分散分析と線形回帰分析

北海道大学　医学統計学
横田　勲

1

今回の内容

2

- 分散分析
  - 一元配置分散分析
  - 検定の多重性
  - 二元配置分散分析
- 線形回帰分析
  - t検定、分散分析の一般線形モデルによる表現

2

Sir. R. A. Fisher

3

- Rothamsted 農事試験場
  - 小麦やジャガイモの収量
    - 多くの収量を得られる品種は？
    - 品種と肥料の最適な組み合わせを知りたい
    - 気象条件や地力の退行による影響を知りたい
- 品種はいくつもあるため、多群の比較に

3

分散分析 (analysis of variance; ANOVA)

4

- データのバラツキを要因別に分解し比較
- データのバラツキ：偏差平方和
  - n個のデータ$y_1, \ldots, y_n$のバラツキは、それぞれの偏差の2乗の和$\sum_i^n (y_i - \bar{y})^2$ で測る
- バラツキの理由
  - 実験にとりあげた因子
  - 実験誤差

4

分散分析の基本用語

5

- 反応 response
  - 問題となっている品質特性、結果、測定値
  - 例：収量、強度、血中ヘモグロビン濃度
- 因子 factor
  - 反応のバラツキに影響を与える変数
  - 例：温度、触媒量、薬剤
- 水準 level
  - 因子の取りうる値
  - 例：40度・50度、実薬・プラセボ

5

例：ラットの反復投与実験

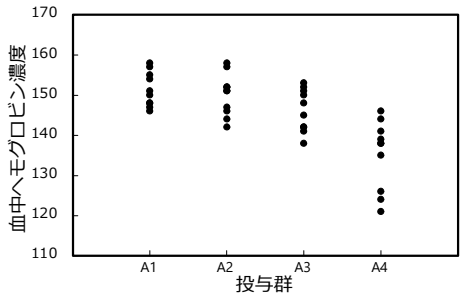
6

- ある薬物Aの影響を調べるために、ラット40匹を10匹ずつの4群に分け、薬剤の静脈内投与を35日続けた
  - A1 : 対照群　　A2 : 薬物Aを5mg/kg
  - A3 :薬物Aを10mg/kg　A4 : 薬物Aを20mg/kg
- 反応：血中ヘモグロビン濃度 (mg/dl)
- 因子：薬物濃度（投与群）
- 水準：A1, A2, A3, A4の4水準

6

ヘモグロビン濃度の分布

7



7

分散分析の前提

8

- データに外れ値はない
- 分布の形、広がりは大体同じ
- 水準間での分布の違い
  ＝水準間での反応の平均値の違い

8

Notation

9

- y_{ij}
  - 第i水準のj番目の反応
    - $i = 1, \ldots, a$(水準の数)
    - $j = 1, \ldots, n$（各水準の人数）
      - nは水準ごとに異なってもよい
- $\bar{y}_{i\cdot} = \sum_j^n y_{ij} \,/n$
  - 第i水準の平均値
- $\bar{y}_{\cdot\cdot} = \sum_i^a \sum_j^n y_{ij} \,/(an)$
  - 全反応の平均値

9

平方和の分解

10

$$\sum_i^a \sum_j^n (y_{ij} - \bar{y}_{\cdot\cdot})^2 = n \cdot \sum_i^a (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 + \sum_i^a \sum_j^n (y_{ij} - \bar{y}_{i\cdot})^2$$

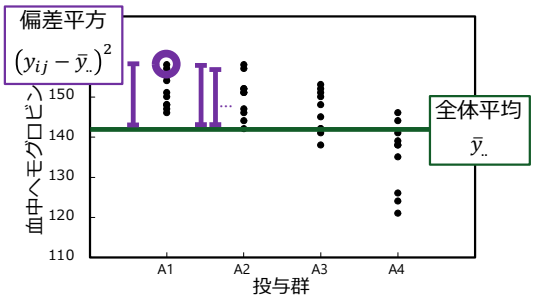
	総平方和S_T	群間平方和S_M	群内平方和S_E
自由度	$an-1$	$a-1$	$a(n-1)$

- 群間平方和：水準間での反応のバラツキ
- 群内平方和：水準内での反応のバラツキ

10

総平方和

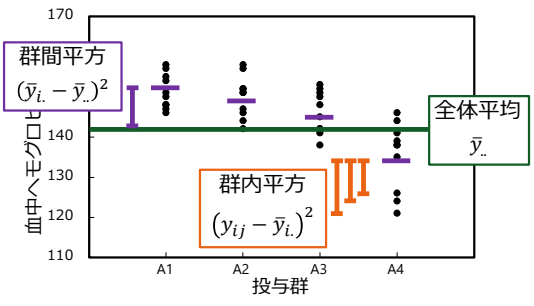
11



11

群間平方和と群内平方和

12



12

平均平方による比較

13

- 平方和を自由度で割った統計量
  - 分散に同じ
- 群間平均平方（モデル平均平方）

$$V_M = \frac{S_M}{a-1}$$

- 群内平均平方（誤差平均平方）

$$V_E = \frac{S_E}{a(n-1)}$$

13

要因効果のF検定

14

- 帰無仮説：水準間で反応に差がない
- 以下の統計量が自由度$(a-1, a(n-1))$のF分布に従う

$$F = \frac{V_M}{V_E}$$

  - 2群比較の場合、F はt検定統計量の2乗に一致

14

分散分析表

15

要因	平方和	自由度	平均平方	F値	P値
モデル	S_M	$a-1$	V_M	V_M/V_E	
誤差	S_E	$a(n-1)$	V_E		
全体	S_T	$an-1$			

15

ラットの例

16

要因	平方和	自由度	平均平方	F値	P値
モデル	1614.8	3	538.3	14.49	<0.01
誤差	1337.6	36	37.2		
全体	2952.4				

- 4群のどこかで
  ヘモグロビン濃度に違いがあった
  - で、どことどこが違ったの？

16

それぞれの群間で比較

17

$$t = \frac{\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}}{\sqrt{\left(1/n_i + 1/n_{i'}\right)V_E}},$$

- それぞれt検定をしてみよう
  - 併合分散s^2を誤差の平均平方V_Eで置き換え
- A1とA2で検定、A1とA3で検定、
  A1とA4で検定、A2とA3で検定.....

17

仮説検定での2つのエラー

18

- αエラー (type-I エラー,第一種の過誤)
  - 本当は差がないのに有意差ありという誤り
    - 消費者リスク
    - (**A**watenbo)あわてんぼうさんの間違い
- βエラー (type-II エラー,第二種の過誤)
  - 本当は差があるのに有意差を出せない誤り
    - 生産者リスク
    - (**B**onyari)ぼんやり者の間違い

18

検定を繰り返すと・・ 19

- 第一種の過誤が保たれない
  - 1回あたりの検定は片側2.5%水準で行っても、「全体の有意水準」がこれを超えてしまう

19

全く効果のない薬を開発 20

- あくまでこの薬は効かない(差はない)
  - 効くかどうか分からないので治験をやる
- いつもは1回だけ検定を行う
  - 40回に1回は誤って有意差あり(αエラー)
- とにかく10回検定をやりまくる
  - 1回でも有意差があれば大喜び！
  - 10回とも有意差なしとなる確率は0.975^{10}
  - 少なくとも1回有意差ありと言ってしまう確率

$$1 - 0.975^{10} = 22.4\% \gg 2.5\%$$

20

有意水準を厳しくしよう！ 21

- Bonferroni法
  - K回検定を繰り返すならば、有意水準をα/K
  - $1 - \left(1 - \frac{\alpha}{K}\right)^K < \alpha$
- 40回に1回くらいなら、効かない薬が世の中に出てもまあいい
  - 10回検定を行うなら、有意水準を10で割ればよい(0.025/10=0.0025)

$$1 - 0.9975^{10} = 2.47\% < 2.5\%$$

21

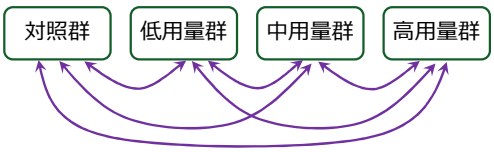
Bonferroni法の欠点 22

- 有意水準を厳しくするため、βエラーをかなり増やしてしまう
  - 本当は差があるのに、有意差ありといえない場合が増えてしまう
- 必要な検定のパターンに応じ、様々な改良法が提案されている
  - なかには第一種の過誤をきちんと制御できない方法もある
    - しかも古い教科書では使用をすすめていることも

22

Tukey法 23

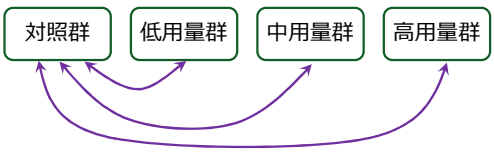
- すべての群に関する対比較に興味
  - この例では${}_4C_2$=6通りの比較



23

Dunnett法 24

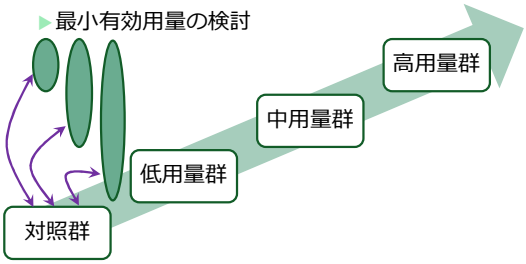
- 対照群との比較に興味
  - この例では3通りの比較



24

Williamsの多重比較 25

- 対照群との比較、かつ、
- どの用量以上で差がみられるかに興味
  - 最小有効用量の検討



25

Fisher's Least Significant Difference 法 26

- 分散分析における要因全体のF検定を利用
  - F検定がある有意水準で有意であった場合、水準間の比較を同じ有意水準で行う
- 3水準(3群)の比較の場合のみ妥当
  - 4水準以上の場合は、水準間の比較をそのままの有意水準で行ってはならない

26

最近使われる多重比較法 27

- ブートストラップ(bootstrap)法
- 並べ替え(permutation)法
  - 計算負荷の大きい手法が利用できるようになったため、近年増加
  - データの正規性からのかい離にも対応

27

一元配置分散分析 28

- 今までの例は、多群の比較
- 因子は1つだった

- 因子が2つ以上になったら？
  - 品種と肥料
  - アスピリンとβカロテン
    - Physicians' Health Study
  - 治療と個体差
    - クロスオーバー試験

28

実験計画法 29

- 与えられた実験目的に対して、どのような実験をするのが最も効果的か
  - 実験回数が少ないが正しい情報を得るには
- 誤差の含まれる実験データから正しい結論を引き出すためには、データをどのように解析すべきか

鷲尾泰俊. 実験の計画と解析. 岩波書店. 1988.

29

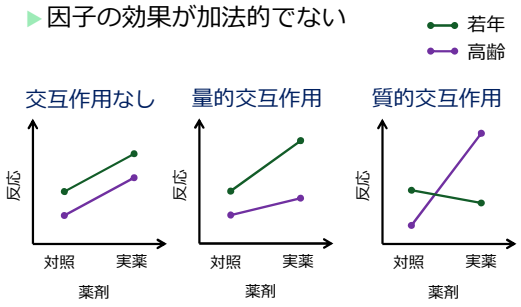
単一因子実験 single factor experiment 30

- 順番に因子ごとの最適水準を決める方法
- 例：年齢と薬剤（対照／実薬）が因子
  - 若年の患者で、薬剤2つを比較
    - 例えば「実薬」が選ばれた
  - 薬剤を実薬に固定し、若年と高齢で比較
- 交互作用がある場合、誤った結論のおそれ
  - 実験パターンが因子の比較する順番により変化してしまう

30

交互作用がある 31

- 因子の効果が加法的でない



31

要因実験 factorial experiment 32

- 全因子の水準の組合せをすべて行う

単一因子実験

	対照	実薬
若年	○	○
高齢	○	×

要因実験

	対照	実薬
若年	○	○
高齢	○	○

32

因子が多い場合 33

- 要因実験では通り数が膨大に
  - 10因子あったら各2水準でも$2^{10} = 1024$通り
  - 多数回の実験を行うコストを抑えたい
  - 多数回の実験を均等な条件で行うことが困難
- 直交法 orthogonal arrayの利用
  - 代表的な、ラテン方格

33

実験順序のランダム化 34

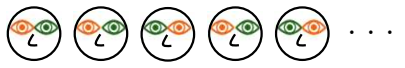
- 対照薬をまず100例に投与して、次の100例に実薬を投与
  - 実薬のほうがよかったとしても、
    - 薬剤の効果
    - 時期効果（投与順序の効果）

    を分離できない
- ランダム化により、実験誤差は確率的変動をするものとして扱える

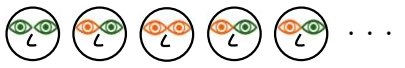
34

実験の場の管理（局所管理） 35

- 実験誤差を小さくするための工夫
- 実験環境が同じになるようなブロック
  - 乱塊法 randomized block design
  - 臨床試験では層別ランダム化 stratified randomized



  - 完全ランダム化 completely randomized design



35

実験計画法の基本的原則 36

- Fisherの3原則とも
- ランダム化
- 局所管理
- 反復
  - 実験誤差と因子の効果（要因効果）を分離するため
    - そのための解析方法として、分散分析

36

多元配置分散分析 37

- 因子の数だけモデル平方和を分割
- 因子の組み合わせによって
  特異的に反応が変わることへも対処可能
  - 交互作用がある

37

対応のあるデータへの対処 38

- 個人やペアを（ブロック）因子に追加
  - 多元配置分散分析のひとつ
  - 因子は群間差とブロック因子の2つ
    - 各ブロック因子の水準で平均値は異なってよい
  - 同じ個人に3回以上測定する場合にも対処可

38

中心角膜厚データ（一部） 39

- 術後24週と2年で違いがあるか

ID	術後24週	術後2年
1	511	532
2	525	538
3	540	546
4	640	710
5	509	529
6	505	525
7	626	550
8	489	503
9	595	543
10	539	523
11	561	572

Kinoshita S, *et al. N Engl J Med.* 2018; 995-1003.

39

対応のあるt検定と分散分析 40

- t検定
  - $t = \frac{2.82}{\sqrt{39.13^2/11}} = 0.24$, 両側P値：**0.82**
- 患者を因子に加えた分散分析

要因	平方和	自由度	平均平方	F値	P値
モデル	48775.7	11	4434.2	5.79	0.005
時点	43.7	1	43.7	0.06	**0.82**
ID	48732.0	10	4873.2	6.37	0.004
誤差	7655.8	10	765.6		
全体	56431.5				

40

分散分析の欠点 41

- 因子は離散的なもののみ
  - 治療群、性別、・・・
- 年齢や血清マーカーのような
  連続量の影響も検討したい

線形回帰分析の利用

41

線形回帰分析 42

- 対象者$k(=1,\dots,n)$に対する単回帰分析
  - k番目の対象者の反応 Y_k
  - 説明変数 X_k
    - 治療群、性別、年齢、血清マーカー、・・・
  - 回帰パラメータ α, β
  - ランダム誤差 $\varepsilon_k \sim N(0, \sigma^2)$
- 以下の足し算を用いた方程式を置く

$$Y_k = \alpha + X_k\beta + \varepsilon_k$$

42

一般線形モデル general linear model 43

- 回帰分析のみならず、t検定や分散分析も統一的に扱うことが可能
- t検定の場合
  - 治療群が1のとき$X_k = 1$、群が2のとき$X_k = 0$

$$Y_k = \alpha + X_k\beta + \varepsilon_k$$

  - 誤差ε_kは期待値がゼロ

43

期待値をとってみよう 44

- 群1であるkさん　$Y_k = \alpha + 1 \times \beta + \varepsilon_k$
- 群1全員の期待値（平均値）は、

$$\frac{(\alpha + \beta + \varepsilon_1) + (\alpha + \beta + \varepsilon_2) + \cdots}{n_1} = \alpha + \beta$$

（∵ランダム誤差は期待値が0）

- 同様に、群2の期待値は、

$$\frac{(\alpha + \varepsilon_1) + (\alpha + \varepsilon_2) + \cdots}{n_2} = \alpha$$

44

t検定の場合 45

- $Y_k = \alpha + X\beta + \varepsilon_k$
  - α：群2の平均値
  - β：群間差
- $\beta = 0$であるかの検定がt検定

45

一元配置分散分析の場合 46

- 治療群が4つ(1,2,3,4)あるならば、
  - 治療群が1のとき$X_1 = 1$、それ以外$X_1 = 0$
  - 治療群が2のとき$X_2 = 1$、それ以外$X_2 = 0$
  - 治療群が3のとき$X_3 = 1$、それ以外$X_3 = 0$
  - 治療群が4のとき$X_4 = 1$、それ以外$X_4 = 0$

$$Y_k = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + \varepsilon_k$$

- $\beta_1 = \beta_2 = \beta_3 = \beta_4$であるかの検定が$F$検定

46

バラツキの分解という観点 47

- 一般線形モデルでも同じ
- でも真値は分からない
  - そもそもあまり興味もない
  - とりあえず基準点を置いて、切片と呼ぼう
    - “一元配置分散分析の場合”のスライドのように、全てモデル因子で分解し、切片を置かなくたっていい

測定値 (measurement) ＝ 切片 (intercept) ＋ モデル因子 (model factor) ＋ 誤差 (error)

$$Y_k = \alpha + X_k\beta + \varepsilon_k$$

47

モデルの構築 48

- 例えば、対照群との比較に興味があれば

$$Y_k = \alpha + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \varepsilon_k$$

  - X_1：治療群がA2の時だけ1、それ以外は0
  - X_2：治療群がA3の時だけ1、それ以外は0
  - X_3：治療群がA4の時だけ1、それ以外は0

48

解析結果の見方① 49

- 解析ソフトSASの出力

パラメータ	推定値	標準誤差	t 値	Pr > |t|	95% 信頼限界	
Intercept	151.4000000	1.92757764	78.54	<.0001	147.4906914	155.3093086
A2	−1.4000000	2.72600644	−0.51	0.6107	−6.9285973	4.1285973
A3	−5.2000000	2.72600644	−1.91	0.0645	−10.7285973	0.3285973
A4	−16.2000000	2.72600644	−5.94	<.0001	−21.7285973	−10.6714027

- 整形して

治療群	推定値	95%信頼区間	両側P値
A2 vs A1	-1.4	(-6.9, 4.1)	0.61
A3 vs A1	-5.2	(-10.7, 0.3)	0.06
A4 vs A1	-16.2	(-21.7, -10.7)	<0.01

49

解析結果の見方② 50

治療群	推定値	95%信頼区間	両側P値
A2 vs A1	-1.4	(-6.9, 4.1)	0.61
A3 vs A1	-5.2	(-10.7, 0.3)	0.06
A4 vs A1	-16.2	(-21.7, -10.7)	<0.01

- 治療群A2はA1に比べ、
  ヘモグロビン濃度の平均の差は-1.4
  - A2のヘモグロビン濃度のほうが低い
  - その信頼区間は(-6.9, 4.1)

50

薬物濃度を連続量として扱う 51

- 薬物Aを与えるごとに、
  どれだけヘモグロビン濃度は低下するか？
  - A1～A4は0,5,10,20mg/kg投与
  - 薬物増加に従い、
    ヘモグロビン濃度は直線的に変化すると仮定
- Xは薬物Aの濃度として、モデルを構築

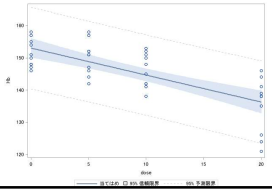
$$Y_k = \alpha + X\beta + \varepsilon_k$$

51

回帰直線の推定 52

パラメータ	推定値	標準誤差	t 値	Pr > |t|	95% 信頼限界	
Intercept	153.0400000	1.49372220	102.46	<.0001	150.0161175	156.0638825
dose	−0.8388571	0.13038276	−6.43	<.0001	−1.1028032	−0.5749110

- 薬物を1mg/kg増やすごとに、
  ヘモグロビン濃度は-0.84mg/dLだけ変化



52

線形回帰モデル 53

- 反応Yの変化を、
  説明変数（共変量）X で説明
  - 説明変数が群（カテゴリカル）であれば、
    群間差が求まる
  - 説明変数が連続量であれば、
    1単位変化あたりの反応の変化が求まる

53

まとめ 54

- 分散分析
  - 平方和をモデル平方和と誤差に分解
- 線形回帰分析
  - t検定や分散分析を含む、解析を広くカバー
  - 説明変数の作り方を工夫して
    回帰係数を解釈可能なものに
    - 回帰係数は
      「説明変数が1単位変化したときの平均値の差」

54