データ解析 第１２回
分散分析(１)

北九州市立大学経済学部

齋藤 朗宏

今日の内容

- 2標本検定では分析不可能な平均値の比較
- １元配置の分散分析

2標本検定では
分析不可能な場合

２標本検定

- 2つのグループの間で平均値に差があるかを調べるのは，応用上重要な問題である．そういった問題に対する検定のことを2標本検定と呼ぶ．
  - ペアとなる標本を用いた分析．
    - ◆1回目のテストと2回目のテストではどちらの方が成績がいいか？
    - ◆1年生の時に測った身長と3年生になってから測った身長では差はあるか？
  - ペアとはならない2つのグループ間で，グループ間で分散が等しい場合．
  - ペアとはならない2つのグループ間で，グループ間で分散が異なる場合．
    - ◆日本人と韓国人で平均身長は等しいのか？
    - ◆A組とB組でテストの平均点は異なるか？

「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．

１元配置の分散分析

分散分析の考え方

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

この３つのクラス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*

*18*人全体：平均*4*点，分散*4.35*

一見すると差がある．

分散分析の考え方

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

7=6（A組の平均点）+1（個人差）

1=3（C組の平均点）-2（個人差）

分散分析の考え方

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

7=6（A組の平均点）+1（個人差）

6 （A組の平均点）=4（18人の平均）
+2（A組の平均からの差）

1=3（C組の平均点）-2（個人差）

3（C組の平均点）=4（18人の平均）
- 1（C組の平均からの差）

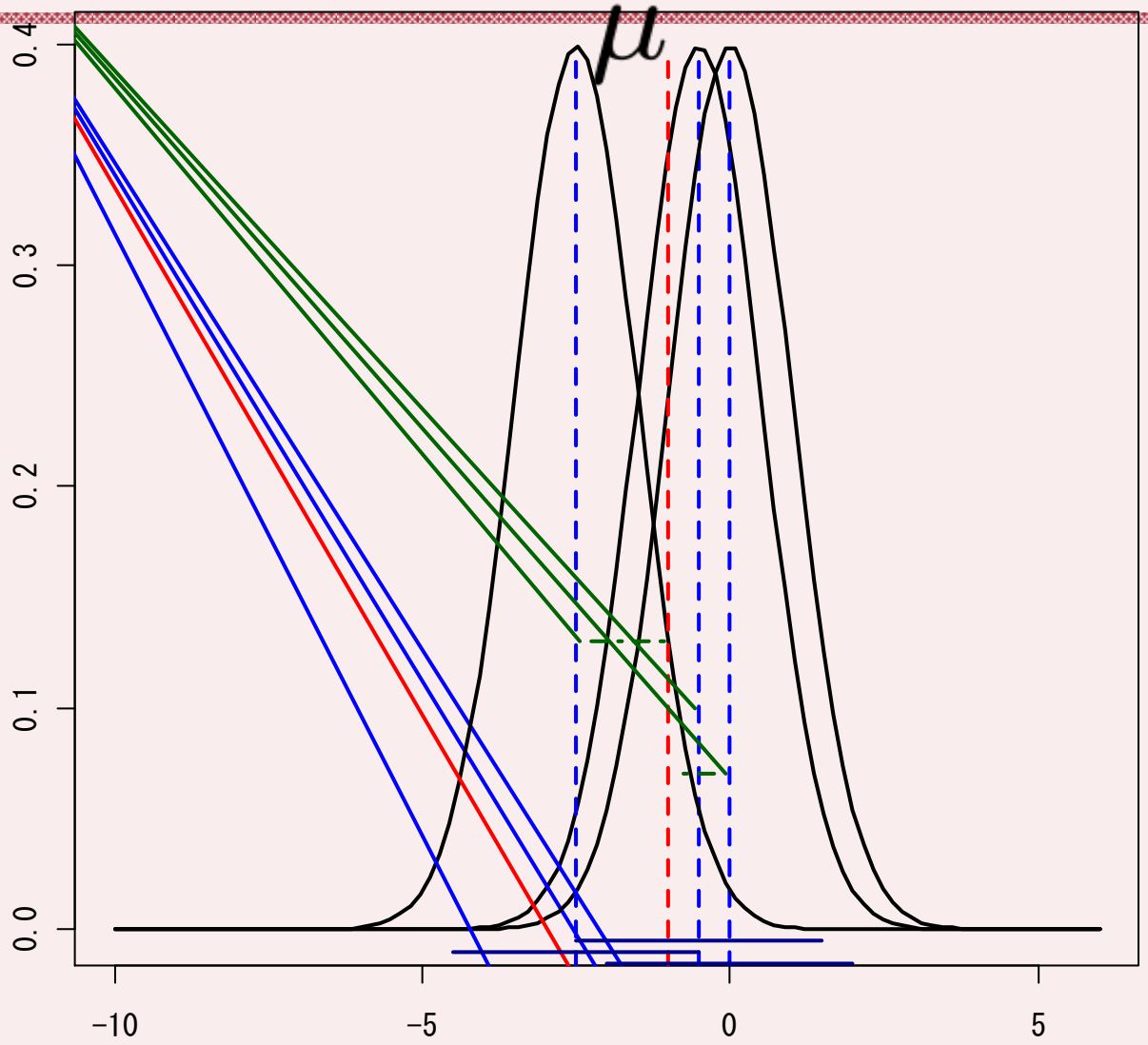
分散分析の考え方

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

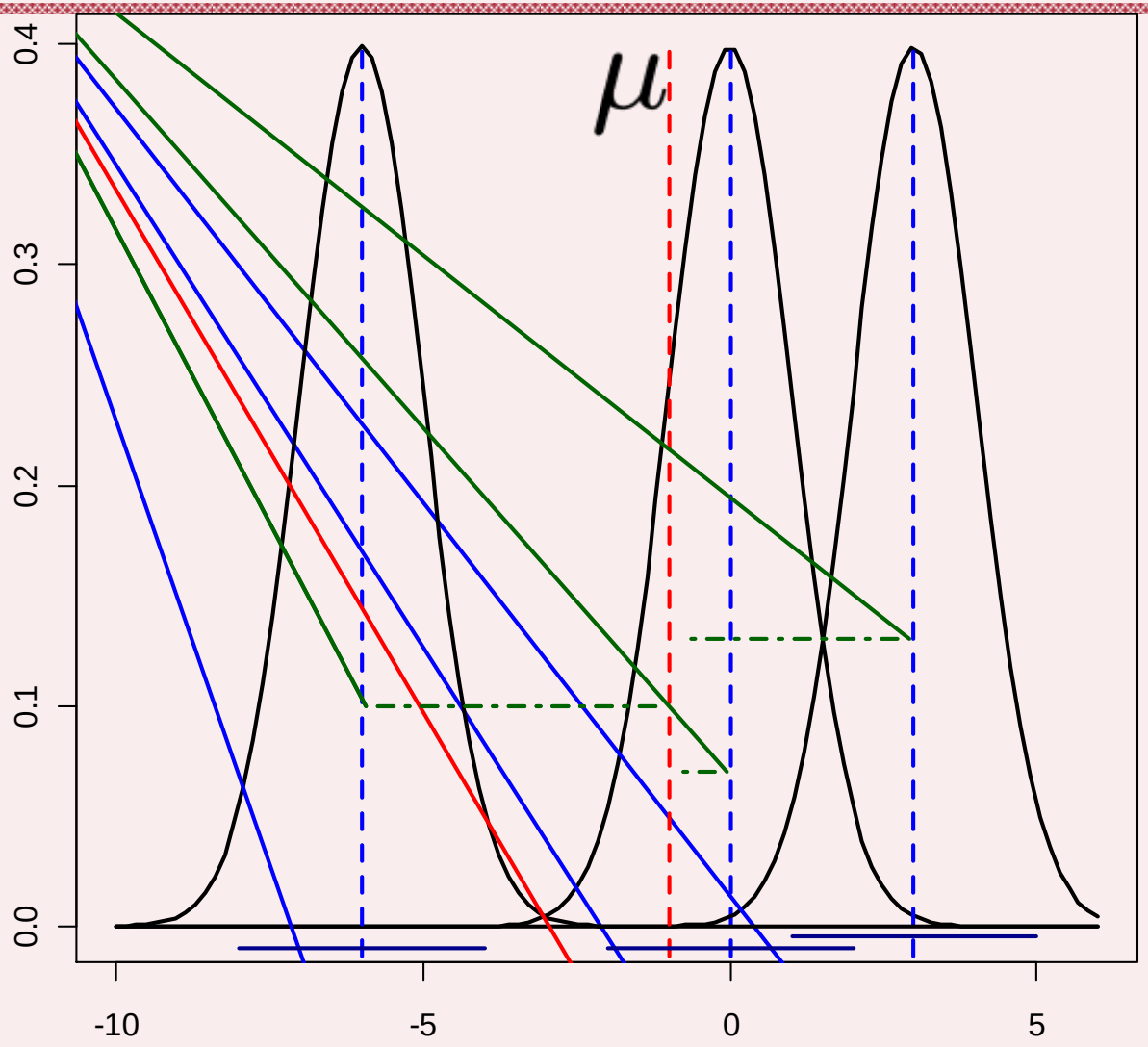
7　　=4（18人の平均点）
+2（A組の平均からの差）
+1（個人差）

ここなら2，つまり各群の平均からの差が十分に大きければ，群ごとに差があると言えそうである．

分散分析のイメージ



分散分析のイメージ



分散分析

- 分散分析とは，「因子（要因）」（クラス，国籍，性別）の「水準」（A組，B組，日本人，韓国人，男，女）が平均値に与える影響を分析する手法．
- A組とB組，　C組のテストの点を比較したい場合には，以下のようなモデルを考える．

$$y_{ij} = \mu + a_j + e_{ij}$$

- ここで，μ は一般平均と呼ばれ，水準を問わない平均的な値である．a_j は水準の効果である．つまり，ある群の平均は，一般平均とどの程度ずれているかを表す値である．e_{ij} は，y_{ij} に対応する誤差である．

分散分析

7　　=4（18人の平均点）
+2（A組の平均からの差）
+1（個人差）

$$y_{ij} = \mu + a_j + e_{ij}$$

$$7 = 4 + 2 + 1$$

分散分析

$$y_{ij} = \mu + a_j + e_{ij}$$

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

$$y_{ij} - \mu = a_j + e_{ij}$$

- 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）と誤差（グループ内のずれ）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）.
   - 証明したいこと（対立仮説）……3つのクラスの間には，得点に差がある.
   - 逆の仮説（帰無仮説）…… 3つのクラスの間には，得点に差がない.

$$y_{ij} = \mu + a_j + e_{ij}$$

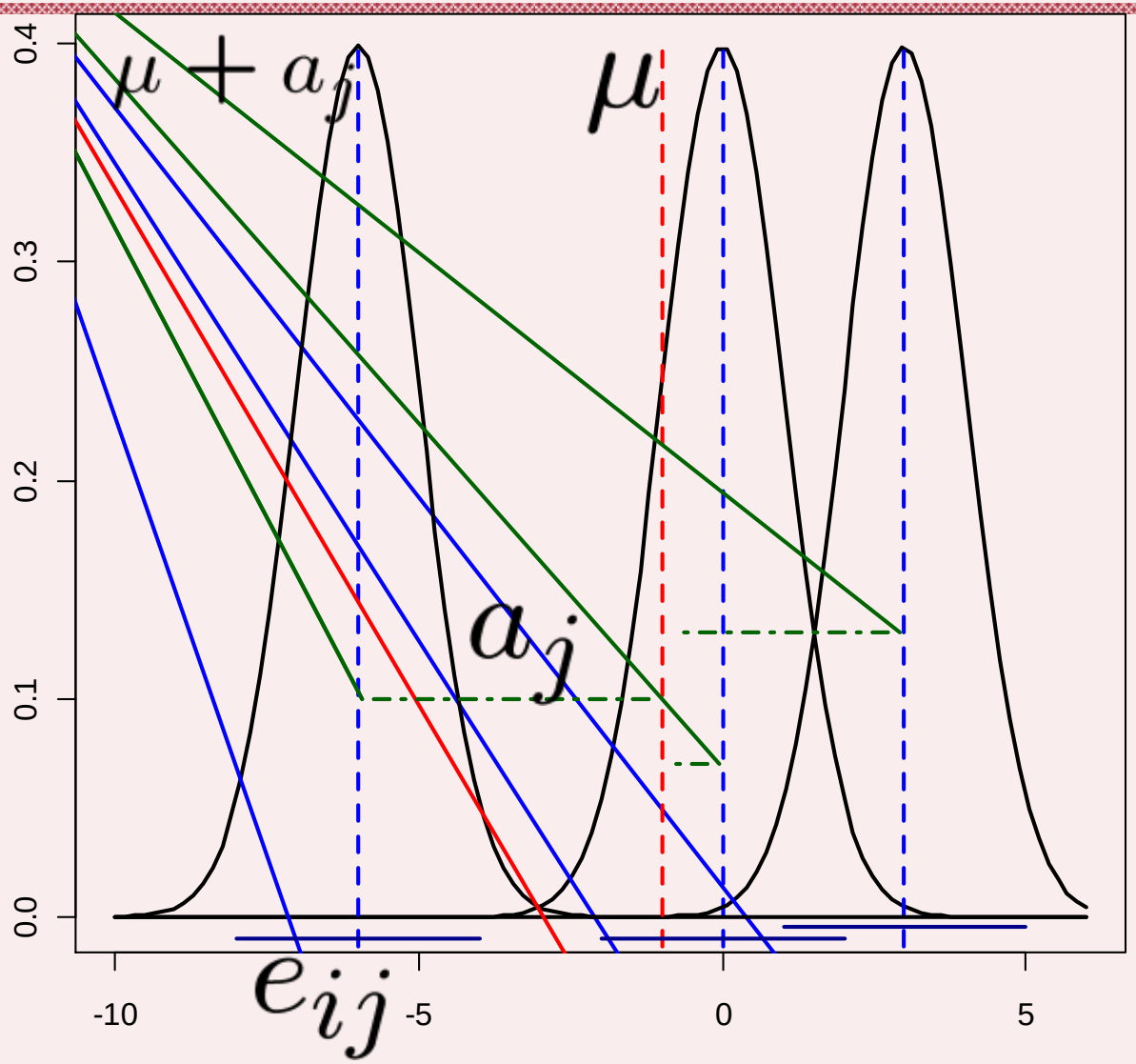
   - 上のモデルを前提とするならば，帰無仮説は以下の通りとなる.

$$a_A = a_B = a_C = 0$$

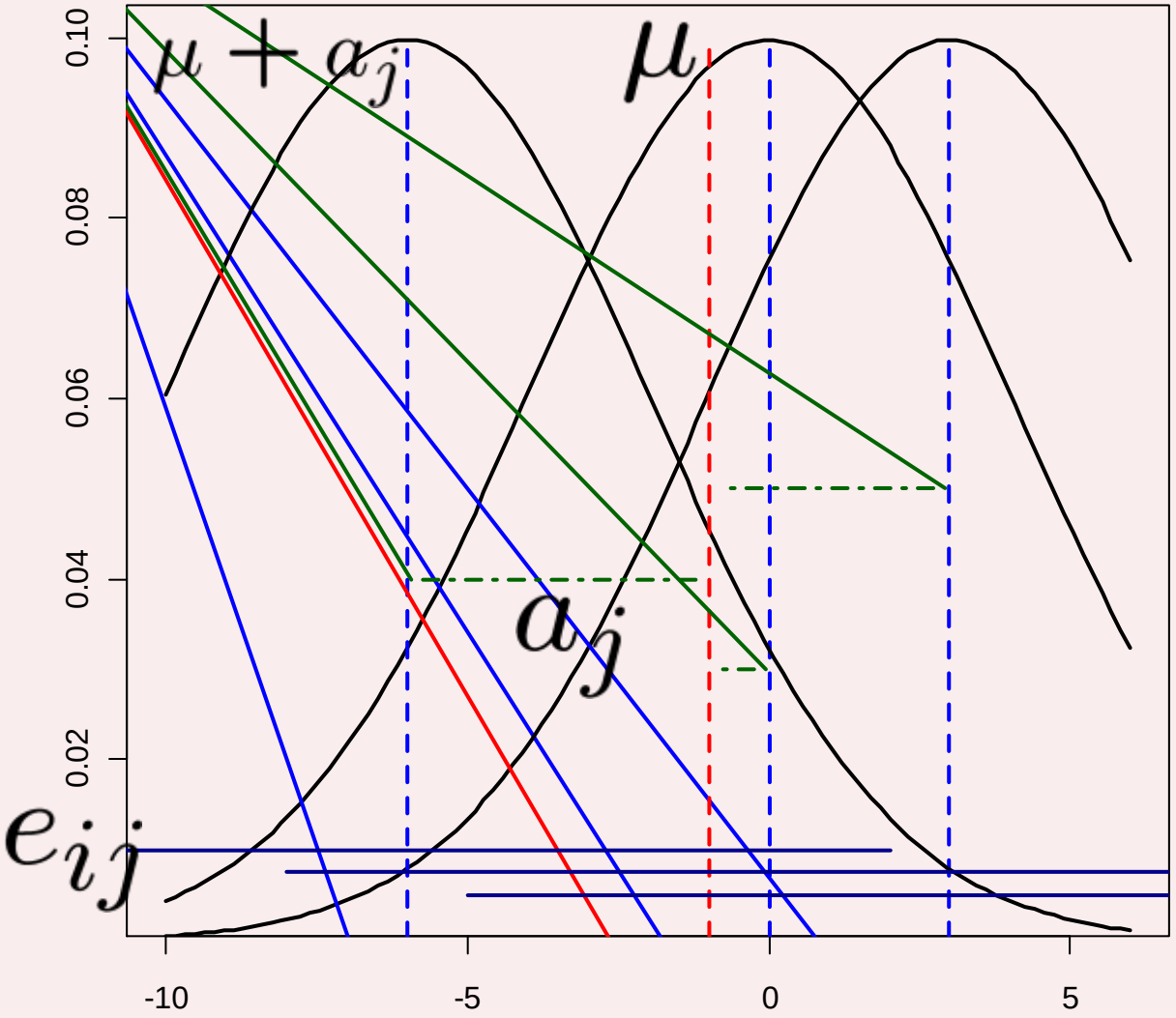
統計的仮説検定

2. 水準の効果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.
   - クラスの中での散らばりと比較して，水準の効果の散らばりが小さいならば，差があるとは言えない.
   - 水準の効果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．この意味で「分散」分析なのである.

分散分析のイメージ



分散分析のイメージ



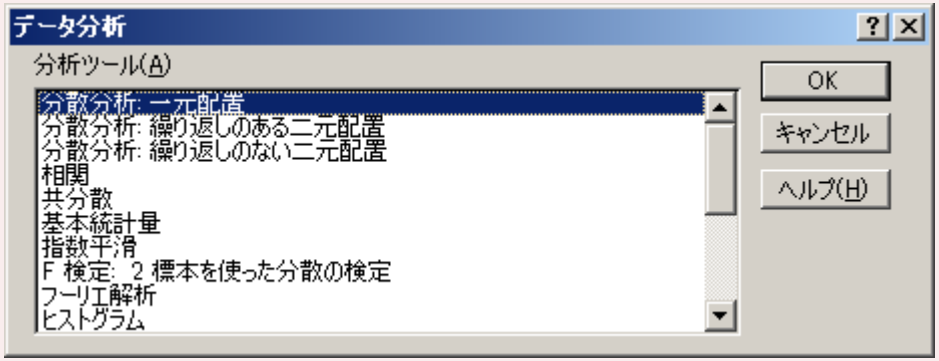
分散分析のイメージ

- 前掲2枚の図は，どちらも一般平均 μ と3つの水準の効果 a_j は等しい．しかし，e_{ij} （誤差）の分散が異なる．誤差の分散が小さい場合の図の方が，3つの集団に差があると言えそうである．
- 一番左のグループに属する人が他のグループの平均点を取る確率を考えると，1枚目では殆どないが，2枚目では決して低くない確率となる．

統計的仮説検定

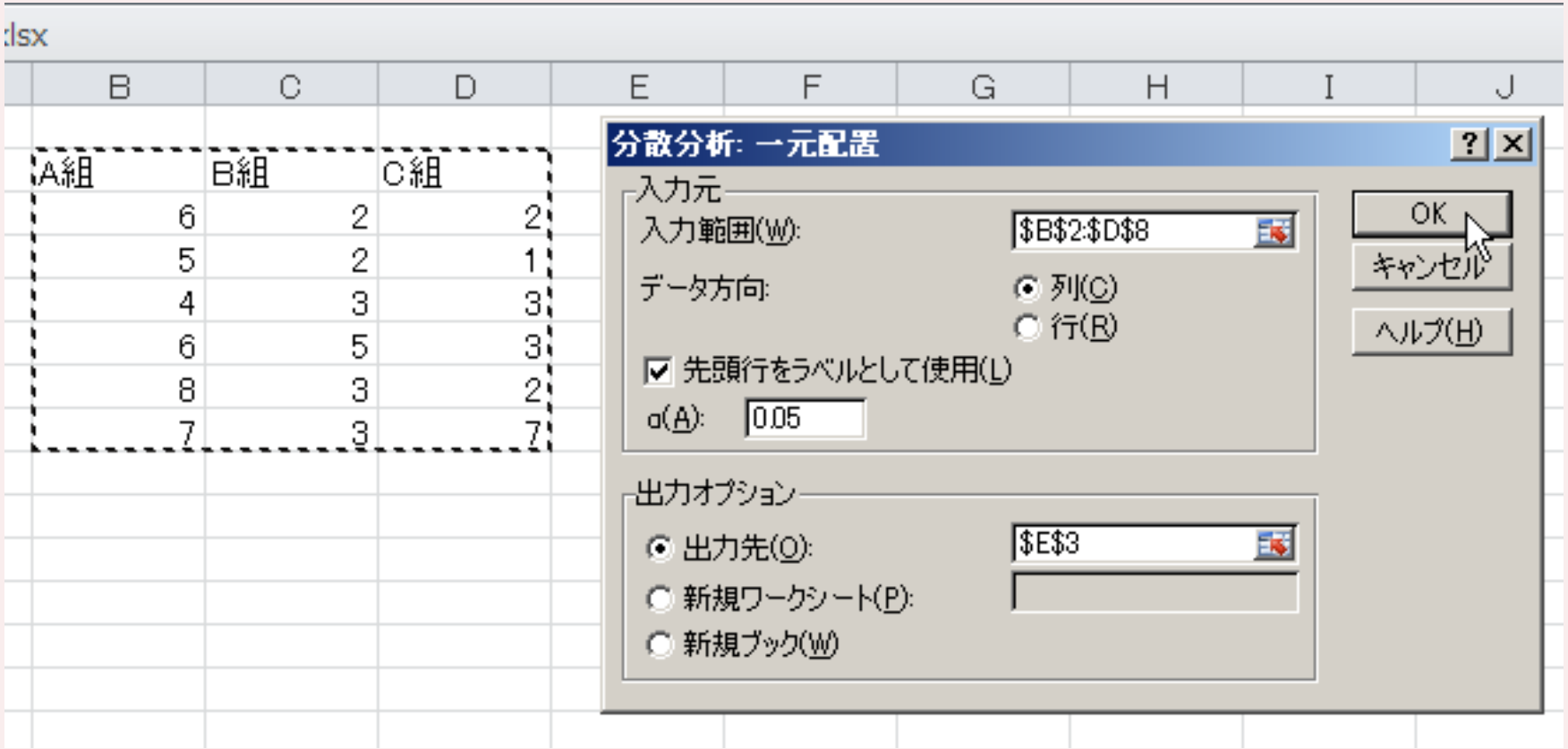
3. 帰無仮説の下で，得られた水準の効果の散らばりと誤差の散らばりの比がどれくらい起こりにくいことなのか検討する.
   - 帰無仮説が正しければ，散らばりの比は F 分布に従う.

分散分析



- Excelでは，3種類の分散分析が可能である．今回のようなケースは，「一元配置の分散分析」を利用する．

分散分析


lsx
B
C
D
E
F
G
H
I
J
A組
B組
C組
6 2 2
5 2 1
4 3 3
6 5 3
8 3 2
7 3 7
分散分析: 一元配置
入力元
入力範囲(W):
$B$2:$D$8
データ方向:
列(C)
行(R)
先頭行をラベルとして使用(L)
α(A):
0.05
出力オプション
出力先(O):
$E$3
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)

分散分析

分散分析: 一元配置

概要

グループ	標本数	合計	平均	分散
A組	6	36	6	2
B組	6	18	3	1.2
C組	6	18	3	4.4

a_jの分散（「平均」の分散の6倍になっている.）

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
グループ間	36	2	18	7.105263158	0.006747	3.68232
グループ内	38	15	2.533333			
合計	74	17				

e_{ij}の分散

18/2.533 ＝ 7.105

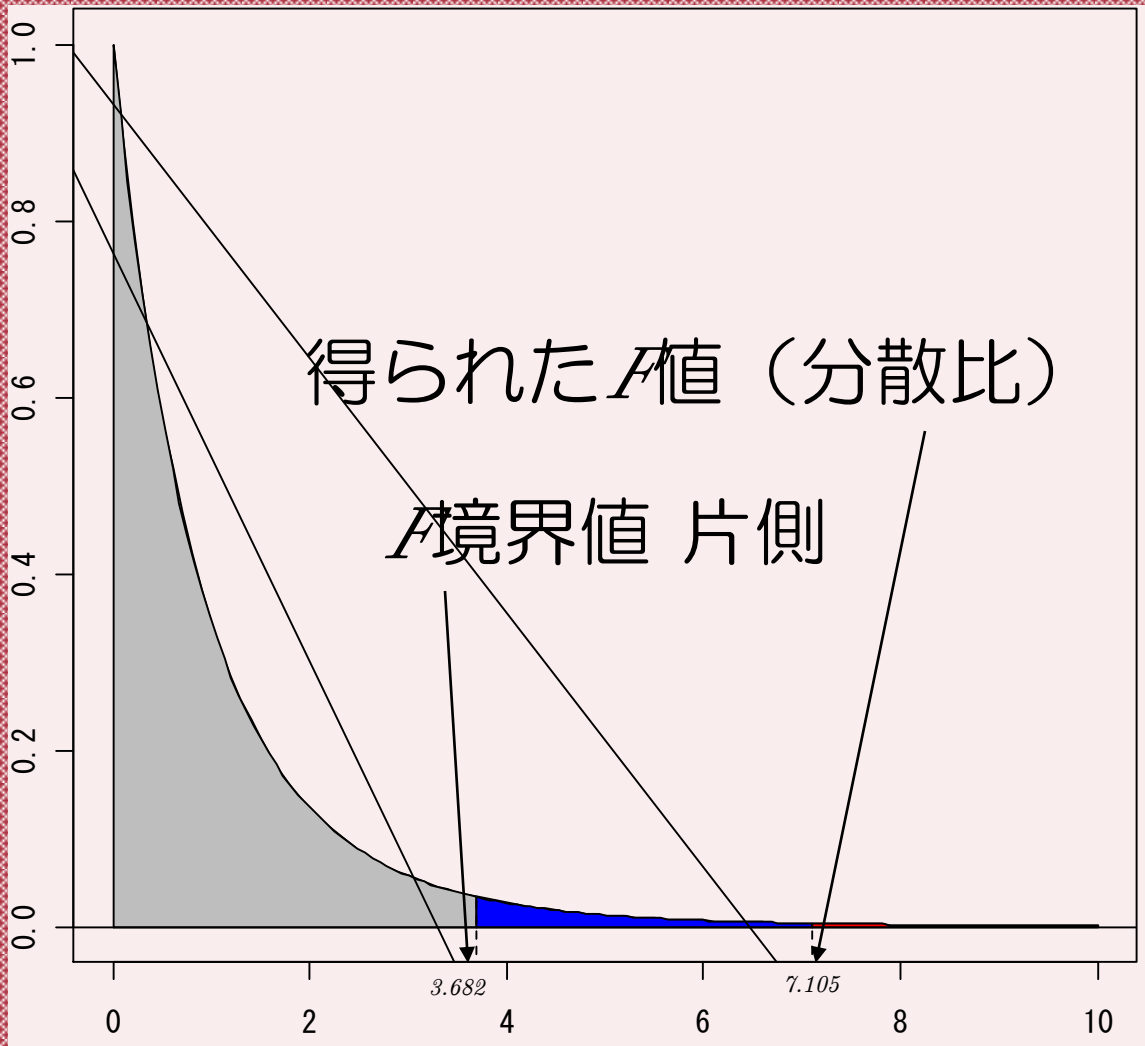
分散分析

観測された分散比	P-値	F 境界値
7.105263158	0.0067	3.68232

帰無仮説が正しい（3つのクラスに差がない）時，F値（観測された分散比）が7.105を超える確率は0.0067．これは5%以下であり，有意水準5%では差があると言える$(1\%$でも同じ$)$．

F値7.105はF境界値3.682を超えている．よって，F値が7.105を超える確率は5%以下であり，有意水準5%では差があると言える．

分散分析



自由度2,15のF分布

灰色の部分は3.682以下の範囲.
これより右側の確率（青＋赤）は5%.

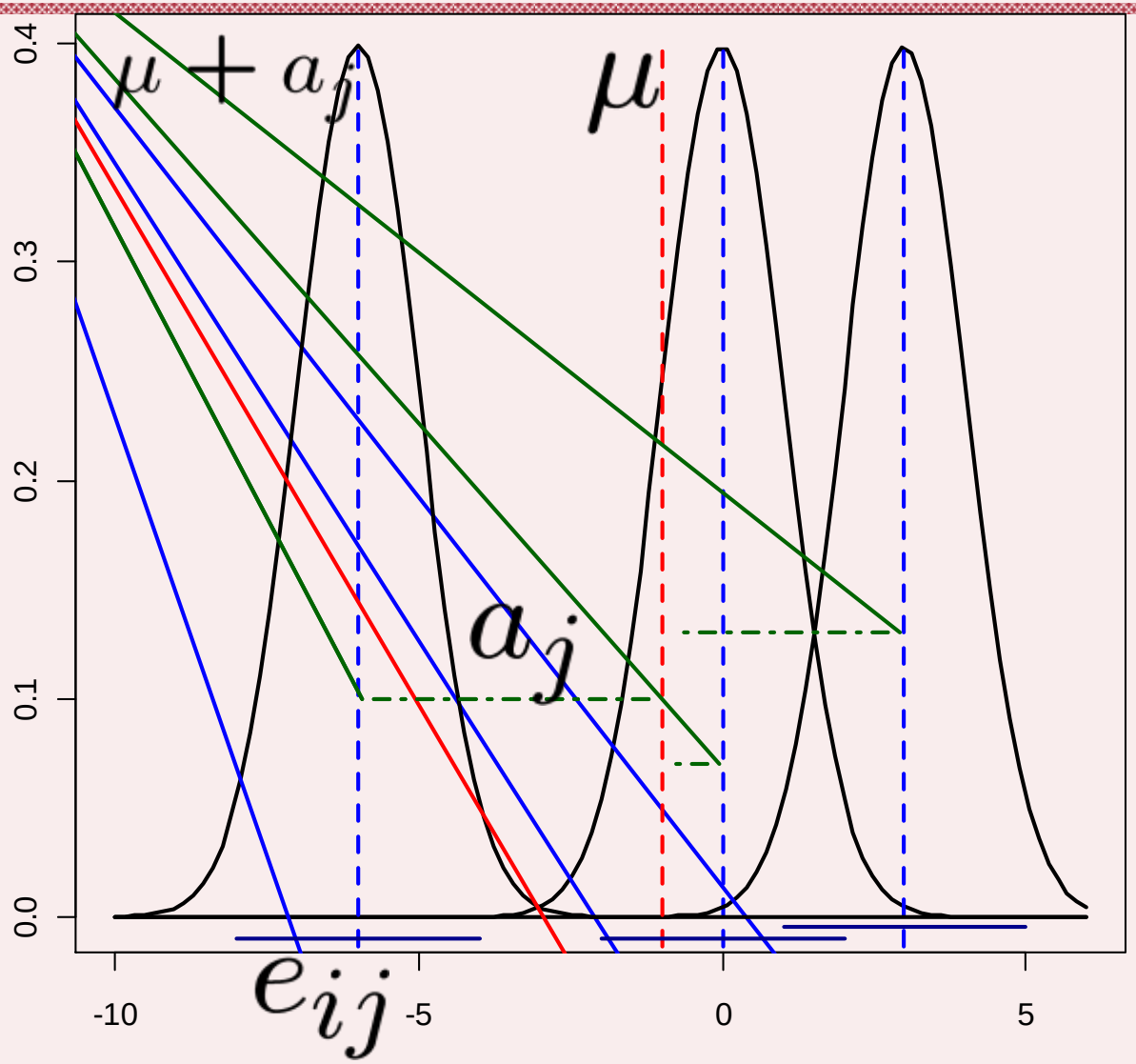
7.105の点より右側の確率は
（赤） 0.0067

統計的仮説検定

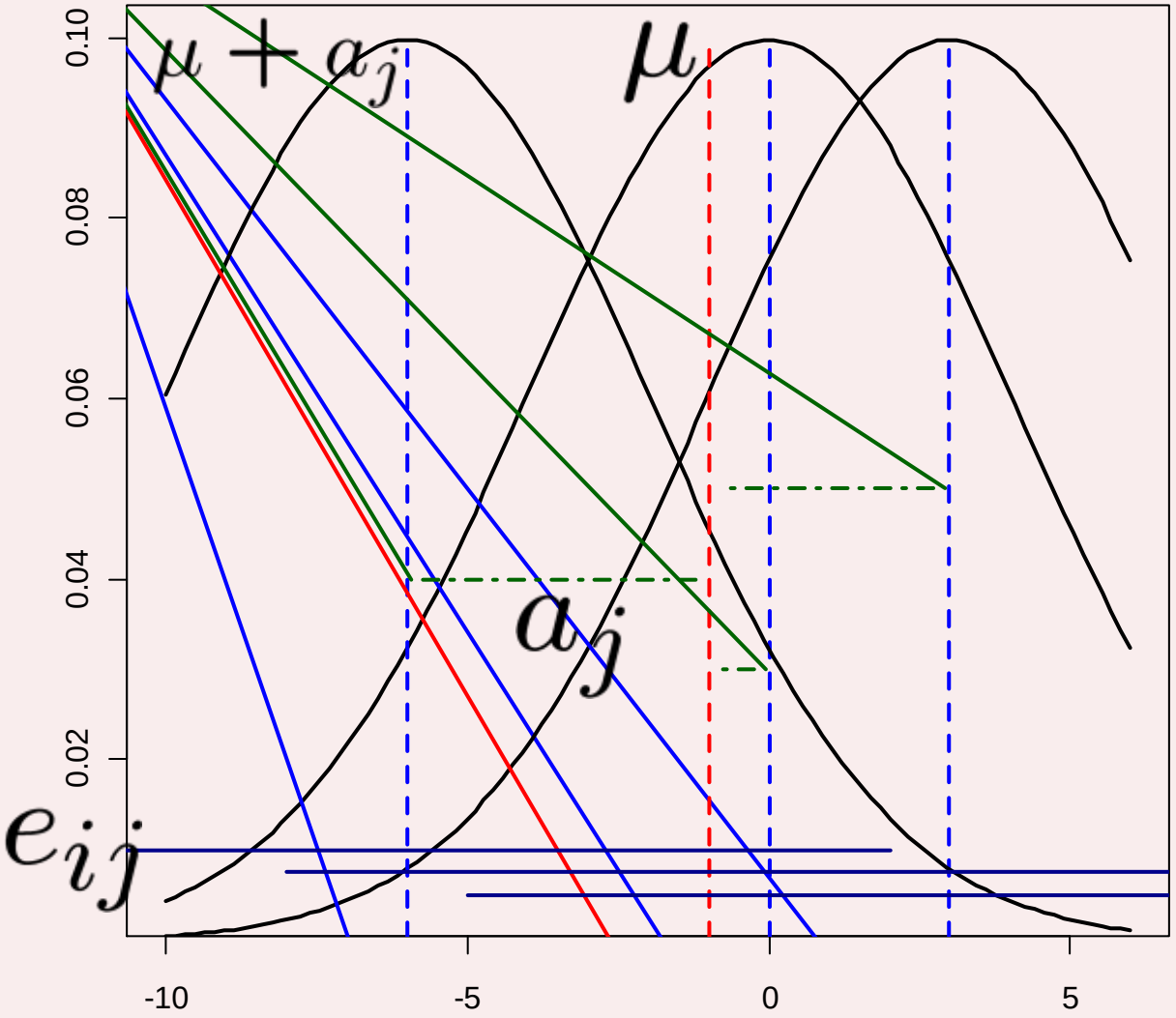
4. 確率的に低い事であれば，確率的に起こりにくいことがたまたま起こったと解釈するのではなく，帰無仮説が間違えていたのだと考える.

- 帰無仮説が正しく，散らばりの大きさの比が *F* 分布に従うとしたとき，*F*=2という値は十分ありえる. この場合，平均値に差があるとは言えない.
- 帰無仮説が正しく，散らばりの大きさの比がＦ分布に従うとしたとき，*F*=7.105という値はまずありえない. この場合，平均値に差がないという仮説（帰無仮説）を棄却し，少なくとも1組の平均値に差があるという仮説（対立仮説）を採択する.
- 「まずありえない」の基準として一般的なのは5%，1%，0.1%.

分散分析のイメージ



分散分析のイメージ



実習

- 先週のデータに加えてもう1カ国の身長のデータがある．身長が国によって異なるか，分散分析で検討せよ．

- 尚，$\alpha = 0.05$とする．