第 1 講　パネルデータ分析の歴史

2　統計学の歴史

統計学（statistics）は、初めから理論体系があったわけではなく様々な実践的な必要に応じて少しずつ蓄積されてきた知識が合流して出来てきたものである。その源流と考えられるものには以下のような分野がある。[1]

(1) ゲーム／ギャンブルに起源を発する確率論（ラプラス）
(2) 軍隊や財政を管理するうえで国家が必要とした統計（エジプト中国ペティ、近代ドイツ国勢学）
(3) 地中海貿易における海上保険の計算 (リスク)
(4) 17 世紀のペストの流行以来の死亡率表の研究（ジョン・グラント）
(5) 天文観測で生じる観測誤差の理論（ガウス、アイリー、ポアンカレ）
(6) 生物等の世代間の相関関係の理論（ゴルトン）
(7) 農学での実験計画理論（フィッシャー）

2. の国家政策の必要上導入された統計は**記述統計学**に相当し[2]、(1)(3)(4) は**確率論**を発展させた。(5)(6)(7) が 19 世紀半ば以降、とくに 20 世紀に入って急速に進歩した統計的推測に関わる分野である。

統計学が学問として進歩するためにはデータとしての統計資料を揃える必要がある。後には全体の一部を切り出してきて全体の統計的性質を知るという**標本理論**が出来てきたがそのためにも定期的に全体の統計量を調べられる必要がある。これは**全数調査 (悉皆調査)** と呼ばれ**国勢調査**がその代表的なものである。

パネルデータ分析の起源は (5) と (7) にあり、とりわけロナルド・フィッシャーが確立した分散分析法がその直接の起源となっている。

政治算術、国勢学、ガウスの正規分布論を総合したのがケトレー (1796 － 1874) であり、**大量観察**に基づいて平均を見つける**普遍的法則**を発見した。

人間集団に限らず物理的観察生物学的観察など全てにおいて観察を重ねデータを蓄積することによってある種の法則性を探索するのが統計学の役割であると考えられている。

ゴルトンは、スイートピーの種子の直径の測定では、親を x 軸に子を y 軸にとると、直接の傾きはおおよそ 1/3 となり、親がばらつくほど子はばらつかない、全体として平均に退行して（回帰して）ゆく法則があることが発見された。「回帰」や「相関」が法則性の表現として始めて意識的に用いられた

[1] 本節の議論は東京大学教養学部統計学教室 (編)(1991) 第 1 章を参照している。
[2] もともと statistics とは今日の統計学を指すというより、国家（state）の状況を歴史的に記述する官庁統計的な意味が強かった。

概念である。[3]

データに則して考えると観察された量には個人差によるばらつきとその背後にあるより一般的な理論上の分布がある。個人差による固定的誤差を**系統誤差**と呼び純粋な確率的誤差と区別すべきであることが認識されこれがパネルデータ分析における誤差構成要素モデルの基本的な考え方につながっている。

2.1 確率論から統計的推測へ

2.1.1 ポアンカレ『科学と仮説』（岩波文庫）第 11 章「確率論」

「法則に関する、つまり不完全な知識に基づいて、或る事象を推察しようと努める代わりに、事象を知っていて法則を推察しようと努めることもよく起こる。原因から結果を導き出すかわりに、結果から原因を導き出そうというのである。それは**原因の確率**といわれる問題であって、科学の応用の見地からすれば最も関心を引くものである。」(p.221)[4]

私は実験的な一つの法則を決定しようとする。この法則を私が知っていれば一つの曲線に表すことができるのである。私は孤立した観察を或る数だけ行う。この一つ一つは一点で表される。私がこれらの相異なる点を得たときにはこれらの点を通る一つの曲線を描く。このさい私はこれらの点からできるだけそれないようにしかもこの曲線に規則的な形を保つようにすなわち角のある点や余り調子の強い彎曲や曲率半径の急激な変化のないように努力する。この曲線は私には確からしい法則を表している。かつ単に観測した値の

[3] このような分野は Biometry（生物測定）と呼ばれ、近代統計学の発展の基礎を提供した。しかしゴルトンは個人の遺伝にも関心をもち、今日でいう多くのパネルデータや世代間の遺伝情報を集め分析した。

[4] ベイズの定理（Bayes' theorem）。A が得られた結果 $H_1, H_2, \ldots H_k$ を原因とする。われわれが知りたいのは A が起こったときに原因が H_i である確率すなわち $P(H_i \mid A)$ であるが、われわれが知ることができるのは原因に対する**結果の確率**$P(A \mid H_i)$ である。ベイズの定理は結果に対する**原因の確率**$P(H_i \mid A)$ を計算する公式を与える。

$H_1, H_2, \ldots H_k$ は互いに背反で、かつ $H_i \cup H_2 \cup \ldots \cup H_k = \Omega$ のごとく全ての場合をつくしているとする。このとき、規則

$$P(H_i \mid A) = \frac{P(H_i) \cdot P(A \mid H_i)}{\sum_j P(H_j) \cdot P(A \mid H_j)}$$

が成り立つ。ここで $P(H_i)$ は H_i の事前確率（prior probability）、$P(H_i \mid A)$ は事後確率（posterior probability）と呼ばれる。

（証明）

$$P(H_i \mid A) = \frac{P(H_i \cap A)}{P(A)} = \frac{P(A \mid H_i) \cdot P(H_i)}{P(A)}$$

ここで

$$\begin{aligned} A &= A \cap \Omega \\ &= A \cap (H_i \cup H_2 \cup \ldots \cup H_k) \\ &= (A \cap H_i) \cup (A \cap H_2) \cup \ldots \cup (A \cap H_k) \end{aligned}$$

それぞれ互いに背反であるから、

$$P(A) = \sum_j P(A \cap H_j) = \sum_j P(H_j) \cdot P(A \mid H_j)$$

これを分母に代入すればよい。

中間にある函数の値を知らせるばかりでなく観測した値でさえも直接観測したより正確に知らせるのだと私は認める。」(pp.234-235)

「これは原因の確率 P の一つの問題である。結果というのは私の記録した測定である。これらは二つの原因すなわち現象の真の法則と観測の誤差の組み合わせに依存している。結果を知ればその現象がこれこれの法則に従う確率と観測がこれこれの誤差を受けた確率とを求めることが問題となる。この場合最も確からしい法則は、えがかれた曲線に相当するし、観測の最も確からしい誤差はその観測に相当する点とこの曲線との距離で表される。(p.235)

「誤差論は直接に原因の確率の問題と結びついている。ここでもまたわれわれは結果から、すなわち一定数の互いに一致しない観測を確認して、原因を推察しようとつとめる。原因というのは一方では測定すべき量の真の値であり、もう一方では個々の観測で行った誤差である。計算すべきことは、一つ一つの誤差の確からしい大きさというものは事象の起こった後から考えてどんなものか。従って測定すべき量の確率値はどんなものかということである。(p.236)

「われわれの結論すべきは何であるか。相変わらず最小二乗法の応用を続けなくてはならないのか。われわれは次のことをはっきりさせておかなければならない、すなわちわれわれは懸念し得るあらゆる系統的誤差を除去したこと、われわれはなおその他の誤差が存在していることをよく知っているが、それを発見することができないこと、しかも決断して確からしい値と認められるような決定的な値を採用しなければならないこと、それにはわれわれのできる**最善はすなわちガウスの方法**[5]を適用することであるのは明らかである。われわれは主観的確率に関する実際的な一つの規則を適用したに過ぎない。これに対して何も文句をいうことはない。しかしさらに進んで、ただ確からしい値はこれこれであるというだけではなく、結果におよぼした確からしい誤差はこれこれだと断定しようとする。それは絶対に不法である。それ

[5]ガウス分布　ガウスは天文学の観測データを数学的に分析するに際して、データの測定誤差がある基本的な法則に従うことを仮定して誤差理論を確立した。この基本的法則というのが誤差関数 (error function) であり、今日正規分布と呼ばれているものの原型である。

正規分布の密度関数は

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-(x-\mu)^2 / 2\sigma^2\right) \quad -\infty < x < \infty$$

で与えられ、定数 $1/\sqrt{2\pi}\sigma$ は

$$\int_{-\infty}^{\infty} \exp\left(-(x-\mu)^2 / 2\sigma^2\right) dx = \sqrt{2\pi}\sigma$$

からきており、$\int_{-\infty}^{\infty} \exp f(x)dx = 1$ と標準化するために与えられている。

確率変数 x が正規分布に従っているとき、その期待値は

$$E(x) = \int_{-\infty}^{\infty} x(1 \cdot /\sqrt{2\pi}\sigma) \exp\left(-(x-\mu)^2 / 2\sigma^2\right) dx = \mu$$

であり、分数は

$$V(x) = \int_{-\infty}^{\infty} (x-\mu)^2 (1/\sqrt{2\pi}\sigma) \exp\left(-(x-\mu)^2 / 2\sigma^2\right) dx = \sigma^2$$

となる。

このことから、x を平均 μ、分散 σ^2 の正規分布といい、$N(\mu, \sigma^2)$ で表す。

はあらゆる系統的誤差を除去してしまったことが確かでなければ真でないし、われわれは除去したかどうかについては絶対に何も知らないからである。われわれが二系列の観測を有していて、これに最小二乗法を適用して、第一の系列に対する確からしい誤差は第二に対するものの半分であることを見出す。しかも第二の系列の方が第一のよりもよいことがあり得る。なぜかといえば、第一のは、ことによると大きな或る系統的誤差が影響しているかも知れないからである。われわれのいえることは第一の系列は第二よりもよいらしい、その付随的誤差が小さいからというだけであって、一方の系列に対する系統的誤差がもう一つの系列に対するものよりも大きいと断定する根拠は少しもない。われわれのこのことに関する無知は絶対的だからである。（pp.238-39）

3 推測統計の思想

主な推計方法

Gauss, *Laplace*	⇒	最小二乗法 (ordinary least square)
Karl Pearson	⇒	積率法 (method of moment)
Ronald Fisher	⇒	最尤法 (maximum likelihood)

3.1 最小二乗法

3.2 積率法

母集合のパラメータの有効推計（大標本推計)

母集合の moments（平均、分散 etc.）を標本集合の moments と一致させることで未知の母集合のパラメータを推計する。

母集合が $f(x \mid \theta)$ という密度関数に従っており、θ が未知のパラメータであるとする。母集合から抽出されたサンプルの期待値は次のように表せる。

$$E(x) = \int_{-\infty}^{\infty} x f(x \mid \theta) dx = g(\theta) \tag{1}$$

すなわち、期待値 $E(x)$ は θ の関数として表せる。これを θ について解くと、

$$\theta = g^{-1}(E(x)) \tag{2}$$

$E(x)$ の代わりにサンプル平均 $\bar{x}$ を代入してやると、

$$\hat{\theta} = g^{-1}(\bar{x}) \tag{3}$$

となる。

もしサンプルが母集合からの無作為抽出であるとすれば、$\hat{\theta}$ は θ と一致すると考えられる。

3.3 最尤法

上の 2 つの方法に比べてモデルが線形モデルであれば、最小二乗法を用いてパラメータを求めることができる。

観察されたサンプル $Y_1, Y_2 \ldots Y_n$ が平均 μ_Y、分散 $var(Y) = \sigma_Y^2$ に従う分布から発生したと考えよう。サンプルは次のような線形モデルに従っていると考えられる。

$$Y_i = \mu_Y + \varepsilon_i \quad E(\varepsilon_i) = 0,\ var(\varepsilon_i) = \sigma_Y^2$$

$E(Y_i) = \mu_Y$、誤差の流列は次のように表せる。

$$e_i = Y_i - \mu_Y \quad i = 1, \ldots, n$$

これを最小化するように μ_Y を選ぶ。すなわち、

$$\min \sum_{i=1}^{n} e_i^2 = \min \sum_{i=1}^{n} (\varphi_{i-}\mu_Y)^2$$

これが最小二乗法であり、残差平方和 (sum of squared errors, SSE) を μ_Y で微分すると

$$\frac{\partial SSE}{\partial \mu_Y} = \frac{d \sum_{i=1}^{n} (\varphi_{i-}\mu_Y)^2}{d\mu_Y} = -2 \sum_{i=1}^{n} (Y_i - \hat{\mu}_Y) = 0$$

$$\Rightarrow \sum_{i=1}^{n} Y_i = n\hat{\mu}_Y \Rightarrow \hat{\mu}_Y = \frac{\sum_{i=1}^{n} Y_i}{n}$$

ここではサンプル平均 $\hat{\mu}_Y$ が最小二乗法推計値となっている。

4 分散分析

2 標本問題[6]

ある製品の品質を他の製品と比較したり、ある薬の効果を他の薬の効果と比較するということは日常的に行われている。

実験計画法では、2 つの標本を扱う場合に、**処理群**(treatment group) と**対照群**(control group) に分けて、その違いを観察する**対象実験**(coutrolled experiment) を行う。

比較の対象となる統計量は平均と平方和である。

$$\bar{Y}_1 = \sum_{j} Y_{1j} / r_1 \quad \bar{Y}_2 = \sum_{j} Y_{2j} / r_2 \tag{4}$$

[6] 東京大学教養学部統計学教室（編）『自然科学の統計学』（東京大学出版会）第 3 章に依拠している。

$$S_1 = \sum_j (Y_{1j} - \bar{Y}_1) \quad S_2 = \sum_j (Y_{2j} - \bar{Y}_2)^2 \tag{5}$$

ここで $\bar{Y}_1$,$\bar{Y}_2$ はそれぞれ $N(\mu_1, \sigma_1^2 / r_1)$, $N(\mu_2, \sigma_2^2 / r_2)$ に従う。また S_1 / σ_1^2, S_2 / σ_2^2 はそれらと独立にそれぞれ自由度 $r_1 - 1$, $r_2 - 1$ のカイ二乗分布に従う。$n = r_1 + r_2$ とする。母集団の分数は次のように表される。

$$\hat{\sigma}_1^2 = S_1 / (r_1 - 1) \quad \hat{\sigma}_2^2 = S_2 / (r_2 - 1) \tag{6}$$

母平均の差の検定

2 つの母平均が等しいという帰無仮説

$$H_0 : \mu_1 = \mu_2$$

を検定する。

対立仮説は

両側対立仮説　$H_1 : \mu_1 \neq \mu_2$

片側対立仮説　$H_2 : \mu_1 > \mu_2$

などが考えられる。

検定方法は分散か既知か未知か未知のとき等しいかどうかにより 3 つに分かれる。

母分散既知の場合

標本平均の差 $\bar{Y}_1 - \bar{Y}_2$ は平均 $\mu_1 - \mu_2$、分散 $(\sigma_1^2 / r_1) + (\sigma_2^2 / r_2)$ の正規分布に従う。標準化変数

$$Z = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{(\sigma_1^2 / r_1) + (\sigma_2^2 / r_2)}} \tag{7}$$

も標準正規分布に従う。$\bar{Y}_1 - \bar{Y}_2$ の評価は Z の分母を基準として行い σ_1^2, σ_2^2 が既知ならば Z も計算できる。

母分散未知だが等しい場合

母分散 σ_1^2, σ_2^2 が未知でも等しいと考えられる場合にはそれを σ^2 とおく。$\bar{Y}_1 - \bar{Y}_2$ は平均 $\mu_1 - \mu_2$, 分数 $(1 / r_1 + 1 / r_2)\sigma^2$ の正規分布に従う。他方 $(S_1 + S_2) / \sigma^2$ が自由度 $(r_1 - 1) + (r_2 - 2) = n - 2$ のカイ二乗分布に従うので σ^2 は併合推定量（pooled variance）

$$\hat{\sigma}^2 = (S_1 + S_2) / (n - 2) \tag{8}$$

で不偏推定できる。

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{(1 / r_1 + 1 / r_2)\hat{\sigma}^2}} \tag{9}$$

は自由度 $n-2$ の t 分布に従う。

母分散未知で等しいとも考えられない場合

（7）の右辺の σ_1^2, σ_2^2 を推定量 (6) で置き換えると

$$t' = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{(\hat{\sigma}_1^2 / r_1) + (\hat{\sigma}_2^2 / r_2)}} \tag{10}$$

となり、自由度

$$\frac{\left((\sigma_1^2 / r_1) + (\sigma_2^2 / r_2)\right)^2}{\sigma_1^4 / r_1^2 (r_1 - 1) + \sigma_2^4 / r_2^4 (r_2 - 1)} \tag{11}$$

として、ν と ν^* に最も近い整数として t 検定を行う。これをウェルテ検定（Welch's Test）と呼ぶ。

分散比の推測

2 つの母分散が等しいという仮説 (等分散仮説)

$$H_0 : \sigma_1^2 = \sigma_2^2$$

を検定する。

S_i / σ_i^2 は自由度 $r_i - 1$ のカイ二乗分布に従うので

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{S_1 / (r_1 - 1)}{S_2 / (r_2 - 1)} \tag{12}$$

の分布は自由度 $\nu_1 = r_1 - 1$、$\nu_2 = r_2 - 1$ の F 分布となる。

対立仮説が $\sigma_1^2 \neq \sigma_2^2$ の場合 $F > F_{\frac{\alpha}{2}}(v_1, v_2) \text{or} F < 1 / F_{\frac{\alpha}{2}}(v_1, v_2)$

対立仮説が $\sigma_1^2 > \sigma_2^2$ の場合、$F > F_\alpha(v_1, v_2)$

対立仮説が $\sigma_1^2 < \sigma_2^2$ の場合 $F < 1 / F_\alpha(v_2, v_1)$

を棄却域とすることで有意水準 α の検定ができる。

一元配置分散分析

3 つ以上の母集団平均 $\mu_1, \mu_2, \ldots \mu_a \quad (a \geq 3)$ の比較には分散分析（analysis of variance: ANOVA）を用いる。

実験結果に影響を及ぼすと考えられる変数を**因子 (factor)** と呼び因子に対して与える条件を**水準 (level)** と呼ぶ。因子と水準を組み合わせて実験を行うことを**処理 (treatment)** と呼ぶ。

処理の比較を目的とする実験では因子として取り上げていないさまざまな要因による**系統誤差**が比較に偏りを生じないように注意する必要がある（完全無作為化法（completely randomized design)、乱魂法 (randomized block design) などを用いる)。

因子を A、水準 $A_1, \ldots, A_a$ を、繰り返し数を $r_1, \ldots, r_a$ とすると、A_i 水準の j 番目のデータを y_{ij} とし次のモデルを想定する。

$$y_{ij} = \mu_i + \varepsilon_{ij} \quad i = 1, 2, \ldots, a; \quad j = 1, \ldots, r_i \tag{13}$$

母数 μ_i は第 i 水準に固有な平均であり実験誤差 ε_{ij} はすべてお互いに独立に $N(0, \sigma^2)$ に従うものとする。

データの総数を $n = \Sigma r_i$ とし、繰り返し数 r_i の重みで μ_i の加重平均

$$\mu = \sum r_i \mu_i / n$$

を一般平均（grand mean）と呼ぶ。

各水準から一般平均を引いたものが正時の効果 (effect) である。

$$\alpha_i = \mu_i - \mu$$

ここで $\sum r_i \alpha_i = 0$

すると (13) 式は次のように書き換えることができる。

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, 2, \ldots ; \quad j = 1, \ldots , r_i \tag{14}$$

これは (共通の効果 μ)+（第 i 水準の効果 α_i）+（それ以外の誤差 ε_{ij}) という形式になっている。これを一元配置（one-way layout）モデルと呼ぶ。

分散分析とは因子 A の全ての水準の平均が等しいという帰無仮説

$$H_0 : \mu_1 = \mu_2 = \ldots \mu_a$$
$$or \quad H_0 : \alpha_1 = \alpha_2 = \ldots \alpha_a = 0$$

を検定する方法である。

データに一元配置モデル (13) をあてはめると残差平方和は

$$\begin{aligned} S_e &= \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 \\ &= \sum_i \sum_j y_{ij}^2 - \sum_i y_i^2 / r_i \end{aligned} \tag{15}$$

となり、S_e / σ^2 は自由度 $\nu_e = n - a$ のカイ二乗分布に従う。

仮説 $H_0 : \mu_1 = \mu_2 = \ldots \mu_a$ のもとで、モデル $y_{ij} = \mu + \varepsilon_{ij}$ をあてはめたときの残差平方和は μ を総平均 $\bar{y}$ で推定して

$$S_T = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 = \sum_i \sum_j y_{ij}^2 - \bar{y}_i^2 / n \tag{16}$$

となる。

仮説 H_0 による残差平方和の増加分は

$$\begin{aligned} S_A &= S_T - S_e \\ &= \sum_i y_i^2 / r_i - y^2 / n \\ &= \sum_i r_i (\bar{y}_i - \bar{y})^2 \end{aligned} \tag{17}$$

となる。

S_A と S_e は独立で自由度 $\nu_A = a - 1$ のカイ二乗分布に従う。
すると

$$F = \frac{S_A / \nu_A}{S_e / \nu_e} \tag{18}$$

が自由度 ν_A、ν_e の F 分布 (ν_A, ν_e) に従うことを利用して仮説検定を行うことができる。これを**分散分析検定 (ANOVA Test)** と呼ぶ。(18) 式から明らかなように平均が等しいという仮説の検定を級間平方和 (S_A) に基づく分散 (S_A / ν_A) と誤差平方和 (S_e) に基づく分散 (S_e / ν_e) の比によって検定することから**分散分析**と呼ばれている。

バートレット検定

分散分析で仮定された分散の一様性を検定する方法にバートレット検定がある。各水準の不偏分散を

$$V_i = \sum_i (y_i - \bar{y}_i)^2 / (r_i - 1) \quad i = 1, \ldots, a$$

としてそれらを併合したものを

$$V_e = \sum_c (v_i - 1) V_i / (n - a) = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 / (n - a)$$

とすると、

$$B = (n - a) \log V_e - \sum (r_i - 1) \log V_i \tag{19}$$

が等分散仮説 $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_a^2$ の下で近似的に自由度 $a - 1$ のカイ二乗分布に従うことを用いる。
このときカイ二乗の近似をよくするために次のように補正する。

$$B' = \frac{B}{1 + \frac{1}{3(a-1)} \left(\sum_i \frac{1}{r_i - 1} - \frac{1}{n-a} \right)} \tag{20}$$

交互作用

2 つ以上の因子を取り上げる場合各因子の単独の効果 (主効果 main effect) だけでなく組み合わせによる効果にも注意を払う必要がある。とりわけ因子間で**加法性**が成り立たず**相殺する効果**(交互効果) がある場合には重要である。
他因子の最適な水準組み合わせを決めるときに各因子別に最適水準を決めていきその集合を最適水準組み合わせとする方法を**単一因子実験**(one factor at a time experiment) というが、交互効果のある場合には因子の全ての水準組み合わせを考慮して最適組み合わせを求める実験 (これを**要因実験**(factorial experiment) という) を行う必要がある。

ただしこの実験では因子と水準のとり方によって実験回数が膨大な数になり実行不可能となることもあることに注意すべきである。[7]

二元配置分散分析

多因子要因実験を完全無作為化法で行うときそれを**多元配置**という。因子 AB の二つを考える二元配置モデルについて考える。

水準組み合わせ A_iB_j での k 番目の観測値をとする。AB の水準数を a, b、繰り返し数を r とする。

$$y_{ijk} = \mu_{ij} + e_{ijk} \quad i = 1, 2, \dots a; \quad j = 1, \dots, b; \quad k = 1, 2, \dots, r \tag{21}$$

μ_{ij} は水準組み合わせ A_iB_j の平均実験誤差 e_{ijk} は独立して $N(0, \sigma^2)$ に従うとする。データ総数は $n = abr$ とする。

A の第 i 水準 A_i、B の第 j 水準 B_j、全体の効果をそれぞれ次のようにおく。

$$\bar{\mu}_i = \sum_j \mu_{ij} \div b, \quad \bar{\mu}_j = \sum_i \mu_{ij} \div a, \quad \mu \sum_i \sum_j \mu_{ij} \div ab \tag{22}$$

μ を一般平均とする。A_i、B_j の効果から μ を引いた正味効果を主効果（main effect）と呼び次のように定義する。

$$\begin{aligned} \alpha_i &= \bar{\mu}_i - \mu, \quad i = 1, \dots a; \\ \beta_j &= \bar{\mu}_j - \mu, \quad j = 1, \dots b \end{aligned} \tag{23}$$

μ_{ij} のうち主効果と一般平均で表せない部分を因子 A, B の相互作用 (interaction) と呼ぶ。

$$\begin{aligned} (\alpha\beta)_{ij} &= \mu_{ij} - (\mu - \alpha_i + \beta_j) = \mu_{ij} - \bar{\mu}_i - \bar{\mu}_j + \mu \\ & i = 1, 2, \dots a; \quad j = 1, 2 \dots, b \end{aligned} \tag{24}$$

(24) に (21) を代入して

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \tag{25}$$

これは (一般平均)+(因子 A の効果)+(因子 B の効果)+(因子 AB の交互作用)+(誤差) という形になっている。

ここで $\Sigma\alpha_i = 0$, $\Sigma\beta_j = 0$, $\sum_i (\alpha\beta)_{ij} = 0$, $j = 1, \dots b$; $\sum_j (\alpha\beta)_{ij} = 0$, $i = 1, \dots a$ である。

一元配置モデルの場合と同様平方和 S_A, S_B, S_{AB} の大きさを S_e を基準として判断するためにはそれぞれの自由度を考慮しなければならない。

自由度は

$$v_T = n - 1, \quad v_A = a - 1, \quad v_B = b - 1, \quad v_{AB} = (a-1)(b-1), \quad v_e = ab(r-1) \tag{26}$$

[7] このような実験組み合わせには直交表を用いて実験の回数を有効な限り最小限に切り下げることを行う。

となり各平方和を自由度で割って平均平方を求める。

$$V_A = S_A \text{Á} v_A, \quad V_B = S_B \text{Á} v_B, \quad V_{AB} = S_{AB} \text{Á} v_{AB}, \quad V_e = S_e \text{Á} v_e \quad (27)$$

各仮説検定に用いられる F 統計量は

H_0： $(\alpha\beta)_{ij} \equiv 0$ の検定は $F_{AB} = V_{AB} \text{Á} V_e$

H_0： $\alpha_i = 0$ の検定は $F_A = V_A \text{Á} V_e$ で表せる。

H_0： $\beta_j = 0$ の検定は $F_B = V_B \text{Á} V_e$

とりわけ第１の仮設は交互作用がないというものでありこれを最初に調べるべきである。

交互作用が有意でありかつ繰り返しのない二元配置モデルは次のように表せる。

$$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij} \quad (28)$$

S_e の中に交互作用と誤差の情報が入り分離できなくなる。この場合には分散分析検定は誤る可能性が出てくる。

参考文献

東京大学教養学部統計学教室（編）(1992)『自然科学の統計室』、東京大学出版会。