医学統計学演習（看護）：資料 6

芳賀昭弘 [*]

1　演習

1. 看護の教育方法として、講義型（L)、実習型（E）の２つの形態による理解度の違いを調べた。それぞれの授業に参加した学生７名（A∼G) をランダムに抽出し、最終テストの得点で比較することにした。
   (1) これは対応ありですか？対応なしですか？ (どちらを選んでも理由が間違いでなければ正答)
   (2) それぞれの母集団が正規性を満たしている（「いる」に下線）として、L と E に差があるかを評価しなさい。
   (3) それぞれの母集団が正規性を満たしていない（「いない」に下線）として、L と E に差があるかを評価しなさい。

Student	A	B	C	D	E	F	G
L	51	66	70	75	73	62	55
E	55	37	47	60	62	53	50

---

[*] Electronic address: haga@tokushima-u.ac.jp

2　復習

被験者全員に同じ薬を使って投薬前後での白血球の量を調べた。なお母集団のデータは正規分布に従っていないものとする。投薬前後の白血球の量に違いがあるかどうかを調べるには、どのような検定が必要でしょう？

第１群は地域 A、第２群は地域 B に住む患者群として、それぞれの血圧を調べた。それぞれの群の母集団は正規分布に従うとする。地域によって血圧が異なるかどうかを調べるには、どのような検定が必要でしょうか？

ここまでは、２群比較のための検定を行ってきました。ですが、上の１つ目の例で言えば、投薬前、投薬直後、投薬後１時間、、、（２つ目の例でも、例えば地域 C, D, $\cdots$）のように２群以上の標本の比較を行いたい場合があります。3 群以上の比較を行いたいときに使う統計的方法として、**分散分析（Analysis of variance: ANOVA）**と呼ばれるものがあります[*1]。

分散分析において有意差が出た（＝どれか２群の間に差が出たということ）場合、次にどの群に差が出たのかを調べたいと思うのが常です。群間の比較を調べるには**多重比較法**を使います（なお、分散分析と絡めずに最初から多重比較法だけで行うほうが良いと主張する研究者も多くいますが、この演習では両方を学んでおきたいと思います)。

分散分析や多重検定を手計算で行うのは大変です。しかし、コンピュータでは**たった一行で**それを計算してくれるコマンドが存在します。簡単です！一方、コンピュータで計算する際に大事なことは、**読み込むデータの書き方**になります。この演習では、**データの書き方（エクセルの作り方）**を学ぶことが実用上一番の目的となります。

3　一元配置分散分析（対応なし）

StatisticsTest.csv というデータを用意しています。まずはどういう構造になっているかエクセルで開いて確認しておきましょう (A∼D はある勉強方法を示しており、数字はそのような勉強方法によるテストの結果（点数）を示しているとします)。確認したら、StatisticsTest.csv を R で読み込んでください。

```
> data <- read.csv(“StatisticsTest.csv”)
> data
```

data に格納されたら、aov 関数を使って分散分析を行います。p 値はいくらでしょう？

```
> summary(aov(data$Statistics ∼ data$Instruction))
```

帰無仮説 H_0:　４群の母平均は等しい。

帰無仮説 H_1:　４群の母平均は等しくない。

有意水準: 5%

[*1] 3 群以上の検定の場合でも、2 群の t-検定を 3 回行うことで同じようなことができると考える人がいるかもしれませんが、これはやってはいけません！3 群以上のグループ比較を行いたいのであれば、必ず分散分析を行ってください。理由は医学統計学の講義で説明しています。

p 値が＿＿＿＿＿＿であるので、統計検定量は有意水準＿＿＿であり、帰無仮説は＿＿＿＿＿＿＿＿＿＿＿＿＿＿

ここでわかったことは、4群の組み合わせのうち、少なくとも1つの組み合わせで差がありそうだということだけで、どの組み合わせに差があるのかまでは教えてくれていません。そのため、次に多重比較を行う流れになります。ここでは Tukey の方法という多重比較法を用いてみます。

```
> TukeyHSD(aov(data$Statistics ∼ data$Instruction))
```

有意水準を 5% とした場合、有意差が生じたのはどの組み合わせでしょうか？

＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

4 一元配置分散分析（対応あり）

SubjectAssesment.csv というデータを用意しています。同じくどういう構造になっているかエクセルで開いて確認しておきましょう（A∼E の学生の代数（Linear Algebra）、微積（Calculus）、確率統計（Probability and Statistics）の好感度と考えることにします）。今度も aov 関数を使って解析します。ただし、前回と違ってデータをそのまま使って解析することはできません。**aov 関数を使う場合、データ構造を図 1 に示すような形式にしないといけないのです**[*2]。

データ形式の修正は、勿論、エクセル上で行ってもよいし R 上で行ってもよいです（データ数が少ない場合、エクセルでも十分ですが、膨大なデータを扱う場合には R や他のプログラム言語を使用しないとできない場合もあります）。今回は、エクセル上で編集してみてください。図 1 右のように修正したら、作業ディレクトリに csv 形式で保存して次を実行してください（ここでは SubjectAssesment2.csv という名前で保存したことにします）。

```
> data <- read.csv(“SubjectAssesment2.csv”)
> summary(aov(data$Assesment ∼ data$Student + data$Subject))
```

帰無仮説 H_0: 3科目の学生評価は等しい。

帰無仮説 H_1: 3科目の学生評価は等しくない。

有意水準: 5%

p 値が＿＿＿＿＿＿であるので、統計検定量は有意水準＿＿＿であり、帰無仮説は＿＿＿＿＿＿＿＿＿＿＿＿＿＿

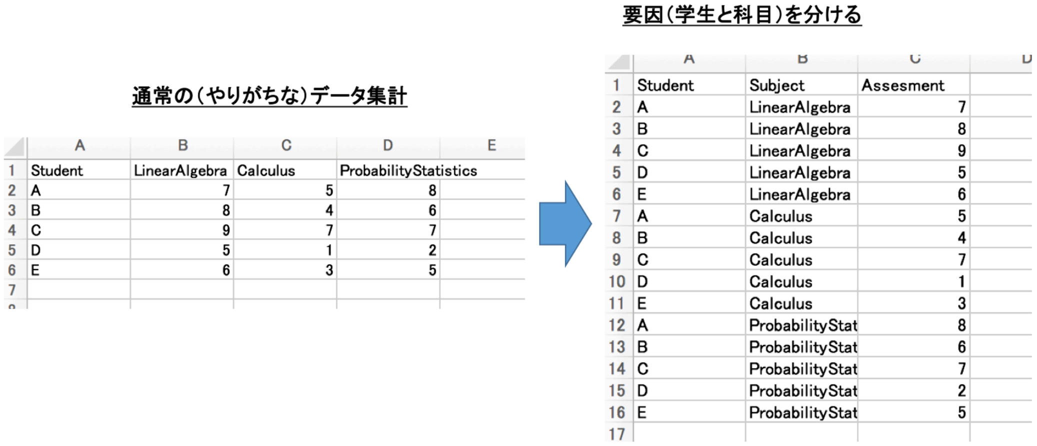
多重比較も行ってみましょう。

```
> TukeyHSD(aov(data$Assesment ∼ data$Student + data$Subject))
```

ところで、今のデータでは、一人の学生が3つの科目を評価しました。これがもし学生が全てばらばらの 15 人で評価したデータであった場合にはどうすれば良いでしょうか？その場合は、対応なしの分散分析を行います。そう思って分析してみましょう。

```
> data <- read.csv(“SubjectAssesment2.csv”)
> summary(aov(data$Assesment ∼ data$Subject))
> TukeyHSD(aov(data$Assesment ∼ data$Subject))
```

[*2] 統計解析のテクニカルな面で最も重要なことは、データをどのような形式で集めて表記しておくか、につきます。卒業研究などで統計解析をする場合には、結果が出た後にどのような統計解析が必要であるかを調べておき、解析がやりやすい形式で結果をまとめておくことが肝要です。



通常の（やりがちな）データ集計

	A	B	C	D	E
1	Student	LinearAlgebra	Calculus	ProbabilityStatistics	
2	A	7	5	8	
3	B	8	4	6	
4	C	9	7	7	
5	D	5	1	2	
6	E	6	3	5	
7					

要因（学生と科目）を分ける

	A	B	C	D
1	Student	Subject	Assesment	
2	A	LinearAlgebra	7	
3	B	LinearAlgebra	8	
4	C	LinearAlgebra	9	
5	D	LinearAlgebra	5	
6	E	LinearAlgebra	6	
7	A	Calculus	5	
8	B	Calculus	4	
9	C	Calculus	7	
10	D	Calculus	1	
11	E	Calculus	3	
12	A	ProbabilityStat	8	
13	B	ProbabilityStat	6	
14	C	ProbabilityStat	7	
15	D	ProbabilityStat	2	
16	E	ProbabilityStat	5	
17				

図 1　データ形式の修正：データ収集の際には左図のように集めるかもしれないが、aov 関数を使いたいのであれば右のような形式で集めておくと便利。

対応ありの分散分析と何が違うかわかりますか？そう、対応ありの場合、
一人の学生が３つの科目を評価することによって、個人差を評価できる
ようになり、個人のばらつきを考慮して科目をより正しく評価できるようになります。

5　二元配置分散分析（対応なし）

今度は評価の要因が 2 つある場合を考えます。TasteAssesment.csv を用意しています。どういう構造になっているかエクセルで開いて確認してください（これは一列目：保存状態 A, B に対し、二列目：水メーカー各社の水（I, V, B）の三列目：美味しさを点数付けしたデータです。美味しさの要因は、保存状態とメーカーの種類の 2 つが考えられますね。なお、修正が必要ないように形式を整えています）。やることは同じですが、要因が複数ある場合には交互作用を加えて評価します。交互作用とは複数の要因が合わさることで生じる効果のことで、解析では（因子）:（因子）という形式で書きます。交互作用を含めて分析する場合には、* を使うこともできます。以下を実行しましょう。

```
> data <- read.csv(“TasteAssesment.csv”)
> summary(aov(data$Assesment ~ data$Temp + data$Maker + data$Temp : data$Maker))
```

もしくは、

```
> summary(aov(data$Assesment ~ data$Temp * data$Maker))
```

帰無仮説 H_0: 保存方法（Temp）によって評価は変化しない。

帰無仮説 H_1: 保存方法（Temp）によって評価は変化する。

有意水準: 5%

p 値が＿＿＿＿＿＿であるので、統計検定量は有意水準＿＿＿であり、帰無仮説は＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

帰無仮説 H_0: 銘柄（Maker）によって評価は変化しない。

帰無仮説 H_1: 銘柄（Maker）によって評価は変化する。

有意水準: 5%

p 値が＿＿＿＿＿＿であるので、統計検定量は有意水準＿＿＿であり、帰無仮説は＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

帰無仮説は＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

帰無仮説 H_0: 保存方法と銘柄の組み合わせによって評価は変化しない。

帰無仮説 H_1: 保存方法と銘柄の組み合わせによって評価は変化する。

有意水準: 5%

p 値が____________であるので、統計検定量は有意水準______であり、帰無仮説は______________________________

保存状態（Temp）に対して味の評価（Assesment）を平均してみます。グラフにすると分かり易いので、次のような関数でプロットしてみましょう。

```
> interaction.plot(data$Maker,data$Temp,data$Assesment)
```

次に銘柄（Maker）に対して味の評価（Assesment）を平均してみます。

```
> interaction.plot(data$Temp,data$Maker,data$Assesment)
```

これらの図を他の人に説明してみてください。横軸と縦軸は何でしょうか？これらの図から何がわかりますか？

____

6　二元配置分散分析（２要因とも対応あり）

上と同じデータを使い、“２要因とも対応あり”の場合に拡張してみたいと思います。“２要因とも対応あり”とは、同じ人が保存方法や銘柄の違いをそれぞれ評価した、ということです（裏を返せば、前節の結果は、全て異なる人で保存方法や銘柄による味のデータを集めていたということです)。TasteAssesment.csv に評価者（Student という項目で A∼E を上から繰り返してください）を加えて TasteAssesment2.csv という名前で保存してください。わからない、もしくはヒントが欲しい人は TasteAssesment2hint.csv をダウンロードしてください。分析は次のように行います。

```
> data <- read.csv(“TasteAssesment2.csv”)
> summary(aov(data$Assesment ∼ data$Temp * data$Maker + Error(data$Student
+ data$Student:data$Temp + data$Student:data$Maker
+ data$Student:data$Temp:data$Maker)))
```

対応なしとの違いは、後半の

Error(data$Student + data$Student:data$Temp + data$Student:data$Maker
+ data$Student:data$Temp : data$Maker))

の部分です。それでは以下を埋めましょう。

帰無仮説 H_0: 保存方法（Temp）によって評価は変化しない。

帰無仮説 H_1: 保存方法（Temp）によって評価は変化する。

有意水準: 5%

p 値が____________であるので、統計検定量は有意水準______であり、帰無仮説は______________________________

帰無仮説 H_0: 銘柄（Maker）によって評価は変化しない。

帰無仮説 H_1: 銘柄（Maker）によって評価は変化する。

有意水準: 5%

p 値が____________であるので、統計検定量は有意水準______であり、帰無仮説は______________________________

帰無仮説は______________________________

帰無仮説 H_0: 保存方法と銘柄の組み合わせによって評価は変化しない。

帰無仮説 H_1: 保存方法と銘柄の組み合わせによって評価は変化する。

有意水準: 5%

p 値が____________であるので、統計検定量は有意水準______であり、帰無仮説は______________________________

前節では、交互効果がみられませんでした。しかし今回、同じデータを使ったのですが、個人差を排除した評価になっているために有意差が出やすくなります。結果はどうだったでしょうか？

7　二元配置分散分析（１要因のみ対応あり）

続いて、“１要因のみ対応あり”という状況を考えてみます。上と同じデータを使いますが、保存方法（Temp）が B のときの評価者（Student）を F～ J に変更し、TasteAssesment3.csv という名前で保存してください（もうヒントは要らないですね？）。次を実行します。

```
> data <- read.csv(“TasteAssesment3.csv”)
> summary(aov(data$Assesment ~ data$Temp * data$Maker +Error(data$Student
+ data$Student:data$Temp + data$Student:data$Temp:data$Maker)))
```

Error() に入れる項目は、別の人が評価しているもの（この場合は、保存方法（Temp））です。

8　演習 1

1. ある大学の法学部（Law)、文学部（Literature)、理学部（Science)、工学部（Technology）の４学部から８名ずつの学生を無作為抽出してそれぞれテストを行った結果が以下で示されている。
   (1) これは対応ありですか？対応なしですか？
   (2) 学部間でテストの母平均に差があるかどうかを有意水準 5% で分散分析で評価しなさい。
   (3) 同じく Tukey の多重比較を実行し、どこに差があるかを評価しなさい。

Department								
Law	75	61	68	58	66	55	65	63
Literature	62	60	66	63	55	53	59	63
Science	65	60	78	52	59	66	73	64
Technology	52	59	44	67	47	53	58	49

2. 看護の教育方法として、講義型（L)、問題型（P)、実習型（E）の３つの形態による理解度の違いを調べた。学生７名（A∼G) を抽出し、全員がその３種類の教育を受けてテストの得点で比較することにした。
   (1) これは対応ありですか？対応なしですか？
   (2) テストに差があるかどうかを有意水準 5% で分散分析で評価しなさい。
   (3) 同じく Tukey の多重比較を実行し、どこに差があるかを評価しなさい。

Student	A	B	C	D	E	F	G
L	51	66	70	75	73	62	55
P	47	54	55	39	60	62	56
E	55	37	47	60	62	53	50