データ解析基礎

第11回 分散分析（１）

三つ以上の母集団の比較

- **三つ以上の母集団のあいだで，平均値に差があるかどうかを検定する．**
  - **帰無仮説** H0：$\mu_1 = \mu_2 = \ldots = \mu_K$
  - **対立仮説** H1：not H0
- **三つの平均値をどのようにして比較するか？**

三つ以上の母集団の比較

3

- 思いつく別の方法
  - 全母集団から二つずつ取り出してペアを作り，

    帰無仮説H0：$\mu_i = \mu_j$

    対立仮説H1：μ_i != μ_j

    の t 検定をすべてのペアについて行なう．
- 残念ながら，これは誤り
- 分散分析と呼ばれる方法を使う．
  - 「分散」という名前がついているが，平均値に関する検定である．

二つずつ比較するとなぜまずいのか？

4

- 例：三つの母集団を二つずつ比較する場合
  - 以下の三つの検定を行ない，少なくとも一つが有意になったときに平均値に差があると判断する．

    1) $\mu_1 = \mu_2$ or not
    2) $\mu_2 = \mu_3$ or not
    3) $\mu_1 = \mu_3$ or not
- こうすると，第1種過誤確率が設定した有意水準より大きくなってしまい，結果として有意水準5%で検定したことにならない．

二つずつ比較したときの第１種過誤確率

5

- 第一種過誤確率を書き下してみる．

  P(μ1=μ2=μ3が正しいが，三つの帰無仮説のうち一つが棄却される)

  = 1 - P(μ1=μ2=μ3が正しく，三つの帰無仮説がすべて棄却されない)

  = 1 - P(μ1=μ2が正しく帰無仮説が棄却されない AND

  μ2=μ3が正しく帰無仮説が棄却されない AND

  μ3=μ1が正しく帰無仮説が棄却されない）

- 仮に三つの検定が独立であるとすると，

  = 1 - P(μ1=μ2が正しくかつ帰無仮説が棄却されない)

  × P(μ2=μ3が正しくかつ帰無仮説が棄却されない)

  × P(μ3=μ1が正しくかつ帰無仮説が棄却されない）

二つずつ比較したときの第１種過誤確率

6

- （前ページの最後の式）

  = 1 - （1 - 有意水準）×（1 - 有意水準）×（1 - 有意水準）

- この値は，最初に設定した有意水準より大きくなる．

  -有意水準を5%とすると，この値は14.2%になる．

- 実際には三つの検定は独立ではないが，それでも
  もとの有意水準より大きくなる．

- 三つの仮説の有意水準を適切に設定することは，
  事実上不可能である．

分布の広がりと平均の差

7

- 三つの違う教育を受けた学生の成績を比較したい．
- 各グループの点数分布が十分に離れていれば，データが与えられたときに，それがどのグループのデータであるかは容易に判断できる．
- 逆に，点数分布が重なっていれば，どのグループに属しているかは確実には判断できない．
- どうやって判断すればよいか？

分布の広がりと平均の差

8

- 複数のグループの平均が離れているかどうかを，各グループの分布の広がりと比較して判断する．
  - ただし，各グループの広がりは一定とする．
- 三つの平均が等しいという仮定（帰無仮説）の下で，その比がどのくらいになるかを評価し，棄却域を設定する．

分散分析(ANOVA: ANalysis Of VAriance)

9

- いま，K個のグループがあるとする．
- 準備として，データに関して以下の量を考える．
  - X_{ij}：j 番目のグループの i 番目のデータ
  - N_j：j 番目のグループのデータ数
  - N：全データ数（すなわち，N_j の和）
  - T_j：j 番目のグループのデータの和
  - T：全データの和（すなわち，T_j の和）
  - M_j：j 番目のグループの平均値
  - M：全データの平均値

分散分析：グループ内での変動

10

- グループ内でのデータのばらつきを示す量を定める．．
  - 各グループ内部でのデータのばらつきの和 *SSW* (Sum of Squares Within groups：水準内変動) は

$$SSW = \Sigma_{ij} \ (X_{ij} - M_j)^2 = \Sigma_{ij} \ X_{ij}^{\ 2} \ - \Sigma_j \ (T_j^{\ 2} \ / N_j)$$

  - *SSW*の自由度は$N - K$ (データ数N，グループ数K)．
- グループ内ばらつきの平均値 (Mean Square Within)

$$MSW = SSW / (N - K)$$

分散分析：グループ間の変動

11

- **グループ間でのデータが離れ具合いを評価する．**
  - **グループ平均のばらつき** *SSB* (Sum of Squares Between groups**：水準間変動**) **は，**

$$SSB = \Sigma_j \ N_j (M_j - M)^2 = \Sigma_j \ (T_j^{\,2} / N_j) \ - T^2 \ / N$$

  - *SSB***の自由度は** K-1

- **グループ間ばらつきの平均値** (Mean Square Between)

$$MSB = SSB / (K - 1)$$

分散分析：全変動

12

- **すべてのデータのばらつきの度合いを評価する．**
- **全変動** (Sum of Squares Total) **は**

$$SST = \Sigma_{ij} \ (X_{ij} - M)^2 = \Sigma_{ij} \ X_{ij}^{\,2} \ - T^2 \ / N$$

- **以上で定義した三つのばらつきの間には次式が成立．**

$$SST = SSB + SSW$$

（全データのばらつき）は（グループ内部のばらつき）と（グループ間のばらつき）に分解できる．

分散分析：変動の分解

13

- *SST* = *SSB* + *SSW*
  - *SSB* = 0 ならば，グループ間に差がない．
  - *SSW* = 0 ならば，データのばらつきはグループ間のばらつきによるものだけになる．
- グループ内ばらつきとグループ間ばらつきを比較する指標の一つとして，*SSB* / *SST* が考えられる．
  - 0と1のあいだの値をとる．
  - 比が大きければグループ間に差があることになる．
  - 独立変数と従属変数の標本相関係数の2乗に等しい．

検定の行ない方

14

- 実際は，*SSB* / *SST* ではなく *MSB* / *MSW* について検定を行なう．
- 以下の条件が成立すると仮定する．
  - すべてのグループの母集団は正規分布に従う．
  - 母集団の分散は等しい．
  - グループはすべて独立である．
- すべての母平均が等しいという帰無仮説の下で，*MSB* / *MSW* は，自由度 $K - 1$，$N - K$ のF分布に従う．

なぜMSB/MSWはF分布に従うのか？

15

- 前ページの仮説の下で，*MSB*と*MSW* はいずれもグループに共通の母分散の不偏推定量になる．
- *MSB/MSW* は，二つのカイ2乗分布の比になる．

$$\frac{MSB}{MSW} = \frac{SSB / (K - 1)}{SSW / (N - K)}$$

$$= \frac{(SSB / \sigma^2) / (K - 1)}{(SSW / \sigma^2) / (N - K)}$$

SSB / σ^2 ，*SSW* / σ^2がχ^2 分布に従うことに注意．

- したがって，*MSB/MSW* はF分布に従うことになる．

例題：

16

- 薬物中毒患者用の三つの処方の成績を比較したい．
- 30人の患者を三グループに分けて各々の処方を適用した結果，テストの成績が以下のように分布した．

処方１：	6, 10, 8, 6, 9, 8, 7, 5, 6, 5
処方２：	4, 2, 3, 5, 1, 3, 2, 2, 4, 4
処方３：	6, 2, 2, 6, 4, 5, 3, 5, 3, 4

- 処方に差があるかどうかを有意水準5%で検定する．
  - 帰無仮説H0： $\mu_1 = \mu_2 = \mu_3$
  - 対立仮説H1: not H0

例題の解法(1)

17

- Step 1: **自由度の計算**
  - **全データ数**N = 30**，グループ数**K = 3**であるから，検定に用いる**F**分布の自由度は**
    K - 1 = 3 - 1 = 2,
    N - K = 30 - 3 = 27
- Step 2: **棄却域の決定**
  - F(2, 27) **の片側**95%**点は**3.354**．これ以上が棄却域．**
  - **平均に違いがあれば，グループ間ばらつきは必ず増加するので，**F**分布上で必ず右片側検定になる．**

例題の解法(2)

18

- Step 3: *MSB***の計算**
  - **各群のデータ数　：**$N_1 = N_2 = N_3 = 10$
  - **各群のデータの和：** $T_1 = 70$, $T_2 = 30$, $T_3 = 40$
  - **各群のデータ平均：** $M_1 = 7$, $M_2 = 3$, $M_3 = 4$
  - **全データの和と平均：** $N = 30$, $T = 140$, $M = 4.67$
- **以上の準備の下で，***SSB***，***MSB***は**

$$SSB = T_1^2/N_1 + T_2^2/N_2 + T_3^2/N_3 - T^2/N$$
$$= 490 + 90 + 160 - 653.3 = 86.67$$

$$MSB = SSB / (K - 1) = 86.67 / 2 = 43.33$$

例題の解法(3)

19

- Step 4: MSWの計算
  - 全データの2乗和 $\Sigma_{ij} X_{ij}^2$ を計算すると800になる．
  - したがって，

$$SSW = \Sigma_{ij} X_{ij}^2 - (T_1^2/ N_1 + T_2^2 / N_2 + T_3^2 / N_3)$$

$$= 800 - (490 + 90 + 160) = 60$$

$$MSW = SSW / (N - K) = 60 / 27 = 2.22$$

となる．

例題の解法(4)

20

- Step5：*MSB*/*MSW*の計算
  - 以上の結果より，この比の値は
    43.33 / 2.22 = 19.52．
- Step6：棄却域との比較
  - 最初に求めた限界点は3.354であった．
    3.354 < 19.52 であるから帰無仮説は棄却される．
- Step7：結論
  - 三つの処方の間には差があることになる．

例題の解法(5)

21

- **補足**
  - **有意水準を**1%, 0.1%**の場合の棄却域はそれぞれ**

    p = 0.01　　F > 5.49

    p = 0.001　　F > 9.02
  - **この例題では，三つの処方間の差は** p < 0.001 **で有意であるといえる．**

分散分析表の作り方

22

- **分散分析では，以上の計算過程を以下のような表（分散分析表という）にまとめて表す．**

Source：	SS	df	MS	F値	p値
（要因）	(平方和)	(自由度)	(平均平方)		
グループ間：	86.67	2	43.33	19.52	***
グループ内：	60.00	27	2.22		
全体：	146.67	29			

多重比較(multiple comparison)

23

- 三つのグループ間に差があることがわかったら，次に，どのグループとどのグループの間に差があるのかを知りたくなるのが人情である．
- しかし，グループを二つずつ取り出して t 検定をすると，先に述べたように有意水準が保てなくなる．
- どうすればよいか？

  → 事後検定(post hoc test) という方法を用いる．

事後検定 (post hoc test)

24

- 「グループ間で平均が等しい」という帰無仮説が棄却されたという前提の下で検定を行なうもの．
- いくつかの方法がある．
  - TukeyのHSD法
  - 修正Tukey法
  - Sheffe検定
- これらの検定をいきなりやってはいけない．

HSD(honestly significant difference) 法

25

- 各グループのデータ数が等しく n であるとする．
- studentized range statistic（student化された範囲）と呼ばれる以下の統計量を用いて検定する．

$$q = \frac{\overline{Xj} - \overline{Xk}}{\mathrm{sqrt}(MSW / n)}$$

- q の分布の値は数表で調べる．
  - 自由度はグループ数 K とグループ内変動の自由度 $N - K$ の二つ．

HSD(honestly significant difference) 法

26

- グループのデータ数が等しくないときは，n を以下の量で置き換える．

$$n' = K / \left(\frac{1}{N1} + \frac{1}{N2} + \ldots + \frac{1}{N_K} \right)$$

例題：

27

- 前出の例に関して，多重比較を行なう．
- 自由度3, 27 のqの数表から棄却域を求める．
  - 有意水準5%とすれば， q > 3.51となる．
  - 検定そのものは両側検定であるが，分散比の検定と同様にq検定では常に右片側検定になる．
- q の分母部分を先に計算しておくと

  sqrt(*MSW* / *n*) = sqrt(2.22 / 10) = 0.47
- 2グループ間の差を求め，対応するq値を求める．

例題：

28

- その結果は

  q_{12} = (7 - 3) / 0.47 = 8.49　> 3.51

  q_{23} = (4 - 3) / 0.47 = 2.13　< 3.51

  q_{13} = (7 - 4) / 0.47 = 6.38　> 3.51
- したがって，平均に有意な差があるのは，「1-2のあいだ」と「1-3のあいだ」と結論できる．
- qの数表は，どんな本にでも載っているわけでない．
  - 参考のため，ホームページに添付しておく．