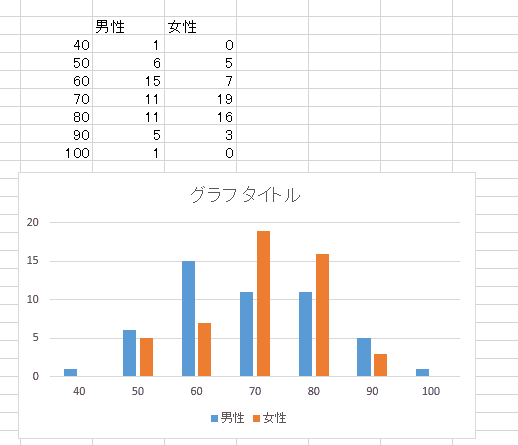
【第7部】様々な分布

性別によるテストの成績の分布から「男女の間にテストの成績の差がある」といえるでしょうか。差がある、もしくは差がないとするならば数学的に検証（検定といいます）する必要があります。

検定するうえでは「分布」を理解する必要があります。特に正規分布の概念を理解し、正規分布から派生するその他の分布と、分散や平均などの「数値の意味」を理解する必要があります。

　理解したうえで分析すると、性別による成績の分布には「差がない」という結論になります。どうしてそのような結論になるか、まずは分布の理解からはじめましょう。



# 二項分布

二項分布とは「n回のベルヌーイ試行（表と裏など、2つの種類しかない試行）の中で何回成功するか」という成功回数の分布です。「サイコロを10回振った時の目の合計が30になる確率」や「コインを10回投げたときに表が5回出る確率」などを考えるときに使います。高校の確率の授業などでも取り上げられる、基本的な分布の１つです。

試行回数をn 、成功する回数をx 、成功確率をpとした場合、n回中x回成功する確率f(x)は

となります。の部分は二項係数として　　と表記されることもあります。 例えばコインを投げて表が出る確率確率を50％としたとき、2回投げて1回しか表が出ない確率は、

2\*25%=50%

となります。

## メレとパスカル

シュバリエ・ド・メレは友人のパスカルに相談しました。「1つのサイコロを4回投げて6が1回でも出れば勝ちという賭けでは勝てた。しかし2つのサイコロを24回投げて6のゾロ目が出れば勝ちという賭けになってから勝てなくなった」というものです。

パスカルは勝率を計算しました。１つのサイコロを4回投げて一度も6が出ない確率は＝0.482なので、最初の賭けの勝率は1－0.482＝0.518＝51.2％となります。２つのサイコロを24回投げて1度も6のゾロ目が出ない確率は＝0.509なので2番目の賭けの勝率は1-0.509=0.491=49.1％となります。確かに最初の賭けは勝率5割を超えますが、2番目の賭けは勝率5割を切ります。メレの勝てなくなったという証言は正しいといえます。

## ギャンブラーの法則

では2番目の賭けはどうすれば有利なるのでしょう。事象の起きる確率をPとしたときに、それをn回したときに勝率が5割を超えれば勝てるとすると、nを1-(1-p)n>50%となるようにすれば良いことになります。確率が50％になるnを求めると

両辺の対数をとると

両辺をで割って01𝑛𝑛𝑗𝑖 =

となります。pにとをそれぞれ入れると、前者や約3.8、後者は約24.6になります。つまりメレは前者であれば4回以上、後者であれば25回以上行うようなゲーム設計にすれば賭けに有利になるといえます。

## 二項分布の練習問題

【問題1】

コインを10回投げたとき、表が3回出る割合を求めなさい。（コインは表と裏が出る確率はそれぞれ50％とする）

【問題2】

１～６までの目があるサイコロを10回投げたとき、合計が35になる確率を求めなさい。（サイコロの目が出る確率はそれぞれとする。）

# ポアッソン分布

ポアソン分布とは、単位当たりにn回の事象が起こる割合の分布です1日あたりの交通事故の数、大量生産された部品において不良品が発生する割合などは、ポアソン分布になります。ポアソン分布は二項分布において成功確率が非常に低い場合の分布です。

## 二項分布からポアソン分布を作る

ポアソン分布の公式は確率λと回数xに対して、eは自然対数の底とすると

とあらわされます。この公式は二項分布からも導き出せます。

1. ポアソン分布の成功回数をλ=npとします。として二項分布に代入します。
2. 一部を入れ替えます
3. nの極限（無限に大きくする）をとります。の極限は1になります。 の極限はに、 は0になります。

1時間に1回しか起こらない出来事が、1時間に2回起こる確率は

eは自然対数の底（2.718281828）なので、e-1は0.367879、これを1/2にすればいいので、0.183939721、つまり1時間に1回しか起こらない出来事が、1時間に2回起こる確率は約18.4％です。

## 待ち時間を計算してみよう（指数分布）

では非常に希な事象が起こる時間間隔を計算してみましょう。時間tが経過した時に事象が起きる確率f(t)と確率密度f(t)dtを求めます。

単位時間当たりで事象が起こる確率はλなので、時間0からtまでに起こる確率はλtとなります。

まずは事象が時間0からtの間にx=0回起こる確率を求めます。ポアソン分布を使って計算すると、

次に時間0~xのあとの微細な時間t~dtの間に事象が起こる確率を求めます。これはλにdtを掛けたλdtです。よって事象が起こる時間間隔xの確率密度関数f(t)Δtは時間0~xまでに起こらなかった確率に時間x~Δxまでに起こった確率を掛ければいいので

両辺からdtを割ると

となります。この分布は指数分布と呼ばれています。

## ポアソン分布の練習問題

【問題1】

学食のカレーには肉が平均2つ入っているとします。カレーに肉が１つしか入っていない確率と、3つ入っている確率をそれぞれ求めなさい

【解答】

ポアソン分布はλ＝2　x=3の時の確率を求める

　答え：18.0%

【問題2】

　1000年間に1回しか出現しないアイドルが10年間に2回出現する確率を求めなさい。

【ヒント】λに0.01、xに2を入れて計算します。答えは0.0050%です。

# 正規分布

正規分布は様々な確率・統計の基本的な概念です。正規分布（またはガウス分布）とは、平均を中心としてグラフにすると「釣り鐘状」「富士山型」などといわれる形になる確率分布です。正規分布は様々な指標にも応用されています。大学入試で難易度の指標などに使われる偏差値（平均50、分散10）や、知的能力を推し量る検査に使われる知能指数(平均100、分散15）などがあります。

平均μ、標準偏差σのとき、正規分布の確率密度は

となります。平均0で分散1の時は

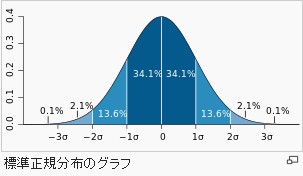
という確率密度関数であらわされ「標準正規分布」といわれています。

## 正規分布と区間確率（シグマ区間）

正規分布では、平均から±σの範囲に68.3％が、平均から±2σの範囲に95％、平均から±3σの間に99.7%が入ります。

偏差値では、±σが40～60、±2σが30～70、±3σが20～80に相当します。知能指数（IQ)では、±σが85～115、±2σが70～130、±3σが55～145に相当します。

1学年400人(1クラス40人×10クラス）とすると、偏差値80・知能指数145の人は上位から0.135%あたりにいるので、人数にして0.5番目、学年に1人いるかいないかの天才です。偏差値70・知能指数130の人は上位から2.28％あたりにいるので9番目、クラスに1人いるかいなかの頭のいい人です。偏差値60だと上位15.87%なので63番目前後、クラスで上位6～7番目ぐらいの人に相当します。



## 広告業界でも利用される正規分布の概念

この概念は広告業界でも使われています。世の中のブームに敏感な人を正規分布にあてはめて分類しています。

１）イノベーター

平均から約＋2σ以上の人（上位2.5％）、最初にブームに乗っかる人を「イノベーター」

２）アーリーアダプター

平均から＋2σ～＋1σ（上位2％～13.5％）、流行始めたときブームを広げる人たちを「アーリーアダプター」

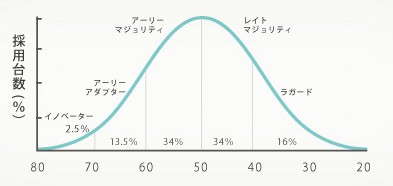
３）マジョリティ

平均から±σ（平均付近の68％）、普通の人をマジョリティといい、平均から上の人をアーリーマジョリティ、平均から下の人を例と「マジョリティ」といいます。

４）ラガード

偏差－σ以下の人たちはブームには全く関心のない「ラガード」といいます。

ジェフリー・ムーアのキャズム理論では偏差＋σが分岐点としています。マジョリティに受け入れられることを「キャズムを超える」と表現して、一般に普及した指標として使っています。



## 正規分布の練習問題

【問題1】

40人のクラスで数学テストを行ったところ、平均点が65点、分散が13点でした。Ａさんはそのテストで78点を取りました。Aさんの成績はクラスで何番目ぐらいか予測してください。

【回答】

79点は平均＋1σなので上位15.87%、40人クラスなので、40人×15.87%で6.4人。つまり6番目～7番目あたりである。

【問題2】

2015年のセンター試験の受験者数は559,131人でした。センター試験が正規分布に従うとします。偏差値は平均50で分散が10の正規分布です。上位1,000人に入る人の偏差値はどれくらいか求めてください。

# t分布と検定

t分布とは正規分布からサンプルを抽出したときの分布です。例えば、高校生の身長を調べるため30人抽出して調査したとします。抽出に細心の注意を払ったとしても、標本平均と母平均、標本分散と母分散は一致するとは言えません。標本の値には必ず「誤差」が入ります。この誤差を表したものがt分布です。

　これを詳しく説明します。平均μで分散σ2の正規分布の母集団から、n個の標本を取り出した場合、標本と母集団の平均にはσ２/nの誤差が出ます。わかりやすく例えると無数の母集団から1人だけ標本を抽出したら、それは分散そのものなので母平均と標本平均の差はσ２です。無数の母集団から1,000人のサンプルを抽出すれば、σ2/1000は0に近づきます。つまり母平均と標本平均の差はほとんどなくなります。t分布はnが無限になった時に正規分布、つまり母分布と一致します。

## t分布の公式（及びガンマ分布の公式）

t分布の公式は下記です。

Γはガンマ分布です。希に起こる現象がn回までの生起に要する時間分布などに使われる分布です。Γ分布の公式は下記です。

## t分布の応用

t分布を利用して、2つの分布が同じ分布か別の分布であるかを検定することができます。検定方法としてはt検定、もしくは回帰分析などがあります。また標本平均から母平均を推定するにあたり、信頼区間を求める際にもt分布の概念を利用します。