一元配置分散分析

渡邊直樹

2021 年 11 月 17 日

分散分析とは，2 群以上のグループに分けて取得されたデータが手許にあるとき,「各データとデータ全体の平均値の乖離」を「各群の平均とデータ全体の平均値の乖離」と「各データとそれが属する群の平均の乖離」の 3 つに分けて評価するデータ解析手法である．分散分析では，件数などの離散変数ではなく，主に連続変数がデータとして取り扱われるので，2 群間の比較を行う場合には t 検定またはウェルチの t 検定が適用されることが多い．実際，2 群間比較における分散分析はそれらの群の間で確率分布の等分散性を仮定した t 検定と本質的には同じである．本教材の目的は，統計手法の適用の際にしばしば出て来る**自由度**（degree of freedom）の概念に慣れること，それに基づいて，一元配置分散分析が実際に行っている F 検定が t 検定と本質的には同じものであることを理解すること，R 本体のコマンドの操作に慣れることである．

統計学における自由度とは「独立に値を決められる変数の数」のことであるが，それをカウントする際にはうっかり見過ごしてしまう事柄も事案に応じてしばしば発生する．F 検定と t 検定が依拠する F 分布と t 分布はともにカイ 2 乗分布から派生して定義された確率分布である．これらの関係をまずまとめておく．

平均 0，分散 1 の正規分布を標準正規分布という．標準正規分布に従う互いに独立な n 個の確率変数 X_i（$i = 1, \ldots, n$）の 2 乗和

$$Y = \sum_{i=1}^{n} = X_1^2 + \cdots + X_n^2$$

が従う確率分布をカイ 2 乗分布という．X_l は自由度 l のカイ 2 乗分布に従う確率変数であり，X_m は自由度 m のカイ 2 乗分布に従う確率変数で

あり，これらが互いに独立であるとすると，

$$F = \frac{X_l/l}{X_m/m}$$

が従う分布を自由度 (l, m) の F 分布という．ここで，自由度 $(1, m)$ の F 分布に従う F の正の平方根を

$$t = \frac{X_1^{1/2}}{(X_m/m)^{1/2}}$$

とすると，t が従う確率分布を自由度 m の t 分布という．このように，F の分子が自由度 1 のカイ 2 乗分布に従う場合には，分布の形状こそ違いこそすれ，実質的に F 検定と t 検定は同じものである．このことに注意して，以下で説明される一元配置分散分析を理解してほしい．

一元配置分散分析

ある薬剤の治験において，その投薬量に応じて A 群，B 群，C 群に分かれて，ある指標データが (1, 3)，(6, 7, 5)，(4, 2) と観察されたとする．（この例では，見た目を簡潔にするために，各群内で平均をとると割り切れるような値を設定してある．）データは全部で 7 つあるので，それらが独立な確率変数であれば，自由度は 7 である．（確率変数 X と Y が独立であるとは，平たくいうと，一方の生起（確率）は他方の生起確率に影響を与えないということである．）個別の被験者に薬剤が別々に投与された場合，観察されたデータは独立であると考えられる．各群内での平均は 2，6，3 であり，データ全体の平均は 4 である．ここで、元のデータであるベクトル $x = (1, 3, 6, 7, 5, 4, 2)$ から，$y = (2, 2, 6, 6, 6, 3, 3)$ と $z = (4, 4, 4, 4, 4, 4, 4)$ を作っておく．

ベクトル y では各群内で平均をとっており，その個数の分だけ y の各変数は制約を受けるので，ベクトル $x-y$ の自由度は $(2-1)+(3-1)+(2-1) = 4$ である．同様に，ベクトル $y-z$ の自由度は $(3-1) = 2$ である．x の各要素が互いに独立に標準正規分布に従う確率変数の実現値だとすると，$x - y$ の各要素の 2 乗和 SS_{within} は自由度 4 のカイ 2 乗分布に，$y - z$ の各要素の 2 乗和 $SS_{between}$ は自由度 2 のカイ 2 乗分布に従う（練習問題）．ここで，SS_{within} と $SS_{between}$ はそれぞれ群内平方和（sum of squares within groups），群間平方和（sum of squares between groups）という．

ベクトル $x-z$ は各データの全体の平均からの差を要素としており，その 2 乗和を SS_{total} ということにすると，

$$SS_{total} = SS_{within} + SS_{between}$$

という関係が得られる．ここで，x の各要素が互いに独立に標準正規分布に従う確率変数であるとき，統計量

$$F = \frac{SS_{between}/2}{SS_{within}/4}$$

は自由度 $(2, 4)$ の F 分布に従う．実際のデータでは $F = 7.333333$ であり，F が 7.333333 以上になる確率は 0.04591837 である．つまり，群内での変動に対する群間の変動が 7.333333 よりも大きくなる確率は約 4.6%である．具体的に言い換えると，ここでは，薬剤投与による指標の平均が A 群，B 群，C 群の間で差がないのであれば，F の分子は 0 となるはずであり，データから得られたその値が 7.333333 であるということは，F の真の値が 0 である確率を約 4.6%であると計算しているのである．よって，3 群間での平均に差がないという帰無仮説は有意水準 5%の下で棄却される．

以上のことを R のコマンドで確認しておこう．

```
> x=c(1, 3, 6, 7, 5, 4, 2)
> g=factor(c(1,1,2,2,2,3,3))
> y=ave(x,g)
> y
[1] 2 2 6 6 6 3 3
> z=ave(x)
> z
[1] 4 4 4 4 4 4 4
> x-y
[1] -1  1  0  1 -1  1 -1
> y-z
[1] -2 -2  2  2  2 -1 -1
> sum((x-y)^2)
[1] 6
> sum((y-z)^2)
[1] 22
> sum((x-z)^2)
```

```
[1] 28
> (sum((y-z)^2)/2)/(sum((x-y)^2)/4)
[1] 7.333333
> 1-pf(7.333333, 2, 4)
[1] 0.04591837
```

上記のデータ処理を簡単に実行するコマンドが関数 anova() であり，線形モデルを当てはめる関数 lm() を使って，次のように打ち込む．

```
> anova(lm(x~g))
Analysis of Variance Table

Response: x
          Df Sum Sq Mean Sq F value  Pr(>F)
g          2     22    11.0  7.3333 0.04592 *
Residuals  4      6     1.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1
```

この例では 3 群での分散分析を行ったが，2 群であれば，F の分子の自由度が 1 となり，自由度 4 の t 分布となっていることが判る．このように，一元配置というのは，薬剤の投与量のみが指標となる数値に影響を与えている因子としているからであり，指標の変動に複数の因子が影響を与えていると想定する場合には，そのデータの解析に多元配置分散分析を適用する．多元配置分散分析に関する説明は別稿に譲ることにする．

分散分析表を出力する必要がないのであれば，次のコマンドでもよい．var.equal=TRUE で群間での等分散性を指定している．oneway.test() はデフォルトでは等分散性を仮定しておらず，ウェルチの方法で分散分析を実行する．この例では，サンプルを極端に少なくしてあるので，等分散性を仮定せずに分析すると，帰無仮説を棄却できなくなっていることに注意してほしい．

```
> oneway.test(x~g, var.equal=TRUE)

One-way analysis of means
```

```
data:  x and g
F = 7.3333, num df = 2, denom df = 4, p-value = 0.04592

> oneway.test(x~g)

One-way analysis of means (not assuming equal variances)

data:  x and g
F = 5.6716, num df = 2.0000, denom df = 1.9608, p-value = 0.153
```

また，分散分析はデータの正規性も前提としているので，その適用前には正規性の検定を実行する必要がある．ここではその説明を省略するが，ベクトル x における各要素の正の平方根を逆サイン関数で変換した $asin(sqrt(x))$ は正規性を満たしやすく，かつ，等分散性も満たしやすいとされている．実務データは正規性と等分散性を満たさないものが少なくないので、この変換を施したものを利用することが多くなる．oneway.test() に相当するノンパラメトリック検定にクラスカル･ウォリス検定（Kruskal-Wallis test）があり，こちらを利用する機会も多い．

```
> kruskal.test(x~g)

Kruskal-Wallis rank sum test

data:  x by g
Kruskal-Wallis chi-squared = 4.7143, df = 2, p-value = 0.09469
```