データ解析基礎

第12回 2元配置の分散分析
(two-way ANOVA)

2要因の分析とは？

2

- 二つの要因がもたらす影響について同時に検定する．
- 例題：
  - 薬物中毒患者に対する三つの処方
  - 男性と女性に対する効き目を同時に調べてみたい．

  要因A：処方　　　　要因B：性別

	処方１	処方２	処方３
男性			
女性			

2元配置

- 例題の続き（それぞれのセルの平均を求める）

	処方１	処方２	処方３	
男性	M=8.4	M=3	M=2.8	M=4.7
女性	M=5.6	M=3	M=5.2	M=4.6
	M=7	M=3	M=4	

  - 行の平均は性別の全体的な効果
  - 列の平均は処方の全体的な効果

    → これらは1元配置の分散分析で調べられる．

- 処方ごとに性別の効果を調べると，性別による効果が処方によって異なることがわかる．

交互作用(interaction)

- 前ページの例のように

  一つの要因がもたらす効果が，別の要因によって変化することを二つの要因の「交互作用」という．

- したがって，2元配置の分散分析では，
  - 要因Aの効果　（単独の効果を「主効果」という）
    (main effect)
  - 要因Bの効果
  - 要因Aと要因Bの交互作用 (interaction)

  の三つについて検定を行なうことになる．

two-way ANOVAで検定する三つの仮説

5

- 仮説1：要因Aの主効果
  - 帰無仮説H0: $\mu_0^A = \mu_1^A = \ldots = \mu_n^A$
  - 対立仮説H1: not H0
- 仮説2 ：要因Bの主効果
  - 帰無仮説H0: $\mu_0^B = \mu_1^B = \ldots = \mu_k^B$
  - 対立仮説H1: not H0
- 仮説3 ：交互作用
  - 帰無仮説H0: 要因AとBの効果に交互作用がない．
  - 対立仮説H1: 交互作用がある．

分散分析における仮定

6

- 分散分析における理論的な前提
  - 各セルに含まれるデータが，正規分布に従う母集団から取り出されたものであること．
  - 各セルに含まれるデータの分散が等しいこと．
- 以下のアルゴリズムを使う上での前提
  - 各セルに含まれるデータ数が等しいか比例関係．

    セルのデータ数をN_{ij}, 各行のデータ数をN_i^A，
    各列のデータ数をN_j^B，総データ数をNとしたとき，
    $N_{ij} = N_i^A N_j^B / N$ が成立する．

分散分析のアルゴリズム：準備

7

- X_{ijk}：セル(i, j) の k 番目のデータ
- N_{ij}：セル(i, j) のデータ数
- N_i^A, N_j^B：要因A (B)が i (j)であるデータ数
- N：全データ数（すなわち，Nij の和）
- T_{ij}, M_{ij}：セル(i, j) のデータの和，および平均
- T_i^A, T_j^B：要因A (B)が i (j)であるデータの和
- M_i^A, M_j^B：要因A (B)が i (j)であるデータの平均
- T, M：全データの和および平均
- K^A, K^B：要因A (B)の水準（グループ）の数

要因Aの効果を調べる

8

- 1元配置のときと同様に，「要因間のばらつき」と「要因内のばらつき」を比較する．
- 要因Aに関してグループ間の平均平方MSAを求める．

$$MSA = SSA / (K^A - 1)$$

$$SSA = \Sigma_i \; N_i^A (M_i^A - M)^2 = \Sigma_i \; (T_i^{A\,2} / N_i^A) - T^2 / N$$

- 各セル内部の平均平方MSWを求める．

$$MSW = SSW / (N - K^A K^B)$$

$$SSW = \Sigma_{ijk} \; (X_{ijk} - M_{ij})^2 = \Sigma_{ijk} \; X_{ijk}{}^2 - \Sigma_{ij} (T_{ij}{}^2 / N_{ij})$$

要因Aの効果を調べる

9

- 帰無仮説の下で，両者の比 *MSA*/*MSW*が 自由度 $(K^A - 1)$, $(N - K^A K^B)$ のF分布に従うことを用いて検定を行なう．
- 注意：
  - 要因A，Bについて別々にone-way ANOVAを行なうことと，two-way ANOVAの主効果を調べることとは異なる．

要因Bによる効果を検定する．

10

- 要因Bについても同じ手続きを行なう．
- グループ間平均平方*MSB*を求める．

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

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

- 帰無仮説の下で，*MSB*/*MSW*が自由度 $(K^B - 1)$, $(N - K^A K^B)$のF分布に従うことを用いて検定を行なう．
  - *MSW*は前ページで求めたものと同じ．

要因AとBの交互作用に関する検定

11

- 要因A, Bの組合せに関する平均平方*MSAB*を求める.

$MSAB = SSAB / (K^A - 1)(K^B - 1)$

$SSAB = \Sigma_{ij}\ N_{ij}\,(M_{ij} - M_i^A - M_j^B + M)^2$

$= \Sigma_{ij}\ (T_{ij}^2 / N_{ij}) - SSA - SSB - T^2 / N$

これは $(Mij - M)$ から $(Mi^A - M) + (Mj^B - M)$ を引いたもの.

- 帰無仮説の下で，*MSAB*/*MSW*が自由度 $(K^A - 1)(K^B - 1)$，$(N - K^A K^B)$のF分布に従うことを用いて検定を行なう.

  -*MSW*は先に求めたものを使う.

分散分析表

12

- 2元配置の場合の分散分析表は次のように作る.

Source：（要因）	SS (平方和)	df (自由度)	MS (平均平方)	F値	p値
A：	*SSA*	$K^A - 1$	*MSA*	*MSA* / *MSW*	
B：	*SSB*	$K^B - 1$	*MSB*	*MSB* / *MSW*	
A×B：	*SSAB*	$(K^A - 1)(K^B - 1)$	*MSAB*	*MSAB* / *MSW*	
Within:	*SSW*	$N - K^A K^B$	*MSW*		

例題：

- 先の例題について有意水準5%で分散分析を行なえ．

	処方１	処方２	処方３
男性患者：	10, 9, 8, 7, 7	5, 4, 3, 2, 1	4, 3, 3, 2, 2
女性患者：	6, 6, 6, 5, 5	4, 4, 3, 2, 2	6, 6, 5, 5, 4

要因A：性別，要因B：処方の種類とする．

- 各種数値を求める．

$N_{11} = N_{12} = N_{13} = N_{21} = N_{22} = N_{23} = 5$

$N_1^A = N_2^A = 15, \quad N_1^B = N_2^B = N_3^B = 10 \quad N = 30$

例題の解法

- 和や平均を求める(ただし，後で使うのは和だけ！)

$T_{11} = 42, \quad T_{12} = 15, \quad T_{13} = 14,$
$T_{21} = 28, \quad T_{22} = 15, \quad T_{23} = 26,$
$T_1^A = 71, \quad T_2^A = 69, \quad T_1^B = 70, \quad T_2^B = 30, \quad T_3^B = 40$

$M_{11} = 8.4, \quad M_{12} = 3, \quad M_{13} = 2.8,$

$M_{21} = 5.6, \quad M_{22} = 3, \quad M_{23} = 5.2,$

$M_1^A = 4.7\ , \quad M_2^A = 4.6, \quad M_1^B = 7, \quad M_2^B = 3, \quad M_3^B = 4$

$T = 140, \quad M = 4.33,$

$K^A = 2, \quad K^B = 3, \quad \Sigma X_{ijk}^2 = 800$

例題の解法（続き）

15

- 平均平方を求める (点線単位で考えるとわかりやすい)

$SSA = \boxed{(T_1^A)^2 / N_1^A + (T_2^A)^2 / N_2^A} - \boxed{T^2 / N}$

$= 71^2 / 15 + 69^2 / 15 - 140^2 / 30 = 0.14$

$MSA = SSA / (K^A - 1) = 0.14 / (2 - 1) = 0.14$

$SSB = \boxed{(T_1^B)^2 / N_1^A + (T_2^B)^2 / N_2^B + (T_3^B)^2 / N_3^B} - \boxed{T^2 / N}$

$= 70^2 / 10 + 30^2 / 10 + 40^2 / 10 - 140^2 / 30 = 86.67$

$MSB = SSB / (K^B - 1) = 86.67 / (3 - 1) = 43.33$

$SSW = \boxed{\Sigma X_{ijk}^2} - \boxed{[T_{11}^2 / N_{11} + T_{12}^2 / N_{12} + T_{13}^2 / N_{13} + T_{21}^2 / N_{21} + T_{22}^2 / N_{22} + T_{23}^2 / N_{23}]}$

$= 26$

例題の解法（続き）

16

- 平均平方を求める．

$MSW = SSW / (N - K^A K^B) = 26 / (30 - 2*3) = 1.08$

$SSAB = \boxed{[T_{11}^2 / N_{11} + T_{12}^2 / N_{12} + T_{13}^2 / N_{13} + T_{21}^2 / N_{21} + T_{22}^2 / N_{22} + T_{23}^2 / N_{23}]}$

$- \boxed{SSA} - \boxed{SSB} - \boxed{T^2 / N}$

$= 774 - 0.14 - 86.67 - 653.33$

$= 33.86$

$MSAB = SSAB / (K^A - 1)(K^B - 1)$

$= 33.86 / (2 - 1)(3 - 1) = 16.93$

- 各セルのデータ数が等しいときはより単純．

例題の解法（要因A（性別）による効果） 17

- 有意水準5%の棄却域は，自由度1, 24のF分布の数表から, F > 4.26であることがわかる．
- 一方，検定統計量は

$$F = MSA / MSW = 0.14 / 1.08 = 0.13$$

- したがって，帰無仮説は棄却されない．
  - 「性別による主効果は有意ではない」と結論される．

例題の解法（要因B（処方）による効果） 18

- 有意水準5%の棄却域は，自由度2, 24のF分布の数表からF > 3.40である．
- 一方，検定統計量を求めると

$$F = MSB / MSW = 43.34 / 1.08 = 40.13$$

- したがって，帰無仮説は棄却される．
  - 「処方による主効果は有意である」と結論される．

例題の解法（交互作用の効果）

19

- 有意水準5%の棄却域は，自由度2, 24のF分布の数表から F > 3.40．
- 一方，検定統計量は次のようになる．

$$F = MSAB / MSW = 16.93 / 1.08 = 15.68$$

- したがって，帰無仮説は棄却される．
  - 「交互作用による効果は有意である」と結論される．

例題の解法（続き）

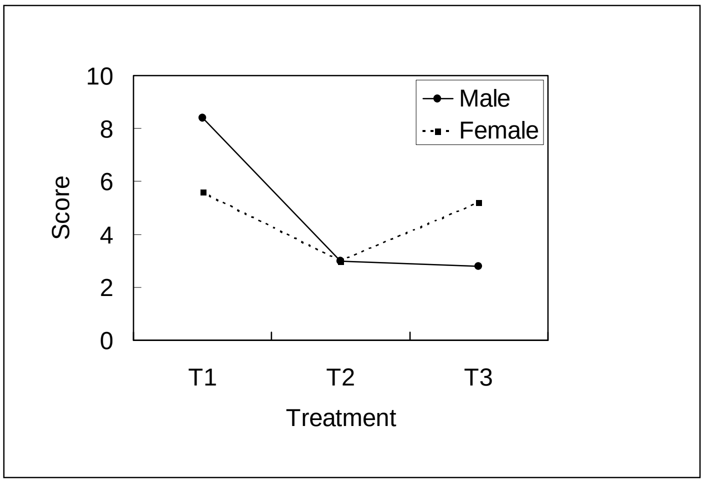
20

- 分散分析表

Source	SS	df	MS	F値	p値
A：	0.14	1	0.14	0.13	> 0.7
B：	86.67	2	43.33	40.13	< 0.001***
A×B：	33.86	2	16.93	15.68	< 0.001***
Within:	26.00	24	1.08		

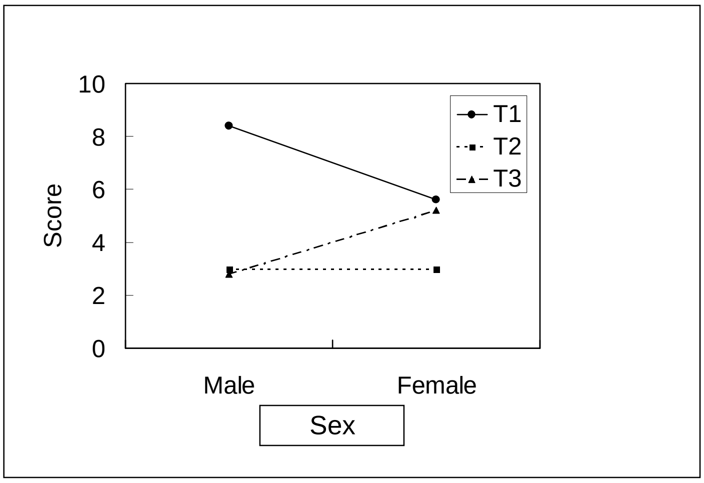
グラフによる結果の理解

- 結果をグラフにかくと交互作用の様子がわかる．



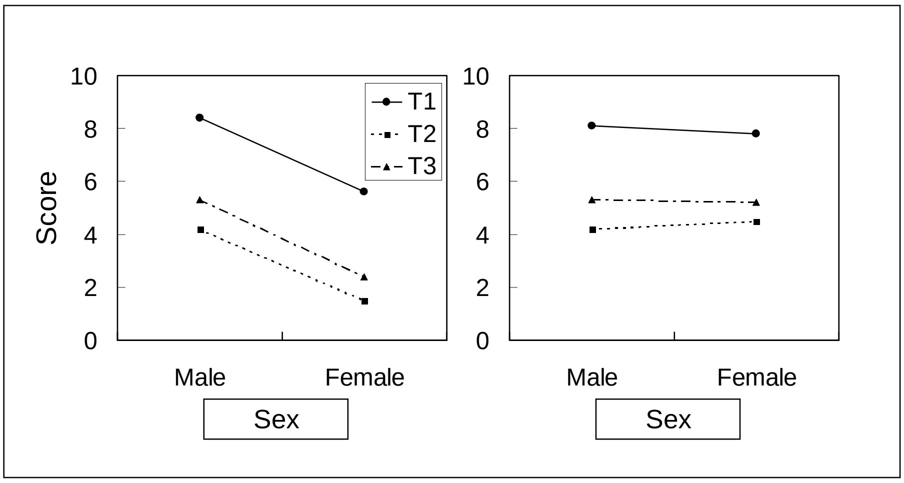
グラフによる結果の理解

- 別の表現



グラフによる結果の理解

- 交互作用がない場合（グラフが平行になる）



関連のあるデータについての分散分析

- 例題：Within-Subject design

  同一の被験者に三つの条件で実験に参加してもらい，条件間で結果に差があるかどうかを調べたい．

- 被験者が一つの要因であると考えられるので，
  - 要因A：実験条件
  - 要因B：被験者

  として，two-way ANOVAを行なえばよい．

- 各セルのデータ数が1なので，セルの分散は使わない．

  →「繰り返しのない分散分析」という．

繰り返しのない分散分析アルゴリズム

- 準備
  - X_{ij}：セル(i, j) のデータ
  - K^A, K^B：要因A (B)の水準（グループ）の数
  - N：全データ数（すなわち，$K^A \times K^B$）
  - T_i^A, T_j^B：要因A (B)が i (j)であるデータの和
  - M_i^A, M_j^B：要因A (B)が i (j)であるデータの平均
  - T, M：全データの和および平均

アルゴリズム

- 要因Aに関する平均平方MSAを求める．

$$MSA = SSA / (K^A - 1)$$

$$SSA = K^B \Sigma_i \ (M_i^A - M)^2 = \boxed{(\Sigma_i T_i^{A\,2}) / K^B} - \boxed{T^2 / N}$$

- 要因Bに関する平均平方MSBを求める（実際は不要）．

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

$$SSB = K^A \Sigma_j \ (M_j^B - M)^2 = \boxed{(\Sigma_j T_j^{B\,2}) / K^A} - \boxed{T^2 / N}$$

- 交互作用に関する平均平方MSABを求める．

$$MSAB = SSAB / (K^A - 1)(K^B - 1)$$

$$SSAB = \Sigma_{ij} \ (X_{ij} - M_i^A - M_j^B + M)^2$$

$$= \boxed{\Sigma_{ij} \ X_{ij}^2} - \boxed{SSA} - \boxed{SSB} - \boxed{T^2 / N}$$

繰り返しのない分散分析

27

- 検定統計量に使って検定を行なう．

  F = *MSA* / *MSAB*

  -*MSA*：実験条件によるばらつき．

  -*MSAB*：実験条件と被験者による交互作用．

  交互作用は被験者ごとに効果に違いがあるときに現れるので，交互作用が大きい場合は，実験条件の効果があるとは考えにくいことになる．

- 検定統計量が，自由度$(K^A\text{-}1)$， $(K^A - 1)(K^B - 1)$のＦ分布に従うことを使って検定する．

繰り返しのない場合の分散分析表

28

Source：（要因）	SS (平方和)	df (自由度)	MS (平均平方)	F値	p値
A：	*SSA*	$K^A - 1$	*MSA*	*MSA/MSAB*	
B：	*SSB*	$K^B - 1$	*MSB*		
A×B：	*SSAB*	$(K^A - 1)(K^B - 1)$	*MSAB*		

- 各セルにはデータが一つしかないので*MSW*がない．
- 検定統計量は，要因Aの効果を交互作用による効果で割ったものに相当する．

分散分析の数理モデル

29

- 1元配置のモデル

$$Y_{ij} = \mu + a_i + \varepsilon_{ij}$$

μ ：全体の平均

a_i ：要因Aによる効果（$a_i = \mu_i - \mu$）

ε_{ij} ：誤差（ノイズ）（$\varepsilon_{ij} = Y_{ij} - \mu_i$）

  - ε_{ij} の分散は等しく互いに独立であることが前提．

- 仮説の設定
  - 帰無仮説 H0：all $a_i = 0$
  - 対立仮説 H1：not all $a_i = 0$

1元配置のモデル

30

- *MSW*と*MSB*は次の量の不偏推定量である．
  - *MSW*：(残差 ε_{ij} の分散)
  - *MSB* ：(残差 ε_{ij} の分散) + ($\underline{\Sigma N_i a_i^2 / (K - 1)}$)

この部分は，要因によって
生じた分散に対応する．

- 検定統計量は*MSB*と*MSW*の比をとっている．
  - 帰無仮説の下で*MSB*の第2項は0になるが，そのとき，分散比（検定統計量）は1になる．

2元配置のモデル

- 繰り返しのある場合の２元配置のモデル

$$Y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk}$$

μ ：全体の平均

a_i ：要因Aによる効果 ($a_i = \mu_i - \mu$)

b_j ：要因Bによる効果 ($b_j = \mu_j - \mu$)

$(ab)_{ij}$ ：交互作用の効果 ($(ab)_{ij} = \mu_{ij} - \mu_i - \mu_j + \mu$)

ε_{ijk} ：誤差（ノイズ）

- 各効果の線形和として表現されていることに注意．

2元配置のモデル

- 主効果Aについての仮説
  - 帰無仮説H0：all $a_i = 0$
  - 対立仮説H1：not all $a_i = 0$
- 主効果Bについての仮説
  - 帰無仮説H0：all $b_j = 0$
  - 対立仮説H1：not all $b_j = 0$
- 交互作用についての仮説
  - 帰無仮説H0：all $(ab)_{ij} = 0$
  - 対立仮説H1：not all $(ab)_{ij} = 0$

2元配置のモデル

33

- MSW, MSA, MSB, $MSAB$ は次の量の不偏推定量．
  - MSW：(残差 ε_{ijk} の分散)
  - MSA ：(残差 ε_{ijk} の分散) + ($\Sigma N_i^A a_i^2 / (K^A - 1)$)
  - MSB ：(残差 ε_{ijk} の分散) + ($\Sigma N_j^B b_j^2 / (K^B - 1)$)
  - $MSAB$: (残差 ε_{ijk} の分散)
    + ($\Sigma N_{ij} (ab)_{ij}^2 / (K^A - 1)(K^B - 1)$)
- 前ページの三つの帰無仮説の下で，それぞれ対応する検定統計量（分散比）は 1になる．

Within-Subject design の場合

34

- 繰り返しのない場合の２元配置のモデル

  $Y_{ij} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ij}$

  μ ：全体の平均

  a_i ：要因Aによる効果

  b_j ：要因Bによる効果（被験者による差）

  $(ab)_{ij}$ ：交互作用による効果

  ε_{ij}：誤差（ノイズ）

  - 帰無仮説 H0：all $a_i = 0$
  - 対立仮説 H1：not all $a_i = 0$

2元配置のモデル

35

- *MSA*, *MSAB* は次の量の不偏推定量である．
- *MSA* ：(残差 ε_{ij} の分散) +
  + (被験者と実験条件の交互作用による分散)
  + (実験条件の効果による分散）
- *MSAB*: (残差 ε_{ij} の分散)
  + (被験者と実験条件の交互作用による分散)
- 帰無仮説の下で，検定統計量（分散比）は1になる．
- 実験条件の効果による分散が大きくなれば有意になる．

分散分析に関連するその他の事項

36

- 等分散性の確認
  - 分散分析では，各セルが等分散であることが前提なので，事前にそれを検定しておくのが望ましい．
- 各セルのデータ数が異なる場合
  - データ数が違う場合の補正法を使う．
- 多元配置の分散分析
  - 要因の種類が三つ以上の場合にも拡張できる．
  - Within-Subject 型についても，要因が二つ以上の場合を議論できる．
- 2元配置での多重比較