【第9部　2つのデータを比べる　その１]

2つのデータを比較して、差が「有意か否か」を検定してみましょう。

# t検定とは

２つのデータを比べてみて「有意な差」があるかを確認するのが「t検定」です。t検定では信頼区間の概念を利用して、2つの分布が同じものか違うものかを判定します。

１つの分布からランダムに２つの標本を抜き出した場合、その平均値はそれぞれt分布になりますが、それぞれの平均値の「差」もt分布になります。この「差」がどれぐらい偶然に起こったものなのかを考えるのがt検定です。

信頼区間の概念と同じく、偶然起きる差が95％（20回に19回）以内、もしくは99％（100回に99回以内）であれば、２つの分布は同じであり「有意な差はない」とします。

逆に20回に1回、もしくは100回に1回起こるような偶然の差以上であれば、それは同じ分布から抜き出した２つの標本ではなく、差がある別の分布から抜き出した標本だと判断します。

## t検定の方法

t検定は標本平均の差をt分布としてt値を求めます。標本Aの標本数を、平均を、分散を、標本Bの標本数を、平均を、分散を、とすると、標本平均の差を標本平均の差の標準誤差で割ればtが求まります。数式にすると

となります。これがを自由度するt分布に従います。

### ｔ値から推測する

t値を求めたら、t分布表に従って分布表の値より小さければ「2つの標本は同じ分布から抜き取られたもの」と判断し「差が生まれたのは偶然である」と判断します。t値が分布表の値より大きければ「２つの標本は同じ分布から抜き取られたものではない」と判断し「差が生まれたのは偶然ではない」として「有意な差がある」と判断します。

サンプル数が多い場合は信頼度95％で約2、信頼度99％で約2.6なので、2以上なら信頼度95％で有意、2.6以上であれば信頼度99％で有意であるとします。

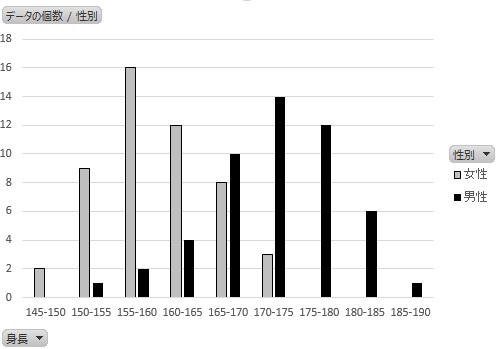
t分布表

|  |  |  |
| --- | --- | --- |
| 自由度 | 信頼度95% | 信頼度99% |
| 1 | 12.706 | 63.657 |
| 2 | 4.303 | 9.925 |
| 3 | 3.183 | 5.841 |
| 4 | 2.776 | 4.604 |
| 5 | 2.571 | 4.032 |
| 6 | 2.447 | 3.707 |
| 7 | 2.365 | 3.500 |
| 8 | 2.306 | 3.355 |
| 9 | 2.262 | 3.250 |
| 10 | 2.228 | 3.169 |
| 15 | 2.132 | 2.947 |
| 20 | 2.086 | 2.845 |
| 30 | 2.042 | 2.750 |
| 40 | 2.021 | 2.705 |
| 60 | 2.000 | 2.660 |
| 120 | 1.980 | 2.617 |
| ∞ | 1.960 | 2.576 |

# 演習

A学校の男女別の身長・体重・テストの成績に有意な差があるかt検定を

利用して検定してみましょう。総務省統計局のデータによれば、18歳の男性の平均身長は171.1cm、女性の平均身長は158.4cmとなっています。下図は18歳男女の身長分布のダミーデータから作ったグラフです。男子は150㎝～155cmに1人いて、170㎝～175㎝の平均の部分が14人と最も多くなります。175㎝以上からは徐々に人数が少なくなります。女子は145㎝～150㎝に2人いて、155㎝～160cmの平均の部分が16人と最も多くなります。160㎝以上から徐々に人数が少なくなるのは男子と同じです。このグラフを見る限り、明らかに2つの違う分布があるのが見て取れます。



## 公式通りの計算をして求める

t検定をするために公式通りにt値を求めて検定してみましょう。

### 計算するための枠を作る

まずt値を入れる枠を作ります。表の右隣に「標本数」「標本平均」「標本分散」「平均偏差の平方和」と縦に順番に入力します。標本数はCOUNT、標本平均はAVERAGE、標本分散はVARPを利用して、それぞれ男子の身長（C2:51）と女子の身長(C52:C101）を「値１」に入力します。

あとは公式にそって計算式を書けばよいのですが、一度に書こうとすると大変なので、分子と、分母の右辺・左辺と分けて計算してみましょう。

### 分母の左辺を作る

まず分母の左辺ですが、１を標本数で割ったものを足します。まず「＝（1/　」を入力した後、男子の標本平均のセルを選択します。そうすると「＝（1/H2　」となるのでカッコで閉じます。同様に「＝（1/H2）+(1/　」と入力し女子の標本平均のセルを選択します。最後にカッコを閉じて完了です。「＝(1/H2）+(1/I2）」と入力され、0.04と表示されます。

### 分母の右辺を作る

次に、分母の右辺です。なので、標本分散を足したものを標本数―１を足したもので割ります。最初に右辺の分子を作ります。

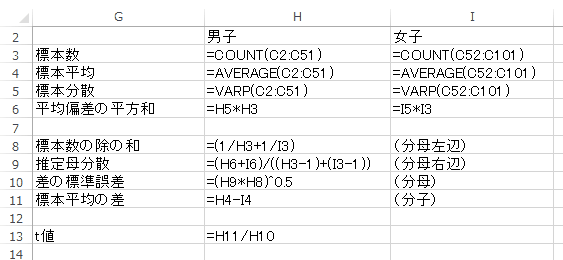
「＝(　」と入力したら男子の標本分散を選択し、「＝(H6」とします、「＋」をいれて女子の標本分散を選択したうえで一度カッコを閉じます。「＝(H6+I6)」これで右辺の分子部分が出来ました。つづけて「 / 」を入力して右辺の分母を作ります。分母部分はまずカッコをいれて「＝(H6+I6)/( 」としたうえで、再度カッコを入れます(かっこが2つになります）。次に男子の標本平均を選択し「＝(H6+I6)/((H3 」となるので「-1」を入力して１つカッコを閉じます。続けて「 +( 」を入力して、女子の標本平均を選択し「＝(H6+I6)/((H3-1)+(I3 」となるので「-1」を入力して１つカッコを閉じます。最後に再度カッコを閉じて完成です。「＝(H6+I6)/((H3-1)+(I3-1))」となって、44.024・・と表示されています。

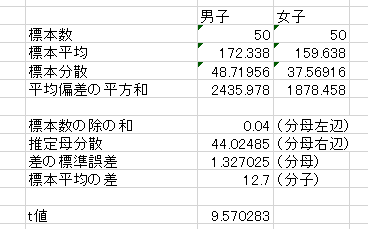
### 分母と分子を作り、t値を求める

分母部分を作ります。分母部分は右辺と左辺をかけて平方根にしたものです。平方根を求めるためには値に0.5を乗じます。「＝(　」と入力したら分母の右辺を選択し、掛け算の記号の「 \* 」を入力したうえで左辺を選択してカッコを閉じます。0.5を乗じるためには「^0.5」もしくは「^(1/2)」と入力します。「「＝(H9\*H8)^0.5」となり、1.327・・・と表示されます。

次に分子部分を作ります。これは男子と女子の平均の差なので「＝」と入力したら男子の標本平均を選択して「＝H4　」として「-」を入力し、女子の平均を選択して「＝H4-I4」とすれば完成です。12.7と表示されます。

最後に分子を分母で割ってt値を求めます。「＝」と入力したら分子部分を選択して「＝H11」としたら、割り算の記号の「/」を入力して分母部分を選択し「＝H11/H10」とします。これでt値が求められます。9.572・・・となっているはずです。





### t分布表と見比べてみる

求められたt値をt分布表で確認します。標本数は50なので、∞と同じと考えて良いです。2.6以上なので、信頼区間99％に入りません。つまり男女の身長は別の分布とみなせます。「男女の身長には有意な差がある」といえます。

## 関数を使って求める

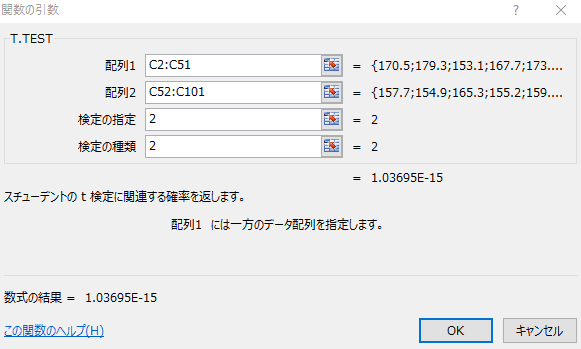
　Excelにはt検定用の関数（t.test)が用意されています。ただしt値を計算してくれる関数ではなく、信頼区間の「α」部分を求める関数です。

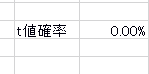
　関数を利用するためには数式バーの左にある「f(x)」を押して「関数の検索」のダイヤログで「t検定」を検索しt.testを選択したうえでダイヤログ下部の「OK」を押して、t.testの関数のダイヤログを開きます。

　値１に男子の身長（C2:51）、値2に女子の身長(C52:C101）を入力します。検定の指定は「両辺」なので「2」、検定の種類は「2」を選びます。でダイヤログ下部の「OK」を押すと関数が完成し、1.03695E-15　と表示されます。

ここで注意しなくてはいけないのが、数字の後ろにE-15とついた数字は小数点以下15桁の数字であるということです。1.03なので、１より大きくなりαとして不適切ではないかと勘違いしてしまいますが、小数点以下の数字です。間違わないようにt.testを使ったときは数値を少数点2桁のパーセント表記にする癖をつけましょう。％にするにはホームタブのリボンの中にある％のマークを押します。％のマークの２つ右隣に←0というマークがあるのでそれを2回押すと小数点2桁が表示されます。

今回は0.00％なので0.01%よりもはるかに小さい確率であることが分かりました。これは男女の身長が同じ分布だと仮定すると、その差が出来る確率は0.01%以下であり、偶然に起こった差としては認められないといえます。つまり男女の身長には有意な差があります。このパーセントが5％以下なら信頼度95％で有意、1％以下なら信頼度99％で有意であるといえます。



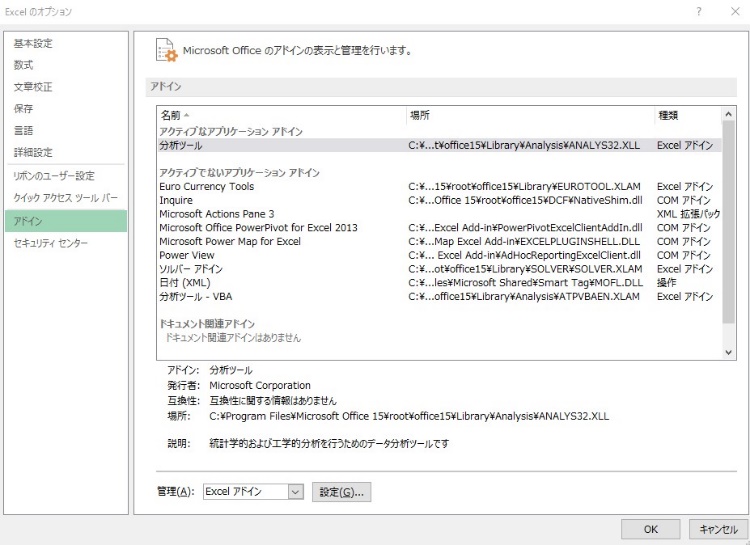
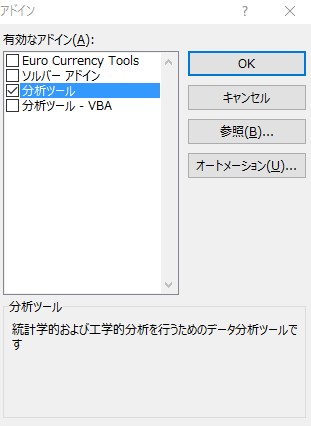


# Excelのデータ分析アドインを利用する

## アドインの準備

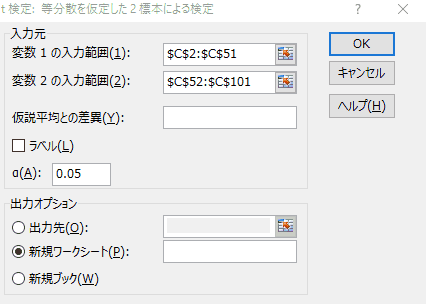
Excelにはデータ分析のアドインがあります。アドインを設定するためには、タブの「ファイル」から情報の画面を表示し、情報の画面の左側にある「オプション」を開きます。

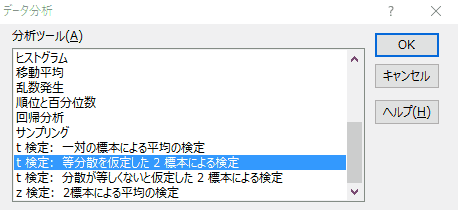
オプションのダイヤログが表示されるので、ダイヤログ左にある「アドイン」を押します。現在設定されているアドイン一覧が出てくるので、その中に「分析ツール」が入っていれば、そのままダイヤログ下部の「OK」を、入っていなければダイヤログ下部中央の「設定」を押します。アドインのダイヤログが開くので「分析ツール」の左側にあるチェックボックスを押してチェックを入れて、ダイヤログ右の「OK」を押します。設定はこれで完了です。



## 分析ツールを起動する

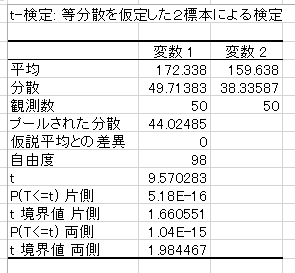
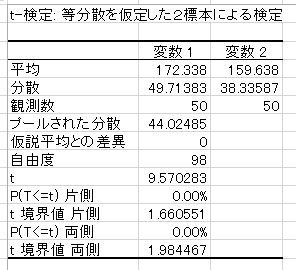
データのタブを押すと、リボンの一番右側に「分析」という項目があり、その中に「データ分析」のアイコンが表示されます。

フレーム内の下の方に「t検定：等分散を仮定した２標本による検定」があるので、選択した状態で「OK」を押します。そうすると回帰分析のダイヤログが開きます。t検定のダイヤログが開いたら、それぞれに項目を入力します。



「変数1の入力範囲（1）」に男子の身長（C2:51）、「変数2の入力範囲（2）」に女子の身長(C52:C101）を入力します。入力がおわったら「OK」を押します。別シートができるので、そこに検定結果が表示されます。

t値が9.57・・・と表示されていて、下の「P（T<=t）片側」と「P（T<=t）両側」の右のセルが確率です。数字の右側にE-15とあるのは小数点15桁の数字という意味なので、これも小数点2桁のパーセント表示に治します。0.00％なので、0.01%以下です。これは男女の身長が同じ分布だと仮定すると、その差が出来る確率は0.01%以下であり、偶然に起こった差としては認められないといえます。つまり男女の身長には有意な差があります。



# 課題

体重やテストの成績に男女の差があるかt検定で確かめてみましょう。