第 4 章 実験計画、分散分析 experimental design and ANOVA(ANalysis Of VAriance)

浅川 伸一

2007 年 01 月 12 日改訂版

1 導入

実験などを行ってデータを得た場合、そのデータを分析することを考える。例えば実験条件が実験群と統制群という 2 群のデータを比較するだけであれば、t 検定を使って条件間の差が有意であるか否かを検討することができる。ここでは、やや複雑な条件を設定した実験データの解析を考える。

条件 1	条件 2	条件 3	
x_{11}	x_{12}	x_{13}	
x_{21}	x_{22}	x_{23}	
x_{31}	x_{32}	x_{33}	
x_{41}	x_{42}	x_{43}	
$\bar{x}_{\cdot 1}$	$\bar{x}_{\cdot 2}$	$\bar{x}_{\cdot 3}$	$\bar{x}_{\cdot\cdot}$

表 1: 1 要因 3 水準の実験データ

表 1 は実験条件が 3 つある（水準が 3 であると言うことがある）場合のデータ例を記号で示したものである。各条件で 4 つのデータが得られたとするとデータを示す行列は 4 行 3 列となる。表中の最下行には各条件の平均が $\bar{x}_{\cdot 1}, \bar{x}_{\cdot 2}, \bar{x}_{cdot3}$ と表現されている。全データを込にした平均、全平均が右下に $\bar{x}_{\cdot\cdot}$ と表記されている。記号中のドット・は「対応する要素について」というような意味合いで用いられる。例えば $\bar{x}_{\cdot 1}$ は一列目（行列の最初の添字は行を表し、二番目の添字は列を表す）について足し合わせてデータ数（この場合は 4) で割った数である。各条件のデータ数が 4 と等しいので $\sum_{j=1}^{3} \bar{x}_{\cdot j} = \bar{x}_{\cdot 1} + \bar{x}_{\cdot 2} + \bar{x}_{\cdot 3} = 3\bar{x}_{\cdot\cdot}$ が成り立つ。

また全平均は加算記号 $\sum$ を 2 つ重ねて $\sum_{i=1}^{4}\sum_{j=1}^{3} x_{ij} = 12\bar{x}_{\cdot\cdot}$ でもある。

ここで、調べたいことは各条件間に差があるか否か、である。換言すれば、帰無仮説 H: $\bar{x}_{\cdot 1} = \bar{x}_{\cdot 2} = \bar{x}_{\cdot 3}$ が棄却できるかどうかを検討することである。

簡単のためデータ個数が各群とも等しく m 個であったとし、水準数は k とすると全分散は

$$S_T^2 = \frac{1}{mk} \sum_{i=1}^{m} \sum_{j=1}^{k} (x_{ij} - \bar{x}_{\cdot\cdot})^2 \tag{1}$$

と表すことができる。

このとき、各データ x_{ij} は次のように表されるとする。

$$x_{ij} = \bar{x}_{\cdot\cdot} + (\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}) + (x_{ij} - \bar{x}_{\cdot j}) \tag{2}$$

右辺は正負の符号に注意すれば、x_{ij} しか残らず、左辺と一致することが判る。右辺第一項を左辺に移項し全てのデータに付いて足しあわせれば式 1 である。すなわち

$$\begin{aligned} S_T^2 &= \frac{1}{mk} \sum_{i=1}^{m} \sum_{j=1}^{k} (x_{ij} - \bar{x}_{\cdot\cdot})^2 & (3) \\ &= \frac{1}{mk} \sum_{i=1}^{m} \sum_{j=1}^{k} \{(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}) + (x_{ij} - \bar{x}_{\cdot j})\}^2 & (4) \\ &= \frac{1}{k} \sum_{j=1}^{k} (\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})^2 + \frac{1}{mk} \sum_{i=1}^{m} \sum_{j=1}^{k} (x_{ij} - \bar{x}_{\cdot j})^2 & (5) \end{aligned}$$

途中の計算は省略したが直上行の左辺第一項を級間分散 (あるいは Between 分散)、第二項を級内分散（あるいは Within 分散）とに分解できることを表してる。

級間分散が十分に大きい場合には帰無仮説が成り立たないことを意味し、従って実験条件間で有意差が認められると推論できる。

これが条件間の平均値の差を比較するために、分散を分析すると書く分散分析を用いる理由である。

2　線形モデルによる定式化

表 1 に表れるデータを一列のベクトルと表現してみる。

$$\begin{pmatrix} x_{11} \\ x_{21} \\ x_{31} \\ x_{41} \\ x_{12} \\ x_{22} \\ x_{32} \\ x_{42} \\ x_{13} \\ x_{23} \\ x_{33} \\ x_{43} \end{pmatrix} = \begin{pmatrix} \bar{x}.. \\ \bar{x}.. \\ \bar{x}.. \\ \bar{x}.. \\ \bar{x}.. \\ \bar{x}.. \\ \bar{x}.. \\ \bar{x}.. \\ \bar{x}.. \\ \bar{x}.. \\ \bar{x}.. \\ \bar{x}.. \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \epsilon_{31} \\ \epsilon_{41} \\ \epsilon_{12} \\ \epsilon_{22} \\ \epsilon_{32} \\ \epsilon_{42} \\ \epsilon_{13} \\ \epsilon_{23} \\ \epsilon_{33} \\ \epsilon_{43} \end{pmatrix} \tag{6}$$

$$= \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \theta_0 = \mu..(= \bar{x}..) \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \epsilon_{31} \\ \epsilon_{41} \\ \epsilon_{12} \\ \epsilon_{22} \\ \epsilon_{32} \\ \epsilon_{42} \\ \epsilon_{13} \\ \epsilon_{23} \\ \epsilon_{33} \\ \epsilon_{43} \end{pmatrix} \tag{7}$$

すなわち、データ行列を一つの列ベクトル $\boldsymbol{x} = (\ x_{11}\ , x_{21},\ x_{31},\ x_{41},\ x_{12},\ x_{22},\ x_{23},\ x_{24},\ x_{13},\ x_{23},\ x_{33},\ x_{43}\ )^T$ で表現し、全ての要素が 1 からなる $(1, 1, \cdots, 1)^T\theta_0$ 列ベクトル $(\theta_0 = \bar{x}..)$ と右辺第二項 0 と 1 とでできた行列を $\boldsymbol{A}$ と表すことにする。さらに右辺第三項のベクトルを $\boldsymbol{\epsilon}$ とすると上式は

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{8}$$

と表現されることになる。行列 $\boldsymbol{A}$ は計画行列と呼ばれることもある。$\boldsymbol{A}$ は実験を計画した段階で定まっている行列で、いわば実験のデザインを行列として表現したとみなせるからである。

$\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \theta_3)^T$ 偏回帰係数とみなせば、分散分析は重回帰分析と形式的に同じものであることがわかる。

ここで分散分析による平均値の差の検定は、帰無仮説 $\theta_1 = \theta_2 = \theta_3 = 0$ すなわち偏回帰係数が 0 とみなしうるか否かを検討することに等しい。

行列 $\boldsymbol{A}$ は 12 行 4 列の行列であるが、12 行 1 列の行列 $\boldsymbol{A}_0$ と 12 行 3 列の行列 $\boldsymbol{A}_1$ とを並べて書いたものである。

$$\boldsymbol{A} = [\boldsymbol{A}_0, \boldsymbol{A}_1] \tag{9}$$

同様に $\boldsymbol{\theta}$ も最初の要素 θ_0 とそれ以外とに分けられる。この場合 θ_0 を一行一列の列ベクトルとみなす。

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\epsilon} & (10) \\ &= \boldsymbol{A}_0\boldsymbol{\theta} + \boldsymbol{A}_1\boldsymbol{\theta}_1 + \boldsymbol{\epsilon} & (11) \end{aligned}$$

行列 $\boldsymbol{A}$ で張られる空間への射影行列 $\boldsymbol{Q}$ が定義できるものとする。$\boldsymbol{A}$ の部分空間である $\boldsymbol{A}_0$ への射影行列を $\boldsymbol{Q}_0$ とすれば、その補空間として $\boldsymbol{Q}_1$ が定義できる。$\boldsymbol{Q}_1 = \boldsymbol{Q} - \boldsymbol{Q}_0$

これらの射影行列を用いてデータベクトル $\boldsymbol{x}$ を射影することを考えれば、$\boldsymbol{x}$ を $\boldsymbol{Q}$ へ射影した $\boldsymbol{Q}\boldsymbol{x}$, $\boldsymbol{Q}_0$ へ射影した $\boldsymbol{Q}_0\boldsymbol{x}$, $\boldsymbol{A}$ で張られる空間へ射影したベクトルのうち、$\boldsymbol{A}_0$ で張られる空間と直交補空間の関係にある空間 $\boldsymbol{Q}_1$ にあるベクトル $(\boldsymbol{Q} - \boldsymbol{Q}_0)\boldsymbol{x}$ の空間関係は 図 1 のようになる。

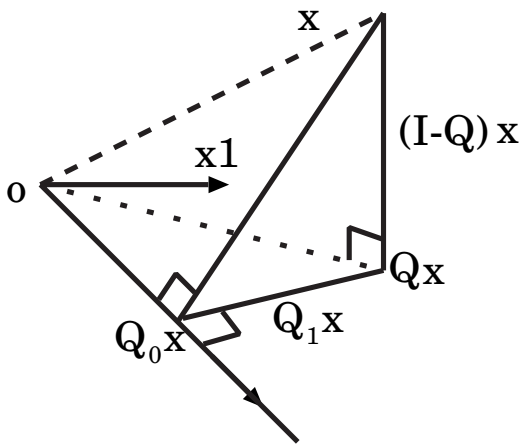

図 1: decomposition of a vector x

定義した射影行列を用いて、x の分散に対応する量であった x の長さ $\boldsymbol{x}^T\boldsymbol{x}$ を分解する。

$$\begin{aligned} \boldsymbol{x}^T\boldsymbol{x} &= \boldsymbol{x}^T\boldsymbol{Q}_0\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{Q}_1\boldsymbol{x} + \boldsymbol{x}^T(\boldsymbol{I} - \boldsymbol{Q})\boldsymbol{x} & (12) \\ &= S_0 + S_1 + S_e & (13) \end{aligned}$$

3　分散分析表

帰無仮説 H が正しいのであれば $\boldsymbol{x}^T\boldsymbol{Q}_1\boldsymbol{x}$ が短くなっているはずである。これは元データ $\boldsymbol{x}^T\boldsymbol{x}$ (全分散) と $(\boldsymbol{I}-\boldsymbol{Q})$(誤差分散) によって影響を受ける。従って次のような表を作る。

要因	平方和	自由度	平均平方	F
A	S_1	m-1	$S_1/(m-1)$	$\dfrac{S_1/(m-1)}{S_e/(mk-m)}$
誤差	S_e	mk - m	$S_e/(mk-m)$	
計	$\boldsymbol{x}^T\boldsymbol{x}$	mk - 1		

表 2: 1 要因の分散分析表

分散分析表に出てくる自由度の定義であるが、それぞれの射影ベクトル長さをその自由度、つまりその空間の次数で割る、というのが定義である。行列 $\boldsymbol{A}$ は 12 行 4 列の行列であるから、この行列から作られる射影行列は 4 次元空間を張ると見なせるかというとそうでは「ない」。すべての要素が 1 である $\boldsymbol{A}_0$ でまず一次元。$\boldsymbol{A}_1$ は 3 列であるが線形独立ではない。式 (7) の右辺第一項を見ると分かるとおり $\boldsymbol{A}_0$ ともう 2 つの列ベクトルが決まると最後の一つのベクトルは線形独立（他のベクトルの線形結合として表すことができる）ではない。従って S_1 の自由度は 2 である。$\boldsymbol{A}_0$ という全ての要素が 1 であるという決まった次元がある以上、$\boldsymbol{x}^T\boldsymbol{x}$ も mk 次元ではなく $mk-1$ 次元空間となる。全ての部分空間の次数の総和が mk 次元となるので、残った S_e の次元は $mk-m$ 次元でなければならない。以上が自由度の考え方である。前期心理統計学１ではデータ数マイナス１が自由度という説明を試みたが、今回の説明の仕方の方が一般性がある。

分散を自由度で除したものは χ^2 分布に従い、その比の分布は F 分布に従うという前期の知識に従えば、この F 分布が有意水準の値を越えていれば帰無仮説を棄却し対立仮説を採択することになる。

Excel で出力される情報は上記で全て説明できる。

射影行列は回帰方程式を解くことによって、以下のように

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{\theta} \tag{14}$$

$$\boldsymbol{A}^T\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta} \tag{15}$$

$$\left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{x} = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta} \tag{16}$$

$$\left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{x} = \boldsymbol{\theta} \tag{17}$$

であることを思い出すこと。式 (17) 左辺を式 (14) の $\boldsymbol{\theta}$ に代入することによって射影行列 $\boldsymbol{Q} = \boldsymbol{A}\left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T$ を得る。手続きは他の射影行列も同様で

ある。ただし $\boldsymbol{A}$ は自由度の説明で述べたとおり行列の次数（ランク rank ともいう）が落ちているので一意に求めることができない点には注意。

4　多重比較

さて、以上のようにして分散分析を行ったとする。結果 F の値が基準を越えて大きく、有意だと認められた場合に、具体的にどの条件間の差異が有意だったのかを検討することが行われる。

2 水準の場合には t 検定でも良かったが 3 水準以上の場合には、t 検定を繰り返し行うと危険率 95 % であっても多数回繰り返すことによって第一種の誤りの可能性が高くなってしまうからである。

多重比較の方法として Sheffe's method(シェッフェの方法)、Turkey's method(チューキーの方法)、Bonferroni の方法などが用いられる。

いずれの方法でも 2 群の平均の比較方法よりもより厳しい基準を設けて危険率の修正を行うことになる。

例えば Scheffe の方法では F の替りに F の値を $m-1$ (水準の数から 1 を引いたもの、すなわち計画行列の次数) 倍して開平した値を越えるか否かで判断する。すなわち自由度 (m-1, m (k -1)) の時の F 分布の 5 % の時の値を分布表から求めるなり、統計パッケージに計算させるなりして値を求めて、その値を越えるかどうかを検討する。

例えば 水準数が k 個あったとき、c_j を任意の定数として

$$\sum_{j=1}^{k} c_j = 0 \tag{18}$$

でかつ、

$$\sum_{j=1}^{k} c_j \mu_j = \mu \tag{19}$$

を検定する場合のことを考える。ただし μ_j は $\bar{x}_{\cdot j}$ の推定値。$\bar{x}_{\cdot 1} = 1, \bar{x}_{\cdot 2} = -1, \bar{x}_{\cdot 3} = 0$ であれば $\mu_1 = \mu_2$ を検定することになる。

Scheffe の方法では任意の c_j に対して

$$\frac{\sum_{j=1}^{k} c_j \bar{x}_j}{\sqrt{\sum_{j=1}^{k} \left(\frac{c_j^2}{m_j}\right) \hat{\sigma}_e^2}} \geq \sqrt{(k-1) F_{k-1,m(k-1)}(\alpha)} \tag{20}$$

が成り立つならば帰無仮説を棄却する。