2009/14th/ANOVA

2024年 6月 28日

目次

一元配置分散分析

- 平均値の差の検定で、3つ以上の標本について平均値の差を比較するときに使用する
  - 2つの標本の平均値の差の検定は「 対応のないt検定 」を使う
  - もし、3つ以上の標本についてt検定を使う（2組ずつのペアで検定をする）と、例えば差がないにもかかわらず差があると検定してしまう危険性がある
- 分散分析（ANOVA ：ANalysis Of VAriance ）は、標本同士の平均値の差の程度がそれぞれの標本内の誤差に比べて大きいかを調べて分析する方法である

▲▼

検定の対象

▲▼

因子と水準

対応のない複数の組の標本について考える。
例えば、3つの血圧降下剤（A薬、B薬、C薬）の効果を調べるために、15人の被験者を無作為に3つのグループに分けて、それぞれのグループにA薬、B薬、C薬いずれかを投与して収縮期血圧を測定したところ、次の表のようになったとする。

番号	A薬	B薬	C薬	全体
1	116	106	108	
2	128	102	100	
3	129	108	108	
4	137	118	114	
5	140	116	110	
合計	650	550	540	1740
平均	130	110	108	116

- 差を調べる変数の要因を「**因子**」という
  - 上の表では「血圧」にあたる
  - ひとつの因子について分析することから「一元配置分散分析」という
- 要因の内容が異なるグループを「**水準**」という
  - 上の表では「A薬、B薬、C薬」のような項目にあたる
- 各水準の標本数（データの個数）を「**繰り返し数**」という

一般には、一元配置分散分析のデータは、次のような表で書くことができる。

番号	1	2	…	p	全体
1	x_{11}	x_{21}	…	x_{p1}	
2	x_{12}	x_{22}	…	x_{p2}	
…					
	x_{1n_1}	x_{2n_2}	…	x_{pn_p}	
繰り返し数	n_1	n_2	…	n_p	n
平均	$\overline{x_1}$	$\overline{x_2}$	…	$\overline{x_p}$	$\overline{x}$

表のデータは、次のことを表している。

- 水準の数 : 1〜p
- それぞれの水準での繰り返し数 :

$$n_1,n_2,\cdots,n_p$$

- 全体の標本数 : 各水準の繰り返し数の合計

$$n=\sum_{i=1}^{p}n_i$$

- 第

$$i$$

 水準の第

$$j$$

 番目のデータ :

$$x_{ij}$$

- それぞれの水準での平均 :

$$\overline{x_1},\overline{x_2},\cdots,\overline{x_p}$$

- 全体の標本の平均 :

$$\overline{x}=\frac{\sum_{i=1}^{p}\sum_{j=1}^{n_i}x_{ij}}{n}$$

▲▼

全変動と級間変動と誤差変動

ここで､第

$$i$$

水準の第

$$j$$

番目のデータ

$$x_{ij}$$

について考えてみる｡

- 全体の平均

  $$\overline{x}$$

  や水準内の平均

  $$\overline{x_i}$$

  を使って考えると､…
- 各データは､「全体の平均」と「全体平均とそのデータの水準の平均とのズレ」と「そのデータの水準の平均とデータとのズレ」に分解することができる｡

  > **（各データの値）**
  > **＝（全体の平均）＋（全体の平均と水準の平均のズレ）＋（各データと水準の平均のズレ）**

- 別の書き方をすると､「各データが全体の平均からどれくらいズレているか」は､「全体平均とそのデータの水準の平均とのズレ」と「そのデータの水準の平均とデータとのズレ」に分解することができる｡

  > **（各データと全体の平均のズレ）**
  > **＝（全体の平均と水準の平均のズレ）＋（各データと水準の平均のズレ）**

- このズレを「**変動**」という
  - 全体の平均と水準の平均とのズレ（差）を「**級間変動**」（または群間変動）といい､次のようにあらわす

    $$T_1=\sum_{i=1}^{p}n_i(\overline{x_i}-\overline{x})^2$$

  - 各データと水準の平均とのズレ（差）を「**誤差変動**」（級内変動または群内変動）といい､次のようにあらわす

    $$T_E=\sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij}-\overline{x_i})^2$$

  - 各データと全体の平均とのズレ（差）を「**全変動**」といい､次のようにあらわす

$$T=\sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2$$
$$=T_1+T_E$$

- また、級間変動と誤差変動について、不偏分散を次のように定義しておく
  - 級間変動の不偏分散

$$V_1$$

$$V_1=\frac{T_1}{p-1}$$

  - 誤差変動の不偏分散

$$V_E$$

$$V_E=\frac{T_E}{n-p}$$

    - もし、各水準の繰り返し数が一定の値

$$n_i$$

      の場合は次のようにも書ける

$$V_E=\frac{T_E}{p(n_i-1)}$$

▲▼

1元配置分散分析

- 3つ以上の標本について平均値の差を調べて、級間変動と誤差変動のどちらの比率が高いか調べる
- 級間変動のほうが大きければ、全変動に与える影響が級間変動の方が大きいと見なし、平均値に差があるとする

▲▼

帰無仮説と対立仮説

対応のない3組以上の標本の平均値に差があるかどうかを調べる。

- 帰無仮説

$$H_0$$

  は「各水準の平均値に差はない」
- 対立仮説

$$H_1$$

は「各水準の（少なくとも1つの組み合わせで）平均値に差がある」

▲▼

検定統計量の算出

- 級間変動（または群間変動）を求める

$$T_1=\sum_{i=1}^{p}n_i(\overline{x_i}-\overline{x})^2$$

- 誤差変動（級内変動または群内変動）を求める

$$T_E=\sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij}-\overline{x_i})^2$$

- 級間変動の不偏分散を求める

$$V_1=\frac{T_1}{p-1}$$

- 誤差変動の不偏分散を求める

$$V_E=\frac{T_E}{n-p}$$

- 第1自由度が

$$df_1=p-1$$

、第2自由度が

$$df_E=n-p$$

のF分布にしたがう、検定統計量

$$F_1$$

を次の式から算出する

$$F_1=\frac{V_1}{V_E}$$

「**分散分析表**」にまとめると、次のようになる。

要因	平方和	自由度	平均平方	F値
級間	T_1	$df_1=p-1$	$V_1=T_1/df_1$	$F_1=V_1/V_E$
誤差	T_E	$df_E=n-p$	$V_E=T_E/df_E$	
全体	$T=T_1+T_E$	$df_T=n-1$	$V_T=T/df_T$	

▲▼

仮説の判定（片側検定）

検定統計量

$$F_1$$

と、第1自由度が

$$df_1=p-1$$

、第2自由度が

$$df_E=n-p$$

のF分布について、有意水準

$$\alpha$$

の有意点の値（F分布表などから求める）を使って、判定をする

- 帰無仮説

  $$H_0$$

  を棄却：

  $$F_1 \geq F(df_1, df_E)$$

  - 「有意に差がある」「検定の結果、有意である」
- 帰無仮説

  $$H_0$$

  を採択：

  $$F_1 < F(df_1, df_E)$$

  - 「有意に差はない」「検定の結果、有意でない」「差があるとはいえない」