経営情報各論ＢⅡ
14．15．分散分析

目次

1. 分散分析とは

- 例えば，養豚場で，子豚の飼育のために飼料A, B, C, Dを用い，さらに豚舎の温度を高中低の３段階に調整しました．このとき，子豚の１月後の体重増加を調べた結果が以下の通りとする．
  この資料を見て，つぎのような両極の意見が考えられる．どちらの意見を採用するかに答えるのが分散分析(Analysis of Variance：ANOVA)です．
  - 「飼料Dを与え，低温にした方が，豚の育ちは良さそうだ」
  - 「全部偶然の結果であり，何も判断できない」
- 工場において改善の効果が得られたか，農場において肥料の効果が得られたか，開発した新薬の効果があるか，さらには他の薬との「飲み合わせ」効果があるかどうか，などを科学的に判断するのが分散分析．
- 因子と水準
  右下左の資料で説明する．飼料に相当するもので，データに影響を与える要因を因子という．
  この飼料の種類Ａ，Ｂ，Ｃ，Ｄのような項目で因子の種類を水準という．
- 一元配置，二元配置の分散分析
  - 1因子の影響を調べる分析を一元配置の分散分析という（右下左）
  - 2因子の影響の有無を調べるのが二元配置の分散分析で，同一条件のデータが１つしかない場合を「繰り返しのない二元配置」（右上），
    同一条件のデータが複数ある場合を「繰り返しのある二元配置」という（右下右）

		飼料			
		A	B	C	D
温度	高	3.11	5.85	8.59	8.81
	中	4.13	8.66	8.30	8.71
	低	7.37	6.36	9.30	12.11

飼料A	飼料B	飼料C	飼料D
8.40	6.79	5.79	7.30
4.44	5.74	8.65	9.20
7.71	5.02	10.38	7.71
7.23	6.71	6.25	8.14
3.57	8.57	8.22	7.14
3.53	8.49	7.15	11.35
3.77	7.80	9.32	6.77
7.31	5.29	5.62	7.97

		飼料			
		A	B	C	D
温度	高	11.03	8.75	9.45	6.18
		13.17	11.25	9.46	8.92
		11:53	6.31	7.97	10.73
	中	13.04	13.70	10.63	8.59
		11.45	11.67	13.66	9.75
		12.76	11.34	13.43	7.59
	低	10.39	12.98	8.02	9.53
		10.06	10.58	8.68	10.42
		13.02	9.98	12.74	8.00

2. 一元配置の分散分析(1/3)

- 一元配置の例（上）
  右表は，４種の飼料A,B,C,Dを各々子豚８頭に与え，１月後の体重増加（単位はkg）を調べたものである．他の要因はできるだけ同一にしている．（データに影響を与える要因を因子，その種類を水準という）
- 一元配置の分散分析
  上の例で，「飼料の効果があるのか」，それとも，「偶然の結果か」の判断を下す手段．
- データの偏差を分解
  飼料の効果は平均値からの「ずれ」＝データの偏差（＝データの値ー平均値）に現れる（下）

  これを次のように分解する．
  偏差＝（水準平均ー全体平均）＋（データの値ー水準平均）
  ＝水準間偏差　＝水準内偏差
  水準間偏差：同一飼料のグループの平均値から資料全体の平均値を引いたもので，飼料の違いの効果を表す量と考えられる
  水準内偏差：当該データから同一飼料のグループ平均を引いたもので，同一条件のもとで得られたデータのバラツキであり，偶然性を表す量，すなわち統計誤差を表す量と考えられる

各データの偏差 （各個体に対する飼料の効果）	＝	水準間偏差 （飼料の効果）	水準内偏差 （統計誤差）

右の偏差の表から水準間偏差と水準内偏差を計算する．ここで次の値を利用する．
$\bar{x} = 7.10,\ \bar{x}_A = 5.75,\ \bar{x}_B = 6.80,\ \bar{x}_C = 7.67,\ \bar{x}_D = 8.20$
結果を次ページに示す．

飼料A	飼料B	飼料C	飼料D
8.40	6.79	5.79	7.30
4.44	5.74	8.65	9.20
7.71	5.02	10.38	7.71
7.23	6.71	6.25	8.14
3.57	8.57	8.22	7.14
3.53	8.49	7.15	11.35
3.77	7.80	9.32	6.77
7.31	5.29	5.62	7.97

飼料A	飼料B	飼料C	飼料D
1.30	-0.31	-1.31	0.20
-2.66	-1.36	1.55	2.10
0.61	-2.08	3.28	0.61
0.13	-0.39	-0.85	1.04
-3.53	1.47	1.12	0.04
-3.57	1.39	0.05	4.25
-3.33	0.70	2.22	-0.33
0.21	-1.81	-1.48	0.87

2. 一元配置の分散分析(2/3)

Aの水準間偏差	Bの水準間偏差	Cの水準間偏差	Dの水準間偏差
-1.36	-0.30	0.57	1.09
-1.36	-0.30	0.57	1.09
-1.36	-0.30	0.57	1.09
-1.36	-0.30	0.57	1.09
-1.36	-0.30	0.57	1.09
-1.36	-0.30	0.57	1.09
-1.36	-0.30	0.57	1.09
-1.36	-0.30	0.57	1.09

水準間偏差(=水準平均－全体平均）：飼料の効果を表す

Aの水準内偏差	Bの水準内偏差	Cの水準内偏差	Dの水準内偏差
2.66	-0.01	-1.88	-0.90
-1.31	-1.06	0.98	1.00
1.97	-1.78	2.71	-0.49
1.49	-0.09	-1.42	-0.06
-2.18	1.77	0.55	-1.06
-2.22	1.69	-0.52	3.15
-1.98	1.00	1.65	-1.43
1.57	-1.51	-2.05	-0.23

水準内偏差(=データ値－水準平均）：統計誤差を表す

- 水準間偏差と水準内偏差を数値化
  水準間偏差（飼料の効果）が大きく，水準内偏差（統計誤差）が小さければ，飼料の違いの効果が認められる．逆だと，偶然性が支配していると考えられる．そのため，これら２つの大小を「変動」を求めることで調べる．変動はデータの散らばり具合を表現することを学びました．水準間変動をQ_1，水準内変動をQ_2とすると，
  $Q_1 = (-1.36)^2 \times 8 + (-0.30)^2 \times 8 + 0.57^2 \times 8 + 01.09^2 \times 8 = 27.66$
  $Q_2 = \{2.66^2 + (-1.31)^2 + \cdots + 1.57^2\} + \{(-0.01)^2 + (-01.06)^2 + \cdots + (-1.51)^2\} + \cdots = 80.93$
  変動Q_1が変動Q_2に比べて大きければ，飼料の違いによる効果が認められ，逆なら，結果は偶然となる．

  飼料の効き具合
  偶然性を表現
- 不偏分散を求める
  F検定を利用したいので，求めた変動を不偏分散に変換する．不偏分散Vは変動Qを自由度f割って求まる．$V = Q/f$
  水準間変動の自由度f_1は，水準間偏差の平均値は０という制約があるので，水準の数-1=4-1=3となる．
  水準内変動の自由度f_2は，水準ごとの偏差の平均値は0という制約があるので，「各水準のデータ数-1」×水準数=(8-1)×4=28
  よって，水準間偏差の不偏分散は$V_1 = Q_1/f_1 = 27.66/3 = 9.22$，水準内偏差の不偏分散は$V_2 = Q_2/f_2 = 80.93/28 = 2.89$

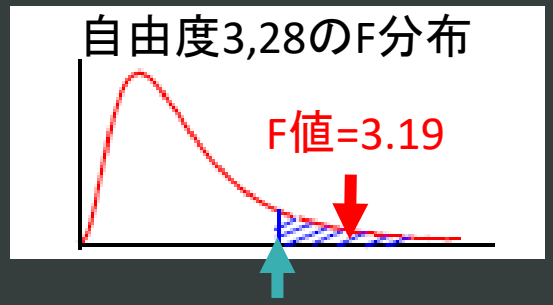
2. 一元配置の分散分析(3/3)

- 分散分析を支えるのはF分布
  次の定理を利用して，飼料の効果を表す不偏分散と偶然性を表す不偏分散との大小を検定する．
  「正規分布に従う同一の母集団から抽出された標本に対して，不偏分散V_1, V_2との比は自由度f_1, f_2のF分布に従う」
  この定理が分散分析のバックボーンである．

- F検定の実行
  F分布による検定（F検定）を実行する．すなわち，次の帰無仮説を有意水準5%で検定する．
  H_0：水準間の差異はない
  つまり，飼料の種類による豚の体重増加の差はない，という帰無仮説を設定する．（対立仮説は，「水準間の差異（飼料の効果）がある」です）
  仮説をF分布で検定するには，上記の定理から，不偏分散の比（F値という）が必要です．

$$F = \frac{V_1}{V_2} = \frac{9.22}{2.89} = 3.19$$

  V_1は飼料の違いの効果を，V_2は偶然性を表しているので，
  このF値が大きければ，飼料の違いの効果が確かめられることになります．
  定理によれば，このF値が自由度3,28のF分布に従う．
  自由度3,28のF分布の上側5%点を求めると，
  上側5%点=2.95
  したがって，上で求めたF値3.19は有意水準5%の棄却域に入る．
  仮説は棄却されたので，
  「飼料の違いの効果があった」ことが有意水準5%で認められた．



F分布表(その1)　上段(黒字):0.05%　下段(赤字):0.01%

分母の分散の自由度 f2	分子の分散の自由度 1	2	3	4	5	6	7	8	9	10	11	12	14
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17	2.13	2.08
	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.99	2.93	2.82
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.06
	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.96	2.90	2.79
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05
	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.93	2.87	2.77
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.04

3. 一元配置の分散分析表(1/3)

■ 分散分析表
以下のフォーム（分散分析表）を埋めていくことで分散分析ができる

変動要因	変動	自由度	不偏分散	分散比	F境界値
水準間変動					
水準内変動					

右の一般的な資料をもとに，分散分析の手順を示す．

W	X	…	Z
w_1	x_1	…	z_1
w_2	x_2	…	z_2
w_3	x_3	…	z_3
…	…	…	…
w_n	x_n	…	z_n

W, X, …, Z: 水準名， n: 各水準のデータ数

① 変動の計算
まず，資料全体の平均値m_Tを求め，次に，各水準ごとの平均値$m_W, m_X, \ldots, m_Z$を求める．そして，水準間変動Q_1，水準内変動Q_2を求める．

$$Q_1 = n\{(m_W - m_T)^2 + (m_X - m_T)^2 + \cdots + (m_Z - m_T)^2\}$$

$$Q_2 = \{(w_1 - m_W)^2 + (w_2 - m_W)^2 + \cdots + (w_n - m_W)^2\} + \cdots + \{(z_1 - m_Z)^2 + (z_2 - m_Z)^2 + \cdots + (z_n - m_Z)^2\}$$

これを2.の例で計算したものが

変動要因	変動	自由度	不偏分散	分散比	F境界値
水準間変動	27.66				
水準内変動	80.93				

3. 一元配置の分散分析表(2/3)

② 自由度の計算
水準の数をlとします．各水準のデータ数はnですので，2.の議論から
水準間の変動の自由度$= l - 1$
水準内の変動の自由度$= l(n-1)$

これを2.の例で計算したものが

変動要因	変動	自由度	不偏分散	分散比	F境界値
水準間変動	27.66	3			
水準内変動	80.93	28			

③ 不偏分散の計算
変動を自由度で割ったものが不偏分散となる．

$V_1 = \frac{Q_1}{l-1}$, $V_2 = \frac{Q_2}{n(l-1)}$

これを2.の例で計算したものが

変動要因	変動	自由度	不偏分散	分散比	F境界値
水準間変動	27.66	3	9.22		
水準内変動	80.93	28	2.89		

3. 一元配置の分散分析表(3/3)

④ F値の計算

不偏分散V_1，V_2の比が自由度$l-1, n(l-1)$のF分布に従うので，次のF値を求める　$F = \frac{V_1}{V_2}$
これを2.の例で計算したものが

変動要因	変動	自由度	不偏分散	分散比	F境界値
水準間変動	27.66	3	9.22	3.19	
水準内変動	80.93	28	2.89		

⑤ F分布のパーセント点を求める
帰無仮説で設定した有意水準に対応する自由度$l-1, n(l-1)$のF分布のパーセント点を求め，表を埋める.
これを2.の例で計算したものが

変動要因	変動	自由度	不偏分散	分散比	F境界値
水準間変動	27.66	3	9.22	3.19	2.95
水準内変動	80.93	28	2.89		

これで分散分析表が完成.

⑥ F値とパーセント点の大小を比較
F値が有意水準に対するパーセント点より大きければ，帰無仮説は棄却されます．対立仮説が採択される.
この例では，分散比（F値）はF境界値（有意水準に対するパーセント点のこと．ここでは5%点）より大きいので，棄却域に入っている．よって，帰無仮説は棄却されることになり，対立仮説が採択される.

【サンプル１】一元配置の分散分析

データサイエンス基礎_分散分析_(1)一元配置の分散分析.ipynb

4. 繰り返しのない二元配置の分散分析(1/3)

- ２因子の効果を判定
ここでも豚の飼育を例にする． ２因子として，与える「飼料」と豚舎の「温度」を取り上げます．それ以外の因子はできるだけ均一化した条件で，１月後の体重増加（単位はkg）を調べた結果が右の資料です．資料には，２因子各組に対して１個のデータしか存在しません．同じ条件下では，１回の実験結果しか得られていない場合です．このような資料に対する分散分析を繰り返しのない二元配置の分散分析といいます．

		飼料			
		A	B	C	D
温度	高	3.11	5.85	8.59	8.81
	中	4.13	8.66	8.30	8.71
	低	7.37	6.36	9.30	12.11

要因（資料と温度）を因子，要因の項目（飼料A,B,C,D，温度の高低）を水準と呼ぶ．

- 因子の効果は偏差に現れる
基本は一元配置の分散分析と同様です．この資料を縦と横から見れば，それぞれ一元配置の分散分析の資料と同様になります．
そこで因子の効果が表れる偏差を次のように分解してみます．偏差＝飼料の効果＋温度の効果＋統計誤差

各データの偏差 （各個体に対する飼料と温度の効果）	＝	飼料因子の水準間偏差	温度因子の水準間偏差	水準内偏差（統計誤差）
		飼料の効果	温度の効果	因子で説明できない部分

- 因子の効果を調べる
温度の効果＝温度の水準平均－全体平均（右上表）
温度の違いの効果の変動$Q_{11} = 4\{(-1.06)^2 + (-0.20)^2 + (1.26)^2\} = 11.00$
この値が相対的に大きければ，「温度の違い」の効果が大きいことになる．

飼料の効果＝資料の水準平均－全体平均（右下表）
飼料の違いの効果の変動$Q_{12} = 3\{(-2.78)^2 + (-0.53)^2 + (1.08)^2 + (2.23)^2\} = 42.39$
この値が相対的に大きければ，「飼料」の効果が大きいことになる．

		飼料			
		A	B	C	D
温度	高	-1.06	-1.06	-1.06	-1.06
	中	-0.20	-0.20	-0.20	-0.20
	低	1.26	1.26	1.26	1.26

		飼料			
		A	B	C	D
温度	高	-2.78	-0.53	1.08	2.23
	中	-2.78	-0.53	1.08	2.23
	低	-2.78	-0.53	1.08	2.23

4. 繰り返しのない二元配置の分散分析(2/3)

- ２因子の効果を引いたものが統計誤差
  統計誤差＝データの偏差ー飼料の効果ー温度の効果（計算結果は左）
  「統計誤差」の変動$Q_2 = (-0.70)^2+(-0.21)^2+\cdots+(-0.69)^2+(0.97)^2 = 10.96$
  「温度の違いの効果」Q_{11}と「飼料の違いの効果」Q_{12}が，統計誤差の変動Q_2に比べて大きいときには，各々の効果が認められることになる.

		飼料			
		A	B	C	D
温度	高	-0.70	-0.21	0.92	-0.01
	中	-0.54	1.74	-0.23	-0.97
	低	1.24	-1.52	-0.69	0.97

- 不偏分散を算出
  変動Q_{11}，Q_{12}，Q_2を自由度で割った数の比，すなわち不偏分散の比はF分布に従う.
  よって，まずは各効果の不偏分散を求める．3.と同様に変動Q_{11}，Q_{12}の自由度は水準数から１引いた値となる.
  温度の効果の自由度=3-1=2
  飼料の効果の自由度=4-1=3
  また，統計誤差の効果Q_2の自由度も3.と同様に
  誤差の自由度=(4-1)(3-1)=6
  でしたがって，温度の効果，資料の効果，統計誤差の効果の不偏分散V_{11}，V_{12}，V_2は以下になる

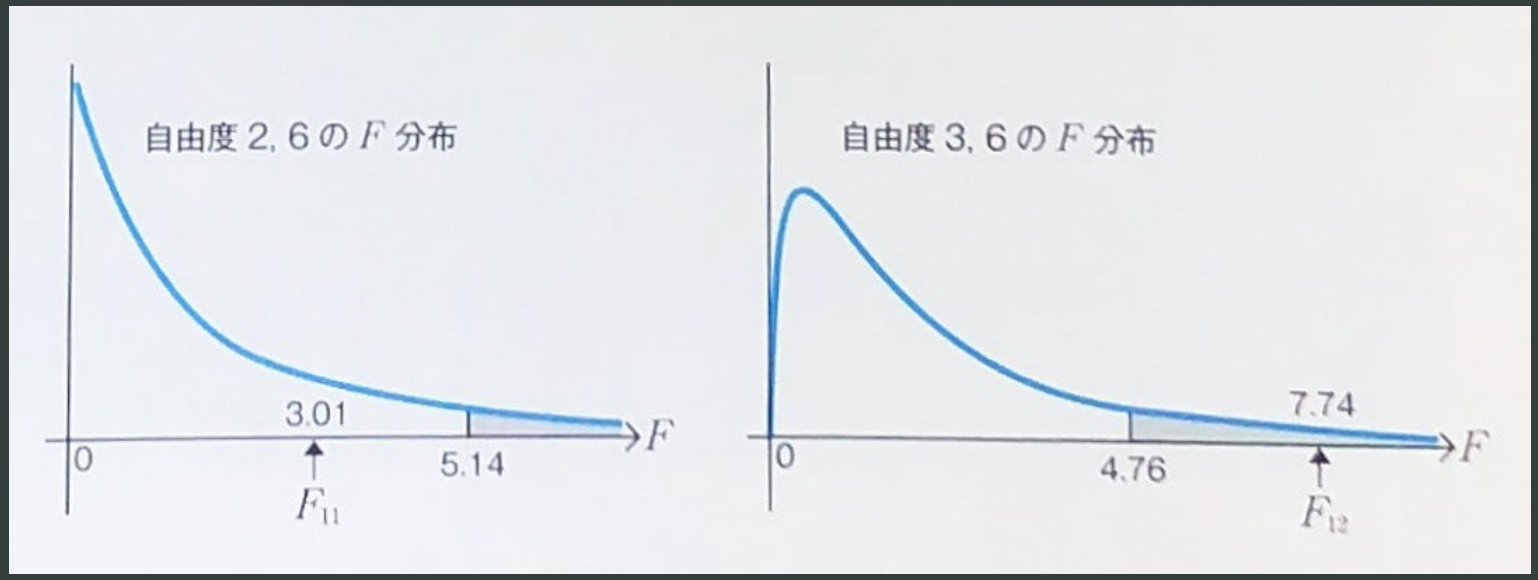
$$V_{11} = \frac{Q_{11}}{2} = \frac{11.0}{2} = 5.50, \quad V_{12} = \frac{Q_{12}}{3} = \frac{42.39}{3} = 14.13, \quad V_2 = \frac{Q_2}{6} = \frac{10.96}{6} = 1.83$$

- 検定開始
  まず，仮説（帰無仮説）を設定
  H_{10}：温度の違いの効果は認められない
  H_{20}：飼料の違いの効果は認められない
  これを5%の有意水準で検定する.
  不偏分散の比はF分布に従う．ですので，次のF値，F_{11}，F_{12}は各々自由度2, 6と自由度3, 6のF分布に従います.

$$F_{11} = \frac{V_{11}}{V_2} = \frac{5.50}{1.83} = 3.01, \quad F_{12} = \frac{V_{12}}{V_2} = \frac{14.13}{1.83} = 7.74$$

4. 繰り返しのない二元配置の分散分析(3/3)

- 検定の続き
  次の図は，自由度2, 6と自由度3, 6のF分布の分布曲線です．
  そこに，これらのF値を記入します．また，有意水準5%の棄却域を網掛けで表示する．



グラフが示すように，自由度2, 6と自由度3, 6のF分布の5%点は
自由度2,6のF分布の5%点=5.14
自由度3,6のF分布の5%点=4.76
よって，「温度の違いの効果」を示すF値は棄却域に入っていないため，帰無仮説H_{10}は棄却できない．よって，統計的なバラツキに比べて，豚舎の温度の違いの効果がある，とは敢えて言えない．
しかし，「飼料の効果」を示すF値は棄却域に入っているので，帰無仮説H_{20}は棄却されます．飼料に関しては，統計的誤差のバラツキに比べて，その効果は十分大きいことが分かった．飼料の違いの効果は認められた．

5. 繰り返しのない二元配置の分散分析表(1/3)

■ 分散分析表
以下のフォーム（分散分析表）を埋めていくことで分散分析ができる

変動要因	変動	自由度	不偏分散	分散比	F境界値
行					
列					
水準内変動					

		因子X			
		X_1	X_2	…	X_l
因子Y	Y_1	a_{11}	a_{21}	…	a_{l1}
	Y_2	a_{12}	a_{22}	…	a_{l2}
	…	…	…	…	…
	…	…	…	…	…
	Y_k	a_{1k}	a_{2k}	…	a_{lk}

X, Y：因子名，$X_1, X_2, \ldots, X_l, Y_1, Y_2, \ldots, Y_k$：水準名，$l, k$: 各水準のデータ数

右の一般的な資料をもとに，分散分析の手順を示す．

① 変動の計算
因子Xの水準間変動Q_{11}，因子Yの水準間変動Q_{12}，純粋な統計誤差の変動Q_2は次のようになる．ここで，全体の平均値をm_T，因子Xの各水準ごとの平均値を$m_{X1}, m_{X2}, \ldots, m_{Xl}$，因子$Y$の各水準ごとの平均値を$m_{Y1}, m_{Y2}, \ldots, m_{Yk}$とする．

$$Q_{11} = k\{(m_{X1} - m_T)^2 + (m_{X2} - m_T)^2 + \cdots + (m_{Xl} - m_T)^2\}$$
$$Q_{12} = l\{(m_{Y1} - m_T)^2 + (m_{Y2} - m_T)^2 + \cdots + (m_{Yk} - m_T)^2\}$$
$$Q_2 = \{(a_{11} - m_T) - (m_{X1} - m_T) - (m_{Y1} - m_T)\}^2 + \cdots + \{(a_{lk} - m_T) - (m_{Xl} - m_T) - (m_{Yk} - m_T)\}^2$$

これを4.の例で計算したものが

変動要因	変動	自由度	不偏分散	分散比	F境界値
行	11.00				
列	42.39				
水準内変動	10.96				

5. 繰り返しのない二元配置の分散分析表(2/3)

② 自由度の計算
因子X, Yの水準の数をl, kとすると，4.の議論から
因子Xの水準間の変動の自由度$= l - 1$
因子Yの水準間の変動の自由度$= k - 1$
水準内変動（誤差変動）の自由度$= (k-1)(l-1)$
これを4.の例で計算したものが

変動要因	変動	自由度	不偏分散	分散比	F境界値
行	11.00	2			
列	42.39	3			
水準内変動	10.96	6			

③ 不偏分散の計算
変動を自由度で割ったものが不偏分散となる.
因子X, Yの水準間変動と水準内変動（誤差変動）の不偏分散を各々V_{11}，V_{12}，V_2とすると，

$$V_{11} = \frac{Q_{11}}{k-1},\ V_{12} = \frac{Q_{12}}{l-1},\ V_2 = \frac{Q_2}{(k-1)(l-1)}$$

これを4.の例で計算したものが

変動要因	変動	自由度	不偏分散	分散比	F境界値
行	11.00	2	5.50		
列	42.39	3	14.13		
水準内変動	10.96	6	1.83		

5. 繰り返しのない二元配置の分散分析表(3/3)

④ F値の計算

不偏分散V_{11}，V_{12}，V_2の比がF分布に従うので，次のF値を求める $F_{11} = \frac{V_{11}}{V_2}$，$F_{12} = \frac{V_{12}}{V_2}$
これを4.の例で計算したものが

変動要因	変動	自由度	不偏分散	分散比	F境界値
行(因子Y)	11.00	2	5.50	3.01	
列(因子X)	42.39	3	14.13	7.74	
水準内変動	10.96	6	1.83		

⑤ F分布のパーセント点を求める
帰無仮説で設定した有意水準に対応するF分布のパーセント点をF分布表やExcelなどの統計ソフトを利用して求め，表を埋める．ここでは5%点を求めます．有意水準に対するパーセント点のことをF境界値という．
これを4.の例で求めたものが

変動要因	変動	自由度	不偏分散	分散比	F境界値
行(因子Y)	11.00	2	5.50	3.01	5.44
列(因子X)	42.39	3	14.13	7.74	4.76
水準内変動	10.96	6	1.83		

これで分散分析表が完成．

⑥ F値とパーセント点の大小を比較
F値が有意水準に対するパーセント点より大きければ，帰無仮説は棄却されます．対立仮説が採択される．

【サンプル２】繰り返しのない二元配置の分散分析

データサイエンス基礎_分散分析_(2)繰り返しのない二元配置の分散分析.ipynb

6. 繰り返しのある二元配置の分散分析(1/4)

- 例で説明
  ここでも豚の飼育を例にする． ２因子として，与える「飼料」と豚舎の「温度」を取り上げます．それ以外の因子はできるだけ均一化した条件で，１月後の体重増加（単位はkg）を調べた結果が右の資料です．資料には，２因子各組に対して複数のデータがあります．同じ条件下で繰り返し実験した結果をまとめた資料です．このような資料に対する分散分析を繰り返しのある二次元配置の分散分析といいます．
- 「繰り返しのある資料」は交互作用が調べられる
  繰り返しのある二元配置の分散分析も，基本は，繰り返しのない二元配置の分散分析と同じ．だが，繰り返しのある場合は，２つの因子の交互作用の情報を得ることができる．
- 偏差の分解
  基本は繰り返しのない二元配置の分散分析と同様．

  偏差＝飼料の効果＋温度の効果＋統計誤差 ・・・・・・（１）

		飼料			
		A	B	C	D
温度	高	11.03	8.75	9.45	6.18
		13.17	11.25	9.46	8.92
		11:53	6.31	7.97	10.73
	中	13.04	13.70	10.63	8.59
		11.45	11.67	13.66	9.75
		12.76	11.34	13.43	7.59
	低	10.39	12.98	8.02	9.53
		10.06	10.58	8.68	10.42
		13.02	9.98	12.74	8.00

同一条件のデータが３つ得られている．
「繰り返し３回」の二元配置の分散分析となる

各データの偏差 (各データ－全体平均) （各個体に対する飼料と温度の効果）	＝	飼料因子の水準間偏差 (飼料の水準平均－全体平均)	温度因子の水準間偏差 (温度の水準平均－全体平均)	水準内偏差 （統計誤差）
		飼料の効果	温度の効果	因子で説明できない部分

繰り返しのある二元配置では，純粋な統計誤差が抽出できる

純粋な統計誤差＝データ値－（同一条件データの平均）

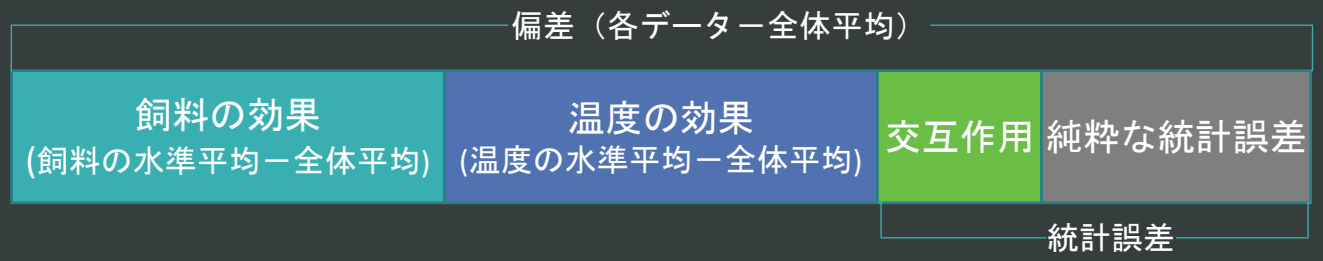
同一条件を持つデータは，決定論的には同一値を取るはずだが，偶然のいたずらで，同一因子・同一水準のデータにもばらつきがある．そこで，

純粋な統計誤差＝データ値－（同一因子・同一水準の平均値） ・・・・・・・（２）

とすることで純粋な統計誤差が抽出できる．

6. 繰り返しのある二元配置の分散分析(2/4)

- 交互作用の算出
  交互作用の効果は，「統計誤差」から「純粋な統計誤差」を引いたものと考える．つまり，各々の因子では説明しきれない「統計誤差」から「純粋な統計誤差」を差し引いたものが交互作用の効果と考えるのです．
  交互作用の効果＝式(1)の統計誤差－式(2)の純粋な統計誤差　・・・・・（３）
  交互作用の効果は，(1)と(3)から次のように偏差に位置づけられる．
  偏差＝温度の効果＋飼料の効果＋交互作用＋純粋な統計誤差　・・・・・（４）



		飼料			
		A	B	C	D
温度	高	-0.90	-0.90	-0.90	-0.90
		-0.90	-0.90	-0.90	-0.90
		-0.90	-0.90	-0.90	-0.90
	中	1.00	1.00	1.00	1.00
		1.00	1.00	1.00	1.00
		1.00	1.00	1.00	1.00
	低	-0.10	-0.10	-0.10	-0.10
		-0.10	-0.10	-0.10	-0.10
		-0.10	-0.10	-0.10	-0.10

温度の効果（＝温度の水準平均－全体平均）

- ２因子の効果を数値化
  以上で全体の枠組みができあがったので，具体的な計算に進む．まず，各因子の効果の部分を求める（右の表）．
  そして，各因子の効果を数値化する．温度の変動をQ_{11}，飼料の変動をQ_{12}とすると，
  $Q_{11} = 12\{(-0.90)^2 + (1.00)^2 + (-0.10)^2\} = 21.95$
  この変動値が大きいと，温度の効果が大きいことになる．
  $Q_{12} = 9\{(1.36)^2 + (0.26)^2 + (-0.02)^2 + (-1.61)^2\} = 40.62$
  この変動値が大きいと，飼料の効果が大きいことになる．

		飼料			
		A	B	C	D
温度	高	1.36	0.26	-0.02	-1.61
		1.36	0.26	-0.02	-1.61
		1.36	0.26	-0.02	-1.61
	中	1.36	0.26	-0.02	-1.61
		1.36	0.26	-0.02	-1.61
		1.36	0.26	-0.02	-1.61
	低	1.36	0.26	-0.02	-1.61
		1.36	0.26	-0.02	-1.61
		1.36	0.26	-0.02	-1.61

飼料の効果（＝飼料の水準平均－全体平均）

6. 繰り返しのある二元配置の分散分析(3/4)

- 純粋な統計誤差を数値化
  式(2)から右の表のように求められる．この純粋な統計誤差も変動Q_2として数値化する．
  $Q_2 = (-0.88)^2+(-0.02)^2+\cdots+(2.93)^2+(-1.32)^2 = 65.78$
  これが相対的に大きければ，純粋な統計誤差の効果が大きいことになる．

- 交互作用を数値化
  式(3)から右の表のように求められる．この交互作用の効果も変動Q_{13}として数値化する．
  $Q_{13} = 3\{(0.99)^2+(-1.06)^2+(-0.59)^2+\cdots+(0.56)^2\} = 21.76$
  この値が大きいと判断されると，交互作用の効果は大きいことになる．

- 不偏分散を算出
  まず，自由度を求める．
  温度の効果の自由度＝温度の水準数－１＝３－１＝２
  飼料の効果の自由度＝飼料の水準数－１＝４－１＝３
  交互作用の自由度＝（温度の水準数－１）×（飼料の水準数－１）＝２×３＝６
  統計誤差の自由度＝全データ数－温度の効果の自由度－飼料の効果の自由度
  －交互作用の自由度－１
  ＝温度の水準数×飼料の水準数×（繰り返し回数－１）＝２４
  よって，各変動を各自由度で割ると不偏分散が求まる．
  温度の効果の不偏分散$V_{11} = Q_{11}/2 = 10.98$
  飼料の効果の不偏分散$V_{12} = Q_{12}/3 = 13.54$
  交互作用の不偏分散$V_{13} = Q_{13}/6 = 3.63$
  統計誤差の不偏分散$V_2 = Q_2/24 = 2.74$

		飼料			
		A	B	C	D
温度	高	-0.88	-0.02	0.49	-2.43
		1.26	2.48	0.50	0.31
		-0.38	-2.46	-0.99	2.12
	中	0.62	1.46	-1.94	-0.05
		-0.97	-0.57	1.09	1.11
		0.34	-0.90	0.86	-1.05
	低	-0.77	1.80	-1.79	0.21
		-1.10	-0.60	-1.13	1.10
		1.86	-1.20	2.93	-1.32

		飼料			
		A	B	C	D
温度	高	0.99	-1.06	-0.59	0.66
		0.99	-1.06	-0.59	0.66
		0.99	-1.06	-0.59	0.66
	中	-0.41	0.51	1.12	-1.22
		-0.41	0.51	1.12	-1.22
		-0.41	0.51	1.12	-1.22
	低	-0.57	0.55	-0.54	0.56
		-0.57	0.55	-0.54	0.56
		-0.57	0.55	-0.54	0.56

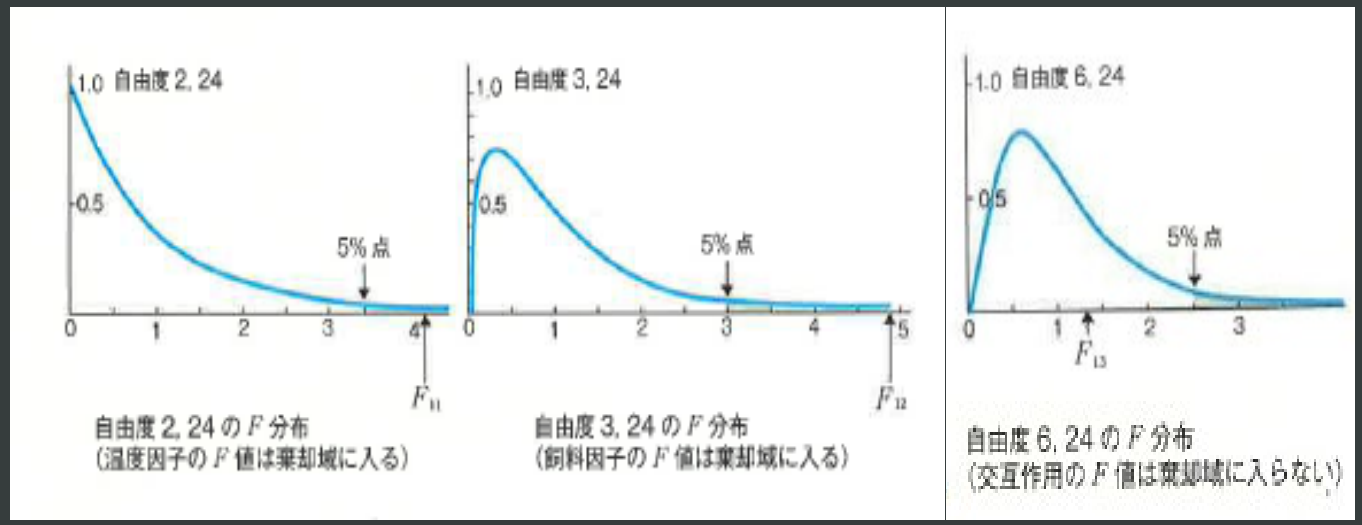
6. 繰り返しのある二元配置の分散分析(4/4)

- 仮説検定の実行
  まず，仮説（帰無仮説）を設定
  H_{10}：温度の効果は認められない
  H_{20}：飼料の効果は認められない
  H_{30}：交互作用は認められない
  これらの帰無仮説を5%の有意水準で検定する.
  不偏分散の比はF分布に従う．ですので，次のF値，F_{11}，F_{12}，F_{13}は各々自由度2, 24，自由度3, 24，自由度6, 24のF分布に従います.

$$F_{11} = \frac{V_{11}}{V_2} = \frac{10.98}{2.74} = 4.01, \quad F_{12} = \frac{V_{12}}{V_2} = \frac{13.54}{2.74} = 4.94, \quad F_{13} = \frac{V_{13}}{V_2} = \frac{3.63}{2.74} = 1.32$$

これらの値をF分布のグラフに記入してみると右のようになる.
グラフに示すように
自由度2,24のF分布の5%点=3.40
自由度3,24のF分布の5%点=3.01
自由度6,24のF分布の5%点=2.51

図からわかるように，温度と飼料の効果は有意水準5%の棄却域に入っています．そこで仮説H_{10}, H_{20}は棄却され，温度と飼料の効果が確かめらた.
これに対して，F_{13}は棄却域に入っていなく，仮説H_{30}は棄却されない．交互作用はこの資料からは検証できなかった.



7. 繰り返しのある二元配置の分散分析表(1/3)

■ 右の一般的な資料をもとに，分散分析の手順を示す．

① 変動の計算

因子Xの水準間変動Q_{11}，因子Yの水準間変動Q_{12}，交互作用の変動Q_{13}，純粋な統計誤差の変動Q_2は次のようになる．ここで，全体の平均値をm_T，因子Xの各水準ごとの平均値を$m_{X1}, m_{X2}, \ldots, m_{Xl}$，因子$Y$の各水準ごとの平均値を$m_{Y1}, m_{Y2}, \ldots, m_{Yk}$とする．また，$m_{X_iY_j}$は因子$X$の水準$i$と因子$Y$の水準$j$にある同一条件の$n$個のデータの平均値とする．

$$Q_{11} = nk\{(m_{X1} - m_T)^2 + (m_{X2} - m_T)^2 + \cdots + (m_{Xl} - m_T)^2\}$$
$$Q_{12} = nl\{(m_{Y1} - m_T)^2 + (m_{Y2} - m_T)^2 + \cdots + (m_{Yk} - m_T)^2\}$$
$$Q_{13} = n[\{(m_{X_1Y_1} - m_T) - (m_{X1} - m_T) - (m_{Y1} - m_T)\}^2 + \cdots + \{(m_{X_lY_k} - m_T) - (m_{Xl} - m_T) - (m_{Yk} - m_T)\}^2]$$
$$Q_2 = [\{(a_{11})_1 - m_{X_1Y_1}\}^2 + \{(a_{11})_2 - m_{X_1Y_1}\}^2 + \cdots + \{(a_{11})_n - m_{X_1Y_1}\}^2] + \cdots + [\{(a_{lk})_1 - m_{X_lY_k}\}^2 + \{(a_{lk})_2 - m_{X_lY_k}\}^2 + \cdots + \{(a_{lk})_n - m_{X_lY_k}\}^2]$$

		因子X			
		X_1	X_2	…	X_l
因子Y	Y_1	$(a_{11})_1$	$(a_{21})_1$	…	$(a_{l1})_1$
		$(a_{11})_2$	$(a_{21})_2$	…	$(a_{l1})_2$
		…	…	…	…
		$(a_{11})_n$	$(a_{21})_n$	…	$(a_{l1})_n$
	Y_2	$(a_{12})_1$	$(a_{22})_1$	…	$(a_{l2})_1$
		$(a_{11})_2$	$(a_{22})_2$	…	$(a_{l2})_2$
		…	…	…	…
		$(a_{11})_n$	$(a_{22})_n$	…	$(a_{l2})_n$
	…	…	…	…	…
	…	…	…	…	…
	Y_k	$(a_{1k})_1$	$(a_{2k})_1$	…	$(a_{lk})_1$
		$(a_{1k})_2$	$(a_{2k})_2$	…	$(a_{lk})_2$
		…	…	…	…
		$(a_{1k})_n$	$(a_{2k})_n$	…	$(a_{lk})_n$

X, Y：因子名，$X_1, X_2, \ldots, X_l, Y_1, Y_2, \ldots, Y_k$：水準名，$l, k$: 各水準のデータ数，繰り返し数$n$回

これを6.の例で計算したものが

変動要因	変動	自由度	不偏分散	分散比	F境界値
行(因子X)	21.95				
列(因子Y)	40.62				
交互作用	21.76				
誤差	65.77				

7. 繰り返しのある二元配置の分散分析表(2/3)

② 自由度の計算
因子X, Yの水準の数をl, k，繰り返し数がnとすると，各変動の自由度は，
因子Xの水準間の変動Q_{11}の自由度$= l - 1$
因子Yの水準間の変動Q_{12}の自由度$= k - 1$
相互作用の変動Q_{13}の自由度$= (l-1)(k-1)$
統計誤差の自由度$= (n-1)lk$
これを6.の例で計算したものが

変動要因	変動	自由度	不偏分散	分散比	F境界値
行(因子X)	21.95	2			
列(因子Y)	40.62	3			
交互作用	21.76	6			
誤差	65.77	24			

② 不偏分散の計算
変動を自由度で割ったものが不偏分散となる.
因子X, Yの水準間変動，交互作用の変動，そして統計誤差の不偏分散を各々V_{11}，V_{12}，V_{13}，V_2とすると，

$$V_{11} = \frac{Q_{11}}{l-1},\ V_{12} = \frac{Q_{12}}{k-1},\ V_{13} = \frac{Q_{13}}{(l-1)(k-1)},\ V_2 = \frac{Q_2}{(n-1)lk}$$

これを4.の例で計算したものが

変動要因	変動	自由度	不偏分散	分散比	F境界値
行(因子X)	21.95	2	10.98		
列(因子Y)	40.62	3	13.54		
交互作用	21.76	6	3.63		
誤差	65.77	24	2.74		

7. 繰り返しのある二元配置の分散分析表(3/3)

④ F値の計算

不偏分散V_{11}，V_{12}，V_{13}，V_2の比がF分布に従うので，次のF値を求める $F_{11}=\frac{V_{11}}{V_2}$，$F_{12}=\frac{V_{12}}{V_2}$，$F_{13}=\frac{V_{13}}{V_2}$
これを4.の例で計算したものが

変動要因	変動	自由度	不偏分散	分散比	F境界値
行(因子X)	21.95	2	10.98	4.01	
列(因子Y)	40.62	3	13.54	4.94	
交互作用	21.76	6	3.63	1.32	
誤差	65.77	24	2.74		

⑤ F分布のパーセント点を求める
帰無仮説で設定した有意水準に対応するF分布のパーセント点をF分布表やExcelなどの統計ソフトを利用して求め，表を埋める．ここでは5%点を求めます．有意水準に対するパーセント点のことをF境界値という．
これを4.の例で求めたものが

変動要因	変動	自由度	不偏分散	分散比	F境界値
行(因子X)	21.95	2	10.98	4.01	3.40
列(因子Y)	40.62	3	13.54	4.94	3.01
交互作用	21.76	6	3.63	1.32	2.51
誤差	65.77	24	2.74		

これで分散分析表が完成．

⑥ F値とパーセント点の大小を比較
F値が有意水準に対するパーセント点より大きければ，帰無仮説は棄却されます．対立仮説が採択される．

【サンプル３】繰り返しのある二元配置の分散分析

データサイエンス基礎_分散分析_(3)繰り返しのある二元配置の分散分析.ipynb

参考書

- 涌井(2010)「統計解析がわかる」 技術評論社
- 涌井(2015)「統計学の図鑑」 技術評論社
- 馬場(2018)「Pythonで学ぶ統計学の教科書」 翔泳社
- 谷合(2018)「Pythonで理解する統計解析の基礎」 技術評論社
- 栗原，丸山(2017)「統計学図鑑」 オーム社
- アクゼル，ソウンデルパンディアン(2007)「ビジネス統計学（上）（下）」 ダイヤモンド社
- 吉田(2006)「直観的統計学」 日経BP社
- 白井(2018)「確率統計学B講義資料」 http://whitewell.sakura.ne.jp/PythonProbStat/Python-statistics6.html
- 馬場「Pythonによる時系列分析の基礎」 https://logics-of-blue.com/python-time-series-analysis/
- 「生物統計学」 https://stats.biopapyrus.jp/
- 石村(1992)「分散分析のはなし」 東京図書
- 石村(2008)「入門はじめての分散分析と多重比較」 東京図書