データ解析 第13回
分散分析(２)

北九州市立大学経済学部

齋藤 朗宏

Copyright (C) 2007-2013 Akihiro Saito's Lab. All Rights Reserved.

今日の内容

- 繰り返しのある２元配置の分散分析
- 繰り返しのない２元配置の分散分析

2標本検定では
分析不可能な場合

「２標本」ではない場合

- 日本人と韓国人と中国人では平均身長は等しいのか？
- A組とB組とC組ではテストの点は違うか？
- A組男子，A組女子，B組男子，B組女子ではテストの点は違うのか？
- これらの問題は，ｔ検定では解決できない．

繰り返しのある
2元配置の分散分析

繰り返しのある２元配置の分散分析

	A組	B組	C組
男	6	2	2
男	5	2	1
男	4	3	3
女	6	5	3
女	8	3	2
女	7	3	7

この３つのクラス間，性別間で
テストの得点に差があるだろうか？
A組：平均6点，分散2
B組：平均3点，分散1.2
C組：平均3点，分散4.4
男子：平均3.11点，女子：平均4.89点
18人全体：平均4点，分散4.35

繰り返しのある２元配置の分散分析

	A組	B組	C組
男	6	2	2
男	5	2	1
男	4	3	3
女	6	5	3
女	8	3	2
女	7	3	7

4点=5（A組男子の平均点）-1（個人差）
5=4（18人の平均点，全平均）
+1（A組男子の全平均からの差）
1=2（A組の全平均からの差）
-0.89（男子の全平均からの差）
-0.11（A組かつ男子であることの効果）

繰り返しのある２元配置の分散分析

	A組	B組	C組
男	6	2	2
男	5	2	1
男	4	3	3
女	6	5	3
女	8	3	2
女	7	3	7

4　　=4（18人の平均点）
　　+2（A組の平均からの差）
　　-0.89（男子の平均からの差）
　　-0.11 （ A組かつ男子であることの効果）
　　-1（個人差）

赤字3つが統計的に意味を持つか？

繰り返しのある２元配置の分散分析

- A組男子，A組女子，B組男子，B組女子， C組男子，C組女子ではテストの点は違うのか？

- ここでは，因子が2つ（クラス，男女）ある．このような場合，2元配置の分散分析を用いる．
- 2元配置の分散分析では，以下のモデルを考える．$y_{ijk} = \mu + a_j + b_k + (ab)_{jk} + e_{ijk}$
- 1因子の場合と比べると，b_k と$(ab)_{jk}$の2つの項が加わっている．$(ab)_{jk}$は交互作用項であり，2つの因子の組み合わせの効果を示す．
  - A組は男子，B組は女子が優秀な場合など．

繰り返しのある２元配置の分散分析

4　=4（15人の平均点）
+2（A組の平均からの差）
-0.89（男子の平均からの差）
-0.11（ A組かつ男子であることの効果）
-1（個人差）

$$y_{ijk} = \mu + a_j + b_k + (ab)_{jk} + e_{ijk}$$

$$4 = 4 + 2 - 0.89 - 0.11 - 1$$

分散分析

$$y_{ijk} = \mu + a_j + b_k + (ab)_{jk} + e_{ijk}$$

- この式は，個々のデータは，全体の平均＋水準の効果＋組み合わせの効果＋誤差の形で表現されることを意味する.

$$y_{ijk} - \mu = a_j + b_k + (ab)_{jk} + e_{ijk}$$

- 以上のように式を変形することで，全体の平均から個々のデータのずれは，水準の効果（グループ間のずれ）と組み合わせの効果，誤差（グループ内のずれ）の和で表現されることも確認できる.

統計的仮説検定

1. 背理法のロジックに従い，証明したいことと逆の仮説を立てる（帰無仮説）．ここでは，帰無仮説，対立仮説は3つある．
   1. 帰無仮説……３つのクラス間には，得点に差がない．
   $$a_A = a_B = a_C = 0$$
   2. 帰無仮説……2つの性別間には，得点に差がない．
   $$b_M = b_F = 0$$
   3. 帰無仮説……クラスと性別については，得点の組み合わせの効果はない．
   $$\begin{aligned}(ab)_{A,M} &= (ab)_{A,F} \\ &= (ab)_{B,M} = (ab)_{B,F} \\ &= (ab)_{C,M} = (ab)_{C,F} = 0\end{aligned}$$

統計的仮説検定

2. 水準の効果，交互作用の散らばりと誤差の散らばりの大きさを利用して仮説を検討する．
   - 誤差の散らばりと比較して，クラス間での散らばりが小さいならば，差があるとは言えない．
   - 誤差の散らばりと比較して，性別の間での散らばりが小さいならば，差があるとは言えない．
   - 誤差の散らばりと比較して，様々な組み合わせの間での散らばりが小さいならば，交互作用があるとは言えない．

   - 水準の効果，交互作用の散らばりが十分に大きいならば，差はないという仮説は棄却される．

統計的仮説検定

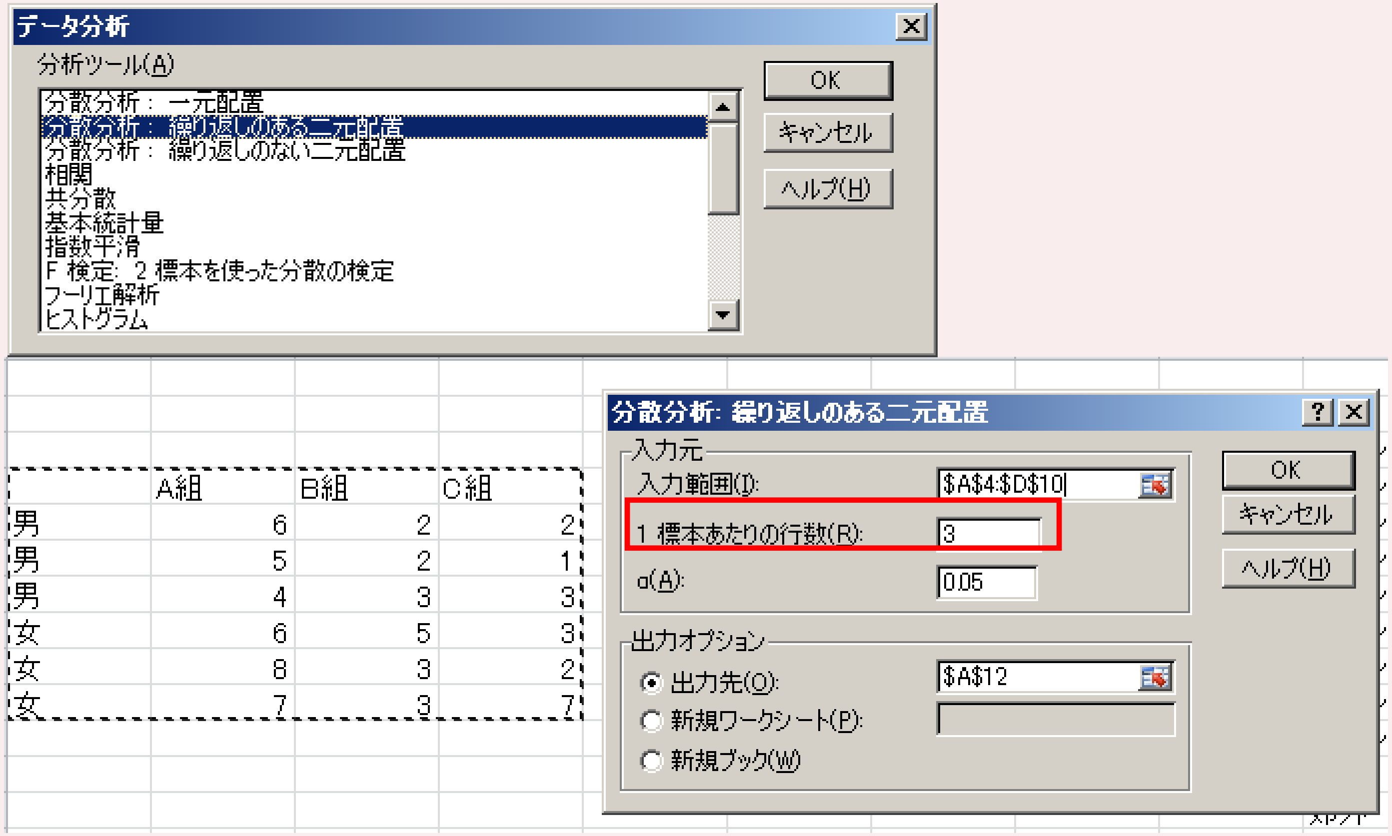
3. 帰無仮説の下で，得られた水準の効果，交互作用の散らばりと誤差の散らばりの比がどれくらい起こりにくいことなのか検討する.
    - 帰無仮説が正しければ，散らばりの比は F分布に従う.

繰り返しのある２元配置の分散分析

	A組	B組	C組
男	6	2	2
男	5	2	1
男	4	3	3
女	6	5	3
女	8	3	2
女	7	3	7

- 「クラス」と「性別」ごとの得点を比較したいなら，たとえばこのようなデータになる．繰り返しは縦に並べる点に注意．
- 「A組」で「男子」のデータは，３つある．

繰り返しのある２元配置の分散分析



- 1つの組み合わせあたり，３つのデータがあるので，「1標本あたりの行数」は３となる.
- 先頭行，列は必ずラベルになるので注意.

分散分析

分散分析：繰り返しのある二元配置						
概要	A組	B組	C組	合計		
男						
標本数	3	3	3	9		
合計	15	7	6	28		
平均	5	2.333333	2	3.111111111		
分散	1	0.333333	1	2.611111111		
女						
標本数	3	3	3	9		
合計	21	11	12	44		
平均	7	3.666667	4	4.888888889		
分散	1	1.333333	7	4.861111111		
合計						
標本数	6	6	6			
合計	36	18	18			
平均	6	3	3			
分散	2	1.2	4.4			
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 境界値
標本	14.22222	1	14.22222	7.314285714	0.0191	4.747225
列	36	2	18	9.257142857	0.0037	3.885294
交互作用	0.444444	2	0.222222	0.114285714	0.893	3.885294
繰り返し誤差	23.33333	12	1.944444			
合計	74	17				

14.22/1.944=7.314

繰り返しのある２元配置の分散分析

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 境界値	
標本	14.22222	1	14.22222	7.314285714	0.0191	4.747225	←性別
列	36	2	18	9.257142857	0.0037	3.885294	←クラス
交互作用	0.444444	2	0.222222	0.114285714	0.893	3.885294	←交互作用
繰り返し誤差	23.33333	12	1.944444				
合計	74	17					

- 帰無仮説（性別で差はない）が正しい時，*F*値（観測された分散比）が7.314を超える確率は0.0191．これは5%以下であり，有意水準5%で差があると言える．
- *F*値7.314は*F*境界値4.747を超えている．よって，*F*値が7.314を超える確率は5%以下であり，有意水準5%で差があると言える．

繰り返しのある２元配置の分散分析

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 境界値	
標本	14.22222	1	14.22222	7.314285714	0.0191	4.747225	←性別
列	36	2	18	9.257142857	0.0037	3.885294	←クラス
交互作用	0.444444	2	0.222222	0.114285714	0.893	3.885294	←交互作用
繰り返し誤差	23.33333	12	1.944444				
合計	74	17					

- 帰無仮説（交互作用はない）が正しい時，*F*値（観測された分散比）が0.114を超える確率は0.893．これは5%以上であり，有意水準5%で交互作用があるとは言えない．
- *F*値0.114は*F*境界値3.885を超えていない．よって，*F*値が0.114を超える確率は5%以上であり，有意水準5%で交互作用があるとは言えない．

交互作用のある場合

	サンフランシスコ	ロスアンジェルス
ダウンタウン	79	95
ダウンタウン	107	99
ダウンタウン	103	70
ダウンタウン	92	116
ダウンタウン	180	170
ダウンタウン	165	145
ダウンタウン	240	205
ダウンタウン	265	200
ダウンタウン	300	210
郊外	75	153
郊外	60	78
郊外	60	75
郊外	94	92
郊外	119	115
郊外	100	155
郊外	102	250
郊外	125	340
郊外	165	380

豊田秀樹著
違いを見抜く統計学
講談社ブルーバックス
より引用.

交互作用のある場合

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 境界値	
標本	2550.25	1	2550.25	0.428120767	0.517588	4.149097	←地区
列	7424.694	1	7424.694	1.246413441	0.272552	4.149097	←都市
交互作用	25546.69	1	25546.69	4.288626767	$\boxed{0.046514}$	4.149097	←組み合わせ
繰り返し誤:	190619.1	32	5956.847				
合計	226140.8	35					

- 交互作用のみ有意になっている.

交互作用のある場合

分散分析：繰り返しのある二元配置			
概要	サンフランシスコ	ロスアンジェルス	合計
ダウンタウン			
標本数	9	9	18
合計	1531	1310	2841
平均	170.1111111	145.5555556	157.8333
分散	6721.611111	2819.277778	4649.441
郊外			
標本数	9	9	18
合計	900	1638	2538
平均	100	182	141
分散	1134.5	13152	8502.941
合計			
標本数	18	18	
合計	2431	2948	
平均	135.0555556	163.7777778	
分散	4998.173203	7867.477124	

- サンフランシスコではダウンタウンの方が平均的に高く，ロスアンジェルスでは郊外の方が平均的に高い.
- これが組み合わせの効果＝交互作用.

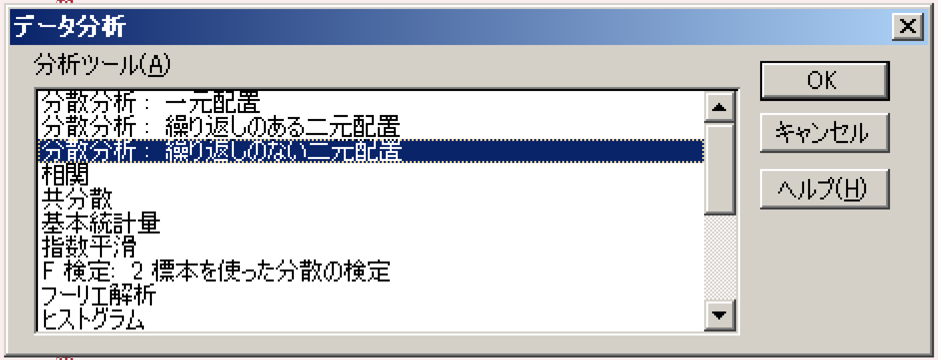
繰り返しのない
2元配置の分散分析

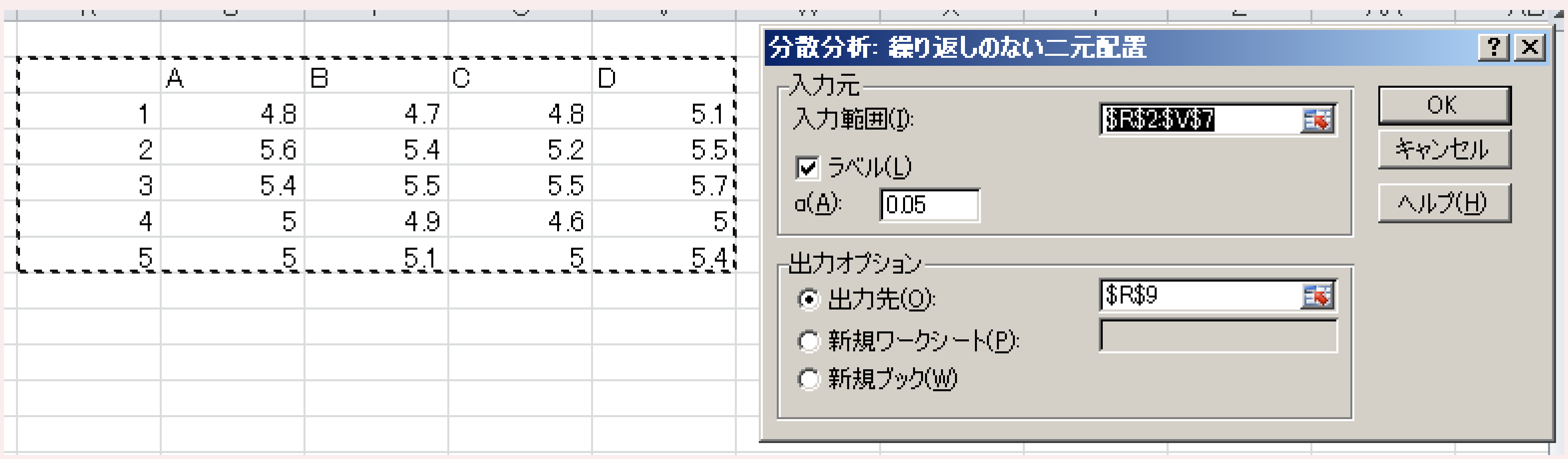
繰り返しのない２元配置の分散分析

	A	B	C	D
1	4.8	4.7	4.8	5.1
2	5.6	5.4	5.2	5.5
3	5.4	5.5	5.5	5.7
4	5	4.9	4.6	5
5	5	5.1	5	5.4

- この表は，ある年のオリンピックにおけるフィギュアスケートの技術点について，AからDまでの審査者が1から5までのスケーターを評価したものである（引用元はホテルデータと同じ，ただし一部抜粋）．
- 要因を「審査者」，「スケーター」と考えると，たとえば審査者Aがスケーター１を評価するのは1回だけで，繰り返しはない．

分散分析


データ分析
分析ツール(A)
分散分析： 一元配置
分散分析： 繰り返しのある二元配置
分散分析： 繰り返しのない二元配置
相関
共分散
基本統計量
指数平滑
F 検定: 2 標本を使った分散の検定
フーリエ解析
ヒストグラム
OK
キャンセル
ヘルプ(H)


A B C D
1 4.8 4.7 4.8 5.1
2 5.6 5.4 5.2 5.5
3 5.4 5.5 5.5 5.7
4 5 4.9 4.6 5
5 5 5.1 5 5.4
分散分析: 繰り返しのない二元配置
入力元
入力範囲(I): $R$2:$V$7
ラベル(L)
α(A): 0.05
出力オプション
出力先(O): $R$9
新規ワークシート(P):
新規ブック(W)
OK
キャンセル
ヘルプ(H)

交互作用

	サンフランシスコ	ロスアンジェルス
ダウンタウン	79	95
ダウンタウン	107	99
ダウンタウン	103	70
ダウンタウン	92	116
ダウンタウン	180	170
ダウンタウン	165	145
ダウンタウン	240	205
ダウンタウン	265	200
ダウンタウン	300	210
郊外	75	153
郊外	60	78
郊外	60	75
郊外	94	92
郊外	119	115
郊外	100	155
郊外	102	250
郊外	125	340
郊外	165	380

繰り返し測定しているので，この中での平均，分散が検討できる．結果，交互作用も調べられる．

	A	B	C	D
1	4.8	4.7	4.8	5.1
2	5.6	5.4	5.2	5.5
3	5.4	5.5	5.5	5.7
4	5	4.9	4.6	5
5	5	5.1	5	5.4

1度しか測定していないので，この中での平均，分散が検討できない．

審査者Aは，他の人と違いスケーター２を一番評価しているが，誤差なのかえこひいき（交互作用）なのかは，1回だけでは判断できない．

繰り返しのない２元配置の分散分析

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 境界値	
行	1.528	4	0.382	26.65116279	6.86E-06	3.259167	←スケーター
列	0.268	3	0.089333	6.23255814	0.008528	3.490295	←審査者
誤差	0.172	12	0.014333				
合計	1.968	19					

- どちらも有意差が見られるので，スケーターごとに技術点の高さに違いがあるし，審査者ごとに技術点の高さに違いがある．つまり，上手いスケーターもいれば下手なスケーターもいるし，点数の甘い審査者もいれば辛い審査者もいる．
- 交互作用が見られない．
  - 一つの組み合わせで1回しか測定しないため，特定の組み合わせが高いのか，誤差なのか区別がつかない．

実習

- 過去6回の夏季五輪の開会式，閉会式の視聴率について，大会ごと，開・閉会式の区分ごとに視聴率に差があるか，分散分析で検討せよ．
- 曜日と工場ごとの不良品の発生数のデータについて，曜日ごと，工場ごとに不良品の発生数は異なるか，分散分析で検討せよ．

- 尚, $\alpha = 0.05$とする.