

# R と分散分析

## 1. 分散分析

実験、観測、調査などでは同じの条件であっても、計測の誤差やノイズなどが混入され、得られているデータにはずれが多かれ少なかれ生じる。また同様な実験、観測、調査を、条件を変えて行ったとき、計測の誤差やノイズなど以外に条件の影響で違いが生み出される可能性がある。実験、観測、調査の結果に影響をおよぼすと考えられる要因はいろいろあるが、その実験、観測、調査で取り上げている要因を因子(factor)と呼び、因子を細分類したグループ(群)を水準(levels)と呼ぶ。

分散分析(analysis of variance; ANOVA)は、得られた各水準の平均値が因子の影響により変化されていると言えるかどうかに関するデータ分析の方法である。本稿では、一元(one-way)分散分析と二元(two-way)分散分析の簡単な例を用いて R による分散分析について説明する。

## 2. 一元分散分析

実験、観測、調査で取り上げている要因・因子(factor)が1つである分散分析を一元分散分析と呼ぶ。そのデータの形式(一元配置)を表1に示す。表1の中の $a_j$  ( $j=1,2,\dots,k$ )は因子Aの水準である。

表1 一元配置データ表

標本	因子 A					
	$a_1$	$a_2$	...	$a_j$	...	$a_k$
1	$y_{11}$	$y_{12}$	...	$y_{1j}$	...	$y_{1k}$
2	$y_{21}$	$y_{22}$	...	$y_{2j}$	...	$y_{2k}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$i$	$y_{i1}$	$y_{i2}$	...	$y_{ij}$	...	$y_{ik}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$n$	$y_{n1}$	$y_{n2}$	...	$y_{nj}$	...	$y_{nk}$
平均	$\bar{y}_{\cdot 1}$	$\bar{y}_{\cdot 2}$		$\bar{y}_{\cdot j}$		$\bar{y}_{\cdot k}$

一元分散分析は対応なし、対応ありに大きく分けることができる。対応とは、 $y_{i1}, y_{i2}, \dots, y_{ij}, \dots, y_{ik}$  は同一の対象  $i$  ( $i=1,2,\dots,n$ ) についてそれぞれの水準の下で得られたデータであるか否かである。具体的な例を用いて説明する。

## 2.1 対応なしの場合

例えば、ある機器の性能について、異なる温度の下でそれぞれ 10 回のテストをランダムに行い、表 2 のようなデータが得られたとする。この場合、各水準(列)内のデータの行の順序には意味がない。例えば、表 2 のセル(1, 1)のデータ 63 を第 1 列内であればどの行と入れ換えてもよい。かつそのとき他の列の第 1 行のデータをセル(1, 1)に連動させなくてもデータが持っている情報には変わりがない。問題を簡単にするためここでは機器の疲労の影響などについては無視する。

表 1 繰り返しの機器のテスト結果(架空)  
(対応なし)

回数	要因・因子 A			
	$a_1$ - 20°C	$a_2$ 0°C	$a_3$ 20°C	$a_4$ 40°C
1	63	64	59	83
2	58	64	87	79
3	64	68	79	65
4	58	61	71	67
5	77	56	65	80
6	66	71	65	72
7	52	64	65	80
8	64	65	71	75
9	49	85	74	72
10	66	75	58	84
平均	61.7	67.3	69.4	75.7

ここでは、機器の平均性能が条件の影響を受けて変化していると言えるかどうかについて興味を持っている。要因・因子の影響による観測値の変化を因子効果と呼ぶ。一元分散分析はこのような多群の平均値の検定に関するデータ分析方法である。

R では、分散分析を行う関数 `aov` が用意されている。関数 `aov` に用いるデータの形式はデータフレーム型で、データ  $y_{ij}$  を 1 列に並べ、それに関連する因子及び水準の情報をそれぞれ異なる列に記録する必要がある。例えば、水準  $a_1$  は `a1`、水準  $a_2$  は `a2`、水準  $a_3$  は `a3`、水準  $a_4$  は `a4` のように表記する。ここの `a1`、`a2`、`a3`、`a4` は水準の群(groups)を識別するラベルである。よって、文字 `A`、`B`、`C`、`D`、`あ`、`い`、`う`、`え`、`1`、`2`、`3`、`4` などを用いてもよい。

次にデータフレームの作成に関するコマンドを示す。データフレーム `bunsan1` の `A` 列が因子の水準に関する情報で、`y` 列がテストで得られた数値である。関数 `data.frame` 中の `factor` は `A` に入力するのは文字であること、`rep("a1",10)` は文字列 `a1` を 10 回生成するコマンドである。

```
>a1<-c(63,58,64,58,77,66,52,64,49,66)
>a2<-c(64,64,68,61,56,71,64,65,85,75)
>a3<-c(59,87,79,71,65,65,65,71,74,58)
>a4<-c(83,79,65,67,80,72,80,75,72,84)
```

```
>bunsan1<-data.frame(A=factor(c(rep("a1",10),rep("a2",10),rep("a3",10),rep("a4",10))),y=c(a1,a2,a3,a4))
> bunsan1
  A  y
1 a1 63
2 a1 58
<中略>
39 a4 72
40 a4 84
```

まず視覚的にデータの大まかな状況を把握するためにデータセット bunsan1 の箱ひげ図を示す。次のコマンドで図1に示す箱ひげ図が作成される。

```
>boxplot(y~A,data=bunsan1,col="lightblue")
```

図1から水準 a1、a2、a3、a4 の中心値が右上がりになっていることがわかる。この差の有意性について分散分析を行ってみよう。

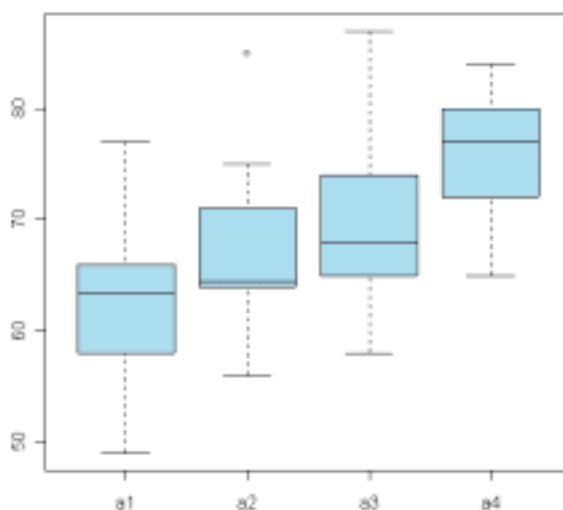


図1 表1のデータの箱ひげ図

関数 aov にデータセット bunsan1 を適したコマンドを次に示す。関数 aov を用いた分散分析の書式は aov(y~A,data=bunsan1)となるが、分散分析の結果の要約を、次のように関数 summary を用いて返すことができる。

```
> summary(aov(y~A,data=bunsan1))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	3	1003.27	334.42	5.302	0.003941 **
Residuals	36	2270.70	63.08		

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

返された結果を分散表と呼ぶ。各データを記号で対応付けた表を次に示す。

表3 分散分析表

	自由度 Df	変動 Sum Sq	不偏分散 Mean Sq	分散比 F value	P 値 Pr(>F)
因子	$k - 1$	$SS_B$	$MS_B$	$\frac{MS_B}{MS_W}$	pf(F)
誤差	$n - k$	$SS_W$	$MS_W$	$MS_W$	

表3の中の記号列について表1の一元配置データ表に基づいてその定義を説明する。データ  $y_{ij}$  は次のように分解することができる。

$$y_{ij} = \bar{y} + (\bar{y}_{\cdot j} - \bar{y}) + (y_{ij} - \bar{y}_{\cdot j})$$

ただし、記号  $\bar{y}$ 、 $\bar{y}_{\cdot j}$  は次のように定義されている。

$$\text{総平均: } \bar{y} = \frac{1}{nk} \sum_{j=1}^k \sum_{i=1}^n y_{ij}$$

$$\text{群(水準)内平均: } \bar{y}_{\cdot j} = \frac{1}{n} \sum_{i=1}^n y_{ij}$$

分解された  $\bar{y}_{\cdot j} - \bar{y}$  は群内の平均と総平均とのずれであり、 $y_{ij} - \bar{y}_{\cdot j}$  は観測値と群内の平均とのずれである。すべての観測値についてこのようなずれを分析するため次のような統計量を導入する。

次の式により定義される水準の違いにより生じるばらつきを群間の変動(variation between groups)と呼び、 $SS_B$  (sums of squares between)で示す。

$$SS_B = \sum_{j=1}^k \sum_{i=1}^n (\bar{y}_{\cdot j} - \bar{y})^2$$

水準・群の数から1を引いた値  $k - 1$  を群間の変動の自由度と言う。群間の変動を自由度で割った値は群間の不偏分散で、 $MS_B$  (mean squares between)で示す。

$$MS_B = \frac{SS_B}{k - 1}$$

水準内のデータのばらつきを次の式で定義したものを群内の変動(variation within groups)と呼び、 $SS_W$  (sums of squares within)で示す。群内の変動を変動誤差・残差(Residuals)とも呼ぶ。

$$SS_W = \sum_{j=1}^k \sum_{i=1}^n (y_{ij} - \bar{y}_{\cdot j})^2$$

群内の変動  $SS_W$  の自由度はデータの総数  $N$  から群数  $k$  を引いた値  $N - k$  である。群内の不偏分散を  $MS_W$  (mean squares within) で示す。

$$MS_W = \frac{SS_W}{N - k}$$

データの群間に差があるかどうかを比較するには、群間の不偏分散と群内の不偏分散について等分散検定を行うことが考えられる。等分散検定が採択されると群の平均値には有意な差がないと言える。分散の比は  $F$  分布にしたがう。群間と群内の不偏分散の比を次に示す。

$$F = \frac{\text{群間の不偏分散}}{\text{群内の不偏分散}} = \frac{MS_B}{MS_A}$$

関数 aov の結果には検定統計量  $F$  に対応する  $P$  値  $[\text{Pr}(>F)]$  が返される。この  $\text{Pr}(>F)$  を限界水準とも呼ぶ。上記の例では  $\text{Pr}(>F)$  値が約 0.004 である。この値は、通常仮説検定に用いる有意水準 0.01 より小さい。よって、有意水準 0.01 で「各水準の平均値が等しい」という仮説が棄却され、群間には差があると統計的に判断される。この結果は機器の性能は安定せず、温度の変化に伴い性能が変化していることを意味する。

各水準のデータの数が一致しない場合でも同様な分析の方法が適応できる。

## 2.2 対応ありの場合

前項の例のデータを 10 台の機器を異なる条件の下で行ったテストの結果として書き直し、表 4 に再掲する。表 4 のデータは見かけ上では表 1 のデータと似ているが、本質的には異なる。表 1 では同水準においてデータの行の順序には意味がないが、表 4 では 1 行が 1 台の機器に対するテストの結果であるので行のデータがセットになっている。このようなデータを対応ありのデータと言う。

表 4 10 台の機器のテスト結果  
(繰り返しなし、対応あり)

機器 の 番号	要因・因子 A			
	$a_1$ - 20° C	$a_2$ 0° C	$a_3$ 20° C	$a_4$ 40° C
No.1	63	64	59	83
No.2	58	64	87	79
No.3	64	68	79	65
No.4	58	61	71	67
No.5	77	56	65	80
No.6	66	71	65	72
No.7	52	64	65	80
No.8	64	65	71	75
No.9	49	85	74	72
No.10	66	75	58	84
平均	61.7	67.3	69.4	75.7

R でこのデータを、関数 `aov` を用いて分析するためには、前項のデータセット `bunsan1` をそのまま用いることはできない。それはデータセット `bunsan1` には対応ありに関する情報が欠けているからである。対応ありに関する情報を付加したデータセットを作成するコマンドを次に示す。ここでは前項で作成したデータベクトル `a1`、`a2`、`a3`、`a4` を用いる。コマンドの中の `A` は水準に関する情報であり、`No` はデータの「対応あり」に関する情報である。

```
>bunsan2<-data.frame(A= factor(c(rep("a1",10), rep("a2",10), rep("a3",10), rep("a4",10))),No=
factor(rep(1:10, 4)),y=c(a1,a2,a3,a4))
> bunsan2
  A No y
1 a1 1 63
2 a1 2 58
<中略>
39 a4 9 72
40 a4 10 84
```

関数 `aov` を用いた書式と返された結果を次に示す。

```
> summary(aov(y ~ A+No, bunsan2))
          Df Sum Sq Mean Sq F value Pr(>F)
A           3 1003.27   334.42   4.3695 0.01243 *
No          9   204.22    22.69   0.2965 0.96952
Residuals  27  2066.48    76.54
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

返された `A` の行が水準・群に関する分散分析の検定統計量であり、その `Pr(>F)` 値・限界水準は約 0.012 である。この `Pr(>F)` 値は通常の仮説検定に用いる有意水準 0.05 より小さい。よって因子 `A` の水準・群の平均値は有意水準 0.05 で差があると言える。`No` の行が機器に関する分散分析の検定統計量で、その `Pr(>F)` 値は約 0.97 である。この値は通常用いられている有意水準よりはるかに大きく、10 台の機器には差があると判断できない。つまり、因子 `A` の水準の平均値には差が認められるが、機器の間には有意の差が認められない結果となっている。

このように、見かけ上は同じデータであってもデータを収集する際の前提条件が異なると分散分析を行う方法や結果の解釈も異なる。分散分析は、基本的な考え方は同じであるがデータセットのタイプによってその解析方法が微妙に異なる。

### 3. 二元分散分析

表 5 のように何らかの 2 つの要因を考慮したデータを二元(Two-way)配置と言う。表 5 には 2 つ形式で示しているが本質的には同じである。二元分散分析では、表 5 のような二元配置データにおける因子 `A` の効果、因子 `B` の効果、因子 `A` と `B` との交互作用効果などについて分析を行う。ここで言う交互作用効果(interaction)とは、異なる因子の効果が絡んで生じる効果で

ある。

表5 二元配置のデータ(対応なし)

表 (a)

		因子 B		
		$b_1$	$b_2$	$b_3$
因子 A	$a_1$	3	6	6
		3	4	7
		5	5	8
		4	6	7
	$a_2$	3	3	3
		5	4	4
		2	5	3
		4	3	2

表 (b)

因子 A	$a_1$			$a_2$		
因子 B	$b_1$	$b_2$	$b_3$	$b_1$	$b_2$	$b_3$
	3	6	6	3	3	3
	3	4	7	5	4	4
	5	5	8	2	5	3
	4	6	7	4	3	2

まず表5の二次元配置データを次のようにR用のデータセット bunsan3 を作成する。

```
>a1<-c(3,3,5,4,6,4,5,6,6,7,8,7)
>a2<-c(3,5,2,4,3,4,5,3,3,4,3,2)
>bunsan3<-data.frame(A=factor(c(rep("a1",12),rep("a2",12))),B=factor(rep(c(rep("b1",4),
rep("b2",4), rep("b3",4)),2)),y=c(a1,a2))
> bunsan3
  A B y
1 a1 b1 3
2 a1 b1 3
<中略>
23 a1 b3 3
24 a1 b3 2
```

交互作用効果を見せず、両因子の効果のみについて分析を行う場合は、関数 aov を次のように用いる。

```
> summary(aov(y~A+B,data=bunsan3))
          Df Sum Sq Mean Sq F value    Pr(>F)
A             1  22.042   22.042  13.8482 0.001349 **
B             2   7.750    3.875   2.4346 0.113161
Residuals    20  31.833    1.592
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

返された結果のAの行が因子Aの分散分析の結果である。そのPr(>F)値は約0.0014で、有

意水準 0.01 より小さい。よって、A 因子の 2 つの水準  $a_1$ 、 $a_2$  は有意水準 0.01 でその平均値には差があると判断される。B の行が因子 B の分散分析の結果であり、その  $\text{Pr}(>F)$  値は 0.113 である。よって、B 因子の 3 つの水準  $b_1$ 、 $b_2$ 、 $b_3$  は有意水準 0.1 でも有意の差があると言えない。

表 5 の因子 A の水準  $a_1$  のみを見た限りでは水準  $b_1$ 、 $b_2$ 、 $b_3$  には異なりがありそうにも関わらずこのような大きな  $\text{Pr}(>F)$  値が得られている。そこで、交互作用効果を考慮した分散分析の統計量を求めてみよう。交互作用効果を考慮した場合は、次のように因子の間に記号\*を用いる。

```
> summary(aov(y~A*B,data=bunsan3))
              Df Sum Sq Mean Sq F value    Pr(>F)
A                1  22.0417  22.0417  23.0000 0.0001447 ***
B                 2   7.7500   3.8750   4.0435 0.0354521 *
A:B               2  14.5833   7.2917   7.6087 0.0040287 **
Residuals       18  17.2500   0.9583
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

返された結果では、因子 B の  $\text{Pr}(>F)$  値は 0.035 で有意水準 0.05 より小さくなっており有意水準 0.05 で  $b_1, b_2, b_3$  に有意な差があると判断される。A:B の行が交互作用効果に関する分散分析の結果である。その  $\text{Pr}(>F)$  値が 0.004 で、0.01 より小さいので有意水準 0.01 で交互作用効果があると判断される。

交互作用効果に関しては折れ線グラフでも視覚的に考察することができる。R には交互作用グラフを作成するための関数 `interaction.plot` が用意されている。次に関数 `interaction.plot` を用いた交互作用図の作成コマンドと結果を示す。

```
> attach(bunsan3)
> interaction.plot(A,B,y)
```

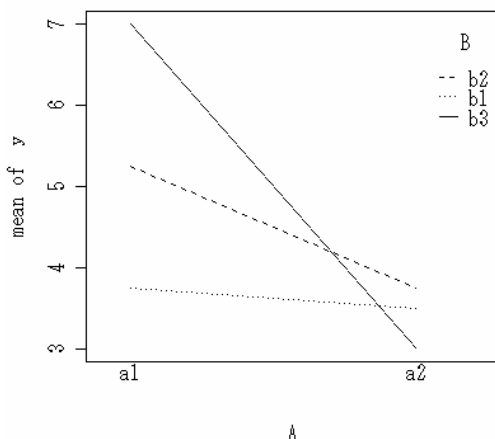


図 2 交互作用図

交互作用効果が存在しない場合は、折れ線が平行になり、交互作用効果が存在する場合は平行にならず、交互作用効果が大きい場合は交差現象が伴う。図2では交互作用効果があることがわかる。交互作用に関しては、本誌の連載「統計数理はいま・・・」で連載中である(第56回(2004年2月)~)分散分析の専門家広津千尋教授の文章が詳しい。

二元分散分析は、データの対応ありとなしで表6に示すように4つのタイプに分けられる。紙面上の都合により、本稿では二元配置についてタイプ の例のみを示している。分散分析は各タイプのデータ構造に基づいて分析を行わなければならない。分散分析のソフトを用いる際には、データ構造のタイプごとに分析方法を選択し、引数などの設定を行わなければならない。初心者にとっては多少煩雑に感じるであろう。

表6 二元配置のタイプ

	因子A	因子B
タイプ	対応なし	対応なし
タイプ	対応あり	対応あり
タイプ	対応なし	対応あり
タイプ	対応あり	対応なし

中部大学人文学部心理学科松井孝雄助教授は「データ解析テクニカルブック、森・吉田編著、北大道書房」の中の十数タイプのデータを用いて「言語Rによる分散分析」というタイトルでR用のデータセットの作成及び分散分析のコマンドを公開している。

<http://mat.isc.chubu.ac.jp/>

### 3. 分散分析の関連関数

本稿では、分散分析についてRに用意された関数 `aov` を用いた分散分析の例を示している。Rでは線形モデルの関数 `lm` と関数 `anova` を用いて同様な分散分析を行うことも可能である。これらの関数による分散分析は、誤差が標準正規分布にしたがうと言う仮定のもとで行っている。しかし、本稿では分散の正規性などについてはまったく触れる余裕がなかった。

一元分散分析はこれらの関数以外に、Rに用意されている関数 `pairwise.t.test`、`oneway.test` などを用いて平均検定を行うことも可能である。また関数 `kruskal.test` による分散の正規性の仮定を置かないノンパラメトリックな分散分析を行うことも可能である。

二元(多重)配置のノンパラメトリックな分散分析方法としてはフリードマン(Friedman)の関数 `friedman.test` がRに用意されている。

また群馬大学社会情報学部青木教授は自作のプログラムを次のホームページで公開している。  
<http://aoki2.si.gunma-u.ac.jp/R/>