二元配置分散分析

渡邊直樹

2022 年 11 月 5 日

分散分析とはデータの分散を使ってグループ（群）間の「平均の差」の検定を行うデータ解析手法である。つまり，2 群間の平均の差を検定するのであれば，データの分布の正規性と等分散性を前提としている点を含めて，t 検定と実質的には同じであり，連続変数の値をとるデータに適用される．よって，分散分析は 3 以上の群または 2 以上の因子での平均の差の検定に用いられる．一元配置分散分析とは，1 つの因子（たとえば，投薬量）について A 群，B 群，C 群といった具合に分け，それらの群に属するデータの平均に統計的な差があるかを検定する．二元配置分散分析とは，2 つの因子（たとえば，性別と投薬量）について，それぞれ A 群，B 群，C 群といった具合にデータを分け，各群に属するデータの平均に差があるかどうかを検定する．因子の数が 2 よりも大きい場合は多元配置分散分析という．分散分析の中では二元配置のものがよく用いられる．本稿では，その二元配置分散分析を EZR で行う場合の手順をノートする．

1　分析手順の概略

まず，統計学における自由度（degree of freedom）について簡単に言及しておく．ある薬剤の投薬量の水準 A 群，B 群，C 群ごとに (1, 3)，(6, 7, 5)，(4, 2) と指標データが観察されたとする．データは全部で 7 つあるので，それらが独立な確率変数であれば，自由度は 7 である．（確率変数 X と Y が独立であるとは一方の生起確率は他方の生起確率に影響を与えないことである．）各群内での平均は 2，6，3 であり，データ全体の平均は 4 である．元のデータであるベクトル $x = (1, 3, 6, 7, 5, 4, 2)$ から，$y = (2, 2, 6, 6, 6, 3, 3)$ と $z = (4, 4, 4, 4, 4, 4, 4)$ を作ると，ベクトル y では各群内で平均をとっており，その個数の分だけ y の各変数は制約を受ける

ので，ベクトル $x-y$ の自由度は $(2-1)+(3-1)+(2-1)=4$ である．同様に，ベクトル $y-z$ の自由度は $(3-1)=2$ である．この例では薬剤の投与量のみを指標の値に影響を与えている因子としている．二元配置分散分析では，その値に別の因子も影響を与えていると想定している．

各種ソフトウェアで分散分析を行うと，結果の一部として「分散分析表」が算出されることが多い．そこには，各群または全体でのデータの自由度も表示されている．分散分析表では「各データとデータ全体の平均値の乖離」,「各群の平均とデータ全体の平均値の乖離」,「各データとそれが属する群の平均の乖離」と各群のデータの自由度から F 値なる統計量が表示されている．

次に，3 以上の群間での比較を行う場合の帰無仮説（null hypothesis）と対立仮説（alternative hypothesis）について説明する．ここでも，因子が 1 つであるとして，その水準を A 群，B 群，C 群に分けたとしよう．帰無仮説と対立仮説は次のようになる．

- 帰無仮説：A 群の平均 =B 群の平均 =C 群の平均
- 対立仮説：A 群の平均 ≠B 群の平均，または，B 群の平均 ≠C 群の平均，または，A 群の平均 ≠C 群の平均

つまり,（3 以上の群での）分散分析において帰無仮説が棄却されたとしても，どの群の間で平均に差があるのかまでは判らない．その場合には，2 群間での「多重比較」を行うことになる．しかし，比較とは一つの要素が異なる群で行うものであるため，何らかの「多重検定の補正」を行わなければならない．特に，因子が 2 つある場合には多重比較に対する注意が必要となる．各種ソフトウェアにおける一元配置分析には多重検定の補正方法に関するオプションが付いているので，3 以上の群または 2 以上の因子を想定する場合には，2 群間の平均の差の検定（t 検定，ウェルチの t 検定，順位和検定，ブルンナー・ムンツェル検定など）ではなく，分散分析が用いられることが多い．

2　EZR での操作

EZR（Easy R）を用いた分散分析の手順をここにノートしておく．ここでは，自分の計算機にデータセット（data5a.csv）が保存されているとして，それを EZR で読み込むことにする．

1. データの読み込み : リボンに表示されているボタンを次の順で押す.

   - データ＞データのインポート＞テキストファイルまたはクリップボード，URL から
   - ポップアップウィンドウにおいて，フィールドの区切り記号として「カンマ [,]」を選択し，OK ボタンを押す．デフォルトでは「空白」が選択されている.（データファイルの場所はデフォルトでローカルファイルシステムが選択されているので，そのままにしておく.）
   - 新たなポップアプウィンドウが現れるので，そこで data5a.csv を選択し、Open ボタンを押す.
   - ホーム画面上で「データセットを表示」ボタンを押して，データが読み込まれていることを確認する.

2. growth(%)（または agrowth）の正規性，等分散性の検定を行う.

   - 統計量＞連続変数の解析＞正規性の検定（Kolmogorov-Smirnov 検定） ··· グラフが表示される.（標準メニュー＞統計量＞要約＞正規性の検定，は 2 群間での検定.）
   - 統計量＞連続変数の解析＞統計解析」3 群以上の等分散性の検定（Bartlett 検定）: グループは dosage か hour のどちらか 1 つを選択する.

3. 分散分析：リボンに表示されているボタンを次の順で押す.

   - 統計量＞連続変数の解析＞複数の因子での平均値の比較（多元配置分析 multi-way ANOVA）
   - ポップアップウィンドウの「データ」の欄にて目的変数 growth(%)（または agrowth）を選択し，因子には dosage と hour を選択する.
   - オプションで「交互作用の解析も行う」の項目にチェックを入れておく.
   - 「適用」（または「OK」）ボタンを押す.
   - 帰無仮説が棄却されたら，どちらかの因子（dosage，hour）を選んで，一元配置分散分析を行う.

- 統計量＞連続変数の解析＞ 3 群以上の間の平均値の比較（一元配置分析 one-way ANOVA）
- ポップアップウィンドウの「データ」の欄にて目的変数 growth(%)（または agrowth）を選択し，因子として dosage または hour を選択する．
- オプションで Bonferroni や Tukey などの補正（多重比較）の項目にチェックを入れておく．
- 「適用」（または「OK」）ボタンを押す．

4. 分析結果の保存と主成分スコアを書き加えたデータセットの出力：リボンに表示されているボタンを次の順で押す．

- ファイル＞出力ファイルに保存．ポップアップウィンドウでファイルに名前をつけ，計算機に csv 形式で分析結果を保存する．（EZR の画面にも分析結果は出力されるが，やや見にくい．）
- データ＞アクティブデータセット＞アクティブデータセットのエクスポート.（アクティヴデータセットの保存を押すと，RData 形式で保存されてしまうので，Excel などで読み込めない.）ポップアップウィンドウにおいて，フィールドの区切り記号として「カンマ [,]」を選択し，OK ボタンを押す．デフォルトでは「空白」が選択されている．

Appendix

F 検定と t 検定が依拠する F 分布と t 分布はともにカイ 2 乗分布から派生して定義された確率分布である．ここでは，これらの関係をまずまとめておく．なお，統計学における自由度とは独立に値を決められる確率変数の数のことである.（本文を参照せよ.）

平均 0，分散 1 の正規分布を標準正規分布という．標準正規分布に従う互いに独立な n 個の確率変数 X_i（$i = 1, \ldots, n$）の 2 乗和

$$Y = \sum_{i=1}^{n} = X_1^2 + \cdots + X_n^2$$

が従う確率分布をカイ 2 乗分布という．X_l は自由度 l のカイ 2 乗分布に従う確率変数であり，X_m は自由度 m のカイ 2 乗分布に従う確率変数で

あり，これらが互いに独立であるとすると，

$$F = \frac{X_l/l}{X_m/m}$$

が従う分布を自由度 (l, m) の F 分布という．ここで，自由度 $(1, m)$ の F 分布に従う F の正の平方根を

$$t = \frac{X_1^{1/2}}{(X_m/m)^{1/2}}$$

とすると，t が従う確率分布を自由度 m の t 分布という．このように，F の分子が自由度 1 のカイ 2 乗分布に従う場合には，分布の形状こそ違いこそすれ，実質的に F 検定と t 検定は同じものである．