2022年　ＫＣＳ基礎

「情報科学」⓪“イントロ“プリント　１年　１組　　番　氏名

この授業の「到達目標」→

「データの違い」についての理解を深める

・数学Ⅰの「データの分析」の内容を踏まえたうえで、「標準偏差」の意味合いを正しく理解する。
・「データ」があるときに、それを正しく読み取るにはどのような点に注意する必要があるか習得する。
・「データ」間の「相関」についてきちんと理解する。
・データの「有意差」についてきちんと理解し、有意差を見極めるための処理の方法を習得する。

ちょっと考えてみましょう

体育大会で、バスケットボールでほかのクラスと対戦することとします。選手候補は７名いて、実力はさほど変わらないので、調子のいい５名をスタメンにして、２名を補欠にしようと思います。

（１）「調子が良い」というのは何を基準にしますか？

（２）「調子が良い」を数値化するには？

（３）数値化したものを比べるにはどうします？

このようなことを考えるためにデータの分析で「平均」や「標準偏差」を習ったのだと思うのですが、どうやって使っていいかわかりますか？
分からなかったらこれからみんなで習得していきましょう。

復習：答え方は自由（文章でも式でも）

（１）平均とは？

（２）分散とは？

（３）標準偏差とは？

2022年　ＫＣＳ基礎

「情報科学」①“偏差値“プリント　１年　１組　　番　氏名＿＿＿＿＿＿＿＿＿＿＿＿

偏差値、って結局何？

３０名が、１００点満点のテストを受けた、とします。(Excel ファイルで配ります)

１．まずは「合計」「最高点」「最低点」を一緒に出してみましょう！

２．この得点一覧を見て、No.１６の生徒は「成績がいい」でしょうか？どのくらいですか？

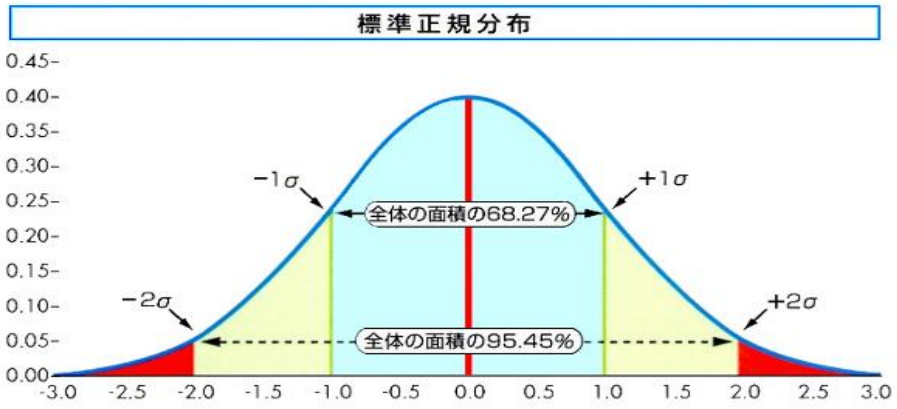
３．「平均」「分散」「標準偏差」を「関数で」求めてみましょう

Excel 関数	平均		分散		標準偏差	

４．平均と標準偏差が分かれば、どのくらいの割合の人がそこに入っているのかが分かります。

ただし、テストを受けた人の点数分布が「正規分布」に従ってるとします

←正規分布とは、平均値に近い分布が多めで、平均から遠いほど分布が少なくなるもので、例えばたくさんの人の身長を調べたりすると、この分布に近くなることが知られています。



下の図では、「σ」が標準偏差の数値です。

※平均＋２σより上には何%いるかな？

◎偏差値とは、テストのたびに違う平均点、標準偏差を比べるのが大変なので、点数を置き換えたもの。

具体的には　平均が＿＿＿＿＿＿、標準偏差が＿＿＿＿＿＿だとすると今回は何点？というものです。

５．偏差値を求めるにはどうしたらいいでしょう？

※たとえば偏差値が「１００」とかってあり得ると思ういますか？

2022年　ＫＣＳ基礎

「情報科学」②“データの相関“プリント　　年　　組　　番　氏名

２つのデータの間にある関係について考えます。

１．都道府県別のデータを用意しました。下の質問について考えてみましょう。

（１）一方が増加すると、もう一方も増加するという傾向がありそうなのはどれとどれですか？

（２）一方が増加すると、もう一方が減少するという傾向がありそうなのはどれとどれですか？

２．「相関」「散布図」について復習しよう。

まずは「県別データ」から、指示された２つの列を「実習」の列１、列２に移す

「散布図」・・・２つの変量からなるデータを平面上に図示したもの

※「グラフ」に散布図があるので、実際に描いてみましょう！

「正の相関」「負の相関」

「相関係数」・・・関数「correl」を使って求めてみる

相関係数	相関

2022年　ＫＣＳ基礎

「情報科学」　都道府県別の相関係数を考える

年　　組　　番　氏名

１．相関係数を調べるデータは？

２．予想（予想した理由も書くこと）

３．実際の相関係数と、考察

まとめ

１．「相関係数」≒「相関関係」≠「因果関係」

２．「疑似相関」に注意！

３．「散布図」を書かずに相関係数を使うなかれ

2022年　ＫＣＳ基礎

基礎的な統計学③“推定と検定”プリント

年　　組　　番　氏名

区間推定で使う用語について

「母集団」：結果のすべて（今後ずっとやっていく結果全部）
「標本」：今回の（１０回の）実験の結果　※「サンプル」とも言います
「サンプルサイズ」：標本の数　　「自由度」：サンプルサイズー１

方法：**「標準誤差」というものを考える**（標準偏差と言葉は似ているけど、意味は全然ちがう）
考え方：まずは今回の１０個分の平均を基準とする（仮にどんなにやってもこの平均値だと思うことにする）といっても、平均は毎回おそらく違っていて、その数字は「正規分布」に従うとして
（だいたいは平均値と似た数字になり、平均から大きく外れることはめったにないはず）
「平均はこの間に入っているはず！」という数値を求める。

ア）母集団の平均は、標本の平均と同じだと考える（予想する）ことにする
イ）母集団の分散は、標本の「分散」を参考にする**（同じじゃないです！）**
※１０回の実験のうち、最初の１個を「基準」とし、「それとのばらつき」と考える（「自由度」）
普通の分散を手計算で求めるとすると「{（偏差の２乗）の合計}÷１０」となるが、最初の１個は「ばらつき」に入れないので、**{（偏差の２乗）の合計}÷９を計算する。これを「不偏分散」という。**
※関数で求める：不偏分散：VAR.S

◎これをもとに「区間推定」を行う
１．「標準誤差」を計算する（標本を取り出したときの、平均の振れ幅の“指標”）

「標準誤差」＝$\sqrt{\text{「不偏分散」}\div\text{サンプルサイズ}}$　（エクセルでルートは関数「SQRT」）

※サンプルサイズが４倍になると、この指標は半分の数値になる**（標本数は多いほうがいい）**
※サンプルサイズが大きい場合には、標本の分散≒不偏分散となるので、この式の替わりに

「標準偏差」÷$\sqrt{\text{サンプルサイズ}}$　という式を使うこともある。

２．実験を繰り返せば、毎回の平均値の分布は「正規分布」に近くなると考えられる。「自由度」と、「例外」として何％を許容するかによって、「ｔ検定値」というもう一つの指標が決まる。
「自由度」９、例外は一般的には５％（厳しくいく場合には１％）
３．「標準誤差」と「ｔ検定値」を掛けると、“振れ幅”を求めることができる。

４．平均±“振れ幅“を「信頼区間」と言う。
実験ごとの「信頼区間」は重なっているか？
（重なってなければ２つの実験に「有意差」：意味のある差があると考えることができる）

2022年　ＫＣＳ基礎

◎改めて「水摂取」と「運動後」をきちんと比べてみよう

「水と運動」シートに、それぞれの「平均」「不偏分散」と「区間（下)、区間（上)」を「値貼付」する
（今回は、同じ人の記録が対応できるため、比べやすい）
※できない例：何かの勉強方法の工夫で、二つの方法を比べる場合

それでは、「誰かひとり」について、２つの実験の「平均の差」を取り、それを区間推定してみる
※「差の推定」：今後この実験を繰り返したとして、数%（今回は５%）の例外を除いて、「差」の範囲はこのくらいに収まるのではないか？ということを推定する。
この区間が「マイナス」になる→「逆転することがありうる」→「有意差があるとは言えない」
という考え方になる。
方法は前回とだいたい一緒　１～１６、どれか好きなところでやってみよう
①「平均差」を出す
②「差の標準誤差」＝$\sqrt{（不偏分散A÷サンプルサイズA）＋（不偏分散B÷サンプルサイズB）}$
Excelの計算式としては =SQRT（不偏分散A／サンプルサイズA＋不偏分散B／サンプルサイズB）
③「ｔ検定値」を求める（今日は右下に「例外%」と「自由度」を入力する形にした）
　※この場合の自由度：それぞれの実験回数－１　なので「１０－１」＋「１０－１」で「１８」
④区間推定を行う

さらに「全員の差」について、推定を取ってみます（シートの下側）

※「水摂取」と「運動後」の実験結果は「有意差」があるか？

◎有意差の確認のためのお手軽な方法：「検定」　今回はｔ検定という方法を使う
区間推定は作業が多いので、もっと簡単に「有意差があるかなしか」だけを調べる方法を「検定」という
（使う数値は推定の時と同じ）
　前提として、「実験方法での差は**ない**（偶然じゃないの？）」と仮定する。**（「帰無仮定」）**
で、計算してそれがひっくり返されたら「差がある」と考える。（有意差）
（統計の基本的な考え方。この考え方を理解すること！）
やり方：標本のｔ検定値を求める。
　標本のｔ検定値＝平均差÷標準誤差
　標本のｔ検定値が関数でのｔ検定値を「上回っていれば」有意差がある、ということになる。
　（上回っている→極端な状態である→有意差がある）

また、「T.TEST」という、ダイレクトにｔ検定結果を出す関数もある
出てきた数字の意味：そこに出た確率で、２つの点数の「母集団」が等しい。