【第10部　2つのデータを比べるその２　相関】

相関とは、一方が増加すると、他方が増加（または減少）する、変量の関係です。例えば身長と体重は相関があります。身長が高くなれば体重も重たくります。そのような関係があるとき「相関がある」といいます。データ分析では頻繁につかうので、基本的な操作として習得しましょう。

# 相関係数

もちろん体重は身長だけではなく体形も影響しますから、同じ身長でも体重にはばらつきがあります。同様に同じ体重でも身長にはばらつきがあります。このばらつき度合いを表すのが「相関係数」です。

## 相関係数の求め方

　身長をx　体重をyとしたとき、「xとyの共分散」を「xとyの分散の幾何平均」で割ったものが相関係数（累積相関係数）となります。

n人の身長と体重を、1人目の身長と体重をx1、y1、2人目の身長と体重をx2,y2、、、n人目の身長と体重をxn,ynとします。身長と体重の平均をX、Yとします。

身長の分散は

体重の分散は

身長と体重の共分散は

になります。相関係数rは

となります。相関係数rを二乗すると決定係数r2になります

## 相関係数の見方

相関係数rは1から-1までの値をとります。xとyが完全に比例しているとき、xが増えるとyが増える場合、相関係数rは１、xが増えるとyが減る場合、相関係数rは-1となります。xとyがまったく比例していないときは0になります。

しかし身長と体重の関係の様に、xとyが完全に比例するということはなかなかありません。誤差や別の要素が混じりますので、相関係数rは0.9であったり、0.2であったり、-0.5であったりします。相関係数には様々な解釈がありますが、おおよそ

|  |  |  |
| --- | --- | --- |
| 相関係数r | (決定係数r2) |  |
| ±0.7～±1 | 0.49～1 | 強い相関がある |
| ±0.4～±0.7 | 0.16～0.49 | 中間程度の相関がある |
| ±0.2～±0.4 | 0.04～0.16 | 弱い相関がある |
| ±0～±0.2 | 0～0.04 | ほぼ相関がない |

といえます。

# 相関係数を計算してみよう。

相関係数は専用の関数をつかって簡単に計算することが出来ます。練習用のシート＊＊＊に練習用の身長・体重・成績のデータがあります。身長と体重、身長と成績、体重と成績のそれぞれの相関係数を出してみましょう。

相関係数の関数は「=CORREL(範囲,範囲)」という関数です。相関の英語である「correlation」から作られています。

### 相関係数のダイヤログを開く

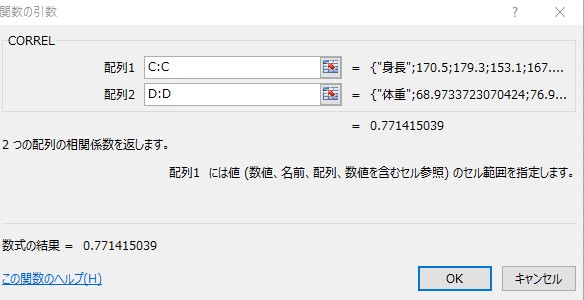
指定したセルに「＝CORREL(」まで入力して数式バーの左にある「f(x)」をおしてダイヤログを開く、もしくは指定したセルに何もない状態でて数式バーの左にある「f(x)」をおして関数のダイヤログを開き、関数の検索のフォームで「相関係数」を検索して関数名にでてきた「CORREL」を指定したうえで「OK」を押してダイヤログを開きます。

## ダイヤログで範囲を指定する

ダイヤログが開いたら配列１と配列2に相関を調べたい２つの数列を指定します。練習用シートにはC列に身長が、D列に体重が入っています。配列１にC列を、配列２にD列を指定することで、身長と体重の相関係数を計算できます。

入力するにはキーボードを使って直接「C:C」「D:D」と入力しても良いですが、慣れるまではフォームの右にあるシートのアイコンを利用する方法が楽です。まずフォームの左にあるシートのアイコンを押します。関数の引数というダイアログが表示されるので、引数に必要なC列、もしくはD列を指定します。指定がおわったらエンターキーをおして入力完了です。これを配列１、配列２で行います。





入力がおわったらダイヤログの「OK」を押して完了です。身長と体重、身長と成績、体重と成績のそれぞれの相関係数をだすと、それぞれ0.77、0.06、0.17という数値になります。身長と体重には強い相関がみてとれます。身長・体重と成績には相関はないといえます。



## グラフにしてわかりやすくしてみよう

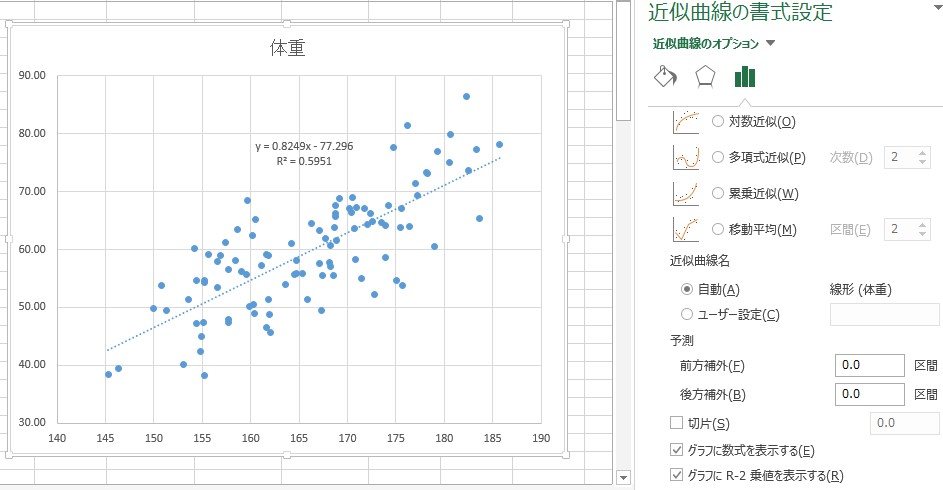
相関は散布図にすると直感的にわかりやすくなります。散布図とは2つの数列の数値を、x軸とy軸に値をとって点描するグラフです。

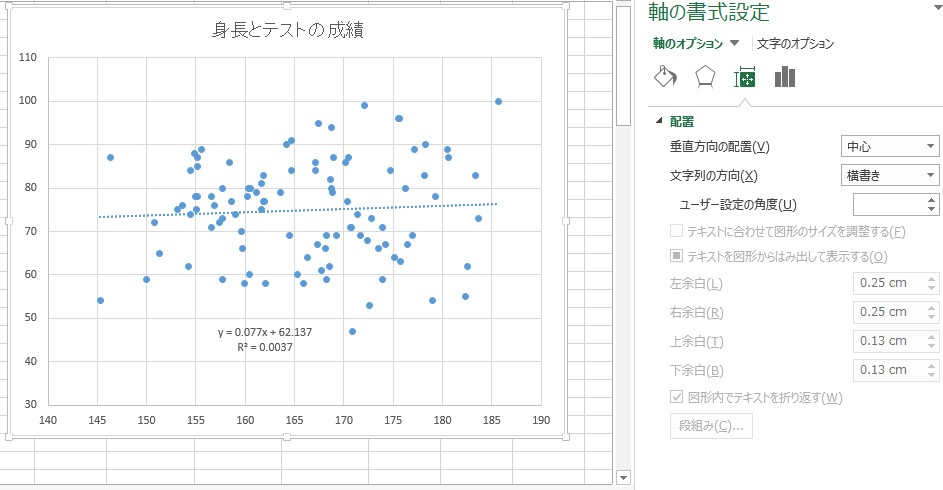
散布図をつくるには、まずx軸とy軸になる列や範囲を指定します。まずx軸なる部分を指定した後、ctrlキーを押しながらy軸になる部分を選択します。選択がおわったら「挿入」タブをクリックし、リボンの中のグラフのエリアから「散布図」を選びます。

　散布図を作成すると、x軸とy軸が値「０」から始まります。グラフが右上にごちゃごちゃっと固まった形になってしまうので最小値を変更します。まずはx軸をクリックします。そうすると右側に修正用の画面が開くので、「軸のオプション」の「境界値」の「最小値」を修正します。身長データであれば140にします。

　次にy軸を修正します。y軸をクリックして右側に開いた画面で「軸のオプション」の「境界値」の「最小値」を修正します。体重であれば30にします。

最後に回帰式と決定係数を表示させます。グラフの「点」の部分を右クリック（macの場合は副ボタンを押す等）して、メニューから「近似曲線の追加（R)」を選択します。右がわに画面が表示されるので「線形近似」を選択したうえで画面下の「グラフに数式を表示する（E）」と「グラフにR-2乗値を表示する（R)」にチェックを入れます。そうすると、回帰直線と回帰式、決定係数（相関係数の二乗値）が表示されます。身長と体重は回帰式がy = 0.8249x - 77.296、R² = 0.5951（Rにすると0.7以上）で、強い相関があるといえます。同様に身長とテストの成績の散布図を作ってみます。y = 0.077x + 62.137　R² = 0.0037なので、身長とテストの成績の間には相関がないといえます。





# 相関と因果の違い

相関を考えるときにもっとも重要のものの１つが「相関と因果は違う」ということです。例えば2000年代前半、県別の一人あたりのジャムの消費量と固定回線のインターネットの普及率の相関を見てみましょう。

これは相関がある、といいうだけで、因果ではありません。パンを食べるとインターネットを使いたくなるわけでも、インターネットを使っているとパンを食べたくなるわけでもありません。

パン食とインターネット利用率に相関がでるのは、都市部のサラリーマン世帯はパン食が多く、且つ仕事でインターネットを使う機会が多いため利用率が高く、地方の農村部では米食のためパンの消費量が少なく、またインターネットを使うような若年層も少ないため利用率が少ないというものです。また都市部ではホテルなどの外食でのパンの消費量が多いということもあります。

同様に朝ごはんを食べる割合が高いと成績が良い、虫歯の多い子は成績が悪いなどの相関もあります。こちらも家庭環境が生活スタイルや成績に影響しているのであって、朝ごはんを食べないと馬鹿になるわけでも、虫歯を直すと頭が良くなるわけでもありません。

【課題】体重と成績の間に相関があるか、散布図を作って確認してみよう。