めてあるのでそちらもご参照ください）。

$$f(18) = 0.007091109\cdots 0.71\%$$
$$f(19) = 0.003732163\cdots 0.37\%$$
$$f(20) = 0.001866081\cdots 0.02\%$$

18 人以上の顧客がある可能性はずいぶん低いようです。その確率は 1 − (17 人までの累積確率) で求められるので、以下のようになります。

$$F(17) = 0.985722386\text{なので、}$$
$$\begin{aligned} F(k \geq 18) &= 1 - F(17) \\ &= 1 - 0.985722386 \\ &= 0.014277614\cdots 1.43\% \end{aligned}$$

1.43%という確率は非常に小さいので、18 人以上の顧客に対応するだけの店員数は必要なさそうです。この場合、17 人までの顧客に対応できる店員数に少し人員を減らしてみる、といった判断が可能です。

実際には、業種にもよりますし（飲食店だともっと顧客数が多いでしょう）、より詳細な場合分け（晴天の場合と雨天の場合の客足の違いやイベント開催日への対応など）も必要になるでしょうが、ポアソン分布は、一定の割合で来客がある場合に**（事象が定常的に発生すると想定されるときに）**、余裕を持った対応のできる人員と人件費との最適な配分を決めるための目安として使えるでしょう。

◇ ◇ ◇ ◇ ◇ ◇ ◇

今回は、まれにしか起こらない事象の確率分布としてポアソン分布を取り上げました。上で見た応用例の他にも、さまざまな分野で活用できそうですね。

次回は、試行を繰り返したときに、**k 回目に成功する確率**（＝成功するまでに k − 1 回失敗する確率）を求めるために使われる幾何分布と、**n 回成功するまでに** k 回失敗する確率を求めるために使われる負の二項分布を取り上げます。次回もお楽しみに！

この記事で取り上げた関数の形式

関数の利用例については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

ポアソン分布の確率質量関数や累積分布関数の値を求めるための関数

POISSON.DIST 関数：ポアソン分布の確率質量関数や累積分布関数の値を求める

形式

POISSIN.DIST(イベント数 , 平均 , 関数形式)

引数

- **イベント数：**目的の事象が起こる回数（k の値）を指定する。
- **平均：**目的の事象が一定期間（一定数）の中で起こる平均の回数を指定する。
- **関数形式：**以下の値を指定する。
  - **FALSE** …… イベント数に対する確率質量関数の値を求める
  - **TRUE** …… イベント数の成功数までの累積分布関数の値を求める

［データ分析］幾何分布と負の二項分布 ～ 三度目の正直の確率は？

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載（確率分布編）の第 5 回。幾何分布とは、k 回目に成功する確率の分布です。一方、負の二項分布は、n 回成功するまでに k 回失敗する確率の分布です。これらの確率分布が利用できる事例を確認した後、確率質量関数や累積分布関数の求め方、可視化の方法などを解説していきます。

羽山博（2024 年 07 月 25 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の確率分布編、第 5 回です。前回は、まれにしか起こらない事象が起こる確率分布であるポアソン分布を取り上げました。今回は、k 回目に成功する確率の分布である幾何分布と、n 回成功するまでに k 回失敗する確率の分布である負の二項分布を取り上げ、その意味や特徴などを見ていきます。

三度目の正直の確率を求めよう ～ 幾何分布

人気アーティストのライブに行きたくても、チケットが抽選でしか手に入らず、なかなかチャンスに恵まれないこともよくありますね。三度目の正直で、ようやくチケットが手に入ったという経験をお持ちの方も多いのではないでしょうか。

そこで、三度目の正直ならぬ k 度目の正直……つまり、k 回目に当選する確率の分布について考えてみましょう。名前だけ先に出すと、この確率分布は**幾何分布**（Geometric distribution）と呼ばれます。図 1 は当選確率が p = 1/4 の場合に、ちょうど k = 3 回目に当選する確率を図解したものです。

図 1　三度目の正直の（3 回目に当選する）イメージ

3 回目に当選するということは、2 回は落選したということ。一般的に言うと、k 回目に当選するということは、k − 1 回は落選したということ。k − 1 回落選する確率と 1 回当選する確率を掛ければ k 回目に当選する確率が求められる。

幾何分布は、前提として各試行が**独立**であることに注意してください。前回の結果が次回の結果に影響を及ぼすことはない、ということですね。

また、毎回の成功確率（ここでは当選確率）は一定であるものとします。成功確率を p とすると、失敗確率は $1-p$ ですね。

幾何分布と以前に説明した超幾何分布は名前は似ていますが、異なるものであることに注意してください。超幾何分布（例えば、くじ引き）では、試行が独立しておらず、前回の結果が次回の結果に影響を及ぼすため、成功確率が試行ごとに変化します。このような抽出の仕方を**非復元抽出**と呼びましたね。一方、幾何分布は**復元抽出**（試行が独立で、成功確率が一定）を前提としています。確率変数の値 k に対する確率質量関数の値についても、超幾何分布は「k 回成功する確率」ですが、幾何分布は「k 回目に成功する確率」と、異なっています。

図 1 の説明からも分かるように、幾何分布では $k-1$ 回がいずれも $1-p$ の確率で失敗し、k 回目に 1 回だけ確率 p で成功するので、その確率質量関数は以下のように表されます。確率変数の値は k です。

$$f(k) = (1-p)^{k-1}p \qquad (1)$$

この式は、図 1 のイメージから比較的素直に理解できると思います。（1）式に説明を加えると以下のようになります。

$$f(k) = \underbrace{(1-p)^{k-1}}_{k-1\text{回失敗する確率}} \times \overbrace{p}^{1\text{回成功する確率}}$$

「成功」というのは必ずしもいい結果が得られるという意味ではなく、目的の事象が起こるということでしたね。製品の検査の例であれば、不良品が見つかることが目的の事象となるので、不良品の発見が「成功」ということになるかもしれません。

では、確認のためにライブチケットの当選確率 p を 1/4 としたとき、$k=3$ の場合（3 回目に当選する確率）と $k=4$ の場合（4 回目に当選する確率）を（1）式に当てはめて求めてみましょう。

- **$k=3$ の場合（3 回目に当選する場合：図 1 の場合）**

$$\begin{aligned} f(3) &= \left(1-\frac{1}{\boxed{ア}}\right)^{\boxed{イ}-1} \times \frac{1}{4} \\ &= \left(\frac{3}{\boxed{ウ}}\right)^{2} \times \frac{1}{4} \\ &= \frac{9}{64} \approx 0.141 \end{aligned}$$

答え：ア＝ 4、イ＝ 3、ウ＝ 4

- **k ＝ 4 の場合（4 回目に当選する場合）**

$$f(4) = \left(1 - \frac{1}{\boxed{ア}}\right)^{\boxed{イ}-1} \times \frac{1}{4}$$
$$= \left(\frac{3}{\boxed{ウ}}\right)^{3} \times \frac{1}{4}$$
$$= \frac{27}{256} \approx 0.105$$

答え：ア＝ 4、イ＝ 4、ウ＝ 4

幾何分布の累積分布関数 F(k) は、**k 回目までに**成功する確率を表します。例えば、3 回目までに成功する確率は、「1 回目に成功する確率、2 回目に成功する確率、3 回目に成功する確率」の累計です。このように、順に値を足していっても求められますが、k 回全てが失敗である確率を **1** から引いた値と等しくなるので、

$$F(k) = 1 - (1 - p)^k \qquad (2)$$

で求めることができます。

ちょっと現実的な例で計算してみましょう。1 等 6 億円のスポーツくじ BIG では、サッカーの試合の勝ち、負け、引き分けを 14 試合について全て当てなくてはいけないので、当選確率 p は以下のようになります（勝ち、負け、引き分けが等確率であるとした場合）。

$$p = \left(\frac{1}{3}\right)^{14} = \frac{1}{4782969}$$

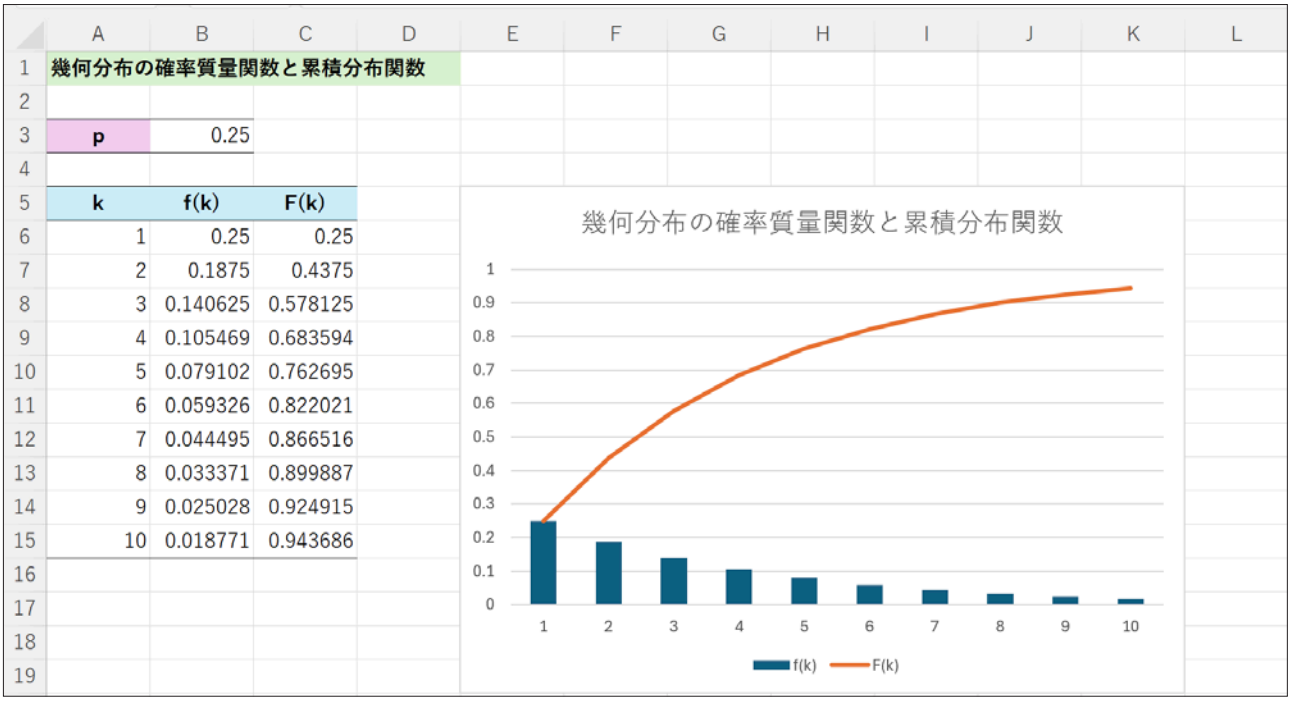
だいたい 480 万分の 1 ですね。では、このくじを 1 枚ずつ 100 回購入したときに、100 回目**までに** 1 等が当選する確率はいくらでしょうか。これは累積分布関数の値なので、（2）式に当てはめると以下のようになります。

$$F(100) = 1 - \left(1 - \frac{1}{4782969}\right)^{100} \approx 0.000021$$

パーセント単位で表すと **0.0021%**です。1 等の期待値（金額）は 0.000021×600,000,000 ＝ 12,600 円です（キャリーオーバーがある場合。なければ、1 等の当選金額は 3 億円が上限）。1 枚 300 円なので、300×100 ＝ 30,000 円を払っても、期待できる見返りは **12,600 円**にしかならないことが分かります（2 等以下も考えるともう少し高くなります）。もちろん、あくまでも平均の話なので、幸運にも 1 等が当選することがあるかもしれません。実際には、筆者のように全てハズレという場合がほとんどですが、外れた分はスポーツ振興に役立てられているということですね。

幾何分布の確率質量関数と累積分布関数を可視化してみよう

BIG の例は確率があまりにも小さく、k の値を大きくしても可視化しづらいので、ライブチケットの例で可視化してみます。Excel には幾何分布の確率質量関数や累積分布関数の値を求める関数がありません。従って、（1）式と（2）式を使って計算します。作成の手順は図 2 の後に記しておきます。



	A	B	C	D	E	F	G	H	I	J	K	L
1	幾何分布の確率質量関数と累積分布関数											
2												
3	p	0.25										
4												
5	k	f(k)	F(k)									
6	1	0.25	0.25									
7	2	0.1875	0.4375									
8	3	0.140625	0.578125									
9	4	0.105469	0.683594									
10	5	0.079102	0.762695									
11	6	0.059326	0.822021									
12	7	0.044495	0.866516									
13	8	0.033371	0.899887									
14	9	0.025028	0.924915									
15	10	0.018771	0.943686									
16												
17												
18												
19												

図 2　幾何分布の確率質量関数と累積分布関数

Excel には幾何分布の確率質量関数と累積分布関数の値を求める関数がないので、（1）式と（2）式に従って計算する。ただし、次項で述べる負の二項分布の確率質量関数と累積分布関数の値を求める関数 NEGBINOM.DIST 関数で代用することはできる（後述）。

グラフ作成の手順は以下の通りです。サンプルファイルをこちらからダウンロードし、［幾何分布］ワークシートを開いて試してみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。具体的な操作方法は、サンプルファイル内に記載しています。

◆ **Excel での操作方法**

- セル **B6** に =(1-B3)^(A6:A15-1)*B3 と入力する
  - ・古いバージョンの Excel でスピル機能が使えない場合は、あらかじめ結果が求められるセル範囲（セル **B6** ～ **B15**）を選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
- セル **C6** に =1-(1-B3)^A6:A15 と入力する
  - ・古いバージョンの Excel でスピル機能が使えない場合は、あらかじめ結果が求められるセル範囲（セル **C6** ～ **C15**）を選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
- セル **B5** ～ **C15** を選択する
- ［挿入］タブを開き、［複合グラフの挿入］ボタンをクリックして［集合縦棒 - 折れ線］を選択する

ここで示した事例は分かりやすさを優先した単純なものですが、幾何分布は、機械が何回の使用で故障するかの確率を求めたり、ある Web サイトにアクセスするユーザーが何回目の訪問で購買活動を行うかの確率を求めたりするのに使えます。

では、少し話題を変えて、次に負の二項分布を見てみましょう。

二匹目の泥鰌（どじょう）が得られるまでの確率を求めよう ～ 負の二項分布

本来、二匹目の泥鰌というのは、うまくいった人や物事のマネをして、うまく事を運ぼう（としてもうまくいかないものだ）という意味ですが、ここでは文字通り、複数回成功するまでの確率の話をします。もう少し正確に言うと、n 回成功するまでに k 回失敗する確率を求めます。毎回の試行は独立であるものとし、各試行での成功確率も一定であるものとします。こちらについても名前だけ先に出しておくと、**負の二項分布**（Negative binomial distribution）と呼ばれる確率分布を利用します。

文献によっては異なる文字を使って「k 回成功するまでに x 回失敗する確率」と表している場合（確率変数の値は x）や、成功と失敗を逆にして「r 回失敗するまでに k 回成功する確率」と表している場合（確率変数の値は k）などがあります。Excel のヘルプでは「r 回成功するまでに x 回失敗する確率」（確率変数の値は x）となっています。慣れないうちは混乱してしまいますが、以下で解説する負の二項分布の意味を理解していれば、どのように表現されていても適切に利用できるはずです。

図 3 のような例で考えてみましょう。またまたライブチケットの例です。チケットの抽選に n 回当選するまでに k 回落選する確率を求めます。この例では、n = 2、k = 1 ですね。

図 3　n=2 回当選するまでに k=1 回落選する例

n = 2 回当選するまでに k = 1 回落選しているということは、「落選―当選―当選」と「当選―落選―当選」の場合。試行の最後は必ず当選であることに注目しよう。試行回数は n + k となる。この例であれば 2 + 1 = 3 回。n 回成功するまでに k + n 回の試行を行う確率とも考えられる。

2 回当選するまでに 1 回落選するのは図 3 に示した 2 つの場合があります。それらの確率を足したものが答えとなるわけですね。ここでは、2 回当選する確率を求めているのではなく、2 回当選する**までに** 1 回失敗する確率を求めていることに注意が必要です。従って、「当選―当選―落選」は図 3 には含まれません（それを含めると、単に 3 回中 2 回当選する確率となり、二項分布で表されることになります）。

負の二項分布の確率質量関数 f(k) は以下の通りです。ここでは、**確率変数の値は失敗回数を表す k である**ということに注意してください。

$$f(k) = \binom{k+n-1}{n-1} p^n (1-p)^k \qquad (3)$$

ちょっと複雑な式なので、この式の意味については後述することとして、まずは具体的な数値を当てはめて計算結果を確かめておきましょう。図 3 とは少し値を変えて試してみます。当選確率 p = 1/4 として、2 回当選するまでに 3 回落選する確率を求めましょう。この場合、n = 2, k = 3 なので、以下のようになります。

$$\begin{aligned} f(x) &= \binom{3+\boxed{ア}-1}{2-1}\left(\frac{1}{\boxed{イ}}\right)^2\left(1-\frac{1}{4}\right)^{\boxed{ウ}} \\ &= \binom{4}{1}\left(\frac{1}{\boxed{エ}}\right)^2\left(\frac{3}{4}\right)^{\boxed{オ}} \qquad (4) \\ &= 4\cdot\frac{3^3}{4^5} \\ &= \frac{27}{256} \approx 0.105 \end{aligned}$$

→目次に戻る

答え：ア＝ 2、イ＝ 4、ウ＝ 3、エ＝ 4、オ＝ 3

（4）式の最初の、

$$\binom{4}{1}$$

は、**4** つの中から **1** つを選ぶ組み合わせ数（${}_4C_1 = 4$）のことで、それ以降のかっこは単に分数を囲んだだけのものです。

では、確率質量関数の意味を見ていきます。図 4 をご覧ください。**最後が必ず成功（当選）になっており、それ以前は二項分布であ**ることに気付けば、上の式が簡単に導き出せます。最後に成功する B の確率はもちろん p です。それ以前の部分 A の試行数は全体の試行数 k＋n から 1 を引いた k＋n－1 ですね。その中で n－1 回成功する確率なので、二項分布の式に試行数 k＋n－1、成功数 n－1、失敗数 k を当てはめれば求められます。

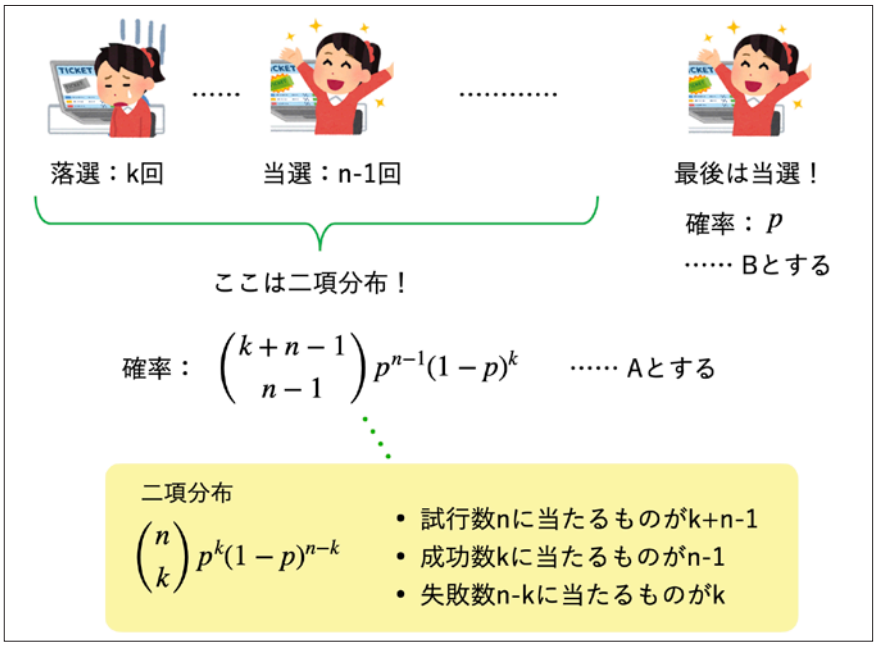


図 4　負の二項分布は、成功する直前までが二項分布になっている

n 回成功（当選）、k 回失敗（落選）なので、全体の試行数は n＋k 回となる。最後の 1 回は必ず成功（確率は p）となり、その前までの n＋k－1 回は、n－1 回が成功で、k 回が失敗となる二項分布。図の A と B を掛ければ負の二項分布の確率質量関数の値が求められる。

f(k) は、図 4 の A と B との積で求められるので、以下のようになります。上に示した（3）式と一致していることを確認しておいてください。

$$f(k) = \overbrace{\binom{k+n-1}{n-1}p^{n-1}(1-p)^k}^{\mathrm{A}} \times \overbrace{p}^{\mathrm{B}}$$
$$= \binom{k+n-1}{n-1}p^n(1-p)^k$$

→目次に戻る

なお、n = 1 の場合は k 回の失敗の後、k + 1 回目に 1 回だけ成功することになります。ということは、n = 1 の場合の負の二項分布の f(k) は、幾何分布の f(k + 1) と等しくなることが分かります。

◆ 負の二項分布で n=1 の場合

$$\begin{aligned} f(k) &= \binom{k+n-1}{n-1} p^n (1-p)^k \\ &= \binom{k+1-1}{1-1} p^1 (1-p)^k \\ &= \binom{k}{0} p(1-p)^k \\ &= p(1-p)^k \end{aligned}$$

◆ 幾何分布

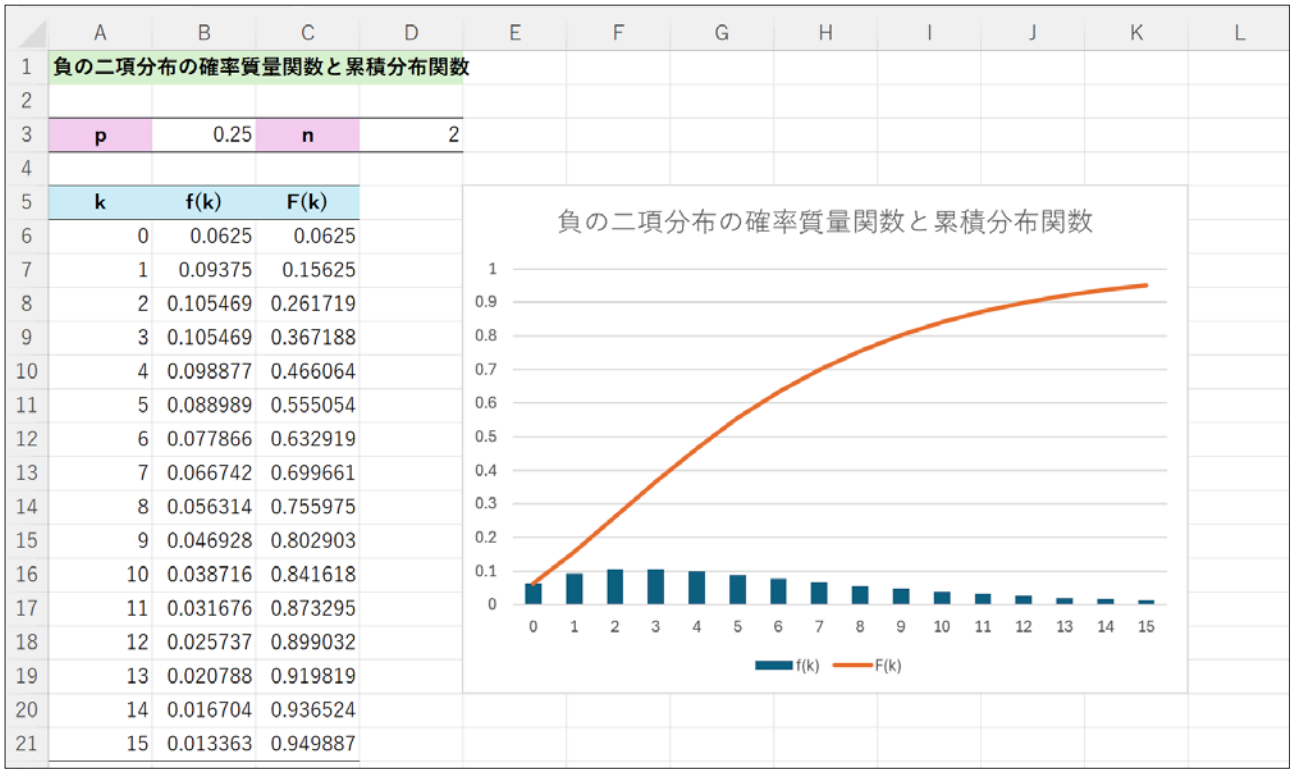
$$\begin{aligned} f(k+1) &= (1-p)^{k+1-1} p \\ &= p(1-p)^k \end{aligned}$$

このことについては、次項で NEGBINOM.DIST 関数について解説した後、またお話しします。

負の二項分布の累積分布関数は、n 回成功するまでに k 回**以下**失敗する確率です。これについては次の項で具体的に見ることにしましょう。

負の二項分布の確率質量関数と累積分布関数を可視化してみよう

こちらもライブのチケットの例で可視化してみます。当選確率 p = 1/4 のとき、n = 2 回当選するまでに、k 回落選する確率を求めます。負の二項分布については、Excel の NEGBINOM.DIST 関数を使って確率質量関数や累積分布関数の値を求めることができます。従って（3）式を使わなくても簡単に値が求められます。



	A	B	C	D
1	負の二項分布の確率質量関数と累積分布関数			
2				
3	p	0.25	n	2
4				
5	k	f(k)	F(k)	
6	0	0.0625	0.0625	
7	1	0.09375	0.15625	
8	2	0.105469	0.261719	
9	3	0.105469	0.367188	
10	4	0.098877	0.466064	
11	5	0.088989	0.555054	
12	6	0.077866	0.632919	
13	7	0.066742	0.699661	
14	8	0.056314	0.755975	
15	9	0.046928	0.802903	
16	10	0.038716	0.841618	
17	11	0.031676	0.873295	
18	12	0.025737	0.899032	
19	13	0.020788	0.919819	
20	14	0.016704	0.936524	
21	15	0.013363	0.949887	

図 5　負の二項分布の確率質量関数と累積分布関数

確率質量関数と累積分布関数の値は（3）式に従って計算してもよいが、いずれも NEGBINOM.DIST 関数で求められる。確率質量関数と累積分布関数の意味についても確認しておこう。例えば k = 5 のときの確率質量関数の値は 2 回当選するまでにちょうど 5 回落選する確率となっており、累積分布関数の値は 2 回当選するまでに、5 回まで（0 ～ 5 回）落選する確率となっている。

グラフ作成の手順は以下の通りです。サンプルファイルをこちらからダウンロードし、［負の二項分布］ワークシートを開いて試してみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。具体的な操作方法は、サンプルファイル内に記載しています。なお、Google スプレッドシートの NEGBINOM.DIST 関数には最後の引数がない（確率質量関数の値しか求められない）ことに注意が必要です。具体的な操作方法は、サンプルファイル内に記載しています。

◆ **Excel での操作方法**

- セル **B6** に =NEGBINOM.DIST(A6:A21,D3,B3,FALSE) と入力する
  - ・古いバージョンの Excel でスピル機能が使えない場合は、あらかじめ結果が求められるセル範囲（セル **B6** ～ **B21**）を選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
- セル **C6** に =NEGBINOM.DIST(A6:A21,D3,B3,TRUE) と入力する
  - ・古いバージョンの Excel でスピル機能が使えない場合は、あらかじめ結果が求められるセル範囲（セル **C6** ～ **C21**）を選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
- セル **B5** ～ **C21** を選択する
- ［挿入］タブを開き、［複合グラフの挿入］ボタンをクリックして［集合縦棒 - 折れ線］を選択する
- ［グラフのデザイン］タブを開き、［データの選択］ボタンをクリックする
- ［データソースの選択］ダイアログボックスで［横（項目）軸ラベル］の下の［編集］ボタンをクリックする
- ［軸ラベル］ダイアログボックスで［軸ラベルの範囲］ボックスをクリックし、セル **A6** ～ **A21** を選択する
- ［OK］をクリックして［軸ラベル］ダイアログボックスを閉じる
- ［OK］をクリックして［データソースの選択］ダイアログボックスを閉じる

NEGBINOM.DIST 関数の引数は図 6 のように指定します。

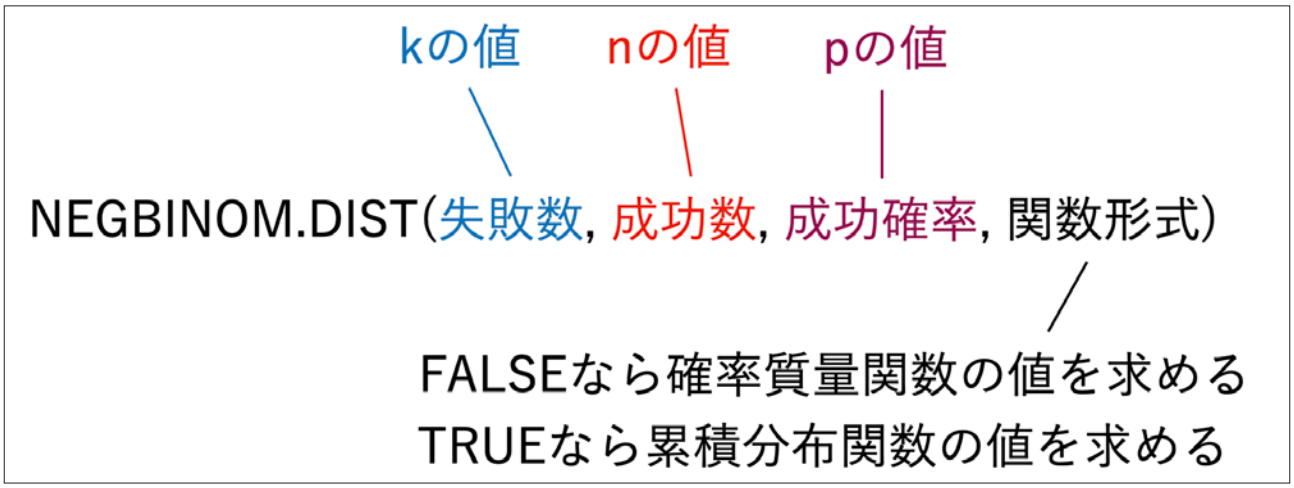


図 6　NEGBINOM.DIST 関数に指定する引数

図 5 の例では、失敗数として、セル A6:A21 というセル範囲を指定しているので、スピル機能によりセル **A6** ～ **A21** の失敗数に対する確率質量関数の値や累積分布関数の値が一度に求められる。つまり、n 回成功するまでに k 回失敗する確率や累積確率が求められる。

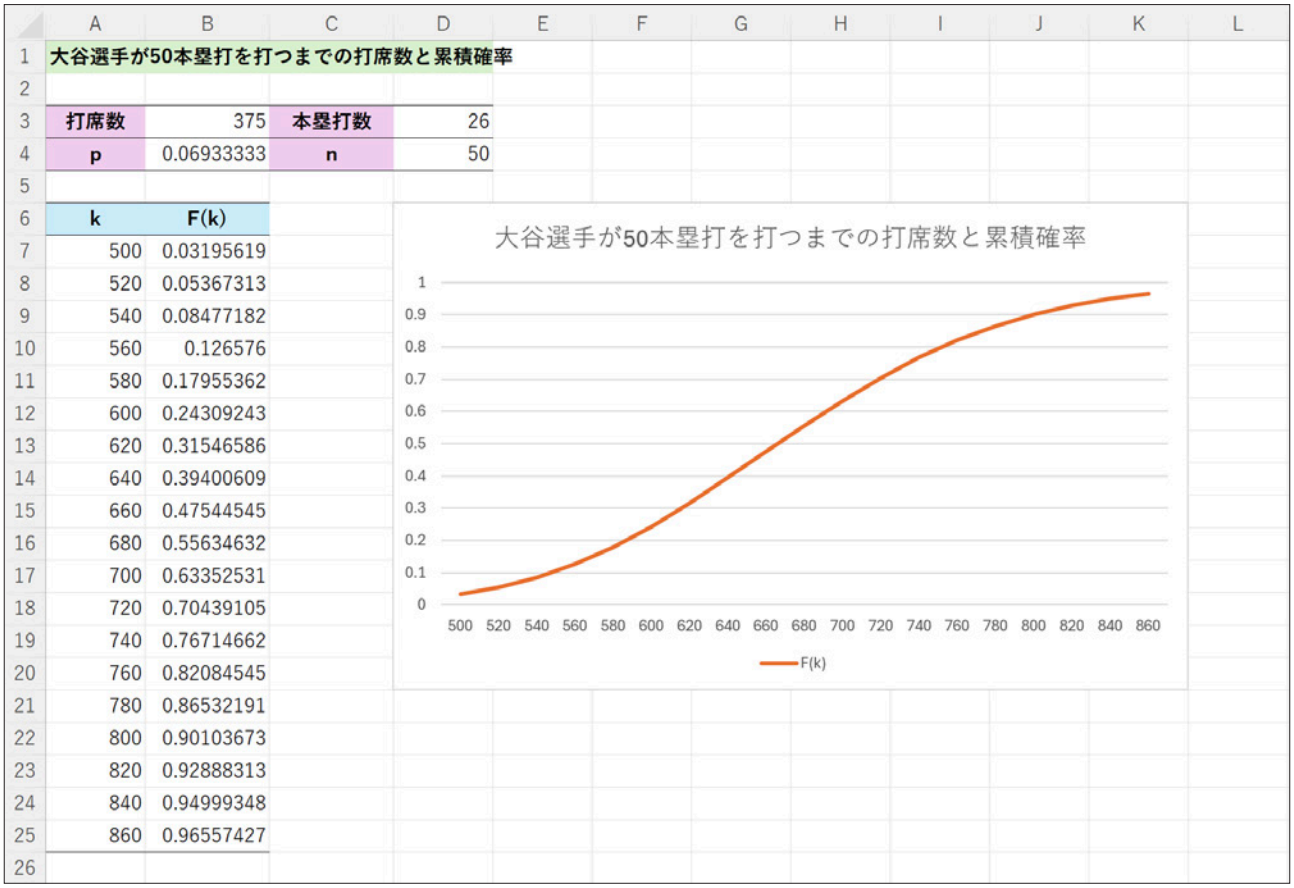
Excel には幾何分布の確率質量関数や累積分布関数の値を求める関数はありませんが、負の二項分布の確率質量関数や累積分布関数の値を求める NEGBINOM.DIST 関数で代用できます。すでに述べた通り、成功数 n ＝ 1 のとき、負の二項分布の f(k) は、幾何分布の f(k ＋ 1) と等しくなります。従って、幾何分布の f(k) を求めたいときには、NEGBINOM.DIST 関数の成功数に **1** を指定し、失敗数に **k-1** の値を指定すれば答えが得られます。サンプルファイルにはそれらの計算を比較した例も含めてあるので、ぜひ参照してみてください。

負の二項分布のちょっとした応用例 ～ 大谷選手は何打席までに 50 本のホームランを打つのか

この執筆を執筆している時点では、ドジャースの大谷翔平選手は今シーズン 375 打席中、26 本のホームランを放っています。そこで、この調子が維持されるものとして、負の二項分布を利用し、50 本目のホームランを打つまでの打席数とその確率を求めてみましょう。単純に考えると、ほぼ倍のホームラン数なので、375×2=750 打席ぐらいで達成できそうな気がしますが、さてどうでしょうか。1 打席でホームランを打つ確率は、これまでの成績を基に 26/375 とします。

750 打席目に 50 本目のホームランを打つということは、成功数 n = 50、失敗数 k = 700 と考えられます。従って、750 打席目**まで**に 50 本目のホームランを打つ確率は、負の二項分布の累積分布関数 =NEGBINOM.DIST(700, 50, 26/375, TRUE) で求められます。実際に計算してみると **0.634** となります。倍の打席数だからといってほぼ確実というわけではなさそうですね。

図 7 は k = 500 ～ 860 までとし、NEGBINOM.DIST 関数で求めた累積分布関数の値とそれをグラフにしたものです。この例もサンプルファイルに含めてあるので、ぜひご参照ください。作成方法は図 5 とほとんど同じです。



	A	B	C	D
1	大谷選手が50本塁打を打つまでの打席数と累積確率			
2				
3	打席数	375	本塁打数	26
4	p	0.06933333	n	50
5				
6	k	F(k)		
7	500	0.03195619		
8	520	0.05367313		
9	540	0.08477182		
10	560	0.126576		
11	580	0.17955362		
12	600	0.24309243		
13	620	0.31546586		
14	640	0.39400609		
15	660	0.47544545		
16	680	0.55634632		
17	700	0.63352531		
18	720	0.70439105		
19	740	0.76714662		
20	760	0.82084545		
21	780	0.86532191		
22	800	0.90103673		
23	820	0.92888313		
24	840	0.94999348		
25	860	0.96557427		
26				

図 7　大谷選手が 50 本目のホームランを打つまでの打席数とその確率

n は成功数（ホームランの数）、k は失敗数（ホームラン以外の打席数）。ホームランの確率はセル **B4** に =D3/B3 と入力されているので、セル **B7** に =NEGBINOM.DIST(A7:A25,D4, B4, TRUE) と入力すれば、それぞれの k に対する累積分布関数の値が求められる。なお、スピル機能が使えないバージョンの Excel であれば、結果が正しく表示されないが、その場合はセル **B7** ～ **B25** を選択し、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押せば配列数式として入力され、結果が表示される。

→目次に戻る

◇ ◇ ◇ ◇ ◇ ◇ ◇

今回は、親しみやすさ重視で宝くじやスポーツの事例を取り上げて解説しましたが、負の二項分布は、n 個の不良品が出るまでに k 個の良品が製造される確率を求めたり、n 回の購買活動が行われるまでに Web サイトに k 回アクセスされる確率を求めたりするのにも使われます。他にもさまざまな分野で活用できそうですね。

今回で、離散型確率分布のお話はひと区切りです。連続型確率分布のお話に移りたいと思いますが、さらなる応用のために、番外編として、累積分布関数の逆関数のお話を挟んでおきたいと思います。例えば、図 7 の例であれば、累積確率が **95%**以上となる k の値（それまでのホームラン以外の打席数）を求めることができます。ちなみに答えは **841 打席**です（もっとも、メジャーリーグの全試合は約 160 試合なので、全ての試合で 4 回の打席に立ったとしても 640 打席しかありません。もちろんムリだということではなく、95%の確率で**ほぼ確実に**というのが難しいということです。2024 年 9 月 20 日追記：現地時間 2024 年 9 月 19 日に 50 本塁打を達成しました！）。というわけで、次回の番外編もお楽しみに！

この記事で取り上げた関数の形式

関数の利用例については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

負の二項分布の確率質量関数や累積分布関数の値を求めるための関数

NEGBINOM.DIST 関数：負の二項分布の確率質量関数や累積分布関数の値を求める

形式

NEGBINOM.DIST(失敗数 , 成功数 , 成功確率 , 関数形式)

引数

- **失敗数**：目的の事象が起こらない回数（k の値）を指定する。
- **成功数**：目的の事象が起こる回数（n の値）を指定する。
- **成功確率**：1 回の試行で目的の事象が起こる確率を指定する。
- **関数形式**：以下の値を指定する。
  - **FALSE** …… 失敗数に対する確率質量関数の値を求める
  - **TRUE** …… 失敗数までの累積分布関数の値を求める

［データ分析］累積分布関数の逆関数 ～ 95%の確率で推しのチケットを入手するまでに何回チャレンジすればいい？

データ分析の初歩からステップアップしながら学んでいく連載（確率分布編）の番外編。代表的な離散型確率分布に対する累積分布関数の逆関数を紹介。例えば、二項分布の累積分布関数では n 回中 k 回まで成功する確率が求められますが、その逆関数では何%か（以上）の確率で成功するまでの回数を求められます。

羽山博（2024 年 08 月 08 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の確率分布編、番外編です。前回までは、代表的な離散型確率分布の確率質量関数や累積分布関数の値を求める方法を見てきました。

今回は、累積分布関数の逆関数を利用して、累積確率を基に、逆に確率変数の値 k を求める方法を紹介します。「95%以上の確率で当選するのは何回までか」といった計算や、「95%以上の確率で当選するまでに何回チャレンジすればいいか」といった計算ができます。

ある確率以上となる確率変数の値 k を求める ～ 累積分布関数の逆関数

もったいぶるわけではありませんが「95%の確率で推しのチケットを入手するまでに何回チャレンジすればいい?」という表題の件については後回しにして、逆関数とは何かということからお話を始めます（逆関数の意味は知っているので、答えを知りたい、という方はこちらに進んでいただいても構いません）。

逆関数ってそもそも何だっけ?

簡単な例から始めましょう。関数とはどのようなものものだったでしょうか。中学の数学では、

$$y = 2x$$

といった関数について学びました。x の値を指定すると y の値が 1 つに決まりますね。例えば、x = 3 なら y = 6 です。**逆関数**は、これとは逆に y の値から x の値を求める関数のことです。この例であれば、x について解いて、

$$x = \frac{1}{2}y$$

とすればいいですね。y = 6 であれば x = 3 となります。

高校の数学では、関数を y = f(x) のように一般的に表すことがあります。このとき、逆関数は、関数名の右肩に－1 を付けて、$x = f^{-1}(y)$ と表されます。具体例は後で見ますが、累積分布関数 F(k) の値、つまり累積確率を P と表すと、k から P = F(k) の値を求めるのとは逆に、P から k の値を求めるので、F(k) の逆関数は $F^{-1}(P)$ と表されます。

二項分布の累積分布関数についてのおさらい

累積分布関数の逆関数も上で見たものと考え方は全く同じです。そこで、二項分布を例に、累積分布関数の逆関数がどのようなものかを見てみましょう。まずは、二項分布の累積分布関数のおさらいから始めます。二項分布の累積分布関数を利用すれば、例えば、$p = 1/4$ の確率で当選するチケットに $n = 10$ 回申し込んだとき、$k = 3$ 回**まで**当選する確率などが求められます。

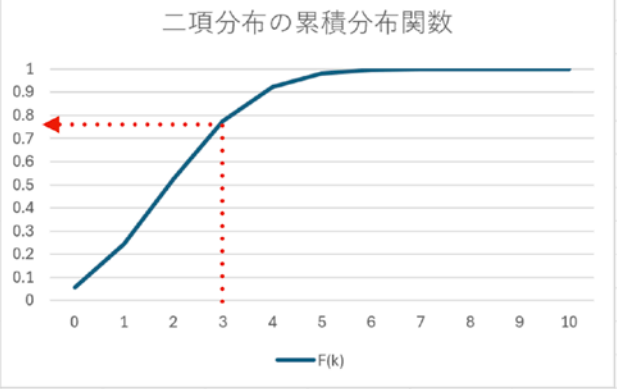
$k = 0 \sim 10$ について、Excel の BINOM.DIST 関数を使って求めた累積分布関数の値とそれを折れ線グラフにした例を図 1 に示しました。セル **B10** を見れば、$k = 3$ 回まで当選する確率 F(k) の値は 0.776(= 77.6%) であることが分かります。**0 回**当選する確率～ **3 回**当選する確率の累計が **77.6%**であるということですね。

二項分布は離散型確率分布なので、確率変数の値 k はとびとびの値を取ります（**0** 以上の整数となります）。「成功回数 k が **3.5 回**」などのような小数になることはないので、可視化する際には折れ線グラフよりも棒グラフの方が適しています。しかし、ここでは前回までに見た複合グラフの形式と合わせるために折れ線グラフにしてあります。

BINOM.DIST 関数の使い方やグラフの作成方法を忘れた方は、こちらでおさらいしてください。また、図 1 のサンプルファイルの［二項分布の累積分布関数］ワークシートにも図 1 の内容と作成手順が含まれているので、そちらもご参照ください Google スプレッドシート用のサンプルファイルはこちらです。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。これらのサンプルファイルには逆関数の例も含まれています（後でまた使います）。

◆ p=1/4の当選確率で、n=10回中、k回**まで**当選する確率：二項分布の累積分布関数

	A	B
1	**二項分布の累積分布関数**	
2		
3	**n**	10
4	**p**	0.25
5		
6	**k**	**F(k)**
7	0	0.056314
8	1	0.244025
9	2	0.525593
10	3	0.775875
11	4	0.921873
12	5	0.980272
13	6	0.996494
14	7	0.999584
15	8	0.99997
16	9	0.999999
17	10	1
18		



■ 3回まで当選する確率は0.776（77.6%）→ 4回以上当選する確率は1-0.776=0.224（22.4%）

図 1　二項分布の累積分布関数で、確率変数の値 k から累積確率を求める

累積分布関数では、k に対する累積確率 F(k) の値が求められる。例えば、k = 3 なら F(3) = 0.776 となり、**3 回**まで当選する確率は **77.6%** ということが分かる。このことから、**4 回**以上当選する確率は 1 − 0.776 = 0.224（= **22.4%**）であることも分かる。

グラフでも確認しておきましょう。グラフの横軸は当選回数 k で、縦軸は k に対する F(k) の値、つまり、二項分布の累積確率となっていますね。横軸の k = 3 に対する縦軸の値は F(k) = 0.776 であることが分かります（横軸の **3** のところから引かれている赤い点線で読み取れます）。いや、そこまで詳しくは分からないだろう、と思われるかもしれませんが、マウスポインターを折れ線の上に位置付けると、k に対する F(k) の値がポップアップ表示されるのでそのことが分かります。この確率を **1** から引くと、**4 回**以上当選する確率（**0.224**）になることも分かりますね。

0 回当選する確率を含めても意味がないので、それは除外して、**1 ～ 3 回**当選する確率を求めたいということもありますね。そのような場合は、**3 回**当選する累積確率から **0 回**当選する累積確率を引けばいいのですが、Excel の BINOM.DIST.RANGE 関数を使えば、その値が簡単に求められます。図 1 の例であれば =B10-B7 でも求められますが、=BINOM.DIST.RANGE(B3,B4,1,3) でも求められます。

→目次に戻る

二項分布の累積分布関数に対する逆関数とは

累積分布関数とは逆の計算をしたいときもあります。つまり、**逆関数**により、累積確率 P = F(k) から確率変数の値 k を求めたいということですね。例えば、n = 10 回申し込んだときに、P = 95%**以上**の確率で当選するのは何回までかを求める、ということです。まずは図 2 のグラフで確認してみましょう。ワークシートは図 1 と同じものです。縦軸の **0.95** から逆にたどればいいですね。ブルーの点線をたどれば、**5 回**までだということが読み取れそうです。B 列の値を見ても、**0.922**（= **92.2%**）に対する k の値が **4**、**0.980**（= **98.0%**）に対する k の値が **5** であることから、間違いありません。

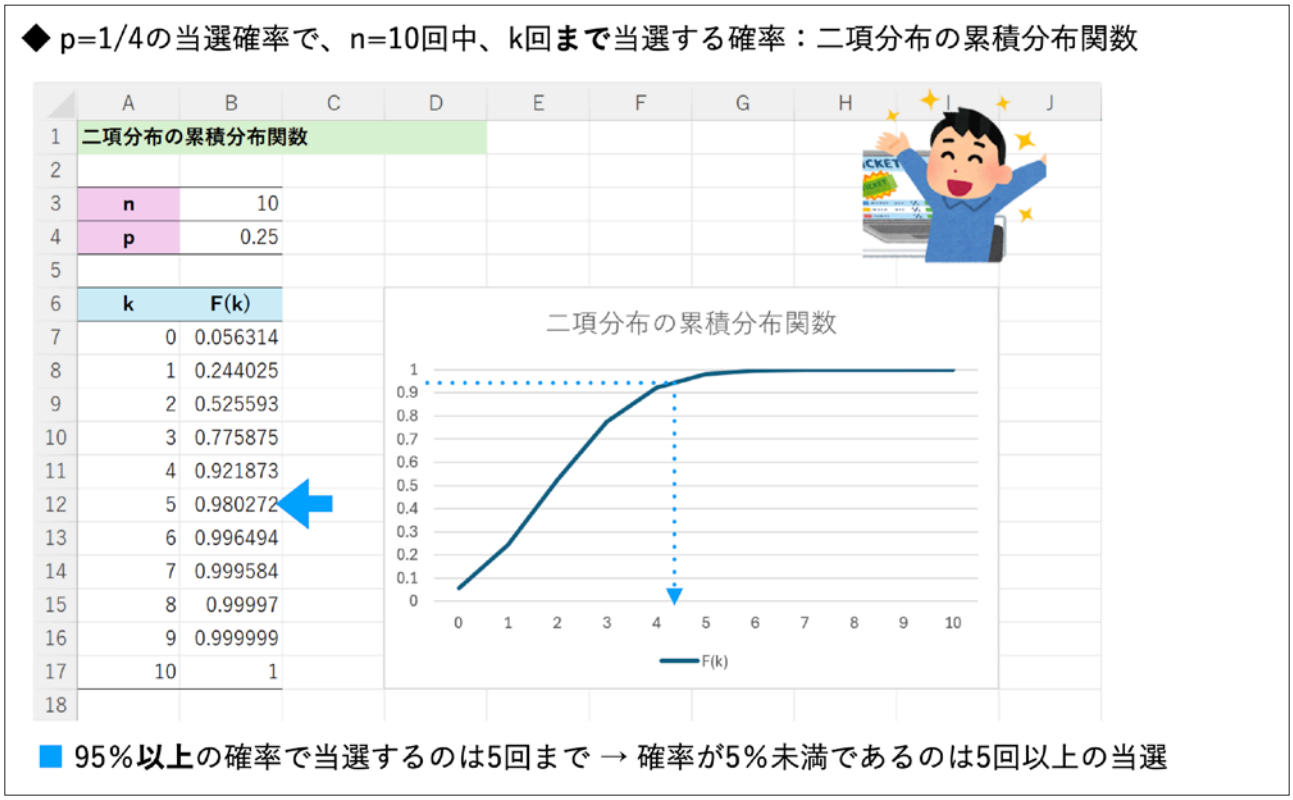


図 2　二項分布の累積分布関数で、累積確率から確率変数の値 k を逆に読み取る

累積分布関数とは逆に、**95%**以上の確率で当選するのは何回までかを知りたいこともある。つまり、F(k) = 0.95 以上となる k の値を求めるということ。グラフや表から k = 5 であることが分かるが、表を作って目視で確認するのではなく、計算で求めたい（後述）。

しかし、図 2 のような表やグラフを作って目視で確認するのではなく、ちゃんと計算で結果を求めたいですね。そこで、朗報です！　Excel では二項分布の累積分布関数に対する逆関数の値を求める BINOM.INV 関数が利用できます。サンプルファイルの［二項分布の逆関数］ワークシートを開いて、セル **E4** に =BINOM.INV(B3,B4,D4) と入力してみてください（図 3）。

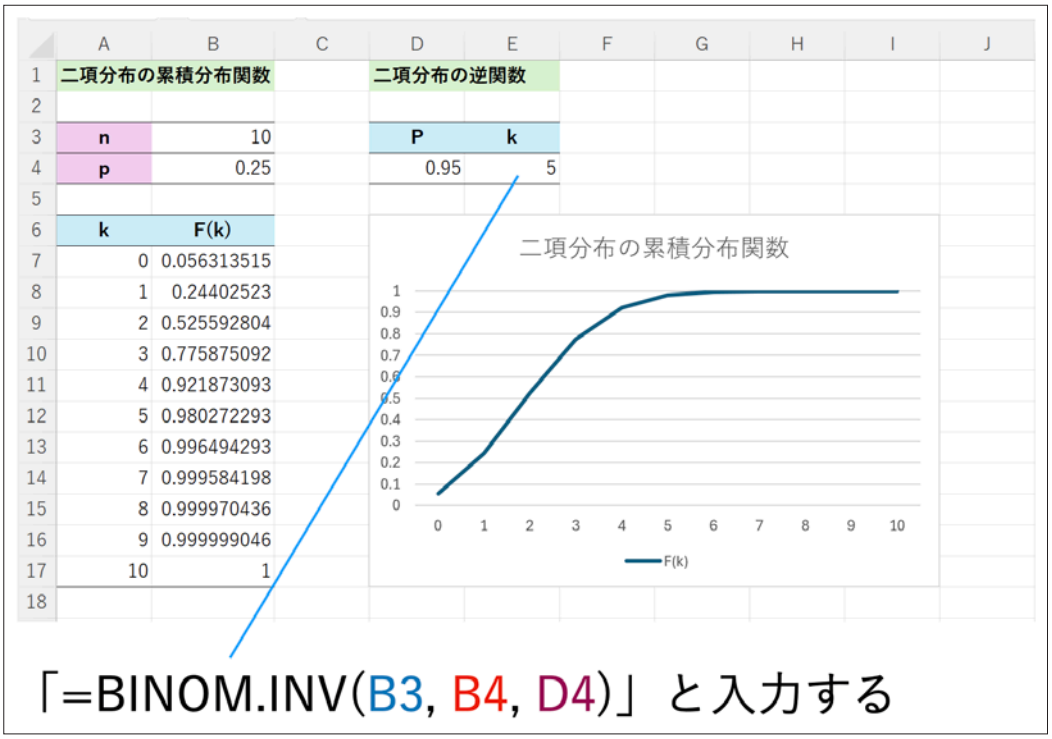


図 3　BINOM.INV 関数により、二項分布の累積分布関数に対する逆関数の値を求める

BINOM.INV 関数には、引数に試行回数 n の値（セル **B3**）、1 回の試行の成功確率 p の値（セル **B4**）、累積確率 P（セル **D4**）を指定する。結果は **5** となる。n = 10 回申し込んだときに、P = 95%以上の確率で当選するのは **5 回**まで。言い換えると、当選する確率が **5%**未満となるのは **5 回**以上当選する場合。

図 3 で入力した BINOM.INV 関数の結果を見れば、n = 10 回申し込んだときに、P = 95%以上の確率で当選するのは **5 回**までということが分かります。従って、確率が **5%**未満であるのは **5 回**以上当選する場合であることも分かります。もし、実際に **5 回**以上当選したとすると、めったに起こらないことが起こった、ということになりますね。このような計算は、統計的検定を行う場合に使うのですが、検定のお話については、この連載の続編となる推測統計編でのお楽しみということにします。

念のため、BINOM.INV 関数に指定する引数を確認しておきましょう（図 4）。

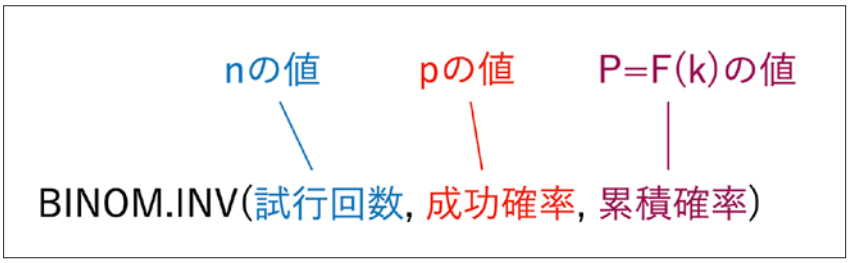


図 4　BINOM.INV 関数に指定する引数

図 3 の例では、試行回数 n の値としてセル **B3**（**10** という値）を指定、**1 回**の試行の成功確率である p の値としてセル **B4**（**0.25** という値）を指定、累積確率 P の値としてセル **D4**（**0.95** という値）を指定している。

→目次に戻る

95%の確率で推しのチケットを入手するまでに何回チャレンジすればいいのか？

ここから、ようやく推しのチケットのお話です。人気アーティストのチケットが抽選販売となっていて、申し込んでもなかなか当選しないような状況を考えてみてください。k 回目に **1 回**成功（当選）する確率は幾何分布で求められました。前回のお話で登場した事例ですね。

「95%の確率で推しのチケットを入手するまでに何回チャレンジすればいいのか」というのは、幾何分布の累積確率が 95%以上となるのは、何回目に成功する**まで**なのか、ということです。従って、幾何分布の累積分布関数に対する逆関数で求められます。

- **幾何分布の累積分布関数：** k 回目までの試行で **1 回**成功する累積確率 P を求める（k から P を求める）
- **幾何分布の累積分布関数に対する逆関数：** 累積確率 P 以上で成功するまでの試行回数 k は何回までかを求める（P から k を求める）

では、Excel の関数を使って……と言いたいところですが、Excel には幾何分布の確率質量関数の値や累積分布関数の値を求める関数もなければ、その逆関数の値を求める関数もありません。確率質量関数と累積分布関数については、負の二項分布の確率を求めるための NEGBINOM.DIST 関数で代用できますが、残念ながら、NEGBINOM.DIST 関数の逆関数はありません。

図 2 と同じような累積分布関数の表を作成して、そこから読み取るというのも手ですが、Python などのプログラミング言語を使うと簡単です。毎回の当選確率を **1/4**（4 分の 1）とすれば、以下のコード（リスト 1）で答えが得られます。SciPy は科学技術計算を行うためのライブラリで、scipy.stats は特に統計的な計算を行うためのモジュールとなっています。

Excel のゴールシーク（＝目標の計算結果を得るために必要な値を逆算する機能）を使うという方法もあります。しかし、NEGBINOM.DIST 関数を使うと残念ながら計算が収束せず、答えが得られません（k の値が整数であるため）。

実は、ベータ分布の確率密度関数や累積分布関数を求めるための BETA.DIST 関数を使って、パラメーター α に **1** を、β に k の値を指定すれば、幾何分布の累積確率が求められます。それに対してゴールシークを適用すれば答えが求められます。

ベータ分布は連続型確率分布なので、今回はこれ以上は触れません（次回以降をお楽しみに！）。入力例やゴールシークの方法については、サンプルファイルの［幾何分布］ワークシートに含めてあるので、興味のある方はご参照ください。

サンプルプログラムはこちらから参照できます。リンクをクリックすれば、ブラウザが起動し、Google Colaboratory の画面が表示されます。

リスト 1 に示した Python コードは 2 つのコードセル（＝プログラムコードを入力して実行できる領域）に入力されています（Google アカウントでのログインが必要です）。コードセルをクリックして［Shift］+［Enter］キーを押せばコードが実行できます。ここで求めたい答えを得るためには 2 つ目のコードセルを実行するだけで構いません。が、以降のコードセルでグラフ作成などを行うので、1 つ目のコードセルを実行した上で、2 つ目のコードセルを実行してください。

```
# 1 つ目のコードセルの内容：
# 以下のモジュールは、これ以降、グラフ作成などで使われるので、最初に必ず 1 回実行しておくこと
import matplotlib.pyplot as plt
import numpy as np
```

```
# 2 つ目のコードセルの内容：
# 毎回の当選確率を 1/4 とする
from scipy.stats import geom

geom.ppf(0.95, 1/4)  # 累積確率、毎回の成功確率を指定
# 出力例：
# 11.0
```

リスト 1　95%の確率で推しのチケットに当選するまでの落選回数を求める

scipy.stats モジュールで利用できる統計関数では、それぞれの分布を表すクラスの、pmf という関数が確率質量関数（連続型確率分布の場合は pdf という関数が確率密度関数）、cdf という関数が累積分布関数、ppf という関数が累積分布関数の逆関数となっている。幾何分布については、geom.ppf という関数が累積分布関数に対する逆関数。この関数に累積確率と毎回の試行での成功確率を指定すれば、その累積確率以上になるまでの失敗数が求められる。

scipy.stats モジュールの geom が、幾何分布を表すクラスです。その geom.ppf 関数が、累積分布関数に対する逆関数となります。この関数に累積確率と毎回の試行での成功確率を指定すれば、その累積確率以上になるまでの失敗数が求められます。

リスト 1 の geom.ppf(0.95, 1/4) 関数を実行すると **11.0** という結果が出力されます。つまり、**95%**の累積確率以上になるまでの失敗数が **11 回**という結果です。

いきなり当選することもあれば、何回申し込んでも落選ということもあるかもしれませんが、この結果から、理論的には **95%**の確率で、**11 回**までの落選の後に（つまり 12 回目までに）当選するものと考えられます。それ以上チャレンジしても当選しないということは、各試行の当選確率である **1/4** という前提が誤っていると考えられます（たまたま、という可能性もあります）。

大谷選手が 95%の確率で 50 本のホームランを打つまでの打席数は？

前回、大谷選手のホームラン数が 50 本になるまでの打席数とその確率を求めました。その際に利用したのが負の二項分布です。負の二項分布は n 回成功するまでに k 回失敗する確率の分布でしたね。これについても、逆関数を利用すれば、**95%**の確率で **50 本**のホームランを打つまでに、ホームランでない打席数は**何回**までになるかが求められます。

残念ながら、負の二項分布についても、累積分布関数の逆関数は Excel にはありません。リスト 1 と同様に、scipy.stats モジュールの nbinom.ppf 関数を使いましょう。上と同じサンプルプログラムの 3 つ目のコードセルにリスト 2 のコードが入力されています。コードセルをクリックし、［Shift］+［Enter］キーを押せば実行できます。

```
from scipy.stats import nbinom

p = 26/375  # 打席数に対するホームランの確率
nbinom.ppf(0.95, 50, p)  # 累積確率、成功数、毎回の成功確率を指定
# 出力例：
# 841.0
```

リスト 2　95%の確率で 50 本のホームランを打つまでの打席数を求める

負の二項分布では、nbinom.ppf 関数を利用すれば累積分布関数の逆関数の値が求められる。引数には、累積確率と成功数、毎回の試行での成功確率を指定する。累積確率以上で指定した回数だけ成功するまでの失敗数が求められる。

前回の原稿執筆時には、375 打席中 26 本のホームランを打っていたので、毎回の成功確率を **26/375** としました。リスト 2 の geom.ppf 関数を実行すると **841.0** という値が出力されます。その結果から、**95%**以上の確率で **50 本**のホームランを打つまでの、ホームランでない打席数は **841** と考えられます。従って、**95%**以上の確率で **50 本**のホームランを打つまでの打席数は、理論上、841 + 50 = 891 打席となります。

もっとも、前回の記事の最後にも記したように、メジャーリーグの全試合は約 160 試合なので、全ての試合で 4 回の打席に立ったとしても、640 打席しかありません。ということは、95%という高い確率でシーズン中に 50 本のホームランを打つということは難しいようです。もちろん、ムリだということではなく、この時点では**ほぼ確実だ**と言うのが難しいということです。（前回の記事にも記しましたが、現地時間 2024 年 9 月 19 日に 50 本塁打を達成しました！）。

ちなみに、この原稿の執筆時点では、ホームラン数は 429 打席中 29 本となっています（オールスターでもホームランを打ちましたが、公式戦での記録には含まれません）。

確率分布と逆関数のいろいろ

Excel では、累積分布関数の逆関数として以下のようなものが利用できます（表 1）。表には、これまでに見てきた離散型確率分布だけでなく、連続型確率分布の逆関数も掲載しておきました。連続型確率分布については、次回以降解説するので、ここでは名前の掲載だけにとどめておきます。太字の分布と関数が既出のものです。

◆ 離散型確率分布

確率分布	Excelでの関数	Excelでの累積分布関数の逆関数
二項分布	**BINOM.DIST, BINOM.DIST.BETWEEN**	**BINOM.INV**
超幾何分布	**HYPGEOM.DIST**	なし
ポアソン分布	**POISSON.DIST**	なし
幾何分布	なし（**NEGBINOM.DISTで代用可能**）	なし（**Pythonでの方法を上で紹介**）
負の二項分布	**NEGBINOM.DIST**	なし（**Pythonでの方法を上で紹介**）
負の超幾何分布	なし	なし

◆ 連続型確率分布

確率分布	Excelでの関数	Excelでの逆関数
正規分布	NORM.DIST	NORM.INV
標準正規分布	NORM.S.DIST	NORM.S.INV
対数正規分布	LOGNORM.DIST	LOGNORM.INV
カイ二乗分布	CHISQ.DIST, CHISQ.DIST.RT	CHISQ.INV, CHISQ.INV.RT
t分布	T.DIST, T.DIST.2T, T.DIST.RT	T.INV, T.INV.2T
F分布	F.DIST, F.DIST.RT	F.INV, F.INV.RT
指数分布	EXPON.DIST	なし
ガンマ分布	GAMMA.DIST	GAMMA.INV
ベータ分布	BETA.DIST	BETA.INV
ワイブル分布	WEIBULL.DIST	なし

表 1　確率分布を求めるための関数と、累積分布関数の逆関数

確率分布と Excel で利用できる関数、逆関数の一覧。離散確率分布のうち、Excel での関数が用意されていないものについては、Python などのプログラミング言語を使って求めるのが簡単。「なし」と書かれているものについては、サンプルプログラムの中に利用例が含まれているので、そちらを参照のこと。

表 1 の離散型確率分布については、累積分布関数のグラフを作成するためのコードと、その逆関数の値を求めるためのコードを、上で紹介した Google Colaboratory でのサンプルプログラムに**全て**含めてあります。興味のある方はぜひご参照ください。

負の超幾何分布については、確率質量関数や累積分布関数を求めるための関数が Excel には用意されていないので、この連載でも触れませんでしたが、そらちについてもサンプルプログラムに含めてあります。負の超幾何分布は、負の二項分布と似ていますが、非復元抽出の場合の確率分布です。

- **負の二項分布：** n 回成功するまでに k 回失敗する確率（復元抽出）
- **負の超幾何分布：** n 回成功するまでに k 回失敗する確率（非復元抽出）

◇ ◇ ◇ ◇ ◇ ◇ ◇

今回は、累積分布関数の逆関数についてお話ししました。例えば、**95%**以上の確率で（ほぼ確実に）成功するのは**何回**までかを求めたり、逆に、**5%**未満の確率でしか成功しないのは**何回**以上なのかを求めたりできるので、応用の幅が広がりますね。

逆関数は、とりわけ、次回以降から始まる連続型確率分布でも重要になってきます。例えば、試験で上位 **5%** に入るには何点取らないといけないか、といったことが求められます。知りたいですよね！　また、上で少し触れたように、統計的検定の計算でも累積分布関数の逆関数が使われます。

というわけで、次回から連続型確率分布のお話に移りたいと思います。連続型確率分布の代表とも言える正規分布からスタートします。では、次回もお楽しみに！

この記事で取り上げた関数の形式

関数の利用例については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

二項分布の累積分布関数の値を求めるための関数（確率変数の範囲を指定）

BINOM.DIST.RANGE 関数：確率変数の値を指定して二項分布の累積分布関数の値を求める

形式

BINOM.DIST.RANGE(試行回数 , 成功確率 , 成功数 1, 成功数 2)

引数

- **試行回数：**試行の回数（n の値）を指定する。
- **成功確率：**1 回の試行で目的の事象が起こる確率を指定する。
- **成功数 1：**成功数の下限を指定する。
- **成功数 2:** 成功数の上限を指定する。省略すると、成功率 1 と同じ値が指定されたものと見なされる（成功数 1 の確率質量関数の値が返される）。

二項分布の累積分布関数に対する逆関数の値を求めるための関数

BINOM.INV 関数：二項分布の累積分布関数に対する逆関数の値を求める

形式

BINOM.INV(試行回数 , 成功確率 , 累積確率)

引数

- **試行回数：**試行の回数（n の値）を指定する。
- **成功確率：**1 回の試行で目的の事象が起こる確率を指定する。
- **累積確率:** 累積確率を指定する。累積分布関数の値がここで指定した累積確率以上になる k（成功数）はいくらまでかが求められる。

［データ分析］正規分布
～ 私より背の高い人はどれぐらいいるの？

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載（確率分布編）の第 6 回。正規分布は平均値を「山」の中心として、標準偏差によって左右対称に「すそ」が広がるような形の連続型確率分布です。正規分布がどのようなものかを確認した後、確率密度関数や累積分布関数の求め方や可視化の方法を解説し、利用例などを紹介していきます。

羽山博（2024 年 08 月 29 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の確率分布編、第 6 回です。前回は、離散型確率分布のまとめとして、さまざまな離散型確率分布の逆関数とその利用例を取り上げました。今回から連続型確率分布のお話に入ります。まずは、連続型確率分布の代表とも言われる正規分布について、その特徴や意味、確率密度関数／累積分布関数の求め方、利用例などを見ていきます。

正規分布で自分の位置を知るには

厚生労働省による国民健康・栄養調査（2019）によると、20 歳以上の日本人男性（サンプルサイズ：**1968 人**）の平均身長は **167.7cm**、標準偏差は **6.9** でした。身長の分布が**正規分布**（後で詳しく見ます）に従っているものと**仮定**すると（2024/08/30 追記：本来は、母集団が正規分布に従っているかどうかを確認する必要がありますが、仮にそうだとすると）、この集団の中で **170cm** の人はどのあたりの位置にいると考えられるでしょうか（図 1）。

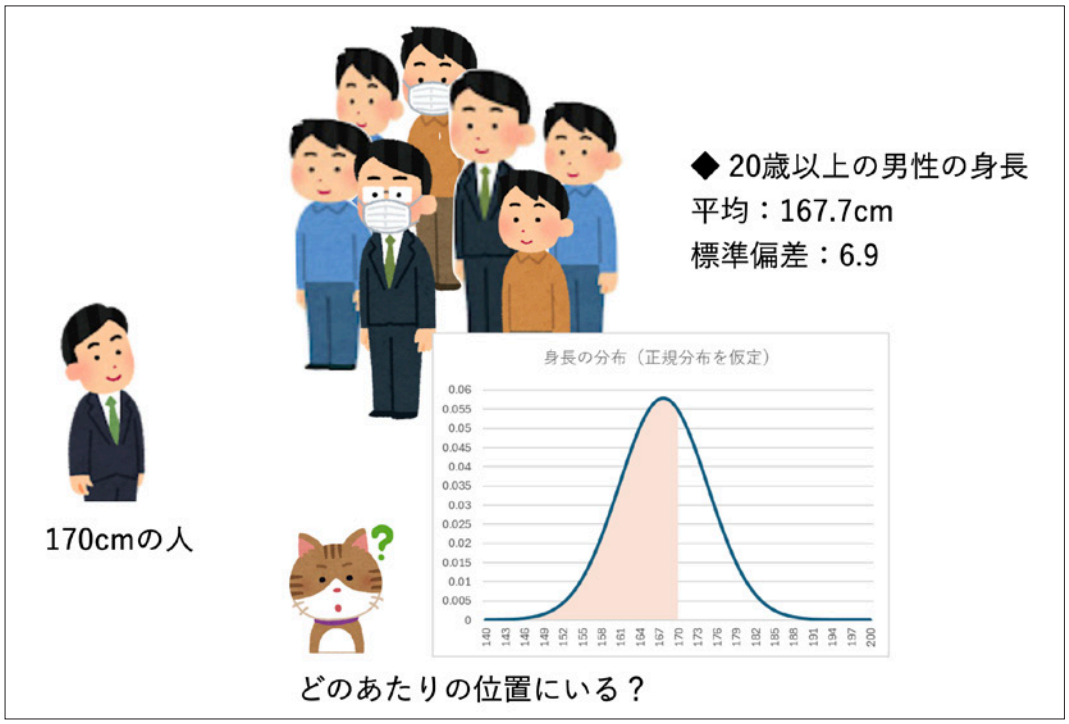


図 1　身長 170cm の人は全体のどのあたりの位置にいるのか

平均 **167.7**、標準偏差 **6.9** の正規分布で、**170** という値は全体のどのあたりだろうか。平均よりも大きいことは分かるが、数値で**何%**と表すならいくらになるだろう。

→目次に戻る

答えから先に言うと、下位から 63.1%（上位から 36.9%）の位置です。が、その値を求めるためには正規分布についての理解を深めておく必要があります。というわけで、今回は正規分布について丁寧に見ていきましょう（求め方だけを先に知りたい方はこちらへ）。まずは、離散と連続のおさらいからです。

当然のことながら、身長が高いから偉いというわけではありません。「上位」「下位」は単に値の大小を表しているだけで、価値判断とは全く関係ありません。

「離散」と「連続」のおさらいから ～「とびとび」か「スキマがない」か

正規分布は連続型確率分布の代表的な分布です。確率分布の意味や**離散**と**連続**の違いについては、この連載の第 1 回で解説しました。簡単におさらいしておきましょう。

まず、確率変数からです。確率変数とはある事象に対して割り当てた値のことで、通常大文字の X で表します（これまでは特に明示していませんでしたが）。離散型確率分布では、確率変数 X の値として k という変数名を使うのが一般的です。つまり、X は確率変数そのものを一般的に表し、k は確率変数の具体的な値を表す変数だというわけです。

離散型確率分布では、確率変数がとびとびの値を取ります。例えば、図 2 のようなサイコロのそれぞれの目が出る確率は離散型一様分布となり、確率変数 X は **1, 2, 3, 4, 5, 6** のいずれかの値を取ります。**1.5** などという目が出ることはありませんね。

さらにおさらいです。ベルヌーイ分布の確率変数 X は **0** か **1** の値を取ります。例えば、サイコロで **1** の目が出る場合を X = 1 とし、1 以外の目が出る場合を X = 0 とします。また、二項分布では確率変数 X が **0**、**1**、**2**、……、**n** という値を取ります。例えば、サイコロを n 回投げて、**1** の出る回数が X に当たります。

一方、連続型確率分布では、確率変数 X は範囲内のどの値でも取れます。連続型確率分布では、確率変数 X の具体的な値を表す変数名として小文字の x を使うのが一般的です。

連続型確率分布の確率変数の例としては、気温や身長などがあります。例えば、気温は下限から上限までのどの値でも取れます。温度計には目盛りが付いていますが、それはあくまでも目安です。目盛りの間はいくらでも細かく（スキマなく）分けられますね。

図 2　離散と連続のイメージ

離散とは値がとびとびになっていること。連続とは範囲内のどの値でも取ることができるということ。温度計には目盛りが付いているが、それはあくまでも目安であることに注意。

では、ちょっとした問題です。デジタル体温計で測定した場合の体温は離散値でしょうか、連続値でしょうか。デジタル体温計での測定値は **36.5℃**のように **0.1℃**刻みになっていて、その間の値（**36.51...℃**など）は表示されないので、離散値と思われるかもしれません。しかし、それはあくまでも体温計の精度（有効数字）の問題です。体温は連続値として分析するのが適切です。

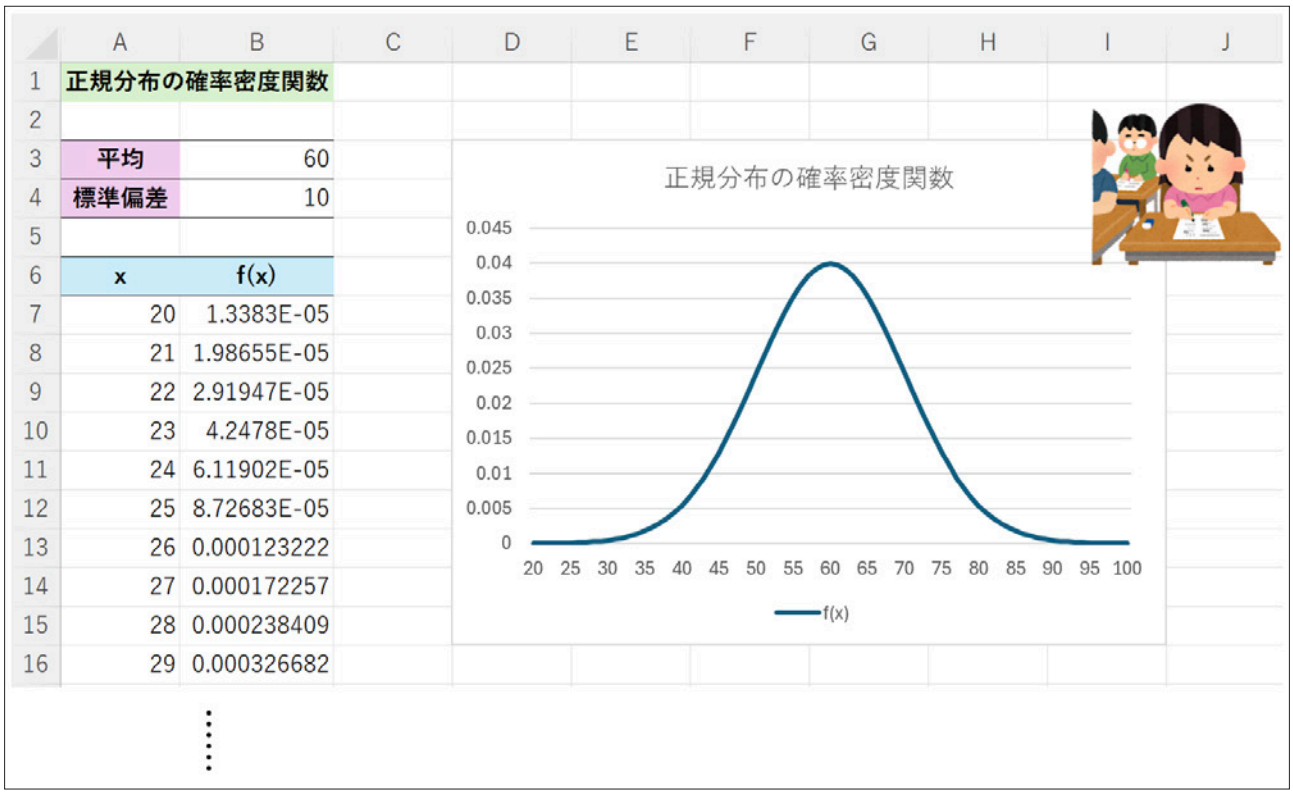
日常の感覚では、1、2、3……という値は連続しているものと捉えられますが、英語では、このような連番は、何らかのモノが順に並んだという意味合いの serial という単語を使って、serial number と呼びます。それに対して、上で見た「連続」を表す単語は、切れ目なく続くといった語感の continuous です。どちらにも「連続」という訳語を当てるので、日本語だと紛らわしいですね。また、整数だから離散、小数だから連続という意味ではないということについても注意しましょう。

正規分布ってどんな感じの分布？ ～ 確率密度関数を可視化してみよう

身長や試験の成績などは、母集団が正規分布に従っているものと仮定して分析を行うことがよくあります。正規分布は平均μと標準偏差σで決まる分布です。分布を一意に決めるこれらの値を**母数（パラメーター）**と呼ぶ、ということはこの連載の第 1 回で説明しましたね。

正規分布の**確率密度関数**を可視化すると、図 3 のようになります。このことについてはすでにご存じの方も多いでしょう。平均のところが一番高くなっていて、左右対称に裾野が広がるような形です。なんとなくですが、成績の分布のように見えますね。

いやいや、それ以前に「確率密度関数って何？」という方も多いと思いますが、理屈は後回しにして、図 3 のグラフを先に作成してみましょう。手順は図の後に記してあります。正規分布の確率密度関数の値は NORM.DIST 関数で求められます。引数には、確率変数の値と、母数（平均と標準偏差）を指定します。ここでは、平均$\mu = 60$、標準偏差$\sigma = 10$ とします。



	A	B
1	正規分布の確率密度関数	
2		
3	平均	60
4	標準偏差	10
5		
6	x	f(x)
7	20	1.3383E-05
8	21	1.98655E-05
9	22	2.91947E-05
10	23	4.2478E-05
11	24	6.11902E-05
12	25	8.72683E-05
13	26	0.000123222
14	27	0.000172257
15	28	0.000238409
16	29	0.000326682

図 3　正規分布の確率密度関数の例

正規分布（平均$\mu = 60$、標準偏差$\sigma = 10$）の確率密度関数。ここでは、$x = 20 \sim 100$ について可視化した。平均値のところが一番高い山になっており、左右対称に裾野が広がるグラフになる。

グラフ作成の手順は以下の通りです。サンプルファイルをこちらからダウンロードし、［正規分布］ワークシートを開いて試してみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。具体的な操作方法は、サンプルファイル内に記載しています。

→目次に戻る

◆ **Excel での操作方法**（タイトルや軸の書式などの細かい設定は省略）

- セル **B7** に =NORM.DIST(A7:A87,B3,B4,FALSE) と入力する
  - ・古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **B7** ～ **B87**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
- セル **B6** ～ **B87** を選択する
- ［挿入］タブを開き、［折れ線／面グラフの挿入］ボタンをクリックして［折れ線］を選択する
- ［グラフのデザイン］タブを開き、［データの選択］ボタンをクリックする
- ［データソースの選択］ダイアログボックスで［横（項目）軸ラベル］の下の［編集］ボタンをクリックする
- ［軸ラベル］ダイアログボックスで［軸ラベルの範囲］ボックスをクリックし、セル **A7** ～ **A87** を選択する
- ［OK］をクリックして［軸ラベル］ダイアログボックスを閉じる
- ［OK］をクリックして［データソースの選択］ダイアログボックスを閉じる

正規分布の確率密度関数 f(x) は以下の式で定義されますが、NORM.DIST 関数で求められるので、この式を無理に暗記する必要はありません。

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad (1)$$

exp(x) は自然対数の底 e = 2.71828... の x 乗を表します。なお、文献によっては、

$$\sqrt{2\pi}\sigma$$

の部分が

$$\sqrt{2\pi\sigma^2}$$

と表記されている場合もあります。

NORM.DIST 関数の引数は以下のように指定します。

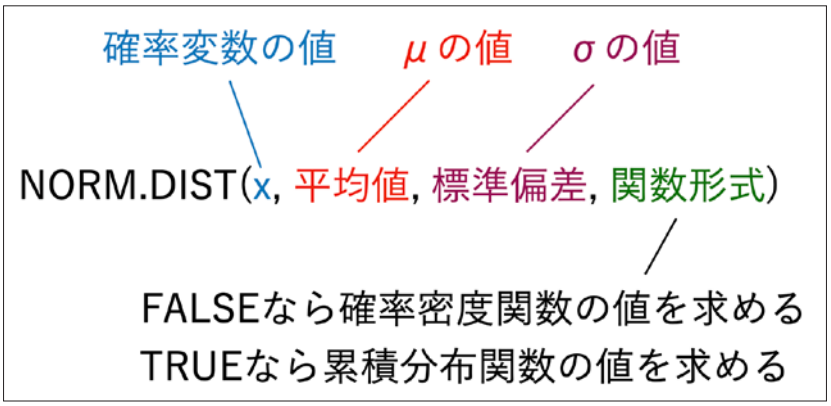


図 4　NORM.DIST 関数に指定する引数

図 3 の例では、関数形式として **FALSE** を指定し、確率変数の値として A7:A87 というセル範囲を指定しているので、スピル機能によりセル **A7** ～ **A87** の値に対する確率密度関数の値が求められる。関数形式に **TRUE** を指定すれば累積分布関数の値が求められる。

→目次に戻る

連続型確率分布や正規分布についての留意点

以下に、連続型確率分布や正規分布についての留意点を箇条書きにしておきます。

- 離散型確率分布での f(k) は**確率質量関数（probability mass function）**と呼ぶが、連続型確率分布での f(x) は**確率密度関数（probability density function）**と呼ぶ
- f(x) の値は、確率変数の値が x のときの確率ではない。この値は累積分布関数の微分係数と考えられる（後述）
- グラフと x 軸とで囲まれた範囲の面積は **1** である
- 正規分布では、台（x の取り得る値の範囲）は $-\infty$〜∞

特に、**確率密度関数 f(x) の値は、確率変数の値が x のときの確率ではない**ことに注意してください。図 3 の例で具体的に言うと、$x = 60$ のとき、f(x) の値は **0.03989** ですが、この値は $x = 60$ である確率ではないということです。**連続型確率分布では、確率変数の特定の値に対する確率を求めることはできません。**しかし、累積分布関数（後述します）によって、一定の範囲（例えば、$x = 40 \sim 60$）に対する累積確率を求めることはできます。**累積分布関数の値は、確率変数の値がある範囲に入る確率です。**

じゃあ、f(x) の値っていったい何、と思われる方も多いでしょう。上の箇条書きでは「累積分布関数の微分係数」と記していますが、日常の言葉で言えば、例えば、「$x = 60$ でどの程度、累積確率が増えるか」という度合い（傾向）を表すもの、といった答えになります。

そもそも正規分布って何？ 〜 二項分布と正規分布の関係

前回までの離散型確率分布では、サイコロを振ったり、ガラガラを回したり、野球選手のヒットの出る確率を使ったり……と、具体的な例を基に確率質量関数がどのような式で表されるのかを見てきました。しかし、今回の正規分布では、実世界での事例を基に確率密度関数を表す（1）式をどう導き出すのかよく分かりませんね。

実は、離散型確率分布の代表格とも言える二項分布と、連続型確率分布の代表格とも言える正規分布には極めて深い関係があります。二項分布の p を変えずに n をどんどん増やしていくと、平均 np、分散 $np(p-1)$ の正規分布に近づくことが分かっています。

ちなみに、二項分布の np を変えずに n をどんどん増やしていくと（その場合、p がどんどん小さくなる）、ポアソン分布になります。この連載の第 4 回で取り上げました。

式の導出は高校の数学でもできますが、かなり大変なので割愛します。その代わり、Excel を使って具体的な例でその様子を可視化してみましょう。図 5 は二項分布の n を大きくした確率質量関数のグラフ（棒グラフ）と、正規分布（平均 np、分散 $np(p-1)$）のグラフ（折れ線グラフ）を重ねてみた例です。$n = 20$ ぐらいでもほぼ重なっていることが分かります。

	A	B	C	D
1	二項分布（n=20, p=0.3）と正規分布（μ=6, σ^2=4.2）			
2				
3	n	20	np	6
4	p	0.3	np(1-p)	4.2
5				
6	k	二項分布	正規分布	
7	0	0.000798	0.002679	
8	1	0.006839	0.009925	
9	2	0.027846	0.028977	
10	3	0.071604	0.066676	
11	4	0.130421	0.120915	
12	5	0.178863	0.172816	
13	6	0.191639	0.194664	
14	7	0.164262	0.172816	
15	8	0.114397	0.120915	
16	9	0.06537	0.066676	
17	10	0.030817	0.028977	
18	11	0.012007	0.009925	
19	12	0.003859	0.002679	

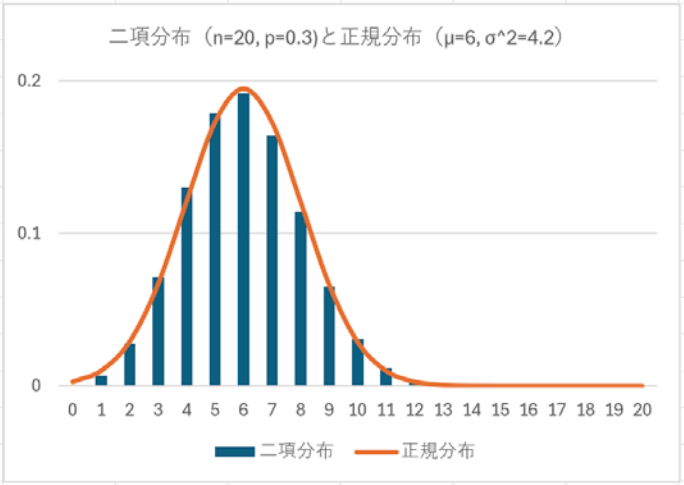


図 5　二項分布と正規分布の関係

n = 20 は **20 回**の打数を、p = 0.3 は **3** 割打者を表すと考えればよい。**20 回**の打数のうちヒットが出る回数とその確率は二項分布で求められる（棒グラフの部分）。$\mu = 6$, $\sigma = 4.2$ の正規分布のグラフ（折れ線グラフ）を描いてみると、それらがほぼ重なることが分かる。

図 4 のグラフを作成する方法については、サンプルファイルの［二項分布と正規分布］ワークシート中に掲載しています。ここでは、各セルに入力されている数式のみを記しておきます。

- セル **D3**：=B3*B4　…… $\mu = np$ の値
- セル **D4**：=D3*(1-B4)　…… $\sigma^2 = np(1-p)$ の値
- セル **B7**：=BINOM.DIST(A7:A27,B3,B4,FALSE)　…… 二項分布の確率質量関数の値
- セル **C7**：=NORM.DIST(A7:A27,D3,SQRT(D4),FALSE)　…… 正規分布の確率密度関数の値

正規分布の平均 μ は $np = 20 \times 0.3 = 6$ とし、分散 σ^2 は $np(1-p) = 6 \times (1-0.3) = 4.2$ としています。NORM.DIST 関数には分散ではなく標準偏差を指定するので、3 番目の引数が SQRT(D4) となっていることに注意してください。

正規分布の累積分布関数を可視化してみよう

すでに述べたように、連続型確率分布では、確率変数の特定の値に対する確率を求めることはできません。しかし、累積分布関数によって、一定の範囲に入る確率（累積確率）を求めることはできます。累積分布関数は基本的には離散型確率分布の場合と同じ考え方で、確率変数 X の値が x となる**まで**の累積確率を関数として表したものです。累積分布関数がどのようなものであるか、図 6 で確認しておきましょう。グラフの作成手順は図の後に記してあります。

→目次に戻る

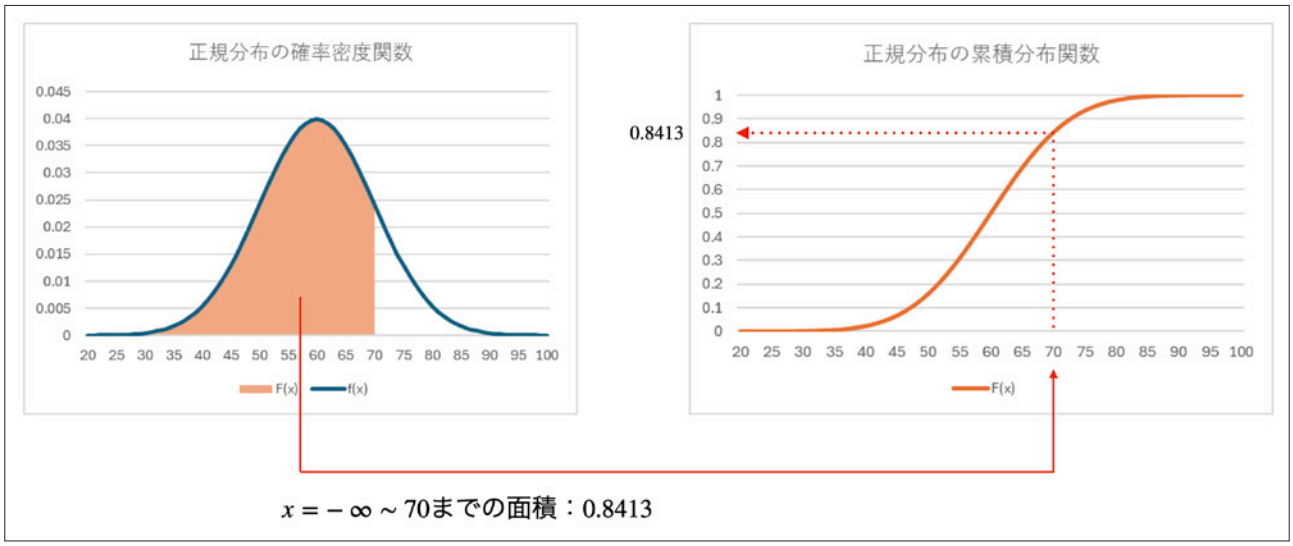


図 6　正規分布の確率密度関数 f(x) と累積分布関数 F(x)

正規分布の台は x = −∞ ～∞だが、x 軸を**−∞～∞**にすることはできないので、目盛りを **20** ～ **100** までとした。左側の確率密度関数のグラフでは、累積確率はグラフと x 軸で囲まれた部分の面積で表される。例えば、x = 70 までの累積確率はオレンジ色で塗りつぶした部分の面積（= **0.8413**）となる。右側の累積分布関数のグラフは x に対する累積確率をプロットしたもの。例えば、x = 70 に対する F(x) の値は **0.8413** になる。

図 6 の左側は正規分布の確率密度関数です。x 軸とグラフで囲まれた面積が累積確率に当たります。例えば、x = −∞ ～ 70 **まで**の累積確率はオレンジ色で塗りつぶされた部分の面積（= **0.8413**）になります。一方、x に対する累積確率（面積）をプロットしていったものが累積分布関数です。こちらは右側のようなグラフになります。右側のグラフでは、x = 70 に対する F(x) の値が **0.8413** になります。

高校数学の記憶がある方は、図 6 を見て、累積分布関数 F(x) は確率密度関数 f(x) を積分したものだと気付くと思います。数式で表すと以下のようになります。

$$F(x) = \int_{-\infty}^{x} f(x)dx = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right) \qquad (2)$$

ただし、erf は誤差関数と呼ばれる以下の関数です。

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt \qquad (3)$$

Excel では、正規分布の累積分布関数の値も NORM.DIST 関数で求められるので、上の（2）式や（3）式を無理に覚える必要はありません（ちなみに、誤差関数の値も ERF.PRECISE 関数で求められますが、これ以上は触れないことにします）。

グラフ作成の手順は以下の通りです。サンプルファイルの［確率密度関数と累積分布関数］ワークシートを開いて試してみてください。Google スプレッドシートでは、サンプルファイル内に手順を記載しています。

◆ **Excel での操作方法**（タイトルや軸の書式などの細かい設定は省略）

- **確率密度関数のグラフ作成**
  - セル **B7** に =NORM.DIST(A7:A87,B3,B4,FALSE) と入力する
    - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **B7** ～ **B87**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
  - セル **B6** ～ **B87** を選択する
  - ［挿入］タブを開き、［折れ線／面グラフの挿入］ボタンをクリックして［折れ線］を選択する
  - ［グラフのデザイン］タブを開き、［データの選択］ボタンをクリックする
  - ［データソースの選択］ダイアログボックスで［横（項目）軸ラベル］の下の［編集］ボタンをクリックする
  - ［軸ラベル］ダイアログボックスで［軸ラベルの範囲］ボックスをクリックし、セル **A7** ～ **A87** を選択する
  - ［OK］をクリックして［軸ラベル］ダイアログボックスを閉じる
  - ［OK］をクリックして［データソースの選択］ダイアログボックスを閉じる
- **累積確率の塗りつぶし**
  - ［グラフのデザイン］で［データの選択］ボタンをクリックする
  - ［データソースの選択］ダイアログボックスで［凡例項目（系列）］の下の［追加］ボタンをクリックする
  - ［系列の編集］ダイアログボックスで［系列名］ボックスをクリックし、セル **B6** を選択する
  - ［系列の編集］ダイアログボックスで［系列値］ボックスをクリックし、すでに入力されている {=1} を削除してから、セル **B7** ～ **B57** を選択する
  - ［OK］をクリックして［系列の編集］ダイアログボックスを閉じる
  - ［OK］をクリックして［データソースの選択］ダイアログボックスを閉じる
  - ［グラフのデザイン］で［グラフの種類の変更］ボタンをクリックする
  - ［グラフの種類の変更］ダイアログボックスで、左のリストから［組み合わせ］を選択する
  - ［系列名］に「f(x)」と表示されている行の［グラフの種類］リストから［折れ線］を選択する
  - ［系列名］に「F(x)」と表示されている行の［グラフの種類］リストから［面］を選択する
  - ［OK］をクリックして［グラフの種類の変更］ダイアログボックスを閉じる
- **累積分布関数のグラフ作成**
  - セル **C7** に =NORM.DIST(A7:A87,B3,B4,TRUE) と入力する
  - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **C7** ～ **C87**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
  - セル **C6** ～ **C87** を選択する
  - ［挿入］タブを開き、［折れ線／面グラフの挿入］ボタンをクリックして［折れ線］を選択する
  - ［グラフのデザイン］タブを開き、［データの選択］ボタンをクリックする
  - ［データソースの選択］ダイアログボックスで［横（項目）軸ラベル］の下の［編集］ボタンをクリックする
  - ［軸ラベル］ダイアログボックスで［軸ラベルの範囲］ボックスをクリックし、セル **A7** ～ **A87** を選択する
  - ［OK］をクリックして［軸ラベル］ダイアログボックスを閉じる
  - ［OK］をクリックして［データソースの選択］ダイアログボックスを閉じる

ずいぶん遅くなりましたが、冒頭の問題に対する答えの求め方を見ておきましょう。平均 **167.7**、標準偏差 **6.9** の正規分布で、**170** という値が全体のどの位置に当たるかは累積分布関数で求められます。具体的には `=NORM.DIST(170, 167.7, 6.9, TRUE)` で求められます。空いているセルに関数を入力して確認してみてください（サンプルファイルの完成例ではセル **E32** に入力されています）。**0.6305...** という値が表示されるはずです。下位から **63.1%**の位置であるということですね。

正規分布はどんなところに現れるのか？ ～ 中心極限定理とは

二項分布の試行を数多く行うと、正規分布に近づくことはすでにお話ししました。正規分布を日常の感覚と直接結び付けて考えるのは難しいですが、その延長線上にあることは理解できたと思います（今回のお話がちょっと理屈っぽくなったのもそのためですね）。そこで、正規分布がどんなところに現れるのかを見ておきましょう。**中心極限定理**と呼ばれる重要な定理についてお話しします。

中心極限定理とは、**母集団の分布がどのような分布であっても**、そこからサンプルを何度も取り出すと、それらの平均値

$$\bar{x}$$

が正規分布に近づくという定理です。具体的には母集団の平均をμ、分散をσ^2としたとき、n 個のサンプルを何度も取り出すと、

$$\bar{x}$$

は平均μ、分散σ^2/n の正規分布に近づくということです。

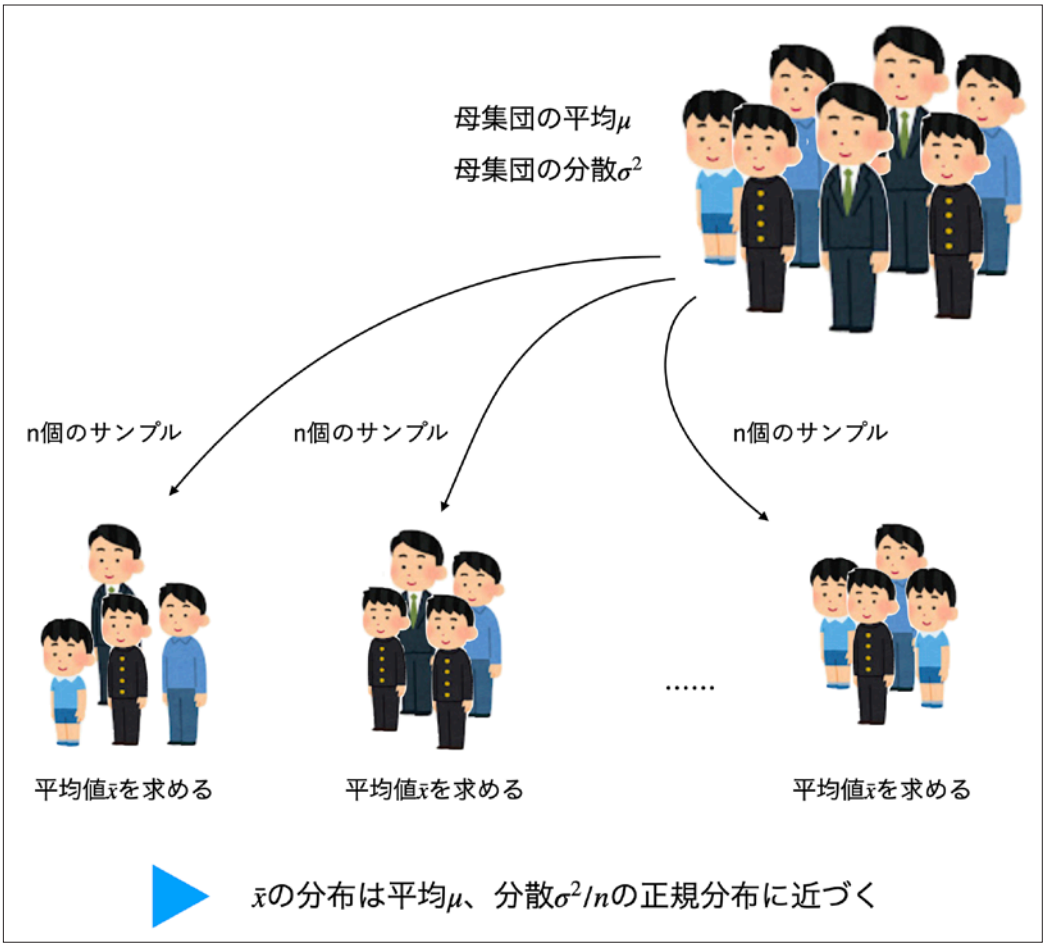


図 7　中心極限定理のイメージ

平均μ、分散σ^2の母集団から、n個のサンプルを何度も取り出すと、
それらの平均$\bar{x}$の分布は、平均μ、分散σ^2/nの正規分布に近づく。

中心極限定理についても証明はかなり難しくなるので、例で見てみましょう。母集団の分布は何でもいいので、ここでは a ≤ x ≤ b の連続型一様分布を使って試します。連続型一様分布の平均μと分散σ^2は以下の通りです。

$$\mu = \frac{a+b}{2}$$
$$\sigma^2 = \frac{(b-a)^2}{12}$$

そこで、a = − 6, b = 6 の一様分布を考えます。ここから n = 12 個のサンプルを **10000 回**取り出して、それぞれの平均値を求め、ヒストグラムを作ってみましょう。この場合、

$$\mu = \frac{-6+6}{2} = 0$$
$$\sigma^2 = \frac{(6-(-6))^2}{12} = 12$$

なので、中心極限定理によって、平均と分散がそれぞれ、

→目次に戻る

$$\mu = 0$$
$$\sigma^2/n = 12/12 = 1$$

の正規分布になるはずです。

このシミュレーションは Excel を使ってもできますが、多数のセルを使う必要があり（地道にやるなら **12×10000 個以上**）、かなり面倒です。一応、サンプルファイルに作成例は含めてあります（［中心極限定理］ワークシート）が、ここでは、Python のプログラムを使うことにします（リスト 1）。

サンプルプログラムはこちらから参照できます。リンクをクリックすれば、ブラウザが起動し、Google Colaboratory の画面が表示されます（Google アカウントでのログインが必要です）。コードセルをクリックし、［Shift］+［Enter］キーを押してコードを実行してみてください。結果は図 8 のようになります。コードの詳細については解説しませんが、コメントとリスト 1 の説明を見れば何をやっているかが大体分かると思います。

```
import numpy as np
from scipy.stats import  norm
import matplotlib.pyplot as plt

# 12 個の一様乱数を 10000 回作る（10000 行× 12 列）
sample = np.random.rand(10000, 12) * 12 - 6  # 0 ～ 1 の乱数なので、12 倍して 6 を引けば -6 ～ 6 の範囲になる
# 各行の平均値を求める
means = sample.mean(axis=1)
# ヒストグラムを作成する（範囲は -4 ～ 4 まで、階級は 100 個、縦軸を確率とする）
plt.hist(means, range=(-4, 4), bins=100, density=True)

# 標準正規分布
x = np.linspace(-4, 4, 100)  # -4 ～ 4 までを 100 個に分けた数列を作る
y = norm.pdf(x, loc=0, scale=1)  # 標準正規分布の確率密度関数の値を求める
plt.plot(x, y) # グラフを作る

# グラフを表示する
plt.show()
```

リスト 1　中心極限定理のシミュレーションを行う

NumPy の random.rand 関数に、行数と列数を指定すれば、一様乱数の配列が作成できる。後は、mean メソッドを使って各行の平均値を求め、matplotlib.pyplot モジュールの hist 関数を使ってヒストグラムを作成するだけ。density=True は「縦軸を確率にする」という指定。さらに、scipy.stats モジュールの norm.pdf 関数を使い、平均 loc=0、標準偏差 scale=1 の正規分布の確率密度関数の値を求め、グラフを重ねて描く。実行例は図 8 を参照。

なお、平均が **0**、標準偏差が **1**（分散も **1** となる）の正規分布を**標準正規分布**と呼びます。a = − 6, b = 6, n = 12 としたのはシミュレーションの結果が標準正規分布に近くなるようにするためです。

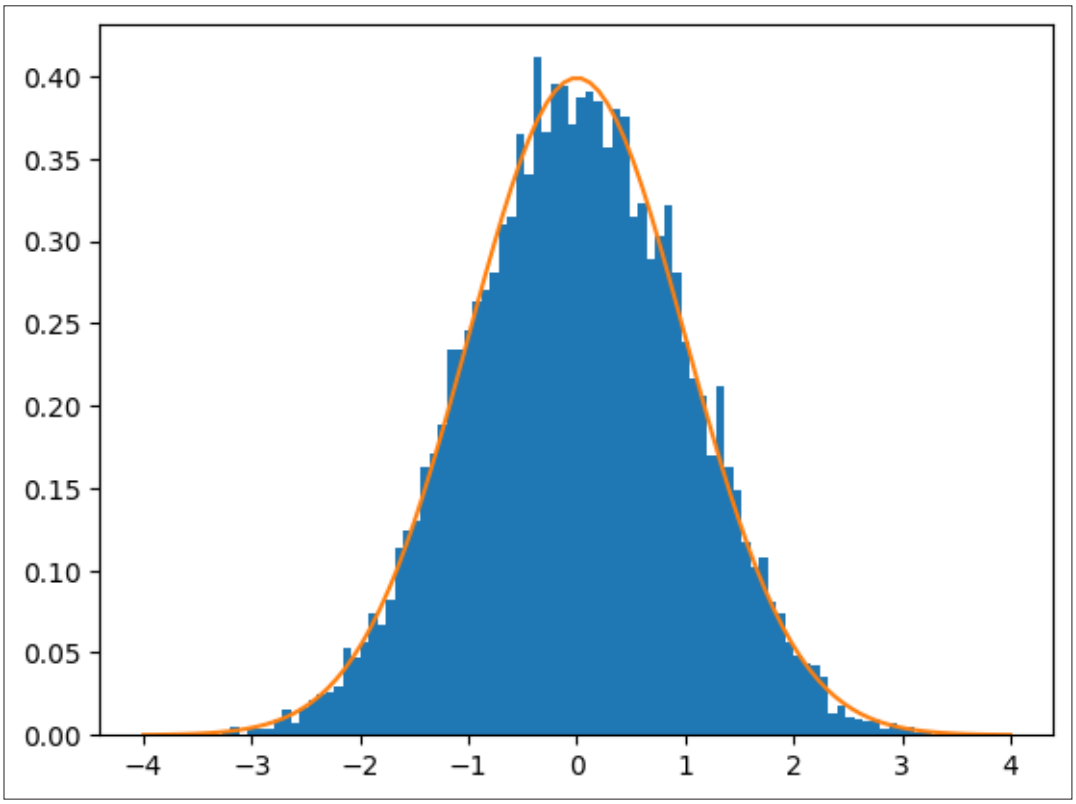


図 8　n 個のサンプルの何度も取り出して求めた平均値のヒストグラムと標準正規分布のグラフ

棒グラフは、**− 6 ～ 6** までの一様乱数から **12 個**のサンプルを **10000 回**取り出して、それらの平均値を基に作成したヒストグラム。折れ線グラフは平均 **0**、標準偏差 **1** の標準正規分布。ほとんど重なっていることが分かる。

一様乱数は毎回異なる値が作成されるので、図 8 のヒストグラムも毎回異なる形になりますが、標準正規分布とほぼ重なります。実は、二項分布の n を大きくしていくと正規分布に近づくというのも中心極限定理によるものです。

母集団が正規分布であると仮定されない場合でも、平均値などの分析のために広く正規分布が使われるのは、この中心極限定理を根拠としています。なお、このように、前提を多少満たさないとしても、妥当な結果が得られることを**頑健性（robustness）**があると言います。

標準正規分布の確率密度関数と累積分布関数の値を Excel で求めるには、NORM.DIST 関数を使って、平均値に **0** を、標準偏差に **1** を指定してもいいのですが、NORM.S.DIST 関数も利用できます。NORM.S.DIST 関数を使えば、標準正規分布であることが明確に分かりますし、引数の指定も簡単になります。関数の形式については、この記事の最後をご参照ください。

異なる集団の間で値の大小を比較するには ～ 標準化の方法と偏差値の求め方

冒頭で示した身長のお話ですが、国民健康・栄養調査（2019）には、年齢層による平均値と標準偏差も掲載されています。それによると、20 歳代の男性は身長の平均が **171.5cm**、標準偏差が **6.6** です。一方、60 歳代の男性は身長の平均が **167.4cm**、標準偏差が **6.0** です。では、これらの集団の中で、身長 **170cm** の 20 代男性と、身長 **170cm** の 60 代男性ではどちらの方が背が高いでしょう？

同じ値なので身長は同じというのはもっともな答えです。しかし、「集団の中で」という条件が入っていると話は変わります。20 歳代の平均身長は **171.5cm** で、60 歳代の平均身長は **167.4cm** なので、集団の中で考えると、20 歳代の **170cm** は低い方で、60 歳代の **170cm** は高い方だろうとも考えられます。このように、平均値や標準偏差が異なる集団の間で、どの位置にいるかを比較したいことがありますね。そのために使われる値が偏差値です。そこで、偏差値の求め方と意味を見ていきましょう。

平均値と標準偏差が異なっていると単純に比較はできないので、まず、それぞれの分布を、平均値 **0**、標準偏差 **1** の標準正規分布になるように値を調整します。そのために、各データ x から平均値μを引き、標準偏差σで割ります。つまり、

$$z = \frac{x - \mu}{\sigma}$$

という計算を行います。このような計算を行うことを標準化と呼び、z の値を標準得点（または、z 値など）と呼びます。

一般に、

母集団の平均はμ、標準偏差はσ

という文字で表され、

サンプルの平均は$\bar{x}$、標準偏差はs

という文字で表されます。ここでは、サンプルとして取り出された集団を母集団そのものと考えているので、μとσを使っています。

例えば、20 歳代の **170cm** という値を標準化すると、

$$z = \frac{170 - 171.5}{6.6} \approx -0.2273$$

となり、60 歳代の **170cm** であれば、

$$z = \frac{170 - 167.4}{6.0} \approx 0.4333$$

となり、60 歳代の **170cm** の方が集団の中での位置としては上位であることが分かります。

Excel では STANDARDIZE 関数の引数に x の値、平均値、標準偏差を指定すると標準得点が求められます。サンプルファイルの［偏差値］ワークシートを開いて、セル **F4** に =STANDARDIZE(E4,B4,C4) と入力して、セル **F5** にコピーしましょう。それぞれ、**-0.22727**...、**0.433333**... という結果が得られるはずです（図 9）。数式を入力しても計算しても大した手間ではありませんが、関数を使った方が標準得点を求めていることが明確に分かりますね。

	A	B	C	D	E	F	G	H
1	**偏差値を求める**							
2								
3	**年代**	**平均**	**標準偏差**		**身長**	**標準得点**	**偏差値**	
4	20歳代	171.5	6.6		170	-0.22727		
5	60歳代	167.4	6.0		170	0.433333		

図 9　標準化を行って異なる集団の値同士を比較する

セル **F4** の値は =(E4-B4)/C4 でも求められるが、=STANDARDIZE(E4,B4,C4) と入力した方が標準得点を求めていることが明確に分かる。セル **F5** にはセル **F4** の式をコピーすればよい。スピル機能を利用するなら、セル **F4** に =STANDARDIZE(E4:E5,B4:B5,C4:C5) と入力するだけでよい

ところで、標準得点は **0** ～ **1** の範囲の小数なので、私たちにとっては実感の湧きにくい値です。そこで、さらに平均が **50**、標準偏差が **10** になるように調整します。そのために、**10** を掛けて、**50** を足します。その値、つまり、**標準得点 ×10 ＋ 50 で求められる値が偏差値**です。

$$\frac{x-\mu}{\sigma} \times 10 + 50$$

この式で計算すると、20 歳代の **170cm** の偏差値は $-0.2273 \times 10 + 50 \approx 47.7$ となり、60 歳代の **170cm** の偏差値は $0.4333 \times 10 + 50 \approx 54.33$ となります。Excel でも計算してみましょう（図 10）。

	A	B	C	D	E	F	G	H
1	**偏差値を求める**							
2								
3	**年代**	**平均**	**標準偏差**		**身長**	**標準得点**	**偏差値**	
4	20歳代	171.5	6.6		170	-0.22727	47.72727	
5	60歳代	167.4	6.0		170	0.433333	54.33333	

図 10　標準得点 ×10+50 が偏差値

セル **G4** に =F4*10+50 と入力し、セル **G5** にコピーすればそれぞれの偏差値が求められる。スピル機能を利用するなら、セル **G4** に =F4:F5*10+50 と入力するだけでよい。平均が **50** になるので、日常的な感覚で比較しやすい値になる。

上の例のように、平均値が異なると偏差値を求めなくてもある程度の察しは付きますが、平均値が等しい場合は判断しづらいですね。例えば、同じ調査で、30 歳代の男性は身長の平均が **171.5cm**、標準偏差が **5.5** でした。40 歳代の男性も身長の平均は **171.5cm** ですが、標準偏差は **5.8** です。それぞれ、**170cm** の人の偏差値を求めてみましょう。

- **30 歳代の 170cm の人：**

$$\frac{170-171.5}{5.5}\times 10+50\approx 47.3$$

- **40 歳代の 170cm の人：**

$$\frac{170-171.5}{5.8}\times 10+50\approx 47.4$$

というわけで、40 歳代の **170cm** の人の方がわずかに上位にいるようです。Excel での計算方法は図 9 と図 10 で見た例と同様です。入力例は［偏差値（完成例）］ワークシートをご参照ください。

上位 10%に入るには何点取らないといけないのか ～ 累積分布関数の逆関数

正規分布の逆関数を利用すれば、例えば、試験で上位**何%**の位置にいるには**何点**を取る必要があるのか、といった計算ができます。最初に見た、平均$\mu = 60$、標準偏差$\sigma = 10$ の正規分布で考えてみましょう。例えば、上位 **10%**に入るために**何点**取らないといけないのかを求めてみます。

上位 **10%**ということは下位から **90%（= 0.9）**ということになりますね。**0.9** という累積確率を基に、累積分布関数の逆関数の値を求めようというわけです（図 11）。

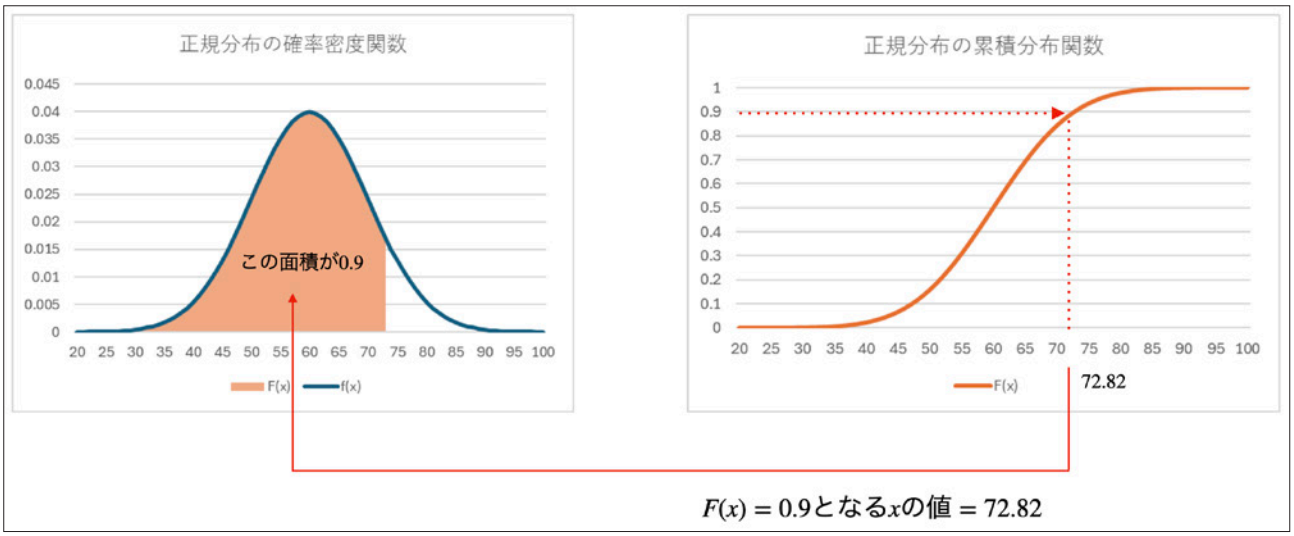


図 11　正規分布の累積分布関数に対する逆関数の値を求める

正規分布の累積分布関数（右側）で、縦軸の **0.9** という値から逆にたどっていけば、そのときの x の値が分かる（**72.82** となる）。この値は、確率密度関数（左側）のグラフと x 軸で囲まれた範囲の面積が **0.9** になるときの x の値。

すでに図 11 の中に答えが書いてありますが、正規分布の累積分布関数に対する逆関数の値を求める NORM.INV 関数を使って答えを求めてみましょう。[累積分布関数の逆関数]ワークシートを開いて、セル **B7** に =NORM.INV(A7,B3,B4) と入力してみてください。

	A	B	C	D
1	正規分布の累積分布関数の逆関数			
2				
3	平均	60		
4	標準偏差	10		
5				
6	F(x)	x		
7	0.9	72.81552		

図 12　NORM.INV 関数を使って累積分布関数に対する逆関数の値を求める

NORM.INV 関数には、累積確率、平均、標準偏差を指定する。下位から **90%**（上位 **10%**）の位置は **72.82** 点であることが分かる。

すでに見たとおり、結果は **72.82 点**となります。**72.82 点**以上取れば、上位 **10%**以内に食い込めるというわけです。NORM.INV 関数の引数は以下の図 13 のように指定します。

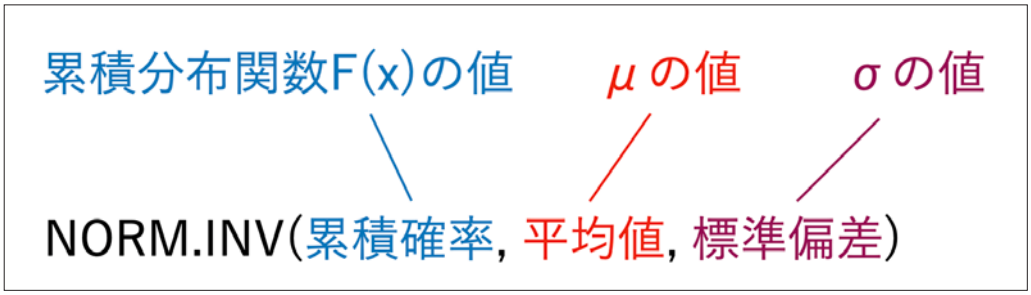


図 13　NORM.INV 関数に指定する引数

図 12 では、累積確率としてセル **A7** の **0.9** を指定し、平均値としてセル **B3** の **60** という値、標準偏差としてセル **B4** の **10** という値を指定している。

正規分布の逆関数は、「母分散が既知の場合の母平均の検定」などでも使われます（推測統計編でお話しします）。なお、標準正規分布であれば、累積分布関数に対する逆関数の値を求めるのに NORM.S.INV 関数が使えます（平均と標準偏差を指定する必要がないので簡単です）。

◇ ◇ ◇ ◇ ◇ ◇ ◇

さて、今回は正規分布の基本的な考え方や確率密度関数と累積分布関数の求め方、また、よく使われる理由や逆関数の求め方などについてお話ししました。正規分布は統計学のさまざまな場面で、これでもかというほど登場します。そういった例については、今後、関連のある箇所で少しずつ触れていくこととして、次回は、カイ二乗分布についてお話しします。

カイ二乗分布は、標準得点を 2 乗和した値の分布で、理論的な値からのズレを求めたりするのに使われます。また、分散の比を表す F 分布や、平均からどれだけ離れているかを表すのに使われる t 分布とも深く関わっています。というわけで、次回もお楽しみに！

この記事で取り上げた関数の形式

関数の利用例については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

正規分布の確率密度関数や累積分布関数の値を求めるための関数

NORM.DIST 関数：正規分布の確率密度関数や累積分布関数の値を求める

形式

NORM.DIST(x, 平均値 , 標準偏差 , 関数形式)

引数

- **x**：確率変数の値を指定する。
- **平均値**：母集団の平均値μを指定する。
- **標準偏差**：母集団の標準偏差σを指定する。
- **関数形式**：以下の値を指定する。
  - **FALSE** …… 確率密度関数の値を求める
  - **TRUE** …… 累積分布関数の値を求める

NORM.S.DIST 関数：標準正規分布の確率密度関数や累積分布関数の値を求める

形式

NORM.S.DIST(x, 関数形式)

引数

- **x**：確率変数の値を指定する。
- **関数形式**：以下の値を指定する。
  - **FALSE** …… 確率密度関数の値を求める
  - **TRUE** …… 累積分布関数の値を求める

備考

※ Google スプレッドシートの NORM.S.DIST 関数には［関数形式］の引数が指定できません（累積分布関数の値のみが求められます）。

正規分布の累積分布関数に対する逆関数の値を求めるための関数

NORM.INV 関数：正規分布の累積分布関数に対する逆関数の値を求める

形式

NORM.INV(累積確率 , 平均値 , 標準偏差)

引数

- **累積確率**：累積分布関数の値を指定する。
- **平均値**：母集団の平均値μを指定する。
- **標準偏差**：母集団の標準偏差 sigma を指定する。

NORM.S.INV 関数：標準正規分布の累積分布関数に対する逆関数の値を求める

形式

NORM.S.INV(累積確率)

引数

累積確率：累積分布関数の値を指定する。

標準化のための関数

STANDARDIZE 関数：値を標準化する（標準得点を求める）

形式

STANDARDIZE(x, 平均値 , 標準偏差)

引数

- **x**：標準化したい値を指定する。
- **平均値**：母集団の平均値μを指定する。
- **標準偏差**：母集団の標準偏差σを指定する。

［データ分析］カイ二乗分布
～ ポテトチップスの内容量のばらつきは改善されたか？

データ分析の初歩から学んでいく連載（確率分布編）の第 7 回。カイ二乗分布は標準得点の二乗和の分布です。標準得点とは何か、二乗することはどういう意味を持つのか、といった基本的なところからカイ二乗分布の姿を明らかにしていきます。続けて、確率密度関数や累積分布関数の求め方や可視化の方法を解説し、利用例などを紹介します。

羽山博（2024 年 09 月 12 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の確率分布編、第 7 回です。前回は、連続型確率分布の代表とも言われる正規分布を取り上げました。今回はカイ二乗分布について、その特徴や意味を基本から解き明かし、確率密度関数／累積分布関数の求め方、利用例などを見ていきます。

ばらつきはどう分布するのか ～ カイ二乗分布につながるイメージ

前回お話しした**中心極限定理**を覚えているでしょうか。母集団がどのような分布であっても、そこから幾つかのサンプル（標本）を取り出して、そのサンプルの平均値を求めることを繰り返すと、それらの平均値の分布が正規分布に近づくというものでしたね。つまり、サンプルの平均値が正規分布に従っているということでした。

であれば、サンプルの「平均値」が正規分布に従うように、**サンプルの「分散」も何らか**の分布に従うのか知りたいですよね。答えから言うと**カイ二乗分布**に従うのですが、順を追って見ていきましょう（正確には、標準正規分布から得られたサンプルの二乗和が確率変数となり、それがカイ二乗分布に従います。そのことを確認するためにも一歩ずつ進めましょう）。

ポイントとなるのはカイ二乗分布の**確率変数はどのようなものであるか**ということと**カイ二乗分布の確率密度関数と累積分布関数はどのようなものであるか**という 2 点です。まず、確率変数からお話ししていきます。

統計学では、○○分布に「従う」という表現がよく出てきます。この「従う」という言葉は「規則に従う」といった意味の「従う」です。例えば「カイ二乗分布に従う」というのは、「確率変数の値が、カイ二乗分布する母集団から取り出されたものである」あるいは「確率変数の分布が理論的にはカイ二乗分布に当てはまるはずだ」といった意味です。なお、統計学では「従う」を ~ という記号で表します（後で登場します）。

具体的な問題で考えてみます。最初に何を知りたいかをイメージしておこうというわけです。あくまで架空の事例ですが、あるお菓子の内容量について、母集団の平均が **100g**、標準偏差が **0.5g** で、正規分布に従っていることが分かっているものとします。さて、このお菓子を n 個取り出して内容量を測定したとき、その分散（を基に得られる確率変数）はどう分布するでしょうか（図 1）。

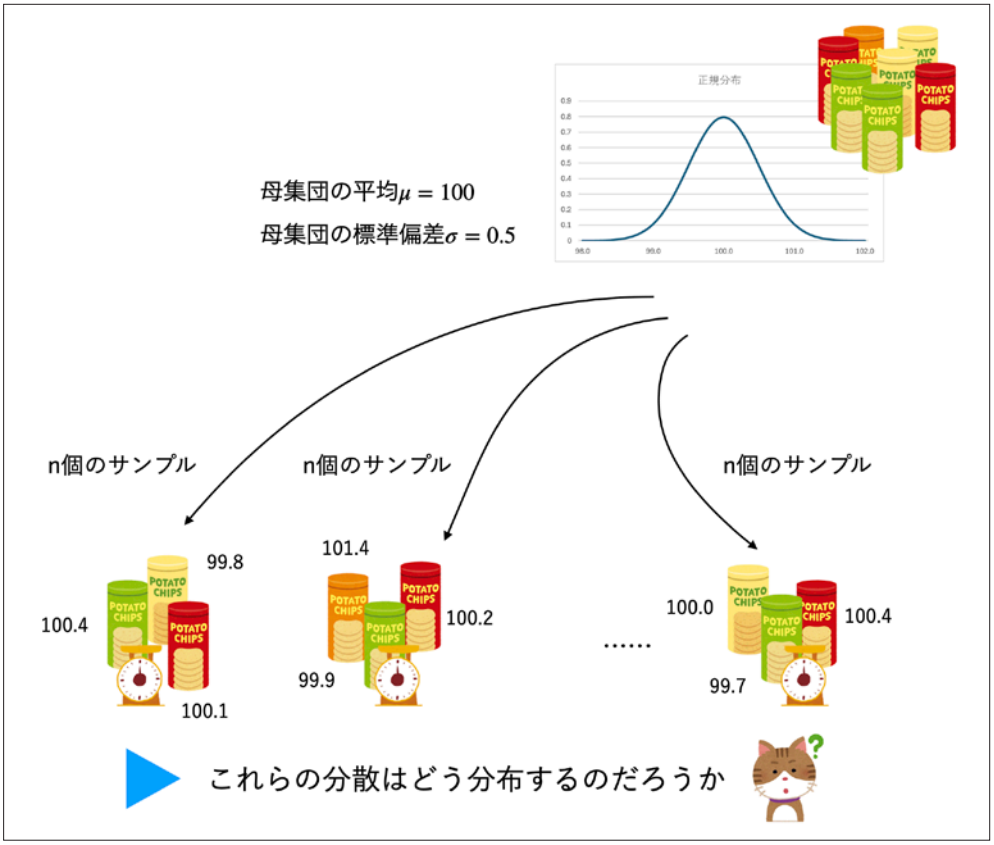


図 1　分散はどんな分布になるのかを考えてみよう

前回は、サンプルの平均は正規分布に従うことを解説した。では、分散はどのような分布に従うのだろうか。それを知るためにはカイ二乗分布の確率変数がどのようなものかを理解することから見ていく必要がある。

前述したように、カイ二乗分布の確率変数がどのようなものであるかについて理解を深めてからでないと答えが出ないので、少しずつお話しします。先にカイ二乗分布の確率密度関数や累積分布関数がどのようなものかを知りたい方はこちらをご覧ください。

カイ二乗分布はどこから現れるのか ～ カイ二乗分布の確率変数

では、カイ二乗分布の**確率変数に**ついて見ていきます。出発点は母集団の分布が正規分布であるという前提からです。

図 1 で見たように、正規分布する母集団から n 個のサンプルを独立に（ある試行が他の試行に影響しない取り出し方で）取り出します。確率変数を X とすると、それらは、$X_1, X_2, \ldots, X_n$ で表されます。小文字の x ではなく、大文字の X を使っているのは、個々の具体的な値を考えているわけではなく、「確率変数である」ということを一般的に表しているからです。

母平均が分かっている場合のカイ二乗値とカイ二乗分布

基準を統一して、各サンプルが平均からどれだけ離れているかを知るには、標準得点を求めればいいですね。以下のように平均値μを引いて、標準偏差σで割れば求められます。一歩ずつ確実に理解できるように、穴埋め問題にしておくので、何が入るかを考えながらゆっくり読み進めてみてください。

$$Z_i = \frac{X_i - \boxed{ア}}{\boxed{イ}} \qquad (1)$$

答え：ア＝　μ　、イ＝　σ

ここで、それぞれの値が平均からどれだけ離れているかを合計したいのですが、標準得点 Z_i は正になる場合もあれば、負になる場合もあるので、そのまま合計すると正と負が相殺されてしまいます。そこで、全てを正にするために**二乗和**を求めることにしましょう。つまり、**二乗して合計する**わけです（いつもの手ですね）。

このようにして求められた値を**カイ二乗値**と呼び、χ^2 と表します。つまり、以下のようになります。Σの部分については、2 行目以降、式を見やすくするために i の値の範囲を省略して表記してあります。

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{n} Z_i^2 \\ &= \sum \left(\frac{X_i - \boxed{ア}}{\boxed{イ}} \right)^2 \\ &= \frac{\sum (X_i - \mu)^2}{\sigma^2}\end{aligned} \qquad (2)$$

答え：ア＝　μ　、イ＝　σ

この**χ^2 という確率変数は自由度 n のカイ二乗分布$\chi_{(n)}^2$ に従います**（自由度とは単にサンプルの個数ということではなく、独立した情報の個数に当たるものです。後で詳しく解説します）。確率変数と分布の名前に同じ文字を使っているので同語反復っぽくなりますが、（2）式の左辺を別の文字で表すと分かりやすくなります。例えば、Y で表すと、

$$Y = \frac{\sum (X_i - \mu)^2}{\sigma^2}$$

となります。この確率変数 Y が自由度 n のカイ二乗分布に従うということですね。統計学では「従う」を ~ という記号で表すので、このことを数式で表すと以下のようになります。左辺がカイ二乗値（確率変数）、右辺の$\chi_{(n)}^2$ が自由度 n のカイ二乗分布です。

$$Y \sim \chi_{(n)}^2$$

なお、標準正規分布では$\mu = 0$、$\sigma = 1$なので、Xがそのまま標準得点になります。従って、カイ二乗値χ^2は以下の式で求められます。（2）式に$\mu = 0$、$\sigma = 1$を代入して確認してみましょう。

$$\chi^2 = \frac{\sum (X_i - \mu)^2}{\sigma^2} = \frac{\sum \left(X_i - \boxed{ア}\right)^2}{\boxed{イ}^2} = \sum {X_i}^2 \qquad (3)$$

答え：ア＝　0　、イ＝　1

標準正規分布から得られたサンプルの二乗和が確率変数となっています。そして、それが自由度nのカイ二乗分布に従います。これで、最初のお話とつながりましたね。

しつこいようですが、（2）式や（3）式で求められる**カイ二乗値χ^2は確率変数です**。つまり、確率密度関数や累積分布関数の**横軸**に当たるものです。要するに、カイ二乗分布に従う確率変数なので、**カイ二乗値**と呼ばれるということです。カイ二乗値に対する確率密度関数の値や累積分布関数の値、つまり縦軸の値は、後で登場するExcelのCHISQ.DIST関数で求められます。

カイ二乗分布の確率密度関数や累積分布関数がどのようなものになるかは、次の項でお話しします。もう少しだけ確率変数（カイ二乗値）のお話をしておきます。

母平均が分からない場合のカイ二乗値とカイ二乗分布

ところで、上の例は

$$\text{母平均}\mu$$

があらかじめ分かっている場合のお話です。母平均μが分からない場合には、

$$\text{サンプルの平均}\bar{X}$$

を代わりに使ってカイ二乗値χ^2を求めます。その場合は自由度がn − 1となります。つまり、カイ二乗値は、

$$\chi^2 = \frac{\sum \left(X_i - \bar{X}\right)^2}{\sigma^2} \qquad (4)$$

となります。**母平均が分からない場合は、（4）式で求められるカイ二乗値χ^2が、自由度n − 1のカイ二乗分布$\chi^2_{(n-1)}$に従います。**

自由度っていったい何？　何が自由なの？

さて、ここでカイ二乗分布の自由度として n を使ったり n − 1 を使ったりするのを、腑（ふ）に落ちないと感じた人も多いと思います。その前にそもそも自由度とはどのような値なのかが謎ですね。

端的に言うと、**自由度**とは独立した情報（変数）の個数のことです。では、独立した情報というのは何なのでしょうか。具体的に見ていきます。

（4）式では、

$$\text{サンプルの平均値}\bar{X}$$

は、n 個のデータ X_1 ～ X_n を基に求められます。つまり、

$$\bar{X}$$

を求めるために、それぞれのデータから 1/n 個ずつの情報をもらっています。

$$(X_1 - \bar{X})$$

の情報の個数は (1 − 1/n) 個、

$$(X_2 - \bar{X})$$

の情報の個数も (1 − 1/n) 個……ということですね。これが n 個あるので情報の個数は全部で (1 − 1/n)× n = n − 1 個になるというわけです。

$$\sum(X_i - \bar{X})$$

の部分には、見た目では n 個の変数がありますが、独立した情報は n − 1 個しかないということになります。

自由度は、「自由に動かせる変数の個数」と説明されることもあります。具体的な数値で考えてみます。

例えば、10, 12, 17 という 3 個のサンプルから平均値を求めると、

$$(10 + 12 + 17)/3 = 13$$

です。平均が 13 のとき、10 と 12 という 2 個の値を決めると、17 という値は自動的に決まってしまいます。ここで仮に 10 と 12 を 11 と 8 に変えたとしても、

$$(11 + 8 + \boxed{?})/3 = 13$$

の「?」の部分は、やはり自動的に **20** に決まります。つまり、自由に動かせるのは 3 − 1 = 2 個となります。

一方、あらかじめ母平均μが分かっている場合は、μはサンプルから求めた値ではなく、それぞれのサンプルとは別に決まっている値なので、（2）式の$\Sigma (X_i - \mu)$の部分には n 個の独立した情報があると考えられます。

自由度はカイ二乗分布だけでなく、次回解説する t 分布や F 分布でも登場します。それぞれの手法や分布での自由度の考え方はなかなか理解しづらいものですが、実用的には「こういう手法や分布の場合には、自由度はいくら（例：$n - 1$）とする」といったルールで覚えておくのが、最も悩まなくて済む付き合い方ではあります。

カイ二乗分布の確率変数と自由度のお話はこれぐらいにして、カイ二乗分布の確率密度関数と累積分布関数がどのようなものであるかを可視化してみましょう。

カイ二乗分布ってどんな感じの分布（1）～ 確率密度関数を可視化してみよう

ここまでは、カイ二乗分布の確率変数であるカイ二乗値χ^2をどのようにして求めるかというお話をしてきました。では、カイ二乗値分布では確率変数がどのように分布するのでしょうか。つまり、確率密度関数と累積分布関数はどのようなものになるのでしょうか。

実は、カイ二乗分布の確率分布関数 f(x; k) の値と累積分布関数 F(x; k) の値は以下の式で求められます。しかし、これらの式を覚える必要は全くありません。詳細については最後のコラムでお話しするので、ここでは式を掲載するだけにとどめます。以下の式は軽くスルーしてもらってけっこうです。

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

$$F(x;k) = \frac{\gamma(k/2, x/2)}{\Gamma(k/2)}$$

上の式を覚えなくても、Excel の CHISQ.DIST 関数を使えば、確率密度関数の値と累積分布関数の値が簡単に求められます。**カイ二乗分布の母数は自由度のみです。**つまり、自由度が決まれば、カイ二乗分布の形も決まります。CHISQ.DIST 関数の形式（図 2）を見てから、確率密度関数の値を求め、カイ二乗分布を可視化しましょう（累積分布関数については次の項で取り扱います）。

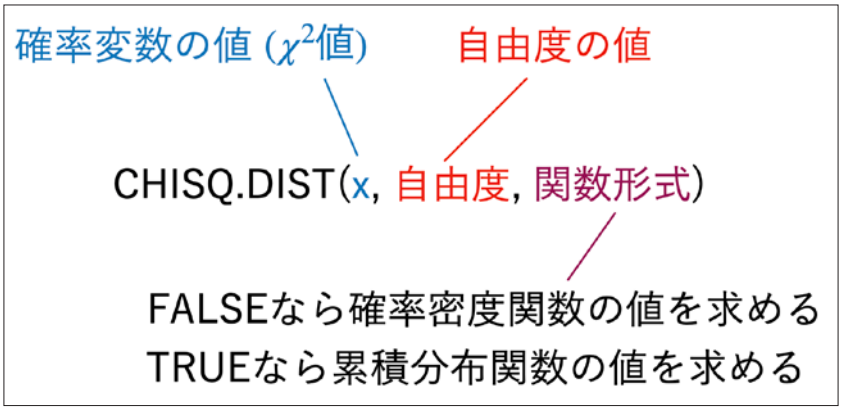
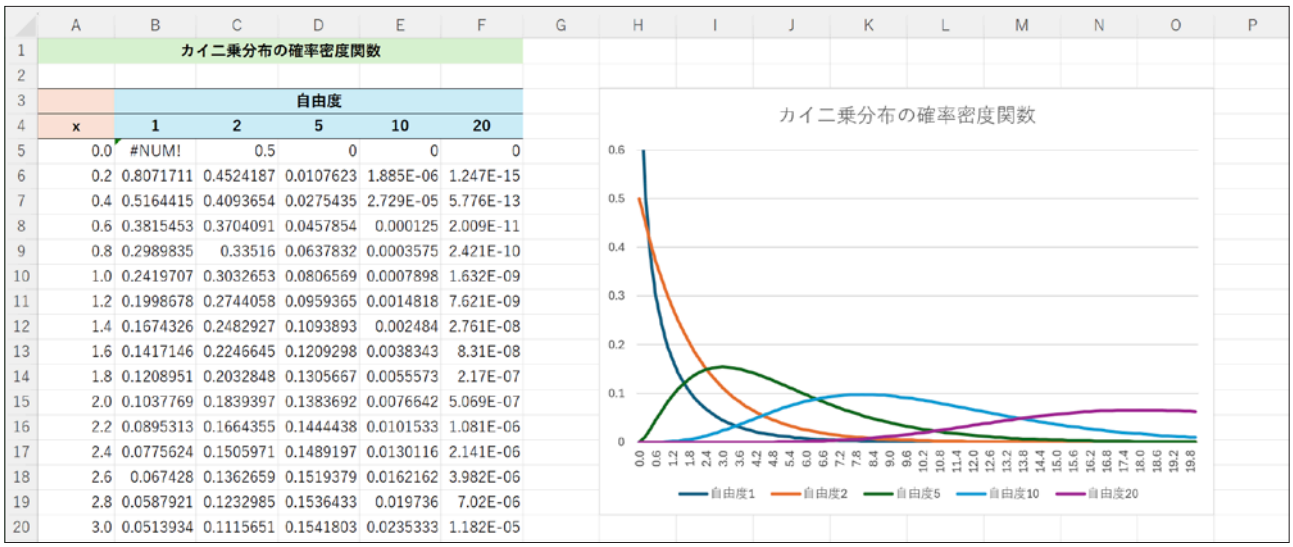


図 2　CHISQ.DIST 関数に指定する引数

CHISQ.DIST 関数には、確率変数の値（χ^2 値）と自由度を指定する。関数形式についてはこれまで見てきた関数と同様、**FALSE** を指定すれば確率密度関数の値が、**TRUE** を指定すれば累積分布関数の値が求められる。

以下に、幾つかの自由度に対する確率密度関数の値を求め、それらのグラフを描いてみます（図 3）。作成の手順は図の後に記しておきます。



カイ二乗分布の確率密度関数

x	自由度 1	2	5	10	20
0.0	#NUM!	0.5	0	0	0
0.2	0.8071711	0.4524187	0.0107623	1.885E-06	1.247E-15
0.4	0.5164415	0.4093654	0.0275435	2.729E-05	5.776E-13
0.6	0.3815453	0.3704091	0.0457854	0.000125	2.009E-11
0.8	0.2989835	0.33516	0.0637832	0.0003575	2.421E-10
1.0	0.2419707	0.3032653	0.0806569	0.0007898	1.632E-09
1.2	0.1998678	0.2744058	0.0959365	0.0014818	7.621E-09
1.4	0.1674326	0.2482927	0.1093893	0.002484	2.761E-08
1.6	0.1417146	0.2246645	0.1209298	0.0038343	8.31E-08
1.8	0.1208951	0.2032848	0.1305667	0.0055573	2.17E-07
2.0	0.1037769	0.1839397	0.1383692	0.0076642	5.069E-07
2.2	0.0895313	0.1664355	0.1444438	0.0101533	1.081E-06
2.4	0.0775624	0.1505971	0.1489197	0.0130116	2.141E-06
2.6	0.067428	0.1362659	0.1519379	0.0162162	3.982E-06
2.8	0.0587921	0.1232985	0.1536433	0.019736	7.02E-06
3.0	0.0513934	0.1115651	0.1541803	0.0235333	1.182E-05

図 3　カイ二乗分布の確率密度関数の例

自由度 1, 2, 5, 10, 20 について、x = 0.0 ～ 20.0 までの確率密度関数の値を求め、グラフを描いてみた。自由度が **1** のとき、x = 0 に対する確率密度関数の値は求められないので、セル **B5** にはエラー値が表示されている。そのため、グラフはセル **B5** を除外して描いている。作成の手順は後の箇条書きを参照。

カイ二乗分布の台（確率変数が取り得る値の範囲）は 0 ～∞です。そもそも、二乗した値の合計なので負になることはありませんね。また、自由度 k が大きくなると、x = k の近くで確率密度関数のピーク（山の頂点）が見られるようになり、左右に裾野が広がる形になります。例えば、自由度が **10** の場合、グラフの横軸が x = 10 となる位置の近くに山のある分布となります（といっても、自由度をもっと大きくしないと x = k の位置がピークにならないのですが、雰囲気はつかめると思います）。実は、**カイ二乗分布は、自由度 k を大きくしていくと、**

$$\text{平均}k\text{、標準偏差}\sqrt{2k}\text{の正規分布}$$

に近づきます（近づき方はそれほど早くないので、k の値をかなり大きくしないと、正規分布のグラフには重なりません）。

→目次に戻る

ここでは、自由度を表すのに k という文字を使いましたが、これは、離散型確率分布の確率変数の値を表すのに使われる k とは全く別のものです（念のため）。なお、自由度を表す文字としては、ギリシア文字の ν（ニュー）などを使うこともあります。ν はラテン文字（アルファベット）の n に当たる文字です。自由度は degree of freedom を略して df と表記されることもあります。

確率密度関数の値を求めるための手順は以下の通りです。可視化については単に折れ線グラフを描くだけなので（前回までですでに何度もやっているので）、今回からは関数の入力にのみ焦点を当てることとして、グラフ作成などの手順についてはサンプルファイル内に掲載しておくこととします。

サンプルファイルをこちらからダウンロードし、［カイ二乗分布］ワークシートを開いて試してみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。

◆ Excel での操作方法

- セル **B5** に =CHISQ.DIST(A5:A105,B4:F4,FALSE) と入力する
  - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **B5** ～ **F105**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す

◆ Google スプレッドシートでの操作方法

- 5 行目のみ手作業で数値を入力する
  - セル **B5** に「#NUM!」というエラー値を入力しておく
  - セル **C5** に **0.5** を入力しておく
  - セル **D5** ～ **F5** に **0** を入力しておく
- セル **B6** に「=ARRAYFORMULA(CHISQ.DIST(A6:A105,B4:F4,FALSE))」と入力する

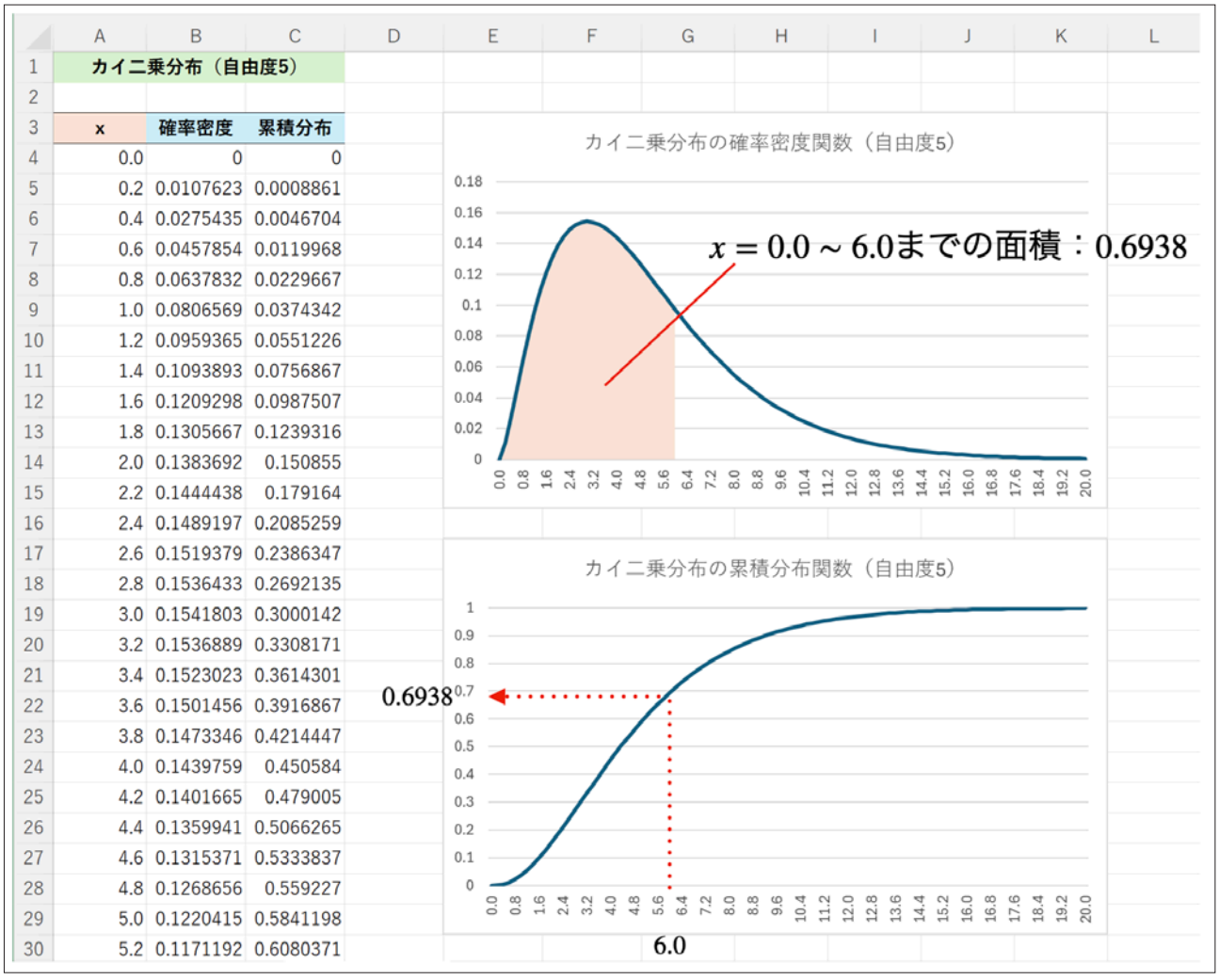
原因は不明ですが、原稿の執筆時点では **x=0.0**, 自由度 **=2** のとき、Google スプレッドシートの CHIDSQ.DIST 関数では確率密度関数の値（**0.5**）が正しく求められず、エラーとなります。上の手順で 5 行目だけを手作業で入力して 6 行目以降を CHISQ.DIST 関数で求めたのはそのためです。

● グラフの作成方法

- サンプルファイル内に掲載しておきます（タイトルや軸の書式などの細かい設定は省略）

カイ二乗分布ってどんな感じの分布（2）～ 累積分布関数を可視化してみよう

続いて、累積分布関数です。こちらは、自由度 **5** の例だけを見ておきます（図 4）。確率密度関数でグラフと x 軸で囲まれた範囲の面積が累積分布関数の値になることを示すために、確率密度関数も併せて作成し、説明をグラフ上に書き加えてあります。［カイ二乗累積分布］ワークシートを開き、図の後に記した手順で試してみてください。グラフ作成などの手順についてはサンプルファイル内に掲載しておくこととします。



x	確率密度	累積分布
0.0	0	0
0.2	0.0107623	0.0008861
0.4	0.0275435	0.0046704
0.6	0.0457854	0.0119968
0.8	0.0637832	0.0229667
1.0	0.0806569	0.0374342
1.2	0.0959365	0.0551226
1.4	0.1093893	0.0756867
1.6	0.1209298	0.0987507
1.8	0.1305667	0.1239316
2.0	0.1383692	0.150855
2.2	0.1444438	0.179164
2.4	0.1489197	0.2085259
2.6	0.1519379	0.2386347
2.8	0.1536433	0.2692135
3.0	0.1541803	0.3000142
3.2	0.1536889	0.3308171
3.4	0.1523023	0.3614301
3.6	0.1501456	0.3916867
3.8	0.1473346	0.4214447
4.0	0.1439759	0.450584
4.2	0.1401665	0.479005
4.4	0.1359941	0.5066265
4.6	0.1315371	0.5333837
4.8	0.1268656	0.559227
5.0	0.1220415	0.5841198
5.2	0.1171192	0.6080371

図 4　カイ二乗分布の累積分布関数の例

上のグラフが確率密度関数のグラフ。グラフと x 軸で囲まれた範囲の面積が累積分布関数の値になる。下のグラフは、その面積、つまり累積分布関数の値をプロットしたもの。例えば、x = 6.0 に対する累積分布関数の値は **0.6938** となる。

→目次に戻る

確率密度関数の値を求める方法はすでに見た通りですが、x = 0.0 ～ 6.0 の範囲を塗りつぶして表示するために使うので、以下に併せて見ておきます。

◆ Excel での操作方法

- 確率密度関数の値を求める
  - ・セル **B4** に =CHISQ.DIST(A4:A104,5,FALSE) と入力する
    - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **B4** ～ **B104**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
- 累積分布関数の値を求める
  - ・セル **C4** に =CHISQ.DIST(A4:A104,5,TRUE) と入力する
    - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **C4** ～ **C104**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す

◆ Google スプレッドシートでの操作方法

- 確率密度関数の値を求める
  - ・セル **B4** に =ARRAYFORMULA(CHISQ.DIST(A4:A104,5,FALSE)) と入力する
- 累積分布関数の値を求める
  - ・セル **C4** に =ARRAYFORMULA(CHISQ.DIST(A4:A104,5,TRUE)) と入力する

● グラフの作成方法

- サンプルファイル内に掲載しておきます（タイトルや軸の書式などの細かい設定は省略）

コラム　カイ二乗分布の確率密度関数をシミュレーションで求める

すでにお話ししたように、カイ二乗値を求めるには、正規母集団（＝正規分布の母集団）から n 個のサンプルを取り出し、それらの標準得点の二乗和を求めます。ということは、その手順に従ってカイ二乗値を幾つも作り、ヒストグラムを描くとカイ二乗分布に近くなるはずですね。冒頭の例で示した平均 $\mu = 100$ と標準偏差 $\sigma = 0.5$ を使ってもいいのですが、結局のところ、標準化した値を使うので、最初から標準正規分布（$\mu = 0$, $\sigma = 1$ の正規分布）を使ってシミュレーションしてみましょう。

シミュレーションは Excel でもできますが、多数のセルを使う必要があり、ちょっと面倒です。一応、サンプルファイルに作成例は含めてありますが、特に解説はしません（［カイ二乗分布のシミュレーション］ワークシート）。そこで、Python のプログラムを使うことにします（リスト 1）。

サンプルプログラムはこちらから参照できます。リンクをクリックすれば、ブラウザが起動し、Google Colaboratory の画面が表示されます（Google アカウントでのログインが必要です）。最初のコードセルをクリックし、［Shift］+［Enter］キーを押してコードを実行してみてください。結果は図 5 のようになります。コードの詳細については解説しませんが、コメントとリスト 1 の説明を見れば何をやっているかが大体分かると思います。

```
# カイ二乗分布の確率密度関数のシミュレーション
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import chi2

# 標準正規分布（μ=0、σ=1）からランダムに 10 個のサンプルを取り出したものを 10000 個作り、それぞれ二乗する
x = np.random.randn(10000, 10)**2
chi2_sample = x.sum(axis=1)  # 各行を合計する。これがカイ二乗値（横軸）

# ヒストグラムを描く（カイ二乗値がどのように現れるかが分かる）
plt.hist(chi2_sample, bins=100, density=True)  # 階級は 100 個、縦軸は確率とする

# カイ二乗分布の確率密度関数を描く
x = np.linspace(0, 40, 100)  # 0 ～ 40 までを 100 個に分けた等差数列（横軸の値）
plt.plot(x, chi2.pdf(x, 10))  # x に対する確率密度関数の値をプロットする
plt.show()
```

リスト 1　カイ二乗分布の確率密度関数をシミュレーションする

標準正規分布（$\mu = 0$、$\sigma = 1$）からランダムに **10 個**の値を取り出すことを **10000 回**繰り返す。10000 行 ×10 列のデータが作られるので、それらの値を全て二乗する。各行を合計すれば、**10 個**のデータを合計したものが **10000 行**作られることになる。この **10000 行**のデータがカイ二乗値になる。続いて、縦軸を確率としてカイ二乗値のヒストグラムを描く。最後に、自由度 **10** のカイ二乗分布の確率密度関数を描けば、ヒストグラムとほぼ重なることが分かる。

リスト 1 の実行例が以下の図 5 です。乱数を使うので結果は毎回少しずつ異なりますが、ほとんど重なっていますね。

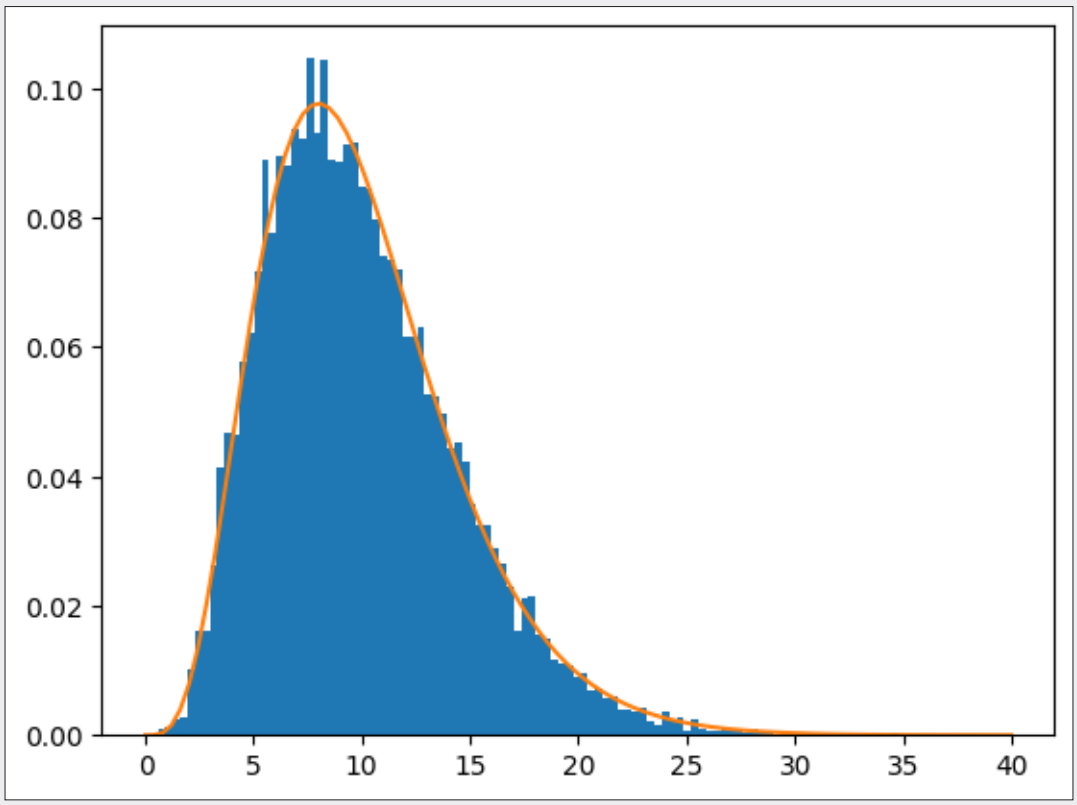


図 5　カイ二乗分布の確率密度関数のシミュレーション結果

棒グラフは、標準正規分布から **10 個**のサンプルを **10000 回**取り出して、それらの二乗和を基に作成したヒストグラム。
折れ線グラフは自由度 **10** のカイ二乗分布の確率密度関数。

カイ二乗分布の応用に向けて 〜 区間推定や検定のための準備

カイ二乗分布は母分散の区間推定や母分散の検定、いわゆるカイ二乗検定（適合度の検定や独立性の検定）などに使われます。詳細は推測統計編でお話ししますが、その準備として、少し発展的なお話をしておきましょう。カイ二乗分布のさらなる理解にもつながります。

母平均が既知の場合、カイ二乗値は以下の式で求められました。（2）式を再掲します。この場合は、定義通りに計算してカイ二乗値を求める必要があります。

$$\chi^2 = \frac{\sum (X_i - \mu)^2}{\sigma^2} \qquad (2)$$

一方、**母平均が未知の場合に**は、もう少し簡単に計算する方法があります。**母平均が未知の場合**、カイ二乗値は以下の式で求められました。（4）式を再掲します。

$$\chi^2 = \frac{\sum (X_i - \bar{X})^2}{\sigma^2} \qquad (4)$$

→目次に戻る

ここで、以下のような s^2 を考えます（これは Excel の VAR.S 関数で求められる不偏分散です）。

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} \qquad (5)$$

（5）式の両辺に $n-1$ を掛けて以下のように変形しましょう。

$$(n-1)s^2 = \sum (X_i - \bar{X})^2 \qquad (6)$$

（6）式の左辺を（4）式に代入すると、以下のようになります。

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \qquad (7)$$

このことから、母平均が未知の場合は、（4）式の通りに計算しなくても、VAR.S 関数によって求めた不偏分散に $n-1$ を掛け、母分散で割るだけで、カイ二乗値が求められることが分かります。このようにして求めたカイ二乗値が自由度 $n-1$ のカイ二乗分布に従うということですね。

ばらつきは小さくなったのか？ ～ 母平均が未知の場合の累積確率を求める

では、冒頭のお話の続きです。これまでの製品は、母平均が **100g**、母標準偏差が **0.5g** でした。そこで、内容量が少な過ぎたり多過ぎたりする製品（不良品）の発生を減らすために製造工程を見直したとします。つまり、製品のばらつきを小さくするために工程を見直したというわけです。工程の見直し後、**10 個**のサンプルを取り出して内容量を測定したところ、以下のような値になったとします。さて、内容量のばらつきは改善されたでしょうか。

100.1, 100.0, 99.8, 100.0, 99.7, 99.8, 99.9, 100.4, 100.3, 100.2

工程の変更によって原料の配分や水分の量なども変わった可能性があるので、母平均も変わったかもしれません。従って、**母平均は既知ではなく、未知であるものと考える**ことにします（実際のところ、母平均が分かっている場合よりも、分からない場合の方が多いですね）。

まず、これらの値を基に**不偏標準偏差（母標準偏差の推定値）**を求めてみましょう。［応用例 1（母平均未知）］ワークシートを開き、セル **A18** に =STDEV.S(A6:A15) と入力してください（図 6）。不偏標準偏差を求めるための STDEV.S 関数や不偏分散（母分散の推定値）を求めるための VAR.S 関数については、「やさしいデータ分析］分散／標準偏差～給与の格差ってどれぐらい？」で解説したので、そちらもぜひご参照ください。

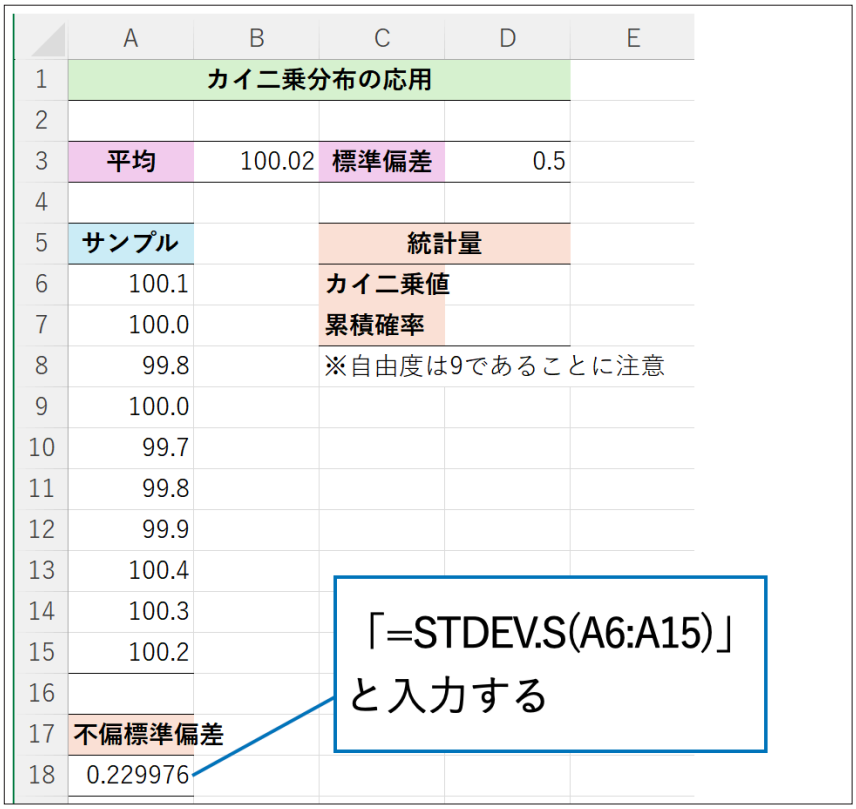


図 6　不偏標準偏差を求めて、母集団の標準偏差を推定する

母集団のばらつきを推定するために、不偏標準偏差を求めるための STDEV.S 関数を使う。セル **A18** に =STDEV.S(A6:A15) と入力すれば、**0.229976（≈ 0.230）**となる。これまでの **0.5** という値よりも小さくなっているように思われるが、本当に小さくなったのだろうか。

不偏標準偏差は **0.230g**（不偏分散はその二乗なので **0.053**）となりました。これまでの標準偏差が **0.5g** だったので、ばらつきは小さくなったように思われます。……が、そう判断していいのでしょうか。そこで、母標準偏差が **0.5g**（母分散は **0.25**）である母集団からサンプルを取り出したときに、上のような値（セル **A6** ～ **A15** の値）が現れる確率を求めてみましょう。そのために、カイ二乗値とカイ二乗分布の累積分布関数の値を求めてみます（あまりにも確率が小さいようであれば、母標準偏差は **0.5g** とは考えにくい、つまり母標準偏差が小さくなった、と言えますね）。

この場合、カイ二乗値は（4）式に従って計算するよりも、（7）式を使った方が簡単です。（7）式を再掲します。

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \qquad (7)$$

［応用例 1（母平均未知）］ワークシートを続けて使います。図 7 の手順に沿って計算してみましょう。s^2 は VAR.S 関数で求めます。n − 1 の値は **10-1=9** で、σの値はセル **D3** に入力されている標準偏差を二乗すれば求められますね。カイ二乗値が求められたら、それを基に CHISQ.DIST 関数を使ってカイ二乗分布の累積分布関数の値（累積確率）を求めます。

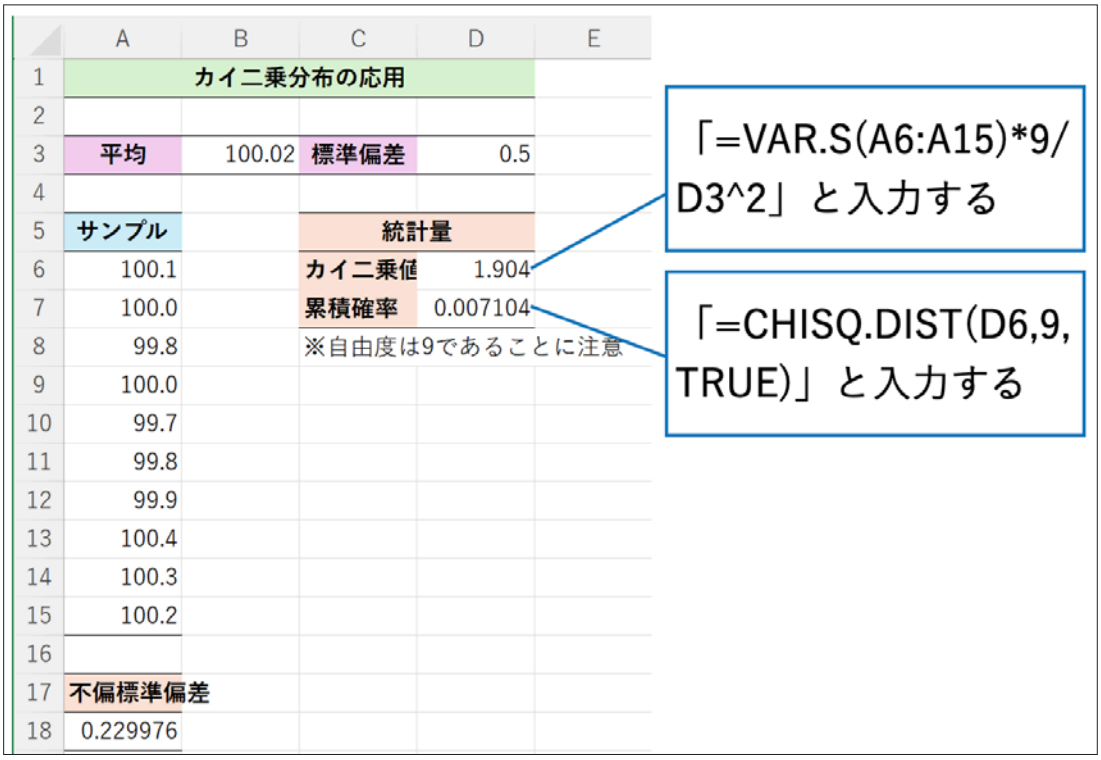


図 7　新しい製造工程での分散は母分散と等しいか（母未知が既知の場合）

（7）式に従って、セル **D6** に =VAR.S(A6:A15)*9/D3^2 と入力するだけで、カイ二乗値が求められる。この場合、自由度は **9** なので、セル **D7** に =CHISQ.DIST(D6,9,TRUE) を入力してカイ二乗分布の累積分布関数の値を求める。

カイ二乗値は **1.904** となり、それに対するカイ二乗分布の累積分布関数の値は **0.0071**（0.71%）になります。この結果を見ると、かなり「まれ」なことが起こったことが分かります。母標準偏差が **0.5**（母分散が **0.25**）だとすると、10 個のサンプルを取り出して、それらの値から求めた不偏標準偏差が **0.230** である（不偏分散が **0.053**）であるというのはそうそう起こらないことです。ということは、製造工程を見直した場合の母分散は **0.25** よりも小さくなったと考えられます。つまり、内容量のばらつきが小さくなったものと考えられます。

上で求めた累積確率を可視化した図（図 8）も［応用例 1（母平均未知完成例)］ワークシートに含めてあります。こちらについてはワークシート内に作成方法を記載しておきます。

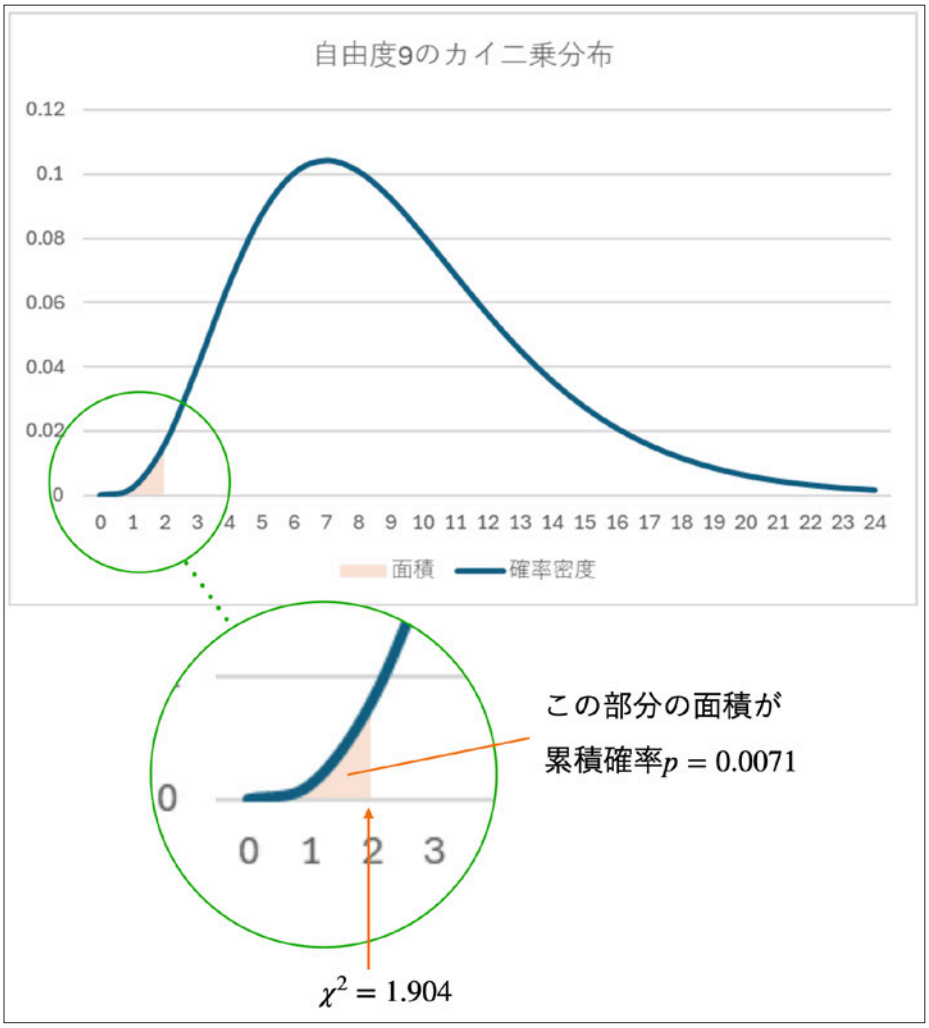


図 8　カイ二乗値 1.904 に対する累積確率

オレンジ色で塗りつぶした部分の面積が累積確率（**0.0071** = 0.71%）。分散が小さくなるとカイ二乗値も小さくなるが、カイ二乗値が **1.904** になるのはかなり「まれ」なこと（確率変数の値がグラフの端の方に位置していて、面積が小さい）。新しい製造工程では内容量のばらつきが小さくなったものと考えられる。

実は、この計算は母分散の検定のための計算にほかなりません。検定の考え方や詳細な方法については、推測統計編でのお話となるので、ここでは、これ以上のお話はしません。帰無仮説や対立仮説などの考え方を十分に知った上で使う必要があるので､この時点で「結論はこうだ！」と言い切るのは少し待ってください。もったいぶるようで申し訳ないのですが、母分散の検定については（さらには適合度の検定、独立性の検定などについても)、推測統計編でお話しします。

なお、(7）式は、F 分布や t 分布にもつながる式です。今回は、カイ二乗分布の確率変数とはどのようなものか、それに対する確率密度関数や累積分布関数の値はどのようにして求めるか、を理解していただければと思います。

逆に、ばらつきが大きくなってしまった場合には、どのような値になるでしょうか。例えば、サンプルの値が以下のようになっていたとします。

100.9, 100.1, 100.2, 100.0, 98.4, 99.6, 99.5, 100.3, 100.2, 101.0

こちらは［応用例 2（分散が大きい場合）］ワークシートに作成してあります。計算方法は全く同じなので、結果だけを記すと、カイ二乗値は **19.824** となり、その値に対するカイ二乗分布の累積確率は **0.98097**（＝ **98.10%**）となります。この確率は下位からの確率なので、上位からだと **0.01903**（**1.90%**）ですね。このようにカイ二乗値が大きな値になることも、やはり「まれ」なことなので、母集団の分散は **0.25** よりも大きいと考えられます。

なお、上位からの累積確率（右側確率または上側確率と呼ばれます）を求めるには **1** から CHISQ.DIST 関数で求めた値を引いてもいいですが、CHISQ.DIST.RT 関数を使うと簡単です。こちらは、セル **D10** に入力した =CHISQ.DIST.RT(D6,9) で求められます（カイ二乗値と自由度を指定するだけです。累積確率を求めることが分かっているので関数形式の指定は不要です）。

ところで、母平均が既知の場合はどうなるでしょうか。つまり、母平均が **100g** であると分かっている場合です（現実には、そういう場合は少ないですが）。その場合は（2）式に従って計算します。サンプルファイルの［応用例 3（母分散既知完成例）］に作成方法と併せて作成例を掲載しています。

カイ二乗分布の累積分布関数に対する逆関数は？

CHISQ.INV 関数や CHISQ.INV.RT 関数を使えば、カイ二乗分布の累積分布関数に対する逆関数の値が求められます。CHISQ.INV 関数には左側（下側）確率と自由度を指定し、CHISQ.INV.RT 関数には右側（上側）確率と自由度を指定します。

例えば、自由度 **10** のカイ二乗分布で、累積確率が **95%**のときの χ^2 値は =CHISQ.INV(95%, 10) または =CHISQ.INV.RT(5%, 10) で求められます。空いているセルにこれらの式を入力してみてください。いずれも **18.307** という値が求められます。これらの例は［カイ二乗分布の逆関数］ワークシートに入力されています。

カイ二乗分布の累積分布関数に対する逆関数の値は、母分散の区間推定などに使われます。詳細は推測統計編でお話しします。

コラム　カイ二乗分布の確率密度関数と累積分布関数を数式で表す

カイ二乗分布の確率密度関数 f(x; k) と累積分布関数 F(x; k) は以下の式で表されます。ここでは、自由度 k を ; で区切って示してあります。例によって、これらの式を無理に覚える必要はありません。上で見たように、いずれも Excel の CHISQ.DIST 関数で答えが求められます。

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} \qquad (8)$$

$$F(x;k) = \frac{\gamma(k/2, x/2)}{\Gamma(k/2)} \qquad (9)$$

ただし、

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$$

$$\gamma(a,x) = \int_0^x t^{a-1} e^{-t} dt$$

Γ は**ガンマ関数**と呼ばれるもので、階乗（n!）の考え方を複素数にまで拡張したものです。γ は**下側不完全ガンマ関数**または**第一種不完全ガンマ関数**と呼ばれるもの（ガンマ関数の定義域である 0 ～∞のうち、下部の 0 ～ x に対する値を求める関数）ですが、数学的な話にはこれ以上深入りしないことにします。

Excel を使って上の定義通りに計算することもできます。ガンマ関数の値については、Excel の GAMMA 関数を使えば求められるので、やや複雑になりますが（8）式を使って確率密度関数の値を求めることができます。一方の累積分布関数については、Excel に不完全ガンマ関数がないので（9）式の定義通りに計算するのはちょっと面倒です（積分の近似値を求める必要があります）。いずれも、サンプルファイルに参考として含めておきますが、ここは Python のプログラムを使った方が簡単なので、その例を紹介します。

（8）式や（9）式に登場するガンマ関数の値を求めるには、scipy.special モジュールの gamma 関数を使い、下側不完全ガンマ関数の値を求めるには scipy.special モジュールの gammainc 関数を使います。実は、gammainc 関数では、正則化された（ガンマ関数で割り算した）下側不完全ガンマ関数の値が求められます。つまり、（9）式の分子の部分ではなく、（9）式そのものの値が求められます。というわけで、**scipy.special モジュールの gammainc 関数の結果はカイ二乗分布の累積分布関数の値と一致します。**

サンプルプログラムはコラムで紹介したシミュレーションの続きに入力されています。リンクをクリックすれば、ブラウザが起動し、Google Colaboratory の画面が表示されます（Google アカウントでのログインが必要です）。2 番目以降のコードセルをクリックし、［Shift］+［Enter］キーを押してコードを実行してみてください。コードの詳細については解説しませんが、コメントとリスト 2 の説明を見れば何をやっているかが大体分かると思います。

```
# 定義に従って、カイ二乗分布の確率密度関数の値を求める
from scipy.special import gamma
import numpy as np

k = 5  # 自由度
x = 3  # 確率変数の値
print(1/(2**(k/2)*gamma(k/2)) * x**(k/2-1) * np.exp(-x/2))  # 定義通りに計算

# 出力例：
# 0.15418032980376925

# 検算
from scipy.stats import chi2

print(chi2.pdf(x, k))  # カイ二乗分布の確率密度関数

# 出力例：
# 0.15418032980376925

# 定義に従って、カイ二乗分布の累積分布関数の値を求める
from scipy.special import gammainc

print(gammainc(k/2, x/2))  # 正則化された下側不完全ガンマ関数を使って定義通りに計算

# 出力例：
# 0.3000141641213724

# 検算
from scipy.stats import chi2

print(chi2.cdf(x, k))  # カイ二乗分布の累積分布関数

# 出力例：
# 0.3000141641213724
```

リスト 2　定義に従ってカイ二乗分布の確率密度関数と累積分布関数の値を求める

ここでは、自由度 k = 5、確率変数の値 x = 3 で計算してみた。確率密度関数については、(9) 式をそのまま書けばよい。累積分布関数については、scipy.special モジュールの gammainc 関数を使って正則化された下側不完全ガンマ関数を求めるとよい。その結果が累積分布関数の値と一致する。

◇ ◇ ◇ ◇ ◇ ◇ ◇

今回はカイ二乗分布での確率変数である χ^2 値の意味やその求め方、さらに、カイ二乗分布の確率密度関数と累積分布関数の求め方などについてお話ししました。推測統計編の内容についても、かなり踏み込んでお話ししてしまいました。カイ二乗分布は、分散の比を表す F 分布や、平均値の推定などに使われる t 分布とも深く関わっているので、ここでは、カイ二乗分布の確率変数と確率密度関数、累積分布関数についての理解を深めていただければ十分かと思います。

というわけで、次回は t 分布についてお話しします。次回もお楽しみに！

この記事で取り上げた関数の形式

関数の利用例については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

カイ二乗分布の確率密度関数や累積分布関数の値を求めるための関数

CHISQ.DIST 関数：カイ二乗分布の確率密度関数や累積分布関数の値を求める

形式

CHISQ.DIST(x, 自由度 , 関数形式)

引数

- **x**：確率変数の値（カイ二乗値）を指定する。
- **自由度**：自由度を指定する。
- **関数形式**：以下の値を指定する。
  - **FALSE** …… 確率密度関数の値を求める
  - **TRUE** …… 累積分布関数の値を求める

備考

※ 累積分布関数の値は**左側確率**（または**下側確率**）とも呼ばれます。

CHISQ.DIST.RT 関数：カイ二乗分布の右側確率の値を求める

形式

CHISQ.DIST.RT(x, 自由度)

引数

x：確率変数の値（カイ二乗値）を指定する。

自由度：自由度を指定する。

備考

※ **1** から、CHISQ.DIST 関数で求められる累積分布関数の値を引いた値が返されます。

カイ二乗分布の累積分布関数に対する逆関数の値を求めるための関数

CHISQ.INV 関数：カイ二乗分布の累積分布関数に対する逆関数の値を求める

形式

CHISQ.INV(累積確率 , 自由度)

引数

- **累積確率：** 累積分布関数の値を指定する。
- **自由度：** 自由度を指定する。

備考

※累積確率には左側確率（下側確率）の値を指定します。

CHISQ.INV.RT 関数：カイ二乗分布の右側確率に対する逆関数の値を求める

形式

CHISQ.INV.RT(右側確率 , 自由度)

引数

- **右側確率：** 累積分布関数の右側確率の値を指定する。
- **自由度：** 自由度を指定する。

備考

※右側確率は 1 －左側確率です。

［データ分析］t 分布
～ 自動車の平均燃費は改善されたか？

データ分析の初歩から学んでいく連載（確率分布編）の第 8 回。t 分布は母分散が分からない場合の平均値に関連する分布です。中心極限定理を出発点とし、正規分布と比較しながら t 分布の姿を明らかにしていきます。続けて、確率密度関数や累積分布関数の求め方や可視化の方法を解説し、利用例などを紹介します。

羽山博（2024 年 10 月 03 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の確率分布編、第 8 回です。前回は、分散に関連する分布として、カイ二乗分布を取り上げました。今回は平均値に関連する t 分布について、その特徴や意味を基本から解き明かし、確率密度関数／累積分布関数の求め方、利用例などを見ていきます。

クルマの燃費は改善されたか ～ t 分布の利用

ある自動車の燃費（**燃料 1 リットル**で走行可能な距離）が、**平均 20km** であったとします。エンジンなどの改良の結果、燃費が改善されたかどうかを知りたい、というのが今回の問題意識です（図 1）。類似の事例としては、売り上げが増えたかどうかを知りたい、機器の操作ミスが減少したかどうかを知りたい、固定資産の使用期間が長くなったかどうかを知りたい……などなど、枚挙にいとまがありません。

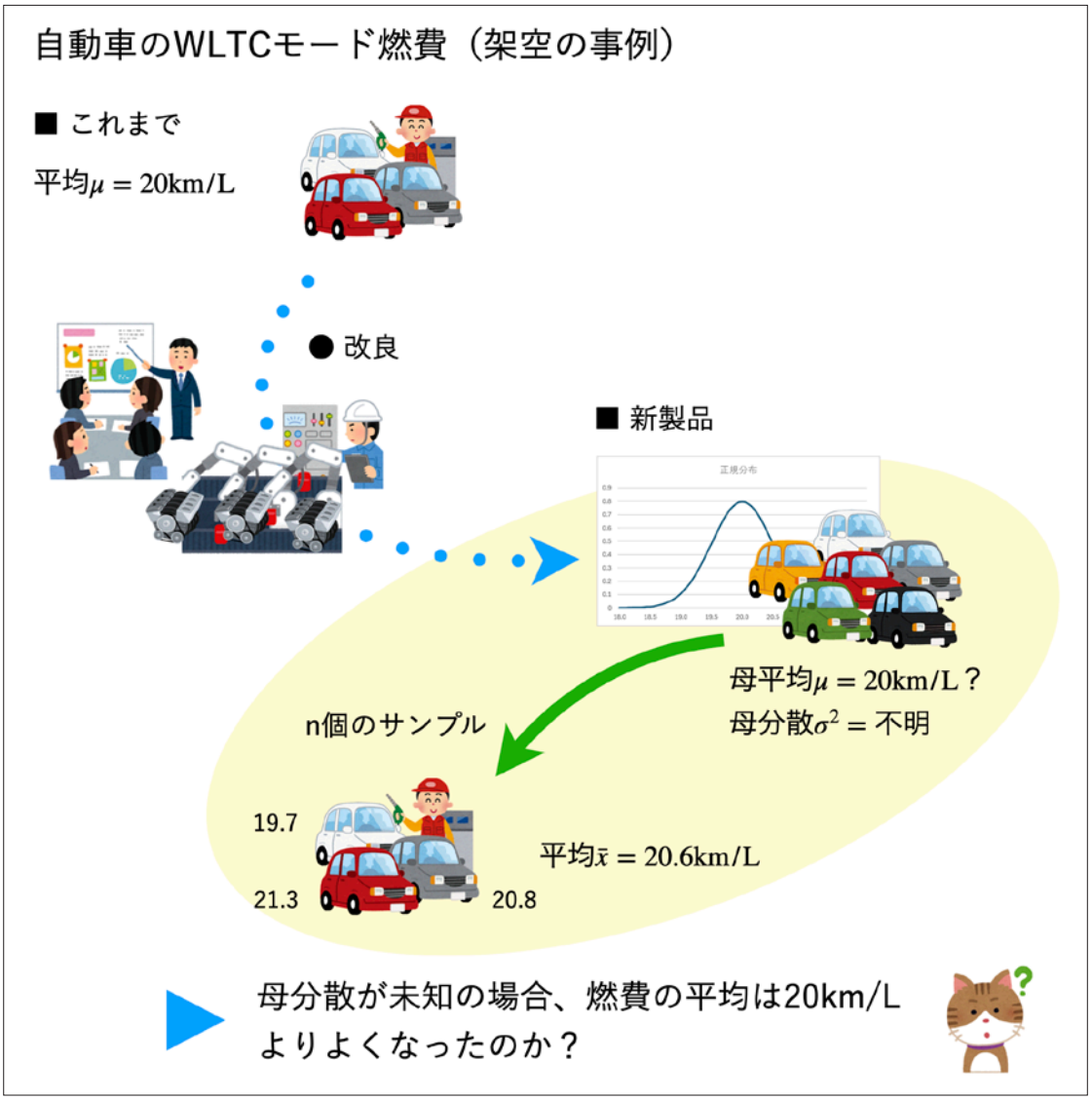


図 1　母分散が分かっていない場合、母平均がある値よりも大きくなったか？

新製品から n 個のサンプルを取り出して燃費を測定した結果から、新製品の母平均がある値よりも大きくなったかどうかを調べたい。母分散が分かっているときには正規分布を使うが、母分散が分からないときは t 分布を使う。WLTC モードの燃費とは市街地や郊外、高速道路の走行を想定した（実際の走行に近い）燃費のこと。ちなみに、実際の車種ごとの燃費は国土交通省のサイトでまとめて見られる。

図 1 の中で、薄い黄色の背景色を付けたところが今回取り扱う内容です。要するに、幾つかのサンプルを取り出したときに、母平均がある値と異なるかどうか、あるいは、ある値より大きくなったか（小さくなったか）を知りたいということです。結論から先に言うと、母集団の分散が分かっている場合は**正規分布**を使うのですが、母集団の分散が分かっていない場合は **t 分布**を使います。

図 1 に示した問いに答えるには、t 分布の確率変数はどのようなものか、t 分布の確率密度関数と累積分布関数はどのようなものかを知っておく必要があります。というわけで、ここから少しずつ見ていきます。母分散が分かっている場合と対比しながらお話しするので、正規分布についてのおさらいと補足も含みます。

母分散が分からない場合の平均値に関する分布 ～ t 分布の確率変数

母平均μと母分散σ^2が分かっている場合、母集団から n 個のサンプルを取り出して

$$\text{平均値}\bar{x}$$

を求めることを繰り返すと、

$$\bar{x}\text{の分布}$$

は平均μ、分散σ^2/nの正規分布に近づきます。……という定理の名前は何だったでしょうか。そう、何度も登場した**中心極限定理**でしたね。

一応、式で表しておきましょう。確率変数を

$$\bar{X}$$

とし、正規分布を N と表して、平均と分散をかっこの中に書くと、

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

となります。~ は「（ある分布に）従う」という意味の記号でしたね。

では、母集団が正規分布に従っているという前提の下で、**母分散σ^2が分かっていない場合**、サンプルの平均値に関する分布はどのようなものになるでしょうか（図 2）。

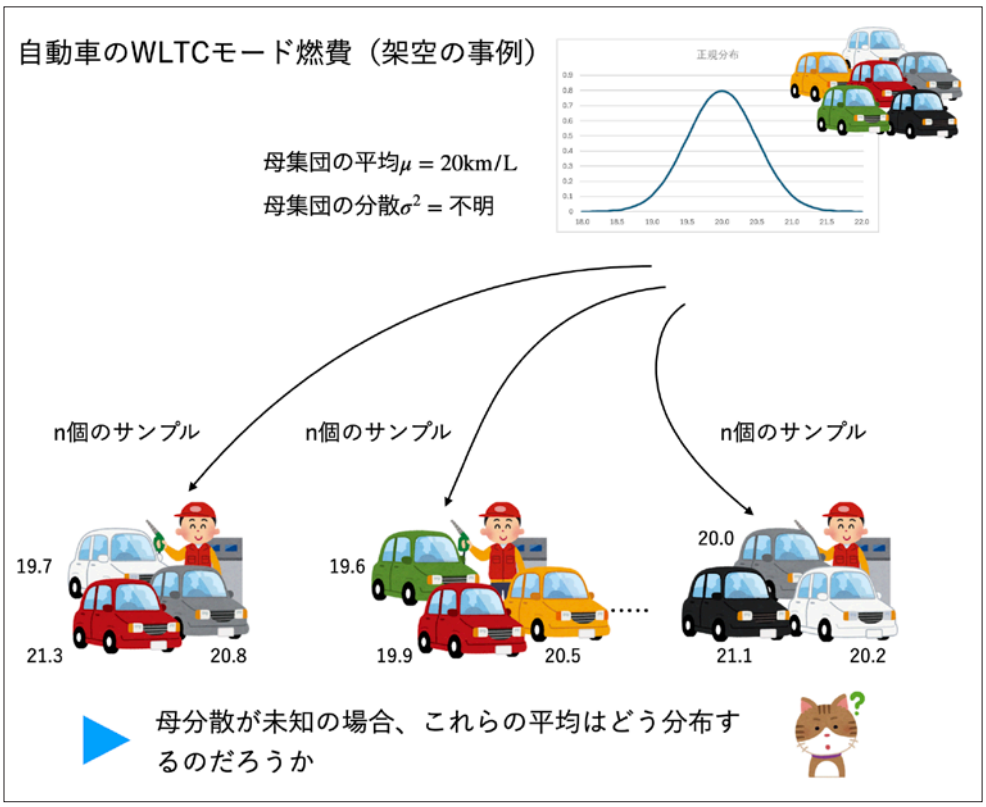


図 2　母分散が分かっていない場合に、平均値に関する分布は？

母集団が正規分布しているという前提の下で、燃費の平均が **20km/L** であるとする。母分散が分かっていない場合には、サンプルから求められた平均値（から求められた確率変数）はどう分布するのだろうか。

→目次に戻る

すでに答えは t 分布であるとお話ししましたが、**t 分布は平均値そのものの分布ではなく、平均値を基に求められた確率変数の分布である**ということに注意してください。

端的に言うことにします。母平均μの正規母集団（＝正規分布の母集団）から独立に取り出した n 個のサンプルの平均値を

$$\bar{X}$$

とすると、**母分散が分かっていない場合**、

$$\frac{\bar{X}-\mu}{s/\sqrt{n}}$$

という式で表される確率変数は、自由度 n − 1 の t 分布に従います。s は不偏標準偏差で、n はサンプルサイズです。自由度とは独立した変数の個数といった意味でしたが、それがなぜ n − 1 になるかは前回詳しく説明したので、そちらをご参照下さい。

t 分布の確率変数がなぜこのような式になるのかは、後の「t 分布と標準正規分布の関係」で説明します。ここでは、

$$\text{サンプルの平均（}\bar{X}\text{）から母平均（}\mu\text{）を引いた「差」を、}\frac{s}{\sqrt{n}}\text{で割ったもの}$$

が t 分布の確率変数になるということだけ確認してください。

ここまでのお話を ~ という記号を使って表すと、以下のようになります。$t_{(k)}$ は自由度 k の t 分布です。

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{(n-1)} \qquad (1)$$

（1）式の左辺が確率変数ですね。この値は **t 値**とも呼ばれます。繰り返しになりますが、t 分布は平均値そのものの分布ではないことに注意してください（大雑把に言うなら、母平均と不偏標準偏差を基にサンプルの平均値を標準化した値です）。（1）式の左辺に書かれた確率変数が自由度 n − 1 の t 分布に従う、ということですね。

では、t 分布の確率密度関数 f(k;x) と累積分布関数 F(k;x) はどのようなものでしょうか。一応、数式を掲載しておくと以下のようになります（以下の式では、自由度を k、確率変数つまり t 値を x と表記し、; で区切って表しています）。しかし、これらの式を覚える必要は全くありません。Excel の T.DIST 関数を使えば簡単に答えが求められるので、軽くスルーしてください。

$$f(k;x)=\frac{\Gamma\left((k+1)/2\right)}{\sqrt{k\pi}\Gamma(k/2)(1+x^2/k)^{(k+1)/2}}$$

$$F(k;z)=I_z(k/2,k/2)$$

ただし、I_z は正則化された不完全ベータ関数で、

$$z = \frac{x + \sqrt{x^2 + k}}{2\sqrt{x^2 + k}}$$

です。……などと言われてもやはり何が何だかですよね。後のコラムでこの定義通りに計算した例も紹介しますが、まずは、Excel の T.DIST 関数を使って t 分布の確率密度関数と累積分布関数を可視化してみましょう。

t 分布ってどんな感じの分布（1）～ 確率密度関数を可視化してみよう

T.DIST 関数を使って、t ＝－ 4.0 ～ 4.0 に対する確率密度関数の値を求め、グラフを描いてみます。**t 分布の母数は自由度のみです。**つまり、自由度が決まれば、t 分布の形も決まります。

先に、T.DIST 関数の形式を見ておきましょう（図 3）。

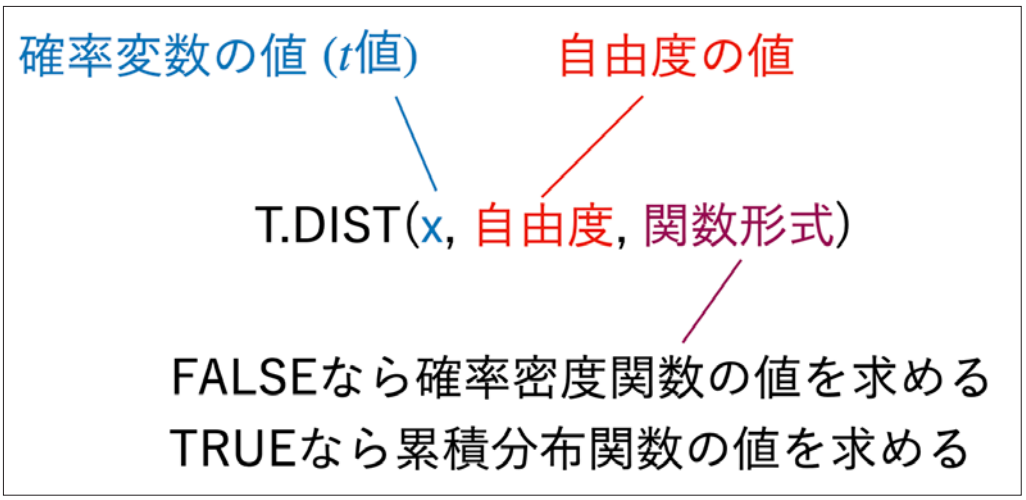
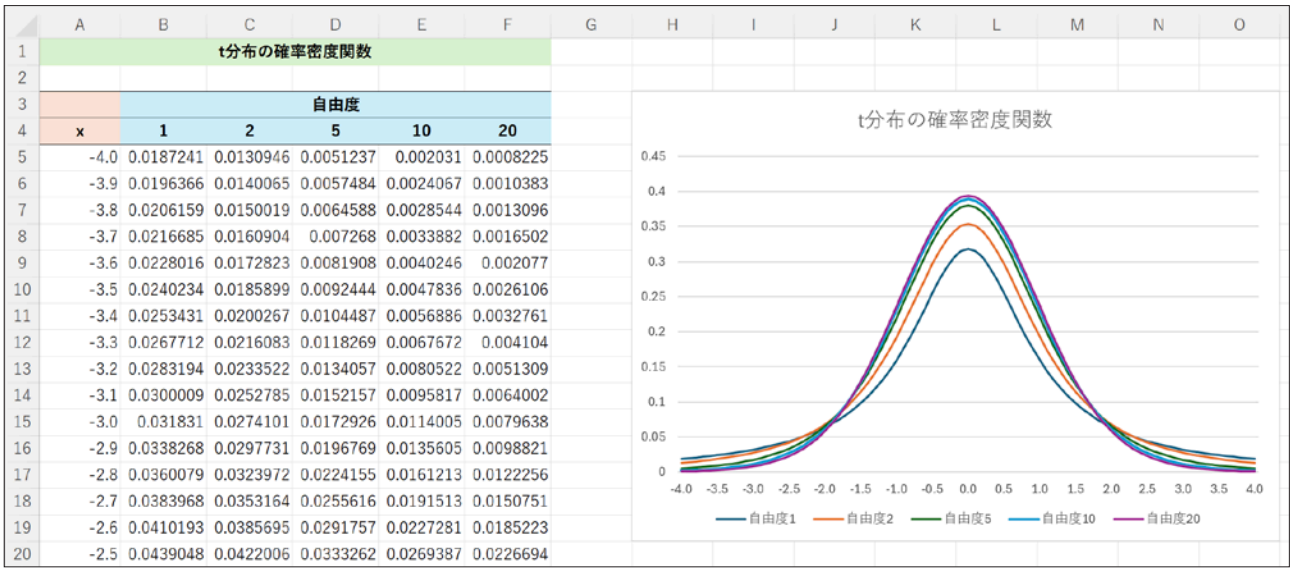


図 3　T.DIST 関数に指定する引数

T.DIST 関数には、確率変数の値（t 値）と自由度を指定する。関数形式についてはこれまで見てきた関数と同様、**FALSE** を指定すれば確率密度関数の値が、**TRUE** を指定すれば累積分布関数の値が求められる。

→目次に戻る

図 4 が、幾つかの自由度に対する t 分布の確率密度関数の値を求めてグラフを描いた例です（累積分布関数については次の項で取り扱います）。作成の手順は図の後に記しておきます。



	A	B	C	D	E	F
1	t分布の確率密度関数					
2						
3		自由度				
4	x	1	2	5	10	20
5	-4.0	0.0187241	0.0130946	0.0051237	0.002031	0.0008225
6	-3.9	0.0196366	0.0140065	0.0057484	0.0024067	0.0010383
7	-3.8	0.0206159	0.0150019	0.0064588	0.0028544	0.0013096
8	-3.7	0.0216685	0.0160904	0.007268	0.0033882	0.0016502
9	-3.6	0.0228016	0.0172823	0.0081908	0.0040246	0.002077
10	-3.5	0.0240234	0.0185899	0.0092444	0.0047836	0.0026106
11	-3.4	0.0253431	0.0200267	0.0104487	0.0056886	0.0032761
12	-3.3	0.0267712	0.0216083	0.0118269	0.0067672	0.004104
13	-3.2	0.0283194	0.0233522	0.0134057	0.0080522	0.0051309
14	-3.1	0.0300009	0.0252785	0.0152157	0.0095817	0.0064002
15	-3.0	0.031831	0.0274101	0.0172926	0.0114005	0.0079638
16	-2.9	0.0338268	0.0297731	0.0196769	0.0135605	0.0098821
17	-2.8	0.0360079	0.0323972	0.0224155	0.0161213	0.0122256
18	-2.7	0.0383968	0.0353164	0.0255616	0.0191513	0.0150751
19	-2.6	0.0410193	0.0385695	0.0291757	0.0227281	0.0185223
20	-2.5	0.0439048	0.0422006	0.0333262	0.0269387	0.0226694

図 4　t 分布の確率密度関数の例

自由度 1, 2, 5, 10, 20 について、x = − 4.0 ～ 4.0 までの確率密度関数の値を求め、グラフを描いてみた。x と表記されている **A 列**の値が確率変数（t 値）であることに注意。**B ～ F 列**はそれぞれの自由度に対する確率密度関数の値。T.DIST 関数を使って確率密度関数の値を求める手順は後の箇条書きを参照。

t 分布の台（確率変数が取り得る値の範囲）は−∞～∞です。そのうちの− 4.0 ～ 4.0 の範囲をグラフ化しています。グラフを見ると、x = 0 の（つまり t 値が **0** である）位置で山が最も高くなり、左右に裾野が広がる形の分布であることが読み取れます。実は、**t 分布は、自由度を大きくしていくと、標準正規分布に近づきます。**

確率密度関数の値を求めるための手順は以下の通りです。可視化については単に折れ線グラフを描くだけなので、関数の入力にのみ焦点を当てることにします。グラフ作成の手順についてはサンプルファイル内に掲載しておきます。

サンプルファイルをこちらからダウンロードし、［t 分布］ワークシートを開いて試してみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。

◆ Excel での操作方法

- セル **B5** に =T.DIST(A5:A85,B4:F4,FALSE) と入力する
  - ・古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **B5** ～ **F85**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す

◆ Google スプレッドシートでの操作方法

- セル **B5** に =ARRAYFORMULA(T.DIST(A5:A85,B4:F4,FALSE)) と入力する

● グラフの作成方法

- サンプルファイル内に掲載しておきます（タイトルや軸の書式などの細かい設定は省略）

t 分布ってどんな感じの分布（2）～累積分布関数を可視化してみよう

続いて、累積分布関数です。こちらは、自由度 **10** の例だけを見ておきます（図 5）。確率密度関数でグラフと x 軸で囲まれた範囲の面積が累積分布関数の値になることを示すために、確率密度関数も併せて作成し、説明をグラフ上に書き加えてあります。[t 分布累積] ワークシートを開き、図の後の手順で試してみてください。グラフ作成の手順についてはサンプルファイル内に掲載しておくこととします。

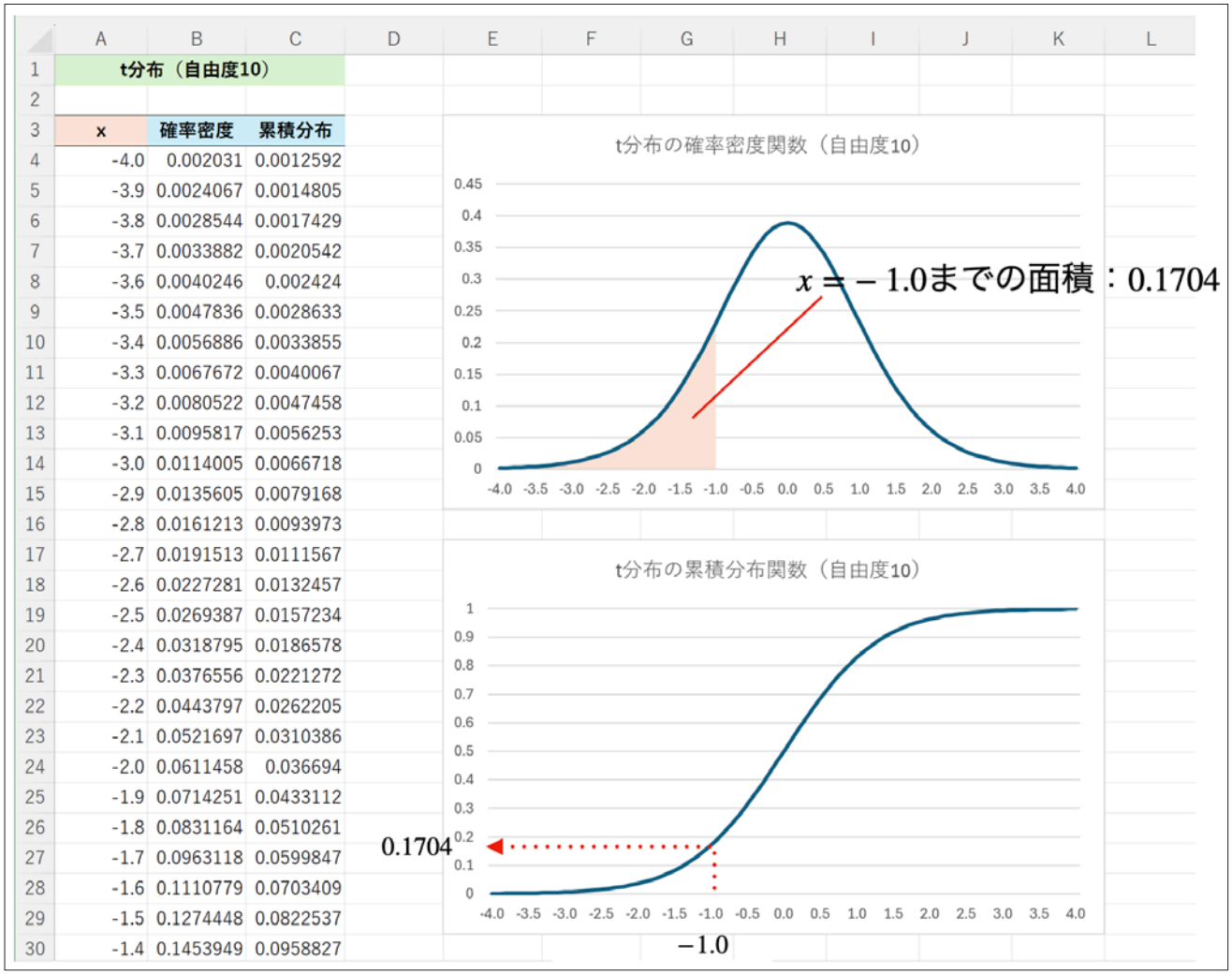


図 5　t 分布の累積分布関数の例

上のグラフが確率密度関数のグラフ。グラフと x 軸で囲まれた範囲の面積が累積分布関数の値になる。下のグラフは、その面積、つまり累積分布関数の値をプロットしたもの。例えば、x ＝－ 1.0 に対する累積分布関数の値は **0.1704** となる。

確率密度関数の値を求める方法はすでに見た通りですが、x ＝－ 4.0 ～ 4.0 の範囲を塗りつぶして表示するために使うので、併せて記しておきます。

→目次に戻る

◆ Excel での操作方法

- 確率密度関数の値を求める
  - ・セル **B4** に =T.DIST(A4:A84,10,FALSE) と入力する
    - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **B4** ～ **B84**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
- 累積分布関数の値を求める
  - ・セル **C4** に =T.DIST(A4:A84,10,TRUE) と入力する
    - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **C4** ～ **C84**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す

◆ Google スプレッドシートでの操作方法

- 確率密度関数の値を求める
  - ・セル **B4** に =ARRAYFORMULA(T.DIST(A4:A84,10,FALSE)) と入力する
- 累積分布関数の値を求める
  - ・セル **C4** に =ARRAYFORMULA(T.DIST(A4:A84,10,TRUE)) と入力する

● グラフの作成方法

- サンプルファイル内に掲載しておきます（タイトルや軸の書式などの細かい設定は省略）

本当に燃費は改善されたのか？ ～ 母分散が未知の場合の累積確率を求める

では、冒頭のお話の続きです。燃費を向上させるためにエンジンの設計などを見直したとします。これまでは燃費が **20km/L** でした。改良後、**10 台**のサンプルを取り出して燃費を測定したところ、以下のような値になったとします。

19.7, 21.3, 20.8, 21.4, 20.1, 19.8, 22.1, 21.1, 20.8, 19.9

これらの値の平均は **20.7** なので、燃費は改善されたように思われます。……が、そう判断していいのでしょうか。そこで、この場合の t 値と t 分布の累積分布関数の値を求めてみます。［応用例 1（母分散未知）］ワークシートを開いて確認してみましょう（図 6）。母平均を **20** と仮定して、サンプルの平均値から t 値を求めてみます。母分散は分からないので（1）式の左辺で t 値を求めればいいですね。（1）式を再掲しておきます。手順は図 6 の後に箇条書きで記しておきます。

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{(n-1)} \qquad (1)$$

さらに、t 値に対する自由度 n − 1 の t 分布の累積分布関数の値を求めれば、累積確率が得られます。ただし、ここでは平均値が大きくなったかどうかを知りたいので、T.DIST.RT 関数を使って右側確率（上側確率）を求めます。

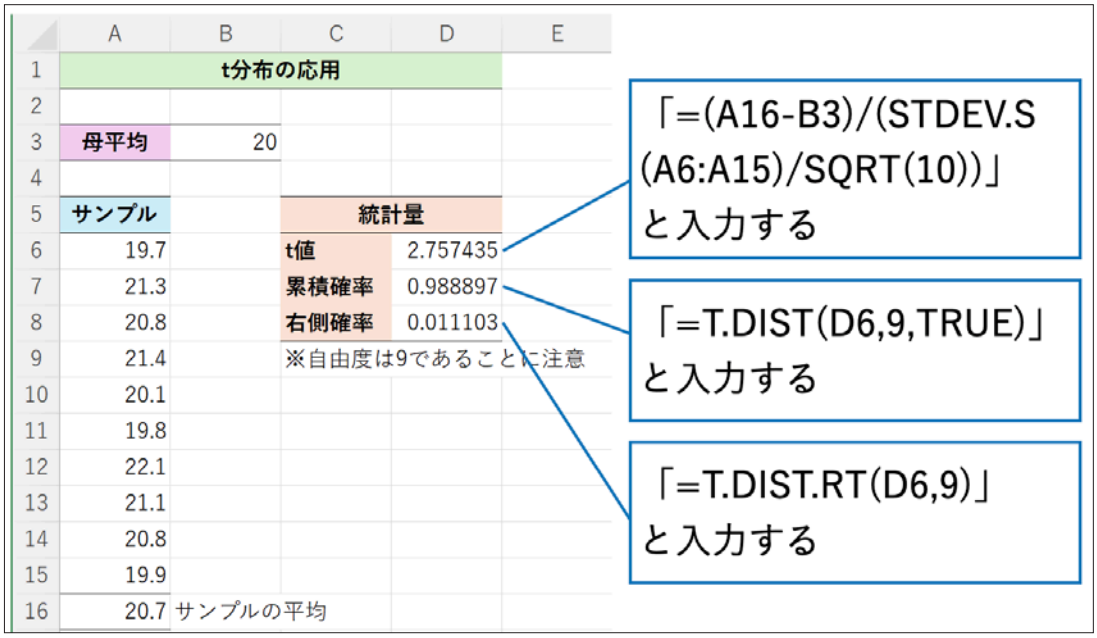


図 6　新しいエンジンの燃費は母平均と等しいか（母分散が未知の場合）

セル **D6** に =(A16-B3)/(STDEV.S(A6:A15)/SQRT(10)) と入力して t 値を求める。結果は t = 2.757 となる。さらにセル **D7** に =T.DIST(D6,9,TRUE) と入力して t = 2.757 に対する t 分布の累積分布関数の値を求めると **0.989**（= 98.9%）となる。ここでは、平均値が大きくなったかどうかを知りたいので、セル **D8** に =T.DIST.RT(D6,9) と入力して右側確率を求める。結果は **0.0111**（= 1.11%）であることが分かる。T.DIST.RT 関数では右側確率（右側の累積確率）を求めることが分かっているので、関数形式の引数はない。

セル **D6** で求めた t 値は（1）式の通りに計算したものです。つまり、サンプルの平均（セル **A16**）と母平均（セル **B3**）の差を求め、それを、

$$\text{不偏標準偏差} / \sqrt{n}$$

で割っています。これが冒頭で見た確率変数の求め方に従って計算した t 値です。結果は t = 2.757 です。セル **D8** でこの値に対する右側確率を求めると **0.0111**（= 1.11%）となります。この結果から、母平均が **20** であるとすると、取り出したサンプルの平均が **20.7** であるというのはかなり「まれ」なことであること分かります。つまり、これらのサンプルは平均が **20** である母集団から取り出されたものではなく、平均がもっと大きい母集団から取り出されたものだ、と言えそうです。というわけで、新しいエンジンの燃費は平均 **20km/L** よりも大きくなったものと考えられます。

この例では、セル **D7** の累積分布関数の値（左側確率）を求める必要はありませんが、参考のために掲載しておきました。なお、両側確率（t 値に対する左側確率と右側確率の小さい方の 2 倍）を求める T.DIST.2T 関数もあります。

上で求めた右側確率を可視化した図（図 7）も［応用例 1（母分散未知完成例）］ワークシートに含めてあります。こちらについてはワークシート内に作成方法を記載しておきます。

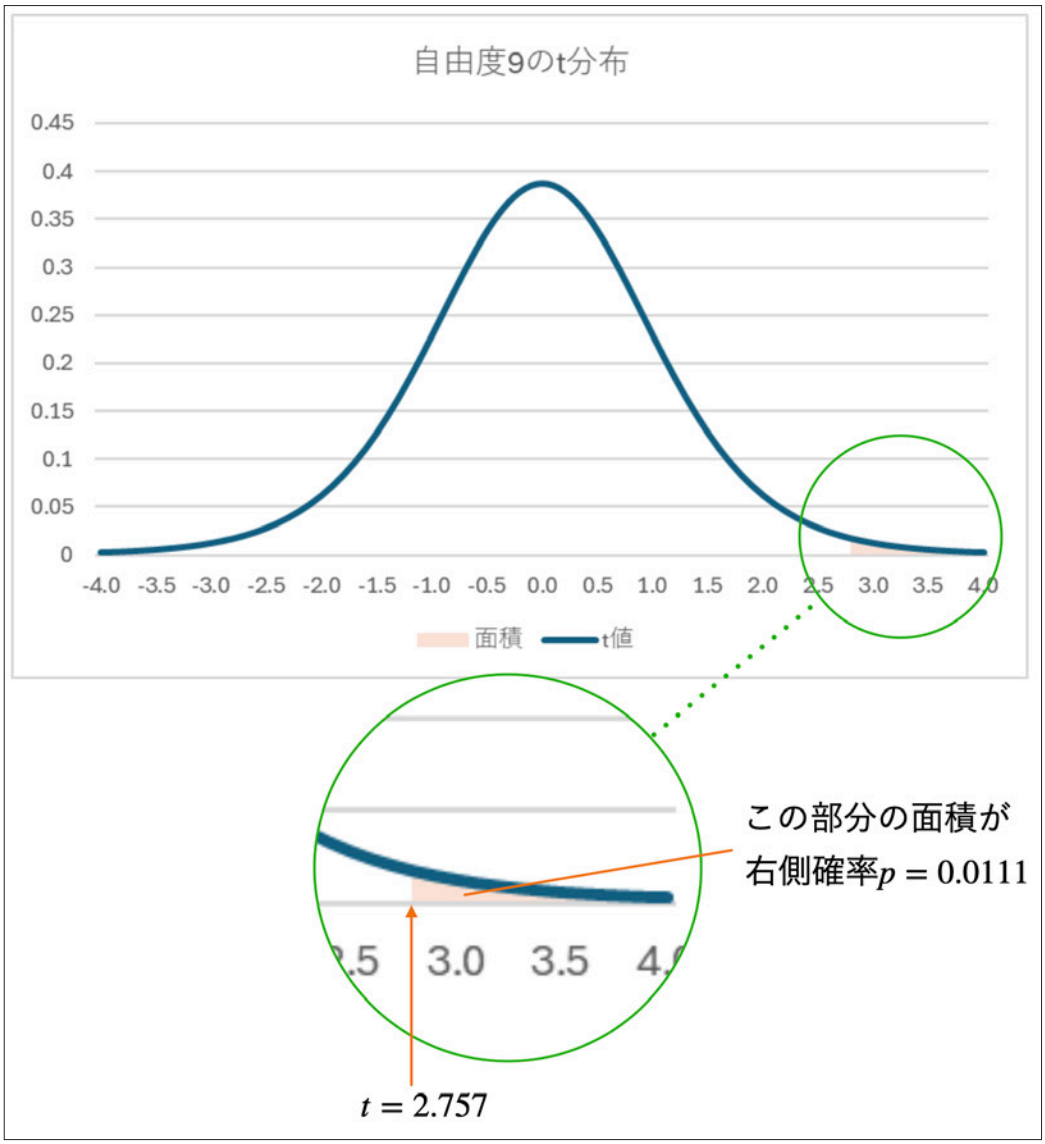


図 7　t=2.757 に対する累積確率

オレンジ色で塗りつぶした部分の面積が右側確率（**1.11%**）。平均が大きくなるとｔ値も大きくなるが、t 値が **2.757** になるのはかなり「まれ」なこと（＝確率変数の値がグラフの端の方に位置していて、面積が小さい）。新しいエンジンでは燃費が向上したものと考えられる。

実は、この計算は母平均の検定のための計算にほかなりません。ただし、帰無仮説や対立仮説などの考え方を十分に知った上で使う必要があるので、この時点で「結論はこうだ！」と言い切るのは少し待ってください。母平均の検定については（さらには平均値の差の検定などについても）、推測統計編でお話しします。

t 分布の累積分布関数に対する逆関数は？

T.INV 関数や T.INV.2T 関数を使えば、t 分布の累積分布関数に対する逆関数の値が求められます。T.INV 関数には左側確率と自由度を指定し、T.INV.2T 関数には両側確率と自由度を指定します。

例えば、自由度 **10** の t 分布で、累積確率（左側確率）が **95%**のときの t 値は =T.INV(95%, 10) で求められます。結果は **1.812** となります。

また、両側確率が **5%**のときの t 値は =T.INV.2T(5%, 10) で求められます。こちらの結果は **2.228** です。

空いているセルにこれらの式を入力してみて確認しておいてください。なお、これらの例は［t 分布の逆関数］ワークシートに入力されています。t 分布の累積分布関数に対する逆関数の値は、母平均の区間推定などに使われます。詳細は推測統計編でお話しします。

ここまで、t 分布の確率変数と確率密度関数、累積分布関数について見てきました。また、応用例として、母平均の検定につながるお話にも（かなり）踏み込みました。ここからは、数式やシミュレーションによって t 分布についての理解を深めていきたいと思います。実用的には知らなくてもあまり問題がないので、数式が苦手な方は、最後の「この記事で取り上げた関数の形式」を確認して、今回はお開きとしてもらっても構いません。

t 分布と標準正規分布の関係

最初に t 分布の自由度を大きくしていけば、標準正規分布に近づくというお話をしました。それを確かめましょう。中心極限定理を表す式は以下の通りでした。

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

N(μ , σ^2/n) は平均μ、分散σ^2/n（＝標準偏差$\sigma/\sqrt{n}$）の正規分布という意味です。上の式を**標準化**するために、

サンプルの平均$\bar{X}$から母平均μを引いて、
母集団の標準偏差$\sigma/\sqrt{n}$で割って

みましょう。

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad (2)$$

この式の意味は、左辺の

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

という確率変数が標準正規分布に従う、ということですね。

一方、t 分布の定義は（1）式の通りでした。再掲します。

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{(n-1)} \qquad (1)$$

（1）式と（2）式はよく似ていますね。左辺を見比べると、違いは標準偏差σと不偏標準偏差 s だけです。ところで、分散と不偏分散の関係は

$$s^2 = \frac{(n-1)\sigma^2}{n}$$

なので、n が大きくなると、(n − 1)/n が **1** に近くなります。つまり、不偏分散が分散とほぼ等しくなります。ということは、不偏標準偏差も標準偏差とほぼ等しくなります。というわけで、（1）式の n を大きくしていくと（2）式に近づく（＝ t 分布で自由度が大きくなると標準正規分布に近づく）ということが納得できると思います。

t 分布の平均（期待値）は **0** です。t 分布は標準正規分布と同様に左右対称の裾野を持ちますが、その山の中心が **0** になるということですね。自由度を k とすると、「t 分布の分散」は

$$\frac{k}{k-2}$$

という式で表されます（その導出は本稿では省略します）。k の値を大きくしていくと、この式の分母と分子の差がほとんどなくなって、分散が **1** に近づくことが分かります。

t 分布とカイ二乗分布の関係

以下のお話は、この時点では特に何かの役に立つわけではありませんが、t 分布と F 分布（次回取り上げます）の関係を知るのに役立つので、その「伏線」として少しお話ししておきます。数式が苦手な方はあまり気にしなくても構いません。

すでに何度も登場していますが、（1）式の左辺が自由度 n − 1 の t 分布の確率変数でした。それを t とすると、

$$t = \frac{\bar{X}-\mu}{s/\sqrt{n}} \qquad (3)$$

と表せます。前回、母平均が分かっていないときのカイ二乗値を簡単に求める式として、以下のような式を紹介しました。

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \qquad (4)$$

（3）式と（4）式からｔ分布とカイ二乗分布の関係が分かります。（4）式をｓ＝の形に変形してみましょう。

$$s = \sqrt{\frac{\sigma^2}{n-1}\chi^2} = \frac{\sigma}{\sqrt{n-1}}\sqrt{\chi^2} \tag{5}$$

（5）式を（3）式のｓに代入して、さらに変形します。

$$t = \frac{\bar{X}-\mu}{\underbrace{\frac{\sigma}{\sqrt{n-1}}\sqrt{\chi^2}}_{s}/\sqrt{n}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}\cdot\frac{\sqrt{\chi^2}}{\sqrt{n-1}}} = \underbrace{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}_{\text{標準正規分布}}\cdot\frac{1}{\sqrt{\frac{\chi^2}{n-1}}} \tag{6}$$

（6）式の右辺の最初の項は標準正規分布の確率変数です。これをＺと置くと、

$$t = \frac{Z}{\sqrt{\frac{\chi^2}{n-1}}} \tag{7}$$

となります。（7）式がｔ分布とカイ二乗分布の関係を表しています。もちろん、（3）式で求められるｔ値と（7）式で求められるｔ値は一致します。サンプルファイルには幾つかの値を指定して、（3）式の結果と（7）式の結果が一致するかを試した例を［t 分布とカイ二乗分布］ワークシートに含めてあるので、興味のある方はご参照ください。

コラム　t 分布の確率密度関数をシミュレーションで求める

すでに何度もお話ししていますが、母分散が分かっていない場合、正規母集団からｎ個のサンプルを取り出して平均値を求めることを繰り返すと、以下の式で求められる確率変数が自由度ｎ－１のｔ分布に従います。

$$\frac{\bar{X}-\mu}{s/\sqrt{n}}$$

これをシミュレーションしてみましょう。計算を簡単にするために、母集団は標準正規分布であるものとします。標準正規分布であれば母分散が$\sigma^2 = 1$と分かっているのですが、ここでは母分散が分からないものとして、サンプルから求めた不偏分散を使うことにします。

→目次に戻る

シミュレーションは Excel でもできますが、多数のセルを使う必要があり、ちょっと面倒です。一応サンプルファイルには作成例（[t 分布のシミュレーション] ワークシート）を含めておきますが、ここでは、Python のプログラムを使うことにします（リスト 1）。

サンプルプログラムはこちらから参照できます。リンクをクリックすれば、ブラウザが起動し、Google Colaboratory の画面が表示されます（Google アカウントでのログインが必要です）。最初のコードセルをクリックし、［Shift］＋［Enter］キーを押してコードを実行してみてください。コードの詳細については解説しませんが、コメントとリスト 1 の説明を見れば何をやっているかが大体分かると思います。

```
# t 分布の確率密度関数のシミュレーション
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import t, norm

# 標準正規分布（μ=0、σ^2=1）からランダムに 10 個のサンプルを取り出したものを 10000 個作り、平均を求める
x = np.random.randn(10000, 10)
t_mean = np.mean(x, axis=1)  # サンプルの平均
t_std = np.std(x, axis=1, ddof=1)  # サンプルから求めた不偏標準偏差 s
t_sample = t_mean / (t_std/np.sqrt(10))  # σではなく、s を使う

# ヒストグラムを描く（t 値がどのように現れるかが分かる）
plt.hist(t_sample, bins=100, range=(-4, 4), density=True)  # 階級は 100 個、縦軸は確率とする

# t 分布の確率密度関数を描く
x = np.linspace(-4, 4, 100)  # -4 ～ 4 までを 100 個に分けた等差数列（横軸の値）
plt.plot(x, t.pdf(x, 9))  # x に対する t 分布の確率密度関数の値をプロットする
plt.show()
```

リスト 1　t 分布の確率密度関数をシミュレーションする

標準正規分布（$\mu = 0, \sigma^2 = 1$）からランダムに n = 10 個の値を取り出すことを **10000 回**繰り返す。10000 行 ×10 列のデータが作られるので、各行の平均を求める。**10 個**のデータの平均が **10000 個**作られることになる。この **10000 個**の平均を基に t 値を求める。続いて、縦軸を確率として t 値のヒストグラムを描く。最後に、自由度 **9** の t 分布の確率密度関数を描けば、ヒストグラムとほぼ重なることが分かる。

リスト 1 の実行例が以下の図 8 です。乱数を使うので結果は毎回少しずつ異なりますが、ほとんど重なっていますね。

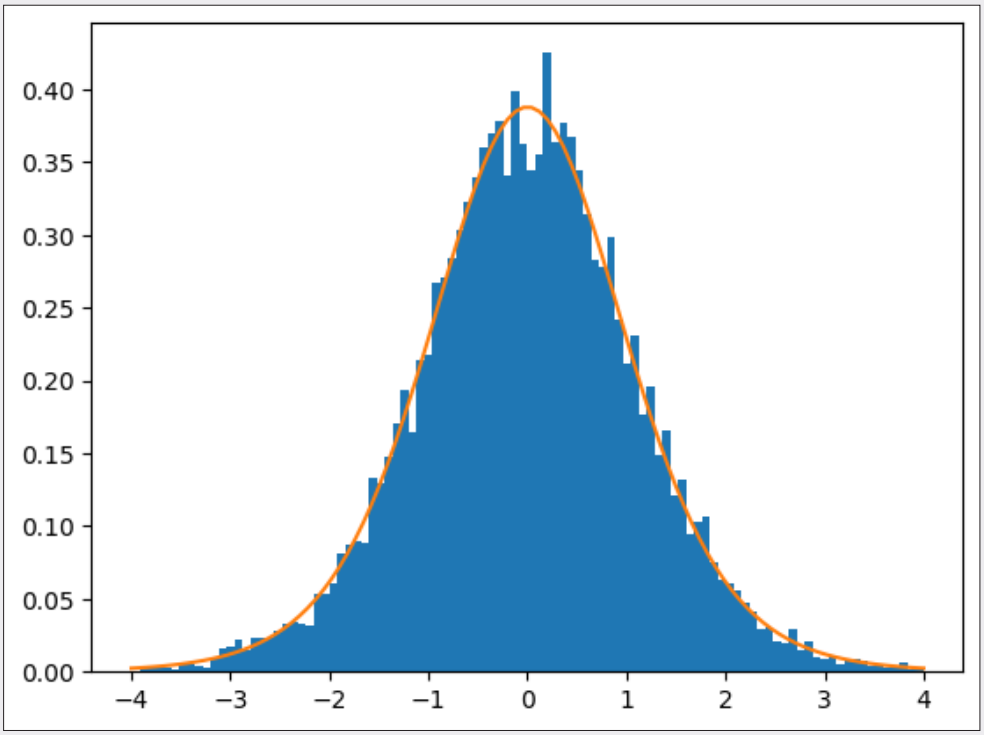


図 8　t 分布の確率密度関数のシミュレーション結果

棒グラフは、標準正規分布から **10 個**のサンプルを **10000 回**取り出し、それらの平均値を基に求めた t 値をヒストグラムにしたもの。折れ線グラフは自由度 **9** の t 分布の確率密度関数。

コラム　t 分布の確率密度関数と累積分布関数を数式で表す

最初に、t 分布の確率密度関数 f(x; k) と累積分布関数 F(x; k) を表す式を掲載しました。これらの式を覚える必要は全くありませんと言いましたが、もう少し詳しく知りたい方のために、定義通りに計算する方法も紹介しておきます。以下の式では、自由度を k、確率変数つまり t 値を x と表記し、; で区切って表しています。

$$f(k;x) = \frac{\Gamma\left((k+1)/2\right)}{\sqrt{k\pi}\Gamma(k/2)(1+x^2/k)^{(k+1)/2}} \qquad (8)$$

$$F(k;z) = I_z(k/2, k/2) \qquad (9)$$

ただし、Iz は正則化された不完全ベータ関数で、

$$z = \frac{x+\sqrt{x^2+k}}{2\sqrt{x^2+k}}$$

とします。Γ は前回も登場したガンマ関数です。以下にガンマ関数の定義や正則化された不完全ベータ関数の定義を記します（ただし、以下の式の t は t 値ではなく、単なる変数です）。しかし、これらの式を使わなくても、Excel の関数だけで（8）式と（9）式の計算ができます。

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$$

$$I_z(a,b) = \frac{B_z(a,b)}{B(a,b)}$$

ただし、

$$B_z(a,b) = \int_0^z t^{a-1}(1-t)^{b-1} dt$$

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

（8）式に登場するガンマ関数の値は、Excel の GAMMA 関数で求められます。

また（9）式に登場する、正則化された不完全ベータ関数 I_z は、ベータ分布の累積分布関数と一致します。

ベータ分布はこの連載では第 12 回で登場する予定ですが、Excel の BETA.DIST 関数で値が求められます。例えば、自由度が **5** で、t = 1.5 に対する確率密度関数の値を（8）式の定義に従って求めたい場合は、=GAMMA((5+1)/2)/(SQRT(5*PI())*GAMMA(5/2)*((1+1.5^2/5)^((5+1)/2))) と入力します。結果は **0.1245** となります。

一方、累積分布関数の値を（9）式に従って求めたい場合は、BETA.DIST 関数に z の値と a = k/2, b = k/2 の値を指定します。従って、=BETA.DIST((1.5+SQRT(1.5^2+5))/(2*SQRT(1.5^2+5)),5/2,5/2,TRUE) と入力します。結果は **0.9030** となります。

サンプルファイルの［t 分布（定義通りに）］ワークシートには上のような値ではなく、セルアドレスを指定して **-4.0** ～ **4.0** までの確率密度関数の値と累積分布関数の値を求めた例を含めてあります。また、不完全ベータ関数の値を求めるために積分の近似計算を行った例も含めてあるので、興味のある方はご参照ください。もっとも、実用的には T.DIST 関数を使った方が便利ですね。

◇ ◇ ◇ ◇ ◇ ◇ ◇

今回はt分布の確率変数であるt値の意味やその求め方、さらに、t分布の確率密度関数と累積分布関数の求め方などについてお話ししました。推測統計編の内容についても、かなり踏み込んでお話ししてしまいました。t分布は、母平均の区間推定や母平均の検定、平均値の差の検定などに使われますが、ここでは、t分布の確率変数と確率密度関数、累積分布関数についての理解を深めていただければ十分かと思います。

というわけで、次回は分散の比に関する分布として、F分布についてお話しします。次回もお楽しみに!

この記事で取り上げた関数の形式

関数の利用例については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

t分布の確率密度関数や累積分布関数の値を求めるための関数

T.DIST 関数：t分布の確率密度関数や累積分布関数の値を求める

形式

T.DIST(x, 自由度 , 関数形式)

引数

- **x**：確率変数の値（t値）を指定する。
- **自由度**：自由度を指定する。
- **関数形式**：以下の値を指定する。
  - **FALSE** …… 確率密度関数の値を求める
  - **TRUE** …… 累積分布関数の値を求める

備考

※累積分布関数の値は**左側確率**（または**下側確率**）とも呼ばれます。

T.DIST.RT 関数：t分布の右側確率の値を求める

形式

T.DIST.RT(x, 自由度)

引数

- **x**：確率変数の値（t値）を指定する。
- **自由度**：自由度を指定する。

備考

※ **1** から、T.DIST 関数で求められる累積分布関数の値を引いた値が返されます。

T.DIST.2T 関数：t 分布の両側確率の値を求める

形式

T.DIST.2T(x, 自由度)

引数

- **x**：確率変数の値（t 値）を指定する。
- **自由度**：自由度を指定する。

備考

※左側確率と右側確率の小さい方の 2 倍の値が返されます。

t 分布の累積分布関数に対する逆関数の値を求めるための関数

T.INV 関数：t 分布の累積分布関数に対する逆関数の値を求める

形式

T.INV(累積確率 , 自由度)

引数

- **累積確率**：累積分布関数の値を指定する。
- **自由度**：自由度を指定する。

備考

※累積確率には左側確率の値を指定します。右側確率に対する逆関数の値を求めたいときには、1 ー右側確率を指定します。

T.INV.2T 関数：t 分布の両側確率に対する逆関数の値を求める

形式

T.INV.2T(両側確率 , 自由度)

引数

- **両側確率**：累積分布関数の両側確率の値を指定する。
- **自由度**：自由度を指定する。

備考

※例えば、両側確率として 5% を指定すると、左側確率が **2.5%**、右側確率が **2.5%**となる t 値の絶対値が求められます。

［データ分析］F 分布
～ 2 つの農法で果物の糖度が安定しているのはどちら？

データ分析の初歩から学んでいく連載（確率分布編）の第 9 回。F 分布は分散の比に関連する分布です。2 つの母集団から取り出されたサンプルを基に「それぞれの母集団の分散に違いがあるのか」を調べる場合などに使われます。F 分布の確率変数と自由度の求め方を見た後、その確率密度関数や累積分布関数について解説します。

羽山博（2024 年 10 月 24 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の確率分布編、第 9 回です。前回は、平均値に関連する分布として、t 分布を取り上げました。今回は分散の比に関連する F 分布について、その特徴や意味を基本から解き明かし、確率密度関数／累積分布関数の求め方、利用例などを見ていきます。

果物の糖度が安定する農法はどちら？ ～ F 分布の利用

「糖度の高い果物」と言えば、皆さんは何を思い浮かべるでしょうか。筆者は巨峰などのぶどうを思い浮かべます。一般に、巨峰の糖度は **18** ～ **20** といわれており、果物の中ではかなり高い方です。果物を育てるに当たっては、糖度などの品質が高いこと、形が良いこと、収量が大きいことなど、考慮すべき事柄が数多くありますが、品質が安定していることも重要かと思います。

図 1 は、異なる 2 つの農法で育てたぶどうの木からサンプルを幾つか取り出して糖度を求めたイメージです（ただし、架空の例です）。サンプルの値を見ると、農法 B の方が糖度の分散が小さい（＝品質が安定している）ように思われます。このように、農法 A と農法 B では分散が異なるのか（農法 B が分散が小さいのか）といったことが、今回の問題意識です。要するに、分散を比較したいというわけです。

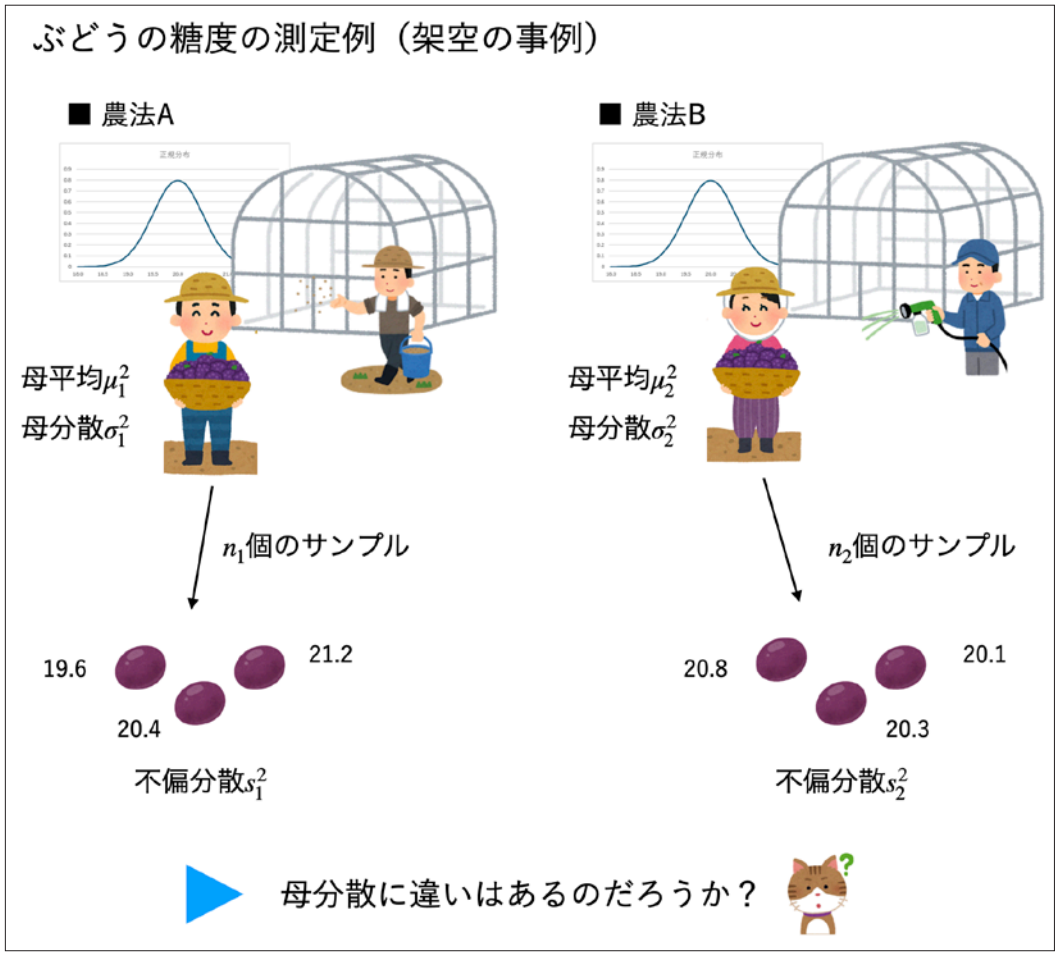


図 1　2 つの農法で育てたブドウの糖度に母分散の違いはあるのか？

図中には記していないが、農法 A の不偏分散は $s_1^2 = 0.64$、農法 B の不偏分散は $s_2^2 = 0.13$ となっている。この結果から、「農法 A よりも農法 B の母分散が小さい＝品質が安定している」と言えそうだが、果たしてどうなのか。それを知るためには F 分布の知識が必要となる。

分散を比較するには、分散の比を求めます。差（引き算）じゃなくて比（割り算）なんですか、と思われる方も多いと思います。それは、分散が二乗された値だからです。ちょっとした例え話をしてみましょう。

二乗した値で最も身近なものは面積ですね。公園などの広さをイメージするのに、よく「東京ドーム何個分」と表現することがあります。例えば、東京都立川市にある昭和記念公園の面積は約 1.8km^2（公開されている場所は約 1.69km^2）で、東京ドームの面積は約 0.047km^2 です。従って、昭和記念公園の面積は、東京ドームの $1.8/0.047 = 38.298$ 個分ということになります。これは、比（割り算）を使った計算ですね。面積を比較するのに、差を取って $1.8 - 0.047 = 1.753\text{km}^2$ だ、などという比較はしません。それと同じ考え方です。

分散の比に関する分布は **F 分布**と呼ばれます。そこで、F 分布の確率変数はどのようなものか、というところから出発し、その後、F 分布の確率密度関数と累積分布関数を可視化していきましょう。

分散の比に関する分布 ～ F 分布の確率変数

理屈は後回しにして、F 分布の確率変数がどのように定義されるかを示しておきます。異なる 2 つの正規母集団から独立にサンプルを取り出すことを考えてみましょう。例えば、農法 A で育てられたぶどうの糖度が母平均μ_1、母分散σ_1^2の正規分布に従い、農法 B で育てられたぶどうの糖度が母平均μ_2、母分散σ_2^2の正規分布に従うといった場合です。

このとき、それぞれの母集団からサンプルを取り出してカイ二乗値χ_1^2, χ_2^2を求めると、確率変数 F は、

$$F = \frac{\chi_1^2/k_1}{\chi_2^2/k_2} \qquad (1)$$

と定義されます。k_1,k_2 はそれぞれの自由度です。この確率変数が自由度 (k_1,k_2) の F 分布 $F(k_1,k_2)$ に従います。取り出したサンプルサイズが n_1 個、n_2 個の場合、$k_1 = n_1 - 1, k_2 = n_2 - 1$ ですね。

（1）式を見ると、F 値はカイ二乗値を自由度で割った値の比になっていることが分かります。カイ二乗分布はこの連載の第 7 回で見た通り、分散に関する分布です。このことから、F 分布は分散の比に関する分布であることが分かります。

さらに、それぞれのサンプルから求めた不偏分散が s_1、s_2 であるとすると、（1）式は以下のように変形できます。

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \qquad (2)$$

（1）式を（2）式に変形する手順については、後のコラムで解説します。s_1^2/σ_1^2 や s_2^2/σ_2^2 は、不偏分散と母分散との比ですね。つまり、サンプルから推定された母分散と、母分散との違いを表します。（2）式はそれらの比となっています。

しかし、これだけではイメージが湧かないと思われるので、（2）式に具体的な数値を幾つか当てはめて意味を考えてみましょう。母集団 A と母集団 B の母分散をそれぞれ$\sigma_1^2 = 0.3$, $\sigma_2^2 = 0.2$ と仮定すると、

ケース 1： $s_1^2 = 0.3, s_2^2 = 0.4$ の場合（母集団 A に比べて、母集団 B のサンプルから求めた不偏分散が母分散に対して大きい場合）は、

$$F = \frac{0.3/0.3}{0.4/0.2} = 1/2 = 0.5$$

ケース 2： $s_1^2 = 0.3, s_2^2 = 0.2$ の場合（いずれも、母分散と不偏分散が等しい場合）は、

$$F = \frac{0.3/0.3}{0.2/0.2} = 1/1 = 1.0$$

ケース 3: $s_1^2 = 0.6, s_2^2 = 0.2$ の場合（母集団 A に比べて、母集団 B のサンプルから求めた不偏分散が母分散に対して小さい場合）は、

$$F = \frac{0.6/0.3}{0.2/0.2} = 2/1 = 2.0$$

となります。

母集団から数多くのサンプルを取り出した場合は、不偏分散は母分散に近くなります。その場合、当然のことながらケース 2 に近くなるはずです。

というわけで、F 分布の確率密度関数は自由度が大きくなると $F = 1.0$ の辺りに山があることが想像されます（これについては後述します）。母集団 A に比べて、母集団 B の不偏分散が大きい場合は F 値が小さくなり、母集団 B の不偏分散が小さい場合は F 値が大きくなることも分かります。

では、次に F 分布の確率密度関数の値を求めて、可視化してみましょう。

F 分布ってどんな感じの分布（1）～ 確率密度関数を可視化してみよう

F 分布の確率密度関数 $f(x;k_1,k_2)$ と累積分布関数 $F(x;k_1,k_2)$ は以下の式で表されます（確率変数 x と自由度 k_1,k_2 を ; で区切って表記しています）。例によって、これらの式を覚える必要は全くありません。Excel の F.DIST 関数を使えば簡単に答えが求められるので、軽くスルーしてください。

$$f(x;k_1,k_2) = \frac{\Gamma((k_1+k_2)/2)(k_1/k_2)^{k_1/2}x^{k_1/2-1}}{\Gamma(k_1/2)\Gamma(k_2/2)(1+(k_1/k_2)x)^{(k_1+k_2)/2}}$$

$$F(x;k_1,k_2) = I_z(k_1/2,k_2/2)$$

ただし、k_1,k_2 は自由度、I_z は正則化された不完全ベータ関数で、

$$z = \frac{k_1 x}{k_1 x + k_2}$$

です。なんだか、目が回りそうな数式ですね。後のコラムでこの定義通りに計算した例も紹介しますが、Excel の F.DIST 関数を使って F 分布の確率密度関数と累積分布関数を可視化してみましょう。

F.DIST 関数を使って、$x = 0.0$ ～ 6.0 に対する確率密度関数の値を求め、グラフを描いてみます。**F 分布の母数は自由度のみです。**つまり、自由度が決まれば、F 分布の形も決まります。

先に、F.DIST 関数の形式を見ておきましょう。

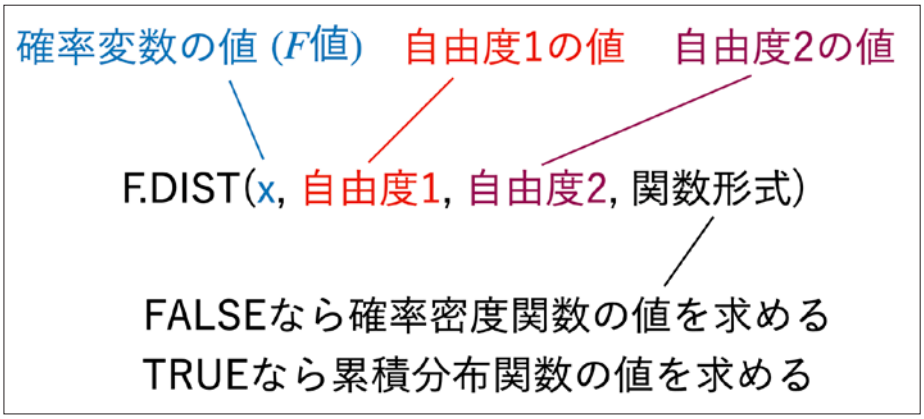


図 2　F.DIST 関数に指定する引数

F.DIST 関数には、確率変数の値（F 値）と「自由度 1」「自由度 2」を指定する（※「自由度が **1**」「自由度が **2**」という意味ではなく、「1 番目の自由度」「2 番目の自由度」という意味です）。関数形式についてはこれまで見てきた関数と同様、**FALSE** を指定すれば確率密度関数の値が、**TRUE** を指定すれば累積分布関数の値が求められる。

図 3 が、幾つかの自由度に対する F 分布の確率密度関数の値を求めてグラフを描いた例です（累積分布関数については次の項で取り扱います）。作成の手順は図の後に記しておきます。

	A	B	C	D	E
1	F分布の確率密度関数				
2					
3		上：自由度1, 下：自由度2			
4		4	5	6	10
5	x	6	8	10	20
6	0.0	0	0	0	0
7	0.1	0.3862381	0.2376807	0.1422971	0.0150468
8	0.2	0.57048	0.4636451	0.3664029	0.1198157
9	0.3	0.6430041	0.5993706	0.5430371	0.3113877
10	0.4	0.6542549	0.6611606	0.6492165	0.519763
11	0.5	0.6328125	0.6728861	0.6950823	0.687882
12	0.6	0.5949902	0.6537194	0.6976447	0.7920249
13	0.7	0.5500975	0.6170609	0.6722524	0.8330474
14	0.8	0.5033914	0.5717074	0.6305642	0.8236136
15	0.9	0.4577637	0.5231975	0.580714	0.7793528
16	1.0	0.41472	0.4748814	0.5280599	0.7143568
17	1.1	0.3749566	0.4286832	0.4759526	0.6395572
18	1.2	0.3387018	0.3856174	0.4263615	0.5626125
19	1.3	0.3059202	0.3461306	0.380338	0.4884255
20	1.4	0.276435	0.3103243	0.3383425	0.4198193
21	1.5	0.25	0.278098	0.300467	0.3581611

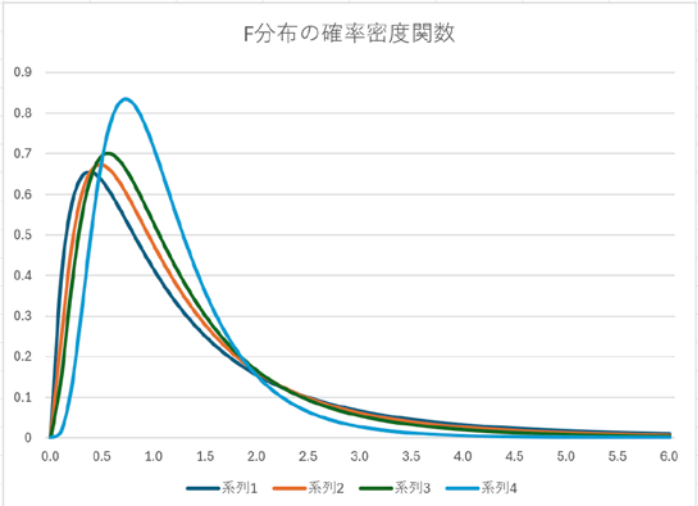


図 3　F 分布の確率密度関数の例

自由度 1 は **4**, **5**, **6**, **10** とし、それらに対応する自由度 2 を **6**, **8**, **10**, **20** として、x = 0.0 ～ 6.0 までの確率密度関数の値を求め、グラフを描いてみた。x と表記されている **A 列**の値が確率変数（F 値）であることに注意。**B ～ E 列**はそれぞれの自由度に対する確率密度関数の値。グラフ作成の手順は後の箇条書きを参照。

確率密度関数の値を求めるための手順は以下の通りです。可視化については単に折れ線グラフを描くだけなので、関数の入力にのみ焦点を当てることにします。グラフ作成の手順についてはサンプルファイル内に掲載しておきます。

サンプルファイルをこちらからダウンロードし、[F 分布] ワークシートを開いて試してみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。

◆ Excel での操作方法

- セル **B6** に =F.DIST(A6:A66,B4:E4,B5:E5,FALSE) と入力する
  - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **B6** ～ **E66**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す

◆ Google スプレッドシートでの操作方法

- セル **B6** に =ARRAYFORMULA(F.DIST(A6:A66,B4:E4,B5:E5,FALSE)) と入力する

● グラフの作成方法

- サンプルファイル内に掲載しておきます（タイトルや軸の書式などの細かい設定は省略）

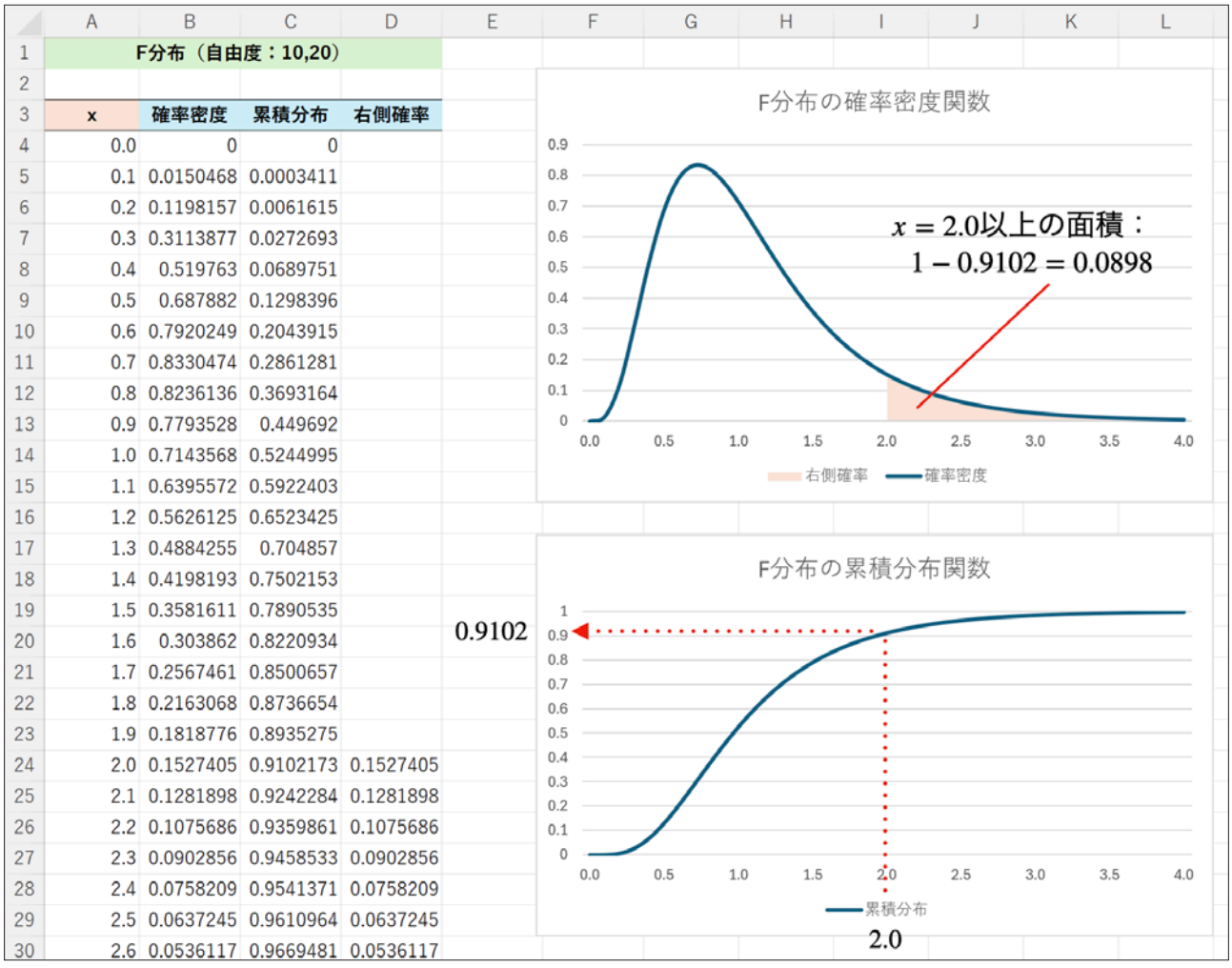
F 分布の台（確率変数が取り得る値の範囲）は 0 ～∞です。図 3 では、そのうちの 0.0 ～ 6.0 の範囲をグラフ化しています。グラフを見て、まず気付くのは F 分布の確率密度関数が左右対称でないことでしょう。また、自由度が大きくなると、山の最も高くなる位置が $x = 1.0$ の（つまり F 値が **1.0** である）位置に近づいていくことも分かります。

F 分布の期待値（平均）は $k_2/(k_2 - 2)$ です。従って、正確には自由度 2 が大きくなると山の最も高くなる位置が $x = 1.0$ に近づくということになります。ただし、グラフが左右対称ではなく、山が左に寄っていて、右側に裾が広がる形になっているので、実際には山の最も高くなる位置は少し左にずれます。

ちなみに自由度 1 が大きくなると山の中心付近が高くなり、裾がより広がります。図 3 でもある程度分かりますが、サンプルファイルで自由度 1 や自由度 2 を大きく変えてみるとそのことが分かるので、いろいろと試してみて下さい。

F 分布ってどんな感じの分布（2）～ 累積分布関数を可視化してみよう

続いて、累積分布関数です。こちらは、自由度 (**10**, **20**)、x = 0.0 ～ 4.0 の例だけを見ておきます（図 4）。確率密度関数でグラフと x 軸で囲まれた範囲の面積が累積分布関数の値になることを示すために、確率密度関数も併せて作成し、説明をグラフ上に書き加えてあります。[F 分布累積] ワークシートを開いて、図の後の手順で試してみてください。グラフ作成の手順についてはサンプルファイル内に掲載しておくこととします。



x	確率密度	累積分布	右側確率
0.0	0	0	
0.1	0.0150468	0.0003411	
0.2	0.1198157	0.0061615	
0.3	0.3113877	0.0272693	
0.4	0.519763	0.0689751	
0.5	0.687882	0.1298396	
0.6	0.7920249	0.2043915	
0.7	0.8330474	0.2861281	
0.8	0.8236136	0.3693164	
0.9	0.7793528	0.449692	
1.0	0.7143568	0.5244995	
1.1	0.6395572	0.5922403	
1.2	0.5626125	0.6523425	
1.3	0.4884255	0.704857	
1.4	0.4198193	0.7502153	
1.5	0.3581611	0.7890535	
1.6	0.303862	0.8220934	
1.7	0.2567461	0.8500657	
1.8	0.2163068	0.8736654	
1.9	0.1818776	0.8935275	
2.0	0.1527405	0.9102173	0.1527405
2.1	0.1281898	0.9242284	0.1281898
2.2	0.1075686	0.9359861	0.1075686
2.3	0.0902856	0.9458533	0.0902856
2.4	0.0758209	0.9541371	0.0758209
2.5	0.0637245	0.9610964	0.0637245
2.6	0.0536117	0.9669481	0.0536117

図 4　F 分布の累積分布関数の例

上のグラフが確率密度関数のグラフ。グラフと x 軸で囲まれた範囲の面積（この例では塗りつぶした部分以外の面積）が累積分布関数の値になる。塗りつぶした部分は右側確率に当たる。下のグラフは累積分布関数の値をプロットしたもの。例えば、x = 2.0 に対する累積分布関数の値は **0.9102** となる。右側確率は 1 − 0.9102 = **0.0898**。

確率密度関数の値を求める方法は既に見た通りですが、x = 2.0 ～ 4.0 の範囲を塗りつぶして表示するために使うので、併せて記しておきます。

→目次に戻る

◆ Excel での操作方法

- 確率密度関数の値を求める
  - ・セル **B4** に `=F.DIST(A4:A44,10,20,FALSE)` と入力する
    - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **B4** ～ **B44**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
- 累積分布関数の値を求める
  - ・セル **C4** に `=F.DIST(A4:A44,10,20,TRUE)` と入力する
    - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **C4** ～ **C44**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
- 右側確率を塗りつぶすための確率密度関数の値を求める
  - ・セル **D4** に `=IF(A4:A44<2,#N/A,B4:B44)` と入力する
    - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **D4** ～ **D44**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
  - ・セル **D4** ～ **D44** にはあらかじめ条件付き書式を設定し、値が **#N/A** のセルの文字色を白にして見えなくしてあります（操作はサンプルファイルを参照）

◆ Google スプレッドシートでの操作方法

- 確率密度関数の値を求める
  - ・セル **B4** に `=ARRAYFORMULA(F.DIST(A4:A44,10,20,FALSE))` と入力する
- 累積分布関数の値を求める
  - ・セル **C4** に `=ARRAYFORMULA(F.DIST(A4:A44,10,20,TRUE))` と入力する
- 右側確率を塗りつぶすための確率密度関数の値を求める
  - ・セル **D4** に `=ARRAYFORMULA(IF(A4:A44<2,#N/A,B4:B44))` と入力する
  - ・セル **D4** ～ **D44** にはあらかじめ条件付き書式を設定し、値が **#N/A** のセルの文字色を白にして見えなくしてあります（操作はサンプルファイルを参照）

● グラフの作成方法

サンプルファイル内に掲載しておきます（タイトルや軸の書式などの細かい設定は省略）

農法の違いにより糖度のばらつきは異なるのか？

さて、冒頭のお話の続きです。農法 A と農法 B とで母分散に違いがあるかどうか（農法 B の母分散の方が小さいのか）を知りたいということでした。まず（2）式を再掲します。

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \qquad (2)$$

この式を基に F 値を求めたいのですが、そもそも母分散が分かっていません。そこで、母分散が等しいと仮定しましょう。それが否定できれば母分散に違いがあることになりますね。つまり、(**2**) 式で $\sigma_1^2 = \sigma_2^2$ であるものとします。すると、

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{s_1^2}{s_2^2} \qquad (3)$$

となります。不偏分散の比がそのまま F 値になるというわけです。

s_1^2 よりも s_2^2 の値が大きければ F 値が小さくなり、累積確率も小さくなります。逆に、s_1^2 よりも s_2^2 の値が小さければ F 値が大きくなり、右側確率が小さくなります。前者は確率密度関数のグラフで、横軸の値が小さくなった場合に当たり、後者は大きくなった場合に当たります（図 **5**）。

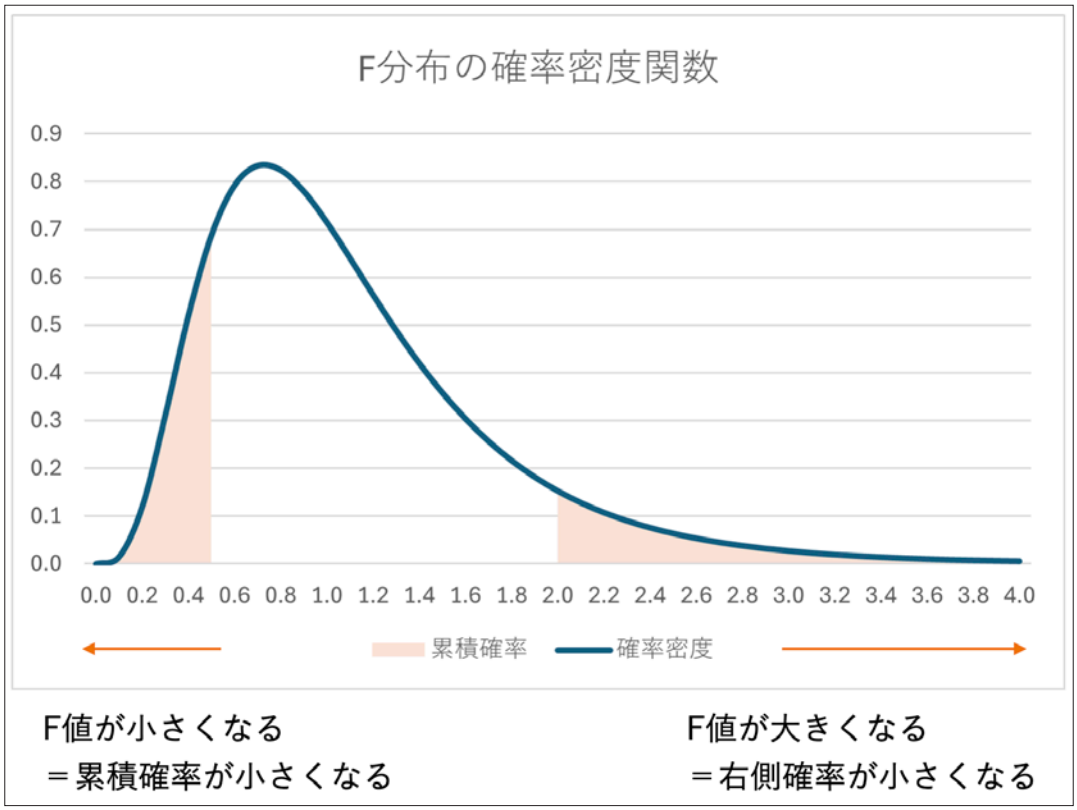


図 5　F 値の大小と累積確率、右側確率

不偏分散の比（F 値）が小さくなると、累積確率が小さくなる。逆に不偏分散の比が大きくなると、右側確率が小さくなる。不偏分散の比が **1** に近いと累積確率はグラフの半分の面積の **0.5** に近くなる。

では、以下の表 1 に示すサンプルで確認してみましょう。母分散が等しいと仮定すると、不偏分散の比、つまり F 値が極端に小さくなる（累積確率が小さくなる）ことはないはずです。逆に、F 値が極端に大きくなる（右側確率が小さくなる）こともないはずです。……が、そういった確率の小さなことが起こったのであれば、母分散が等しいという仮定を棄（す）てなければなりませんね。

農法	サンプル									
A	19.3	20.6	20.7	19.7	19.4	20.2	20.3	20.2	20.7	19.5
B	20.1	19.9	19.7	20.4	20.2	20.7	20.4	19.9	20.9	20.4

表 1　農法 A と農法 B から取り出したサンプルの値（糖度）

図 6 は、これらの値を基にそれぞれの不偏分散を求め、F 値とそれに対する累積確率と右側確率を求めたものです。［応用例］ワークシートを開き、図中に示した手順で試してみてください。

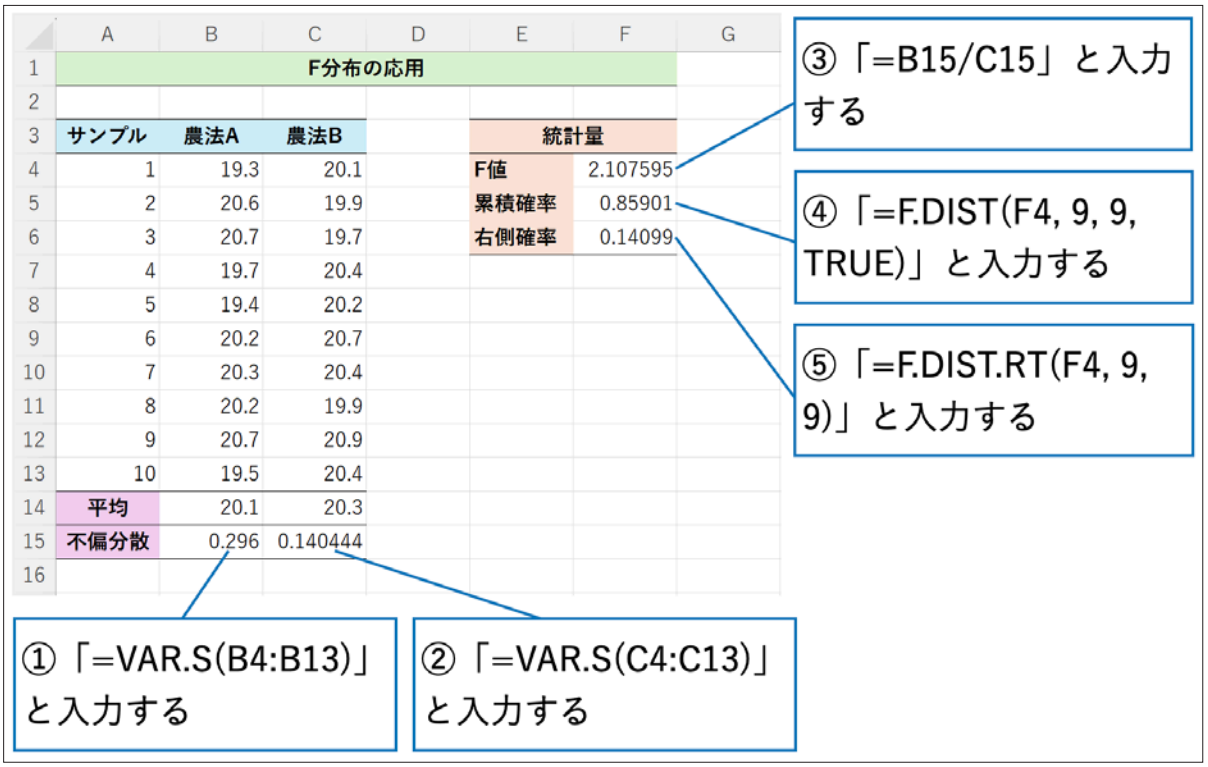


図 6　F 値の大小と累積確率、右側確率

農法 A と農法 B のサンプルの値は縦方向に入力されている。不偏分散を求めると農法 B のばらつきが小さいように見える。セル **F4** で求めた不偏分散の比が F 値となる。F 値を基にセル **F6** で右側確率を求めると **0.141**（＝ 14.1%）となる。自由度は **9** であることに注意。

結果は **14.1%**となります。一般に、**5%**あるいは **1%**以下であれば「まれにしか起こらないこと」と考えられ、母分散が等しいと仮定するには無理がある（農法 B の母分散が小さい）ということになります。しかし、この場合、比較的小さな値であるとはいえ、母分散が等しいという仮定を捨てるほどではないようです。つまり、農法 B の母分散が小さいとは言えないということになります（といっても、仮説を棄（す）てられないというだけなので、等しいと断定することはできません）。

実は、この計算は等分散の検定のための計算にほかなりません。ただし、帰無仮説や対立仮説などの考え方を十分に知った上で使う必要があるので、この時点で「結論はこうだ！」と言い切るのは少し待ってください。等分散の検定については推測統計編でお話しします。

F 分布の累積分布関数に対する逆関数は？

F.INV 関数や F.INV.RT 関数を使えば、F 分布の累積分布関数に対する逆関数の値が求められます。

F.INV 関数には左側確率と自由度 **1**、自由度 **2** を指定し、F.INV.RT 関数には右側確率と自由度 1、自由度 2 を指定します。例えば、自由度 1 が **6**、自由度 2 が **10** の F 分布で、累積確率（左側確率）が **95%**のときの F 値は =F.INV(95%, 6, 10) で求められます。結果は **3.217** となります。この確率は、右側確率が **5%**のときの F 値と等しくなるので、=F.INV.RT(5%, 6, 10) の結果と一致します。空いているセルにこれらの式を入力してみて確認しておいてください。

なお、これらの例は［F 分布の逆関数］ワークシートに入力されています。F 分布の累積分布関数に対する逆関数の値は、母分散の比の区間推定などに使われます。詳細は推測統計編でお話しします。

ここまで、F 分布の確率変数と確率密度関数、累積分布関数について見てきました。また、応用例として、等分散の検定につながるお話にも（かなり）踏み込みました。ここからは、数式やシミュレーションによって F 分布についての理解を深めていきたいと思います。実用的には知らなくてもあまり問題がないので、数式が苦手な方は、最後の「この記事で取り上げた関数の形式」を確認して、今回はお開きとしてもらっても構いません。

コラム　F 分布とカイ二乗分布、不偏分散の関係

冒頭で見た（**1**）式を（**2**）式に変形する手順を解説します。（**1**）式と（**2**）式をそれぞれ再掲しておきましょう。

$$F = \frac{\chi_1^2/k_1}{\chi_2^2/k_2} \qquad (1)$$

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \qquad (2)$$

母平均が未知の場合にカイ二乗分布の確率変数を簡単に求める方法をこの連載の第 7 回で紹介しました。自由度 n − **1** を k と表すと、以下のようになります。

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{ks^2}{\sigma^2}$$

左辺がそれぞれχ_1, χ_2であれば、

$$\chi_1^2 = \frac{k_1 s_1^2}{\sigma_1^2} \qquad (3)$$

$$\chi_2^2 = \frac{k_2 s_2^2}{\sigma_2^2} \qquad (4)$$

となります。（3）式と（4）式を（1）式に代入してみましょう。

$$F = \frac{\frac{\cancel{n_1} s_1^2}{\sigma_1^2} / \cancel{n_1}}{\frac{\cancel{n_2} s_2^2}{\sigma_2^2} / \cancel{n_2}} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

となります。ちゃんと（**2**）式になっていますね。

コラム　F 分布の確率密度関数をシミュレーションで求める

母分散がσ_1, σ_2 の 2 つの正規母集団から n_1,n_2 個のサンプルを取り出して不偏分散を求め、上で見た（**2**）式で F 値を繰り返し求めてみましょう。ヒストグラムを作成すると、自由度 1 が $n_1 - 1$、自由度 2 が $n_2 - 2$ の F 分布に従うはずです。

というわけで、シミュレーションしてみましょう。シミュレーションは Excel でもできますが、多数のセルを使う必要があり、ちょっと面倒です。一応サンプルファイルには作成例（[F 分布のシミュレーション] ワークシート）を含めておきますが、ここでは、Python のプログラムを使うことにします（リスト 1）。

サンプルプログラムはこちらから参照できます。リンクをクリックすれば、ブラウザが起動し、Google Colaboratory の画面が表示されます（Google アカウントでのログインが必要です）。最初のコードセルをクリックし、[Shift] ＋ [Enter] キーを押してコードを実行してみてください。結果は図 7 のようになります。コードの詳細については解説しませんが、コメントとリスト 1 の説明を見れば何をやっているかが大体分かると思います。

```
# F 分布の確率密度関数のシミュレーション
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import f

# 正規分布 N1（μ=3、σ^2=3）からランダムに 10 個のサンプルを取り出したものを 10000 個作る
x1_varp = 3
x1 = np.random.normal(loc=3, scale=np.sqrt(x1_varp), size=(10000, 10))
# 正規分布 N2（μ=4、σ^2=5）からランダムに 10 個のサンプルを取り出したものを 10000 個作る
x2_varp = 5
x2 = np.random.normal(loc=4, scale=np.sqrt(x2_varp), size=(10000, 10))

# サンプルの不偏分散を計算
x1_vars = np.var(x1, axis=1, ddof=1)
x2_vars = np.var(x2, axis=1, ddof=1)

# F 値を求める
F = (x1_vars/x1_varp) / (x2_vars/x2_varp)
# ヒストグラムを描く（F 値がどのように表れるかが分かる）
```

```
plt.hist(F, bins=100, range=(0, 6), density=True) # 階級は 100 個、縦軸は確率とする

# F 分布の確率密度関数を描く
x = np.linspace(0, 6, 100)  # 0～6までを 100 個に分けた等差数列（横軸の値）
plt.plot(x, f.pdf(x, 9, 9))  # x に対する F 分布の確率密度関数の値をプロットする
plt.show()
```

リスト 1　F 分布の確率密度関数をシミュレーションする

正規分布 N1（$\mu = 3, \sigma^2 = 3$）と正規分布 N2（$\mu = 4, \sigma^2 = 5$）からランダムに n = 10 個の値を取り出すことを **10000 回**繰り返す。10000 行 ×10 列のデータが作られるので、各行の不偏分散を求める。**10 個**のデータの不偏分散が **10000 個**作られることになる。この **10000 個**の不偏分散と母分散を基に F 値を求める。続いて、縦軸を確率として F 値のヒストグラムを描く。最後に、自由度 1 = **9**、自由度 2 = **9** の F 分布の確率密度関数を描けば、ヒストグラムとほぼ重なることが分かる。

リスト 1 の実行例が以下の図 7 です。乱数を使うので結果は毎回少しずつ異なりますが、ほとんど重なっていますね。

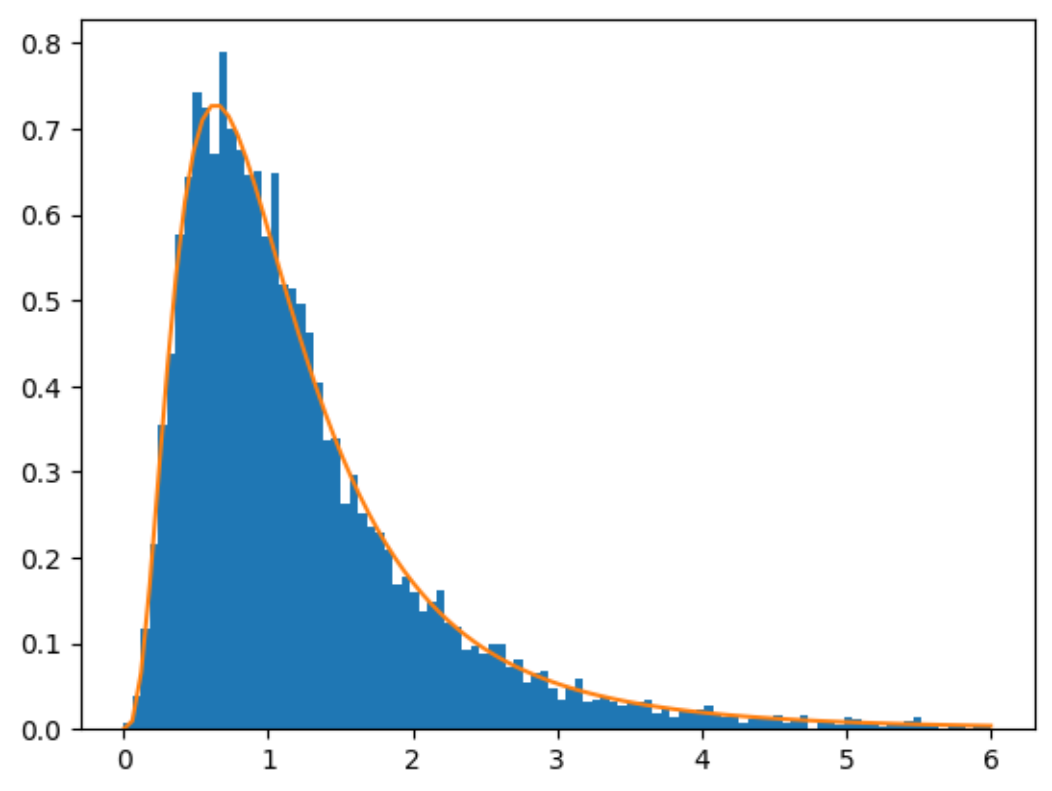


図 7　F 分布の確率密度関数のシミュレーション結果

棒グラフは、2 つの正規分布から **10 個**のサンプルを **10000 回**取り出し、それらの不偏分散と母分散を基に求めた F 値をヒストグラムにしたもの。折れ線グラフは自由度 1 = **9**、自由度 2 = **9** の F 分布の確率密度関数。

コラム　F 分布の確率密度関数と累積分布関数を数式で表す

最初に、F 分布の確率密度関数 $f(x;k_1,k_2)$ と累積分布関数 $F(x;k_1,k_2)$ を表す式を掲載しました（確率変数 x と自由度 k_1,k_2 を ; で区切って表記しています）。もう少し詳しく知りたい方のために、定義通りに計算する方法も紹介しておきます。

$$f(x;k_1,k_2)=\frac{\Gamma((k_1+k_2)/2)(k_1/k_2)^{k_1/2}x^{k_1/2-1}}{\Gamma(k_1/2)\Gamma(k_2/2)(1+(k_1/k_2)x)^{(k_1+k_2)/2}} \qquad (5)$$

$$F(x;k_1,k_2)=I_z(k_1/2,k_2/2) \qquad (6)$$

ただし、k_1,k_2 は自由度、I_z は正則化された不完全ベータ関数で、

$$z=\frac{k_1x}{k_1x+k_2}$$

とします。

Γはこの連載で何度も登場したガンマ関数です。ガンマ関数と不完全ベータ関数の定義については、前回紹介したので、そちらを参照していただくこととして、ここでは、Excel の関数を使った計算方法だけを示しておきます。ガンマ関数の値は GAMMA 関数で求められ、不完全ベータ関数の値は BETA.DIST 関数の累積分布関数で求められます。

例えば、自由度 1 が **6** で、自由度 2 が **10** の場合、F = 2 に対する確率密度関数の値を（5）式の定義に従って求めたい場合は、k_1 に **6**、k_2 に **10**、x に **2** を代入するだけです。分子を

```
=GAMMA((6+10)/2)*(6/10)^(6/2)*2^(6/2-1)
```

で求め、分母を

```
=GAMMA(6/2)*GAMMA(10/2)*(1+6/10*2)^((6+10)/2)
```

で求めて割り算すればいいですね。分子は **4354.56**、分母は **26340.42** となるので、結果は **0.1653** となります。一方、累積分布関数の値を（6）式に従って求めたい場合は、BETA.DIST 関数に z の値と a = k_1/2, b = k_2/2 の値を指定します。従って、

```
=BETA.DIST(6*2/(6*2+10), 6/2, 10/2, TRUE)
```

と入力します。結果は **0.8411** となります。サンプルファイルの［F 分布（定義通りに）］ワークシートには上のような値ではなく、セルアドレスを指定して **0.0** ～ **6.0** までの確率密度関数の値と累積分布関数の値を求めた例を含めてあります。もっとも、実用的には F.DIST 関数を使った方が便利ですね。

◇ ◇ ◇ ◇ ◇ ◇ ◇

今回は F 分布の確率変数である F 値の意味やその求め方、さらに、F 分布の確率密度関数と累積分布関数の求め方などについてお話ししました。推測統計編の内容についても、かなり踏み込んでお話ししてしまいました。F 分布は、母分散の比の区間推定や等分散の検定などに使われますが、ここでは、F 分布の確率変数と確率密度関数、累積分布関数についての理解を深めていただければ十分かと思います。

次回は一定時間内に何らかの事象が起こる確率を求める場合などに使われる指数分布についてお話しします。次回もお楽しみに！

この記事で取り上げた関数の形式

関数の利用例については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

F 分布の確率密度関数や累積分布関数の値を求めるための関数

F.DIST 関数：F 分布の確率密度関数や累積分布関数の値を求める

形式

F.DIST(x, 自由度 1, 自由度 2, 関数形式)

引数

x：確率変数の値（F 値）を指定する。

自由度 1：自由度 1 を指定する。

自由度 2：自由度 2 を指定する。

関数形式：以下の値を指定する。

- **FALSE** …… 確率密度関数の値を求める
- **TRUE** …… 累積分布関数の値を求める

備考

※累積分布関数の値は**左側確率**（または**下側確率**）とも呼ばれます。

F.DIST.RT 関数：F 分布の右側確率の値を求める

形式

F.DIST.RT(x, 自由度 1, 自由度 2)

引数

x：確率変数の値（F 値）を指定する。

自由度 1：自由度 1 を指定する。

自由度 2：自由度 2 を指定する。

備考

※ 1 から、F.DIST 関数で求められる累積分布関数の値を引いた値が返されます。

F 分布の累積分布関数に対する逆関数の値を求めるための関数

F.INV 関数：F 分布の累積分布関数に対する逆関数の値を求める

形式

F.INV(累積確率 , 自由度 1, 自由度 2)

引数

累積確率：累積分布関数の値を指定する。

自由度 1：自由度 1 を指定する。

自由度 2：自由度 2 を指定する。

備考

※累積確率には左側確率の値を指定します。右側確率に対する逆関数の値を求めたいときには、F.INV.RT 関数を使うか、1 －右側確率を指定します。

F.INV.RT 関数：F 分布の右側確率に対する逆関数の値を求める

形式

F.INV.RT(右側確率 , 自由度 1, 自由度 2)

引数

右側確率：累積分布関数の右側確率の値を指定する。

自由度 1：自由度 1 を指定する。

自由度 2：自由度 2 を指定する。

備考

※例えば、右側確率が 5% の場合の F 値は、F.INV 関数で左側確率を 95% と指定した場合の F 値と一致します。

［データ分析］指数分布
～ 5 分以内に次の顧客が到着する確率は？

データ分析の初歩から学んでいく連載（確率分布編）の第 10 回。指数分布は待ち行列の分析などに使われる分布です。一定期間に起こる事象の数が分かっているときに、ある期間内にその事象が起こる確率が求められます。今回は具体例を基に、確率を求めたり、指数分布の確率密度関数や累積分布関数の形を見ていきます。

羽山博（2024 年 11 月 07 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の確率分布編、第 10 回です。前回は、分散の比に関連する分布として、F 分布を取り上げました。今回は待ち行列の分析などに使われる指数分布について、その特徴や意味を基本から解き明かし、確率密度関数／累積分布関数の求め方、利用例などを見ていきます。

一定時間内に顧客が来る確率は？ ～ 指数分布の利用

指数分布は、一定の時間内に事象（何らかの出来事）が何回か起こることが分かっているときに、ある時間間隔でその事象が起こる確率を表す分布です。これは、具体的な例で説明した方が分かりやすいでしょう。例えば、ある店舗で、顧客が **10 分**当たり **3 人**来ることが分かっている場合、**次の 5 分間**に顧客が来る確率を求める、といった場合に使われます（図 1）。

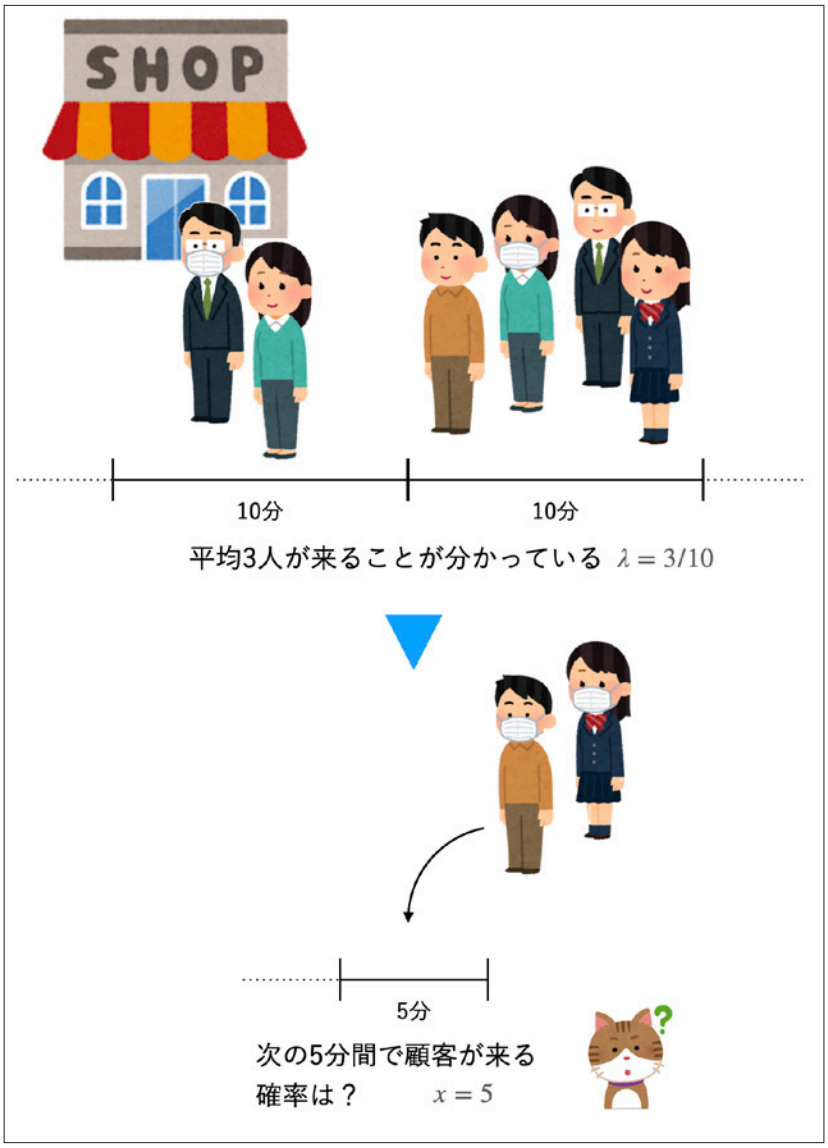


図 1　ある時間内に顧客が来る確率を知りたい！

単位時間に事象が起こる回数λが分かっているとき、目的の時間間隔 x の間にその事象が起こる確率を求めるのに指数分布（の累積分布関数）が使える。顧客が **10** 分当たり **3** 人来る場合、単位時間は「分」なので、λの値は **3/10**。目的の時間間隔が **5** 分なら x = **5** となる。

指数分布の確率変数 x は目的の時間間隔です。単位時間内で事象が起こる回数をλ（ラムダ）とすると、指数分布の確率密度関数と累積分布関数は以下の式で定義されます。e は自然対数の底（= **2.71828**...）です。

◆ 指数分布の確率密度関数

$$f(x) = \lambda e^{\lambda x} \qquad (1)$$

◆ 指数分布の累積分布関数

$$F(x) = 1 - e^{-\lambda x} \qquad (2)$$

→目次に戻る

母数はλだけなので、λの値が決まると関数の形が決まります。指数分布の**累積分布関数**を使えば、目的の時間内に次の顧客が来る確率が求められるので、図 1 の例で計算してみましょう。

単位時間は「分」とします。**10 分**当たり **3 人**の顧客が来るということは、**1 分**当たり$\lambda = 3/10$ 人が来るということです。目的の時間間隔は **5 分**なので、$x = 5$ です。（2）式に当てはめてみましょう。簡単な穴埋め問題にしてあります。

$$
\begin{aligned}
F(x) &= 1 - e^{-\lambda x} \\
&= 1 - e^{-\frac{\boxed{ア}}{10} \times \boxed{イ}} \\
&= 1 - 2.71828^{-1.5} \\
&= 0.7769
\end{aligned}
$$

答え：ア＝　**3**　、イ＝　**5**

なお、指数分布の累積分布関数は、目的の時間内に次の顧客がちょうど 1 人来る確率ではないことに注意してください。顧客は何人来てもいいので、顧客が **0 人**ではない（**＝少なくとも 1 人来る**）確率であると考えられます。

指数分布が問題としているのは、何人の顧客が来るかということではなく、あくまでも、次の顧客が来るのが目的の時間内である確率です（目的の時間間隔が指数分布の確率変数であることからも明らかです）。

指数分布は一定時間内に機械が故障する確率を求めたり、一定期間内に株価が暴落する確率を求めたりするなど、さまざまな場面で応用されます。

では、次に指数分布の確率密度関数と累積分布関数を可視化してみましょう。

指数分布ってどんな感じの分布（1）～ 確率密度関数を可視化してみよう

指数分布の確率密度関数や累積分布関数を可視化するには（1）式や（2）式をそのまま使ってもいいですが、Excel の EXPON.DIST 関数を使うのが簡単です。EXPON.DIST 関数の形式を見ておきましょう。

図 2　EXPON.DIST 関数に指定する引数

EXPON.DIST 関数には、目的の時間間隔 x と、単位時間内で事象が起こる回数λを指定する。関数形式についてはこれまでの連載で見てきた関数と同様、**FALSE** を指定すれば確率密度関数の値が、**TRUE** を指定すれば累積分布関数の値が求められる。

まず、確率密度関数から見ていきます。図 3 は、幾つかのλの値を想定し、x = 0.0 ～ 15.0 までの確率密度関数の値をプロットしたものです。表の作成手順は図の後に記しておきます。

	A	B	C	D	E
1	指数分布の確率密度関数				
2					
3		λ			
4	x	0.1	0.3	0.5	0.8
5	0.0	0.1	0.3	0.5	0.8
6	0.5	0.0951229	0.2582124	0.3894004	0.536256
7	1.0	0.0904837	0.2222455	0.3032653	0.3594632
8	1.5	0.0860708	0.1912884	0.2361833	0.2409554
9	2.0	0.0818731	0.1646435	0.1839397	0.1615172
10	2.5	0.0778801	0.14171	0.1432524	0.1082682
11	3.0	0.0740818	0.1219709	0.1115651	0.0725744
12	3.5	0.0704688	0.1049813	0.086887	0.0486481
13	4.0	0.067032	0.0903583	0.0676676	0.0326098
14	4.5	0.0637628	0.0777721	0.0526996	0.021859
15	5.0	0.0606531	0.066939	0.0410425	0.0146525
16	5.5	0.057695	0.057615	0.0319639	0.0098219
17	6.0	0.0548812	0.0495897	0.0248935	0.0065838
18	6.5	0.0522046	0.0426822	0.0193871	0.0044133
19	7.0	0.0496585	0.0367369	0.0150987	0.0029583
20	7.5	0.0472367	0.0316198	0.0117589	0.001983
21	8.0	0.0449329	0.0272154	0.0091578	0.0013292
22	8.5	0.0427415	0.0234245	0.0071321	0.000891

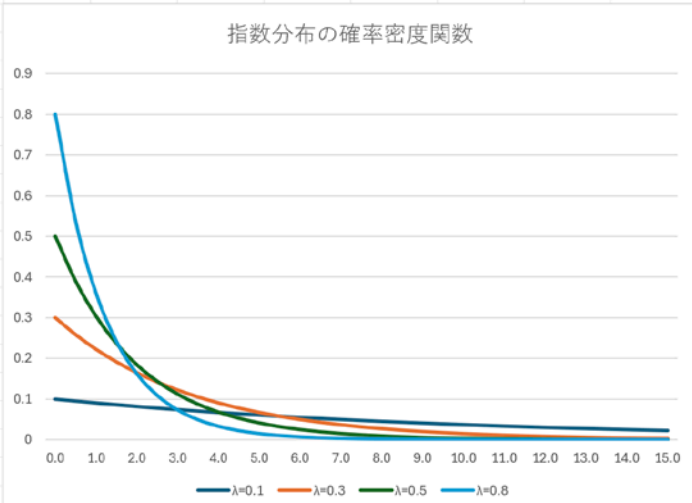


図 3　指数分布の確率密度関数の例

λは **0.1**, **0.3**, **0.5**, **0.8** とし、x = 0.0 ～ 15.0 までの確率密度関数の値を **0.5** 刻みで求め、グラフを描いてみた。x と表記されている **A** 列の値が確率変数。**B** ～ **E** 列はそれぞれのλに対する確率密度関数の値。

→目次に戻る

確率密度関数の値を求めるための手順は以下の通りです。可視化については単に折れ線グラフを描くだけなので、関数の入力にのみ焦点を当てることにします。グラフ作成の手順についてはサンプルファイル内に掲載しておきます。

サンプルファイルをこちらからダウンロードし、［指数分布］ワークシートを開いて試してみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。なお、（1）式の通りに計算した例も［指数分布（定義通りに作成）］ワークシートに含めてあります。

◆ Excel での操作方法

- セル **B5** に `=EXPON.DIST(A5:A35,B4:E4,FALSE)` と入力する
  - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **B5** ～ **E35**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す

◆ Google スプレッドシートでの操作方法

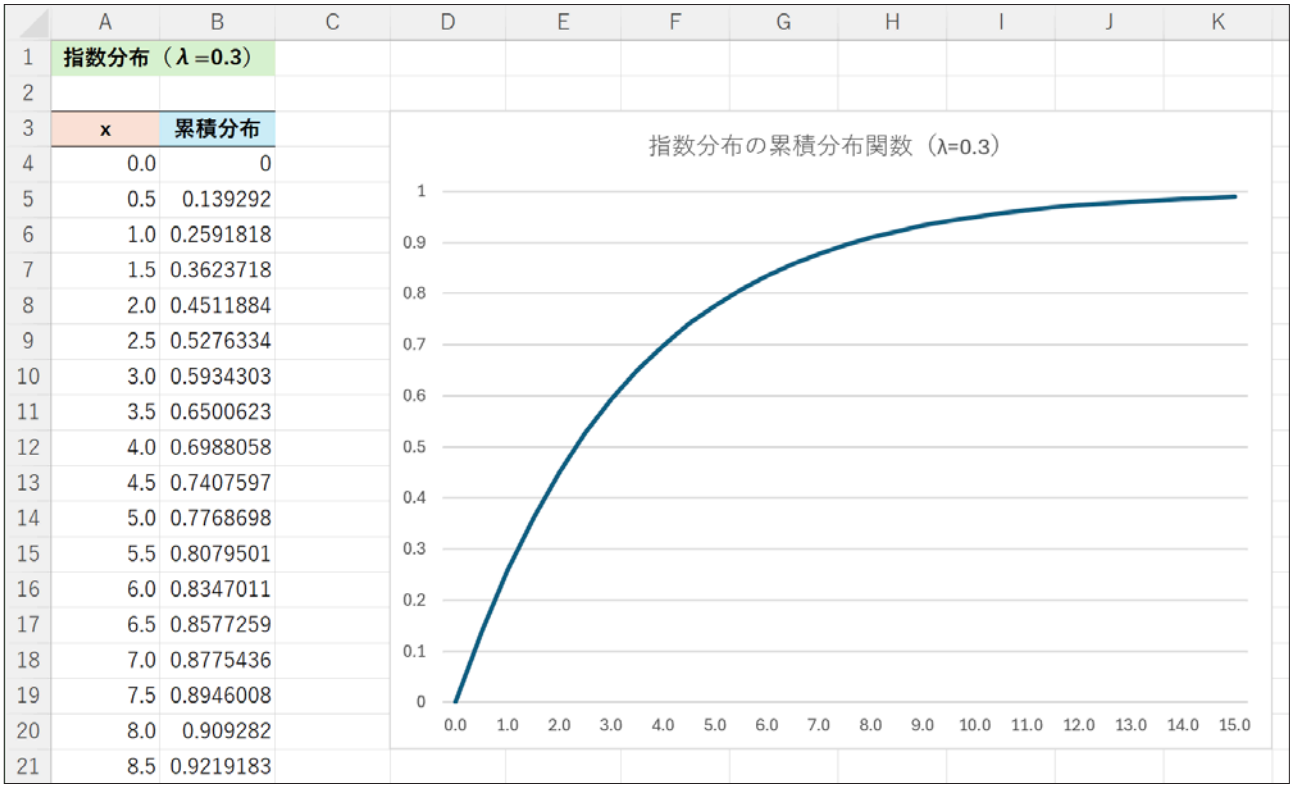
- セル **B5** に `=ARRAYFORMULA(EXPON.DIST(A5:A35,B4:E4,FALSE))` と入力する

● グラフの作成方法

サンプルファイル内に掲載しておきます（タイトルや軸の書式などの細かい設定は省略）

指数分布ってどんな感じの分布（2）～累積分布関数を可視化してみよう

続いて、累積分布関数です。こちらは、$\lambda = 0.3$ の例だけを見ておきます（図 4）。［指数分布累積］ワークシートを開いて、図の後の手順で試してみてください。



	A	B
1	指数分布（λ=0.3）	
2		
3	x	累積分布
4	0.0	0
5	0.5	0.139292
6	1.0	0.2591818
7	1.5	0.3623718
8	2.0	0.4511884
9	2.5	0.5276334
10	3.0	0.5934303
11	3.5	0.6500623
12	4.0	0.6988058
13	4.5	0.7407597
14	5.0	0.7768698
15	5.5	0.8079501
16	6.0	0.8347011
17	6.5	0.8577259
18	7.0	0.8775436
19	7.5	0.8946008
20	8.0	0.909282
21	8.5	0.9219183

図 4　指数分布の累積分布関数の例

$\lambda = 0.3$ とし、$x = 0.0$ ～ 15.0 までの累積分布関数の値を **0.5** 刻みで求め、グラフを描いてみた。最初に見た顧客の例（$\lambda = 0.3$, $x = 5.0$ の値）は、セル **B14** で求められている（**0.7769** となっている）。

累積分布関数の値を求めるための手順は以下の通りです。ここでも、関数の入力にのみ焦点を当てることとし、グラフ作成の手順についてはサンプルファイル内に掲載しておきます。

◆ Excel での操作方法

- セル **B4** に =EXPON.DIST(A4:A34,0.3,TRUE) と入力する
  - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **B4** ～ **B34**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す

◆ Google スプレッドシートでの操作方法

- セル **B4** に =ARRAYFORMULA(EXPON.DIST(A4:A34,0.3,FALSE)) と入力する

● グラフの作成方法

- サンプルファイル内に掲載しておきます（タイトルや軸の書式などの細かい設定は省略）

指数分布では、確率変数を求めるための計算などもなく、また、累積分布関数が直接応用に結び付くので、比較的すんなりと理解できるのではないかと思います。ここからは、少し発展的なお話としてポアソン分布との関係や幾何分布との関係を見てみます。実用的には知らなくてもあまり問題がないので、数式が苦手な方は、最後の「この記事で取り上げた関数の形式」を確認して、今回はお開きとしてもらっても構いません。

コラム　指数分布とポアソン分布の関係

指数分布のお話を読みながら、何か前にやったことがあるような……という既視感にとらわれた方もいるのではないでしょうか。この連載の第 4 回で紹介したポアソン分布の確率質量関数に似ていますよね。100 年に 1 人の天才が今後 100 年間に何人か現れる確率を求めるのにポアソン分布の確率密度関数を使いました。具体例で比較してみましょう。

- ポアソン分布の確率質量関数の例：100 年に 1 人の天才が、今後 100 年間に k 人現れる確率はいくらか
- 指数分布の累積分布関数の例：100 年に 1 人の天才が、期間 x 内に少なくとも 1 人現れる確率はいくらか

一般化するために、**100 年**を単位時間として書き直してみましょう。単位時間当たりに起こる事象の数は λ と表せます（ここでは $\lambda = 1$ です）。

- ポアソン分布の確率質量関数の例：単位時間に λ 人の天才が、単位時間内に **k** 人現れる確率はいくらか
- 指数分布の累積分布関数の例：単位時間に λ 人の天才が、**期間** x 内に少なくとも 1 人現れる確率はいくらか

上の箇条書きで太字にした部分、つまり、ポアソン分布の事象の数を $k = 1$ とし、指数分布の期間を単位時間 $x = 1$ とすれば、どちらもほとんど同じです。が、重要な違いがあります。

それは、ポアソン分布では $k = 1$ 人現れる確率はいくらか、ということであるのに対し、指数分布では**少なくとも** 1 人現れる確率はいくらか、つまり 0 人でない確率はいくらかということです。……ということは、ポアソン分布で $k = 1$ の確率ではなく、$k = 0$ **でない確率**を求めれば、指数分布の累積分布関数と同じ値になるはずですね。

ポアソン分布の確率質量関数は以下の式で定義されます。

$$f(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

k = 0 でない確率は 1 − f(0) なので、

$$
\begin{aligned}
1 - f(0) &= 1 - \frac{\lambda^0 \times e^{-\lambda}}{0!} \\
&= 1 - \frac{1 \times e^{-\lambda}}{1} \\
&= 1 - e^{-\lambda} \qquad (3)
\end{aligned}
$$

一方、指数分布の累積分布関数は以下の通りでした。

$$F(x) = 1 - e^{-\lambda x}$$

x = 1 を代入すると、

$$F(x) = 1 - e^{-\lambda} \qquad (4)$$

となり、（3）式と（4）式がちゃんと一致します。上の例では、**100 年**を単位時間として（**100 年**を **1** として）、その期間に **1 人**の天才が現れるのでλ = 1 です。その場合、今後 100 年間に少なくとも **1 人**の天才が現れる確率は、

$$
\begin{aligned}
1 - e^{-\lambda} &= 1 - e^{-1} \\
&\approx 0.632
\end{aligned}
$$

となります。

というわけで、λが等しく、x が単位時間と等しい場合、次の単位時間内に少なくとも **1** 回の事象が起こる確率は、ポアソン分布の確率質量関数を使っても、指数分布の累積分布関数を使っても求められるというわけです。なお、上の事例を EXPON.DIST 関数と POISSION.DIST 関数で計算した例、定義に従って計算した例の両方を［指数分布とポアソン分布］ワークシートに含めてあるので、興味のある方はご参照ください。

コラム　指数分布と幾何分布の関係

既視感と言えば、指数分布は第 5 回で紹介した幾何分布の累積分布関数にも似ているような気がしませんか。幾何分布の累積分布関数 F(k) は、k 回目までに目的の事象が起こる確率を表します。

$$F(k) = 1 - (1 - p)^k \qquad (5)$$

連載の第 5 回では、当選確率が p = 1/4 のライブチケットの抽選で、k = 3 回目に当選する確率を求めるのに幾何分布の確率質量関数を利用しました。k = 3 回目**まで**に当選する確率を求めるなら、幾何分布の累積分布関数を利用すればいいですね。（5）式に p と k の値を代入すると、

$$F(3) = 1 - (1 - 1/4)^3$$
$$= 0.5781$$

となります。

上の例は、指数分布では$\lambda = 1/4$、x = 3 となります（**4 回**に **1 回**当選する抽選に、**3 回**のうち少なくとも 1 回当選する確率ということですね）。既に見た、指数関数の累積分布関数の（2）式に当てはめると、

$$F(x) = 1 - e^{-\lambda x}$$
$$= 1 - (2.71828)^{-1/4 \times 3}$$
$$= 0.5276$$

となります。違いが生じるのは、指数分布が、幾何分布の回数を細かく分けた場合の近似値となっているからです。ここでは回数をあまり細かく分けていないので、違いが大きくなっていますが、p やλを小さくし、k や x を大きくすれば、ほぼ等しくなります。

上の事例を EXPON.DIST 関数と NEGBINOM.DIST 関数で計算した例、定義に従って計算した例の両方を［指数分布と幾何分布］ワークシートに含めてあるので、興味のある方はご参照ください。例えば、p やλの値として **0.01**、k や x の値として **100** などを指定すると、結果はほぼ一致します。

Excel には幾何分布の確率質量関数や累積分布関数を求めるための関数がありませんが、負の二項分布の確率質量関数や累積分布関数の値を求める NEGBINOM.DIST 関数で、成功数に **1** を、失敗数に **k-1** を指定すれば求められます。

指数分布は連続型確率分布で、幾何分布は離散型確率分布です。幾何分布の試行を細かく分けてしていくと指数分布になります。ちょっと面倒ですが、定義から確認してみましょう。

試行を細かく分けると個々の試行の確率は小さくなります。例えば n 個に分けると、個々の確率λは 1/n 倍になるので、

$$p = \frac{\lambda}{n}$$

となり、試行回数は n 倍になるので、

$$k = nx$$

と表せます。幾何分布の累積分布関数を表す（5）式にこれらを代入し、n を無限大に近づけると、

$$\lim_{n \to \infty} (1 - (1 - p)^k) = 1 - \lim_{n \to \infty} \left(1 - \frac{\lambda}{n}\right)^{nx} \qquad (6)$$

となります。ここで、以下のような、e のべき乗の値を求める公式を適用します（この公式の証明は省略します）。

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda} \qquad (7)$$

以下の [] で囲んだ部分は（7）式に示した通り、$e^{-\lambda}$に収束します。そこで、（6）式に（7）式を代入すれば、指数分布の累積分布関数である（2）式になります。

$$\begin{aligned}\lim_{n\to\infty}(1-(1-p)^k) &= 1-\left[\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n\right]^x \\ &= 1-e^{-\lambda x}\end{aligned}$$

◇ ◇ ◇ ◇ ◇ ◇ ◇

今回は指数分布について、具体的な利用例を基に考え方や、確率密度関数と累積分布関数の求め方などについてお話ししました。広く応用できそうだということもご理解いただけたと思います。

次回は、一定時間内に事象が少なくとも何回か起こる確率を求める場合などに使われるガンマ分布についてお話しします。次回もお楽しみに！

この記事で取り上げた関数の形式

関数の利用例については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

指数分布の確率密度関数や累積分布関数の値を求めるための関数

EXPON.DIST 関数：指数分布の確率密度関数や累積分布関数の値を求める

形式

EXPON.DIST(x, λ , 関数形式)

引数

x：目的の時間間隔（確率変数）の値を指定する。

λ：単位時間に起こる事象の数を指定する。

関数形式：以下の値を指定する。

- **FALSE** …… 確率密度関数の値を求める
- **TRUE** …… 累積分布関数の値を求める

［データ分析］ガンマ分布とアーラン分布 ～ 5 分以内に 2 匹以上の猫が通る確率は？

データ分析の初歩から学んでいく連載（確率分布編）の第 11 回。ガンマ分布やアーラン分布は、待ち行列の分析などに使われる分布です。ある事象が起こる平均の間隔が分かっているときに、ある期間内にその事象が何回か以上起こる確率が求められます。今回は具体例を基に、その確率を求め、ガンマ分布の確率密度関数や累積分布関数の形を見ていきます。

羽山博（2024 年 11 月 21 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の確率分布編、第 11 回です。前回は、待ち行列の分析などに使われる指数分布を取り上げました。今回も同様に待ち行列の分析などに使われるガンマ分布について、その特徴や意味を基本から解き明かし、確率密度関数／累積分布関数の求め方、利用例などを見ていきます。

一定時間内に何匹か以上の猫が通る確率は？ ～ ガンマ分布とアーラン分布の利用

ガンマ分布は、ある事象（出来事）が起こる平均の間隔βが分かっているときに、ある期間 x 内にその事象がα回以上起こる確率を表す分布です。ガンマ分布の母数αが正の整数である場合は、特に**アーラン分布**と呼ばれます。

少し前のことですが、筆者の仕事場の前が野良猫の通り道になっていて、仕事をしているとよく猫が目の前を横切りました。このことを例に、ガンマ分布がどのように適用できるか、具体例を見てみましょう（図 1）。

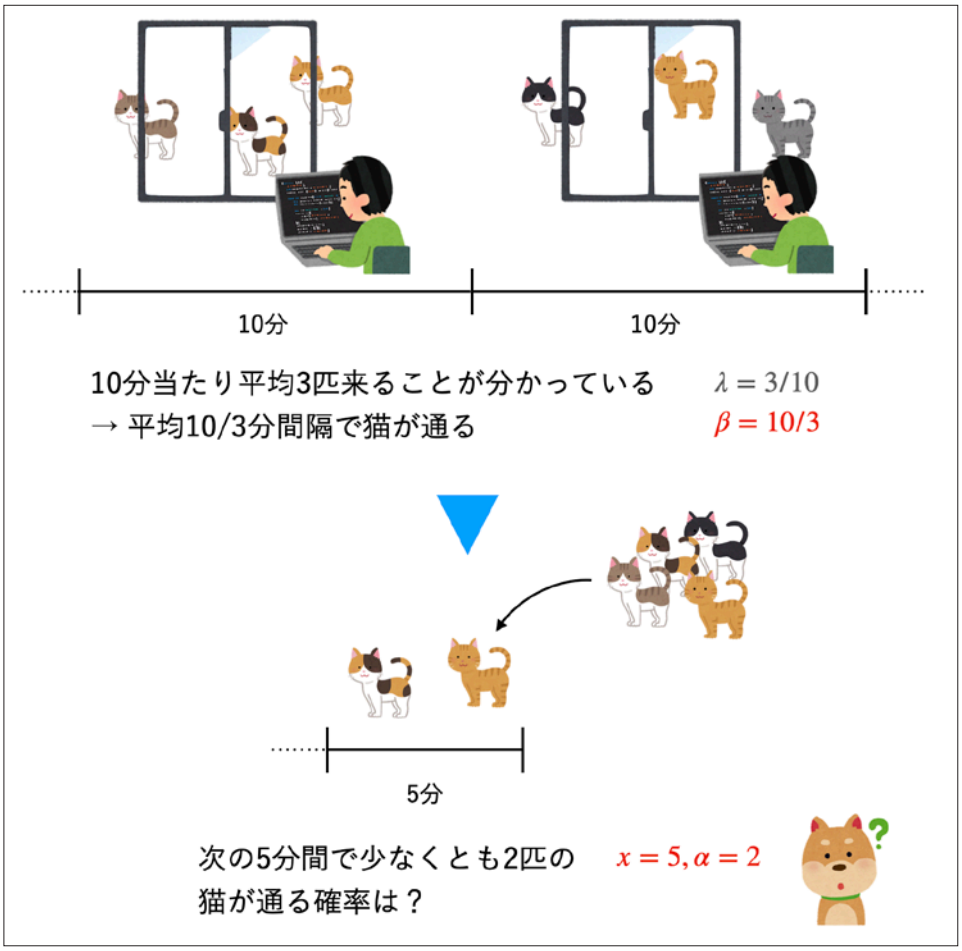


図 1　ある期間に何匹か以上の猫が通る確率を知りたい！

ガンマ分布の累積分布関数は、事象が起こる平均の間隔βが分かっているとき、目的の期間 x 内にその事象が少なくともα回起こる（＝α回以上起こる）確率を求めるのに使う。

図 1 に示された問題意識は、**10 分**当たり平均 **3 匹**の猫が通ることが分かっている場合に、次の **5 分間**に少なくとも **2 匹**の猫が通る（2 匹以上の猫が通る）確率を求めたい、ということです。そんなことをして何の役に立つのか、と思われるかもしれませんが、「猫」を「顧客」や「車」に置き換えると、一定時間内に何人かの顧客が到着する確率や一定時間内に車が何台か以上通る確率を求めるということになり、がぜん、実用性が感じられるようになると思います（車の通行量の例についても、後ほど見ることにします）。

ところで、図 1 で「猫」を「顧客」に置き換えると、前回見た指数分布とそっくりですね。でも違いがあります。まずは指数分布との違いを確認するところから始めましょう。

指数分布とガンマ分布との違いを具体的かつ端的に言うと、指数分布の累積分布関数は、ある期間内に少なくとも **1 回**の事象が起こる（**1** 回以上起こる）確率を求めるのに使いますが、ガンマ分布の累積分布関数は、ある期間内に少なくとも**α回**の事象が起こる（α回以上起こる）確率を求めるために使われます。

ガンマ分布で$\alpha = 1$の場合、指数分布になります。また、ガンマ分布は、独立した指数分布の和と考えられます。具体的にどういうことなのかは、最後のコラムで解説します。

ガンマ分布の母数はαとβで、指数分布の母数はλです。これらを表にまとめて比較してみます（表 1）。いずれも確率変数は目的の期間です。

		ガンマ分布	指数分布
母数	目的の期間内で事象が何回以上起こるか	α	-
	一定期間内で事象が起こる平均の間隔	β	-
	一定期間内で事象が起こる平均の回数	-	λ
確率変数	目的の期間	x	x

表 1　ガンマ分布と指数分布の比較

- は指定する必要がないことを意味する。図 1 の例であれば、事象は「猫が通ること」に当たる。従って「目的の期間内に事象が何回以上起こるか」は「目的の時間内に猫が何匹以上通るか」と読み替えられる。これがαの値となる。図 1 の例であれば$\alpha = 2$。また、「一定期間内で事象が起こる平均の間隔」は「猫が通る平均の時間間隔」となる。これがβの値。図 1 の例なら **10/3**（分）。

指数分布の母数λは「一定期間内で事象が起こる回数」です。一方、ガンマ分布の母数βは「一定期間内で事象が起こる平均の間隔」なので、λの逆数となります。つまり、$\beta = 1/\lambda$となります。例えば、**10 分**間に **3 匹**の猫が通るのであれば、$\lambda = 3/10$ です。この場合、

$$\begin{aligned}\beta &= 1/\lambda \\ &= 10/3\end{aligned}$$

となります。**10 分**間に **3 匹**の猫が通るということは、平均して **10/3 分**間に **1 匹**の猫が通るということですね。

指数分布と比較しながらガンマ分布の意味が確認できたので、ガンマ分布の確率密度関数と累積分布関数の定義を見た後、図 1 の確率を計算してみましょう。

ガンマ分布の確率密度関数と累積分布関数

以下に示した式がガンマ分布の確率密度関数と累積分布関数の定義です。e は自然対数の底（＝ **2.71828**...）です。かなり複雑な式ですが、例によって、これらの式を覚える必要は全くありません。Excel の GAMMA.DIST 関数を使えば、簡単に答えが求められます。

◆ ガンマ分布の確率密度関数

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} \qquad (1)$$

◆ ガンマ分布の累積分布関数

$$F(x) = \frac{\gamma(\alpha, \frac{x}{\beta})}{\Gamma\alpha} \qquad (2)$$

ただし、Γはガンマ関数、γは下側不完全ガンマ関数（第一種不完全ガンマ関数）です。これらの関数の定義は、第 7 回のコラムで紹介しました（繰り返しになりますが、それらの定義を覚えたりする必要はありません）。

GAMMA.DIST 関数の書き方は以下の通りです。

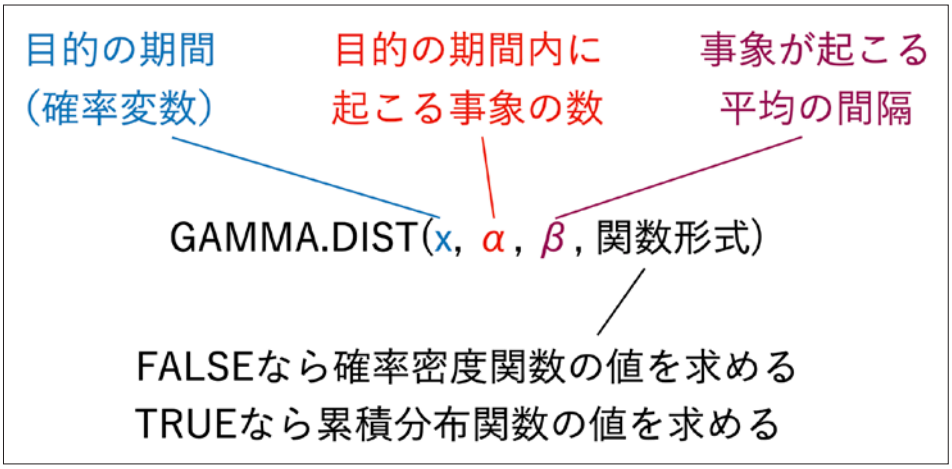


図 2　GAMMA.DIST 関数に指定する引数

GAMMA.DIST 関数には、目的の期間 x と、何回以上その事象が起こるかを表すα、事象が起こる平均の間隔βを指定する（βは、あらかじめ分かっている値）。関数形式についてはこれまで見てきた関数と同様、**FALSE** を指定すれば確率密度関数の値が、**TRUE** を指定すれば累積分布関数の値が求められる。

冒頭の図 1 で紹介した例について、GAMMA.DIST 関数を使って確率を求めてみましょう。**10 分**間に平均して **3 匹**の猫が通る＝平均して **10/3** 分間に **1 匹**の猫が通ることが分かっている場合に、次の **5 分**以内に **2 匹**以上の猫が通る確率を求めます。この場合、確率変数 x の値は **5** で、$\alpha = 2, \beta = 10/3$ となります。

サンプルファイルをこちらからダウンロードし、［ガンマ分布の利用］ワークシートを開いて試してみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。

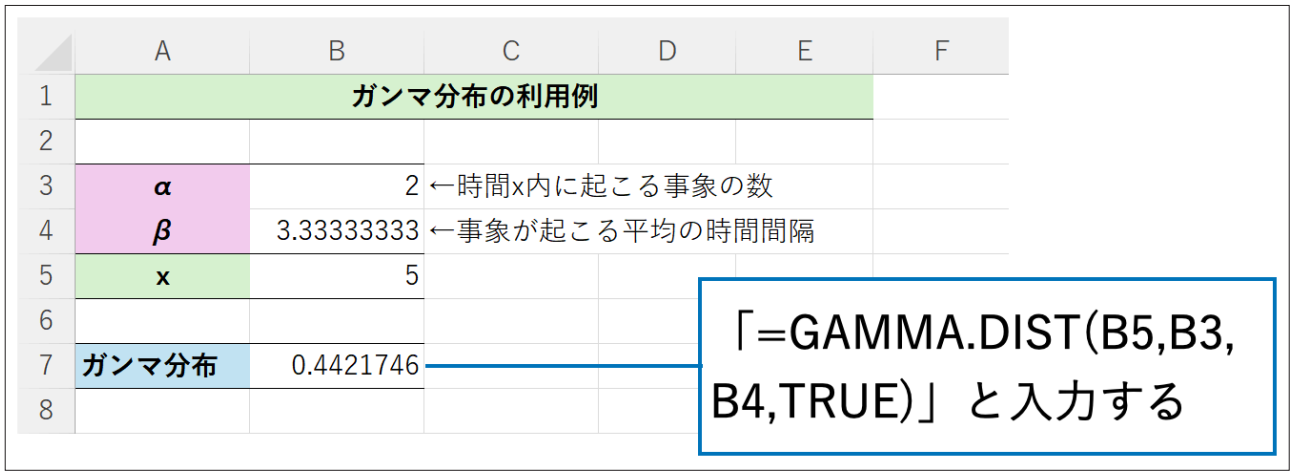


図 3　GAMMA.DIST 関数の利用例

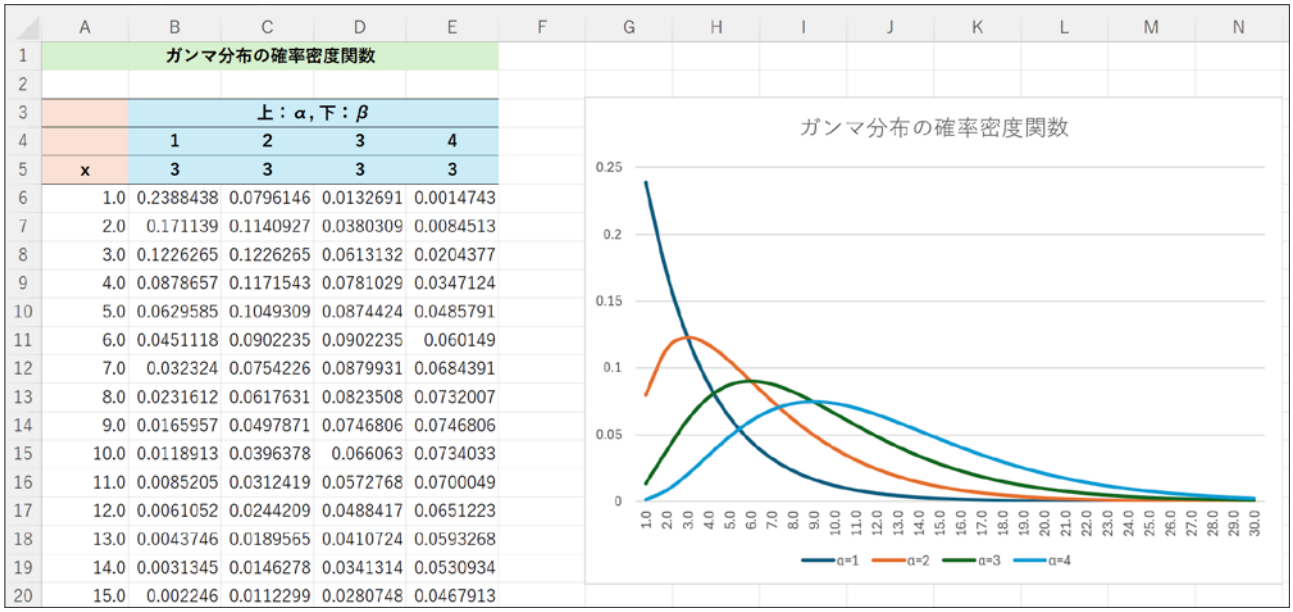
GAMMA.DIST 関数には x の値、αの値、βの値を順に指定する。ここでは累積分布関数の値を求めるので、最後の引数には **TRUE** を指定する。よってセル **B7** に =GAMMA.DIST(B5,B3,B4,TRUE) と入力すればよい。猫が通る平均の時間間隔が$\beta = 10/3$ 分であるとき、x ＝ 5 分以内に$\alpha = 2$ 匹以上の猫が通る確率は **0.442** であることが分かる。

私たちは、無意識のうちに「期間」を分単位や年単位などで考えてしまいます。例えば、**10 分**であれば、期間は **10** であると考えるのが普通です。しかし、ここでの期間とは一定の間隔のことです。問題に合わせて、例えば、**10 分**を 1 つの期間と考えたり、**100 年**を 1 つの期間と考えたりしても構いません。例えば、上の例で **10 分**を 1 つの期間と考えれば、1 つの期間に猫が **3 匹**通ることになります。その場合、$\lambda = 3$ なので、$\beta = 1/3$ となります。目的の間隔は **5 分**なので、1 つの期間とした **10 分**の半分、つまり **1/2** が x の値となります。αについては **2** のままですね。空いているセルに「=GAMMA.DIST(1/2, 2, 1/3, TRUE)」と入力してみてください。図 3 の結果と一致します。

続いて、ガンマ分布の確率密度関数と累積分布関数を可視化してみましょう。

ガンマ分布ってどんな感じの分布（1）～ 確率密度関数を可視化してみよう

上で見た GAMMA.DIST 関数の引数として母数α , βに幾つかの値を指定し、x の値を変化させていったグラフを描けば、ガンマ分布の確率密度関数や累積分布関数を可視化できます。ただし、以下の例ではβ＝ 3 に固定するものとします。ガンマ分布の台（確率変数が取り得る値の範囲）は 0 ～∞ですが、α＝ 1, x ＝ 0 のときには GAMMA.DIST 関数がエラーとなるので、ここでは x ＝ 1 ～ 30 までの値を求めることとします（図 **4**）。



	A	B	C	D	E
1	ガンマ分布の確率密度関数				
2					
3		上：α, 下：β			
4		1	2	3	4
5	x	3	3	3	3
6	1.0	0.2388438	0.0796146	0.0132691	0.0014743
7	2.0	0.171139	0.1140927	0.0380309	0.0084513
8	3.0	0.1226265	0.1226265	0.0613132	0.0204377
9	4.0	0.0878657	0.1171543	0.0781029	0.0347124
10	5.0	0.0629585	0.1049309	0.0874424	0.0485791
11	6.0	0.0451118	0.0902235	0.0902235	0.060149
12	7.0	0.032324	0.0754226	0.0879931	0.0684391
13	8.0	0.0231612	0.0617631	0.0823508	0.0732007
14	9.0	0.0165957	0.0497871	0.0746806	0.0746806
15	10.0	0.0118913	0.0396378	0.066063	0.0734033
16	11.0	0.0085205	0.0312419	0.0572768	0.0700049
17	12.0	0.0061052	0.0244209	0.0488417	0.0651223
18	13.0	0.0043746	0.0189565	0.0410724	0.0593268
19	14.0	0.0031345	0.0146278	0.0341314	0.0530934
20	15.0	0.002246	0.0112299	0.0280748	0.0467913

図 4　ガンマ分布の確率密度関数の例

αは 1, 2, 3, 4 とし、βは **3** に固定。x ＝ 1.0 ～ 30.0 までの確率密度関数の値を **1** 刻みで求めてグラフを描いてみた。x と表記されている **A** 列の値が確率変数（確率変数は整数でなくてもよい）。**B** ～ **E** 列の 6 行目以降はそれぞれのα , βでの x に対する確率密度関数の値。

確率密度関数の値を求めるための手順は以下の通りです。サンプルファイルの［ガンマ分布］ワークシートを開いて試してみてください。可視化については単に折れ線グラフを描くだけなので、関数の入力のみに焦点を当てることにします。グラフ作成の手順についてはサンプルファイル内に掲載しておきます。

◆ Excel での操作方法

- セル **B6** に =GAMMA.DIST(A6:A35,B4:E4,B5:E5,FALSE) と入力する
  - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **B6** ～ **E35**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す

◆ Google スプレッドシートでの操作方法

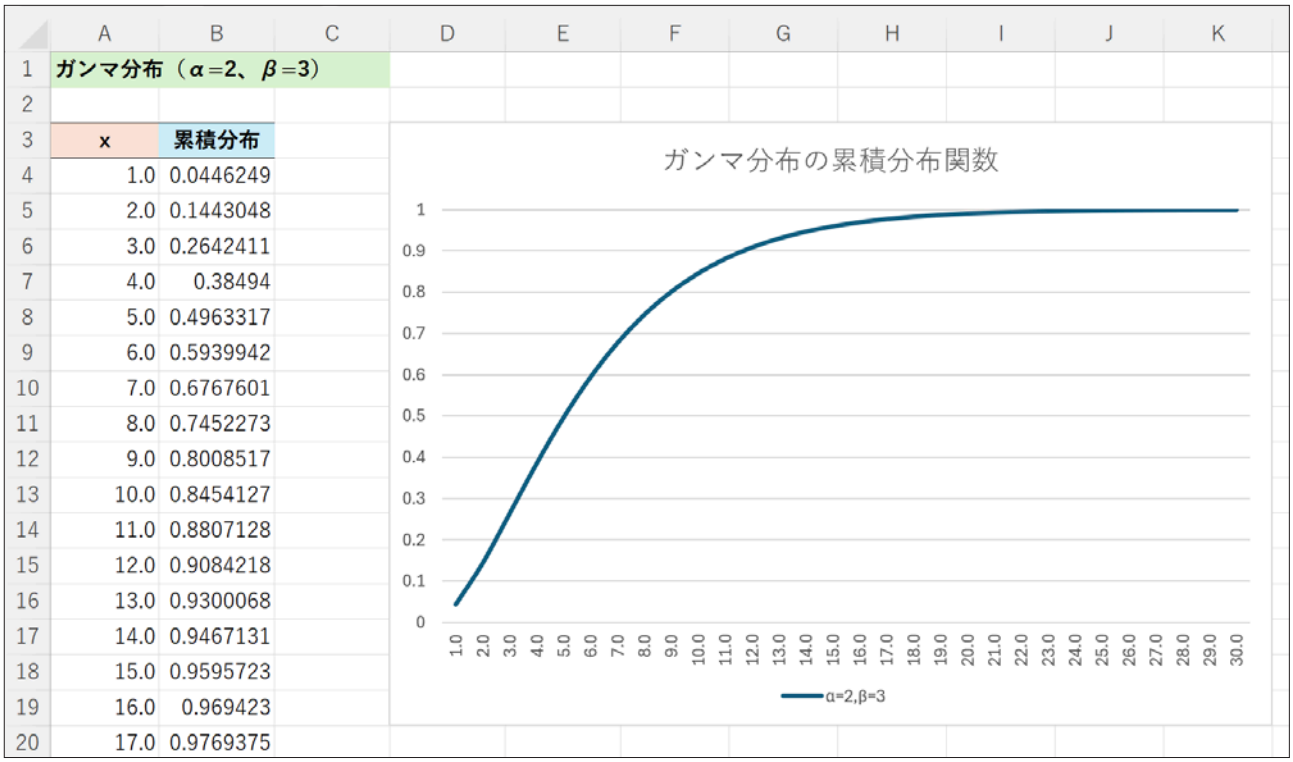
- セル **B6** に =ARRAYFORMULA(GAMMA.DIST(A6:A35,B4:E4,B5:E5,FALSE)) と入力する

● グラフの作成方法

- サンプルファイル内に掲載しておきます（タイトルや軸の書式などの細かい設定は省略）

ガンマ分布ってどんな感じの分布（2）～ 累積分布関数を可視化してみよう

続いて、累積分布関数です。こちらは、$\alpha = 2, \beta = 3$ の例だけを見ておきます（図 5）。［ガンマ分布累積］ワークシートを開いて、図の後の手順で試してみてください。



	A	B	C	D	E	F	G	H	I	J	K
1	ガンマ分布（α=2、β=3）										
2											
3	x	累積分布									
4	1.0	0.0446249									
5	2.0	0.1443048									
6	3.0	0.2642411									
7	4.0	0.38494									
8	5.0	0.4963317									
9	6.0	0.5939942									
10	7.0	0.6767601									
11	8.0	0.7452273									
12	9.0	0.8008517									
13	10.0	0.8454127									
14	11.0	0.8807128									
15	12.0	0.9084218									
16	13.0	0.9300068									
17	14.0	0.9467131									
18	15.0	0.9595723									
19	16.0	0.969423									
20	17.0	0.9769375									

図 5　ガンマ分布の累積分布関数の例

α を **2**、β を **3** として、x = 1.0 ～ 30.0 までの累積分布関数の値を **1** 刻みで求めてグラフを描いてみた。x = 20 になると累積確率は **99%** を超える。例えば、猫が通る間隔が平均 **3 分**であるとき、**20 分**以内にはほぼ確実に **2 匹**（以上）の猫が通ることが分かる。

累積分布関数の値を求めるための手順は以下の通りです。ここでも、関数の入力のみに焦点を当てることとし、グラフ作成の手順についてはサンプルファイル内に掲載しておきます。

◆ Excel での操作方法

- セル **B4** に =GAMMA.DIST(A4:A33,2,3,TRUE) と入力する
  - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **B4** ～ **B33**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す

◆ Google スプレッドシートでの操作方法

- セル **B4** に =ARRAYFORMULA(GAMMA.DIST(A4:A33,2,3,TRUE)) と入力する

● グラフの作成方法

- サンプルファイル内に掲載しておきます（タイトルや軸の書式などの細かい設定は省略）

ガンマ分布も指数分布と同様、確率変数を求めるための計算などがなく、また、累積分布関数が直接応用に結び付くので、比較的すんなりと理解できるのではないかと思います。ここからは、少し発展的なお話としてガンマ分布とポアソン分布との関係を見てみます。また、ガンマ分布が指数分布の和であるということをシミュレーションで確認します。実用的には知らなくてもあまり問題がないので、数式が苦手な方や変数の対応関係を細かく確認するのが面倒だと思われる方は、最後の「この記事で取り上げた関数の形式」を確認して、今回はお開きとしてもらっても構いません。

コラム　ガンマ分布と指数分布、ポアソン分布の関係

前回、指数分布とポアソン分布の関係をコラムで解説しました。ガンマ分布は指数分布の和である（詳細は次のコラムで見ます）ということなので、ガンマ分布と指数分布、ポアソン分布が互いに関係しているということも想像できると思います。そこで、それらの関係を図で表し、その後、もう一度意味を考えていくことにします。

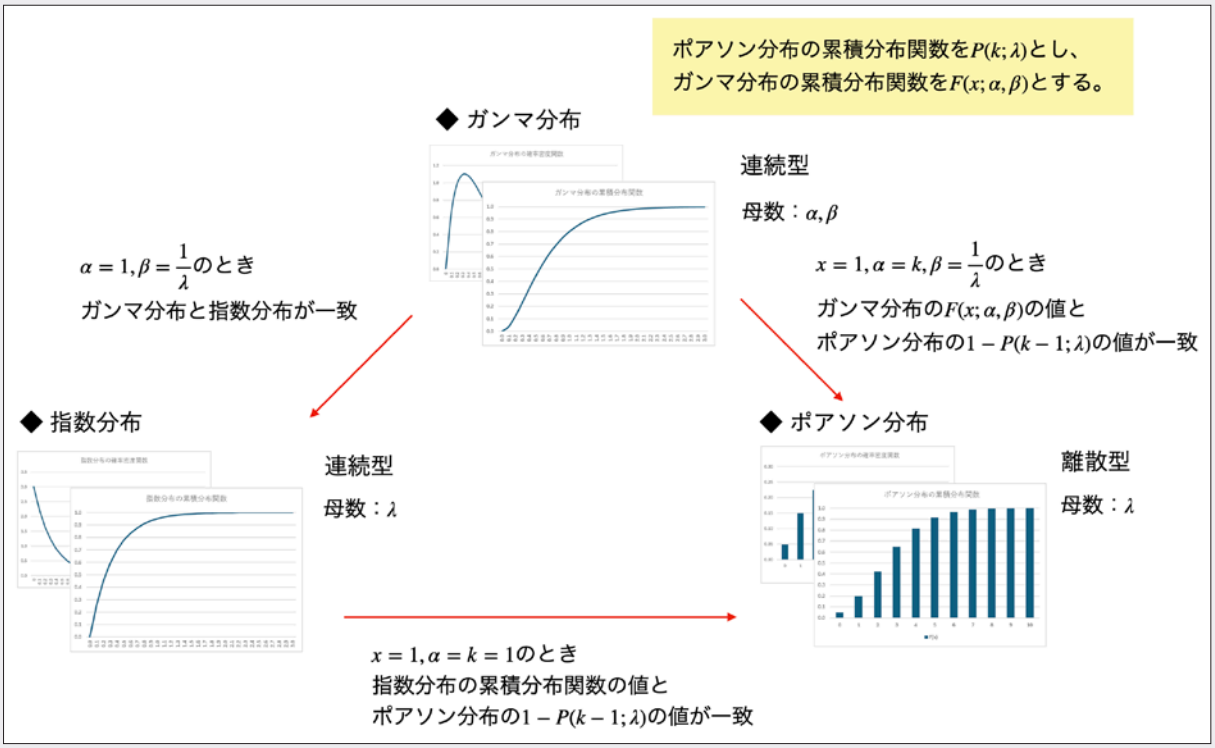


図 6　ガンマ分布と指数分布、ポアソン分布の関係

$\alpha=1$, $\beta=1/\lambda$のとき、ガンマ分布と指数分布は一致する。また、ポアソン分布の累積分布関数を $P(k;\lambda)$ とし、$x=1$, $\alpha=k$, $\beta=1/\lambda$としたとき、ガンマ分布の累積分布関数 $F(x,\alpha,\beta)$ は、$1-P(k-1;\lambda)$ と一致する。詳細についてはこの後解説する。

→目次に戻る

ガンマ分布の累積分布関数とポアソン分布の累積分布関数の違いを端的に言うなら、以下のようになります。

- ガンマ分布の母数βは一定期間内に事象が起こる平均の間隔であるが、ポアソン分布の母数λは一定期間内に事象が起こる回数（いずれもあらかじめ分かっている値）…… 従って、$\beta = 1/\lambda$となる
- ガンマ分布では目的の期間が確率変数 x となるが、ポアソン分布では事象が起こる期間（あらかじめ分かっている期間）と目的の期間は同じなので、式の中には現れない …… ガンマ分布で $x = 1$ とすると、ポアソン分布の期間と同じになる
- ガンマ分布の累積分布関数では、目的の期間に事象がα回**以上**起こる確率が求められるのに対し、ポアソン分布の累積分布関数では、目的の期間に事象が k 回**まで**起こる確率が求められる。

これらの関係を数式で解き明かすのはかなり難しいので、ここでは、図 6 や上に示した箇条書きの意味を具体的な例で考えてみましょう。猫の事例よりも実用的な例として、車の通行量で考えることにします（さらに言えば、コールセンターに入る電話の数や機器の故障の数などでも基本的な考え方は同じです）。理屈よりも実際に計算した例を先に見たいという方は、こちらまで読み飛ばしていただいても構いません。

離散型確率分布であるポアソン分布が分かりやすいので、ポアソン分布を中心に見ていくことにしましょう。以下、母数λの表記は省略して、ポアソン分布の確率質量関数を $p(k)$、累積分布関数を $P(k)$ と簡単に表すことにします。例えば、**1 分**間に平均して$\lambda = 3$ 台の車が通ることが分かっている場合、**1 分**間に車が、

- $k = 5$ 台通る確率は、確率質量関数 $p(k = 5)$ の値 ……（**1**）
- $k = 5$ 台**まで**通る確率は、累積分布関数 $P(k = 5)$ の値 ……（**2**）

で求められます。従って、**1 分**間に車が

- $k = 5$ 台**以上**通る確率は $1 - P(k - 1 = 4)$ の値 ……（**3**）

となります（全体から $k - 1$ 台まで通る確率を引けばよい）。

一方、ガンマ分布の累積分布関数では、車が通る平均の間隔をβとするので、$\beta = 1/\lambda = 1/3$ です。

- 期間 x 内に車が$\alpha = 5$ 台**以上**通る確率は、累積分布関数 $F(x; \alpha, \beta)$ の値

ということは、ガンマ分布で $x = 1$, $\alpha = k$, $\beta = 1/\lambda$としたときの累積分布関数の値 $F(1; \alpha, \beta)$ と、ポアソン分布での（3）の計算、つまり $1 - P(k - 1)$ の値が同じになるはずです。

また、指数分布の累積分布関数の値は、ガンマ分布で$\alpha = 1$, $\beta = 1/\lambda$を指定した場合の累積分布関数の値や、ポアソン分布で $k = 1$ を指定した場合の $1 - P(k - 1)$ の値と同じになります。

では、具体的な値を当てはめて確認してみましょう。［分布同士の関係］ワークシートを開いて、以下のように入力すると同じ結果が得られることが分かります。

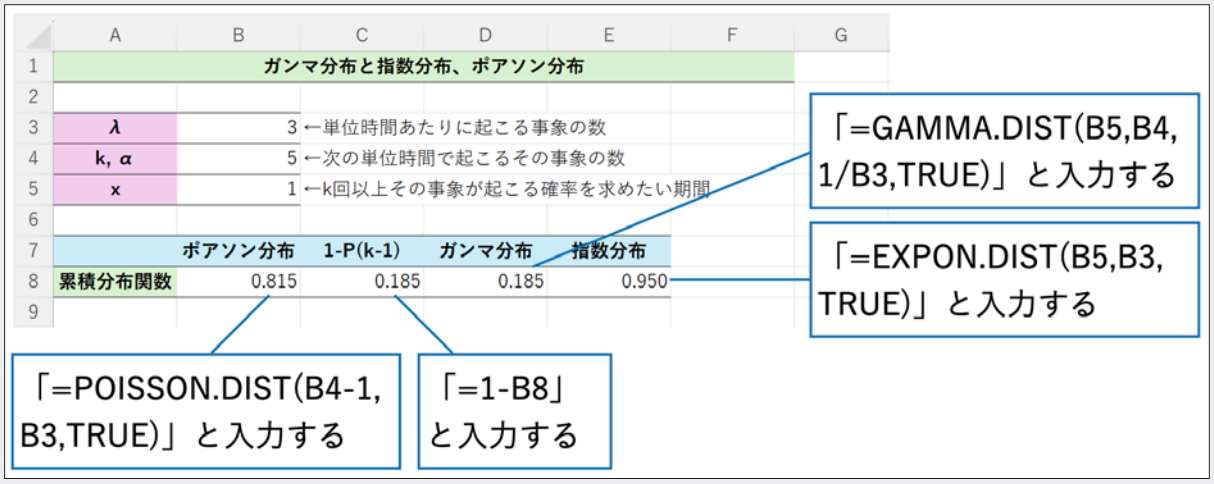


図 7　ガンマ分布と指数分布、ポアソン分布の関係を確認する

セル **B8** の POISSON.DIST 関数では、ポアソン分布の k − 1 に当たる値として B4-1 を指定して累積分布関数の値を求めている（k − 1 回までの確率が求められる）。
セル **C8** では **1** からその値を引いて、事象が k 回以上起こる確率を求める。
一方、セル **D8** の GAMMA.DIST 関数では、時間間隔 x に当たる値としてセル **B5** を指定し、αに当たる値としてセル **B4** を、$\beta = 1/\lambda$に当たる値として 1/B3 を指定している。この場合、セル **B5** に **1** を入力すると、x = 1 となるので、セル **C8** とセル **D8** の値が一致する。
さらに、セル **B4** の値が **1** のとき、k = α = 1 なので、セル **C8**、**D8** の値とセル **E8** の EXPON.DIST 関数の値が一致する（セル **B4** に **1** を入力して試してみるとよい）。

セル **B5** に入力された x の値が **1** であれば、ポアソン分布の 1 − P(k − 1) の値に当たるセル **C8** の値と、ガンマ分布の $F(x; \alpha, \beta)$ の値に当たるセル **D8** の値が一致します。さらに、セル **B4** に入力された k, αの値が **1** であれば、指数分布の累積分布関数の値とも一致します。セル **B3** ～ **B5** にいろいろな値を入力して確かめてみてください。

なお、参考として、サンプルファイルの［分布同士の関係 (詳細版)］ワークシートに、条件付き書式を使ってそれぞれの分布の関係を見やすくした例を含めてあります。k, αの値によって、どの値が同じになるかが視覚的に確認できます。ぜひ試してみてください。

コラム　ガンマ分布が指数分布の和ってどういうこと？

「ガンマ分布は、独立した指数分布の和である」というお話が何回か登場しました。それはいったいどういうことでしょうか。これについても、数式で解き明かすのはやや難しいので、Python のプログラムを使ってシミュレーションしてみましょう。

サンプルプログラムはこちらから参照できます。リンクをクリックすれば、ブラウザが起動し、Google Colaboratory の画面が表示されます（Google アカウントでのログインが必要です）。最初のコードセルをクリックし、［Shift］＋［Enter］キーを押してコードを実行してみてください。結果は図 8 のようになります。コードの詳細については解説しませんが、コメントとリスト 1 の説明を見れば何をやっているかが大体分かると思います。

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import expon, gamma

p = 3/10  # 10 分間に平均 3 匹の猫が通るものとする

# 2 つの指数分布からランダムにサンプルを抽出
sample1 = expon.rvs(scale=p, size=10000)
sample2 = expon.rvs(scale=p, size=10000)

# 指数分布の和を求める
sample_sum = sample1 + sample2

# 指数分布の和をヒストグラムにする
plt.hist(sample_sum, bins=100, density=True)

# ガンマ分布の確率密度関数を描く
x = np.linspace(0, 6, 100)  # 0 ～ 6 を 100 個に分けた値
gamma_pdf = gamma.pdf(x, a=2, scale=p)  # α =2 のガンマ分布
plt.plot(x, gamma_pdf)
plt.show()
```

リスト 1　指数分布の和とガンマ分布の関係をシミュレーションする

$\lambda = 3$ での例（コードではλの値を p という変数に代入している）。
expon.rvs 関数は指数分布からランダムにサンプルを抽出するための関数。引数 scale にはλの値を指定する。expon.rvs 関数を使って抽出した **10000 個**のサンプルを 2 組作り、それぞれの値の和を求めた sample_sum を基にヒストグラムを作成する。
次に、gamma.pdf 関数を使ってガンマ分布の確率密度関数の値を求める。確率変数 x の値は **0** ～ **6** までを **100** 個に分割した値。引数 a がガンマ分布の母数αに当たる。これには、独立した指数分布の個数である **2** を指定する。gamma.pdf 関数では、引数 scale にβの値ではなくλの値をそのまま指定する。
このようにして求められた確率密度関数の値の並びを折れ線グラフにし、確率密度関数のグラフを描く。

実行結果は以下の通りです（図 8）。独立した指数分布の和がガンマ分布と一致することが分かります。

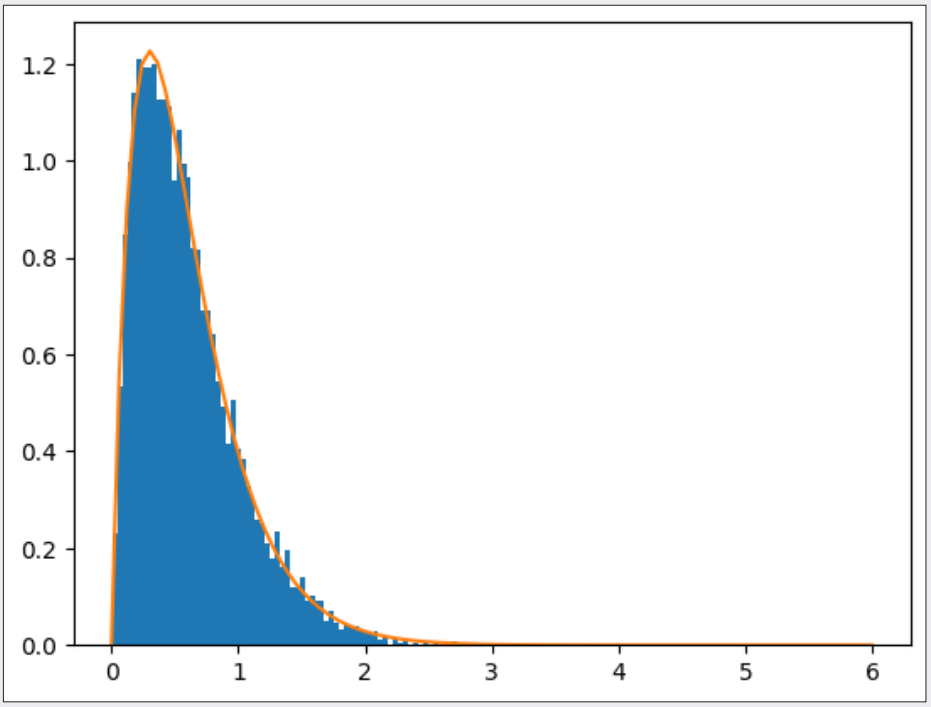


図 8　指数分布の和をヒストグラムにしたものとガンマ分布の確率密度関数

独立した指数分布の和を求め、それをヒストグラムにすると、ガンマ分布の確率密度関数と一致する。

◇ ◇ ◇ ◇ ◇ ◇ ◇

今回はガンマ分布について、指数分布との違いを見た後、具体的な利用例を基に考え方や、確率密度関数と累積分布関数の求め方などについてお話ししました。また、ポアソン分布との関係についても解説しました。ガンマ分布も指数分布と同様、広く応用できそうだということもご理解いただけたと思います。

次回は、いわゆる A/B テスト（2 つの Web サイトのデザインを作成し、どちらがより購買に結び付くかを分析するなど）やベイズ統計の事前分布としてよく使われたりするベータ分布についてお話しします。次回もお楽しみに！

この記事で取り上げた関数の形式

関数の利用例については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

ガンマ分布の確率密度関数や累積分布関数の値を求めるための関数

GAMMA.DIST 関数：ガンマ分布の確率密度関数や累積分布関数の値を求める

形式

GAMMA.DIST(x, α , β , 関数形式)

引数

- **x**：目的の時間（確率変数）の値を指定する。
- **α**：目的の時間内に事象が何回以上起こるかを指定する。
- **β**：単位時間内で事象が起こる平均の間隔を指定する。
- **関数形式**：以下の値を指定する。
  - **FALSE** …… 確率密度関数の値を求める
  - **TRUE** …… 累積分布関数の値を求める

[データ分析] ベータ分布
～ 3 ポイントシュート成功率 100%に信ぴょう性はあるか？

データ分析の初歩から学んでいく連載（確率分布編）の第 12 回。ベータ分布は「確率の確率」とも呼ばれる分布です。ある事象の成功数と失敗数が分かっているときに、成功率が一定の範囲に入っている確率を求めるのに使われます。今回も具体例を基に、ベータ分布の利用例や、確率密度関数と累積分布関数の形を見ていきます。

羽山博（2024 年 12 月 12 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の確率分布編、第 12 回です。前回は、待ち行列の分析などに使われるガンマ分布を取り上げました。今回はどちらの方法が優れているかを知るための A/B テストや、ベイズ統計での事前分布としてよく使われるベータ分布について、その特徴や意味を基本から解き明かし、確率密度関数／累積分布関数の求め方、利用例などを見ていきます。

3 ポイントシュートの成功率はどれぐらいか ～ ベータ分布の利用

最近は、人気のアニメや、日本国内での B リーグの流行、オリンピックや米国の NBA（National Basketball Association）での日本人選手の活躍などでバスケットボール（以下、バスケ）がニュースに取り上げられることも多くなってきたので、バスケをよく知らない方でも「3 ポイントシュート」という言葉を耳にしたことのある人は多いと思います。バスケのフィールドゴールは通常 2 点ですが、3 ポイントラインの外側から放った場合には点数が 3 点入るというものです。遠くからシュートしないといけないので、失敗するリスクも高いですが、成功すると効果は抜群です。

筆者はバスケの素人ですが、3 ポイントシュートにチャレンジして、いきなり成功したとします。**1 回**シュートして **1 回**成功したので成功率は **100%**です。筆者としては鼻高々ですが、この **100%**という数字には信ぴょう性があるでしょうか。そんなのはたまたまだ、と軽くあしらわれるのがオチです。

では、プロの選手が何試合かで **25 回**中 **11 回**、3 ポイントシュートを決めたとしましょう。成功率は 11/25 = **44%**となりますが、その数字にはどれぐらいの信ぴょう性があるでしょうか（図 1）。

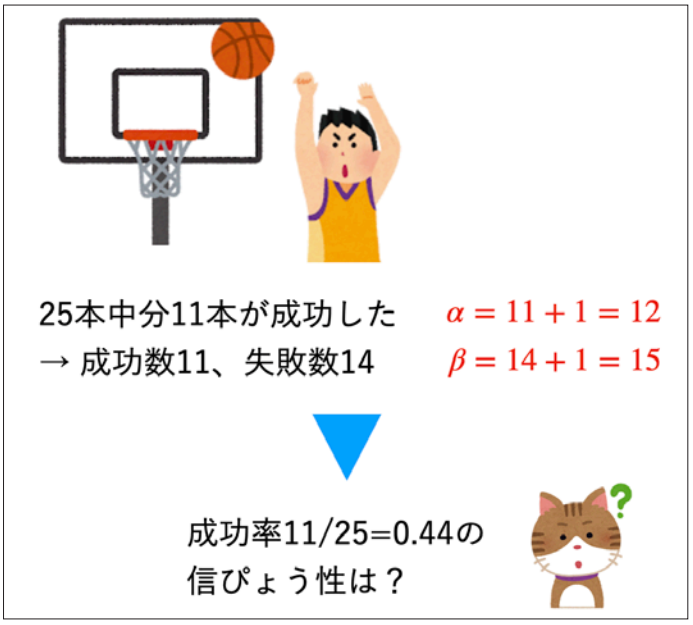


図 1　シュートの成功率にどの程度信ぴょう性があるかを知りたい！

これまでの成功数と失敗数が分かっているとき、単純計算した成功率ではなく、確率分布を基に、成功率がどのあたりの範囲にあるのかを知りたい。そのために使えるのがベータ分布。母数のαは成功数を、βは失敗数を表す。実際の数に **1** を足している理由は後述する。

ただし、ここで言う「信ぴょう性」とは、「ちょうど **44%**であることが、どれだけ信じられるか」という意味ではなく、**「成功率がどのように分布していて、その分布の中で " 真の成功率 " が特定の範囲に入っている確率はどのくらいか」**という意味です。例えば、観測されたデータ（**44%**）を基に、" 真の成功率 " が **40** ～ **50%**という特定の範囲に入っている確率はどの程度なのか、と評価することです。

もう少しくだけた言い方をすれば、何試合かの 3 ポイントシュート成功率は **44%**だったけれど、その選手の実力を表す " 真の成功率 " はどの範囲に（高い確率で）入っているか、ということです。

この問題に対するアプローチは、**頻度主義**と呼ばれる古典的な統計学の考え方とベイズ主義と呼ばれる考え方では異なります。今回は**ベイズ主義**の考え方にそって話を進めます。頻度主義とベイズ主義の違いについては後ほど簡単に説明することとして、取りあえず、基本的な考え方と計算の方法を続けて見ていくことにします。

プロの選手で何回かシュートを打った結果だから、真の成功率は **44%**前後の狭い範囲に（高い確率で）入っているんじゃないの、と思われますね。でも、実力はもっと上で、たまたま調子が良くなかっただけかもしれません。あるいは、実力はそれほどでもないのにたまたま調子が良かっただけかもしれません。

そこで、さらに数試合を戦ったとして、以下のようなシナリオを考えてみましょう。

- **シナリオ 1：**その後、**100 回**中 **44 回**の成功という成績を収めた。この場合も、単純計算した成功率は 44/100 ＝ **44%**、最終的な通算の成功数は **125 回**中 **55 回**で失敗数 **70 回**だったので通算でも 55/125 ＝ **44%**となる。
  - →　ここまでで、かなりの数をこなしているので、真の成功率は **44%**の近くのより狭い範囲に分布していると考えられる。
- **シナリオ 2：**その後、**100 回**中 **59 回**の成功という成績を収めた。単純計算した成功率は 59/100 ＝ **59%**となる。最終的な通算の成功数は **125 回**中 **70 回**で失敗数 **55 回**だった。
  - →　しかし、真の成功率が **59%**の近辺になったと考えるのは性急。通算だと 70/125 ＝ **56%**なので、**56%**の近くに分布すると考えるのが妥当。

ちょっと先走りになりますが、真の成功率がどのように分布するかを上の例に合わせて可視化すると図 **2** のようになります。上側のグラフがシナリオ 1 に、下側のグラフがシナリオ 2 に対応しています。ここではあくまでイメージを捉えるだけで構いません。グラフの作り方は後で解説します。

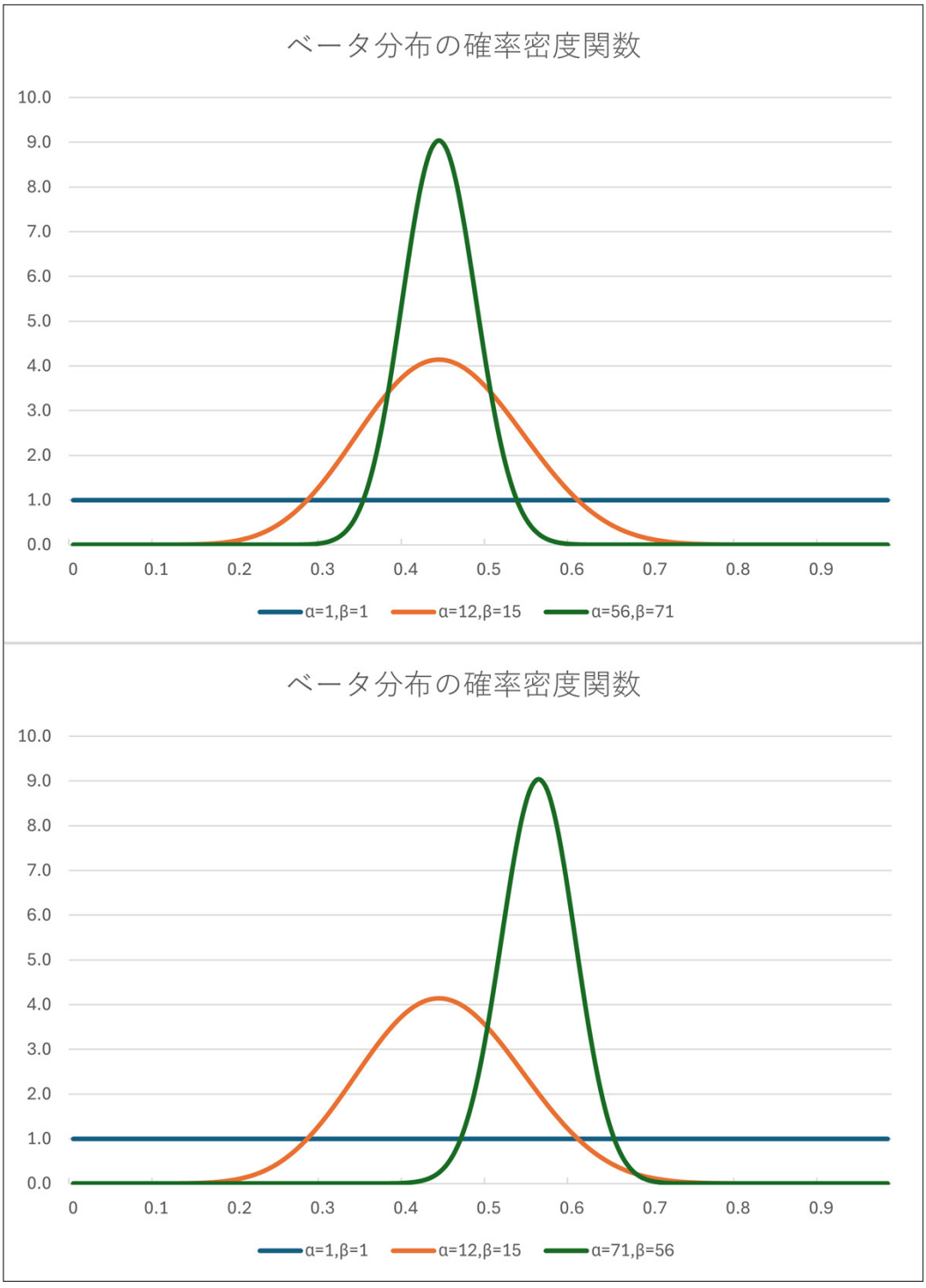


図 2　真の成功率はどのあたりの範囲にあるか

オレンジの曲線が **25 回**中 **11 回**成功の場合のグラフ。上の図では、緑の曲線がシナリオ 1（通算の成功回数 **55 回**）のグラフ。真の成功率が存在する確率の高い範囲がかなり狭くなった。下の図では、緑の曲線がシナリオ 2（通算の成功回数 **70 回**）のグラフ。確率の高い範囲がかなり右にずれた。αやβの値が成功数＋ 1 と失敗数＋ 1 になっている理由は後述する。

各シナリオにおいて、**0 回目**（青の直線）、**25 回目**（オレンジの曲線）、……、**125 回目**（緑色の曲線）のようにして定期的に試行（図 2 の可視化）を繰り返し、それまでの成功率の分布（事前分布）を、「○回目」時点のデータが得られた後の成功率の分布（事後分布）に更新します。この更新をさらに繰り返せば、真の成功率がどのあたりの範囲に入るかが徐々に明確になってくるはずです。

さて、ここでようやくベータ分布の登場です。バスケの例のようなベルヌーイ試行（成功するかしないか）の場合、事前分布をベータ分布であるとすれば、事後分布もベータ分布となることが分かっています。

ベータ分布の確率密度関数では、ある成功率（例えば **44%**）に対する確率密度が求められます。ただし、**確率密度は確率そのものではありません。**つまり、成功率が **44%**である確率が求められるわけではありません。確率密度とは、累積確率がどの程度増えるか、といった「傾向」を表すものです。このことについては、第 6 回で解説した通りです。

一方、ベータ分布の累積分布関数では成功率が何らかの値までである確率が求められます。例えば、成功率が **50%**までである確率が求められます。従って、成功率が **40** ～ **50%**である確率も求められます。成功率が **50%** までの確率から成功率が **40%**までの確率を引けばいいですね。これは図 2 での、グラフと x 軸の **0.4** ～ **0.5** までで囲まれた範囲の面積に当たります。

3 ポイントシュートの成功率が 40 ～ 50%である確率を求める

では、成功率の分布を求めてみましょう。具体的には、BETA.DIST 関数を使って、3 ポイントシュートの成功率が **40** ～ **50%**である確率を求めてみます。BETA.DIST 関数の書き方を見てから操作に進みます。

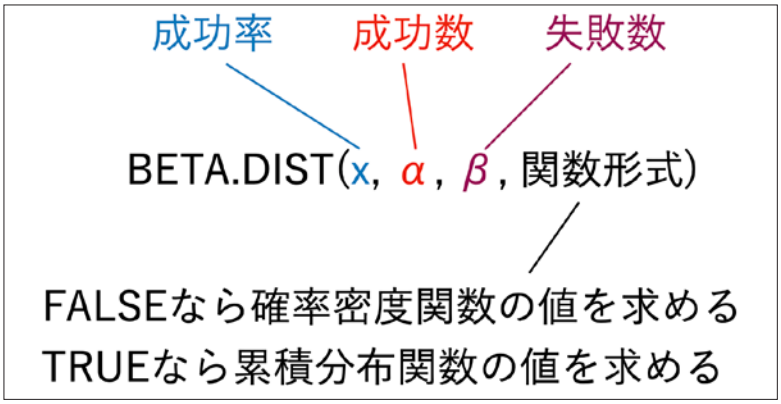


図 3　BETA.DIST 関数に指定する引数

BETA.DIST 関数には、成功率 x（目的の確率）と、成功数を表すα、失敗数を表すβを指定する。関数形式についてはこれまで見てきた関数と同様、**FALSE** を指定すれば確率密度関数の値が、**TRUE** を指定すれば累積分布関数の値が求められる。

成功率がベータ分布の確率変数 x に当たります。台（確率変数 x が取り得る値の範囲）は **0** ～ **1** です。

ベータ分布の母数は成功数を表すαと失敗数を表すβですが、図 2 ではαとβを実際の値＋ 1 としました。その理由は、最初の試行以前の状態が分からない場合、取りあえず成功率が **0** ～ **1** である確率が全て等しいものとするためです。ベータ分布はα＝ 1, β＝ 1 の場合は連続型一様分布になるので、そのような場合の事前分布として使われます。成功数 **1**、失敗数 **1** の状態から、**11 回**の成功、**14 回**の失敗という情報が得られた、と考えるわけです。このように事前の情報が得られない場合に使われる分布のことを**無情報事前分布**と呼びます。

ただし、それ以前の成績が分かっている場合やどの程度であるかが想定される場合には、無情報事前分布ではなく、妥当だと思われる分布を使った方が適切です。

例えば、昨シーズンの成績が **200 回**中 **60 回**の成功、**140 回**の失敗（＝成功率 **30%**）だったとします。シーズンオフに大幅な成長が見られたとするなら、昨シーズンの成績をベースとしつつも控えめに考慮するのがいいでしょう。そこで、成功数と失敗数の割合を保ったまま、値そのものを小さくし、成功数に＋ 6、失敗数に＋ 14 して計算するといったことも考えられます。

なお、αとβにそのまま成功数と失敗数を指定すると、成功数あるいは失敗数の値が **0** の場合にベータ分布の値が求められないというやや消極的な理由もあります（BETA.DIST 関数も **#NUM** エラーを返します）。試行の回数が増えると **1** を足した影響はほとんどなくなります。

では、図 4 で利用例を確認してみましょう。**50%まで**の累積確率から **40%まで**の累積確率を引いて、**40** ～ **50%**の範囲に入る確率を求めます。

サンプルファイルをこちらからダウンロードし、［ベータ分布の利用］ワークシートを開いて試してみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。具体的な操作方法については、図中の説明や図の後の説明を参照してください。

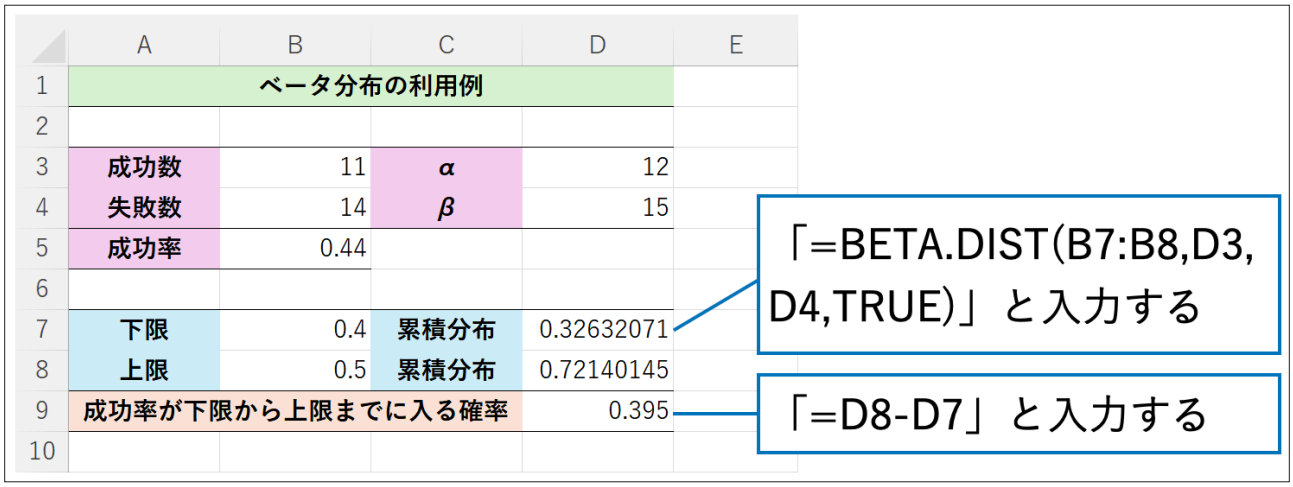


図 4　BETA.DIST 関数の利用例

BETA.DIST 関数には x の値、αの値、βの値を順に指定する。ここでは累積分布関数の値を求めるので、最後の引数には **TRUE** を指定する。x ＝ 50 の場合の累積確率から x ＝ 40 の場合の累積確率を引くと、**40** ～ **50%**の範囲に入る確率が求められる。

◆ **Excel での操作方法**

- セル **D7** に「=BETA.DIST(D7:B8,D3,D4,TRUE)」と入力する
  - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **D7** ～ **D8**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す
- セル **D9** に「=D8-D7」と入力する

◆ **Google スプレッドシートでの操作方法**

- セル **D7** に「=ARRAYFORMULA(BETA.DIST(B7:B8,D3,D4,TRUE))」と入力する
- セル **D9** に「=D8-D7」と入力する

セル **D9** の値は **0.395**（＝ **39.5%**）となりました（図 2 の例で説明するなら **25 試合目**時点での確率）。

ここで、セル **B3** に **22**、セル **B4** に **28** と入力してみてください（**50 試合目**時点）。データから得られた成功率は **44%**のままですが、セル **D9** の、**40** ～ **50%**の範囲に入る確率は **0.527**（＝ **52.7%**）となります。確率が **39.5%**から **52.7%**に高まっているので、成功率が **40** ～ **50%**の範囲内にあることがより確からしくなりました。

また、セル **B3** に **70**、セル **B4** に **55** と入力してみてください（**125 試合目**時点）。データから得られた成功率は **56%**となります。このとき、**40** ～ **50%**の範囲に入る確率（セル **D9** の値）はぐっと下がって **0.090**（＝ **9.0%**）となります。真の成功率がある範囲は **56%**あたりだと考えられるので当然ですね。そこで、セル **B7** に **0.5**、セル **B8** に **0.6** と入力してみましょう。すると、**0.733**（＝ **73.3%**）という値が得られます。

このことは、図 2 のグラフと照らし合わせてみるとよく分かります。というわけで、そろそろ、ベータ分布の確率密度関数と累積分布関数を可視化する方法を見ておきたいですね。その前に、ベータ分布の確率密度関数と累積分布関数の定義だけを掲載しておきます。例によって、定義を覚える必要は全くありません。さらっとスルーして、ベータ分布の可視化に進んでいただいて結構です。

ベータ分布の確率密度関数と累積分布関数

以下に示した式がベータ分布の確率密度関数と累積分布関数の定義です。これらの式を覚える必要は全くありません。Excel の BETA.DIST 関数を使えば、簡単に答えが求められます。

◆ ベータ分布の確率密度関数

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \qquad (1)$$

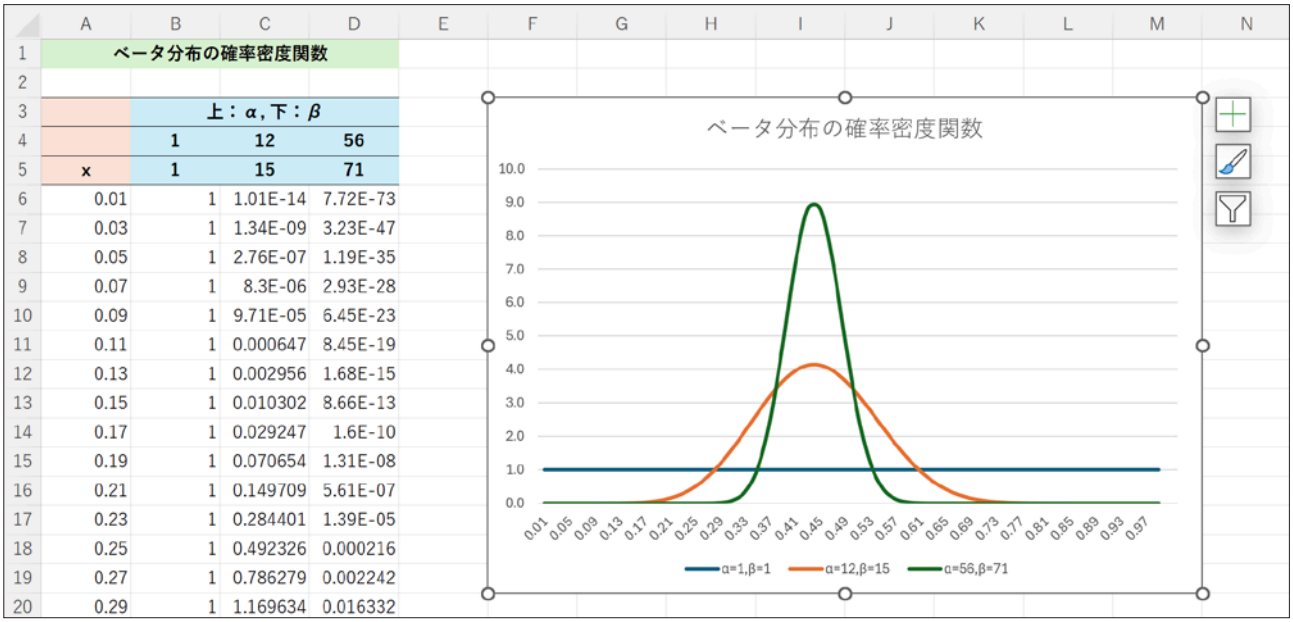
◆ ベータ分布の累積分布関数

$$F(x) = I_x(\alpha,\beta) \qquad (2)$$

ただし、B はベータ関数、I は正則化ベータ関数です。これらの関数の定義は、第 8 回のコラムで紹介しました。

ベータ分布ってどんな感じの分布（1）～ 確率密度関数を可視化してみよう

BETA.DIST 関数の引数として母数α , βに幾つかの値を指定し、x の値を変化させていったグラフを描けば、ベータ分布の確率密度関数や累積分布関数を可視化できます。今回は、図 1 の例で使った値を指定して、成功率がどの範囲にあるのかを可視化したいと思います。なお、x = 0 や x = 1 の場合、α , βの値によっては BETA.DIST 関数がエラーになることがあるので、x = 0.01 ～ 0.99 のグラフとします（図 5）。



	A	B	C	D
1	ベータ分布の確率密度関数			
2				
3		上：α, 下：β		
4		1	12	56
5	x	1	15	71
6	0.01	1	1.01E-14	7.72E-73
7	0.03	1	1.34E-09	3.23E-47
8	0.05	1	2.76E-07	1.19E-35
9	0.07	1	8.3E-06	2.93E-28
10	0.09	1	9.71E-05	6.45E-23
11	0.11	1	0.000647	8.45E-19
12	0.13	1	0.002956	1.68E-15
13	0.15	1	0.010302	8.66E-13
14	0.17	1	0.029247	1.6E-10
15	0.19	1	0.070654	1.31E-08
16	0.21	1	0.149709	5.61E-07
17	0.23	1	0.284401	1.39E-05
18	0.25	1	0.492326	0.000216
19	0.27	1	0.786279	0.002242
20	0.29	1	1.169634	0.016332

図 5　ベータ分布の確率密度関数の例

前述のシナリオ 1 に沿ったグラフ。αは **1**、**12**、**56** とし、βは **1**、**15**、**71** とする。x = 0.01 ～ 0.99 までの確率密度関数の値を **0.02** 刻みで求めてグラフを描いてみた。x と表記されている **A** 列の値が確率変数（横軸の値）。**B** ～ **D** 列の **6** 行目以降はそれぞれのα , βでの x に対する確率密度関数の値。シナリオ 2 に沿ったグラフにするには、セル **D4** に **71**、セル **D5** に **56** を入力すればよい。

確率密度関数の値を求めるための手順は以下の通りです。可視化については単に折れ線グラフを描くだけなので、関数の入力にのみ焦点を当てることにします。グラフ作成の手順についてはサンプルファイル内に掲載しておきます。［ベータ分布］ワークシートを開いて、図の後の手順で試してみてください。

◆ Excel での操作方法

- セル **B6** に =BETA.DIST(A6:A55,B4:D4,B5:D5,FALSE) と入力する
  - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **B6** ～ **D55**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す

◆ Google スプレッドシートでの操作方法

- セル **B6** に =ARRAYFORMULA(BETA.DIST(A6:A55,B4:D4,B5:D5,FALSE)) と入力する

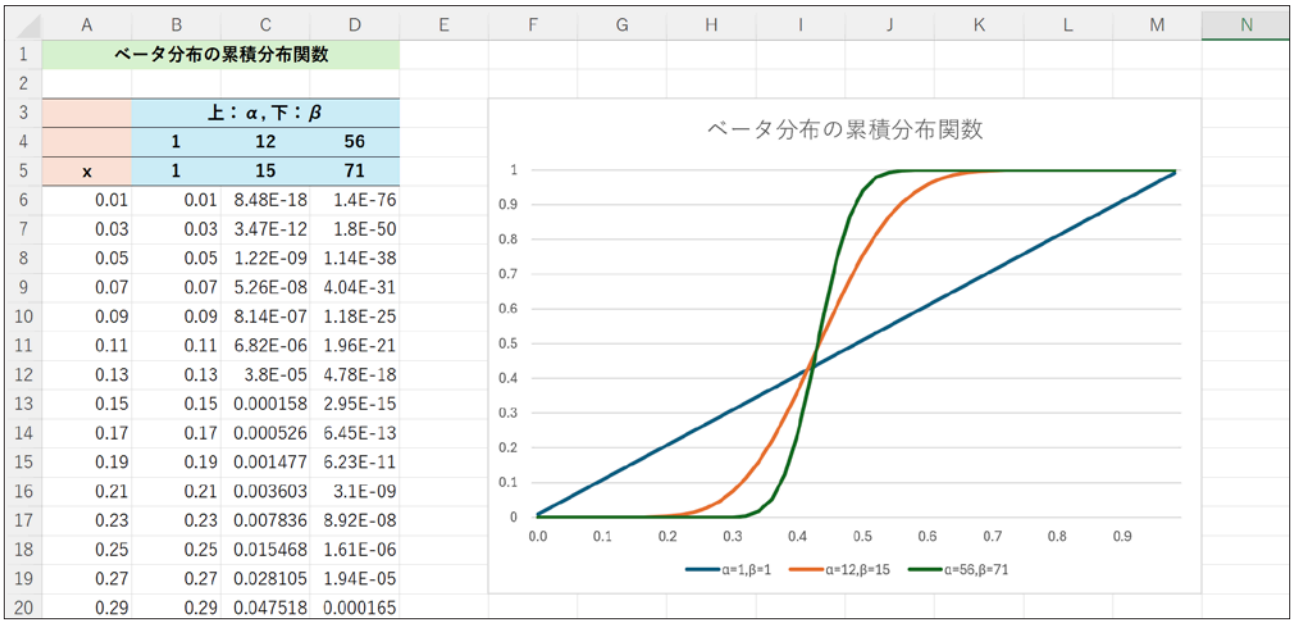
● グラフの作成方法

- サンプルファイル内に掲載しておきます（タイトルや軸の書式などの細かい設定は省略）

図 5 のグラフを見ると、$\alpha = 1, \beta = 1$ の場合は、連続型一様分布になっていることが分かります。また、$\alpha = 12, \beta = 15$ の場合は、山が **0.25** ～ **0.65** ぐらいの広い範囲にあり、$\alpha = 56, \beta = 71$ の場合は、**0.35** ～ **0.55** ぐらいの狭い範囲にあることも分かります。後者の方が、観測された値（**0.44**）にぐっと近くなりますね。ただし、山の高さが確率を表しているわけではないということに注意してください。グラフと x 軸で囲まれた範囲の面積が確率（累積確率）を表します。

ベータ分布ってどんな感じの分布（2）～ 累積分布関数を可視化してみよう

続いて、累積分布関数です。x, α , βの値は図 5 と同様とします（図 6）。［ベータ分布累積］ワークシートを開いて、図の後の手順で試してみてください。



	A	B	C	D
1	ベータ分布の累積分布関数			
2				
3		上：α, 下：β		
4		1	12	56
5	x	1	15	71
6	0.01	0.01	8.48E-18	1.4E-76
7	0.03	0.03	3.47E-12	1.8E-50
8	0.05	0.05	1.22E-09	1.14E-38
9	0.07	0.07	5.26E-08	4.04E-31
10	0.09	0.09	8.14E-07	1.18E-25
11	0.11	0.11	6.82E-06	1.96E-21
12	0.13	0.13	3.8E-05	4.78E-18
13	0.15	0.15	0.000158	2.95E-15
14	0.17	0.17	0.000526	6.45E-13
15	0.19	0.19	0.001477	6.23E-11
16	0.21	0.21	0.003603	3.1E-09
17	0.23	0.23	0.007836	8.92E-08
18	0.25	0.25	0.015468	1.61E-06
19	0.27	0.27	0.028105	1.94E-05
20	0.29	0.29	0.047518	0.000165

図 6　ベータ分布の累積分布関数の例

前述のシナリオ 1 に沿ったグラフ。αは **1**、**12**、**56** とし、βは **1**、**15**、**71** とする。x = 0.01 ～ 0.99 までの累積分布関数の値を **0.02** 刻みで求めてグラフを描いてみた。α = 56, β = 71 の場合は、**0.35** あたりまでは累積確率がほぼ **0** で、それ以降、急激に **1** に近づく。なお、横軸は成功率で、縦軸は累積確率。シナリオ **2** に沿ったグラフにするには、セル **D4** に **71**、セル **D5** に **56** を入力すればよい。

操作の手順は図 4 とほぼ同じです。違いは、BETA.DIST 関数の［関数形式］に累積分布関数を表す **TRUE** を指定することだけです。グラフ作成の手順についてはサンプルファイル内に掲載しておきます。

◆ Excel での操作方法

- セル **B6** に =BETA.DIST(A6:A55,B4:D4,B5:D5,TRUE) と入力する
  - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **B6** ～ **D55**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す

◆ Google スプレッドシートでの操作方法

- セル **B6** に =ARRAYFORMULA(BETA.DIST(A6:A55,B4:D4,B5:D5,TRUE)) と入力する

● グラフの作成方法

- サンプルファイル内に掲載しておきます（タイトルや軸の書式などの細かい設定は省略）

図 5 のグラフから、α = 56, β = 71 の場合は、**0.35** あたりまでは累積確率がほぼ **0** で、それ以降急激に高くなることが分かります。また、**0.55** あたりでほぼ **1** に近づいていることも分かります。このことからも、成功率が **35** ～ **55%**の範囲にあることがかなり「確からしい」と言えそうです。

確率が 95%となる成功率の範囲を求める ～ ベータ分布の逆関数

BETA.DIST 関数で関数の形式に **TRUE** を指定すると、成功率に対するベータ分布の累積分布関数の値、つまり累積確率が求められました。例えば、成功率に **50%**（＝ **0.5**）を指定すると、成功率が **50%まで**である確率が求められました。

ここでは、その逆の計算を行います。つまり、ある累積確率になるのは成功率が何%までの場合かということを求めます。そのためには、BETA.INV 関数を使って、累積確率から確率変数の値（成功率）を求めます。まず、BETA.INV 関数の書き方を見ておきましょう。

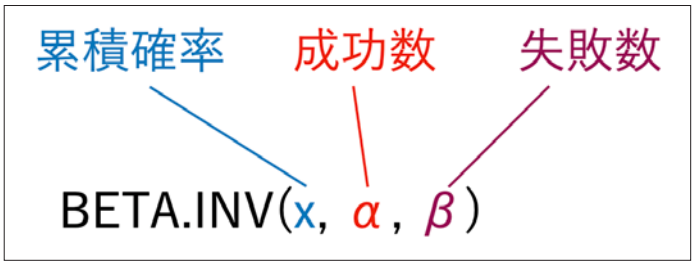


図 7　BETA.INV 関数に指定する引数

BETA.INV 関数には、累積確率と、成功数を表すαと、失敗数を表すβを指定する。累積分布関数の値を基に、確率変数の値を求めることができる。

BETA.INV 関数を使えば、例えば、全体の **95%**を占める成功率がどの範囲であるかを求めることもできます。**2.5%**までの位置から **97.5%**までの位置ですね。［ベータ分布の逆関数］ワークシートを開いて、図の中に記された手順で試してみてください。

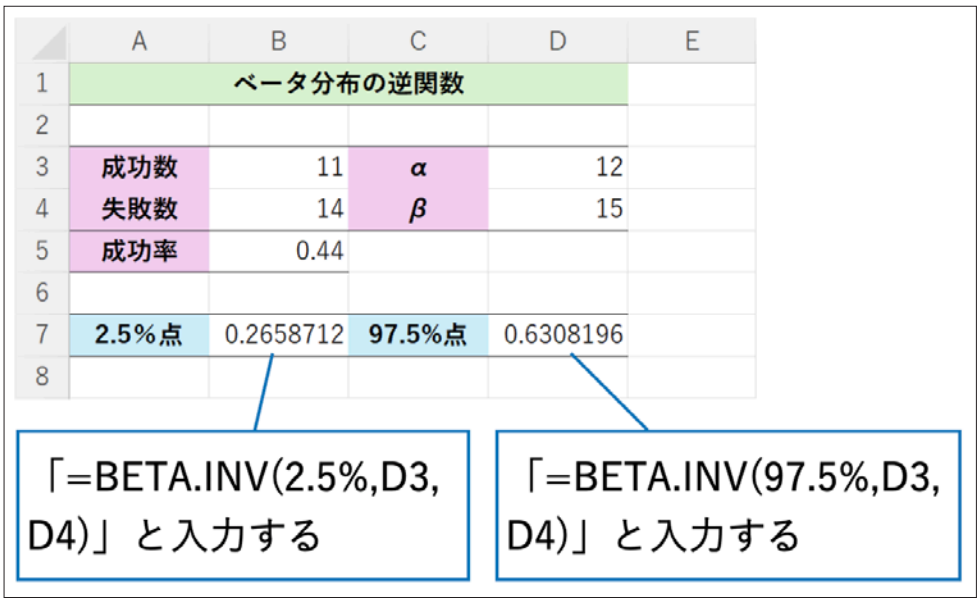


図 8　ベータ分布の逆関数で全体の 95%を占める範囲を求める

BETA.INV 関数には、累積確率と成功数αと、失敗数βを指定する。**2.5%**の確率である成功率の範囲は **0.2659** まで、**97.5%**の確率である成功率の範囲は **0.6308** まで、という結果が求められた。全体の **95%** となる成功率の範囲は **0.2659** ～ **0.6308** ということになる。

結果は、**2.5%**点が **0.2659**、**97.5%**点が **0.6308** となりました。全体の **95%**（＝ 97.5%－ 2.5%）を占める成功率の範囲が **0.2659** ～ **0.6308** であるということですね。

ここで、シナリオ 1 を試してみましょう。セル **B3** に **55**、セル **B4** に **70** と入力してみてください。**2.5%**点が **0.3560**、**97.5%** 点が **0.5277** という結果になります。**0.3560** ～ **0.5277** ということなので、かなり範囲が狭くなりました。

また、シナリオ 2 も試してみてください。セル **B3** に **70**、セル **B4** に **55** と入力すれば、全体の **95%**を占める成功率の範囲が **0.4723** ～ **0.6440** となることが分かります。

これらの数字をどう解釈するかは、頻度主義とベイズ主義とでは異なります。頻度主義では、このような **95%**に含まれる範囲のことを **95%信頼区間**と呼びます。ただし、信頼区間は「真の値（真の成功率）」が、**95%**の確率でその範囲の中に入っているという意味ではありません。サンプルを取り出して信頼区間を求めることを何度も繰り返すと、それらの信頼区間のうちの **95%**が真の値を含んだものになっている、ということです。

一方のベイズ主義ではこの範囲を**確信区間**または**信用区間**と呼びます。ベイズ統計では、その範囲の中に真の成功率が **95%**の確率で含まれている、と素直に解釈できます。なお、今回の例では、簡単な計算で事後分布が求められましたが、分布によっては計算が複雑になり、簡単に求められない場合もあります。そのような場合には乱数を使ったシミュレーションにより事後分布を求めます。詳細については、この連載の続編である推測統計編でお話しする予定です。

ここでは、分布の累積確率が小さい部分を左右から **2.5%**ずつ切り取って **95%**の範囲を求めました。このような方法で求めた確信区間を**等裾事後確信区間（ETI：Equal-Tailed posterior credible Interval）**と呼びます。

一方、山の高いところから **95%**を求めることもあります。そのような確信区間を**最大事後密度確信区間（HDI または HPDI：Highest posterior Dencity Interval）**と呼びます。

A/B テストによりどちらの広告が有効かを判定する

最後に、実用的な例として、A/B テストの例も紹介しておきましょう。Web サイトの広告をユーザーがクリックした後、資料請求や購買などの行動を取ったことを**コンバージョン（CV）**といいます。

ここで、広告を 2 種類を用意し、ランダムに表示して、どちらの効果が高いかを見極めることとしましょう。一定期間の実験を行ったところ、以下のような結果になったとします。CV 数は成功数に当たるものです、クリック数は文字通り広告をクリックした回数です。こちらは、失敗数ではなく、試行数に当たることに注意してください。

	広告A	広告B
CV数	11	24
クリック数	250	275
CV率	0.044	0.0872

表 1　Web サイトの広告に対する CV 率（= CV 数 ÷ ユーザークリック数）

一見して、広告 B の CV 率の方が良さそうです。そこで、CV 率の分布をそれぞれ可視化してみましょう。確率密度関数とグラフの作成方法は図 5 の例とほぼ同じです。CV 率の値が小さいので、確率変数 x の範囲を **0.01** ～ **0.2** に限定して作成してみます。［A-B テスト］ワークシートを開いて図 9 の後の手順で試してみてください。

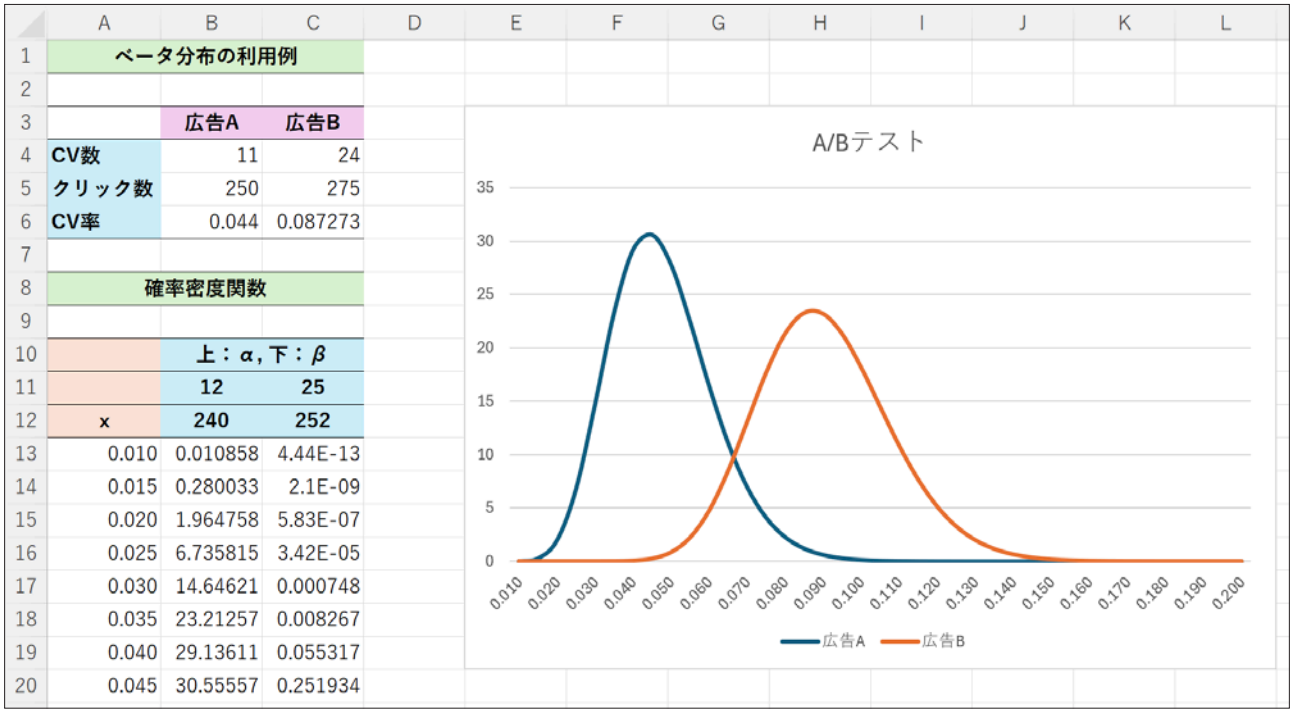


図 9　A/B テストの実行結果

ここでは、事前分布を連続一様分布とするので、αは CV 数に **1** を加えた値とし、βはクリック数－ CV 数＋ 1 とする。確率密度関数を見ても広告 B が優れていることが分かる。ただし、全ての場合において広告 B が優れているというわけではない。例えば、広告 A の CV 率が **0.08** 以上になる確率や、広告 B の CV 率が **0.06** 以下になる確率は小さいが、**0** というわけではない。

→目次に戻る

◆ **Excel での操作方法**

- セル **B13** に `=BETA.DIST(A13:A51,B11:C11,B12:C12,FALSE)` と入力する
  - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **B13** ～ **C51**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す

◆ **Google スプレッドシートでの操作方法**

- セル **B13** に `=ARRAYFORMULA(BETA.DIST(A13:A51,B11:C11,B12:C12,FALSE))` と入力する

● **グラフの作成方法**

- サンプルファイル内に掲載しておきます（タイトルや軸の書式などの細かい設定は省略）

図 9 を見ると、広告 B の方が優れていることが分かります。ただし、グラフには重なっている部分があるので、絶対的に広告 B が優位であるというわけではありません。例えば、広告 A の CV 率が **0.08** で、広告 B の CV 率が **0.06** になることもあり得ます。

そこで、広告 B がどの程度優れているかを簡単なシミュレーションで求めてみましょう。それぞれのベータ分布に従う乱数を作成して、A の CV 率より B の CV 率の方が大きい割合を求めればいいですね。Excel でももちろんできますが、乱数を 1000 個作るとすれば少なくとも 1000×2 個のセルが必要になります。Excel での例も［A-B テスト (完成例)］ワークシートに含めてありますが、Python であれば数行で済むので、ここでは簡単なコードを紹介しておきます。

サンプルプログラムはこちらから参照できます。リンクをクリックすれば、ブラウザが起動し、Google Colaboratory の画面が表示されます（Google アカウントでのログインが必要です）。最初のコードセルをクリックし、［Shift］＋［Enter］キーを押してコードを実行してみてください。コードの詳細については解説しませんが、コメントとリスト 1 の説明を見れば何をやっているかが大体分かると思います。

```
from scipy.stats import beta
import numpy as np

cv_a = 11  # A の CV 数
click_a = 250  # A のクリック数
cv_b = 24  # B の CV 数
click_b = 275  # B のクリック数
a = beta.rvs(cv_a+1, click_a-cv_a+1, size=1000)  # ベータ分布の乱数を 1000 個
b = beta.rvs(cv_b+1, click_b-cv_b+1, size=1000)  # ベータ分布の乱数を 1000 個
np.mean(a < b)  # b の方が大きくなる割合を求める

# 出力例（乱数を使っているので実行のたびに結果が少し変わります）: 0.971
```

リスト 1　広告 A よりも広告 B の方が優っている割合を求める

`beta.rvs` 関数はベータ分布からランダムにサンプルを抽出するための関数。引数には、α、β、そして作成する乱数の個数を指定する。**1000 個**の乱数を作り、`a < b` という比較を行うと **True** か **False** が返されるが、Python では **True** を **1**、**False** を **0** と見なせるので、NumPy の `mean` 関数で平均値を求めると **True** の割合（`b` が大きい割合）が求められる。

結果を見ると、**97.1%**の確率で広告 B の CV 率が高くなっていることが分かります。なお、サンプルプログラムの中には図 9 と同様のグラフを描画するコードも含めてあります。

◇ ◇ ◇ ◇ ◇ ◇ ◇

今回はベータ分布について、簡単な利用例を紹介した後、確率密度関数と累積分布関数の求め方についてお話しし、さらに A/B テストへの応用例を見ました。その中で、ベイズ統計における事前分布と事後分布についても簡単に触れました。確信区間などについてやや深入りしたかもしれませんが、詳細については推測統計編でお話しします。

次回は、機械の寿命や故障率の分析などに使われるワイブル分布のお話をします。次回もお楽しみに！

この記事で取り上げた関数の形式

関数の利用例については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

ベータ分布の確率密度関数や累積分布関数の値を求めるための関数

BETA.DIST 関数：ベータ分布の確率密度関数や累積分布関数の値を求める

形式

BETA.DIST(x, α , β , 関数形式)

引数

- **x**：確率を求めたい事象の成功率を指定する。
- **α**：事象の成功数を指定する。
- **β**：事象の失敗数を指定する。
- **関数形式**：以下の値を指定する。
  - **FALSE** …… 確率密度関数の値を求める
  - **TRUE** …… 累積分布関数の値を求める

BETA.INV 関数：ベータ分布の累積分布関数の値に対する逆関数の値を求める

形式

BETA.INV(x, α , β)

引数

- **x**：ベータ分布の累積確率を指定する。
- **α**：事象の成功数を指定する。
- **β**：事象の失敗数を指定する。

［データ分析］ワイブル分布 ～ 15 年以内にエアコンが故障する確率は？

データ分析の初歩から学んでいく連載（確率分布編）の第 13 回。ワイブル分布は機械の寿命や故障率の分析に使われる分布です。今回も具体例を基に、ワイブル分布の利用例や、確率密度関数と累積分布関数の形を見ていきます。母数（パラメーター）として指定するαやβの適切な値の決め方も解説します。

羽山博（2025 年 01 月 09 日）

データ分析の初歩から応用まで少しずつステップアップしながら学んでいく連載の確率分布編、第 13 回（最終回）です。前回は、成功率の分布を求めたり、A/B テストによる優劣を判定したりするのに使われるベータ分布を取り上げました。今回は機械などの寿命や故障率の分析に使われるワイブル分布の特徴や意味を基本から解き明かし、確率密度関数／累積分布関数の求め方、利用例などを見ていきます。

15 年以内にエアコンが故障する確率を求める ～ ワイブル分布の利用

2024 年の夏は記録的な猛暑でした。そんな中で、エアコンの故障という悲劇が筆者に襲いかかりました。冷房のボタンを押しても、暖房になってしまうのです！　調べてみると四方弁（しほうべん）と呼ばれる、冷房と暖房を切り替える弁が固着して動かなくなるのが原因のようでした。かなり古い機種だったので、修理ではなく買い換えを選択しましたが、今度は、新しいエアコンが猛暑のせいで上手く排熱できず、すぐに止まってしまうという事態に……。

そんなわけで、私事からのお話でしたが、今回はエアコンの寿命について考えてみたいと思います（図 1）。事例として取り上げるのはエアコンの寿命ですが、さまざまな機器や生命体の寿命などについても同様に応用できるお話です。

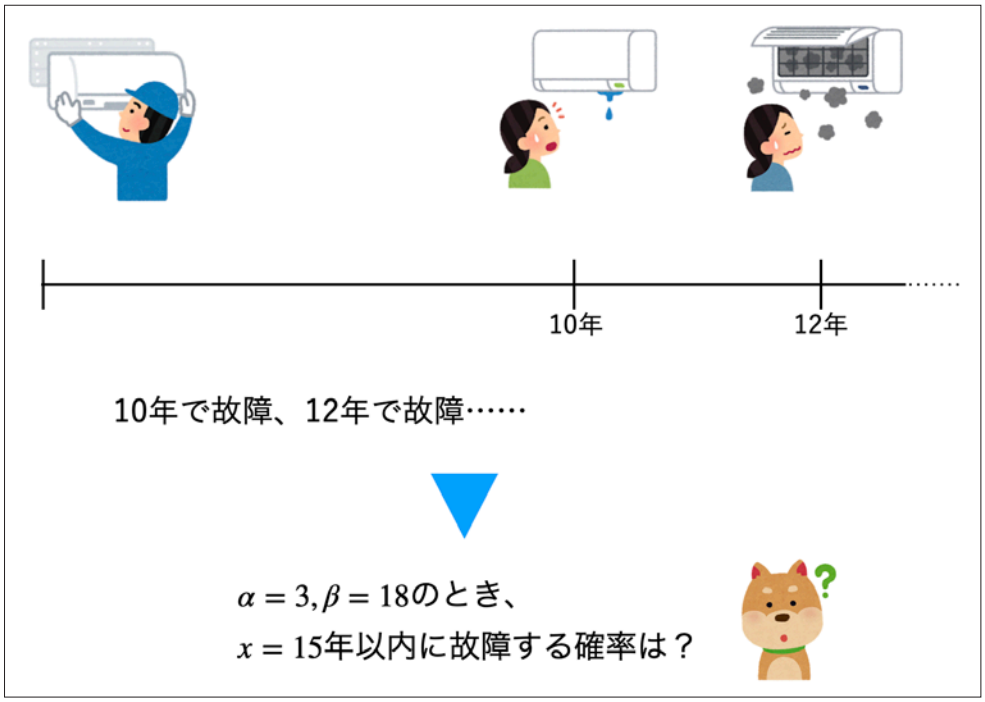


図 1　エアコンなどの機器が一定時間以内に故障する確率を求めたい！

故障するまでの年数のデータが何件かあるとき、ワイブル分布の母数α , βを求めて、x = 15 年以内に故障する確率を求めたい。ただし、α , βの求め方については後述することとして、以下の例では、α= 3, β= 18 で計算してみよう。

すでに述べたように、機械の寿命などの分析にはワイブル分布が利用されます。ワイブル分布の母数は、分布の形を決める形状パラメーターαと分布の広がりを表すのに使われる尺度パラメーターβで、確率変数は時間を表す x となります。

ここでは、α= 3, β= 18 であるものとして、x = 15 年以内に故障する確率を計算するところから始めます。α , βが変わるとグラフの形がどう変わるのか、また、α , βをどうやって決めたのかが気になるところですが、それについては後述します。理屈は後回しにして、Excel の WEIBULL.DIST 関数を使って計算する方法を見ておきましょう。

ここでは、Excel のヘルプに合わせて、形状パラメーターをα、尺度パラメーターをβという文字で表しています。ただし、文献によっては、形状パラメーターとして m やβ、尺度パラメーターとしてαやη（イータ）などの文字が使われることがあります。α , βの意味が逆になっている場合もあるので、ご注意ください。

WEIBULL.DIST 関数の形式を見てから操作に進みます（図 2）。

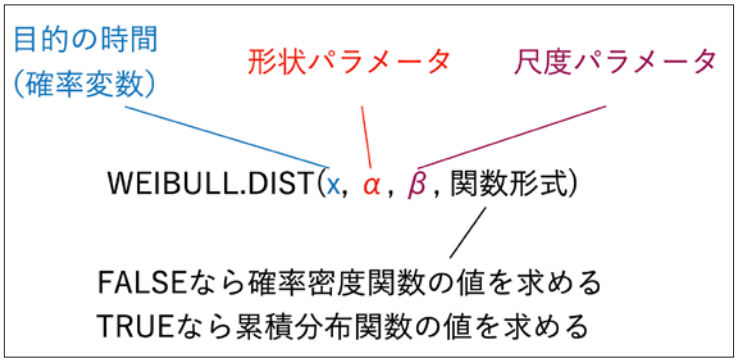


図 2　WEIBULL.DIST 関数に指定する引数

WEIBULL.DIST 関数には、確率変数 x（目的の時間）と、形状パラメーター α、尺度パラメーター β を指定する。関数形式についてはこれまで見てきた関数と同様、**FALSE** を指定すれば確率密度関数の値が、**TRUE** を指定すれば累積分布関数の値が求められる。

ワイブル分布の確率変数 x には時間を指定します。累積分布関数では、指定した時間までに故障などの事象が起こる確率が求められます。ここでは故障を例としましたが、故障に限らず、目的の事象が起こるまでの時間が確率変数 x となるわけです。台（確率変数 x が取り得る値の範囲）は loc ～∞です（loc は指定した開始位置）。

Excel の WEIBULL.DIST 関数では、loc の値が **0** に固定されているので、**x<0** の場合はエラーになります。WEIBULL.DIST 関数で台の開始位置を変えたい場合は、x から loc の値を引いて WEIBULL.DIST 関数を適用します。

では、ワイブル分布の母数を $\alpha = 3$, $\beta = 18$ として、累積確率を求めてみましょう。

サンプルファイルをこちらからダウンロードし、［ワイブル分布の利用］ワークシートを開いて試してみてください。Google スプレッドシートのサンプルはこちらから開くことができます。メニューから［ファイル］－［コピーを作成］を選択し、Google ドライブにコピーしてお使いください。具体的な操作方法については、図中の説明を参照してください。

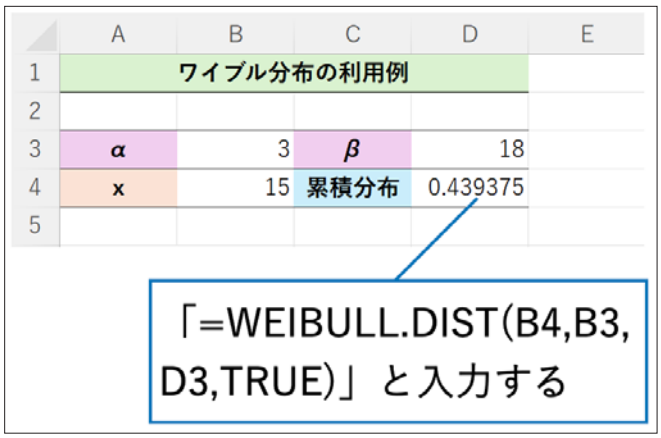


図 3　WEIBULL.DIST 関数の利用例

WEIBULL.DIST 関数には x の値、α の値、β の値を順に指定する。ここでは、**15 年**以内に故障する確率を求めたいので、累積分布関数の値を求める。従って、最後の引数には **TRUE** を指定する。

→目次に戻る

図 **3** のように、セル **D4** に「=WEIBULL.DIST(B4,B3,D3,TRUE)」と入力すると、セル **D4** の値は **0.439**（＝ **43.9%**）となりました。**15** 年以内に故障する確率は **43.9%**であるということが分かりました。

続いて、ワイブル分布の確率密度関数と累積分布関数を可視化してみます。特に、累積分布関数を可視化すれば、何年か以内に故障する確率がよく分かります。が、その前に、ワイブル分布の確率密度関数と累積分布関数の定義だけを掲載しておきます。例によって、定義を覚える必要は全くありません。さらっとスルーして、ワイブル分布の可視化に進んでいただいて結構です。

ワイブル分布の確率密度関数と累積分布関数

以下に示した式がワイブル分布の確率密度関数と累積分布関数の定義です。繰り返しになりますが、これらの式を覚える必要は全くありません。Excel の WEIBULL.DIST 関数を使えば、簡単に答えが求められます。

◆ ワイブル分布の確率密度関数

$$f(x) = \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-(\frac{x}{\beta})^{\alpha}} \qquad (1)$$

◆ ワイブル分布の累積分布関数

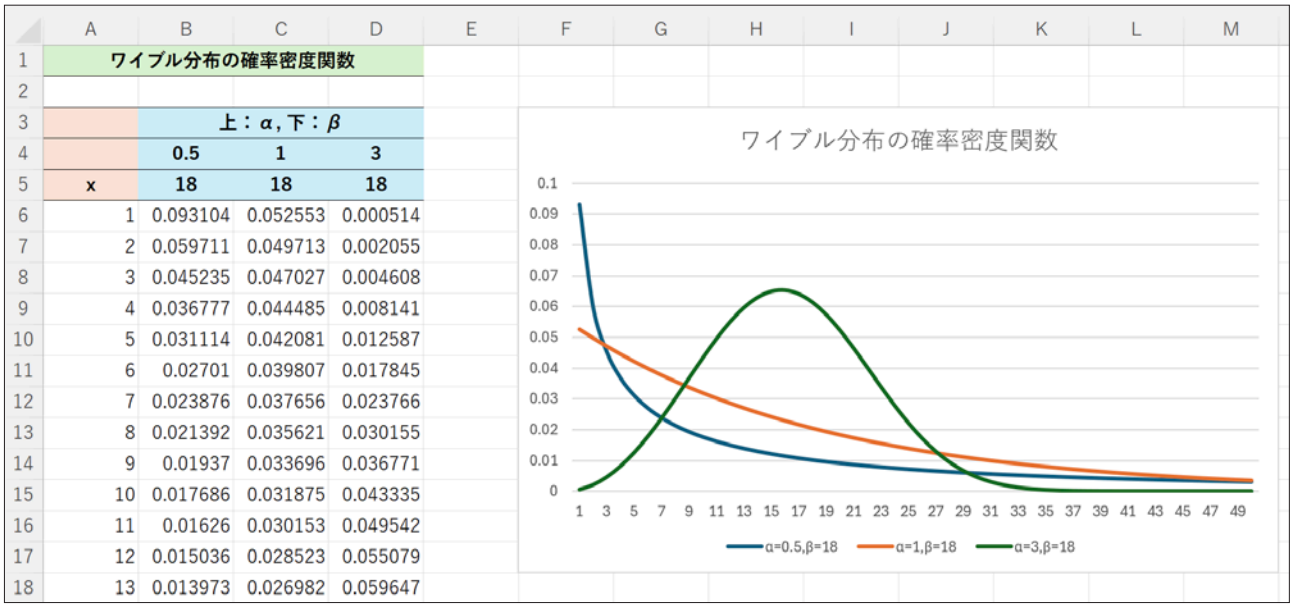
$$F(x) = 1 - e^{-(\frac{x}{\beta})^{\alpha}} \qquad (2)$$

ただし、e は自然対数の底（＝ **2.718**...）です。

ワイブル分布ってどんな感じの分布（1）～ 確率密度関数を可視化してみよう

WEIBULL.DIST 関数の引数として母数α , βに幾つかの値を指定し、x の値を変化させていったグラフを描けば、ワイブル分布の確率密度関数や累積分布関数を可視化できます。今回は、βについては図 1 の例で使った **18** という値を指定し、αの値を変えて可視化したいと思います。尺度パラメーターであるβを変えても分布の高さや幅が変わるだけですが、形状パラメーターであるαを変えるとグラフの形が変わるので、違いがよく分かると思います。

なお、x ＝ 0 の場合、WEIBULL.DIST 関数の値が **0** となり、グラフがきれいにつながらないことがあるので、x ＝ 1 ～ 50 のグラフとします（図 4）。



	A	B	C	D
1	ワイブル分布の確率密度関数			
2				
3		上：α, 下：β		
4		0.5	1	3
5	x	18	18	18
6	1	0.093104	0.052553	0.000514
7	2	0.059711	0.049713	0.002055
8	3	0.045235	0.047027	0.004608
9	4	0.036777	0.044485	0.008141
10	5	0.031114	0.042081	0.012587
11	6	0.02701	0.039807	0.017845
12	7	0.023876	0.037656	0.023766
13	8	0.021392	0.035621	0.030155
14	9	0.01937	0.033696	0.036771
15	10	0.017686	0.031875	0.043335
16	11	0.01626	0.030153	0.049542
17	12	0.015036	0.028523	0.055079
18	13	0.013973	0.026982	0.059647

図 4　ワイブル分布の確率密度関数の例

αは **0.5**、**1**、**3** とし、βは **18** とする。x = 1 ～ 50 までの確率密度関数の値を **1** 刻みで求めてグラフを描いてみた。x と表記されている **A** 列の値が確率変数（横軸の値）。**B** ～ **D** 列の **6** 行目以降はそれぞれのα , βでの x に対する確率密度関数の値。

確率密度関数の値を求めるための手順は以下の通りです。可視化については単に折れ線グラフを描くだけなので、関数の入力にのみ焦点を当てることにします。グラフ作成の手順についてはサンプルファイル内に掲載しておきます。［ワイブル分布］ワークシートを開いて、図の後の手順で試してみてください。

◆ Excel での操作方法

- セル **B6** に =WEIBULL.DIST(A6:A55,B4:D4,B5:D5,FALSE) と入力する
  - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **B6** ～ **D55**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す

◆ Google スプレッドシートでの操作方法

- セル **B6** に =ARRAYFORMULA(WEIBULL.DIST(A6:A55,B4:D4,B5:D5,FALSE)) と入力する

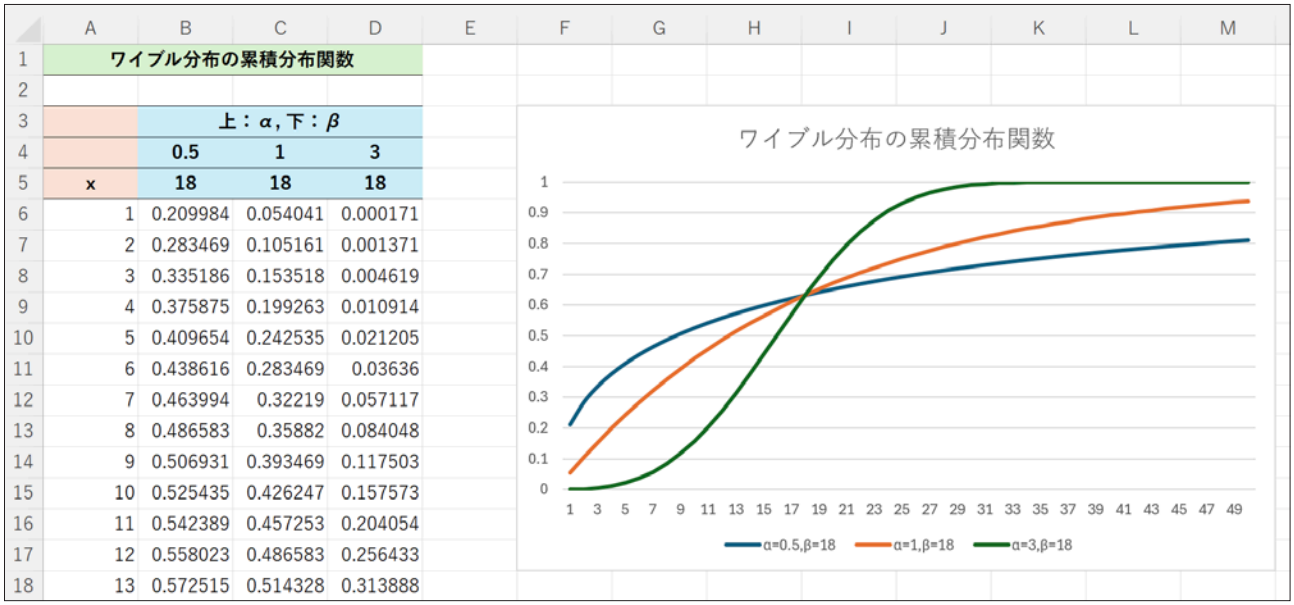
● グラフの作成方法

- サンプルファイル内に掲載しておきます（タイトルや軸の書式などの細かい設定は省略）

図 4 のグラフを見ると、形状パラメーターαが **1** より小さいとき（例：α = 0.5）は、最初の部分（例えば x = **3 まで**）の値が大きく、徐々に落ち着いていくことが分かります。この場合は、初期不良のモデル化に使われます。形状パラメーターαが **1** のときは、$\lambda = 1/\beta$の指数分布となります。この場合は、一定の割合で（偶発的に）故障が起こるようなモデルとなります。形状パラメーターαが **1** より大きいとき（例：α = 3）は、徐々に故障が増えていく経年劣化のモデルとなります。このことは、図 4 のグラフよりも、**故障率**を可視化した図の方が分かりやすいので、後ほど掲載する図 9 であらためて見ることとします。

ワイブル分布ってどんな感じの分布（2）～ 累積分布関数を可視化してみよう

続いて、累積分布関数です。x, α , βの値は図 4 と同様とします（図 5）。［ワイブル分布累積］ワークシートを開いて、図の後の手順で試してみてください。



	A	B	C	D
1	ワイブル分布の累積分布関数			
2				
3		上：α, 下：β		
4		0.5	1	3
5	x	18	18	18
6	1	0.209984	0.054041	0.000171
7	2	0.283469	0.105161	0.001371
8	3	0.335186	0.153518	0.004619
9	4	0.375875	0.199263	0.010914
10	5	0.409654	0.242535	0.021205
11	6	0.438616	0.283469	0.03636
12	7	0.463994	0.32219	0.057117
13	8	0.486583	0.35882	0.084048
14	9	0.506931	0.393469	0.117503
15	10	0.525435	0.426247	0.157573
16	11	0.542389	0.457253	0.204054
17	12	0.558023	0.486583	0.256433
18	13	0.572515	0.514328	0.313888

図 5　ワイブル分布の累積分布関数の例

図 4 と同様、αは **0.5**、**1**、**3** とし、βは **18** とする。x = 1 ～ 50 までの累積分布関数の値を **1** 刻みで求めてグラフを描いてみた。x と表記されている **A** 列の値が確率変数（横軸の値）。**B** ～ **D** 列の **6** 行目以降はそれぞれのα , βでの x に対する累積分布関数の値。

操作の手順は図 4 とほぼ同じです。違いは、WEIBULL.DIST 関数の［関数形式］に累積分布関数を表す **TRUE** を指定することだけです。図 4 の場合と同様、グラフ作成の手順についてはサンプルファイル内に掲載しておきます。

◆ Excel での操作方法

- セル **B6** に =WEIBULL.DIST(A6:A55,B4:D4,B5:D5,TRUE) と入力する
  - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **B6** ～ **D55**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す

◆ Google スプレッドシートでの操作方法

- セル **B6** に =ARRAYFORMULA(WEIBULL.DIST(A6:A55,B4:D4,B5:D5,TRUE)) と入力する

● グラフの作成方法

- サンプルファイル内に掲載しておきます（タイトルや軸の書式などの細かい設定は省略）

図 5 のグラフから、α = 3, β = 18 の場合は、x = 30 あたりで累積確率がほぼ **1** に近づくことが分かります。つまり、最初の図 1 で見たエアコンは、長く使えたとしても **30 年**程度でほぼ寿命が来るものと考えてよさそうです。寿命を求めるには、次の項で見るワイブル分布の逆関数が便利です。

→目次に戻る

故障する確率が 95%となる時間を求める ～ ワイブル分布の逆関数

ワイブル分布の累積分布関数は、ある時間までに故障する確率を求めるのに使えます。ということは、その逆関数を利用すれば、故障する累積確率が何%かになる時間を求めることができます。例えば、故障する確率が**95%**以上になる時間が求められます。

しかし、Excel にはワイブル分布の逆関数を求めるための関数がありません。そこで、累積分布関数の値を求める（2）式を x について解いた式を使います。（2）式は以下の通りでした。

$$F(x) = 1 - e^{-(\frac{x}{\beta})^{\alpha}} \qquad (2)$$

これを x について解くと、以下のようになります（解き方については、後のコラムで解説します）。対数 log は底が e の自然対数です。

$$x = \beta(-\log(1 - F(x)))^{\frac{1}{\alpha}} \qquad (3)$$

では、（**3**）式を使って、故障する累積確率 F(x) の値が **95%**以上になる時間を求めてみましょう。自然対数は LN 関数で求められるので、図 6 のようになります。［ワイブル分布の逆関数］ワークシートを開いて試してみてください。

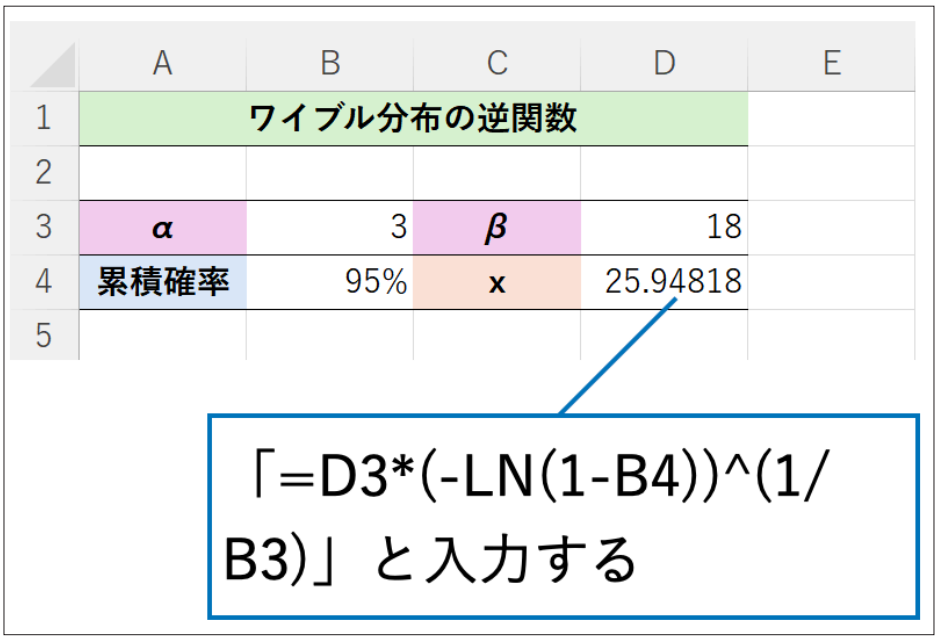


図 6　ワイブル分布の累積分布関数に対する逆関数を利用する
累積確率 F(x) の値はセル **B4** の **95%**、αに当たるのがセル **B3** の値、βに当たるのがセル **D3** の値。

セル **D4** に（3）式の計算を行う数式（「=D3*(-LN(1-B4))^(1/B3)」）を入力すると、**25.95** という値が得られます。これは、ほぼ **26 年**を過ぎると故障率が **95%以上**になることを意味します。

コラム　ワイブル分布の逆関数を自作する

Excel の LAMBDA 関数と名前機能を使うと、関数が自作できます。そこで、ワイブル分布の累積分布関数に対する逆関数の値を求める WEIBULL.INV 関数を自作してみましょう。操作方法は図 7 に示した通りです。［ワイブル分布の逆関数］ワークシートを開いて試してみてください。なお、LAMBDA 関数は Microsoft 365 の Excel やデスクトップ版の Excel 2021 以降でサポートされており、Excel 2019 以前のバージョンでは利用できません。

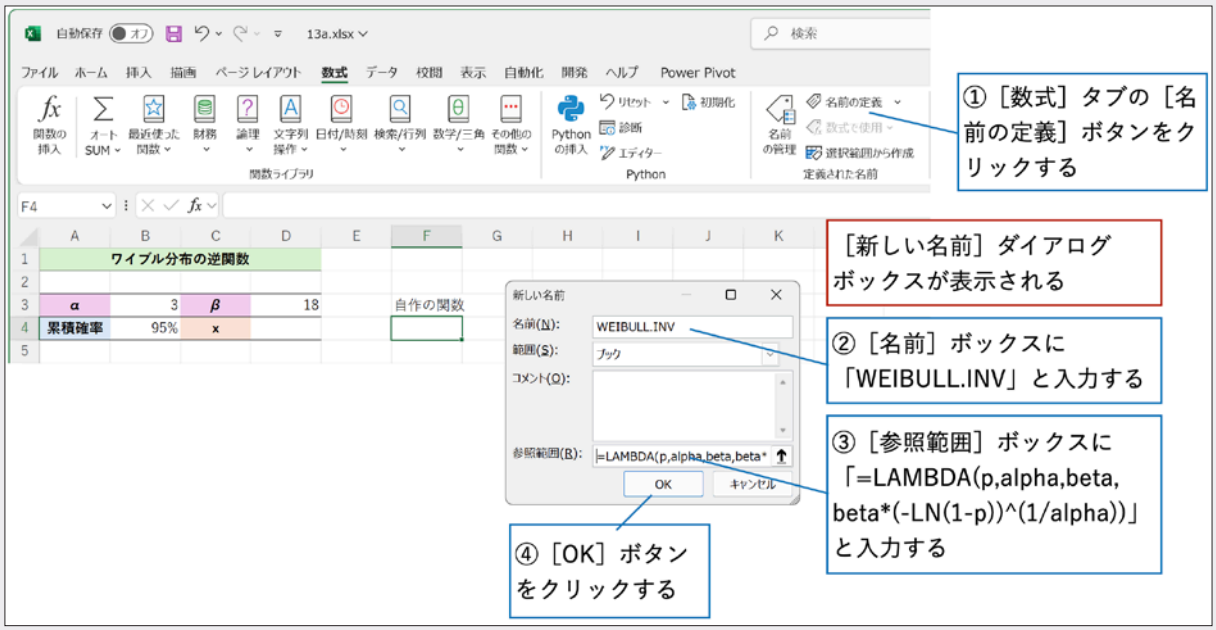


図 7　ワイブル分布の累積分布関数に対する逆関数を自作する

LAMBDA 関数を使って自作の関数を作成し、分かりやすい名前を付けておけばよい。ここでは、［範囲］を「ブック」としているが、特定のワークシートだけで使う関数であれば、［範囲］の一覧からワークシート名を選択すればよい。なお、［ワイブル分布の逆関数(完成例)］ワークシートでは WEIBULL.INV 関数をワークシート内だけで使えるものとして定義してある。

LAMBDA 関数は、この連載の記述統計編第 16 回でも取り上げました。上の例に則して簡単に解説しておきます（図 **8**）。

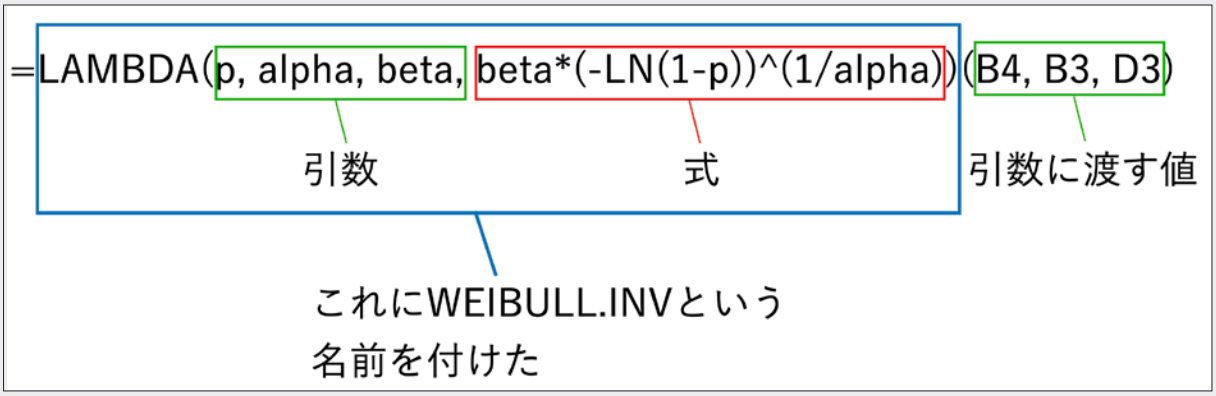


図 8　LAMBDA 関数の使い方

LAMBDA 関数の定義には「引数」と「式」を指定する。引数には分かりやすい名前を付けておけばよい。指定した引数を使って「式」の計算を行うという意味になる。実際に計算するときには、定義の後に () に囲んで指定した値が順に、引数に渡される。この例であれば、セル **B4** の参照が p に、セル **B3** の参照が alpha に、セル **D3** の参照が beta に渡される。図 7 では名前機能を使って、LAMBDA 関数の定義に **WEIBULL.INV** という名前を付けているので、=WEIBULL.INV(B4,B3,D3) と入力すれば「式」の計算ができる。

→目次に戻る

LAMBDA 関数を使えば、式の中で使われる引数（変数）と、式を記述するだけで、新しい関数が作成できます。さらに、名前機能を使って分かりやすい名前を付けておけば、自作関数の出来上がりです。自作の関数が作成できたら、空いているセルに =WEIBULL.INV(B4,B3,D3) と入力してみてください。図 6 と同じ **25.94818**... という結果が得られます。

なお、Google スプレッドシートでは、名前付き関数の機能を使って同じことができます。ただし、名前付き関数では関数名に「.」が使えないので、サンプルファイルでは「WEIBULL_INV」という名前にしてあります。名前付き関数の作り方はサンプルファイル内に記しておきます。

コラム　ワイブル分布の逆関数を求める

図 6 で使ったワイブル分布の累積分布関数に対する逆関数の式は、以下のようにして求めます。単に数式を変形しているだけなので、数式が苦手な方はスルーしてもらっても構いません。

累積分布関数 F(x) は以下の通りでした。

$$F(x) = 1 - e^{-(\frac{x}{\beta})^{\alpha}} \qquad (2)$$

（2）式の F(x) を右辺に、

$$e^{-(\frac{x}{\beta})^{\alpha}}$$

を左辺に移項します。

$$e^{-(\frac{x}{\beta})^{\alpha}} = 1 - F(x)$$

両辺の自然対数を求めると以下のようになります。

$$\log_e(e^{-(\frac{x}{\beta})^{\alpha}}) = \log_e(1 - F(x))$$

対数の公式 $\log_e e^a = a$ を適用すると、左辺の指数部分が取り出せます。

$$-\left(\frac{x}{\beta}\right)^{\alpha} = \log_e(1 - F(x))$$

両辺に − 1 を掛けて、

$$\left(\frac{x}{\beta}\right)^{\alpha} = -\log_e(1 - F(x))$$

両辺を 1/ α 乗すると、以下のようになります。

$$\frac{x}{\beta} = (-\log_e(1 - F(x)))^{\frac{1}{\alpha}}$$

両辺に β を掛ければ、(3) 式となります。

$$x = \beta(-\log_e(1 - F(x)))^{\frac{1}{\alpha}} \qquad (3)$$

故障率を求めてバスタブ曲線を確認する

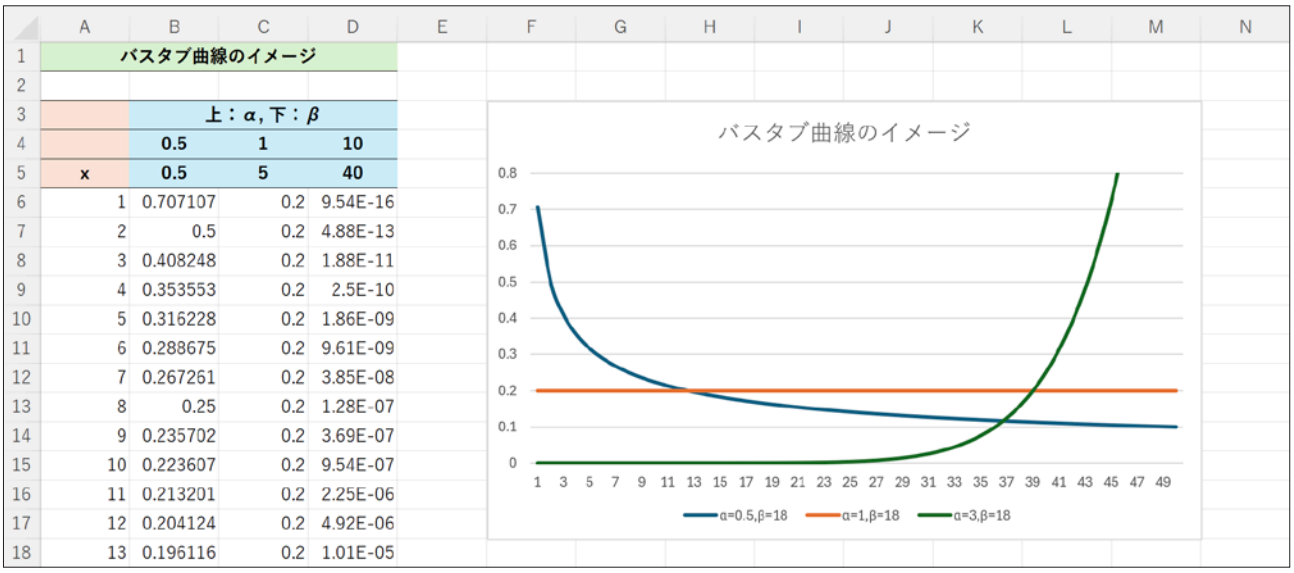
すでに述べたように、ワイブル分布は形状パラメーター α の値を変えることによって、以下のようなモデルの記述に使われます。

- $\alpha < 1$ …… 初期不良のモデル
- $\alpha = 1$ …… 偶発的な故障のモデル（指数分布）
- $\alpha > 1$ …… 経年劣化のモデル

故障率を可視化してこのことを確認してみましょう。故障率は以下の式で求められます。

$$\frac{\alpha}{\beta} \cdot \left(\frac{x}{\beta}\right)^{\alpha-1} \qquad (4)$$

ここでは、**バスタブ曲線**と呼ばれるグラフのイメージが分かるように、α と β にやや極端な値を指定した例を見てみます（図 9）。［故障率］ワークシートを開いて試してみてください。操作方法は図 9 の後に記しておきます。



	A	B	C	D
1	バスタブ曲線のイメージ			
2				
3		上：α, 下：β		
4		0.5	1	10
5	x	0.5	5	40
6	1	0.707107	0.2	9.54E-16
7	2	0.5	0.2	4.88E-13
8	3	0.408248	0.2	1.88E-11
9	4	0.353553	0.2	2.5E-10
10	5	0.316228	0.2	1.86E-09
11	6	0.288675	0.2	9.61E-09
12	7	0.267261	0.2	3.85E-08
13	8	0.25	0.2	1.28E-07
14	9	0.235702	0.2	3.69E-07
15	10	0.223607	0.2	9.54E-07
16	11	0.213201	0.2	2.25E-06
17	12	0.204124	0.2	4.92E-06
18	13	0.196116	0.2	1.01E-05

図 9　バスタブ曲線のイメージ

α <1 の場合、グラフの左側に見られるような初期不良の故障率が可視化できる。、$\alpha = 1$ の場合、グラフの中央付近で見られるような一定の故障率（偶発的な故障の率）となる。α >1 の場合、グラフの右端で見られるように、時間とともに故障率が増えていく。これは経年劣化を表しているものと考えられる。全体ではバスタブ（浴槽）のような形になっていることが分かる。

◆ **Excel での操作方法**

- セル **B6** に =B4:D4/B5:D5*(A6:A55/B5:D5)^(B4:D4-1) と入力する
  - 古いバージョンの Excel でスピル機能が使えない場合は、結果が求められるセル範囲（セル **B6** ～ **D55**）をあらかじめ選択しておき、関数を入力した後、入力の終了時に［Ctrl］＋［Shift］＋［Enter］キーを押す

◆ **Google スプレッドシートでの操作方法**

- セル **B6** に =ARRAYFORMULA(B4:D4/B5:D5*(A6:A55/B5:D5)^(B4:D4-1)) と入力する

● **グラフの作成方法**

- サンプルファイル内に掲載しておきます（タイトルや軸の書式などの細かい設定は省略）

セル **B6** に入力する =B4:D4/B5:D5*(A6:A55/B5:D5)^(B4:D4-1) という式はスピル機能にかなり慣れていないと分かりにくいかもしれません。取りあえず、（4）式に従って、セル **B6** に数式を素直に入力するなら =B4/B5*(A6/B5)^(B4-1) となります。αがセル **B4** に、βがセル **B5** に、x がセル **A6** に当たるので、（**4**）式そのままですね。ただし、この式ではセル **B6** の値しか求められません。そこで、全ての値を一度に求めるために、αとして B4:D5 を、βとして B5:D5 を、x として A6:A55 を指定したわけです。

図 9 に示されたバスタブのような形のグラフでは、故障率が時間とともにどのように変化するかが示されています。α <1 の場合は「初期不良」を表し、グラフの左側で故障率が高くなることが分かります。一方、α = 1 の場合は故障率が一定となり、「偶発的な故障」が起こる率を示しています。α >1 の場合は「経年劣化」が進み、時間の経過とともに故障率が増加する様子が右側に表れています。

形状パラメーターαと尺度パラメーターβの推定

ここまでは、ワイブル分布を利用して故障の確率や故障率を求めてきました。計算の方法は分かったと思いますが、αやβをどうやって決めるのか、という問題が残っています。これらの母数を推定する方法は、推測統計編の話題になるのですが、簡単に紹介しておきます。

母数を推定するためには、測定されたサンプルの値を使います。その方法には、最尤（さいゆう）推定法や最小二乗法（回帰分析）による推定法などがあります。

最尤推定法とは、サンプルを基に、母数として「最も尤もらしい」（もっとも、もっともらしい）値を推定する方法です。簡単に言うと、各試行の確率の積を母数の関数（尤度関数、「ゆうどかんすう」と読む）とし、その関数を最大にする母数の推定値を求めるというものです。ただし、積（掛け算）の形だと、微分などの計算が面倒なので、和（足し算）にするために対数を取った対数尤度関数を求め、対数尤度関数を最大にする母数の推定値を求めるのが普通です。後のコラムで簡単な例を紹介します。

Excel ではいずれの方法もかなりの手間がかかるので、ここでは、Python のプログラムを使って、最尤推定法による母数の推定を行ってみましょう。

サンプルプログラムはこちらから参照できます。リンクをクリックすれば、ブラウザが起動し、Google Colaboratory の画面が表示されます（Google アカウントでのログインが必要です）。最初のコードセルをクリックし、［Shift］+［Enter］キーを押してコードを実行してみてください。コードの詳細については解説しませんが、コード内のコメントとリスト 1 の説明を見れば何をやっているかが大体分かると思います。

```
from scipy.stats import weibull_min

# 故障した年数
data = [5, 10, 13, 13, 16, 17, 18, 21, 23, 26]

# 母数の推定
shape, loc, scale = weibull_min.fit(data, floc=0)
print(shape, scale)

# 出力例
3.0420091560120834 18.139000902259845
```

リスト 1　最尤法によりワイブル分布の母数を推定する

scipy.stats モジュールの weibull_min は、故障時間や寿命などの分析に使われるクラス。fit 関数にデータを指定すれば、形状パラメーター α、分布の開始位置、尺度パラメーター β の推定値が求められる。floc=0 は分布の開始位置を **0** に固定するという意味。ここでは、形状パラメーターと尺度パラメーターのみを表示した。

結果は形状パラメーターの推定値が **3.04**、尺度パラメーターの推定値が **18.14** となっています。最初に示した事例で $\alpha = 3$、$\beta = 18$ としたのは、このようなデータから推定された母数を想定したものだったから、というわけです。

なお、上記のサンプルプログラムのノートブックファイルには、最小二乗法により母数の推定を行った例も含めてあります。ただし、サンプル数が少ない場合は最尤推定法との違いが大きくなるので、**1000 個**のサンプルデータとしてあります。Google Colaboratory の 2 つ目のコードセルをクリックし、［Shift］+［Enter］キーを押して実行すると、回帰分析による推定と最尤推定法の結果を比較できます（ほぼ同じ値になります）。

参考として、Excel での最尤推定法による母数の推定と最小二乗法による母数の推定を行った例も、Excel のサンプルファイルに含めてあります。ただし、最尤推定法の場合は、対数尤度関数を最大化する母数を求めるために、あらかじめソルバーアドインを組み込んでおく必要があります。詳細については割愛しますが、サンプルファイル中に簡単な説明を掲載しています。

コラム　最尤推定法のごく簡単な例

最尤推定法がどのようなものであるか、実感が湧かないかもしれないので、ここでは、できるだけ簡単な例としてベルヌーイ試行を取り上げ、最尤推定法による母数の推定を手計算でやってみたいと思います。例えば、コインを投げて、表が出たことを **1**、裏が出たことを **0** と表すものとして、**5 回**の試行が [1, 1, 0, 1, 0] という結果になったとします。

ベルヌーイ分布の母数 p（表が出る確率）を最尤推定してみます。表が出る確率は p、裏が出る確率は 1 − p です。最初に出た「**1**」という値が現れる確率として「もっともらしい」と考えられる値は、p ですね。この値を「尤度（ゆうど）」と呼びます。次に出た「**1**」の尤度も p です、その次に出た「**0**」の尤度は 1 − p ですね。このようにして求めた全ての試行の尤度（確率）を掛け合わせ、それを関数（尤度関数）と見なして、f(p) と表すと、

$$\begin{aligned} f(p) &= p \times p \times (1-p) \times p \times (1-p) \\ &= p^3(1-p)^2 \end{aligned}$$

となります。この尤度関数の値を最大にする p の値を母数の推定値とします。この例は簡単なので、このままでも 0<p<1 の範囲で f(p) を最大化する p の値を求めることはできますが、普通は対数を取った対数尤度関数とします。

$$\begin{aligned} \log(f(p)) &= \log(p^3(1-p)^2) \\ &= \log p^3 + \log(1-p)^2 \\ &= 3\log p + 2\log(1-p) \end{aligned}$$

この関数は 0<p<1 の範囲で上に凸となっているので、最大値を持ちます。そこで、微分して **0** と置きます。

$$(3\log p + 2\log(1-p))' = 0 \qquad (5)$$

対数関数の微分の公式

$$(\log p)' = \frac{1}{p}$$

を適用して、左辺を変形すると、以下のようになります。ただし、log(1 − p) の微分は、

$$\frac{1}{1-p}$$

ではありません。1 − p を u と置いて合成関数の微分を行う必要があります。計算の過程は省略しますが、結果は、

$$(\log(1-p))' = -\frac{1}{1-p} = \frac{1}{p-1}$$

となります。従って、（5）式は、以下のようになります。

$$\frac{3}{p}+\frac{2}{p-1}=0$$

0<p<1 なので、分母が **0** にならないことも確認できますね。通分すると、

$$\frac{3(p-1)+2p}{p(p-1)}=0$$

となるので、両辺に p(p − 1) を掛けて、

$$\begin{aligned}3(p-1)+2p&=0\\5p-3&=0\\p&=\frac{3}{5}\end{aligned}$$

となります。結局のところ、**5 回**のうち **3 回**表が出たので、表が出る確率 p の最尤推定値は 3/5 で求められますが、最尤推定法による母数の推定を手計算でやるとこのようになるというわけです。

◇ ◇ ◇ ◇ ◇ ◇ ◇

今回はワイブル分布について、簡単な利用例を紹介した後、確率密度関数と累積分布関数の求め方についてお話しし、さらに、推測統計の領域に少し踏み込んで、母数の推定方法も紹介しました。

さて、離散型確率分布の二項分布から始まったこの連載も、超幾何分布やポアソン分布、連続型確率分布の代表とも言える正規分布、さらには、カイ二乗分布、t 分布、F 分布、ベイズ統計などでよく使われるベータ分布を経て、今回で最終回となりました。

この連載では、確率分布の利用例を紹介する中で、必要に応じて推測統計についても個別に説明しました。しかし、具体的な事例の紹介や体系的な解説はしていません。そこで、この連載の続編として『社会人 1 年生から学ぶ、やさしい推測統計』で、推測統計についての詳しい解説を準備しています。どうぞお楽しみに！

この記事で取り上げた関数の形式

関数の利用例については、この記事の中で紹介している通りです。ここでは、今回取り上げた関数の基本的な機能と引数の指定方法だけを示しておきます。

ワイブル分布の確率密度関数や累積分布関数の値を求めるための関数

WEIBULL.DIST 関数：ワイブル分布の確率密度関数や累積分布関数の値を求める

形式

WEIBULL.DIST(x, α , β , 関数形式)

引数

- **x**：故障するまでの時間などを指定する。
- **α**：形状パラメーターを指定する。
- **β**：尺度パラメーターを指定する。
- **関数形式**：以下の値を指定する。
  - **FALSE** …… 確率密度関数の値を求める
  - **TRUE** …… 累積分布関数の値を求める

LN 関数：自然対数の値を求めるための関数

形式

LN(x)

引数

- **x**：自然対数を計算する際の対象となる数値（真数）を指定する。

LAMBDA 関数：独自の関数を作るための関数

形式

LAMBDA(引数 1, 引数 2, ..., 式)

引数

- **引数 1, 引数 2, ...**：式に与える引数に対応する変数名を定義する。
- **式**：変数を使った計算の方法を指定する。

備考

例えば、「=LAMBDA(x, y, x+y)」であれば、最初の x に第 1 引数の値が渡され、y に第 **2** 引数の値が渡される。最後の x+y の値が答えとして返される。従って、セルに「=LAMBDA(x, y, x+y)(A1,A2)」と入力すると、セル **A1** の値が x に渡され、セル **A2** の値が y に渡され、x+y の値つまり **A1** と **A2** の和が返される。さらに、名前機能を利用して「=LAMBDA(x, y, x+y)」に myadd といった名前を付ければ、「=myadd(A1, A2)」でセル **A1** とセル **A2** の和が求められる。このようにして関数を自作できる。

なお、LAMBDA 関数は、Excel 2019 以前では使えない。

筆者紹介

羽山博

IT 系ライターの傍ら、非常勤講師として東大で情報・プログラミング関連の授業を、一橋大で AI 関連の授業を担当。趣味の献血は心拍数が基準を超えてしまい 99 回で中断。心肺機能を高めるために水泳を始めるも、一向に上達せず。また、リターンライダーとして何十年ぶりかに大型バイクにまたがるも、やはり体力不足を痛感。足腰を鍛えるために最近は四股を踏む日々。超安全運転なので、原付やチャリに抜かされることもしばしば（すり抜けキケン、制限速度守ってね！）。

編集：@IT 編集部
発行：アイティメディア株式会社