2016 Singularity math. private school #3

浅川 伸一 asakawa@ieee.org

1 単回帰分析の拡張

二つの独立変数 independent variable(説明変数、予測変数とも言う) x_1, x_2 から、従属変数 dependent variable (目的変数とも言う) y を予測することを考える。

個々の独立変数 x_1, x_2 と従属変数 y_i との間に $y_i = b_0 + b_1 x_1 + b_1 + e_i$ という (線形な) 関係があると仮定されるときパラメータ $b_0, b_1,$ および b_2 を推定する問題を考える [1]。

1.1 説明変数が 2 つの場合

この場合、相関係数が 3 つある。それぞれの独立変数 x_1, x_2 と、従属変数 y との相関係数を $r_{x_1 y}, r_{x_2 y}$ と表す。独立変数 x_1 と x_2 との相関係数を $r_{x_1 x_2}$ とする。

x_1, x_2 それぞれ単独で y を予測するよりも、x_1 と x_2 との両者を使って従属変数 y を予測した方が予測の精度が上がると期待される。この他に x_1, x_2 の両者と y との相関係数を r_{yx1x2} あるいは r_{y12} と表記する。単回帰の場合にならって次の用語を用いる。

重回帰係数 multiple regression coefficient x_1 と x_2 の合成得点から y を予測するときの回帰係数

重相関係数 multiple correlation coefficient x_1 と x_2 の合成得点と y との相関係数

重相関係数の定義は次のとおり。

$$r_{y\cdot 12} = \sqrt{\frac{r_{y1}^2 + r_{y2}^2 - 2 r_{y1} r_{y2} r_{12}}{1 - r_{12}^2}} \tag{1}$$

この他に、3 者関係に独特な概念がある。x_1 から x_2 の影響を取り除いた成分 $x_1|x_2$ と y との相関係数 $r_{y(1|2)}$ あるいは $r_{y(x_1|x_2)}$ を

部分相関係数 semipartial correlation coefficient[2] と呼ぶ。$r_{y(x_1|x_2)}$ のように下付き添字に下付き添字がついて小さくなりすぎる場合には省略して $r_{y(1|2)}$ と表記することがある。

部分相関係数 $r_{y(1|2)}$ では x_1 から x_2 の影響を取り除いてあるが、目的変数 y については x_2 の影響を取り除いていない。y からも x_2 の影響を取り除き x_1 からも x_2 の影響を取り除いた上で相関係数を求めることができれば x_2 という変数に影響されない x_1 と y との関係が現われることになる。これを

偏相関係数 partial correlation coefficient と呼び $r_{y1|2}$ と表す。

$$r_{y1|2} = \frac{r_{y1} - r_{y1} r_{12}}{\sqrt{1 - r_{y2}^2}\sqrt{1 - r_{12}^2}} \tag{2}$$

偏相関係数 $r_{y1|2}$ と部分相関係数 $r_{y(1|2)}$ との間には

$$r_{y1|2} = \frac{r_{y(1|2)}}{\sqrt{1 - r_{y2}^2}} \tag{3}$$

という関係がある。

[1] 独立変数が一つときは、中学数学との対応から $y = ax + b$ と記していたが、この例ではパラメータを a, b から b に変更していることに注意

[2] part correlation coefficient とも

次に相関係数の算出でなく、回帰について考える。単回帰の場合の回帰係数は

$$b = r\frac{s_y}{s_x} \tag{4}$$

であった。誤差 (ベクトル、あるいは残差ベクトル) と予測ベクトルとは直交し、長さの 2 乗は三平方の定理から $s_y^2 = s_e^2 + s_{\hat{y}}^2$ が成り立つことに注意すれば、$s_e^2 = s_y^2\left(1 - r^2\right)$ となる。この式は $\boldsymbol{y}$ から $\boldsymbol{x}$ の影響を取り去った後の分散 (距離の 2 乗) は、元の分散に 1 から相関係数の 2 乗を引いた値の開平をかけたものである、と言い表せる。相関係数の 2 乗のことを**決定係数** coefficient of determination と呼び、従属変数の全分散のうち、独立変数で説明できる割合を表す量となる (前章のハンドアウトも参照のこと)。

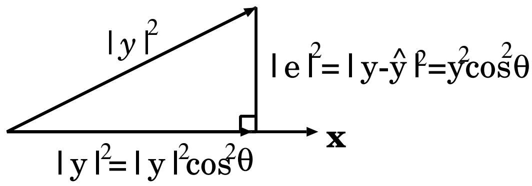

Figure 1: Pythagorean theorem

このことから、

$$s_{1|2} = s_1\sqrt{1 - r_{12}^2} \tag{5}$$

である。この値を回帰式に代入すると

$$b_{y(1|2)} = r_{y(1|2)}\frac{s_y}{s_{1|2}} \tag{6}$$

$$= \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{1 - r_{12}^2}}\,\frac{s_y}{s_2\sqrt{1 - r_{12}^2}} \tag{7}$$

$$= \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\,\frac{s_y}{s_1} \tag{8}$$

となる。これを**偏回帰係数** partial correlation coefficient と呼ぶ。

単回帰係数の場合は問題にならないことが多いが、回帰係数 $r_{xy}\frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2}$ (前回のハンドアウト (14) 式) は標準偏差の単位に依存して大きくなったり小さくなったりするので、あらかじめ各変数を分散 1 となるように変換しておいてから回帰係数を求めることがある。この場合回帰係数は $\frac{s_{xy}}{s_x^2}$ で相関係数に一致する。同様に重回帰において、あらかじめ各変数を基準化しておけば偏回帰係数は

$$b_{y(1|2)^*} = \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\,\frac{s_y}{s_1} \tag{9}$$

$$= \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2} \tag{10}$$

となる。これを**標準化偏回帰係数** standard partial regression coefficient と呼ぶ。

1.2 解析的な解

y_i の値の予測値 $\hat{y}_i$ を x_1 と x_2 を用いて $\hat{y}_i = b_0 + b_1x_1 + b_1$ としたとき、i 番目のデータとその予測値との差 (誤差) を

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1x_{1i} + b_2x_{2i}) \tag{11}$$

と表記する。n 組のデータが得られたとき全ての誤差の自乗和を最小にするパラメータの値を推定することを考える。誤差の自乗和を $Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ とし、これを各パラメータについて微分して 0 とおくことで以下の式を得る。

$$\begin{pmatrix} \frac{\partial Q}{\partial b_0} \\ \frac{\partial Q}{\partial b_1} \\ \frac{\partial Q}{\partial b_2} \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial b_0}\sum_{i=1}^n (y - \hat{y}_i)^2 \\ \frac{\partial}{\partial b_1}\sum_{i=1}^n (y - \hat{y}_i)^2 \\ \frac{\partial}{\partial b_2}\sum_{i=1}^n (y - \hat{y}_i)^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{12}$$

推定すべき未知パラメータの数と方程式の数が同じ 3 つであるので、この連立方程式は特別な場合を除いて解くことができる。これを解くと以下の解が得られる。

$$b_0 = \bar{y} - (b_1\bar{x_1} + b_2\bar{x_2}) \tag{13}$$

$$b_1 = \frac{r_{x_1y} - r_{x_2y}r_{x_1x_2}}{1 - r^2_{x_1x_2}}\frac{s_y}{s_{x_1}} \tag{14}$$

$$b_2 = \frac{r_{x_2y} - r_{x_1y}r_{x_1x_2}}{1 - r^2_{x_1x_2}}\frac{s_y}{s_{x_2}} \tag{15}$$

ここで r_{x_1y} は独立変数 x_1 と従属変数 y との相関係数を表す。同様に $r_{x_1x_2}$ は 2 つの独立変数間の相関係数を、s_{x_1} は x_1 の標準偏差を表すものとする。

2　グラフによる表現

各変数からそれぞれの平均を引いた値を要素とするベクトル (平均偏差ベクトル) を考えれば、2 つの変数間の相関係数は平均偏差ベクトル間のなす角の余弦 cos であった。今 3 つのベクトル y, x_1, x_2 を考える

2 つの独立変数と説明変数とからそれぞれの平均を引いた平均偏差ベクトルをそれぞれ $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}$ とする。

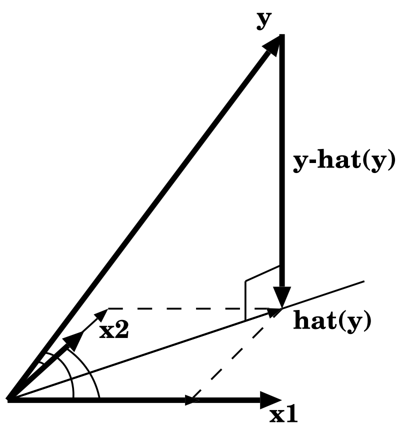

Figure 2: 2 つのベクトル x_1 と x_2 とによって表現される 2 次元空間へベクトル y を射影したときの模式図

$\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ とによって一つの空間が定義できる。$\boldsymbol{y}$ がこの空間内にあれば任意の定数 a と b とによって $\boldsymbol{y} = a\boldsymbol{x}_1 + b\boldsymbol{x}_2$ と表現できる。3 つのベクトルが同一空間 (この場合同一平面) 上にあるときにはこのことが成り立つが、一般に 3 つのベクトルは同一空間内にあるとは限らない。

そこで、ベクトル $\boldsymbol{y}$ をベクトル $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ とによって定義される空間へ射影し、この射影してできたベクトルを $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ とによって記述することを考える。射影ベクトルを $\hat{\boldsymbol{y}}$ とすれば、$\boldsymbol{y}$ から射影ベクトルを引いた残差ベクトル $\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ と残差ベクトルは直交する $\hat{\boldsymbol{y}} \perp \boldsymbol{r}$。

説明変数 $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ とから目的変数 $\boldsymbol{y}$ を予測する重回帰分析とは、$\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ とによって張られる空間へ $\boldsymbol{y}$ を射影し、できたベクトル $\boldsymbol{y}$ を $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ に分解すること、およびそのときの $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ にかかる係数 b_1, b_2 を求めることである。

3　ベクトルと行列による表現

従属変数 y_i $(i = 1, 2, \cdots, n)$ データをベクトル $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$ とし、1 を n 個並べたベクトル $\mathbf{1}'$ と 2 つの独立変数ベクトル $\boldsymbol{x}_1, \boldsymbol{x}_2$ を合わせて n 行 3 列の行列を $\boldsymbol{X} = \left(\mathbf{1}', \mathbf{1}, \boldsymbol{x}'_1, \boldsymbol{x}'_2\right)$ とする。このようにすると (11) は次のように書く

ことができる。ただし $\mathbf{1} = \underbrace{(1, 1, \cdots, 1)}_{n\text{ 個}}'$ である。

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{b} && (16)\\ \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} &= \begin{pmatrix} 1, & x_{11}, & x_{12} \\ 1, & x_{21}, & x_{22} \\ \vdots & \vdots & \vdots \\ 1, & x_{n1}, & x_{n2} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} && (17) \end{aligned}$$

代数の操作からの類推を使ってこれを形式的に解くと、

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{b} && (18)\\ \boldsymbol{X}'\boldsymbol{y} &= \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} && (19)\\ (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} && (20)\\ (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} &= \boldsymbol{b} && (21) \end{aligned}$$

求めるべき回帰係数と定数でできたベクトルは、従属変数ベクトルと定数からできた行列と目的変数ベクトルとを使って代数の演算のように解くことができる。(21) の左辺が 3 行 1 列のベクトルとなることを確認せよ。

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

$$S_T^2 = \frac{1}{mk}\sum_{i=1}^{m}\sum_{j=1}^{k}\left(x_{ij} - \bar{x}_{\cdot\cdot}\right)^2 \qquad (22)$$

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。

$$x_{ij} = \bar{x}_{\cdot\cdot} + \left(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}\right) + \left(x_{ij} - \bar{x}_{\cdot j}\right) \qquad (23)$$

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 22 である。すなわち

$$S_T^2 = \frac{1}{mk}\sum_{i=1}^{m}\sum_{j=1}^{k}\left(x_{ij}-\bar{x}_{..}\right)^2 \tag{24}$$

$$= \frac{1}{mk}\sum_{i=1}^{m}\sum_{j=1}^{k}\left\{\left(\bar{x}_{.j}-\bar{x}_{..}\right)+\left(x_{ij}-\bar{x}_{.j}\right)\right\}^2 \tag{25}$$

$$= \frac{1}{k}\sum_{j=1}^{k}\left(\bar{x}_{.j}-\bar{x}_{..}\right)^2+\frac{1}{mk}\sum_{i=1}^{m}\sum_{j=1}^{k}\left(x_{ij}-\bar{x}_{.j}\right)^2 \tag{26}$$

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。

4 線形モデルによる定式化

表 2 に表れるデータを一列のベクトルと表現してみる。

$$\begin{pmatrix} x_{11}\\ x_{21}\\ x_{31}\\ x_{41}\\ x_{12}\\ x_{22}\\ x_{32}\\ x_{42}\\ x_{13}\\ x_{23}\\ x_{33}\\ x_{43} \end{pmatrix} = \begin{pmatrix} \bar{x}_{..}\\ \bar{x}_{..}\\ \bar{x}_{..}\\ \bar{x}_{..}\\ \bar{x}_{..}\\ \bar{x}_{..}\\ \bar{x}_{..}\\ \bar{x}_{..}\\ \bar{x}_{..}\\ \bar{x}_{..}\\ \bar{x}_{..}\\ \bar{x}_{..} \end{pmatrix} + \begin{pmatrix} 1&0&0\\ 1&0&0\\ 1&0&0\\ 1&0&0\\ 0&1&0\\ 0&1&0\\ 0&1&0\\ 0&1&0\\ 0&0&1\\ 0&0&1\\ 0&0&1\\ 0&0&1 \end{pmatrix} \begin{pmatrix} \theta_1\\ \theta_2\\ \theta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_{11}\\ \epsilon_{21}\\ \epsilon_{31}\\ \epsilon_{41}\\ \epsilon_{12}\\ \epsilon_{22}\\ \epsilon_{32}\\ \epsilon_{42}\\ \epsilon_{13}\\ \epsilon_{23}\\ \epsilon_{33}\\ \epsilon_{43} \end{pmatrix} \tag{27}$$

$$= \begin{pmatrix} 1&1&0&0\\ 1&1&0&0\\ 1&1&0&0\\ 1&1&0&0\\ 1&0&1&0\\ 1&0&1&0\\ 1&0&1&0\\ 1&0&1&0\\ 1&0&0&1\\ 1&0&0&1\\ 1&0&0&1\\ 1&0&0&1 \end{pmatrix} \begin{pmatrix} \theta_0 = \mu_{..}(=\bar{x}_{..})\\ \theta_1\\ \theta_2\\ \theta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_{11}\\ \epsilon_{21}\\ \epsilon_{31}\\ \epsilon_{41}\\ \epsilon_{12}\\ \epsilon_{22}\\ \epsilon_{32}\\ \epsilon_{42}\\ \epsilon_{13}\\ \epsilon_{23}\\ \epsilon_{33}\\ \epsilon_{43} \end{pmatrix} \tag{28}$$

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

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{29}$$

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。

$$\boldsymbol{A} = [\boldsymbol{A}_0, \boldsymbol{A}_1] \tag{30}$$

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。

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{A\theta} + \boldsymbol{\epsilon} && (31) \\ &= \boldsymbol{A}_0\boldsymbol{\theta} + \boldsymbol{A}_1\boldsymbol{\theta}_1 + \boldsymbol{\epsilon} && (32) \end{aligned}$$

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 $\boldsymbol{Qx}$, $\boldsymbol{Q}_0$ 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

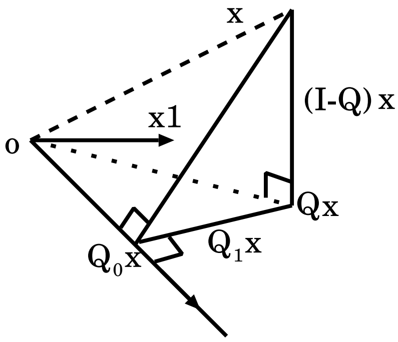

Figure 3: decomposition of a vector x

図 3 のようになる。

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。

$$\begin{aligned} \boldsymbol{x}^T\boldsymbol{x} &= \boldsymbol{x}^T\boldsymbol{Q}_0\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{Q}_1\boldsymbol{x} + \boldsymbol{x}^T(\boldsymbol{I} - \boldsymbol{Q})\boldsymbol{x} && (33) \\ &= S_0 + S_1 + S_e && (34) \end{aligned}$$

5　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 (28) 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

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{\theta} \tag{35}$$

$$\boldsymbol{A}^T\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta} \tag{36}$$

$$\left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{x} = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta} \tag{37}$$

$$\left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{x} = \boldsymbol{\theta} \tag{38}$$

であることを思い出すこと。式 (38) 左辺を式 (35) 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。

6　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

$$\sum_{j=1}^{k} c_j = 0 \tag{39}$$

でかつ、

$$\sum_{j=1}^{k} c_j\mu_j = \mu \tag{40}$$

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

$$\frac{\sum_{j=1}^{k} c_j\bar{x}_j}{\sqrt{\sum_{j=1}^{k}\left(\frac{c_j^2}{m_j}\right)\hat{\sigma}_e^2}} \geq \sqrt{(k-1)F_{k-1,m(k-1)}(\alpha)} \tag{41}$$

が成り立つならば帰無仮説を棄却する。

7 多重比較の実際

データ例として向後先生のハンバーガーショップを用いる
http://kogolab.jp/elearn/hamburger/chap6/sec3.html

	わくわく	もぐもぐ	ぱくぱく	
	80	75	80	
	75	70	80	
	80	80	80	
	90	85	90	
	95	90	95	
	80	75	85	
	80	85	95	
	85	80	90	
	85	80	85	
	80	75	90	
	90	80	95	
	80	75	85	
	75	70	98	
	90	85	95	
	85	80	85	
	85	75	85	
	90	80	90	
	90	80	90	
	85	90	85	
	80	80	85	
データ数	20	20	20	60
群平均	84.00	79.50	88.15	83.88
分散	29.0	29.75	28.53	41.57

分散分析表は表 3 のとおり。

変動要因	変動	自由度	分散	F	p
要因	748.63	2	374.37	12.22	0.0001
誤差	1745.55	57	30.62		
計	2494.19	59			

表 3: 分散分析表

であるから、ハンバーガーショップによる違いは有意である。

このデータによれば、(一袋あたりのポテトの本数) はもぐもぐ店が最も少く、次にわくわく店である。そこで帰無仮説 H0: $\bar{x}_1 = \bar{x}_2$ を scheffe の方法によって検討してみる。$c_1 = 1, c_2 = -1, c_3 = 0$ とおくと scheffe の式、左辺は

$$\frac{|c_1\bar{x}_1 + c_2\bar{x}_2 + c_3\bar{x}_3|}{\sqrt{\left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \frac{c_3^2}{n_3}\right)\hat{\sigma}_e^2}} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\left(\frac{1^2}{20} + \frac{(-1)^2}{20} + \frac{0^2}{20}\right)30.62}} \tag{42}$$

$$= \frac{|84 - 79.5|}{1.75} \tag{43}$$

$$= 2.571 \tag{44}$$

である。一方右辺は

$$
\begin{aligned}
\sqrt{(k-1)F_{k-1,m(k-1)}(\alpha)} &= \sqrt{(3-1)F_{2,3(20-1)}(0.05)} && (45)\\
&= \sqrt{2\times \mathrm{Finv}(0.05,2,57)} && (46)\\
&= 2.51 && (47)
\end{aligned}
$$

であるから、右辺の 5 % の有意水準の値よりも左辺の方が大きい。したがって、もぐもぐ店とわくわく店には違いがあると言える。

同様の計算をモグモグ店とわくわく店で比較するには $c_1 = 1, c_2 = 0, c_3 = -1$ として計算すれば良い。右辺の値は変らず、左辺の分母の値も変らないので左辺分子の値のみ入れ替えて $|84.0 - 88.15| = 4.15$ であるから右辺は $4.15/2.571 = 2.37$ となって 5 % の有意水準を示す右辺の値 2.51 よりも小さくなる。このことから、モグモグ店とわくわく店の間には有意な差が認められるとは言えないことになる。

同じようにして例えば 5 水準の分散分析について $\bar{x}_1$ と $\bar{x}_3$ についてだけ検討したければ、$c_1 = 1, c_2 = 0, c_3 = -1, c_4 = 0, c_5 = 0$ と置いて計算すれば良い。条件 1 と条件 2 が等しく、それらは条件 3 と条件 4 とは異なるという仮説を検討したければ $c_1 = 1, c_2 = 1, c_3 = -1, c_4 = -1, c_5 = 0$ と置いて計算すれば良い。

8 一因子の実験計画の復習

線形モデルによる実験計画とは以下のようにまとめられた。以下のデータに対して、

条件 1	条件 2	$\cdots$	条件 k	
x_{11}	x_{12}	$\cdots$	x_{1k}	
x_{21}	x_{22}	$\cdots$	x_{2k}	
$\vdots$	$\vdots$	$\cdots$	$\vdots$	
$x_{n_1 1}$	$x_{n_2 2}$	$\cdots$	$x_{n_k k}$	
$\bar{x}_{\cdot 1}$	$\bar{x}_{\cdot 2}$	$\cdots$	$\bar{x}_{\cdot k}$	$\bar{x}_{\cdot\cdot}$

表 4: 1 要因 k 水準の実験データ

$$
\begin{pmatrix} x_{11}\\ x_{21}\\ \vdots\\ x_{n_1 1}\\ x_{12}\\ x_{22}\\ \vdots\\ x_{n_2 2}\\ \vdots\\ x_{1k}\\ x_{2k}\\ \vdots\\ x_{n_k k} \end{pmatrix}
=
\begin{pmatrix}
1 & 1 & 0 & \cdots & 0\\
1 & 1 & 0 & \cdots & 0\\
\vdots & \vdots & \cdots & \cdots & \vdots\\
1 & 1 & 0 & & \vdots\\
1 & 0 & 1 & & \vdots\\
1 & 0 & 1 & & \vdots\\
\vdots & \vdots & \cdots & \cdots & \vdots\\
1 & 0 & 1 & & \vdots\\
\vdots & \vdots & \cdots & \cdots & \vdots\\
1 & 0 & 0 & \vdots & 1\\
1 & 0 & 0 & \vdots & 1\\
\vdots & \vdots & \cdots & \cdots & \vdots\\
1 & 0 & 0 & \cdots & 1
\end{pmatrix}
\begin{pmatrix} \theta_0\\ \theta_1\\ \theta_2\\ \vdots\\ \theta_k \end{pmatrix}
+
\begin{pmatrix} \epsilon_{11}\\ \epsilon_{21}\\ \vdots\\ \epsilon_{n_1 1}\\ \epsilon_{12}\\ \epsilon_{22}\\ \vdots\\ \epsilon_{n_2 2}\\ \vdots\\ \epsilon_{1k}\\ \epsilon_{2k}\\ \vdots\\ \epsilon_{n_k k} \end{pmatrix}
\qquad (48)
$$

これをデータベクトル $\boldsymbol{x}$, 計画行列 $\boldsymbol{A} = [\boldsymbol{A}_0, \boldsymbol{A}_1]$, 平均ベクトル $\boldsymbol{\theta}$, および誤差ベクトル $\boldsymbol{\epsilon}$ を用いて

$$
\boldsymbol{x} = \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\epsilon} \qquad (49)
$$

と書くことができた。このとき計画行列のランクは rann$\boldsymbol{A} = k$ なので

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{pmatrix} N & n_1 & n_2 & \cdots & n_k \\ n_1 & n_1 & 0 & \cdots & 0 \\ n_2 & 0 & n_2 & & 0 \\ \vdots & \vdots & 0 & \cdots & \vdots \\ n_k & 0 & 0 & \cdots & n_k \end{pmatrix} \tag{50}$$

の逆行列は存在しない。

一つの解法としては $\boldsymbol{A}_0$ で張られる空間の補空間を考えてその上で直交補空間への分割を考えれば良い。

すなわち、データ行列を一つの列ベクトル $\boldsymbol{x} = (\, x_{11}\ ,x_{21}, \cdots ,x_{n_1 1}, x_{12}, x_{22}, \cdots, x_{n_2 2}, \cdots, x_{1k}, x_{2k}, \cdots, x_{n_k k}\,)^T$ で表現し、全ての要素が 1 からなる $(1, 1, \cdots, 1)^T\theta_0$ 列ベクトルで張られる空間への射影行列を用いて射影し、その直交補空間への射影行列

$$\boldsymbol{I} - \boldsymbol{Q}_0 = \boldsymbol{I} - A_0\left(\boldsymbol{A}_0^T\boldsymbol{A}_0\right)^{-1}\boldsymbol{A}_0^T \tag{51}$$

を用いて射影し、その射影されたベクトルをさらに直和分解すれば良かった。ここで $\boldsymbol{I}$ は単位行列を表す。

9　2 要因の場合

測定すべき実験変数が複数の場合を考える。

9.1　繰り返しのない場合

各条件で一度しか測定が行われない場合である。

	条件 A1	条件 A2	条件 A3	
条件 B1	x_{11}	x_{12}	x_{13}	$\bar{x}_{1\cdot}$
条件 B2	x_{21}	x_{22}	x_{23}	$\bar{x}_{2\cdot}$
条件 B3	x_{31}	x_{32}	x_{33}	$\bar{x}_{3\cdot}$
条件 B4	x_{41}	x_{42}	x_{43}	$\bar{x}_{4\cdot}$
	$\bar{x}_{\cdot 1}$	$\bar{x}_{\cdot 2}$	$\bar{x}_{\cdot 3}$	$\bar{x}_{\cdot\cdot}$

表 5: 2 元配置の分散分析に用いられるデータ

表 5 では列方向に条件 A (3 水準)、行方向に 条件 B (4 水準) が表現されている。実際には各条件毎に一つしかデータが得られないことは稀である。しかし、例えば、条件 B を繰り返しの順序と考えれば実験の最初に得られた条件 B1 と最後に得られた B4 とを比較することで、練習効果、疲労効果などを考慮にいれた分析をすることになり、しばしば用いられる手法である。

この時得られたデータは次のモデルに従うと考える

$$x_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \tag{52}$$

すなわち、データは全平均の推定値と条件 A の効果と条件 B の効果と誤差とで表される。このときの線形モデルは

$$\begin{pmatrix} x_{11} \\ x_{21} \\ x_{31} \\ x_{41} \\ x_{12} \\ x_{22} \\ x_{32} \\ x_{42} \\ x_{13} \\ x_{23} \\ x_{33} \\ x_{43} \end{pmatrix} = \begin{pmatrix} 1 & 1 & & & 1 & & & \\ 1 & 1 & & & & 1 & & \\ 1 & 1 & & & & & 1 & \\ 1 & 1 & & & & & & 1 \\ 1 & & 1 & & 1 & & & \\ 1 & & 1 & & & 1 & & \\ 1 & & 1 & & & & 1 & \\ 1 & & 1 & & & & & 1 \\ 1 & & & 1 & 1 & & & \\ 1 & & & 1 & & 1 & & \\ 1 & & & 1 & & & 1 & \\ 1 & & & 1 & & & & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \epsilon_{31} \\ \epsilon_{41} \\ \epsilon_{12} \\ \epsilon_{22} \\ \epsilon_{32} \\ \epsilon_{42} \\ \epsilon_{13} \\ \epsilon_{23} \\ \epsilon_{33} \\ \epsilon_{43} \end{pmatrix} \tag{53}$$

式中、要素が書かれていない個所は 0 である。これは

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{54}$$

と書くことができて、形式的には一要因（一元配置）の分散分析と全く同じ式である。

$$\boldsymbol{A}^T\boldsymbol{x} = \begin{pmatrix} 12\bar{x}_{..} \\ 4\bar{x}_{.1} \\ 4\bar{x}_{.2} \\ 4\bar{x}_{.3} \\ 3\bar{x}_{1.} \\ 3\bar{x}_{2.} \\ 3\bar{x}_{3.} \\ 3\bar{x}_{4.} \end{pmatrix} \tag{55}$$

である。また、

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{pmatrix} 12 & 4 & 4 & 4 & 3 & 3 & 3 & 3 \\ 4 & 4 & & & 1 & 1 & 1 & 1 \\ 4 & & 4 & & 1 & 1 & 1 & 1 \\ 4 & & & 4 & 1 & 1 & 1 & 1 \\ 3 & 1 & 1 & 1 & 3 & & & \\ 3 & 1 & 1 & 1 & & 3 & & \\ 3 & 1 & 1 & 1 & & & 3 & \\ 3 & 1 & 1 & 1 & & & & 3 \end{pmatrix} \tag{56}$$

統計的検討をする場合には帰無仮説が二つあって $H1 : \alpha_1 = \alpha_2 = \alpha_3 = 0$, $H2 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ である。

式 (53), (54) の計画行列 $\boldsymbol{A}$ の中で第一列を $\boldsymbol{A}_0$, 第 2 列から 4 列までを $\boldsymbol{A}_1$, 第 5 列から第 8 列までを $\boldsymbol{A}_2$ とする。

$$\boldsymbol{A}_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \tag{57}$$

$$
A_1 = \begin{pmatrix} 1 & & \\ 1 & & \\ 1 & & \\ 1 & & \\ & 1 & \\ & 1 & \\ & 1 & \\ & 1 & \\ & & 1 \\ & & 1 \\ & & 1 \\ & & 1 \end{pmatrix} \tag{58}
$$

$$
A_2 = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \\ 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \\ 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix} \tag{59}
$$

さらに $\boldsymbol{\alpha} = (\theta_1, \theta_2, \theta_3)^T, \boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^T$, とおくとデータは

$$
\boldsymbol{x} = \boldsymbol{A}_0\mu + \boldsymbol{A}_1\boldsymbol{\alpha} + \boldsymbol{A}_2\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{60}
$$

と表現できる。ここで μ は全平均を表す。

一要因の分散分析で行ったように、全分散 (の総データ数倍) は

$$
12S_T^2 = \sum\sum\left(x_{ij} - \bar{x}_{\cdot\cdot}\right)^2 \tag{61}
$$

で表されるが、これを変形し

$$
\begin{aligned}
12S_T^2 &= \sum\sum\left(x_{ij} - \bar{x}_{\cdot\cdot}\right)^2 && (62)\\
&= \sum\sum\left(x_{ij} - \bar{x}_{\cdot j} - \bar{x}_{i\cdot} + \bar{x}_{\cdot\cdot} + \bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot} + \bar{x}_{i\cdot} + \bar{x}_{\cdot\cdot}\right)^2 && (63)\\
&= \sum\sum\left\{\left(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}\right) + (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})\right. && (64)\\
&\quad \left.+\left(\bar{x}_{ij} - \bar{x}_{\cdot j} - \bar{x}_{i\cdot} + \bar{x}_{\cdot\cdot}\right)\right\}^2 \\
&= 4\sum\left(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}\right)^2 + 3\sum(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2 && (65)\\
&\quad + \sum\sum\left(x_{ij} - \bar{x}_{\cdot j} - \bar{x}_{i\cdot} + \bar{x}_{\cdot\cdot}\right)^2 \\
&= S_0^2 + S_1^2 + S_2^2 + S_e^2 && (66)
\end{aligned}
$$

従って上記のごとき分散の分解によって次の分散分析表を得る。

要因	平方和	自由度	平均平方	F
A	S_1	m-1	$S_1/(m-1)$	$\dfrac{S_1/(m-1)}{S_e/(m-1)(k-1)}$
B	S_2	k-1	$S_2/(k-1)$	$\dfrac{S_2/(k-1)}{S_e/(m-1)(k-1)}$
誤差	S_e	$(m-1)(k-1)$	$S_e/(m-1)(k-1)$	
計	$\boldsymbol{x}^T\boldsymbol{x}$	mk - 1		

表 6: 繰り返しのない 2 要因の分散分析表

線形モデルに即して考えれば、計画行列 $\boldsymbol{A}$ の部分行列によって構成される空間 $\boldsymbol{A}_0, \boldsymbol{A}_1^{\perp}, \boldsymbol{A}_2^{\perp}$, および $\boldsymbol{A}_e^{\perp}$
によって直和分解され $L(A) = A_0 \oplus A_1^{\perp} \oplus A_2^{\perp}$ と表される。$\boldsymbol{A}^{\perp}$ は $\boldsymbol{A}_0$ で張られる空間とは直交補空間の関係にある空間を表す。このとき各部分空間への射影行列を $\boldsymbol{Q}_0, \boldsymbol{Q}_1, \boldsymbol{Q}_2, \boldsymbol{Q}_e$ とすると $\boldsymbol{I} = \boldsymbol{Q}_0 + \boldsymbol{Q}_1 + \boldsymbol{Q}_2 + \boldsymbol{Q}_e$ である。

従って、これらの射影行列を用いれば、一要因の分散分析の場合と全く同じようにしてベクトルの長さ $\boldsymbol{x}^T\boldsymbol{x}$ を分解することができる。

$$\begin{aligned} \boldsymbol{x}^T\boldsymbol{x} &= \boldsymbol{x}^T\boldsymbol{Q}_0\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{Q}_1\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{Q}_2\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{Q}_e\boldsymbol{x} && (67) \\ &= \boldsymbol{x}^T\boldsymbol{Q}_0\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{Q}_1\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{Q}_2\boldsymbol{x} + \boldsymbol{x}^T(\boldsymbol{I} - \boldsymbol{Q})\boldsymbol{x} && (68) \\ &= S_0 + S_1 + S_2 + S_e && (69) \end{aligned}$$

9.2 繰り返しのある 2 要因の分散分析

表 5 のように、各条件あたりのデータが 1 つしかない場合を複数個のデータを持つことができるように拡張する。分析に用いられるデータは次の表 7 のようになる。

	条件 A1	条件 A2	条件 A3	
条件 B1	x_{111}	x_{121}	x_{131}	$\bar{x}_{1\cdot\cdot}$
	x_{112}	x_{122}	x_{132}	
条件 B2	x_{211}	x_{221}	x_{231}	$\bar{x}_{2\cdot\cdot}$
	x_{212}	x_{222}	x_{232}	
	$\bar{x}_{\cdot 1\cdot}$	$\bar{x}_{\cdot 2\cdot}$	$\bar{x}_{\cdot 3\cdot}$	$\bar{x}_{\cdot\cdot\cdot}$

表 7: 繰り返しのある 2 元配置の分散分析に用いられるデータ

対応する線形モデルは

$$x_i jk = \mu + \alpha_i + \beta_j + (\gamma)_{ij} + \epsilon_{ijk} \tag{70}$$

これを書き下すと次のようになる。

$$\begin{pmatrix} x_{111} \\ x_{112} \\ x_{121} \\ x_{122} \\ x_{131} \\ x_{132} \\ x_{211} \\ x_{212} \\ x_{221} \\ x_{222} \\ x_{231} \\ x_{232} \end{pmatrix} = \begin{pmatrix} 1 & 1 & & & 1 & & 1 & & & & & \\ 1 & 1 & & & 1 & & 1 & & & & & \\ 1 & & 1 & & 1 & & & 1 & & & & \\ 1 & & 1 & & 1 & & & 1 & & & & \\ 1 & & & 1 & 1 & & & & 1 & & & \\ 1 & & & 1 & 1 & & & & 1 & & & \\ 1 & 1 & & & & 1 & & & & 1 & & \\ 1 & 1 & & & & 1 & & & & 1 & & \\ 1 & & 1 & & & 1 & & & & & 1 & \\ 1 & & 1 & & & 1 & & & & & 1 & \\ 1 & & & 1 & & 1 & & & & & & 1 \\ 1 & & & 1 & & 1 & & & & & & 1 \end{pmatrix} \begin{pmatrix} \theta_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{13} \\ \gamma_{21} \\ \gamma_{22} \\ \gamma_{23} \end{pmatrix} + \begin{pmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \epsilon_{121} \\ \epsilon_{122} \\ \epsilon_{131} \\ \epsilon_{132} \\ \epsilon_{211} \\ \epsilon_{212} \\ \epsilon_{221} \\ \epsilon_{222} \\ \epsilon_{231} \\ \epsilon_{232} \end{pmatrix} \tag{71}$$

これまでと全く同様にして

$$\boldsymbol{x} = \boldsymbol{A\theta} + \boldsymbol{\epsilon} \tag{72}$$

と書くことができて、計画行列が $\boldsymbol{A} : [\boldsymbol{A}_0, \boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_3]$ となっただけである。$\boldsymbol{\theta} = (\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})^T$ であり、$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$, $\boldsymbol{\gamma} = (\gamma_1 1, \gamma_1 2, \gamma_1 3, \gamma_2 1, \gamma_2 2, \gamma_2 3)^T$ である。

空間の直和分解については

$$R = L(A_0) \oplus L^{\perp}(A_1) \oplus L^{\perp}(A_2) \oplus L^{\perp}(A_3) \oplus L^{\perp}(A_e) \tag{73}$$

帰無仮説は 3 つあって、それぞれ対応する空間へと射影したベクトルの長さが誤差ベクトルと比べて十分に長いと言えるか否かを検討することになる。

分散分析表は表 8 のようになる。

要因	平方和	自由度	平均平方	F
A	S_1	m-1	$S_1/(m-1)$	$\dfrac{S_1/(m-1)}{S_e/(m-1)(k-1)}$
B	S_2	k-1	$S_2/(k-1)$	$\dfrac{S_2/(k-1)}{S_e/(m-1)(k-1)}$
A×B	S_3	(m-1)(k-1)	$S_3/(m-1)(k-1)$	$\dfrac{S_3/(m-1)(k-1)}{S_e/mk(t-1)}$
誤差	S_e	$mk(t-1)$	$S_e/mk(t-1)$	
計	$\boldsymbol{x}^T\boldsymbol{x}$	mk - 1		

表 8: 繰り返しのある 2 要因の分散分析表

10　交互作用

2 要因以上の実験計画で繰り返しが存在する場合には、各条件毎に分散を求めることができる。このとき、各個の条件毎に差が存在する場合がある。

むしろ、積極的に交互作用の効果を検討するような実験が立案されることも多い。

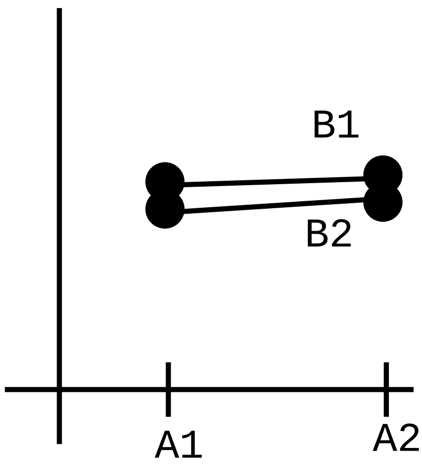

Figure 4: いずれの効果も認められない例

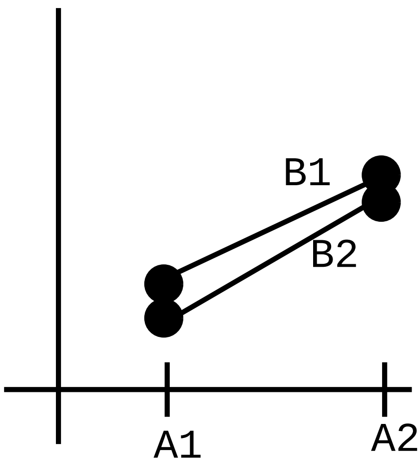

Figure 5: 要因 A の効果が認められる例

図 4 は横軸に条件 (要因) A(2 水準) をとり、条件 (要因) B (同じく 2 水準) を異なる線として表現したものである。この図 4 の場合は条件 A にも条件 B にも差異が認められず、すべてのデータが全平均の近くに集まっている。

一方、図 5 では要因 B の効果が認められず (B1, B2 間で差がない) 要因 A の効果のみが有意であるような場合である。

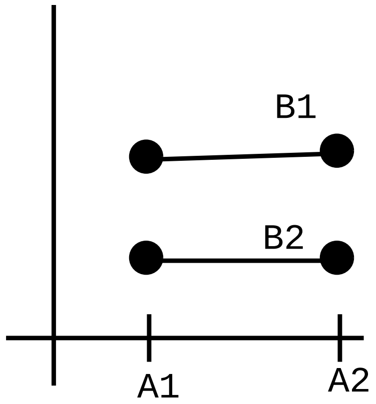

Figure 6: 要因 B の効果が認められる例

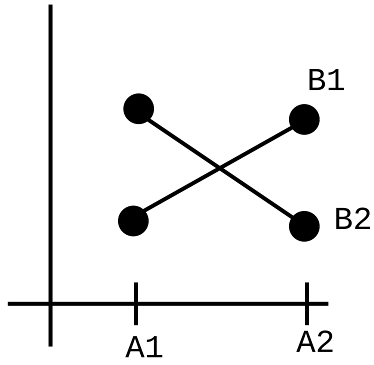

Figure 7: A と B の交互作用効果が認められる例

図 6 は反対に要因 B の効果は有意であるが、要因 A では A1 と A2 との間に差が認められない場合である。

図 7 は交互作用が有意となった場合である。条件毎に要因 A の効果と要因 B の効果の現れ方が異なる。

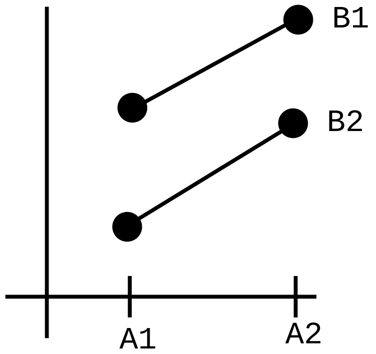

Figure 8: 2 つの要因の主効果が共に有意であると考えられる例

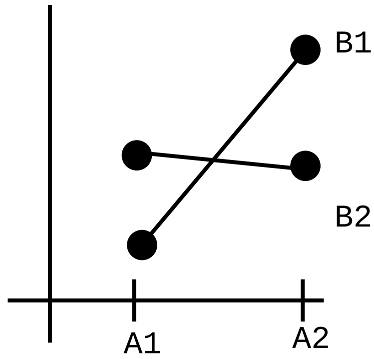

Figure 9: 要因 A の主効果と交互作用が認められる例

図 8 は A1 と A2 との間にも、B1 と B2 との間にも共に有意な差が認められる場合の例である。

図 9 は A1 と A2 との間に有意な差が認められ、かつ交互作用も存在するが要因 B の効果は認められない例である。あるいは A1 と A2 との差が有意となったのは、B2 条件のみであったということもできる。

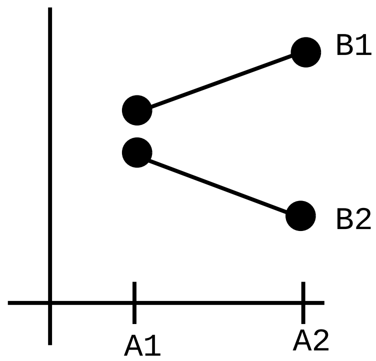

Figure 10: B の主効果と交互作用が認められる例

図 10 は B1 と B2 との間に有意な差が認められ、かつ交互作用も存在するが要因 A の効果は認められない例である。これは B1 と B2 との差が認められたのは A2 条件においてのみであったと解釈することもできる。

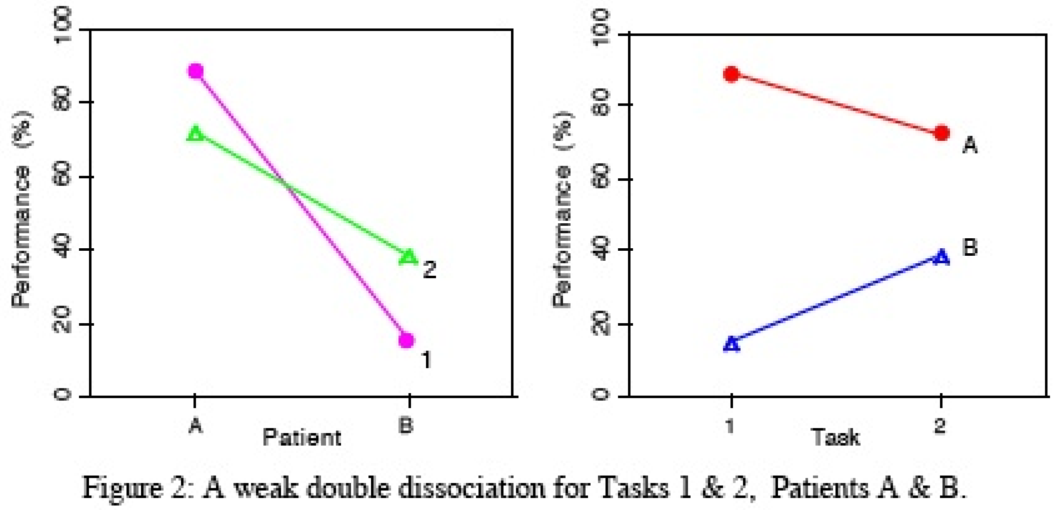

Figure 11: Bullinaria(1999), Figure 2

図 11 左は横軸に二人の意味記憶に障害を持つ患者の語彙判断課題の成績を示している (Bullinaria,1999)。患者 A は課題 1, 課題 2 の成績とも良好であり、患者 B は相対的に成績が落ちている。

同じデータを書き直したものが図 11 右である。横軸を課題とし患者の違いを異なる線として表現している。

このように同じデータであっても作図の仕方によって印象が異なる。

11　さらに複雑な実験計画

3 要因以上の実験計画であっても基本的な考え方は変らない。どれか一つの要因について繰り返しがなければ、繰り返しのない 2 要因の分散分析と同じく交互作用の検討をすることができない。最も一般的な形で 3 要因の実験計画を表現すると

$$x_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + (\gamma\alpha)_{ki} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl} \tag{74}$$

となる。ここで $(\alpha\beta)_{ij}, (\beta\gamma)_{jk}, (\gamma\alpha)_{ki}$ は 2 次の交互作用を表し、$(\alpha\beta\gamma)_{ijk}$ は 3 次の交合作用を表す。データの性質や実験の目的によってはいくつかの交互作用を無視して考えることも行われる。

エクセルの統計ツールで用意されているのは 2 元配置 (2 要因) の分散分析までであるが、どんな複雑な実験計画でも計画行列を作り、射影行列による直和分解を用いれば原理的には計算可能である。

SAS, SPSS などの統計パッケージソフトウェアでは GLM(一般線形モデル: generalized linear model) などの名前で利用可能である。

つまり、すべての実験計画の基本的なアイデアはすべて同じであり、計画行列を如何に作るかによってデータの解釈が異なってくる。計画行列を元にデータを直和分解して、その部分空間内での線分の長さに対応する量を考えれば良い。

また、一般に検討すべき要因が増えれば増えるほど手続きは煩雑になり、結果の解釈も難しくなる。高次交互作用の解釈も難しい。

そこで、主効果 (2 次の交互作用も加えることがある) のみを検討する方法が開発されている。

11.1　一部実施と直交計画表

例えば A, B, C, D という 4 つの要因があり、すべて 2 水準であるとする。交互作用については AB についてだけ考慮したいが、それ以外の交互作用については理論的に無視できるものと考えよう。すべての実験条件について $2^4 = 16$ 条件あるが、この一部だけを実施するものとし、以下の表 11.1 のうち x_i の条件だけ実施する。

	A	B	C	D	
1	1	1	1	1	x_1
2	1	1	1	2	
3	1	1	2	1	
4	1	1	2	2	x_2
5	1	2	1	1	
6	1	2	1	2	x_3
7	1	2	2	1	x_4
8	1	2	2	2	
9	2	1	1	1	
10	2	1	1	2	x_5
11	2	1	2	1	x_6
12	2	1	2	2	
13	2	2	1	1	x_7
14	2	2	1	2	
15	2	2	2	1	x_8
16	2	2	2	2	

A の主効果を α_1, α_2, B, C, D の主効果をそれぞれ β, γ, δ で表すものとする。A と B との交互作用については τ とする。

線形モデルを用いて

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & -1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \\ \beta \\ \gamma \\ \delta \\ \tau \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \end{pmatrix} \tag{75}$$

計画行列のうち 2 列目が A, 3 列目が B, 4 列目が C, 5 列目が D, 6 列目が A と B との交互作用を表している。これを

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{76}$$

として解くことを考える。

計画行列 $\boldsymbol{A}$ の列ベクトルは直交していることは

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{pmatrix} 8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 8 \end{pmatrix} \tag{77}$$

となることから確認できる。$\boldsymbol{A}^T\boldsymbol{x}$ を計算すると

$$\begin{pmatrix} x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 \\ x_1 + x_2 + x_3 + x_4 - x_5 - x_6 - x_7 - x_8 \\ x_1 + x_2 - x_3 - x_4 + x_5 + x_6 - x_7 - x_8 \\ x_1 - x_2 + x_3 - x_4 + x_5 - x_6 + x_7 - x_8 \\ x_1 - x_2 - x_3 + x_4 - x_5 + x_6 + x_7 - x_8 \\ x_1 + x_2 - x_3 - x_4 - x_5 - x_6 + x_7 + x_8 \end{pmatrix} \tag{78}$$

各要素がそれぞれ全平均、A の効果、B の効果、C の効果、D の効果、A と B との交互作用の効果に成っている。$(\boldsymbol{A}^T\boldsymbol{A})^{-1}$ は対角要素が 1/8 で非対角要素が全て 0 である（直交しているので）単純な行列になっているので計算が極端に簡単となる。

$$\begin{aligned} \boldsymbol{x}^T\boldsymbol{x} &= \boldsymbol{x}^T\boldsymbol{A}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{x} && (79) \\ &= \boldsymbol{x}^T\boldsymbol{A}_0\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{A}_1\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{A}_2\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{A}_3\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{A}_4\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{A}_5\boldsymbol{x} && (80) \\ &= S_0^2 + S_\alpha + S_\beta^2 + S_\gamma^2 + S_\delta^2 + S_\tau^2 + S_e^2 && (81) \end{aligned}$$

従って次の分散分析表 9 を得る。

要因	平方和	自由度	
A	S_α	1	S_α/S_e
B	S_β	1	S_β/S_e
C	S_γ	1	S_γ/S_e
D	S_δ	1	S_δ/S_e
AB	S_τ	1	S_τ/S_e
誤差	S_e	1	

表 9: 分散分析表

この例からわかるとおり計画行列が直交している場合には多数の因子を少い測定で得ることができる。この直交表にそれぞれの因子（主効果）および交互作用を割り当てればよい。例えば 3 要因の実験計画で全ての交互作用を考慮しなければならないとすれば主効果 3 つ（ A, B, C)、2 次の交互作用 3 つ (AB, BC, CA) 3 次の交互作用 1 つ (ABC) を含む実験に対して次の表 10

データ	μ	A	B	AB	C	AC	BC	ABC
x_1	1	1	1	1	1	1	1	1
x_2	1	1	1	1	-1	-1	-1	-1
x_3	1	1	-1	-1	1	1	-1	-1
x_4	1	1	-1	-1	-1	-1	1	1
x_5	1	-1	1	-1	1	-1	1	-1
x_6	1	-1	1	-1	-1	1	-1	1
x_7	1	-1	-1	1	1	-1	-1	1
x_8	1	-1	-1	1	-1	1	1	-1
(75)A の列	1	2	3	6	4			5
上の例での要因	μ	α	β	τ	δ	e	e	γ

表 10: 直交計画表 $L_u(2^7)$

を直交計画表という。直交計画表の全ての列は互いに直交することが本質的な点である。交互作用 AB の列の各要素は A と B との列の対応する要素同士の積になっている。

直交行列の 1 列を水準 A と考えることを直交行列に因子（あるいは水準）A を割り付けると言う。直交行列には $L_{16}(2^{15})$ などいくつか考案されている。考え方は全て同じである。