［利用の手引き］

# 分散分析を扱う：統計解析ソフトウエア SPSS の利用

犬塚裕樹†
Hiroki Inutsuka†

†久留米大学　医学部看護学科
† Kurume University, School of Nursing

## 1. はじめに

試験の得点や気温などの連続尺度の値が、あるグループ間や処理方法によって違いがあるかどうかを調べるために、統計処理方法として分散分析が利用されます。この分散分析は有名で、ANOVA (ANalysis Of VAriance)とよばれています。大学学部での学習でもよく利用されている分析方法の 1 つです。

本稿では、統計解析ソフトウエア SPSS を利用して、分散分析により手元にもっている実際のデータを分析することを主目的として、基本的な利用法について説明しました。

## 2. 分析の目的

成績の得点などの連続尺度のデータが複数のグループで保管されているとします。各グループの連続尺度のデータの母平均値がグループ間で差があるだろうか。この問いに答えるために、分散分析の統計手法が役にたちます。

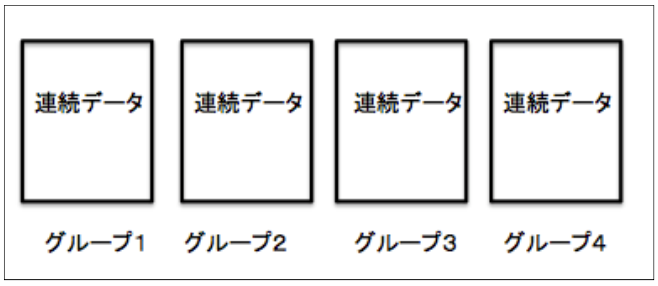


**図 1** 連続データのグループ

ここでは、4 つのグループにおいて、連続数値データがあるとしましょう。各グループでの連続数値データの母平均値が、4 つのグループ間で差があるかどうかを調べます。

これは、分散分析のなかでは、「1 つの要因について複数のグループ間の平均値の差を検定する、一元配置・多水準の場合の分析」とよばれるものに分類されます。一般的には、分散分析と、ひとことでいっても、二元配置分散分析、三元配置分散分析などがあります。

グループ間での平均値の差を調べようとするのに、分散ということばがつかわれた手法をつかうというのは、一見、不思議な感じがするかもしれません。それは、グループの平均値はグループの数値の分散に影響し、また、全体の数値データのばらつきは、グ

ループ間の平均値のばらつきと、グループ内の平均値からの偶然によるばらつきの 2 つに分離できるという便利な性質を利用するためです。

## 3. データの入力

1 次データはエクセルに入力するとして、その後、SPSS にデータを移しデータ分析をするという手順で説明をします。

まず、エクセルのシートにデータを入力します。入力形式は図 2 のように、「グループ」の項目と「数値」の項目に入力します。各グループのデータ数は、グループ 1 が 7 個、グループ 2 が 6 個、グループ 3 が 7 個、グループ 4 が 5 個となっています。グループで、データ数が同じでなくてもかまいません。

| | A | B |
|---|---|---|
| 1 | グループ | 数値 |
| 2 | 1 | 33 |
| 3 | 1 | 24 |
| 4 | 1 | 26 |
| 5 | 1 | 22 |
| 6 | 1 | 30 |
| 7 | 1 | 37 |
| 8 | 1 | 29 |
| 9 | 2 | 39 |
| 10 | 2 | 27 |
| 11 | 2 | 30 |
| 12 | 2 | 36 |
| 13 | 2 | 41 |
| 14 | 2 | 45 |
| 15 | 3 | 58 |
| 16 | 3 | 61 |
| 17 | 3 | 63 |
| 18 | 3 | 68 |
| 19 | 3 | 58 |
| 20 | 3 | 73 |
| 21 | 3 | 74 |
| 22 | 4 | 40 |
| 23 | 4 | 28 |
| 24 | 4 | 33 |
| 25 | 4 | 45 |
| 26 | 4 | 43 |

**図 2** グループデータ

このデータを SPSS に読み込みます。

SPSS を起動し「メニュー／開く」をクリックします。

このエクセルファイルを指定して開くと、図 3 のような画面が開きます。

入力値が小数点になっていますので、整数値表示にします。

画面左下の「変数　ビュー」ボタンをクリックします。すると、図 4 のような「変数」画面にかわります。

この画面で図 4 のように、小数桁数を 0 に変更します。「データ　ビュー」のボタンをクリックして、データ画面を表示します。図 5 の画面になり、これで、「グループ」も「数値」のデータは整数値表示となりました。

| | グループ | 数値 |
|---|---|---|
| 1 | 1.0 | 33.0 |
| 2 | 1.0 | 24.0 |
| 3 | 1.0 | 26.0 |
| 4 | 1.0 | 22.0 |
| 5 | 1.0 | 30.0 |

**図 3** SPSS のデータビュー画面

| | 名前 | 型 | 幅 | 小数桁数 | ラベル | |
|---|---|---|---|---|---|---|
| 1 | グループ | 数値 | 12 | 0 | | な |
| 2 | 数値 | 数値 | 12 | 0 | | な |
| 3 | | | | | | |

**図 4** SPSS の変数ビュー画面

| | グループ | 数値 |
|---|---|---|
| 1 | 1 | 33 |
| 2 | 1 | 24 |
| 3 | 1 | 26 |
| 4 | 1 | 22 |
| 5 | 1 | 30 |
| 6 | 1 | 37 |
| 7 | 1 | 29 |

**図 5** SPSS のデータビュー画面

(整数への返還後)

## 4. 分析の進め方

解析を始める前に、まずは、入力データがどのような分布をしているかを、グラフを描いて様子をみることがだいじです。

「メニュー／グラフ／レガシーダイアログ」にマウス・ポインターをのせると図 6 が表示されます。

そこで、「散布図／ドット」の項目でクリックします。

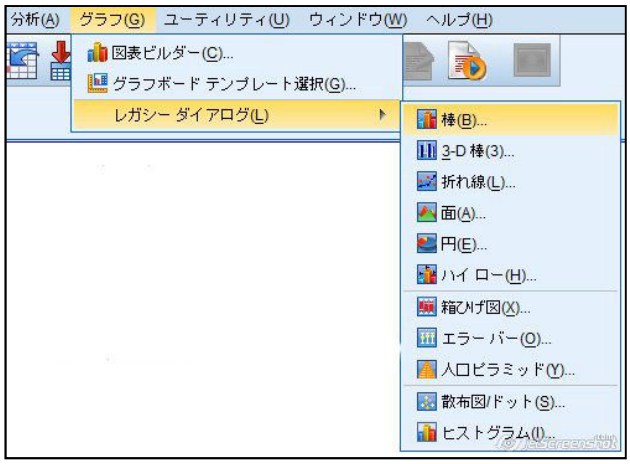


**図 6** レガシーダイアログメニュー

すると、図 7 が表示されます。

このウィンドウで、「シンプルドット」ボタンをクリックし、「定義」ボタンをクリックします。


**図 7** 散布図/ドット画面

すると、図 8 のウィンドウが表示されます。「X 軸変数」のウィンドウに「数値」をいれ、「行」のウィンドウに「グループ」を入力します。


**図 8** シンプルドットプロットの定義画面

すると、分布を示す図 9 が表示されます。この図から次のことがわかります。どのグループのデータ値のばらつきも同じぐらいで、値はグループ 3 以外、同じような値になっていることがわかります。グループ 3 では、値が大きいようです。

この分析の目的は、グループ内の平均値を、グループ間で比較することです。分析の進め方は、グループのペアごとに平均値が異なるかどうかを検定していくことも考えられます。しかし、まずすべきことは、4 つのグループ全体ですべてのグループの平均値が等しいかどうかの検定です。

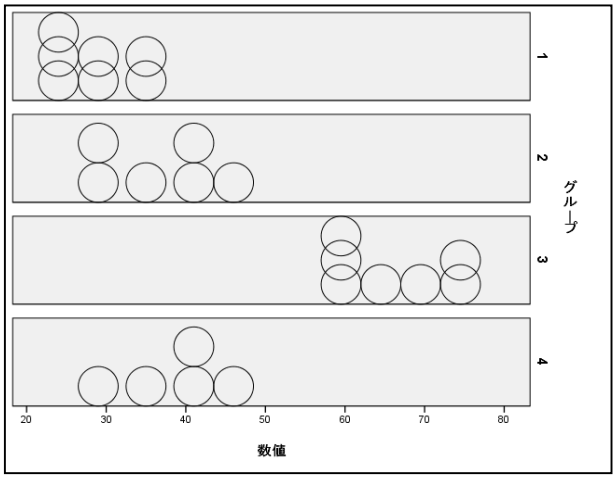


**図 9** 分布図

基本的な考え方として、統計検定はなんどもおこなってはいけない、ということがあります。グループのペアごとに平均値の検定をしていく場合、ここでは、グループが 4 個あります。そのため、グループ間ですべてのペアをつくり検定をしていくと、4 つの中で 2 つの組み合わせをつくる組み合わせの数は 6 個になります。

それぞれの検定で有意水準を 5%に設定した場合、この数値は平均値が等しいとする帰無仮説が棄却されたとしたとき、その結果が誤る確率をあらわします。

6 回の検定をしたとき、その中で少なくとも 1 回の検定で誤るという確率を求めてみましょう。

6 回とも全く誤らない確率は $(1 - 0.05)6 = 0.74$ ですから、少なくとも 1 回誤る確率は $1 – 0.74 = 0.26$ という大きな確率となります。これほど大きい確率で誤ってしまうことになります。

分散分析に進んでみましょう。「ファイル／分析／一元配置分散分析」でクリックします。すると、図 10 のウィンドウが表示されます。

「従属変数リスト」のウィンドウには、「数値」を入力します。「因子」ウィンドウには、「グループ」を入力します。


**図 10** 一元配置分散分析画面

つぎに、「オプション」をクリックすると図 11 のウィンドウが表示されます。

「記述統計量」をクリックしてチェックを入れます。

さらに、グループ内の分散がグループ間で等しいことが重要です。そのために、「等分散性の検定」にチェックをいれます。

また、グループの平均値のグラフも見たいので、「平均値のプロット」にもクリックしてチェックを入れましょう。

上記の 3 カ所にチェックをいれて、最後に「続行」のボタンをクリックします。

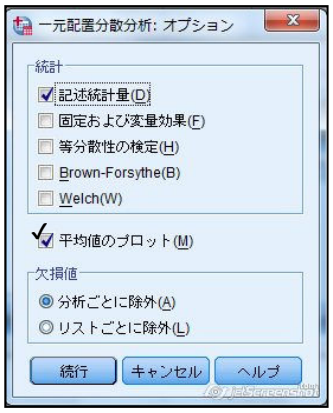


**図 11** 一元配置分散分析オプション画面

図 10 のウィンドウにもどります。このウィンドウで、「OK」をクリックします。

すると、一元配置分散分析の結果が表示されます。これとどうじに、記述統計の結果、グループの平均値のグ

**表 1** 記述統計

記述統計

数値

| | 度数 | 平均値 | 標準偏差 | 標準誤差 | 平均値の 95% 信頼区間 | | 最小値 | 最大値 |
|---|---|---|---|---|---|---|---|---|
| | | | | | 下限 | 上限 | | |
| 1 | 7 | 28.71 | 5.219 | 1.973 | 23.89 | 33.54 | 22 | 37 |
| 2 | 6 | 36.33 | 6.802 | 2.777 | 29.20 | 43.47 | 27 | 45 |
| 3 | 7 | 65.00 | 6.733 | 2.545 | 58.77 | 71.23 | 58 | 74 |
| 4 | 5 | 37.80 | 7.120 | 3.184 | 28.96 | 46.64 | 28 | 45 |
| 合計 | 25 | 42.52 | 15.919 | 3.184 | 35.95 | 49.09 | 22 | 74 |

ラフと等分散性の検定のウィンドウも表示されます。

記述統計の表 1 より、各グループの平均値や標準偏差などがわかります。図 12 の平均値のグラフからは、平均値の違いの程度が直感的にわかります。

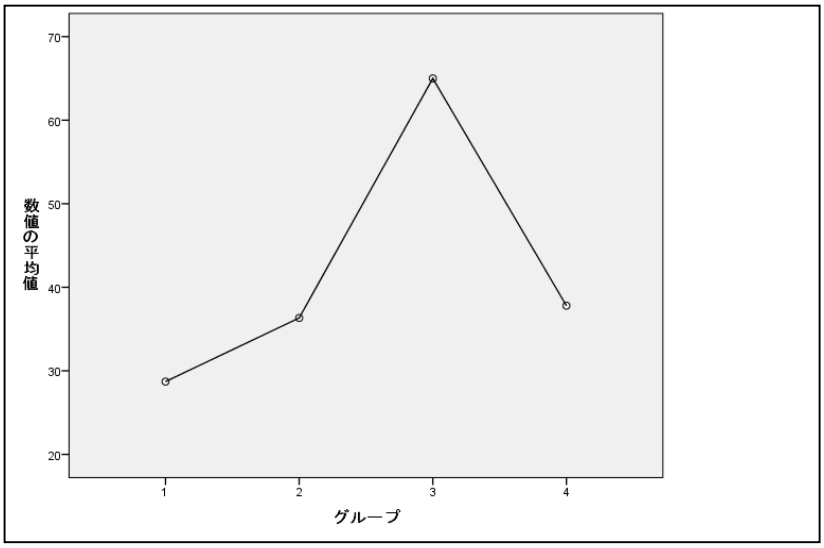


**図 12** 平均値のグラフ

表 2 の等分散性の検定の表からは、有意確率が 0.638 であることがわかります。すなわち、帰無仮説の「グループ間の母分散は等しい」が 5%の有意水準でも棄却されなかいことがわかります。

**表 2** 等分散性の検定

等分散性の検定

数値

| Levene 統計量 | 自由度1 | 自由度2 | 有意確率 |
|---|---|---|---|
| .506 | 3 | 21 | .683 |

とはいうものの、このことは、等分散性の証明になってはいません。しかし、明確に異なるということが示されない限り、等分散ということがなりたっているとみなして、分散分析にすすむことが一般になされています。

そこで、表 3 の一元配置分散分析の結果をみることにします。一元配置分散分析の表には、F 値や有意確率が示されています。したがって、この表から、検定結果を知ることができます。

**表 3** 一元配置分析

一元配置分析

[データセット1]

分散分析

数値

| | 平方和 | 自由度 | 平均平方 | F 値 | 有意確率 |
|---|---|---|---|---|---|
| グループ間 | 5212.678 | 3 | 1737.559 | 41.962 | .000 |
| グループ内 | 869.562 | 21 | 41.408 | | |
| 合計 | 6082.240 | 24 | | | |

有意水準1%で、あるいは有意水準5%でも、帰無仮説である「すべての母平均値が同じである」が棄却されていることがわかりました。この結果から、グループ 3 の平均値が他のグループからは突出したも

のであることがわかりました。

そこで、こんどは、グループ間での平均値の違いのようすを調べていくことにします。

そのためには、図 10 において、「その後の検定」ボタンをクリックします。すると、図 13 のようなウィンドウが表示されます。このウィンドウにはさまざまな解析法がリストされています。「Dunnet」では、1 つのグループを対照として、他のグループとペアで検定がなされます。「最小有意差」では、全ペアに対して t 検定がなされます。


**図 13** 一元配置分散分析：その後の多重比較画面

ここでは、「最小有意差」をクリックして、チェックしてみましょう。そして「続行」のボタンをクリックします。

すると、多重比較の結果の表 4 があらわれます。

**表 4** その後の検定の多重比較

その後の検定

多重比較

従属変数: 数値

LSD

| (I) グループ | (J) グループ | 平均値の差 (I-J) | 標準誤差 | 有意確率 | 95% 信頼区間 | |
|---|---|---|---|---|---|---|
| | | | | | 下限 | 上限 |
| 1 | 2 | -7.619* | 3.580 | .045 | -15.06 | -.17 |
| | 3 | -36.286* | 3.440 | .000 | -43.44 | -29.13 |
| | 4 | -9.086* | 3.768 | .025 | -16.92 | -1.25 |
| 2 | 1 | 7.619* | 3.580 | .045 | .17 | 15.06 |
| | 3 | -28.667* | 3.580 | .000 | -36.11 | -21.22 |
| | 4 | -1.467 | 3.897 | .710 | -9.57 | 6.64 |
| 3 | 1 | 36.286* | 3.440 | .000 | 29.13 | 43.44 |
| | 2 | 28.667* | 3.580 | .000 | 21.22 | 36.11 |
| | 4 | 27.200* | 3.768 | .000 | 19.36 | 35.04 |
| 4 | 1 | 9.086* | 3.768 | .025 | 1.25 | 16.92 |
| | 2 | 1.467 | 3.897 | .710 | -6.64 | 9.57 |
| | 3 | -27.200* | 3.768 | .000 | -35.04 | -19.36 |

*. 平均値の差は 0.05 水準で有意です。

グループ 2 とグループ 4 間は、「平均値が等しい」という帰無仮説が有意水準 5%では有意に棄却されていませんが、それ以外のグループ対は有意に棄却され、平均値に差があるといえる結果となっています。この結果は、図 12 の平均値のグラフで状況を知ることができます。

## 5. 分散分析の原理

一般に A という因子を考え、水準が r 個あるとし、それぞれの観測値が表 5 のように示します。

**表 5** それぞれの観測値

| A の水準 | 観 測 値 | | | | 計 | 平均 |
|---|---|---|---|---|---|---|
| $A_1$ | $x_{1l}$ | $x_{12}$ | … | $x_{1n1}$ | $T_1$ | $\bar{x}_1$ |
| $A_2$ | $x_{21}$ | $x_{22}$ | … | $x_{2n2}$ | $T_2$ | $\bar{x}_2$ |
| … | … | … | … | … | | |
| $A_r$ | $x_{r1}$ | $x_{r2}$ | | $x_{rnr}$ | $T_r$ | $\bar{x}_r$ |

すなわち、i 番目の水準 $A_i$ で、j 番目の観測値を $x_{ij}$ とします。Ai の観測値の数は $n_i$ 個とします。

そして、4 つの量について次のように定義します。

$$T_i = \sum_j x_{ij}$$

$$\bar{x}_i = \frac{1}{n_i} T_i$$

$$n = \sum_i n_i$$

$$\bar{x} = \frac{1}{n} \sum_{i,j} x_{ij}$$

さて、全観測値の全平方和 $S_T$ はつぎのように書き換えることができます。

$$S_T = \sum_{i,j} (x_{ij} - \bar{x})^2$$
$$= \sum_{i,j} (x_{ij} - \bar{x}_i)^2 + \sum_i n_i (\bar{x}_i - \bar{x})^2$$

ここで、

$$S_A = \sum_i n_i (\bar{x}_i - \bar{x})^2$$

$$S_E = \sum_{i,j} (x_{ij} - \bar{x}_i)^2$$

とおくと、全平方和 $S_T$ は

$$S_T = S_A + S_E$$

のように 2 つの平方和に分解できることがわかります。$S_A$ は級間平方和、$S_E$ は級内平方和とよばれます。

つぎに、モデルを考えることにします。$x_{ij}$ は下記のような確率変数 $X_{ij}$ の実現値であるとします。

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

ここで、$\varepsilon_{ij}$ は n 個の互いに独立な $N(0, \sigma^2)$ にしたがう確率変数です。さらに、$\alpha_i$ は下の等式をみたすものとします。

$$n_1\alpha_1 + n_2\alpha_2 + \ldots + n_r\alpha_r = 0$$

すると、

$$\overline{X}_i = \mu + \alpha_i + \overline{\varepsilon}_i$$

が得られます。ここで、

$$\overline{\varepsilon}_i = \frac{\sum_j \varepsilon_{ij}}{n_i}$$

とおいています。

また、

$$\overline{X} = \mu + \overline{\varepsilon}$$

が得られます。ここで、

$$\overline{\varepsilon} = \frac{\sum_{i,j} \varepsilon_{ij}}{n}$$

とおいています。

そこで、$S_A$ と $S_E$ を統計量として書きかえると、つぎの 2 式が得られます。

$$S_A = \sum_i^s n_i(\overline{X}_i - \overline{X})^2 = \sum_i^s n_i(\alpha_i + \overline{\varepsilon}_i - \overline{\varepsilon})^2$$

$$S_E = \sum_i^s (\overline{X}_i - \overline{X})^2 = \sum_i^s (\varepsilon_{ij} - \overline{\varepsilon}_i)^2$$

そこで、$S_A$ の期待値 $E[S_A]$ と、$S_E$ の期待値 $E[S_E]$ を求めます。その結果、次の 2 式が得られます。

$$E[S_A] = (r-1)\sigma^2 + \sum_i^s n_i \alpha_i^2$$

$$E[S_E] = (n-1)\sigma^2$$

ここで、各水準 i による効果がない、ということを検定することを考えましょう。すなわち、このことは、帰無仮説 $H_0$:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$$

を、対立仮説：0 でない $\alpha_i$ が存在する、に対して検定することにします。

上の式より

$$\frac{S_A}{r-1} \quad と \quad \frac{S_E}{n-r}$$

は母分散 $\sigma^2$ の不偏推定量であることがわかります。これに基づいて、これらの比を F と定義すると

$$F = \frac{\dfrac{S_A}{r-1}}{\dfrac{S_E}{n-r}}$$

は自由度 (r - 1, n - r) の F 分布にしたがうことが知られています。このことから、帰無仮説 $H_0$ を検定することができます。

## 6. おわりに

本稿では、統計検定の 1 つである分散分析について、ある 1 つの特定の場合の利用法を簡単に説明しました。

統計検定では、帰無仮説に基づき、ある統計量の確率分布が理論的に計算されます。その確率分布に対して、手元の標本データから帰無仮説が棄却されるかどうかを調べます。統計検定をおこなう場合に注意すべきことがあります。確率分布が計算される際に、ふつう、簡単化のためにいくつかの前提となる仮定がなされます。そのために、標本データがこの仮定をみたしておく必要があります。

本稿であつかった分散分析では、

(1) 標本は正規分布する母集団から抽出されたもの、

(2) 標本の母分散は比較するグループ間で等しいこと、

(3) 標本は独立に抽出されたものであることが仮定されています。

(1)については、図 9 の分布図でおおまかなところを確認することができます。極端に、両端にデータが局在し 2 つの峰をもつ分布をしていないかどうかを確認します。

(2)については本文に記載しています。

(3)については、標本を抽出する際に標本間で明らかに相関みられると推測される状況では、この検定をあきらめるか、なにか対処法を工夫することが必要になります。

## 参考文献

[1] 「すぐわかる分散分析」　内田治、牧野泰江、西澤英子　共著、東京図書（2007）

[2] 「統計入門」　和田秀三　著、サイエンス社（1982）