**相澤裕介『統計処理に使うExcel活用法』カットシステム**

**理論的な要点　(補足と第四章)**

2010/11/30　柳瀬陽介

**相関関係について**

この教科書には（なぜか）ないのだけれど、相関関係（散布図と相関係数）については知っておく必要がある。

■**相関関係**：「英語ができる生徒は、数学もできる」という関係はどういう関係だろう。

「『英語ができる生徒は、数学もできる』と聞いたので、一生懸命英語を勉強したら、英語の成績は上がったが、数学の成績は一向に上がらなかった。先生は嘘つきだ！」と息巻く生徒がいたら、あなたはどう説得しますか？

参考サイト

<http://www.ec.kansai-u.ac.jp/user/amatsuo/probstat/CResultCorr.pdf>

■**散布図**：直感的な把握として散布図を考えてみよう。(下図はウィキペディアより)



■**相関係数**：散布図の関係を一つの数値で表現したもの。

-1 から 1 の間の実数値をとり、1 に近いときは2 つの確率変数には正の相関があるといい、-1 に近ければ負の相関があるという。0 に近いときはもとの確率変数の相関は弱い。因みに 1 もしくは -1 となる場合は 2 つの確率変数は線形従属の関係にある。

参考サイト

<http://statistics.web.fc2.com/correlation_coefficient.html>

※やってみよう。

散布図はエクセルのグラフ機能で簡単に出せる。また、相関係数(correlation)はエクセル関数のCORRELを使うか、ツール→分析ツール→相関を使うことで出せる。ダウンロードデータで散布図を描き、相関係数を求めてみよう。

分散分析について（教科書第四章）

■**分散分析**（1要因）

二つの平均値を比べるのはt検定だが、三つ以上だと分散分析を使う。

<http://kogolab.jp/elearn/hamburger/chap6/sec2.html>

の説明をそのまま使う（上記リンクをクリックしてください）。

モグモグハンバーガー、ワクワクハンバーガー、パクパクハンバーガーの三種類のハンバーガーのおいしさの得点（100点満点の主観的評価）を比較したところ、個別データの集計は表のようになり、分布は下図のようになった。

|  |  |  |
| --- | --- | --- |
| ワクワク | モグモグ | パクパク |
| 80 | 75 | 80 |
| 75 | 70 | 80 |
| 80 | 80 | 80 |
| 90 | 85 | 90 |
| 95 | 90 | 95 |
| 80 | 75 | 85 |
| 80 | 85 | 95 |
| 85 | 80 | 90 |
| 85 | 80 | 85 |
| 80 | 75 | 90 |
| 90 | 80 | 95 |
| 80 | 75 | 85 |
| 75 | 70 | 98 |
| 90 | 85 | 95 |
| 85 | 80 | 85 |
| 85 | 75 | 85 |
| 90 | 80 | 90 |
| 90 | 80 | 90 |
| 85 | 90 | 85 |
| 80 | 80 | 85 |



・**群間のズレ**

「全体の平均とモグモグの平均のズレ」は、これは全体の平均から各群（各標本集団。この例ならモグモグ、ワクワク、パクパク）がどれほどズレているかである。これを「群間のズレ」呼ぶ。「群間のズレ」は、各群の違いを示す。

・**群内のズレ**

一方、「モグモグの平均からのズレ」は、群（標本集団）の中で、個々のデータがどれほどズレているかを示す。これを「群内のズレ」と呼ぶ。群内のズレは、「個体差」「個人差」「誤差」と考えられる。

・**全体の平均からのズレ　＝　群間のズレ　＋　群内のズレ**

すべてのデータについて、

全体の平均からのズレ　＝　群間のズレ　＋　群内のズレ

ということが成り立つ。

・**分散分析の考え方**

群間のズレ（群ごとの差）の方が、群内のズレ（誤差）より大きいならば、標本集団の間には差があると考えるべきだろう。

逆に群間のズレ（群ごとの差）の方が、群内のズレ（誤差）より小さいならば、標本集団の間には差がないと考えるべきだろう。

・ズレは**「データ　–　平均」の平方和**で出す。

ここでも

全体の平方和　＝　群内の平方和　＋　群間の平方和

という関係がなりたつ。

・**平方和をそれぞれの自由度で割って「平均平方」を出す（=「分散」）**

自由度は次のように決められる。

群間の自由度　=　群の数　－　1

群内の自由度　=　（群1のデータ数－1）＋（群2のデータ数－1）+（群3のデータ数－1）

全体の自由度　=　データの総個数　－　1

・分散の比を**F検定**する。

F=　　$\frac{群間の分散（=群間の平均平方）}{群内の分散（=群内の平均平方）}$

このFの値が大きければ（=偶然でこのF値が出る確率が5%以下だったら）、群間の分散は､群内の分散よりも､非常に大きいのだから､群間の分散(つまりばらつき)には統計的な意味がある､つまり､標本集団の間には差がある（=標本集団の間に差がないとは言い難い）ということになる。

■**多重比較**

しかし、分散分析でわかるのは（例えば）三つの平均値が全て等しい(A=B=C)とは言えない、というだけで、A>B>C, A>B=C, A=B>Cなのかはわからない。

この細かな差を検定するには「多重比較」という分析が必要だが、エクセルではそれは不可能。SPSSなどのソフトを使って行う（<http://www.spss.co.jp/>）。

注：二要因の分散分析（主効果、交互作用など）についてはこの授業では省略します。