分散分析

第 12 講 - 複数のグループ間の違いを検証する

村田 昇

講義概要

- 分散分析とは
- 一元配置
  - 一元配置のモデル
  - 検定の構成
- 二元配置
  - 二元配置のモデル
  - 検定の構成

分散分析とは

分散分析

- 平均の差の検定
  2 つのグループ間 で平均の差があるか否か
- 分散分析
  2 つ以上のグループ間 で平均の差があるか否か
- 分散分析における仮説の例
  - ある小売店について “売上高は月によって差があるか”
  - ある銘柄の株価について “収益率は曜日によって差があるか”

基本的な考え方

- データの変動 (**分散**) =不確定性 を分解
  - グループ間での変動
  - 観測誤差のみに起因する変動
- 変動の大きさを比較
  - 平均に差がない : 自由度を除いて両者に本質的な差がない
  - 平均に差がある : グループ間の変動が増してより大きくなる
- 分散分析は **“データの変動の分析”**

分散の比の検定 (再掲)

分散の比の検定

- 2 種類のデータの分散が等しいか否かを検定する

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

考え方

- $X_1, \ldots, X_m$ および $Y_1, \ldots, Y_n$ の不偏分散: s_1^2, s_2^2
  - このとき s_1^2, s_2^2 は独立でそれぞれ
    - $(m-1)s_1^2/\sigma_1^2$ は自由度 $m-1$ の χ^2 分布に従う
    - $(n-1)s_2^2/\sigma_2^2$ は自由度 $n-1$ の χ^2 分布に従う
  - **F 検定**

$$\text{検定統計量}: F = \frac{s_1^2}{s_2^2} \left(= \frac{((m-1)s_1^2/\sigma_1^2)/(m-1)}{((n-1)s_2^2/\sigma_2^2)/(n-1)} \right)$$

    - 帰無分布は自由度 $m-1, n-1$ の F 分布

両側検定の棄却域

- 有意水準を選択 : $\alpha \in (0, 1)$
- H_0 の下では以下が成立

$$P(F < F_{\alpha/2}(m-1, n-1) \text{ または } F > F_{1-\alpha/2}(m-1, n-1)) = \alpha$$

  - 自由度 $m-1, n-1$ の F 分布
    - $F_{\alpha/2}(m-1, n-1)$: $\alpha/2$ 分位点
    - $F_{1-\alpha/2}(m-1, n-1)$: $1-\alpha/2$ 分位点
- 第一種過誤の上限が α の棄却域

$$R_\alpha = \left(-\infty, F_{\alpha/2}(m-1, n-1)\right) \cup \left(F_{1-\alpha/2}(m-1, n-1), \infty\right)$$

- データから検定統計量 F の値を計算
- 以下の場合，帰無仮説を棄却

$$F < F_{\alpha/2}(m-1, n-1) \quad \text{または} \quad F > F_{1-\alpha/2}(m-1, n-1)$$

一元配置

一元配置の問題設定

- グループ分けが 1 種類 : p グループ $A_1, A_2, \ldots, A_p$
- 各グループ i ごとに n_i 個のデータを観測

$$Y_{i1}, Y_{i2}, \ldots, Y_{in_i}, \quad (i = 1, 2, \ldots, p)$$

- 小売店の売上高の問題
  - $A_1, A_2, \ldots, A_p$: 月
  - $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$: i 月の各日における売上高

一元配置のモデル

- 分散分析の用語
  - **因子** : グループ分けのこと (月ごと)
  - **水準** : 因子内の各グループのこと (1 月，2 月，3 月，…)
- 観測データのモデル

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad (i = 1, \ldots, p;\ j = 1, \ldots, n_i).$$

- モデルの仮定
  - 水準 A_i における観測データの平均値 μ_i は定数
  - ε_{ij} は独立同分布 (平均 0, 分散 σ^2 の正規分布)

一元配置の検定

- 検定問題

  各水準 $A_1, A_2, \ldots, A_p$ の平均 $\mu_1, \mu_2, \ldots, \mu_p$ に差があるか否かを検定する
  - 帰無仮説 : 全ての水準で平均に差はない
  - 対立仮説 : 平均の異なる水準がある

$$H_0 : \mu_1 = \cdots = \mu_p \quad \text{vs} \quad H_1 : \text{ある } i, j \text{ に対して} \mu_i \neq \mu_j.$$

分析の考え方

- データの変動から因子間での変動と観測誤差の変動を抽出して比較
- 各種平均 ($n = \sum_{i=1}^{p} n_i$ は全サンプル数)

$$\text{(全データの標本平均)} \qquad \bar{Y}_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n_i} Y_{ij},$$

$$\text{(水準 } A_i \text{ の標本平均)} \qquad \bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \quad (i = 1, \ldots, p).$$

- 各種変動

$$\text{(全変動)} \qquad SS_T = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2,$$

$$\text{(級内変動)} \qquad SS_W = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2,$$

$$\text{(級間変動)} \qquad SS_B = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{p} n_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2.$$

  - **級内変動** : 各水準内でのデータの変動の合計 (観測誤差に起因)
  - **級間変動** : 水準間でのデータの変動の合計 (水準ごとのデータ数で重みづけ)
- 変動の分解

$$\underbrace{(\text{全変動})\ SS_T}_{\text{自由度 } n-1} = \underbrace{(\text{級内変動})\ SS_W}_{\text{自由度 } n-p} + \underbrace{(\text{級間変動})\ SS_B}_{\text{自由度 } p-1}$$

- 帰無仮説 H_0 が正しい場合
  - 水準内・水準間でのデータの変動はともに観測誤差が原因
  - 自由度を除けば本質的な違いはない
  - $SS_T/(n-1)$ は σ^2 の不偏推定量
  - $SS_W/(n-p), SS_B/(p-1)$ もともに σ^2 の不偏推定量
- 変動の分解

$$\underbrace{(\text{全変動})\ SS_T}_{\text{自由度 } n-1} = \underbrace{(\text{級内変動})\ SS_W}_{\text{自由度 } n-p} + \underbrace{(\text{級間変動})\ SS_B}_{\text{自由度 } p-1}$$

- 対立仮説 H_1 が正しい場合
  - 水準間での変動 SS_B は観測誤差と平均値の差に影響される
  - SS_B は SS_W より本質的に大きくなる

一元配置の検定

- 検定統計量

$$F = \frac{SS_B/(p-1)}{SS_W/(n-p)}$$

- 帰無仮説の下で次の事実が成り立つ
  - SS_B, SS_W は独立
  - SS_B は自由度 $p-1$ の χ^2 分布に従う
  - SS_W は自由度 $n-p$ の χ^2 分布に従う
- 帰無分布は自由度 $p-1, n-p$ の F 分布に従う
- 対立仮説の下 F は大きな値をとるので右片側検定

棄却域を用いる場合

- 有意水準を選択 : $\alpha \in (0,1)$
- 自由度 $p-1, n-p$ の F 分布
  - $F_{1-\alpha}(p-1, n-p)$: $1-\alpha$ 分位点
- H_0 の下で以下が成立

$$P(F > F_{1-\alpha}(p-1, n-p)) = \alpha$$

- 第一種過誤の上限が α となる棄却域

$$R_\alpha = (F_{1-\alpha}(p-1, n-p), \infty)$$

- データから検定統計量 F の値を計算
- 以下の場合，帰無仮説を棄却

$$F > F_{1-\alpha}(p-1, n-p)$$

p 値を用いる場合

- p 値を計算 (右片側検定の場合の計算方法)

$$(p \text{ 値}) = \int_F^\infty f(x)dx$$

  - $f(x)$ は自由度 $p-1, n-p$ の F 分布の確率密度

- p 値が α 未満なら帰無仮説を棄却

分散分析表 (一元配置の場合)

	自由度	平方和	平均平方和	F 値	p 値
級間	$p-1$	SS_B	$SS_B/(p-1)$	F	$\int_F^\infty f(x)dx$
級内	$n-p$	SS_W	$SS_W/(n-p)$		
全変動	$n-1$	SS_T			

- $f(x)$ は自由度 $p-1, n-p$ の F 分布の確率密度

モデルの別表現

- 各水準の平均値の相対効果による定式化
  - 因子 A 全体の平均効果 : μ
  - 平均 μ を基準とした各水準 A_i の相対的な効果 : α_i
  - 平均値の別表現

$$\mu_i = \mu + \alpha_i, \quad \mu = \frac{1}{n}\sum_{i=1}^{p} n_i\mu_i, \quad \sum_{i=1}^{p} n_i\alpha_i = 0$$

- 帰無仮説 H_0 は以下と同等

$$H_0 : \alpha_1 = \cdots = \alpha_p = 0$$

分散分析の計算

- 基本書式

```
aov(formula, data)
#' formula: 式, 一元配置の場合は (観測値 ~ 因子)
#' data: データフレーム
#' 分析の結果を参照する関数がいくつかあるので, 多くの場合適当なオブジェクトに代入
```

- 分散分析表の表示

```
aov(formula, data) |> summary()      # 分析表形式での表示
aov(formula, data) |> anova()        # データフレーム (data.frame 形式)
aov(formula, data) |> broom::tidy() # データフレーム (tibble 形式)
aov(formula, data) |> model.table(type = "means")   # 平均値
aov(formula, data) |> model.table(type = "effects") # 効果 (既定値)
```

分散分析の計算 (検定のみ)

- 基本書式

```r
oneway.test(formula, data, subset, na.action, var.equal = FALSE)
#' formula: 式
#' data: データフレーム
#' subset: subset の指定
#' na.action: 欠損値の扱い
#' var.equal: 等分散を仮定する場合は TRUE, 標準では Welch の近似が用いられる
```

実習

練習問題

- 一元配置分散分析について確率シミュレーションを行いなさい.
- 東京の気候データ (tokyo_weather.csv) の気温の項目について以下の問に答えよ.
  - 曜日ごとの気温の平均と分散を求めよ.

```r
#' 週日を因子 (factor) とするには例えば以下のような項目を加えればよい
tw_data <- read_csv("data/tokyo_weather.csv") |>
  mutate(day_of_week = ordered(day_of_week, # 曜日を順序付き因子にする
                               levels = wday(1:7, label = TRUE)))
#' 関数 wday() の出力は使用言語に依存するので, 強制的に英語にするには locale を指定して
#' wday(..., locale = "en_US") とすれば良い
#' 省略形を用いない場合は abbr を指定する. 以下はドイツ語の例
#' wday(1:7, label = TRUE, abbr = FALSE, locale = "de_DE")
#' 利用可能な locale(言語) を調べるには stringi::stri_locale_list() を用いる
```

  - 曜日ごとに平均が異なるといえるかどうか分散分析を用いて検定しなさい.

二元配置

二元配置の問題

- 因子が 2 種類ある場合
  - 一方の因子の水準の平均値に差があるか否かを検定
  - もう一方の因子の水準で平均値に差があるかは不問
- 複数の薬の治験の問題

  複数の薬の効能を複数の被験者に投与する実験
  - “薬の種類” と “被験者番号” の 2 種類の因子
  - “薬の種類” という因子での薬の効能の差を検証したい
  - “被験者番号” という因子で効能に差があることは許容したい (薬の効き目には個人差があると考えられるため)

二元配置のモデル

- 2 種類の因子 A, B
  - 因子 A には a 個の水準 : $A_1, \ldots, A_a$
  - 因子 B には b 個の水準 : $B_1, \ldots, B_b$
  - 因子 A, B の水準がそれぞれ A_i, B_j である観測値 : Y_{ij}
  - 各水準ごとに 1 つの観測値が得られる一番簡単な場合を想定
- 観測データのモデル

$$Y_{ij} = \alpha_i + \beta_j + \varepsilon_{ij} \quad (i = 1, \ldots, a;\ j = 1, \ldots, b).$$

- モデルの仮定
  - 因子 A, B の水準 A_i, B_j における効果 α_i, β_j は定数
  - ε_{ij} は独立同分布 (平均 0, 分散 σ^2 の正規分布)

二元配置の例

- 複数の薬の治験の問題
  - 因子
    - 因子 A : “薬の種類”
    - 因子 B : “被験者番号”
  - 効果
    - α_i : 薬 A_i の効能
    - β_j : 被験者 B_j 固有の薬の効きやすさ
  - 薬の効能に差があるか否かという検定

二元配置の検定

- 検定問題

  因子 A の各水準の効果に差があるか否かを検定する

  (因子 B の効果は除いて検定したい)
  - 帰無仮説 : 因子 A の効果 α_i に差はない
  - 対立仮説 : 因子 A の効果に差のあるものがある

$$H_0 : \alpha_1 = \cdots = \alpha_a \quad \text{vs} \quad H_1 : \text{ある } i_1, i_2 \text{ に対して} \alpha_{i_1} \neq \alpha_{i_2}.$$

分析の考え方

- データの変動から因子間での変動と観測誤差の変動を抽出して比較
- 各種平均 ($n = ab$)

$$\begin{aligned}
&\text{(全データの標本平均)} && \bar{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b} Y_{ij}, \\
&\text{(水準 } A_i \text{ の標本平均)} && \bar{Y}_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b} Y_{ij} \quad (i = 1, \ldots, a), \\
&\text{(水準 } B_j \text{ の標本平均)} && \bar{Y}_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a} Y_{ij} \quad (j = 1, \ldots, b).
\end{aligned}$$

- 因子効果の推定量としての標本平均

$$\begin{aligned}
&\bar{Y}_{i\cdot} \to \alpha_i + \frac{1}{b}\sum_{j=1}^{b}\beta_j, \\
&\bar{Y}_{\cdot j} \to \frac{1}{a}\sum_{i=1}^{a}\alpha_i + \beta_j, \\
&\bar{Y}_{\cdot\cdot} \to \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha_i + \beta_j) = \frac{1}{a}\sum_{i=1}^{a}\alpha_i + \frac{1}{b}\sum_{j=1}^{b}\beta_j \\
&Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot} \to \varepsilon_{ij}
\end{aligned}$$

- 各種変動

$$\text{(行間変動)} \qquad SS_A = b\sum_{i=1}^{a}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2,$$

$$\text{(列間変動)} \qquad SS_B = a\sum_{j=1}^{b}(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2,$$

$$\text{(誤差変動)} \qquad SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot})^2.$$

  - **行間変動** : 水準 A 内でのデータの変動
  - **列間変動** : 水準 B 内でのデータの変動

- 変動の分解

$$\underbrace{\text{(全変動) } SS_T}_{\text{自由度 } ab-1} = \underbrace{\text{(行間変動) } SS_A}_{\text{自由度 } a-1} + \underbrace{\text{(列間変動) } SS_B}_{\text{自由度 } b-1} + \underbrace{\text{(誤差変動) } SS_E}_{\text{自由度 } (a-1)(b-1)}$$

- 帰無仮説 H_0 が正しい場合
  - 変動 SS_A, SS_E はともに観測誤差のみが原因で生じる
  - 自由度を除けば本質的な違いはない
  - $SS_A/(a-1), SS_E/(a-1)(b-1)$ は σ^2 の不偏推定量

- 変動の分解

$$\underbrace{\text{(全変動) } SS_T}_{\text{自由度 } ab-1} = \underbrace{\text{(行間変動) } SS_A}_{\text{自由度 } a-1} + \underbrace{\text{(列間変動) } SS_B}_{\text{自由度 } b-1} + \underbrace{\text{(誤差変動) } SS_E}_{\text{自由度 } (a-1)(b-1)}$$

- 対立仮説 H_1 が正しい場合
  - 因子 A 内の水準間での効果 $\alpha_1, \ldots, \alpha_a$ の差に影響される
  - SS_A は SS_E より本質的に大きくなる

二元配置の検定

- 検定統計量

$$F_A = \frac{SS_A/(a-1)}{SS_E/(a-1)(b-1)}$$

- 帰無仮説の下で次の事実が成り立つ
  - SS_A, SS_E は独立
  - SS_A は自由度 $a-1$ の χ^2 分布に従う
  - SS_E は自由度 $(a-1)(b-1)$ の χ^2 分布に従う
- 帰無仮説の下 F_A は自由度 $a-1, (a-1)(b-1)$ の F 分布に従う
- 対立仮説の下 F_A は大きな値をとるので右片側検定

棄却域を用いる場合

- 有意水準 : $\alpha \in (0,1)$
- 自由度 $a-1,(a-1)(b-1)$ の F 分布
  - $F_{1-\alpha}(a-1,(a-1)(b-1))$: $1-\alpha$ 分位点
- H_0 の下で以下が成立

$$P(F_A > F_{1-\alpha}(a-1,(a-1)(b-1))) = \alpha$$

- 第一種過誤の上限が α となる棄却域

$$R_\alpha = (F_{1-\alpha}(a-1,(a-1)(b-1)),\infty)$$

- データから検定統計量 F_A の値を計算
- 以下の場合，帰無仮説を棄却

$$F_A > F_{1-\alpha}(a-1,(a-1)(b-1))$$

p 値を用いる場合

- p 値を計算 (右片側検定の場合の計算方法)

$$(p \text{ 値}) = \int_{F_A}^{\infty} f(x)dx$$

  - f は自由度 $a-1,(a-1)(b-1)$ の F 分布の確率密度
- p 値が α 未満なら帰無仮説を棄却

分散分析表 (二元配置の場合)

	自由度	平方和	平均平方和	F値	p値
因子A	$a-1$	SS_A	$\dfrac{SS_A}{a-1}$	F_A	$\displaystyle\int_{F_A}^{\infty} f(x)dx$
因子B	$b-1$	SS_B	$\dfrac{SS_B}{b-1}$	F_B	$\displaystyle\int_{F_B}^{\infty} f(x)dx$
誤差	$(a-1)(b-1)$	SS_E	$\dfrac{SS_E}{(a-1)(b-1)}$		

- f は自由度 $a-1,(a-1)(b-1)$ の F 分布の確率密度

モデルの別表現

相対効果による定式化

- モデルの書き換え

$$\mu^* = \bar{\alpha} + \bar{\beta}, \qquad \alpha_i^* = \alpha_i - \bar{\alpha}, \qquad \beta_j^* = \beta_j - \bar{\beta}$$

$$\bar{\alpha} = \frac{1}{a}\sum_{i=1}^{a}\alpha_i, \qquad \bar{\beta} = \frac{1}{b}\sum_{j=1}^{b}\beta_j$$

$$Y_{ij} = \mu^* + \alpha_i^* + \beta_j^* + \varepsilon_{ij} \quad (i = 1, \ldots, a;\ j = 1, \ldots, b)$$

ただし $\sum_{i=1}^{a} \alpha_i^* = \sum_{j=1}^{b} \beta_j^* = 0$

- 帰無仮説 H_0 は以下と同等

$$H_0 : \alpha_1^* = \cdots = \alpha_a^* = 0$$

分散分析の計算

- 基本書式

```
aov(formula, data)
#' formula: 式, 二元配置の場合は (観測値 ~ 因子 + 因子)
#' data: データフレーム
```

- 分散分析表なども同様

実習

練習問題

- package::datarium に含まれている jobsatisfaction データについて以下の問に答えよ.

  性別 (gender) と学歴 (education_level) の違いによる仕事の満足度 (score) を収集したデータ

  - データを適当な方法で可視化しなさい.
  - それぞれの因子で満足度の平均に違いがあるか二元配置の分散分析を用いて検討しなさい.

```
#' パッケージのインストールと読み込みは以下のように行うことができる
install.packages("datarium") # インストール
library("datarium") # 読み込み
```

次回の予定

- 回帰分析
- 回帰係数の推定
  - 点推定
  - 区間推定
- 回帰係数の検定
  - 係数の有意性
- 決定係数