人間行動システム統計演習 B

第 3 回：Analysis of Variance 2　担当：前川，栗山，荒井

2012 年 6 月 20 日

＊ OCWi から，本日分のデータを[*1]を自分の実習用フォルダへコピーしておいてください。

1　一元配置分散分析

3 種類の紅茶があり，そのおいしさを評定してもらったという状況を考えます。紅茶の種類により，おいしさに違いはあるでしょうか。紅茶（Tea）の種類を A，B，C とし，おいしさの点数を Score とします。

- 帰無仮説 H_0：3 種類の Tea において，おいしさの点数（Score）の母平均が等しい。
- 対立仮説 H_1：3 種類の Tea において，おいしさの点数（Score）の母平均が等しくない（少なくとも 1 対の群間で母平均に差がある）。

3 つ以上の標本について，平均値の差の有無を検定する場合には，t 検定ではなく分散分析を行います[*2]。

1.1　一元配置分散分析　対応なし

対応のない一元配置分散分析では，全体の平方和 SS_{total} を群間平方和 SS_A(between-groups sum of squares) と群内平方和 SS_e(within-group sum of squares) とに分けて考えます。

要因を A，その水準数を a とします。第 j 群に n_j 人の被験者を無作為に割り当て，第 j 群の第 i 番目の被験者の観測値を x_{ij} とします。全被験者数を $N = \sum_{j=1}^{a} n_j$，全体の平均値を $\bar{x}$，第 j 群における平均値を $\bar{x}_j$ とすると，

$$\sum_{j=1}^{a}\sum_{i=1}^{n_j}(x_{ij}-\bar{x})^2 = \sum_{j=1}^{a} n_j(\bar{x}_j-\bar{x})^2 + \sum_{j=1}^{a}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2$$

であり，

$$SS_{total} = SS_A + SS_e$$

が成り立ちます。

このとき，検定統計量 F は次のように求められ，

$$F = \frac{SS_A/(a-1)}{SS_e/(N-a)}$$

F は帰無仮説が真のとき，第 1 自由度 $a-1$，第 2 自由度 $N-a$ の F 分布に従います。この F の値が自由度 $((a-1),(N-a))$ の F 分布における上側 5 %点より大きな値ならば，帰無仮説を棄却し，対立仮説を採択することになります。

[*1] 今回の実習で用いるデータは『データ解析・R 言語』渡辺利夫著 p35 に掲載されているデータを用いさせていただきました。

[*2] t 検定を繰り返してはいけない理由については，統計演習 A でも扱っていると思います。

データについて

まず，Tea1way.txt を data1 として読み込みます。被験者の変数 Subj を因子型に変更しておきます。続いて，箱ひげ図 (Boxplot) を描いてデータの確認を行います (図??)。

```
> data1 <- read.table("Tea1way.txt", header=TRUE, sep="")
> data1$Subj <- as.factor(data1$Subj) #因子型に変換
> data1
   Subj Tea Score
1     1   A     9
2     2   A     9
3     3   A     7
4     4   A     8
5     5   A     8
6     6   A     7
7     7   A     6
8     8   A     5
9     9   A     6
10   10   A     5
11    1   B     8
  ...
29    9   C     4
30   10   C     5
>
> #データの確認
> boxplot(Score ~ Tea, data=data1)
```

対応なしの場合には，次の表のように 30 人の異なる被験者の結果とみなします。

Tea		
A	B	C
9	8	6
9	7	7
7	6	6
8	4	5
8	5	7
7	7	6
6	6	5
5	7	3
6	5	4
5	8	5

では，分析を行ってみましょう。aov() 関数を用います。関数 aov はバランスのとれた実験デザイン，すなわち，要因のどの水準にも同数ずつのデータが得られている場合に利用できる関数です。書式は次の通りです。

aov(目的変数 ~ 群分けのための変数)

★分散分析では，各要因の各水準のデータ数が同じになるように被験者を割り当てるのが原則です。

```
> anova.1 <- aov(Score ~ Tea, data=data1)
> summary(anova.1)
            Df Sum Sq Mean Sq F value Pr(>F)
Tea          2 12.867  6.4333  3.4396 0.0467 *
Residuals   27 50.500  1.8704
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```

分散分析表が出力されます。p 値が 0.0467 なので，5 %水準で「紅茶 Tea の種類により，おいしさの点数

Score の平均値に有意な差がある」と言うことができます。

続いて，どの紅茶に差があるのかを見るために，多重比較を行います。ここでは，Tukey の方法を用います。関数 TukeyHSD() を利用します。

```
> TukeyHSD(anova.1, "Tea")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Score ~ Tea, data = data1)

$Tea
    diff       lwr         upr     p adj
B-A -0.7 -2.216451  0.81645141 0.4957759
C-A -1.6 -3.116451 -0.08354859 0.0371031
C-B -0.9 -2.416451  0.61645141 0.3200947

> plot(TukeyHSD(anova.1, "Tea"))
```

- diff グループ間の平均値差
- lwr,upr 95% 信頼区間の下限と上限
- p adj p 値

C-A のところの p 値が有意水準 0.05 よりも小さいことから，紅茶 A と紅茶 C の間に 5% 水準で有意な差があると言えます。

TukeyHSD の結果を plot すると信頼区間の図が描かれます (図??)。

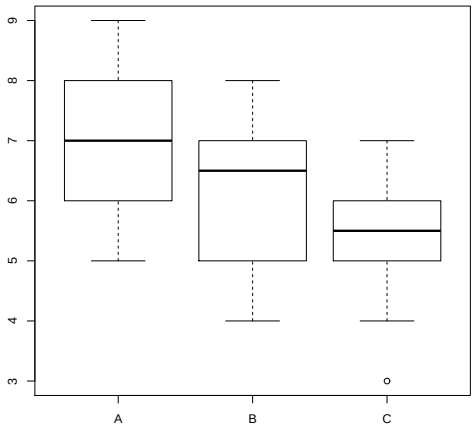

図 1

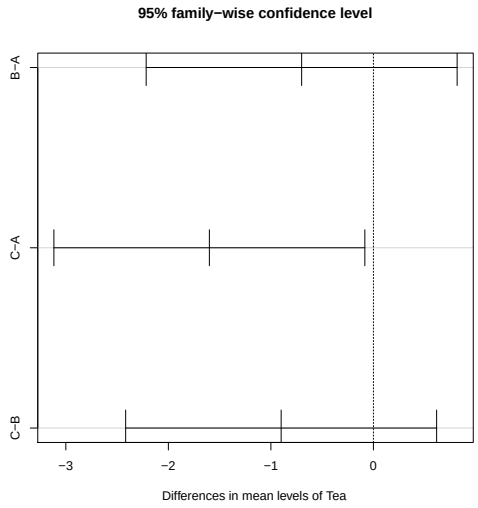

図 2

1.2 一元配置分散分析 対応あり

続いて，上の表のように 10 人の被験者 Subj1〜10 が，それぞれ紅茶 A〜C を飲み，その得点のデータがあると考えます。このようなデータを対応のあるデータと言います。

	Tea		
Subj	A	B	C
1	9	8	6
2	9	7	7
3	7	6	6
4	8	4	5
5	8	5	7
6	7	7	6
7	6	6	5
8	5	7	3
9	6	5	4
10	5	8	5

対応ありの場合には，データ全体のばらつき SS_{total} をさらに被験者の平均のばらつき SS_S を用いて次のように分けます。

$$SS_{total} = SS_A + SS_S + SS_e$$

各群に割り当てる人数を n とすると，次の検定統計量 F は

$$F = \frac{SS_A/(a-1)}{SS_e/(n-1)(a-1)}$$

帰無仮説が真のとき，第 1 自由度 $a-1$，第 2 自由度 $(n-1)(a-1)$ の F 分布に従います。この F の値が自由度 $((a-1),(n-1)(a-1))$ の F 分布における上側 5 %点より大きな値ならば，帰無仮説を棄却し，対立仮説を採択することになります。

```
> anova.1w <- aov(Score ~ Tea + Error(Subj), data=data1)
> summary(anova.1w)

Error: Subj
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  9   24.7  2.7444

Error: Within
          Df Sum Sq Mean Sq F value  Pr(>F)
Tea        2 12.867  6.4333  4.4884 0.02621 *
Residuals 18 25.800  1.4333
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```

aov の引数のところで Error(Subj) が追加されています。個人差に相当する要因を Error で指定しています。

分散分析表を見ると，Tea に関する検定において p 値が 0.02621 なので，「紅茶の種類によりおいしさの母平均に有意な差がある」と言うことができます。

ちなみに，aov(Score ~ Tea + Subj, data=data1) でも同じ結果が得られます。

```
> anova.1w2 <- aov(Score ~ Tea + Subj, data=data1)
> summary(anova.1w2) #同じ結果
            Df Sum Sq Mean Sq F value  Pr(>F)
Tea          2 12.867  6.4333  4.4884 0.02621 *
Subj         9 24.700  2.7444  1.9147 0.11513
Residuals   18 25.800  1.4333
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

多重比較

続いて，どの紅茶間に差があるのかを見るために，多重比較を行います。

```
> pairwise.t.test(data1$Score, data1$Tea,
+                 p.adjust.method="bonferroni", paired=TRUE )


        Pairwise comparisons using paired t tests

data:  data1$Score and data1$Tea

  A      B
B 0.9650 -
C 0.0016 0.4401

P value adjustment method: bonferroni
```

各組合せの p 値が出力されます。この結果から，紅茶 A と紅茶 C の間に有意な差があると言うことができます。

2　二元配置分散分析

次に，紅茶（Tea) の違いと，その温度 (Temp) の違いにより，得点 (Score) に違いがあるかを調べます。Tea は，紅茶 A，B，C の 3 種類の値を取り，Temp は H，C の 2 種類とします。

データについて

Tea2way1.txt を data2 として読み込みます。このデータは data1 に Temp の変数が加わったものです。先ほどと同様，被験者の変数 Subj を因子型に変更しておきます。続いて，箱ひげ図 (Boxplot) を描いてデータの確認を行います。

```
> data2 <- read.table("Tea2way1.txt", header=TRUE, sep="")
> data2$Subj <- as.factor(data2$Subj) #因子型に変換
> data2
   Subj Tea Temp Score
1     1   A    H     9
2     2   A    H     9
 ...
30   10   C    C     5
```

```
#データの確認
> boxplot(Score ~ Tea, data=data2)
> boxplot(Score ~ Test, data=data2)
```

※ R コマンダーというパッケージを用いると，次の図のような水準ごとの平均値のグラフを描いてくれます。研究室に戻ったら試してみて下さい。

```
#データの確認
> library(Rcmdr)
> plotMeans(data2$Score, data2$Tea,data2$Test, error.bars="sd")
```

2.1　二元配置分散分析 対応なし

下の表のように 30 人の異なる被験者の結果とみなして，紅茶の違いやと温度の違いによって成績に違いがあるかを分析してみます。

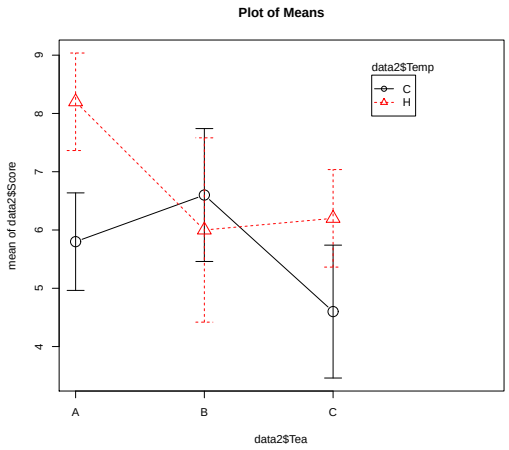

図 3　平均値の推移図

Tea A		Tea B		Tea C	
Temp H	Temp C	Temp H	Temp C	Temp H	Temp C
9	7	8	7	6	6
9	6	7	6	7	5
7	5	6	7	6	3
8	6	4	5	5	4
8	5	5	8	7	5

```
> anova.2 <- aov(Score ~ Tea*Temp, data=data2)
> summary(anova.2)
            Df  Sum Sq Mean Sq F value   Pr(>F)
Tea          2 12.8667  6.4333  5.3611 0.011892 *
Temp         1  9.6333  9.6333  8.0278 0.009188 **
Tea:Temp     2 12.0667  6.0333  5.0278 0.015006 *
Residuals   24 28.8000  1.2000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

モデル式は，`Score ~ Tea*Temp` のように指定します。モデル式の記法には次のようなものがあります。

~	左辺の変数を、右辺の変数でモデル化する
+ A	説明変数 A を加える
A:B	説明変数 A, B の交互作用項
A * B	A + B + A:B と同じ

Tea，Temp による主効果，及び交互作用が有意であることが分かります。続いて，主効果が有意なので，主効果に関する多重比較を，また，交互作用が有意なので、単純主効果（ある要因のある水準における，別の要因の効果）の有意差検定を行うことになりますが，ここでは省略します。

2.2　二元配置分散分析 1 要因のみ対応あり

次に 10 人の被験者 1〜10 が，紅茶 A,B,C の全てを飲んでいるが、温度については H か C のどちらかのみを飲んでいる場合を考えます。

	Temp H				Temp C		
Subj	Tea A	Tea B	Tea C	Subj	Tea A	Tea B	Tea C
1	9	8	6	6	7	7	6
2	9	7	7	7	6	6	5
3	7	6	6	8	5	7	3
4	8	4	5	9	6	5	4
5	8	5	7	10	5	8	5

```
> anova.2w1 <- aov(Score ~ Tea*Temp+Error(Subj/Temp), data=data2)
 警告メッセージ:
In aov(Score ~ Tea * Temp + Error(Subj/Temp), data = data2) :
   Error() モデルは特異です
> summary(anova.2w1)

Error: Subj
          Df  Sum Sq Mean Sq F value  Pr(>F)
Temp       1  9.6333  9.6333   5.115 0.05358 .
Residuals  8 15.0667  1.8833
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Error: Within
          Df Sum Sq Mean Sq F value   Pr(>F)
Tea        2 12.867  6.4333  7.4951 0.005048 **
Tea:Temp   2 12.067  6.0333  7.0291 0.006445 **
Residuals 16 13.733  0.8583
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

aov を実行すると「警告メッセージ」が出ていますが，分析自体は正しく行われます。

分析の結果，Temp の主効果は 5% 水準で有意ではありません。Tea の主効果及び Tea と Temp の交互作用は有意であることが分かります。

モデル式の Error 項は，Error(Subj:Temp+Subj:Temp:Tea) のように指定しても同じ結果となります。今回の例では，同じ温度に関して，同じ人の得点が 3 つずつあり，ここに個人差があらわれることになります。そこで，個人差 Subj と温度 Temp が関係する組合せを Error で指定しています。

2.3 二元配置分散分析　２要因とも対応あり

最後に，5 人の被験者が全員、紅茶 A,B,C を飲み、温度 H と C の両方の得点を出している場合を考えましょう。

	Tea A		Tea B		Tea C	
Subj	Temp H	Temp C	Temp H	Temp C	Temp H	Temp C
1	9	7	8	7	6	6
2	9	6	7	6	7	5
3	7	5	6	7	6	3
4	8	6	4	5	5	4
5	8	5	5	8	7	5

Tea2way2.txt を data3 として読み込みます。

```
> data3 <- read.table("Tea2way2.txt", header=TRUE, sep="")
> data3$Subj <- as.factor(data3$Subj) #因子型に変換
> data3
   Subj Tea Temp Score
1     1   A    H     9
2     2   A    H     9
    ...
30    5   C    C     5
```

続いて分析を行います。

```
> anova.2w2 <- aov(Score ~ Tea*Temp+Error(Subj/(Tea*Temp)),
+                  data=data3)
> summary(anova.2w2)

Error: Subj
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  4   13.2     3.3

Error: Subj:Tea
          Df Sum Sq Mean Sq F value  Pr(>F)
Tea        2 12.867  6.4333  7.5686 0.01429 *
Residuals  8  6.800  0.8500
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Error: Subj:Temp
          Df Sum Sq Mean Sq F value  Pr(>F)
Temp       1 9.6333  9.6333  20.643 0.01047 *
Residuals  4 1.8667  0.4667
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Error: Subj:Tea:Temp
          Df  Sum Sq Mean Sq F value  Pr(>F)
Tea:Temp   2 12.0667  6.0333  6.9615 0.01773 *
Residuals  8  6.9333  0.8667
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

モデル式の Error 項は，Error(Subj+Subj:Tea+Subj:Temp+Subj:Tea:Temp) のように指定しても同じ結果となります。

Tea の主効果，Temp の主効果，Tea と Temp の交互作用がいずれも有意であると言えます。

3　本日の課題

`ANOVA-kadai.txt` は，時間帯 (Time)(朝：1，夕：2）ごとに 3 種類のテスト (Test) を行ったときの成績についての架空のデータです。このデータを読み込んでください。

1. 12 人の被験者のうち，1〜6 までは朝に，7〜12 は夕方に，3 種類のテストを受けたデータだとみなして，1 要因のみ対応のある二元配置分散分析を行ってください。検定結果とともに，簡単な解釈も付けてください。（ヒント：被験者として Subj2 を使います）
2. 6 人の被験者 1〜6 が，朝夕ともに，3 種類のテストを受けたデータだとみなして，2 要因ともに対応のある二元配置分散分析を行ってください。検定結果とともに，簡単な解釈も付けてください。（ヒント：被験者として Subj1 を使います）
3. 1 要因のみ対応のある場合（課題 1）と 2 要因とも対応のある場合（課題 2）の違いを簡単に述べてください。

・注意　独立変数を `as.factor()` 関数で因子型に変換してから分析してください。

課題提出について

課題は，6 月 26 日 1：00AM までに，OCWi で提出してください。今回は，結果の解釈についての記述を含みますので，「R の出力結果と解釈」を txt 形式あるいは PDF 形式で提出して下さい。簡単な解釈は 2〜3 行で構いません。

- ファイル名：「学籍番号-B3.txt」あるいは「学籍番号-B3.pdf」
- 1 行目：学籍番号，名前
- 2 行目以降：課題

担当者メールアドレス：sayarai@rd.dnc.ac.jp