

2020/4/20 北大SPH・統計解析の基礎⑤

## 分散分析と 線形回帰分析

北海道大学 医学統計学  
横田 勲

## 今回の内容

- ▶ 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

## Sir. R. A. Fisher



- ▶ 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

## 分散分析 (analysis of variance; ANOVA)

- ▶ 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  $y_1, \dots, y_n$  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

## 分散分析の基本用語

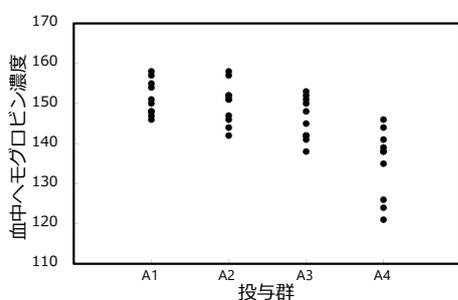
- ▶ 反応 response
  - ▶ 問題となっている品質特性、結果、測定値
  - ▶ 例：収量、強度、血中ヘモグロビン濃度
- ▶ 因子 factor
  - ▶ 反応のバラツキに影響を与える変数
  - ▶ 例：温度、触媒量、薬剤
- ▶ 水準 level
  - ▶ 因子の取りうる値
  - ▶ 例：40度・50度、実薬・プラセボ

## 例：ラットの反復投与実験

- ▶ 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

## ヘモグロビン濃度の分布

7



## 分散分析の前提

8

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

## Notation

9

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

## 平方和の分解

10

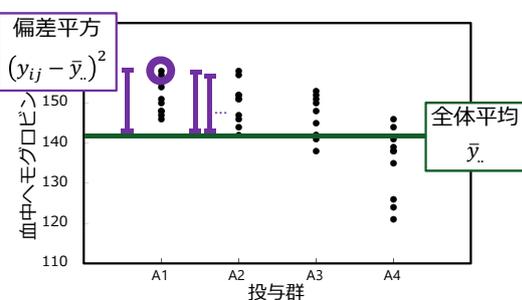
$$\sum_i^a \sum_j^n (y_{ij} - \bar{y})^2 = n \cdot \sum_i^a (\bar{y}_i - \bar{y})^2 + \sum_i^a \sum_j^n (y_{ij} - \bar{y}_i)^2$$

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

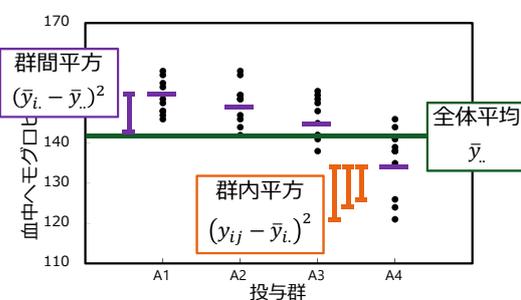
## 総平方和

11



## 群間平方和と群内平方和

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)

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

## 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

## 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	$S_T$	$an-1$			

## 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?

## それぞれの群間で比較

17

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

- ▶ 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.....

## 仮説検定での2つのエラー

18

- ▶  $\alpha$ 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
    - ▶ (Awatenbo)あわてんぼうさんの間違い
- ▶  $\beta$ 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
    - ▶ (Bonyari)ぼんやり者の間違い

## 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

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

## 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 $\alpha/K$
  - ▶  $1 - \left(1 - \frac{\alpha}{K}\right)^K < \alpha$
- ▶ 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( $0.025/10=0.0025$ )  
 $1 - 0.9975^{10} = 2.47\% < 2.5\%$

## Bonferroni法の欠点

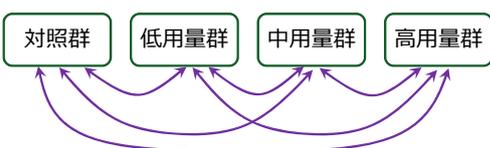
22

- ▶ 有意水準を厳しくするため、 $\beta$ エラーをかなり増やしてしまう
  - ▶ 本当は差があるのに、有意差ありとし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

## Tukey法

23

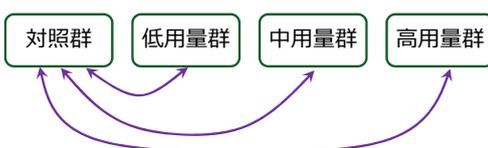
- ▶ すべての群に関する対比較に興味
  - ▶ この例では ${}_4C_2=6$ 通りの比較



## Dunnnett法

24

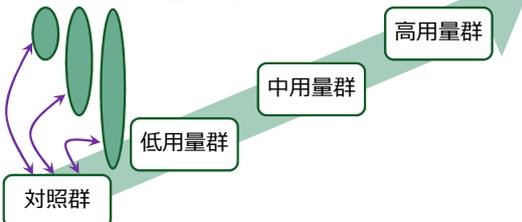
- ▶ 対照群との比較に興味
  - ▶ この例では3通りの比較



## 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



## 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、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

## 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

## 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 $\beta$ カロテン
    - ▶ Physicians' Health Study
  - ▶ 治療と個体差
    - ▶ クロスオーバー試験

## 多元配置分散分析

29

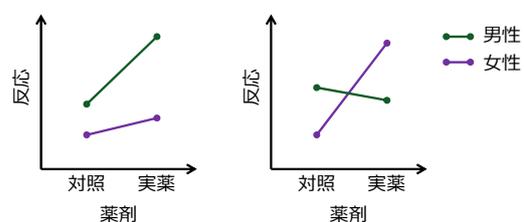
- ▶ 因子の数だけモデル平方和を分割
- ▶ 因子の組み合わせによって特異的に反応が変わることへも対処可能
  - ▶ 交互作用がある

## 交互作用

30

量的交互作用

質的交互作用



## 対応のあるデータへの対処

31

- ▶ 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

## 中心角膜厚データ（一部）

32

- ▶ 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.

## 対応のあるt検定と分散分析

33

- ▶ t検定

$$\text{▶ } t = \frac{2.82}{\sqrt{39.13^2/11}} = 0.24, \text{ 両側P値: } \mathbf{0.82}$$

- ▶ 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	<b>0.82</b>
ID	48732.0	10	4873.2	6.37	0.004
誤差	7655.8	10	765.6		
全体	56431.5				

## 分散分析の欠点

34

- ▶ 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

## 線形回帰分析

35

- ▶ 対象者 $k(= 1, \dots, n)$ に対する単回帰分析
  - ▶  $i$ 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  $\alpha, \beta$
  - ▶ 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

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

## 一般線形モデル general linear model

36

- ▶ 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 $\varepsilon_k$ は期待値がゼロ

## 期待値をとってみよう

37

- ▶ 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、  

$$\frac{(\alpha + \beta + \varepsilon_1) + (\alpha + \beta + \varepsilon_2) + \dots}{n_1} = \alpha + \beta$$
（∵ランダム誤差は期待値が0）
- ▶ 同様に、群2の期待値は、  

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

## t検定の場合

38

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

## 一元配置分散分析の場合

39

- ▶ 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

## バラツキの分解という観点

40

- ▶ 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
    - ▶ 90枚目のように、全てモデル因子で分解して、切片を置かなくたっていい

測定値 = 切片 + モデル因子 + 誤差

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

## モデルの構築

41

- ▶ 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

## 解析結果の見方①

42

- ▶ 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

## 解析結果の見方②

43

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)

## 薬物濃度を連続量として扱う

44

- ▶ 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

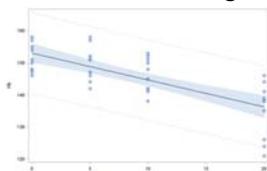
$$Y_k = \alpha + X\beta + \varepsilon_k$$

## 回帰直線の推定

45

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



## 線形回帰モデル

46

- ▶ 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

## まとめ

47

- ▶ 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」

F分布  $\alpha=0.05$ 

分母の 自由度	分子の自由度											
	1	2	3	4	5	6	7	8	9	10	15	20
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	245.95	248.01
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.45
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.78
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.84	1.75
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.81	1.72
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.79	1.70
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.78	1.69
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77	1.68
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.72	1.62
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.70	1.61
400	3.86	3.02	2.63	2.39	2.24	2.12	2.03	1.96	1.90	1.85	1.69	1.60
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.69	1.59
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.68	1.58