繰返し数の違う一元配置分散分析の不偏分散の期待値 $E(V)$ の求め方について

文責：稲葉太一（神戸大学）

Q. 質問内容

繰返し数の同じ一元配置分散分析の不偏分散の期待値 $E(V)$ の求め方については分かるのですが、繰返し数の違う場合の計算方法について教えて下さい。

A. 回答

まず、繰返し数の異なる一元配置分散分析のデータの構造式と、その制約式の意味を説明し、平方和の分解を証明してから、不偏分散（平均平方）V_A の期待値の計算をする。というのは、不偏分散の期待値の計算に、各記号の意味や制約式の意味が使われるので、この順に説明する。

A.1. 一元配置分散分析の構造式

一元配置分散分析とは、３つ以上の正規分布に従うと考えられる群の母平均に違いがあるかどうか検討する方法論である。群の数を a とおき、第 i 群から取られたデータ数を $n_i(i=1,\cdots,a)$ とおくと、データの構造式は以下のようになる。

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i=1,\cdots,a, \quad j=1,\cdots,n_i \tag{1.1}$$

ただし、$\sum_{i=1}^{a} n_i\alpha_i = 0, \quad \varepsilon_{ij} \sim N(0,\sigma^2)$.

A.2. 制約式の意味

第 i 群の第 j 番目のデータ x_{ij} が、第 i 番目の正規母集団 $N(\mu_i,\sigma^2)$ に従っていると仮定する。

$$x_{ij} \sim N(\mu_i,\sigma^2)$$

ここで、誤差 ε_{ij} を、$\varepsilon_{ij} := x_{ij} - \mu_i$ とおくと、誤差の４条件（不偏性、等分散性、独立性、正規性）が成り立つ。ここで、:= とは、「右辺で左辺を定義する」という記号である。また、総データ数を $N := \sum_{i=1}^{a} n_i$ とし、総平均 μ を

$$\mu := \frac{1}{N}\sum_{i=1}^{a} n_i\mu_i$$

とおき、$\alpha_i := \mu_i - \mu$ とすれば、(1.1) 式が制約式を含んで得られる。

A.3. 平均、平方和、自由度等の記号の定義

1) 和と平均

各群のデータの和 $x_{i\cdot}$ と平均 $\overline{x}_{i\cdot}$ は以下で計算される。

$$x_{i\cdot} := \sum_{j=1}^{n_i} x_{ij}, \quad \overline{x}_{i\cdot} := \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}, \quad (i = 1, \cdots, a)$$

次に、データの総和 $x_{\cdot\cdot}(= T)$ と総平均 $\overline{\overline{x}}$ は以下で計算される。

$$T = x_{\cdot\cdot} := \sum_{i=1}^{a}\sum_{j=1}^{n_i} x_{ij}, \quad \overline{\overline{x}} := \frac{1}{N}\sum_{i=1}^{a}\sum_{j=1}^{n_i} x_{ij}, \quad (i = 1, \cdots, a)$$

2) 平方和と自由度

$$\text{群間平方和：} S_A = \sum_{i=1}^{a}\sum_{j=1}^{n_i} (\overline{x}_{i\cdot} - \overline{\overline{x}})^2, \quad \phi_A = a - 1$$

$$\text{群内平方和：} S_E = \sum_{i=1}^{a}\sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{i\cdot})^2, \quad \phi_E = N - a$$

$$\text{総平方和：} S_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i} (x_{ij} - \overline{\overline{x}})^2, \quad \phi_T = N - 1$$

A.4. 平方和の分解

総平方和は、群内平方和と群間平方和の合計に分解される。まず、

$$x_{ij} - \overline{\overline{x}} = (x_{ij} - \overline{x}_{i\cdot}) + (\overline{x}_{i\cdot} - \overline{\overline{x}})$$

である。このことから、総平方和 S_T は、次のように展開できる。

$$\begin{aligned} S_T &= \sum_{i=1}^{a}\sum_{j=1}^{n_i} (x_{ij} - \overline{\overline{x}})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i} \{(x_{ij} - \overline{x}_{i\cdot}) + (\overline{x}_{i\cdot} - \overline{\overline{x}})\}^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{i\cdot})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n_i} (\overline{x}_{i\cdot} - \overline{\overline{x}})^2 + 2\sum_{i=1}^{a}\sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{i\cdot})(\overline{x}_{i\cdot} - \overline{\overline{x}}) \\ &= S_E + S_A + 2\sum_{i=1}^{a}\sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{i\cdot})(\overline{x}_{i\cdot} - \overline{\overline{x}}) \end{aligned}$$

ここで、最後の積和の項がゼロであるから、$S_T = S_A + S_E$ が分る。

（最後の項がゼロになる理由）

まず、$(\overline{x}_{i\cdot} - \overline{\overline{x}})$ は、j に関して定数だからくくり出せる。

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{i\cdot})(\overline{x}_{i\cdot} - \overline{\overline{x}}) = \sum_{i=1}^{a} (\overline{x}_{i\cdot} - \overline{\overline{x}}) \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{i\cdot})$$

ここで、$\sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{i\cdot})$ は、第 i 群の各データ x_{ij} とそれらの平均 $\overline{x}_{i\cdot}$ の差、即ち「偏差 $x_{ij} - \overline{x}_{i\cdot}$ 」の和だから、以下のようにゼロになる。

$$\sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{i\cdot}) = \sum_{j=1}^{n_i} x_{ij} - n_i \overline{x}_{i\cdot} = \sum_{j=1}^{n_i} x_{ij} - \sum_{j=1}^{n_i} x_{ij} = x_{i\cdot} - x_{i\cdot} = 0.$$

A.5. 不偏分散（平均平方）V_A の期待値

まず、$V_A = S_A/\phi_A$ であるから、S_A の期待値を考える。S_A は $\overline{x}_{i\cdot} - \overline{\overline{x}}$ の平方和であるから、データの構造式 (1.1) 式より、第 i 群の平均 $\overline{x}_{i\cdot}$ と全平均 $\overline{\overline{x}}$ は、

$$\overline{x}_{i\cdot} = \sum_{j=1}^{n_i} x_{ij}/n_i = \sum_{j=1}^{n_i}(\mu + \alpha_i + \varepsilon_{ij})/n_i = (n_i\mu + n_i\alpha_i + \varepsilon_{i\cdot})/n_i = \mu + \alpha_i + \overline{\varepsilon}_{i\cdot}$$

$$\overline{\overline{x}} = \sum_{i=1}^{a}\sum_{j=1}^{n_i} x_{ij}/N = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(\mu + \alpha_i + \varepsilon_{ij})/N = (N\mu + \sum_{i=1}^{a} n_i\alpha_i + \varepsilon_{\cdot\cdot})/N = \mu + \overline{\overline{\varepsilon}}$$

と表される。

次に、これらのことから S_A の期待値が計算される。つまり、

$$\overline{x}_{i\cdot} - \overline{\overline{x}} = (\mu + \alpha_i + \overline{\varepsilon}_{i\cdot}) - (\mu + \overline{\overline{\varepsilon}}) = \alpha_i + (\overline{\varepsilon}_{i\cdot} - \overline{\overline{\varepsilon}})$$

より、

$$S_A = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(\overline{x}_{i\cdot} - \overline{\overline{x}})^2 = \sum_{i=1}^{a} n_i(\overline{x}_{i\cdot} - \overline{\overline{x}})^2 = \sum_{i=1}^{a} n_i\{\alpha_i + (\overline{\varepsilon}_{i\cdot} - \overline{\overline{\varepsilon}})\}^2$$

を計算したい。ここで、以下の３つに注目すると、計算が見通し良く行える。

1) $Var(x) = E(x^2) - \{E(x)\}^2$ より、$E(x^2) = Var(x) + \{E(x)\}^2$.
2) ２つの確率変数 x, y に対して、$Var(x - y) = Var(x) + Var(y) - 2Cov(x, y)$.
3) ２つの確率変数 x, y が独立であれば、共分散 $Cov(x, y) = 0$.

群間平方和 S_A の $\{\alpha_i + (\overline{\varepsilon}_{i\cdot} - \overline{\overline{\varepsilon}})\}^2$ の期待値を計算すると

$$E\Big[\{\alpha_i + (\overline{\varepsilon}_{i\cdot} - \overline{\overline{\varepsilon}})\}^2\Big] = Var\Big\{\alpha_i + (\overline{\varepsilon}_{i\cdot} - \overline{\overline{\varepsilon}})\Big\} + \Big[E\{\alpha_i + (\overline{\varepsilon}_{i\cdot} - \overline{\overline{\varepsilon}})\}\Big]^2$$
$$= Var(\overline{\varepsilon}_{i\cdot} - \overline{\overline{\varepsilon}}) + \alpha_i^2$$

ここで、第１項は、

$$Var(\overline{\varepsilon}_{i\cdot} - \overline{\overline{\varepsilon}}) = Var(\overline{\varepsilon}_{i\cdot}) + Var(\overline{\overline{\varepsilon}}) - 2Cov(\overline{\varepsilon}_{i\cdot}, \overline{\overline{\varepsilon}})$$

であり、$\overline{\varepsilon}_{i\cdot}$ は n_i 個の平均だから $Var(\overline{\varepsilon}_{i\cdot}) = \sigma^2/n_i$, $\overline{\overline{\varepsilon}}$ は N 個の平均だから $Var(\overline{\overline{\varepsilon}}) = \sigma^2/N$ が分っている。また、

$$\overline{\overline{\varepsilon}} = \frac{1}{N}\sum_{i=1}^{a}\sum_{j=1}^{n_i}\varepsilon_{ij} = \frac{1}{N}\sum_{i=1}^{a}\varepsilon_{i\cdot} = \frac{1}{N}\sum_{i=1}^{a} n_i\overline{\varepsilon}_{i\cdot}$$

であり、群番号が違えば $\overline{\varepsilon}_{i\cdot}$ 同士は独立になるから、

$$Cov(\overline{\varepsilon}_{i\cdot}, \overline{\overline{\varepsilon}}) = Cov(\overline{\varepsilon}_{i\cdot}, \frac{1}{N}\sum_{i'=1}^{a} n_{i'}\overline{\varepsilon}_{i'\cdot}) = \sum_{i'=1}^{a}\frac{n_{i'}}{N}Cov(\overline{\varepsilon}_{i\cdot},\ \overline{\varepsilon}_{i'\cdot})$$

$$= \frac{n_i}{N}Cov(\overline{\varepsilon}_{i\cdot},\ \overline{\varepsilon}_{i\cdot}) = \frac{n_i}{N}Var(\overline{\varepsilon}_{i\cdot}) = \frac{n_i}{N}\frac{\sigma^2}{n_i} = \frac{\sigma^2}{N}$$

と計算でき、

$$Var(\overline{\varepsilon}_{i\cdot} - \overline{\overline{\varepsilon}}) = \frac{\sigma^2}{n_i} + \frac{\sigma^2}{N} - 2\frac{\sigma^2}{N} = \frac{\sigma^2}{n_i} - \frac{\sigma^2}{N}$$

と分る。

よって、

$$E(S_A) = \sum_{i=1}^{a} n_i E\{\alpha_i + (\overline{\varepsilon}_{i\cdot} - \overline{\overline{\varepsilon}})\}^2 = \sum_{i=1}^{a} n_i \left[\frac{\sigma^2}{n_i} - \frac{\sigma^2}{N} + \alpha_i^2\right]$$

$$= \sum_{i=1}^{a} \left[\sigma^2\left(1 - \frac{n_i}{N}\right) + n_i\alpha_i^2\right] = \sigma^2\left(a - \frac{1}{N}\sum_{i=1}^{a} n_i\right) + \sum_{i=1}^{a} n_i\alpha_i^2 = (a-1)\sigma^2 + \sum_{i=1}^{a} n_i\alpha_i^2.$$

したがって、$V_A = S_A/\phi_A = S_A/(a-1)$ より、

$$E(V_A) = \sigma^2 + \frac{1}{a-1}\sum_{i=1}^{a} n_i\alpha_i^2$$

が導かれる。