分散分析

樋口さぶろお http://hig3.net

龍谷大学理工学部数理情報学科

生活の中の統計技術　L09(2018-12-03 Mon)

最終更新: Time-stamp: "2018-12-03 Mon 10:08 JST hig"

今日の目標

- 対応のない/ある 1 要因の分散分析の考え方が説明できる
- 手で, 自由度, 級内分散, 級間分散が計算できる

分散分析とは

分散分析=ANOVA=Analysis of Variance
ある一群の試験紙.

- 多群の間に平均値の差があるかを判定する.
- (カテゴリ変数, 量的変数) のデータで, 量的変数がカテゴリ変数に依存してるかどうかを判定する.

この 2 つは同じこと.

学級	点数
カテゴリ	量的
A	78
A	79
A	79
A	80
B	78
⋮	⋮

質的変数と量的変数

量的変数

実数 整数. これまで扱ってた変数 or データ X.

質的変数

質的変数=カテゴリ (カル) 変数

例 $Y \in \{$A 型, B 型, AB 型, O 型 $\}$. 名義尺度

例 $Y \in \{$ 優, 良, 可, 不可 $\}$. 順序尺度

ここまで来たよ

7 分散分析
- 対応のない 1 要因の分散分析
- 対応のある 1 要因の分散分析

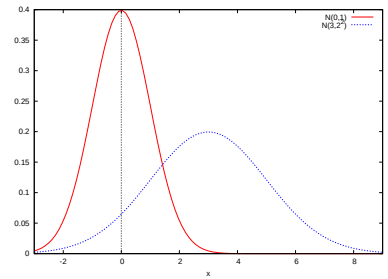
対応のない 1 要因の分散分析

いろいろやった後, 最終的に分散の大きさを比較するために, F 検定という統計的仮説検定を行う手続き.

与えられたデータ 各組で異なる教え方をした. ランダムに何人かよんできてテストをした.

j	学級	点数
1	A 組	78 79 79 80
2	B 組	78 86 81 83 82
3	C 組	86 85 87

仮定 各学級のデータは, 正規分布 $\mathrm{N}(\mu + a_j, \sigma^2)$ にしたがう. (学級番号 $j = 1, 2, 3$).



問 母集団の点数の平均値は, 学級によって異なるか?

分散分析の問

問 母集団の点数の平均値は, 学級によって異なるか?
試験紙で言うと,

変色しない=No=帰無仮説　平均値はすべての学級で同じ
変色した=Yes=検定で帰無仮説を棄却　平均値が異なる学級がある (どのペアか, またはぜんぶか, は問わない)

もし 2 群=2 学級だけだったら, 平均値の差の検定
多群=2 学級以上何学級でも, 分散分析

分散分析表を作る準備

学級別の表

i	学級	データ	個数	級内平均	不偏標本分散
1	A 組	78 79 79 80	4	79	$\frac{1}{4-1}[(78-79)^2+\cdots]$
2	B 組	78 86 81 83 82	5	82	
3	C 組	86 85 87	3	85	
計			12	82	

分散分析の用語と記号

「学級」＝水準＝level＝群＝級.

データ y_{ji} ＝ $y_{\text{水準番号 データ番号}}$

水準	データ	個数	級内平均	
A_1	$y_{11}, \ldots, y_{1n_1}$	n_1	$\overline{y}_{1\bullet}$	$\sum_i (y_{1i} - \overline{y}_{1\bullet})^2$
A_2	$y_{21}, y_{22}, \ldots, y_{2n_2}$	n_2	$\overline{y}_{2\bullet}$	$\sum_i (y_{2i} - \overline{y}_{2\bullet})^2$
$\vdots$				
A_ℓ	$y_{\ell 1}, y_{\ell 2}, \ldots, y_{\ell n_\ell}$	n_ℓ	$\overline{y}_{\ell\bullet}$	$\sum_i (y_{\ell i} - \overline{y}_{\ell\bullet})^2$
		n	全平均 $\overline{y}_{\bullet\bullet}$	級内平方和 S_{E}
		$= \sum_j n_j$	$= \frac{1}{n} \sum_j y_{ji}$	$= \sum_j \sum_i (y_{ji} - y_{j\bullet})^2$

• はその添字で平均したという意味.

n:データ (点数) の総数

ℓ:学級＝級の個数

分散分析表を作る手順

級内平均 $\overline{y}_{j\bullet} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ji}$.

全平均 $\overline{y}_{\bullet\bullet} = \frac{1}{n}\sum_{j=1}^{\ell}\sum_{i=1}^{n_j} y_{ji}$.

級間平方和 (学級の効果=縦のちらばりの合計)

$$S_{\mathrm{A}} = \sum_{j=1}^{\ell}\sum_{i=1}^{n_j}(\overline{y}_{j\bullet} - \overline{y}_{\bullet\bullet})^2 = \sum_{j=1}^{\ell} n_j \times (\overline{y}_{j\bullet} - \overline{y}_{\bullet\bullet})^2$$

残差 (級内) 平方和 (E_{ij} の効果=横のちらばりの合計)

$$S_{\mathrm{E}} = \sum_{j=1}^{\ell}\sum_{i=1}^{n_j}(y_{ji} - \overline{y}_{j\bullet})^2$$

全平方和 (すべてのちらばりの合計) 実は $S_{\mathrm{A}} + S_{\mathrm{E}} = S_{\mathrm{T}}$.

$$S_{\mathrm{T}} = \sum_{j=1}^{\ell}\sum_{i=1}^{n_j}(y_{ji} - \overline{y}_{\bullet\bullet})^2$$

1 元配置の分散分析表

分散分析表

変動要因	平方和	自由度	平均平方	F
級間	S_{A}	$\phi_{\mathrm{A}} = \ell - 1$	$V_{\mathrm{A}} = S_{\mathrm{A}}/\phi_{\mathrm{A}}$	V_{A}/V_E
残差	S_{E}	$\phi_{\mathrm{E}} = n - \ell$	$V_{\mathrm{E}} = S_{\mathrm{E}}/\phi_{\mathrm{E}}$	
全	S_{T}	$\phi_{\mathrm{T}} = n - 1$		

級間に差がない $\Leftrightarrow$ 級間平均平方も級内平均平方も大差ない $\Leftrightarrow$ F が大きくない.

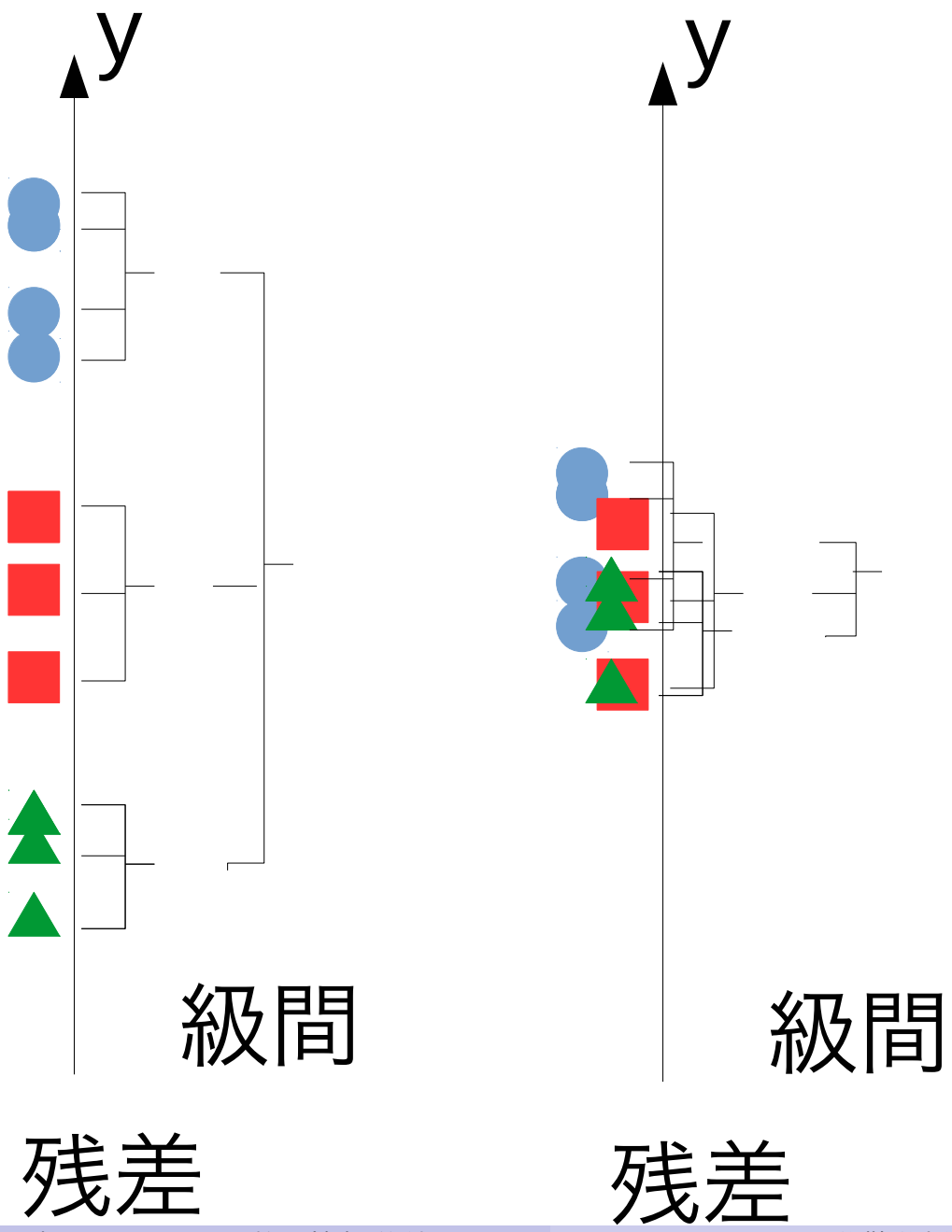
帰無仮説のもとで, $F = \frac{V_{\mathrm{A}}}{V_{\mathrm{E}}} = \frac{S_{\mathrm{A}}/(\ell-1)}{S_{\mathrm{E}}/(n-\ell)}$ は自由度 $(\ell - 1, n - \ell)$ の F 分布にしたがう (**).

$\boxed{\text{全平方和}} = \boxed{\text{級間平方和}} + \boxed{\text{残差平方和}}$.
第 1 項と第 2 項の大きさの比較をするのが分散分析.

L09-Q1

分散分析

上の場合に対して, 3 つの平方和を求め, 分散分析表を作ろう


y
級間
残差
y
級間
残差

Excel で分散分析

準備:統計ツールを有効化 動画

ファイル > オプション > アドイン > Excel のアドイン > 設定 > 分析ツール にチェックを入れて OK する.

平均値の信頼区間 データ > データ分析 > 分散分析: 一元配置

データ方向: このスライドと同じく級内のデータが横に並んでたら 行,

逆なら 列

Excel の出力例

「グループ」＝級

分散分析: 一元配置

概要

グループ	標本数	合計	平均	分散
列 1	4	10	2.5	1.666666667
列 2	4	26	6.5	1.666666667
列 3	4	42	10.5	1.666666667

分散分析表

変動要因	変動	自由度	分散	観測された分散比	P–値	F境界
グループ間	128	2	64	38.4	3.92101E–05	4.2564
グループ内	15	9	1.666666667			
合計	143	11				

「P-値 $<$ あらかじめ設定した有意水準」なら変色. 帰無仮説を棄却.

L09-Q2

Quiz(分散分析)

次のデータに対して, 1 元配置の分散分析表を作ろう. 有意水準 $\alpha = 0.05$ で F 検定しよう.

	水準				
A_1	11	9	12	9	9
A_2	10	17	18	20	10
A_3	25	23	21	22	24

統計検定での出題例

2017 年 6 月 2 級問 14	2017 年 11 月 2 級問 16

ここまで来たよ

対応のある 1 要因の分散分析

十分にそろばんの訓練を積んだ 4 人の生徒に, 昇級試験の模擬試験を 3 回受けさせた. 問題は同じだが, 3 回とも違うタイプのそろばんを使った.

i	そろばんの種類	生徒 1	2	3	4	個数	級内平均	不
1	そろばん A	78	79	79	80	4	79	$\frac{1}{4-1}[(78 -$
2	そろばん B	78	81	83	86	5	82	
3	そろばん C	85	86	86	87	3	86	
計						12	82	

そろばんの違いは点数に影響するか?

生徒の実力による差は, 誤差平均和に入れない処理.

Excel では 分散分析: 繰り返しのない二元配置の

$$\boxed{\text{全平方和}} = \boxed{\text{級間平方和}} + \boxed{\text{ブロック間平方和}} + \boxed{\text{残差平方和}}.$$

級間と残差の大小の比較をするのが「対応のある」1 要因の分散分析

お知らせ

- 来週 2018-12-10 月 2 は実習室で
- レポート 1(長くない)
  - Manaba で振り返りの作文的なもの https://manaba.ryukoku.ac.jp
  - 2018-12-11 火 まで
- 期末試験計画
  - 30 ピーナッツ/科目 100 ピーナッツ
  - 60 分
  - 2019-01-28 月
- レポート計画