# 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。

【信頼区間：サンプル調査によって全体の平均値を求める】

# サンプルから全体を推測するには

10万個のリンゴが収穫されたとします。このリンゴの重さの平均値を知りたいという場合を考えてみましょう。平均値を知りたいならば、10万個全ての重さを計れば正確な平均値を知ることができます。しかし全部を計るのは大変な手間がかかり、現実的ではありません。

そのような場合は、全体から一部のリンゴを抜き取って調査し、平均値を推測します。もちろん、一部のリンゴの平均値が、全体の平均値を正確に表すわけではありません。サンプル調査には必ず「誤差」が生じます。この誤差を考慮して「どうも平均値は〇〇～〇〇の間にあるようだ」と推定します。この〇〇～〇〇の区間のことを「信頼区間」といい、このような推定方法を「区間推定」といいます。

# 標本調査の誤差と信頼区間

100個のリンゴを抜き取って重さを計ったところ、平均値が300gでした。さて全体の平均値はどれくらいになるでしょうか。

統計的には、抜き取った際の「誤差」を考慮して「一定の間隔の間に全体の平均値がある」と考えます。誤差が1gと計算されたなら、その誤差を考慮して300g±1g、つまり「全体の重さの平均は299gから301gの間である」とします。

これを統計的な用語で説明すると、

* 10万個のリンゴ 「母集団」
* すべての重さの調査 「全量調査」
* 全体の平均　 「母平均」
* 抜き取られた一部 「標本」または「サンプル」
* 一部を抜きとる調査 「標本調査」または「サンプル調査」
* 一部の平均値 「標本平均」または「サンプル平均」
* 考慮された誤差の範囲 「信頼区間」
* このような推定方法 「区間推定」

といいます。

## 区間推定・信頼区間の考え方

「母集団の平均は300gだ」という決め打ちではなく「母集団の平均はこの幅の中にある」と数値に幅を持たせた推定（区間推定）には「信頼度」という概念が必要になります。

数値幅を大きくすると、母平均がその幅の中に入っている確率は高くなりますが、幅があまりに大きすぎるとその幅にあまり意味がなくなってしまいます。逆に幅が狭すぎると母平均がその幅の中に入っている確率が低くなります。そのため「その幅にどの程度信頼性があるか」ということをあらかじめ定めておきます。

標本調査においては、20回に19回ぐらい推定が当てはまる程度の幅を「信頼度95％の信頼区間」、100回に1回ぐらい推定が当てはまる程度の幅を「信頼度99％の信頼区間」といいます。

通常は信頼度95％の信頼区間を使い、より正確な推定をしたい場合は信頼度99％の信頼区間を使います。

# 信頼区間の求め方

信頼区間を求めるには

1. サンプル数が多いか少ないか（30以上か30未満か）
2. 母集団の分散が既に判明しているか否か
3. 母集団が正規分布と判明しているか否か

によって求め方が変わってきます。

サンプル数については経験則として30以上か否かで判断をします。30以上であればサンプル数が多い（大標本）としてみなします。サンプル数が多い場合は母集団が正規分布かどうか判明していない場合でも、中心極限定理を援用して、母集団が正規分布で近似できるとみなして推定します。

サンプル数30未満の場合は母集団が正規分布で分散もあらかじめ判明しているときのみ推定が可能です。分散が未知の場合や、分布が正規分布になっているか否かが未知の場合、推定はできません。

### 分散σ2が判明している（既知）の場合

母集団の分散（母分散）σ2が判明している場合を考えてみましょう。n個のサンプルの値をx1,x2,x3,x4・・・xｎとすると、

標本平均

標本分散

となります。平均μで分散σ2の正規分布の確率変数XはX~N(μ, σ2）と表記されるので、サンプルの分布の確率変数,は

となります。標本平均からμを引くと平均が0になるので

さらに分布を（偏差）で割ると分散は１になるので

となります。これは標準正規分布と同じです。

標準正規分布とみなせるようになったので、標準正規分布の確率変数がαから１－α、つまり信頼区間95％ならば確率変数が2.5％～97.5％の間、信頼区間99％なら確率変数が0・5％～99.5％の間にある場合を求めます。

標準正規分布の確率xの時の確率変数をZ(x)とすると

と表すことが出来ます。この不等式を平均μで解くと、すべての辺をで割ると

となるので、を引いて、マイナスをかけると

となります。正規分布で2.5％、0.5％となるはそれぞれ約1.96と2.58なので、

信頼度95％の信頼区間は

信頼度99％の信頼区間は

となります。

### 分散σ2が判明していない（未知）の場合

母集団の分散（母分散）σ2が判明していない場合を考えてみましょう。n個のサンプルの値をx1,x2,x3,x4・・・xｎとすると、

標本平均

標本分散　＝

となります。既知の場合とちがって、分散を分布の中から計算する必要があります。これは自由度(n-1)のt分布になるが、tが十分大きい時は正規分布とほぼ一致するので

となります。正規分布で2.5％、0.5％となるはそれぞれ約1.96と2.58なので、

信頼度95％の信頼区間は

信頼度99％の信頼区間は

となります。

## リンゴの重さの信頼区間を求める（分散が既知の場合）

10万個のリンゴから100個のリンゴを抜き出して調査をしました。過去の調査からリンゴの重さの分散が9gだったとします。

1. 標本の数 n　　＝100個
2. 標本平均 　　＝300g
3. 母分散 σ2 ＝9g

信頼区間95％であれば

なので、それぞれに数字を代入して、300g±(1.95×3/10)= 300g±0.59gとなります。

信頼区間99％であれば

なので、それぞれに数字を代入して、300g±(×3/10)= 300g±0.77gとなります。

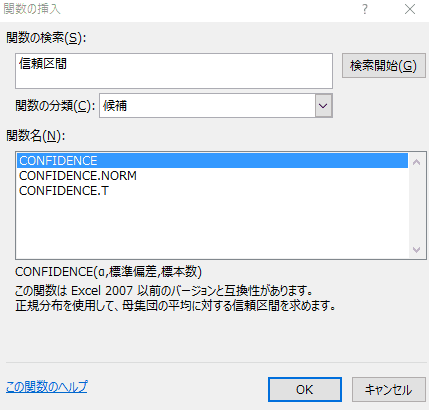
# EXCELでの信頼区間の求め方

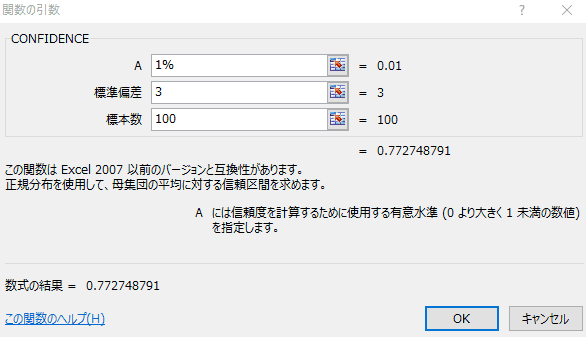
## 分散や平均が分かっているとき

Excelには信頼区間を求める関数が用意されています。CONFIDENCEという関数です。セルに直接「= CONFIDENCE(　」と入力しても良いですが、判らないときは、セルを選択した状態で数式バーの左にある「f(x)」をクリックし、関数の検索のダイヤログを開きます。

ダイヤログの「関数の検索(S)」に「信頼区間」と入力して「検索開始(G)」を押します。すると「関数名(N)」にいくつか候補が出てきます。CONFIDENCEはexce2007からあるもの、CONFIDENCE.NORMは正規分布として計算するもの、CONFIDENCE.Tはt分布として計算するものです。今回はCONFIDENCEを利用して演習してみましょう。

関数名のCONFIDENCEを選択した状態でダイヤログ下部の「OK」を押します。「関数の引数」のダイヤログが開くのでAにαに相当する数字、信頼区間95％を求めるなら5％もしくは0.05、信頼区間99%を求めるなら1％もしくは0.01を入力します。標準偏差と標本数はそれぞれ数字を入れてダイヤログ下部の「OK」を押します。するとセルに結果が表示されます。





## 平均や分散も一緒に求めたいとき

ある数列から平均や分散も一緒に求めたうえで信頼区間を求める場合は、分散、標本数、平均をそれぞれ求めたうえでCONFIDENCEや、CONFIDENCE.T（母集団の分散・偏差が分からないとき）を利用します。

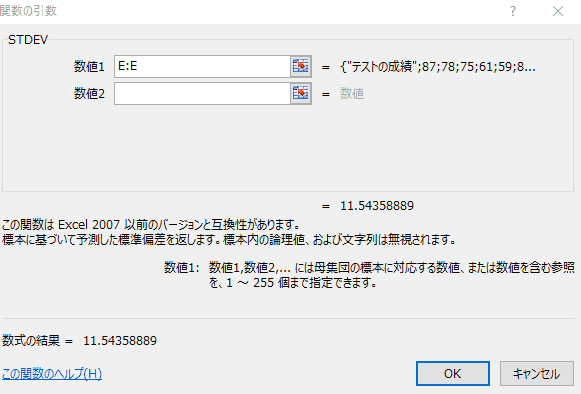
演習シートのA学校データを開いてみましょう。テストの成績を例に信頼区間を求めてみます。

A学校データには100人分のテストのデータがあります。これのデータからその学年全体の平均点を推定してみましょう。

必要な数字はAと標準偏差と標本数です。G列に2行目から順番に「α」「偏差」「標本数」「平均」「信頼区間」と入力してそれぞれを求めましょう

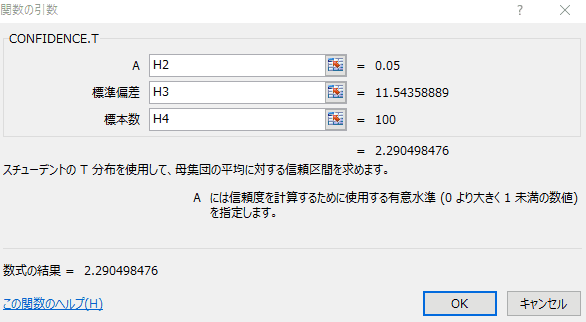
### Excelを利用した求め方

Aは信頼区間なので5％もしくは1％（0.05か0.01）と入力します1。次に偏差を求めます。偏差はSTDEVという関数を利用します。「=STDEV(　」と入力したら数式バーの左にある「f(x)」を押して関数のダイヤログを開きます。「数値１」に範囲を指定します。文字は無視してくれるのでE列全体を選びます。同様に標本数は「COUNT」を、平均は「AVERAGE」を利用してE列全体を数値１に選びます。



最後に信頼区間を求めます。「=CONFIDENCE.T(　」と入力して数式バーの左にある「f(x)」を押して関数のダイヤログを開きます（もしくは関数の検索から信頼区間を検索してダイヤログを開きます）

A、標準偏差、標本数には先ほど求めたそれぞれのセルを入れます。H2,H3.,H4と直接入力しても良いですし、カーソルが入った状態でそれぞれのセルを選択することでも入力できます。完成したらダイヤログ下部の「OK」を押して完成です。信頼区間は2.29となり、平均から考えるとこの学校全体の平均点は74.9±2.3といえます。



# 課題

　K大学の卒業生120名の平均年収を調べたところ、平均1500万円、偏差が700万円であった。このとき卒業生全体の平均年収の95％の信頼区間を求めよ。ただし年収の分布は正規分布と仮定する。

【回答】

　1500±（1.96×700÷）＝126.53　なので、1500万円±127万円、つまりK大学卒業生の平均年収は1373万円～1627万円の間にあるといえる。

# 【コラム】サイコロは記憶を持たない？

喫茶店で働いているドジっ子なMさん、2日に1回はお皿を割ってしまいます。しかしなぜかここ10日間は1枚もお皿を割っていません。ドジっ子ぶりは相変わらずで、周囲はまたお皿を割ってしまうのではないかとハラハラしています。そこでMさんが次にお皿を割るまでの時間の確率分布を求めてみましょう。

確率分布は「2日に1回起こる事象」が「10日間も起こらなかったとき」、それ以降にその事象が起こるまでの時間の確率を求めればよいことになります。これは時間S（ここでは10日）まで無事である確率で、時間S~S+tで起こる確率を割ると求められます。

まず時間Sまで無事の（事象が起こらない）確率を求めます。時間Sまで無事ということは時間S~∞までの間で壊してしまう(事象が起こる)ということですから時間Sから∞まで指数分布を積分した

という式になります。逆に時間Sまでに無事でない確率は成功しない確率なので

となります。

10日間まで何事もなかったとき、そのあと事象が起こる割合は、時間S（ここでは10日）まで無事である確率で、時間S~S+tで起こる確率を割るので

となります。これは時間Sまでに無事ではない確率とまったく一緒です（時間Sを時間tに変えただけ）。それまで事象が起こったか起こらなかったかに関わらず、次の事象が起こるのはその時点からみて時間tに対しての割合で発生します。

この、事象が起こる確率が、事前に事象が起こったか起こらなかったかに関わらないことを「無記憶性」といいます。無記憶性をもつ絶対連続な分布は指数分布だけです。サイコロは記憶を持たない、とも言いますが、数式上でもそれは証明されています。