統計的方法のメモ

前川眞一

130607,140411,0417,24,0530

1　はじめに

2　標本分布

2.1　はじめに

母集団から大きさ n の標本を抽出する。
観測されるデータは、母集団からサンプルされたものであると考える。
n が大きくなればなるほど、そのデータは母集団の状態をより正確に反映していると考えられる。

2.2　標本統計量

標本平均

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$

標本分散

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{2}$$

（$n-1$ で割る流儀もある。）分散の平方根を標準偏差と呼ぶ。
標本共分散

$$c_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \tag{3}$$

標本相関係数

$$r_{XY} = \frac{c_{XY}}{S_X S_Y} \tag{4}$$

$$= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{5}$$

2.3　標本平均の分布

母集団において

$$X_i \sim N(\mu, \sigma^2), \quad i = 1, 2, \ldots, n, \quad iid \tag{6}$$

である場合、独立な n 個の観測値の標本平均

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{7}$$

の分布は

$$\overline{X} \sim N(\mu, \sigma^2/n) \tag{8}$$

となる。

これは、母集団から、大きさ n の標本を採り、その平均を計算したものを $\overline{X}_k$ とする、という作業を m 回繰り返した時の $\overline{X}_k,\ k = 1, 2, \ldots, m$ の分布である。

これは、正規分布の性質から

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \tag{9}$$

とも書ける。

母分散が未知の場合は, 上記の Z の中の σ を標本標準偏差

$$s^2 \quad = \quad \sum_{i=1}^{n} \frac{1}{n-1}(X_i - \overline{X})^2 \tag{10}$$

で置き換えたもの T が自由度 $n-1$ の t 分布をすること

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t(n-1) \tag{11}$$

が知られている。

したがって、母数の値 (μ, σ) がわかっていれば、母平均がある区間に入る確率を計算できる。

正規分布の場合、μ, σ がわかっていれば、標準正規分布表から

$$\beta \quad = \quad \Pr\left(Z \leq z_\beta\right) \tag{12}$$

となる点 z_β が計算できる。t 分布の場合も同様に

$$\beta \quad = \quad \Pr\left(T \leq t_{\beta,n-1}\right) \tag{13}$$

となる点 $t_{\beta,n-1}$ が計算できる。

（両側検定の場合は $\beta = \alpha/2$ とする。）

2.4 標本比率の分布

標本比率の分布は

$$p \sim \ N(\pi, \frac{\pi(1-\pi)}{n}) \tag{14}$$

で近似することが出来る。したがって、

$$Z = \frac{p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1) \tag{15}$$

を利用する。

3　区間推定（信頼区間）

母集団から n 個のサンプルを取り、標本平均 $\overline{X}$ と標本標準偏差 s を観測した時の母平均 μ の $100 \times (1-\alpha)\%$ 信頼区間は以下の通りである。

$$m_L = \overline{X} - y_{(1-\alpha/2)} \frac{s}{\sqrt{n}} \tag{16}$$

$$m_U = \overline{X} + y_{(1-\alpha/2)} \frac{s}{\sqrt{n}} \tag{17}$$

ただし、$y_{(1-\alpha/2)}$ はそれぞれの分布において $100 \times (1-\alpha/2)\%$ パーセンタイル、すなわち、$100 \times (\alpha/2)\%$ の上側確率を与える点である。すなわち、
正規分布の場合

$$y_{(1-\alpha/2)} = z_{(1-\alpha/2)} \tag{18}$$

t 分布の場合

$$y_{(1-\alpha/2)} = t_{(1-\alpha/2),n-1} \tag{19}$$

である。

母集団標準偏差 σ がわかっている時、もしくは n が大きい時は正規分布を用いても良い。

3.1　導出

$$\begin{aligned}
1-\alpha &= \Pr\left(\mu - y_{(1-\alpha/2)}\frac{s}{\sqrt{n}} \le \bar{X} \le \mu + y_{(1-\alpha/2)}\frac{s}{\sqrt{n}}\right) && (20)\\
&= \Pr\left(-y_{(1-\alpha/2)}\frac{s}{\sqrt{n}} \le \bar{X} - \mu \le y_{(1-\alpha/2)}\frac{s}{\sqrt{n}}\right) && (21)\\
&= \Pr\left(-\bar{X} - y_{(1-\alpha/2)}\frac{s}{\sqrt{n}} \le -\mu \le -\bar{X} + y_{(1-\alpha/2)}\frac{s}{\sqrt{n}}\right) && (22)\\
&= \Pr\left(\bar{X} + y_{(1-\alpha/2)}\frac{s}{\sqrt{n}} \ge \mu \ge \bar{X} - y_{(1-\alpha/2)}\frac{s}{\sqrt{n}}\right) && (23)\\
&= \Pr\left(\bar{X} - y_{(1-\alpha/2)}\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + y_{(1-\alpha/2)}\frac{s}{\sqrt{n}}\right) && (24)
\end{aligned}$$

3.2　標本比率

$$m_L = p - z_{(1-\alpha/2)}\sqrt{\frac{p(1-p)}{n}} \tag{25}$$

$$m_U = p + z_{(1-\alpha/2)}\sqrt{\frac{p(1-p)}{n}} \tag{26}$$

4 統計的仮説検定の一般論

4.1 手続き

- 母集団の統計モデルを作る。
  正規分布、２項分布
- 検定仮説を立てる。
  H_0 帰無仮説（何らかの制約が付いているもの）
  H_1 制約を取ったものでふつうはこれを採択したい場合が多い。
- H_0 の下での検定統計量の分布を求める。
  標本平均、標本比率、平均平方和 その他
  正規分布、t 分布、 F 分布 etc
- 棄却域を決める。
  H_0 の下での検定統計量の分布で起こりそうもないこと（確率が小さいこと）を定義する。
- データを取る。$X_i,\ i = 1, 2, \ldots, n$
- 観測された検定統計量が棄却域に入るかどうかを調べる。
  棄却域に入ったということは H_0 の下で起こりそうもないことが起こったということなので
  H_0 を棄却する。
  もしくは p-値を計算する。p-値とは、H_0 が真の時に、検定統計量が、実際にデータから計算された検定統計量の値よりも極端な値をとる確率。

4.2 検定統計量の選び方

Table 1: 仮説検定に伴う誤り

		真の状態	
		H_0	H_1
採択された仮説	H_0	正解	第２種の誤り Type II Error β
	H_1	第１種の誤り Type I Error α	正解

検定力とは H_1 が正しい時にそれを正しく採択する確率のこと、すなわち $1 - \beta$。

以下の図では、右上から左下への斜線部分が 第１種の誤り Type I Error の確率 α を表し、左上から右下への斜線部分が 第２種の誤り Type II Error の確率 β を表している。したがって、検定力は右側の分布において、白い部分からこの斜線部分を引いたものとなる。

この図では Type I Error の大きさは同じだが、下の統計量の方が検定力が高いことがわかる。

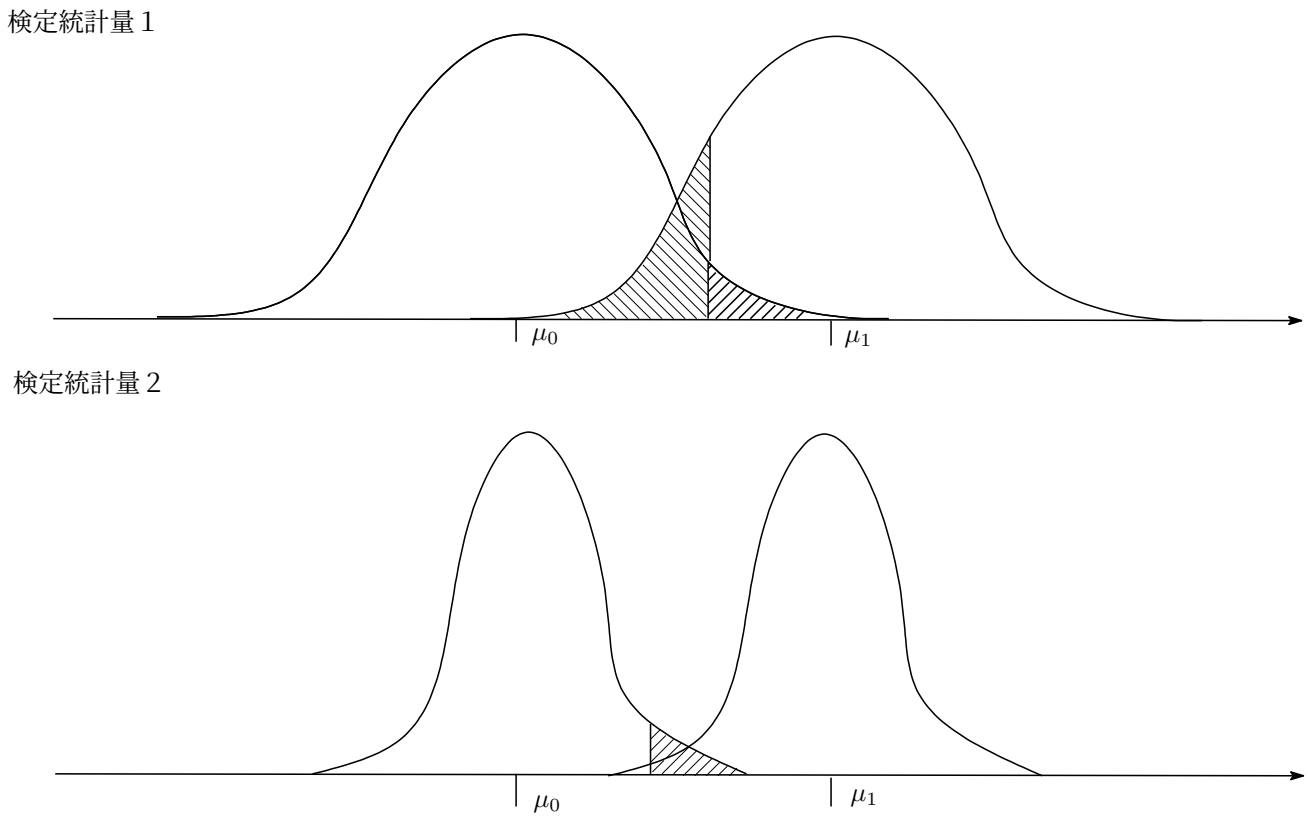

Figure 1: Power Comparison

5 仮説検定

5.1 一つの母集団の母数に関する仮説検定

- 母集団の統計モデルを作る。
  母集団においてデータは正規分布をする。

$$X_i \sim N(\mu, \sigma^2), \quad i = 1, 2, \ldots, n, \quad iid \tag{27}$$

- 検定仮説を立てる。
  両側検定
  $H_0 : \mu = \mu_0$　　母平均 μ は μ_0 に等しい。
  $H_1 : \mu \neq \mu_0$　　母平均 μ は μ_0 に等しくない。
  片側検定 1
  $H_0 : \mu = \mu_0$　　母平均 μ は μ_0 に等しい。
  $H_1 : \mu \geq \mu_0$　　母平均 μ は μ_0 より大きい。
  片側検定 2
  $H_0 : \mu = \mu_0$　　母平均 μ は μ_0 に等しい。
  $H_1 : \mu \leq \mu_0$　　母平均 μ は μ_0 より小さい。

- H_0 の下での検定統計量の分布を求める。

$$\overline{X} \sim N(\mu_0, \sigma^2/n) \qquad \text{or} \qquad Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \tag{28}$$

もしくは σ を s で置き換えて

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t(n-1) \tag{29}$$

- 棄却域を決める。
  H_0 の下での検定統計量の分布で起こりそうもないこと（確率が小さいこと）を定義する。
  両側検定
  棄却域 $= (T < t_{(\alpha/2),\, n-1}) \quad \text{or} \quad (t_{(1-\alpha/2),\, n-1} < T) \qquad (t_{(\alpha/2)} = -t_{(1-\alpha/2),\, n-1})$
  片側検定
  棄却域 $= (t_{(1-\alpha),\, n-1} < T)$

- データを取る。$X_i,\ i = 1, 2, \ldots, n$ から z や t を計算する。（小文字はデータから計算された Z や T の値を示す。）

- 観測された検定統計量が棄却域に入るかどうかを調べる。
  棄却域に入ったということは H_0 の下で起こりそうもないことが起こったということなので
  H_0 を棄却する。
  両側検定の場合、$t < -t_{(1-\alpha/2,\, n-1)}$ または $t_{(1-\alpha/2,\, n-1)} < t$ ならば H_0 を棄却する。

  p-値の計算
  H_0 の下で $T > t$ または $T < t$ となる確率。この確率が小さい場合、 H_0 の下で起こりにくいことが起こったとしてH_0 を棄却する。

5.2 二つの母集団の母数（の差）に関する仮説検定 $\sigma_1 = \sigma_2 (= \sigma^2)$ が仮定出来る場合

母集団の統計モデルは。

$$X_{ij} \sim N(\mu_j, \sigma^2), \quad i = 1, 2, \ldots, n_j, \quad j = 1, 2 \quad iid \tag{30}$$

両側検定
$H_0 : \mu_1 = \mu_2$　　ふたつ母平均 μ_1, μ_2 は等しい。
$H_1 : \mu_1 \neq \mu_2$　　ふたつ母平均等しくない。
片側検定 1
$H_0 : \mu_1 = \mu_2$　　ふたつ母平均 μ_1, μ_2 は等しい。
$H_1 : \mu_1 \geq \mu_2$　　母平均 μ_1 は μ_2 より大きい。
片側検定 2
$H_0 : \mu_1 = \mu_2$　　ふたつ母平均 μ_1, μ_2 は等しい。
$H_1 : \mu_1 \leq \mu_2$　　母平均 μ_1 は μ_2 より小さい。

検定統計量は

$$\overline{X_1} \sim N(\mu_1, \sigma^2/n_1) \qquad \text{and} \qquad \overline{X_2} \sim N(\mu_2, \sigma^2/n_2) \tag{31}$$

であるから

$$\overline{X_1} - \overline{X_2} \quad \sim \quad N(\mu_1 - \mu_2, \sigma^2/n_1 + \sigma^2/n_2) \;=\; N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right) \tag{32}$$

したがって

$$Z = \frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1) \tag{33}$$

もしくは σ を s で置き換えて

$$T = \frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} = \frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2) \tag{34}$$

ただし、ふたつのサンプルのデータをまとめて（プールして）推定した σ^2 の推定値を

$$s^2 \quad = \quad \frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{n_1}(X_{i1} - \overline{X_1})^2 + \sum_{i=1}^{n_2}(X_{i2} - \overline{X_2})^2\right) \tag{35}$$

とする。

棄却域の計算は以下の通り。ただし、 $t_{(\beta)}$ を自由度 $n_1 + n_2 - 2$ の t 分布において下側確率が β となる点とする。なお、$t_{(1-\beta)} = -t_{(\beta)}$ である。

両側検定

棄却域 $= (T < t_{(\alpha/2)}) \quad \text{or} \quad (t_{(1-\alpha/2)} < T)$

片側検定 1

棄却域 $= (t_{(1-\alpha)} < T)$

片側検定 2

棄却域 $= (T < t_{(\alpha)})$

5.3 二つの母集団の母数（の差）に関する仮説検定 $\sigma_1 = \sigma_2$ が仮定出来ない場合

$$\overline{X_1} \sim N(\mu_1, \sigma_1^2/n_1) \qquad \text{and} \qquad \overline{X_2} \sim N(\mu_2, \sigma_2^2/n_1) \tag{36}$$

したがって、

$$\overline{X_1} - \overline{X_2} \sim \;\; N\left(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2\right) \tag{37}$$

しかし、これの σ_1, σ_2 を標本標準偏差 s_1, s_2 で置き換えたものは t 分布をしない。したがって、検定統計量とその自由度を以下のように変更する。 (Welch)

$$T = \frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim\sim t(\nu) \tag{38}$$

ただし、修正された自由度を

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}} \tag{39}$$

とする。

棄却域の計算は上記の通り。ただし、 $t_{(\beta)}$ を自由度 ν の t 分布において下側確率が β となる点とする。

6　3 母集団以上の平均値の差の検定

t 検定は、一つの母集団の平均値が、特定の値であるという仮説や、二つの母集団の平均値が同じであるという仮説を検定する際に用いられる。

母集団が三つ以上になった場合、t 検定は使わない。

その理由は第 1 種の過誤の確率をコントロールできないからである。

いま、三つの母集団があり、その平均を μ_1, μ_2, μ_3 とする。数学的には $\mu_1 = \mu_2,\quad \mu_2 = \mu_3$ が $\mu_1 = \mu_2 = \mu_3$ を意味するので、検定仮説として次の二つを考える。

$$
\begin{aligned}
&H_{01} : \mu_1 = \mu_2 \\
&H_{11} : \mu_1 \neq \mu_2 \qquad (40) \\
&\qquad \text{and} \\
&H_{02} : \mu_2 = \mu_3 \\
&H_{12} : \mu_2 \neq \mu_3 \qquad (41)
\end{aligned}
$$

それぞれの仮説を有意水準 α で検定すると、二つの検定で少なくとも一つの type I error を起こす確率は、これらの仮説が独立であるとしても、$1 - (1-\alpha)^2 \approx 2\alpha$ となる。独立でない場合はどうなるか解らない。

一般的に、母集団が p 個あり、$p-1$ 個の仮説群の検定を行った場合、それらが独立の場合には全体としての type I error の確率（少なくとも一つの type I error を起こす確率）は $1 - (1-\alpha)^{p-1} \approx (p-1)\alpha$ となる。

したがって、母集団が三つ以上ある場合には、別の方式を考えなければならない。

一般的に p 個の母集団があり、母分散は共通であると仮定できる場合、

$$
\begin{aligned}
H_0 :&\quad \mu_1 = \mu_2 =, \cdots, = \mu_p \quad \text{（すべての母平均は等しい）} &(42)\\
H_1 :&\quad \text{not } H_0 \quad \text{（少なくとも一つの母平均は他とことなる）} &(43)
\end{aligned}
$$

を適切な検定統計量を用いて有意水準 α で検定することになる。（通常は σ の値を未知として、F 統計量を用いる。）

このような手法を（一元配置の）分散分析と呼ぶ。（後述）

7　回帰分析

7.1　はじめに

単一の連続的な変数 Y （従属変数、基準変数、被説明変数）を一つ、または、複数の変数群 X （独立変数、説明変数、予測変数)、で説明する方法。

一般的な回帰曲線の定義は、説明変数 X を与えた時の基準変数 Y の条件付き期待値 $\boldsymbol{\mathcal{E}}(Y|X)$ の描く曲線（曲面）のこと。この曲線が、直線（平面）であると仮定したものが、以下に示す回帰分析である。

$$y_i = \alpha + \beta\, x_i + \varepsilon_i \qquad \text{単回帰分析} \tag{44}$$

$$y_i = \beta_0 + \beta_1\, x_{i1} + \beta_2\, x_{i2} + \varepsilon_i \qquad \text{重回帰分析（説明変数ふたつ）} \tag{45}$$

$$y_i = \beta_0 + \sum_{j=1}^{q} \beta_j\, x_{ij} + \varepsilon_i \qquad \text{重回帰分析（一般形）} \tag{46}$$

ただし、$i = 1, 2, \ldots, N,$ として

$$\varepsilon_i \sim\ N(o, \sigma^2) \tag{47}$$

また、誤差を除いた部分、すなわち、X を与えた時の Y の期待値の部分

$$\widehat{y}_i = \beta_0 + \sum_{j=1}^{q} \beta_j\, x_{ij} \tag{48}$$

を予測値と呼ぶ。

回帰係数 β_j は、他の説明変数の値を固定した時に、第 j 番目の変数 x_{ij} を一単位だけ動かした時の y の変化量を表す。

7.2　回帰係数の推定

単回帰分析の場合、回帰係数の推定値は

$$\widehat{\beta} = \frac{\boldsymbol{\mathcal{C}}ov\,(Y, X)}{var\,(X)} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{49}$$

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\,\overline{X} \tag{50}$$

となり、一般的には

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}`\boldsymbol{y} \tag{51}$$

となる。

7.3　分散説明率・決定係数・R^2

$$\mathrm{R}^2 = \frac{var\left(\widehat{Y}\right)}{var\,(Y)} \tag{52}$$

を分散説明率と呼ぶ。この値が大きい方がよい。ちなみに、単回帰の場合、 R^2 は Y と X の相関係数の 2 乗となっている。重回帰の場合は $\boldsymbol{\mathcal{C}}orr\,(Y, \widehat{y})$ の 2 乗である。

どのような説明変数を用いても、説明変数を増やすことにより、 R^2 は増加する。

7.4 回帰分析の行列表記

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} && (53) \\ &= \widehat{\boldsymbol{y}} \quad + \boldsymbol{\varepsilon}, \qquad \widehat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta} && (54) \end{aligned}$$

$\boldsymbol{y} \to N \times 1$	基準変数
$\boldsymbol{X} \to N \times (q+1)$	説明変数 + $\mathbf{1}$
$\boldsymbol{\beta} \to (q+1) \times 1$	回帰係数（切片を含む）
$\widehat{\boldsymbol{y}} \to \boldsymbol{X}\boldsymbol{\beta} = N \times 1$	モデルの予測値

$$\begin{pmatrix} \widehat{y}_1 \\ \widehat{y}_{21} \\ \vdots \\ \widehat{y}_N \end{pmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1q} \\ 1 & x_{21} & x_{22} & \dots & x_{2q} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{N1} & x_{N2} & \dots & x_{Nq} \end{bmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{pmatrix} \tag{55}$$

$$y_i = \beta_0 + \sum_{i=1}^{q} x_{ij}\beta_j + \varepsilon_{ij} \tag{56}$$

ここで

$$x_{i0} = 1 \tag{57}$$

とすると

$$y_i = \sum_{i=0}^{q} x_{ij}\beta_j + \varepsilon_{ij} \tag{58}$$

$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 && (59) \\ \hat{y}_i &= \sum_{j=0}^{q} x_{ij}\beta_j && (60) \end{aligned}$$

残差の 2 乗和 RSS を回帰係数 $\boldsymbol{\beta}$ の要素で偏微分すると

$$\begin{aligned} \frac{\partial\,\mathrm{RSS}}{\partial\,\beta_k} &= 2\sum_{i=1}^{N}(y_i - \hat{y}_i) \times \frac{\partial\,(y_i - \hat{y}_i)}{\partial\,\beta_k} && (61) \\ &= 2\sum_{i=1}^{N}(y_i - \hat{y}_i) \times \frac{\partial \sum_{j=0}^{q} x_{ij}\beta_j}{\partial\,\beta_k} && (62) \\ &= 2\sum_{i=1}^{N}(y_i - \hat{y}_i) \times x_{ik} && (63) \\ &= 2\sum_{i=1}^{N}(y_i - \sum_{j=0}^{q} x_{ij}\beta_j) \times x_{ik} && (64) \end{aligned}$$

$$= \quad 2\sum_{i=1}^{N} y_i x_{ik} - 2\sum_{i=1}^{N}(\sum_{j=0}^{q} x_{ij}\beta_j)x_{ik} \tag{65}$$

$$= \quad 2\sum_{i=1}^{N} y_i x_{ik} - 2\sum_{i=1}^{N}(\sum_{j=0}^{q} x_{ij}x_{ik}\beta_j) \tag{66}$$

$$= \quad 2\sum_{i=1}^{N} y_i x_{ik} - 2\sum_{j=0}^{q}(\sum_{i=1}^{N} x_{ij}x_{ik})\beta_j \tag{67}$$

$\sum_{i=1}^{N} y_i x_{ik}$ は $\boldsymbol{X}'\boldsymbol{y}$ の第 k 要素、$\sum_{i=1}^{N} x_{ij}x_{ik}$ は $\boldsymbol{X}'\boldsymbol{X}$ の第 jk 要素
したがって、方程式 $\frac{\partial\,\mathrm{RSS}}{\partial\,\beta_k} = 0, k = 0, 1, 2, \ldots, q$ は、まとめて

$$\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0} \tag{68}$$

と書ける。従って、解は

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{69}$$

となる。

7.5 仮説検定

モデル全体の検定、すなわち、H_0 : $\beta_0 = 0, \beta_1 = 0, \beta_2 = 0, \ldots, \beta_q = 0$ の検定には F 分布を使う。個々の回帰係数が 0 出ないという仮説 H_0 : $\beta_j = 0$ の検定には t 分布を使う。

R^2 が小さくてもモデルの帰無仮説が棄却されることがある。(統計的には回帰モデルを作る意味はあるが、あまり役に立つモデルではない。)

統計的仮定：誤差項は正規分布（平均は 0、分散は共通）

$$y_i \quad = \quad \beta_0 + \sum_{j=1}^{q} x_{ij}\beta_j + \varepsilon_{ij} \;=\; \sum_{j=0}^{q} x_{ij}\beta_j, \quad \varepsilon_{ij} \sim \; N(0, \sigma^2) \tag{70}$$

行列で書くと

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \tag{71}$$

このモデルに関して、以下の仮説検定を考える。

$$H_0 : \quad \beta_{(1)} = \beta_{(2)} = \ldots = \beta_{(r)} = 0 \tag{72}$$

$$H_1 : \qquad \text{上記の制約がない} \tag{73}$$

すなわち、ある特定の r 個の回帰係数が 0 であるという帰無仮説である。このとき、帰無仮説の下で

$$F \quad = \quad \frac{\left(\mathrm{RSS}_0 - \mathrm{RSS}_1\right)/r}{\mathrm{RSS}_1/(N-q-1)} \tag{74}$$

が第 1 自由度 r、第 2 自由度 $N-q-1$ の F 分布をすることが分かっている。ただし、RSS_k は、仮説 H_k の元でパラメタを推定した際の残差の 2 乗和

$$\mathrm{RSS}_k = \sum_{i=1}^{N}(y_i - \widehat{y_{(k)}}_i)^2, \quad k = 0 \text{ or } 1 \tag{75}$$

である。(RSS_0 は H_0 で 0 と指定された回帰係数に対応する説明変数を $\boldsymbol{X}$ 行列から除いて回帰係数を推定した時の残差の 2 乗和。その時の予測値が $\widehat{y_{(0)}}_i$ 。 また、RSS_1 は全ての説明変数を用いて計算した残差の 2 乗和で、その時の予測値が $\widehat{y_{(1)}}_i$ 。

これは H_0 が切片を除くすべての回帰係数が 0 という r=q 個の制約の場合

$$F = \frac{\sum_{i=1}^{N}(\widehat{y}_i - \overline{y})^2/q}{\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2/(N-q-1)} \tag{76}$$

となる。ただし $\widehat{y}_i$ は H_1 の下での予測値、$\overline{y}$ は H_0 の下での予測値である。

また、一つの回帰係数 $\beta_{(1)}$ のみが 0 という r=1 個の制約の場合も同様の F 検定（第 1 自由度 1、第 2 自由度 $N-q-1$ ）が使えるが、通常は、それと同値の t 検定（自由度 $N-q-1$ ）を用いることが多い。

$\mathrm{RSS}_1 = \sum_{i=1}^{N}(y_i - \widehat{y}_i)^2$ を誤差平方和 (error SS)、$\mathrm{RSS}_0 - \mathrm{RSS}_1$ を回帰平方和 (model SS)、$\sum_{i=1}^{N}(y_i - \overline{y})^2$ を全平方和 (total SS) と呼ぶことがある。また、

$$R^2 = \frac{\sum_{i=1}^{N}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} = \frac{var\,(\widehat{\boldsymbol{y}})}{var\,(\boldsymbol{y})} = 1 - \frac{\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} \tag{77}$$

を決定係数と呼ぶ。

7.6 回帰分析のバリエーション

1. 単回帰分析
   基準変数 1 個、説明変数 1 個

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \beta \boldsymbol{x} + \boldsymbol{\varepsilon} \tag{78}$$

2. 重回帰分析
   基準変数 1 個、説明変数 q 個

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{79}$$

3. 多変量回帰分析
   基準変数 p 個、説明変数 q 個

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}' + \boldsymbol{\mathcal{E}}, \quad \boldsymbol{\beta} = (p \times q) \tag{80}$$

$$\widehat{\boldsymbol{\beta}'} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{81}$$

4. 多項式回帰分析

$$\widehat{y_i} = \sum_{m=0}^{M} \beta_m x_i^m \tag{82}$$

$$\boldsymbol{Y} = \boldsymbol{G}^{(M)}(\boldsymbol{x})\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{G}^{M}(\boldsymbol{x}) = (n \times M+1) \tag{83}$$

$$\boldsymbol{G}(\boldsymbol{x}) = [\boldsymbol{g}_m(\boldsymbol{x})], \quad m = 0, 1, \ldots, M \tag{84}$$

$$\boldsymbol{g}_m(\boldsymbol{x}) = [x_i^m] \tag{85}$$

5. 非線形回帰としてのスプライン回帰分析（多項式 + 切断冪関数）

$$\widehat{y_i} = \sum_{m=0}^{M} \beta_m x_i^m + \sum_{\ell=1}^{K} \beta_\ell (x_i - c_\ell)_+^M \tag{86}$$

$$(x_i - c_\ell)_+^M = \begin{cases} 0 & (x_i < c_\ell) \\ (x - c_\ell)^M & (x_i \geq c_\ell) \end{cases} \tag{87}$$

$$\boldsymbol{Y} = \boldsymbol{G}^{(K,M)}(\boldsymbol{x})\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \boldsymbol{G}^M(\boldsymbol{x}) = (n \times M + K + 1) \tag{88}$$

ただし、 K は節点の数、c_ℓ は節点

6. ニューラルネットワーク

$$y_i \quad = \quad \beta_0 + \sum_{j=1}^{Q} \psi(a_j(x_i - b_j))\beta_j + \varepsilon_i, \quad i = 1, 2, \cdots, n \tag{89}$$

ただし、

$$\psi(x) = \frac{1}{(1 + \exp(-x))} \tag{90}$$

は係数 a_j, b_j を持つロジスティク関数である。従って

$$\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{x})\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{91}$$

の形で書くと

$$\boldsymbol{G}(\boldsymbol{x}) = [\mathbf{1},\ \psi(a_1(\boldsymbol{x} - b_1)),\ \psi(a_2(\boldsymbol{x} - b_2)),\ \ldots\ \psi(a_Q(\boldsymbol{x} - b_Q))] \tag{92}$$

となる。

スプライン回帰や多項式と似ている点は、両者とも、本来の説明変数を G という形に展開し、それらの線形結合で $\boldsymbol{y}$ を予測しようと試みることである。

両者の違いは、スプラインの場合は、展開に際して用いられる接点 $c_\ell, \ell = 1, 2, \ldots, K$ は定数であるのに対し、ニューラルネットワークにおいては、その係数 $(a_j, b_j), j = 1, 2, \ldots, Q$ はパラメタとして推定されるものである点である。

また、多項式回帰の場合、中間層を増やすことは、結局、最初に展開された説明変数の線形回帰を作ることになり、意味がないが、スプラインや、特にニューラルネットワークの場合は、意味がある。

7.7 デザイン行列

水準数が A の 1 元配置の分散分析モデルは、各水準での観測値数を $n_j, j = 1, 2, \ldots, A$ とすると、

$$y_{ij} = \widehat{y}_{ij} + \varepsilon_{ij}, \quad \widehat{y}_{ij} = \mu_j, \quad i = 1, 2, \ldots, n_j,\ j = 1, 2, \ldots, A \tag{93}$$

である。これを線形モデル $\widehat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta}$ の形で書くと

$$\begin{pmatrix} \widehat{y}_{11} \\ \widehat{y}_{21} \\ \vdots \\ \widehat{y}_{n_1 1} \\ \widehat{y}_{12} \\ \widehat{y}_{22} \\ \vdots \\ \widehat{y}_{1A} \\ \vdots \\ \widehat{y}_{n_A A} \end{pmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \times \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_A \end{pmatrix} \tag{94}$$

となる。ただし、 μ_j は、第 j 群で観測される $y_{ij}, i = 1, 2, \dots, n_j$、の母平均である。（これをセル平均と呼ぶことがある。）

この同じモデルを、以下のように書き表すことも出来る。ただし、μ は全水準の平均 (grand mean) を表す母数、 α_j は第 j 水準の効果 (effect) を表す母数である。

$$y_{ij} = \widehat{y}_{ij} + \varepsilon_{ij}, \quad \widehat{y}_{ij} = \mu + \alpha_j, \quad i = 1, 2, \dots, n_j, \ j = 1, 2, \dots, A \tag{95}$$

である。これを線形モデル $\widehat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta}$ の形で書くと

$$\begin{pmatrix} \widehat{y}_{11} \\ \widehat{y}_{21} \\ \vdots \\ \widehat{y}_{n_1 1} \\ \widehat{y}_{12} \\ \widehat{y}_{22} \\ \vdots \\ \widehat{y}_{1A} \\ \vdots \\ \widehat{y}_{n_A A} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{bmatrix} \times \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_A \end{pmatrix} \tag{96}$$

しかし、このデザイン行列は、その第 2 列目から最後までを加えると第 1 列目の単位ベクトルとなる。このように、行列のある列が、残りの列の線形結合として表されるような場合にその行列の各列は 1 次独立ではない（1 次従属である）と言う。そして、その場合、$\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ は存在しない。従って、パラメタ $\boldsymbol{\beta}$ の推定値は一意には定まらない。

そこで、通常は、効果の和が 1 であるという制約条件

$$\sum_{j=1}^{A} \alpha_j = 0 \tag{97}$$

をもちいて、母数の数を一つ減らすことを考える。すると、$\alpha_A = -\sum_{j=1}^{A-1}\alpha_j$ であるから、同じモデルを

$$\begin{pmatrix} \widehat{y}_{11} \\ \widehat{y}_{21} \\ \vdots \\ \widehat{y}_{n_1 1} \\ \widehat{y}_{12} \\ \widehat{y}_{22} \\ \vdots \\ \widehat{y}_{1A} \\ \vdots \\ \widehat{y}_{n_A A} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & -1 & -1 & \dots & -1 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & -1 & -1 & \dots & -1 \end{bmatrix} \times \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{A-1} \end{pmatrix} \tag{98}$$

と書くこともできるが、このデザイン行列の各列は 1 次独立となる。

この二つのセル平均の表現

$$\widehat{y_{ij}} = \mu_j \quad \text{and} \quad \widehat{y_{ij}} = \mu + \alpha_j, \quad \sum_{j=1}^{A}\alpha_j = 0 \tag{99}$$

を比べると、

$$\mu = \frac{1}{A}\sum_{j=1}^{A}\mu_j \quad \text{（母平均の平均）} \tag{100}$$

$$\alpha_j = \mu_j - \mu \quad \text{（母平均の } \mu \text{ からのズレ）} \tag{101}$$

という関係があることがわかる。

2 元配置の場合のモデルは

$$y_{ijk} = \widehat{y}_{ijk} + \varepsilon_{ijk}, \quad \widehat{y}_{ijk} = \mu + \alpha_j + \beta_k, \quad i = 1, 2, \dots, n_{ij},\ j = 1, 2, \dots, A\ k = 1, 2, \dots, B \tag{102}$$

であるが、全ての n_{ij} を 1 とした場合のデザイン行列は

$$\begin{pmatrix} \widehat{y}_{111} \\ \widehat{y}_{112} \\ \vdots \\ \widehat{y}_{11B} \\ \widehat{y}_{121} \\ \widehat{y}_{122} \\ \vdots \\ \widehat{y}_{1A1} \\ \vdots \\ \widehat{y}_{1AB} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 1 \\ 1 & 0 & 1 & \dots & 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 0 & \dots & 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 \end{bmatrix} \times \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_A \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_B \end{pmatrix} \tag{103}$$

となる。このデザイン行列の場合、要因 A に対応する列の和、並びに要因 B に対応する列の和がそれぞれ単位ベクトルになり、二つの線形従属関係がある。従って、パラメタの推定値は一意には定まらない。なお、このデザイン行列を $\boldsymbol{X} = \{\mathbf{1},\ \boldsymbol{A},\ \boldsymbol{B}\}$ と分割し、すべてのデータの分だけ生成すると、二つの要因の水準間の同時度数行列 $\boldsymbol{F}$ は

$$\boldsymbol{F} = \boldsymbol{A}'\boldsymbol{B} \tag{104}$$

という AxB の行列で与えられる。

通常の制約条件 $\sum_{j=1}^{A}\alpha_j=0,\ \sum_{k=1}^{B}\beta_k=0$ を入れると、$\beta_B=-\sum_{k=1}^{B-1}\beta_k$ であるから、同じモデルを

$$\begin{pmatrix} \widehat{y}_{111} \\ \widehat{y}_{112} \\ \vdots \\ \widehat{y}_{11B} \\ \widehat{y}_{121} \\ \widehat{y}_{122} \\ \vdots \\ \widehat{y}_{1A1} \\ \vdots \\ \widehat{y}_{1AB} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 1 & 0 & \dots & 0 & -1 & -1 & \dots & -1 \\ 1 & 0 & 1 & \dots & 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & -1 & -1 & \dots & -1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & -1 & -1 & \dots & -1 & -1 & -1 & \dots & -1 \end{bmatrix} \times \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{A-1} \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{B-1} \end{pmatrix} \tag{105}$$

と書くこともできる。このデザイン行列の各列は１次独立である。、

なお、モデルに交互作用が組み込まれている場合には、交互作用のパラメタ $\alpha\beta_{jk}$ に対応するデザイン行列の列を、 α_j と β_k に対応する $\boldsymbol{X}$ の２つの列 $\boldsymbol{x}_{(1+j)},\boldsymbol{x}_{(1+A+k)}$ の要素毎の積を要素として持つベクトルとして定義していけばよい。

例えば、$A=2, B=3$ の場合には以下のようになる。

$$\boldsymbol{y}=\begin{pmatrix} \underline{\quad y \quad} \\ y_{111} \\ y_{211} \\ \vdots \\ y_{N_{11}11} \\ y_{112} \\ y_{212} \\ \vdots \\ y_{N_{12}12} \\ y_{113} \\ y_{213} \\ \vdots \\ y_{N_{13}13} \\ y_{121} \\ y_{221} \\ \vdots \\ y_{N_{22}21} \\ y_{122} \\ y_{222} \\ \vdots \\ y_{N_{22}22} \\ y_{123} \\ y_{223} \\ \vdots \\ y_{N_{23}23} \end{pmatrix} \quad \boldsymbol{X}=\left[\begin{array}{cccccc} \mu & \alpha_1 & \beta_1 & \beta_2 & \alpha\beta_{11} & \alpha\beta_{12} \\ \hline 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & -1 & -1 & 1 & 1 \end{array}\right] \quad \boldsymbol{\beta}=\begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ \alpha\beta_{11} \\ \alpha\beta_{12} \end{pmatrix} \tag{106}$$

ただし、たとえば y_{113} の要素のところ入っている -1 は、本来は $y_{113}=\mu+\alpha_1+\beta_3+\alpha\beta_{13}+e_{113}$ であるが、そこに、制約条件を使って $\beta_3=-\beta_1-\beta_2$ を代入したためである。同様にまた、 y_{121} の要素のところ入っている -1 は、 $\alpha_2=-\alpha_1$ を代入したためである。また、交互作用 $\alpha\beta_{ij}$ に対応するデザイン行列の列は α_1 と $\beta_j, j=1,2$ に対応するデザイン要列の列の積として定義されている。

２要因モデルの場合のセル平均と効果を表す母数の関係は、交互作用の効果に関しても

$$\sum_{j=1}^{A}\alpha\beta_{jk}=\sum_{k=1}^{B}\alpha\beta_{jk}=0 \tag{107}$$

を仮定すると、

$$\mu = \frac{1}{AB}\sum_{j=1}^{A}\sum_{k=1}^{B}\mu_{jk} \quad \text{（母平均の平均）} \tag{108}$$

$$\alpha_j = \mu_{j.}-\mu \quad \text{（行平均の } \mu \text{ からのズレ）} \tag{109}$$

$$\beta_k = \mu_{.k}-\mu \quad \text{（列平均の } \mu \text{ からのズレ）} \tag{110}$$

$$\alpha\beta_{jk} = \mu_{jk}-(\mu+\alpha_j+\beta_k) \quad \text{（加算モデルからのズレ）} \tag{111}$$

となる。ただし、

$$\mu_{j.} \quad = \quad \frac{1}{B}\sum_{k=1}^{B}\mu_{jk}, \quad \text{（行平均）} \quad \mu_{.k} \; = \; \frac{1}{A}\sum_{j=1}^{A}\mu_{jk} \quad \text{（列平均）} \tag{112}$$

である。

8 実験計画・分散分析

8.1 用語

- 要因（因子）factor
  実験結果に影響を与えると考えるもの。（名義尺度の説明変数）
- 水準 level
  要因として用いられた変数の値
- 処理 treatment
  要因の各水準の組み合わせ
- 効果 effect 各要因の各水準の効果を数値化したもの
- 実験計画 design of experiment, n 元配置デザイン
  n は要因の数
- 被験者間要因 between subject factor、被験者内要因 within subject factor
  その要因の各水準が異なる被験者に割り当てられるか否か。
  被験者内要因は繰り返し測定、反復測定、対応のある測定、対比較とも呼ばれる。
- 繰り返し
  各処理の下での測定数。
- 平方和 Sum of Squares (SS)、平均平方等 Mean Square (MS)
  F 統計量を計算するために使われる統計量で、平均平方は平方和を対応する自由度で割ったもの。
  F 統計量は平均平方の比となる。
- 自由度 degrees of freedom
  平均平方の分母で、主効果の自由度は要因の水準の数 - 1。
- 平方和、自由度の分解
  恒等式による平方和等の加算的分解。
- 主効果 main effect、交互作用 interaction effect
- 固定効果（母数モデル） fixed effect, 変量効果（変量モデル） random effect
- 線形モデル linear model, 加算モデル additive model
- crossed, nested factors
- 多重比較 multiple comparison
- フィッシャーの３原則 （反復、無作為化、局所管理）
- 制御因子、標示因子、ブロック因子

8.2 １元配置分散分析

要因 A					
水準 1	水準 2	水準 3		水準 A	
μ_1	μ_2	μ_3		μ_A	μ

8.2.1 モデル

$$
\begin{aligned}
Y_{ij} &= \mu_j + \varepsilon_{ij} && (113)\\
&= \mu + \alpha_j + \varepsilon_{ij} && (114)\\
\varepsilon_{ij} &\sim N(0, \sigma_e^2) && (115)
\end{aligned}
$$

ただし

$$
\begin{aligned}
\mu &= \frac{1}{A}\sum_{j=1}^{A}\mu_j \qquad \text{or} \qquad \mu = \frac{1}{\sum_{j=1}^{A} n_j}n_j\mu_j && (116)\\
\alpha_j &= \mu_j - \mu && (117)
\end{aligned}
$$

8.2.2 検定仮説

$$
\begin{aligned}
H_0 &: \mu_1 = \mu_2 = \ldots = \mu_A && (118)\\
H_0 &: \alpha_1 = \alpha_2 = \ldots = \alpha_A = 0 && (119)
\end{aligned}
$$

8.2.3 データ、自由度、平方和の分解

$$
\begin{aligned}
Y_{ij} - \mu &= \mu_j - \mu \quad + \quad Y_{ij} - \mu_j && (120)\\
Y_{ij} - \overline{Y_{..}} &= \overline{Y_{.j}} - \overline{Y_{..}} \quad + \quad Y_{ij} - \overline{Y_{.j}} && (121)\\
\sum_{j=1}^{A}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y_{..}})^2 &= \sum_{j=1}^{A} n_j(\overline{Y_{.j}} - \overline{Y_{..}})^2 \quad + \quad \sum_{j=1}^{A}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y_{.j}})^2 && (122)\\
SS_{total} &= SS_A \quad + \quad SS_{error} && (123)\\
&= SS_{between} \quad + \quad SS_{within} && (124)\\
\sum_{j=1}^{A} n_j - 1 &= A - 1 \quad + \quad \sum_{j=1}^{A} n_j - A && (125)\\
df_{total} &= df_A \quad + \quad df_{error} && (126)\\
&= df_{between} \quad + \quad df_{within} && (127)
\end{aligned}
$$

8.2.4 分散分析表

source	SS	df	MS	EMS	F
A	SS_A	df_A	$\frac{SS_A}{df_A}$	$\sigma_e^2 + \frac{\sum_{j=1}^{A} n_j\alpha_j^2}{A-1}$	$\frac{MS_A}{MS_e}$
error	SS_{error}	df_{error}	$\frac{SS_{error}}{df_{error}}$	σ_e^2	
total	SS_{total}	df_{total}			

8.3 2元配置分散分析 （全ての n_{jk} が等しい、バランスが取れている、直交している場合）

		要因 B (k)					
		水準 1	水準 2	水準 3		水準 B	
要因 A (j)	水準 1	μ_{11}	μ_{12}	μ_{13}		μ_{1B}	$\mu_{1.}$
	水準 2	μ_{21}	μ_{22}	μ_{23}		μ_{2B}	$\mu_{2.}$
	水準 A	μ_{A1}	μ_{A2}	μ_{A3}		μ_{AB}	$\mu_{A.}$
		$\mu_{.1}$	$\mu_{.2}$	$\mu_{.3}$		$\mu_{.B}$	μ

8.3.1 モデル

$$
\begin{aligned}
Y_{ijk} &= \mu_{jk} + \varepsilon_{ijk} && (128)\\
&= \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + \varepsilon_{ijk} && (129)\\
\varepsilon_{ijk} &\sim N(0, \sigma_e^2) && (130)
\end{aligned}
$$

ただし

$$
\begin{aligned}
\mu &= \frac{1}{AB}\sum_{j=1}^{A}\sum_{k=1}^{B}\mu_{jk} && (131)\\
\alpha_j &= \mu_{j.} - \mu && (132)\\
\beta_k &= \mu_{.k} - \mu && (133)\\
\alpha\beta_{jk} &= \mu_{jk} - (\mu + \alpha_j + \beta_k) \;=\; \mu_{jk} - \mu_{j.} - \mu_{.k} + \mu && (134)
\end{aligned}
$$

（交互作用は加算モデルからのズレを表している。）

8.3.2 検定仮説

$$
\begin{aligned}
H_{0A} &: \alpha_1 = \alpha_2 = \ldots = \alpha_A = 0 \quad (\mu_{1.} = \mu_{2.} = \ldots = \mu_{A.}) && (135)\\
H_{0B} &: \beta_1 = \beta_2 = \ldots = \beta_B = 0 \quad (\mu_{.1} = \mu_{.2} = \ldots = \mu_{.B}) && (136)\\
H_{0AB} &: \alpha\beta_{11} = \alpha\beta_{12} = \ldots = \alpha\beta_{AB} = 0 && (137)
\end{aligned}
$$

8.3.3 データ、自由度、平方和の分解

$$
\begin{aligned}
Y_{ijk} - \mu &= \mu_{jk} - \mu \;+\; Y_{ijk} - \mu_{jk} && (138)\\
&= \mu_{j.} - \mu \;+\; \mu_{.k} - \mu \;+\; \mu_{jk} - \mu_{j.} - \mu_{.k} + \mu \;+\; Y_{ijk} - \mu_{jk}\\
Y_{ijk} - \overline{Y_{...}} &= \overline{Y_{.jk}} - \overline{Y_{...}} \;+\; Y_{ijk} - \overline{Y_{.jk}} && (139)\\
&= \overline{Y_{.j.}} - \overline{Y_{...}} \;+\; \overline{Y_{..k}} - \overline{Y_{...}} \;+\; \overline{Y_{.jk}} - \overline{Y_{.j.}} - \overline{Y_{..k}} + \overline{Y_{...}} \;+\; Y_{ijk} - \overline{Y_{.jk}}\\
\sum_{j=1}^{A}\sum_{k=1}^{B}\sum_{i=1}^{n}(Y_{ijk} - \overline{Y_{...}})^2 &= nB\sum_{j=1}^{A}(\overline{Y_{.j.}} - \overline{Y_{...}})^2 \;+\; nA\sum_{k=1}^{B}(\overline{Y_{..k}} - \overline{Y_{...}})^2 \;+ && (140)\\
&\quad n\sum_{j=1}^{A}\sum_{k=1}^{B}(\overline{Y_{.jk}} - \overline{Y_{.j.}} - \overline{Y_{..k}} + \overline{Y_{...}})^2 \;+\; \sum_{i=1}^{n}\sum_{j=1}^{A}\sum_{k=1}^{B}(Y_{ijk} - \overline{Y_{.jk}})^2 && (141)\\
SS_{total} &= SS_A \;+\; SS_B \;+\; SS_{AB} \;+\; SS_{error} && (142)\\
nAB - 1 &= A - 1 \;+\; B - 1 \;+\; (A-1)(B-1) \;+\; AB(n-1) && (143)\\
df_{total} &= df_A \;+\; df_B \;+\; df_{AB} \;+\; df_{error} && (144)
\end{aligned}
$$

8.3.4 分散分析表

source	SS	df	MS	EMS	F
A	SS_A	df_A	$\frac{SS_A}{df_A}$	$\sigma_e^2 + \frac{nB\sum_{j=1}^{A}\alpha_j^2}{A-1}$	$\frac{MS_A}{MS_e}$
B	SS_B	df_B	$\frac{SS_B}{df_B}$	$\sigma_e^2 + \frac{nA\sum_{k=1}^{B}\beta_k^2}{B-1}$	$\frac{MS_B}{MS_e}$
AB	SS_{AB}	df_{AB}	$\frac{SS_{AB}}{df_{AB}}$	$\sigma_e^2 + \frac{n\sum_{j=1}^{A}\sum_{k=1}^{B}\alpha\beta_{ij}^2}{(A-1)(B-1)}$	$\frac{MS_{AB}}{MS_e}$
error	SS_{error}	df_{error}	$\frac{SS_{error}}{df_{error}}$	σ_e^2	
total	SS_{total}	df_{total}			