第８回　Ｆ分布・分散分析その１　一元配置

Ａ．Ｆ分布

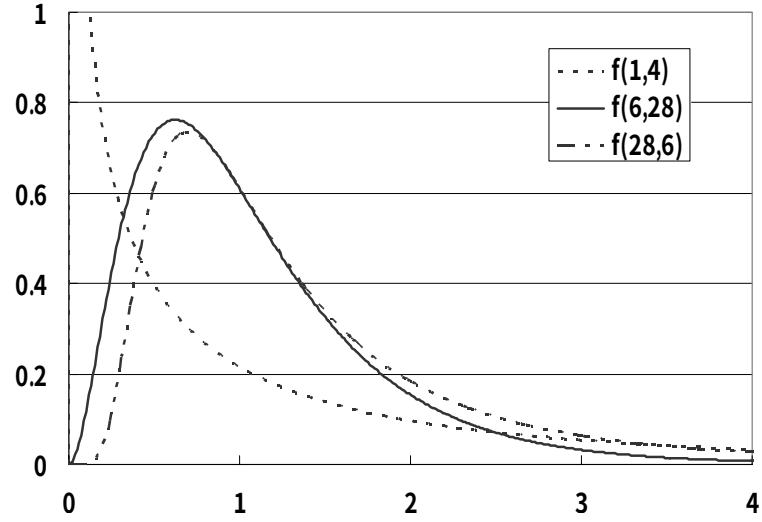
１．Ｆ分布

２つの母集団から得られた標本の分散比はＦ分布で検定する．分散分析という手法はこのＦ分布によって行うので，Ｆ検定は重要な手法である．カイ二乗分布と違い，２つの標本を比較するので，２つの標本それぞれの自由度からＦ分布が決まる．

F 分布：$x_{11}, x_{12}, \cdots, x_{1n_1}$ が互いに独立に $N(\mu_1, {\sigma_1}^2)$ に従い，さらにそれらと独立に

$x_{21}, x_{22}, \cdots, x_{2n_2}$ が互いに独立に $N(\mu_2, {\sigma_2}^2)$ に従うとき，

$F = \dfrac{V_1 / \sigma_1^2}{V_2 / \sigma_2^2}$ は自由度 $(n_1 - 1, n_2 - 1)$ の F 分布に従う．



Ｆ分布もカイ二乗分布と同じく，左右対称ではない．

２．Ｆ検定の手順

Ｆ分布に基づいて，２つの母集団の分散の比を推定することもできるが，あまり使わないので，ここでは述べない．

Ｆ検定は分散分析という手法の骨格をなす方法である．まずここでは基本的なＦ検定だけを練習しよう．

例：Ａ，Ｂ２種類の飼料を与えて一定期間飼育したハムスターの体重増加量を測定した結果，右のようなデータを得た（単位　g）．飼料による体重増加量のばらつきに差があるかを検定せよ．

飼料A	飼料B
73	71
74	68
70	73
77	71
72	71
72	72
69	69
73	72
72	68
75	

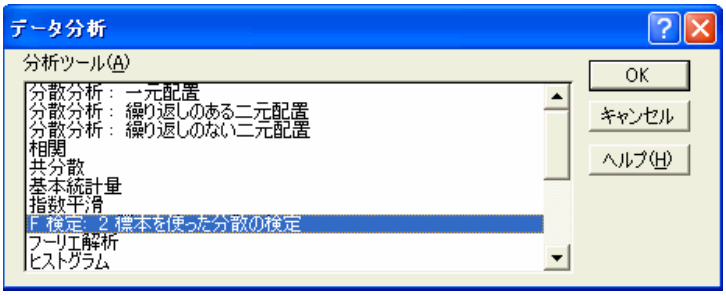
帰無仮説：$\sigma_A^2 = \sigma_B^2$　２つの標本の母分散は同じである．

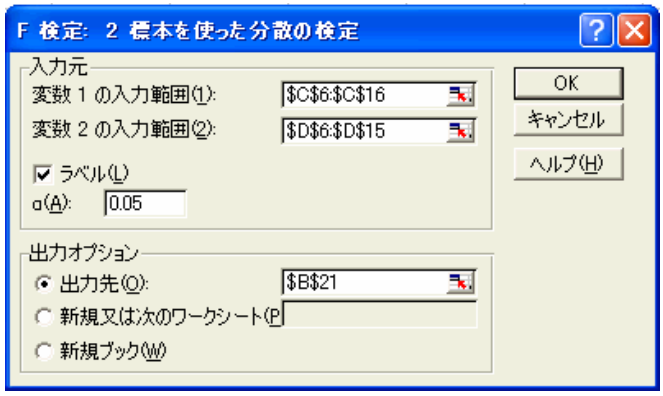
飼料 A と飼料 B で飼育したときにハムスターの体重増加量の分散に差がない．

対立仮説：$\sigma_A^2 \neq \sigma_B^2$　２つの標本の母分散は異なる．

飼料 A と飼料 B で飼育したときにハムスターの体重増加量の分散に差がある．

帰無仮説が成り立つと仮定したときに今回のデータが得られる確率 P 値はエクセルの分析ツールで以下のように計算できる．





F−検定：2 標本を使った分散の検定

	飼料A	飼料B
平均	72.7	70.55556
分散	5.344444	3.277778
観測数	10	9
自由度	9	8
観測された分散比	1.630508	
P(F<=f) 両側	0.251207	
F 境界値 両側	3.388124	

検定の結果

P 値が 0.2512 であるから，5%の有意水準で帰無仮説は棄却できない．したがって，5%の有意水準で飼料 A と飼料 B でハムスターの体重増加量のばらつきに差があるとはいえない．

練習　商社Ｓはタイでタマネギを栽培している．スーパーの基準は厳しく，ほとんど同じ直径(cm)のタマネギでないと納入させてくれない．今まで栽培していた品種Ａに代わり，多収で耐病性の強い品種Ｂを導入したいが直径のばらつきは品種Ａと同じではないのかもしれない．無作為に選んだ標本から右のようなデータを得た．品種Ｂのばらつきは品種Ａと異なるのかを有意水準 5%で両側検定せよ．

品種A	品種B
10.2	11.0
10.8	11.7
10.6	10.9
10.5	11.6
10.7	11.5
10.9	12.1
10.4	11.8
10.5	12.4
10.6	11.5
10.5	

B．統計的検定についての補足

★　正規分布を仮定できない場合の検定

ｔ分布，カイ二乗分布，Ｆ分布は母集団が正規分布するときに標本の統計量（それぞれ母平均，母分散，２つの母分散の比）がどのように分布するかを表したものである．したがって，母集団が正規分布しないときにこれらの分布を用いた推定・検定をすると誤りをおかす可能性がある．ただし多少正規分布にずれた分布でもこれらの分布で検定してもおおむね大きな間違いはないことがわかっている．しかし，正規分布を仮定することが危険である，あるいは正規分布に当てはめようのない分布ではノンパラメトリック検定（授業では説明しない）が利用できる．

C．分散分析とは？

１．３つ以上の標本平均から母平均が異なるかを同時に比較したい

例：ヤギの成長がよくなるという薬Ａ，Ｂ，Ｃ，Ｄのうち，どれが効果があるかあるいはないかを比較したい．したがって，対照区（コントロール，薬を与えない区），Ａ，Ｂ，Ｃ，Ｄの５つを比較することになる．処理：薬，水準：５つである．

次のような結果を得た．

処理	対照区	A	B	C	D
	100	105	96	100	115
	102	108	97	97	112
	104	110	100	95	100
	105	106	102	104	105
	103	104	99	103	106
平均	102.8	106.6	98.8	99.8	107.6
標準偏差	1.92	2.41	2.39	3.83	5.94

①　一番値の高いものと次に高いもの，２番目に高いものと３番目に高いものを次々とｔ検定すればよいのか？

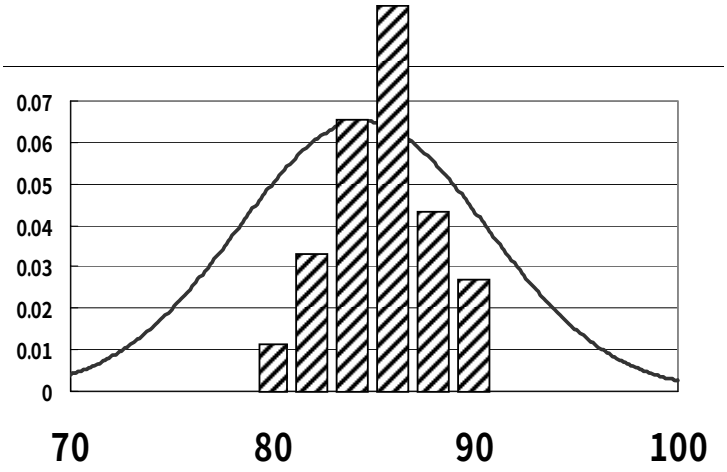
ｔ検定をいくども行う問題点

ｔ検定は決められた一組のデータについて行うと，決められた有意水準のもとで帰無仮説が棄却できるかを検討できる．上のように５つのデータセットの場合，${}_5C_2$＝１０回のｔ検定をすることになる．したがって，全体で見たら有意水準（危険率）は５％以上になってしまう．

同一の正規分布から 10 個のデータを次々と取って，一番平均の大きいものと一番平均の小さいものを選び，ｔ検定するとどうなるか？　20 回データを取ったときに実験すると・・・

★　ウズラの雄の体重の模擬実験

1000 羽のウズラ（平均体重 84.43g，標準偏差 6.12g）のデータから無作為（ランダム）に 10 羽を選んで，平均と標準偏差を計算した．一番平均の大きいものと一番平均の小さいものをｔ検定すると，５％の有意水準で有意差があった．



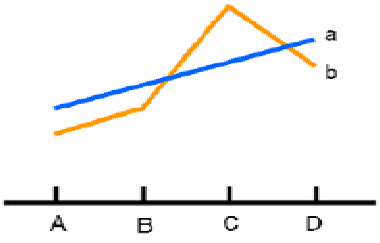
２．２つ以上の要因を同時に比較したい

例：ヤギの成長がよくなる薬A，B，C，Dはえさと混ぜると効果が高いかもしれない．そう考えて，麦わら，稲わら，濃厚飼料の３種類と組み合わせて試験をすることを考えた．処理は２つ，薬と飼料であり，薬は５水準，飼料は３水準ある．このとき比較するべき試験は，薬A＋麦わら，薬A＋稲わら，薬A＋濃厚飼料，薬B＋麦わら，薬B＋稲わら，薬B＋濃厚飼料，薬C＋麦わら，薬C＋稲わら，薬C＋濃厚飼料，薬D＋麦わら，薬D＋稲わら，薬D＋濃厚飼料，対照区＋麦わら，対照区＋稲わら，対照区＋濃厚飼料となる．

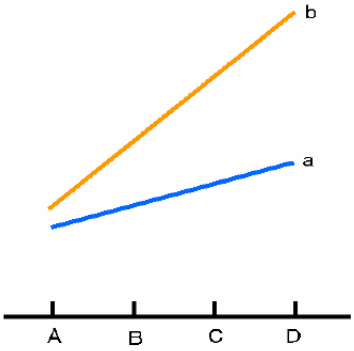
薬Aでは稲わらが一番よいえさだったが，薬Bでは麦わらが一番よいえさだったという結果が出るかもしれない．ｔ検定では２つの要因が絡み合った結果を解析できない．２ つ以上の要因が絡んだ結果（**交互作用**）はｔ検定では解析できない．

ｔ検定ではわからないこと　交互作用

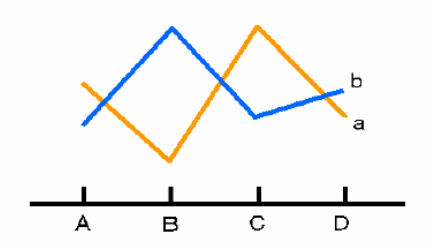
①　組み合わせの妙



②　相乗効果



③　打ち消しあう場合



交互作用のあるときは１つの要因（処理）だけを見て結論づけるわけには行かない．要因（処理）の組み合わせを考える必要が出てくる．

さらに要因を一つ一つ順番に調べるやり方では要因を組み合わせた場合の最適な条件にたどり着けるとは限らない．

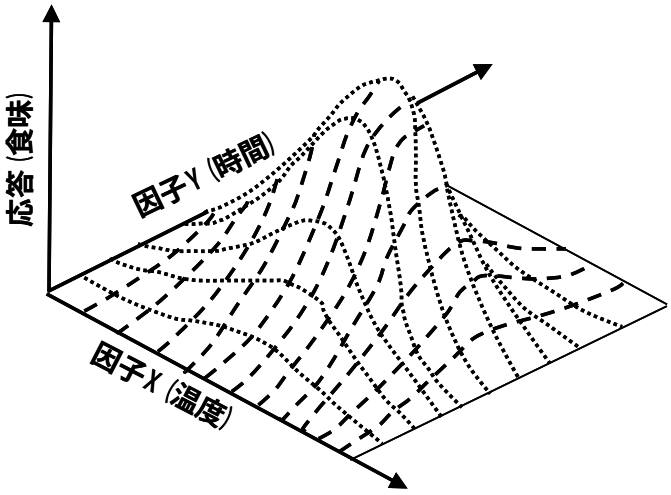


図　２因子に対する応答曲面

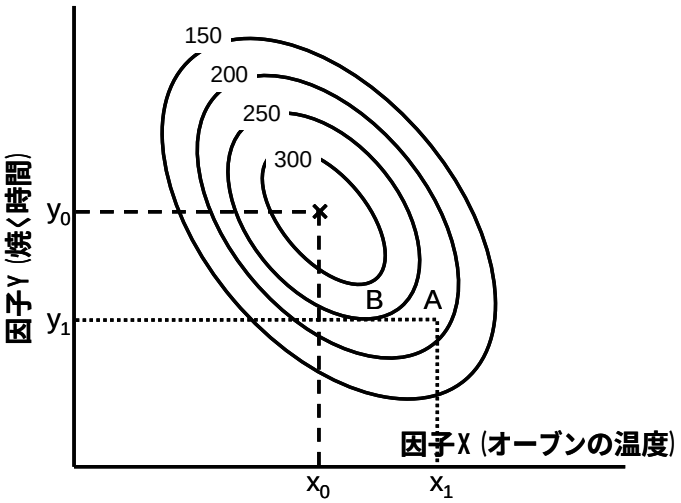


図　２因子応答曲面の等高線図

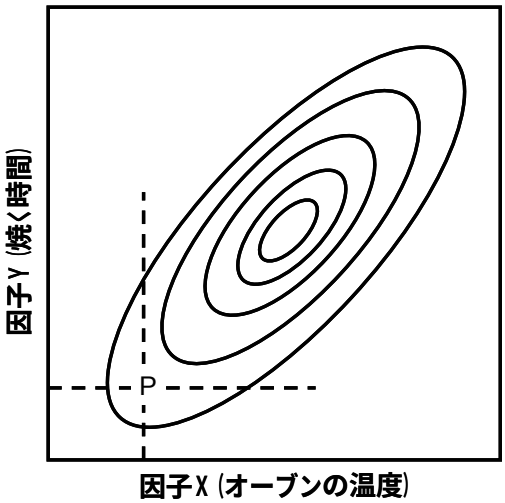


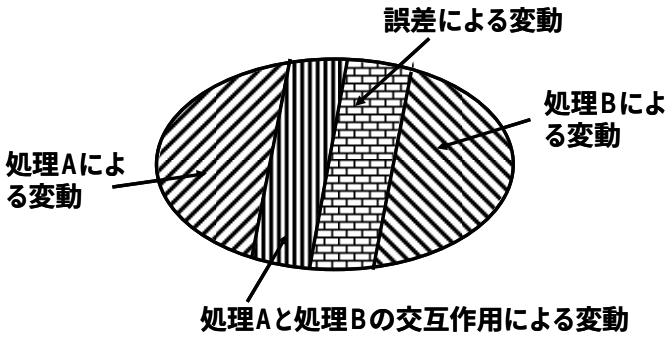
図　最適な水準に到達しない実験例
　上の例では因子Ｘを固定したまま，因子Ｙの水準を変えても応答は下がる．しかし，真の最適水準の組み合わせは遙かに右上にある．

３．ばらつき（分散）を偶然誤差と意味のあるものに分解する

処理の間の違い，交互作用は本当に意味のあるものなのか？（有意であるか？）

実験データのばらつきは処理によって生じたもの（意味のある部分）と誤差によるばらつき（意味のない部分）に分けることができる．

分散の加法性



データの総変動＝処理による変動＋誤差変動

＝個々の処理による変動（主効果）＋交互作用＋誤差変動

２つの標本の分散の検定　Ｆ検定

処理による変動が誤差による変動に比べて，十分に大きいのか（正確には処理による変動と誤差による変動は等しいという帰無仮説が棄却できるのか）を検討する．

Ｄ．分散分析の理論（一元配置の場合）

例：ハムスターをひまわり，大豆，人工餌の３種類のどれで育てるのが一番よいかを実験した．

①　実験結果に全く誤差がなく，餌の効果だけが現れたらどうなるか？

全く差がない場合，実験結果は下のようになる

	ひまわり	大豆	人工餌
1	15	15	15
2	15	15	15
3	15	15	15
4	15	15	15
5	15	15	15

餌の効果に差があるなら，（効果の合計=0）

	ひまわり	大豆	人工餌
1	0	-3	3
2	0	-3	3
3	0	-3	3
4	0	-3	3
5	0	-3	3

餌の効果に差があるなら，（効果の合計=0）

	ひまわり	大豆	人工餌
1			
2			
3			
4			
5			

② 実験結果に誤差がランダムに適当に混ざっているならばどうなるか？

餌の効果が誤差なく発揮されると
実験結果は下のようになる

	ひまわり	大豆	人工餌
1			
2			
3			
4			
5			

誤差があるなら,
（誤差の合計=0）

	ひまわり	大豆	人工餌
1	2	3	1
2	0	0	-4
3	3	-2	-1
4	-3	0	-1
5	1	-2	3

左の２つを足すと
実験結果は下のようになる

	ひまわり	大豆	人工餌
1			
2			
3			
4			
5			

誤差によって，データの処理による違いがわかりにくくなった.
餌の効果が±３に対して，誤差も±４もあるからである.

③ 得られた実験結果から誤差を分離してみる

実際に得られた実験結果は
下の通りである

	ひまわり	大豆	人工餌
1			
2			
3			
4			
5			

列の合計
列の平均
列の効果

左の実験結果から列の平均を
それぞれ引くと誤差を分離できる

	ひまわり	大豆	人工餌
1			
2			
3			
4			
5			

誤差の合計=0

④ 列の効果を判定する:誤差に比べて十分に大きいのか？

ばらつきのうち,誤差と効果によるものを分けて，比較してみよう（右の表).

誤差と効果を比較する

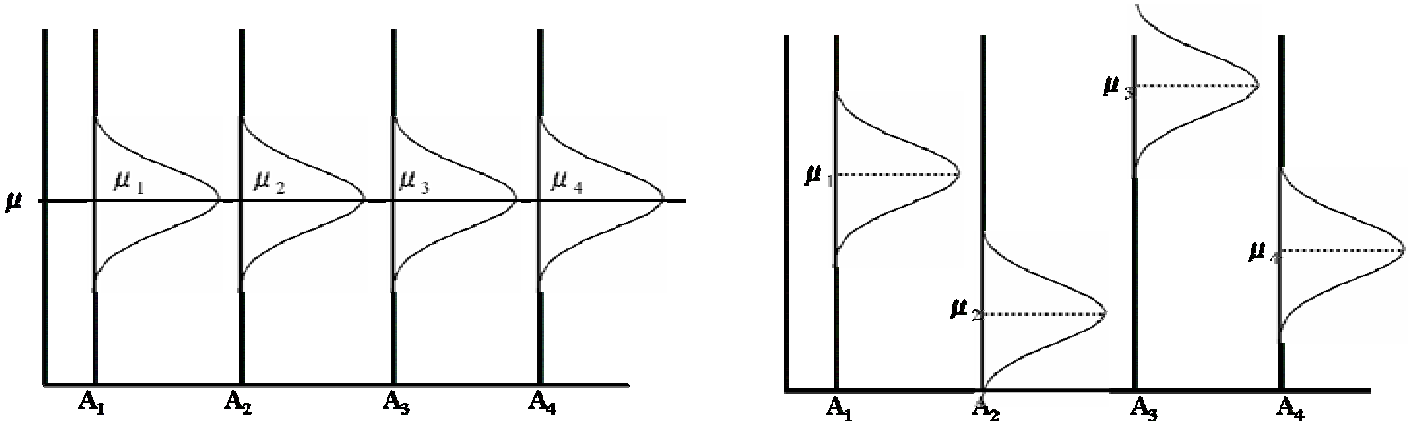
	繰り返し	ひまわり	大豆	人工餌
効果	1			
	2			
	3			
	4			
	5			
誤差	1			
	2			
	3			
	4			
	5			

⑤　**分散分析**を行う：効果による変動が誤差による変動に比べて十分に大きいか？をＦ検定で検定する.

帰無仮説：　効果による変動と誤差による変動には差がない.

誤差変動よりも処理の変動の方が大きいとかんがえてよいから片側検定となる.

上の帰無仮説のいうことは下の図のように読み替えることもできる



帰無仮説H_0の状態　　対立仮説H_1の状態

図　分散分析における帰無仮説と対立仮説

帰無仮説：$\mu_1 = \mu_2 = \mu_3 = \mu_4$　どの水準でも母平均は同じである

対立仮説：水準（処理）間の母平均のどれか一つは異なる

次の分散比Ｆを計算し，この分散比が得られる確率Ｐ値を計算する（実際はエクセルなどのソフトがデータを入力しただけでＰ値を計算する）

$$F = \frac{\text{効果のばらつきの大きさ}}{\text{誤差のばらつきの大きさ}} = \frac{\text{効果の分散}}{\text{誤差の分散}} = \frac{V_1}{V_2}$$

ここで　分散　$V = \dfrac{\sum (x_i - \bar{x})^2}{\phi}$

φは自由度：データの数から使用した平均値の数を引いたもの

効果には 0.6，-3.2，2.6 の３つしかなく，しかもこれらを作り出すために全体の平均を使っているので，自由度はφ＝３－１＝２となる.

したがって，右の表のように効果の分散を計算でき，
$V_1 = 43.4$であることがわかった.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
0.6	0.6	0.36
0.6	0.6	0.36
0.6	0.6	0.36
0.6	0.6	0.36
0.6	0.6	0.36
-3.2	-3.2	10.24
-3.2	-3.2	10.24
-3.2	-3.2	10.24
-3.2	-3.2	10.24
-3.2	-3.2	10.24
2.6	2.6	6.76
2.6	2.6	6.76
2.6	2.6	6.76
2.6	2.6	6.76
2.6	2.6	6.76

$$S = \sum (x_i - \bar{x})^2 = 86.80$$

$$V_1 = \frac{86.80}{2} = 43.4$$

誤差のデータは 15 個ある．使用した平均は各列にあるので，15.6，11.8，17.6 の３つであり，自由度は
φ＝１５－３＝１２となる．

x_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$
1.4	1.4	1.96
-0.6	-0.6	0.36
2.4	2.4	5.76
-3.6	-3.6	12.96
0.4	0.4	0.16
3.2	3.2	10.24
0.2	0.2	0.04
-1.8	-1.8	3.24
0.2	0.2	0.04
-1.8	-1.8	3.24
1.4	1.4	1.96
-3.6	-3.6	12.96
-0.6	-0.6	0.36
-0.6	-0.6	0.36
3.4	3.4	11.56

したがって，右の表のように誤差の分散を計算でき，$V_2 = 5.43$ であることがわかった．

以上からＦ値は $F = \dfrac{V_1}{V_2} = \dfrac{43.4}{5.43} = 7.99$

このようなＦ値が得られる確率Ｐ値を計算すると，0.006229 であることから（片側検定）

有意水準 1%で帰無仮説は棄却される．
処理と誤差のばらつきには有意水準1%で有意な差がある．すなわちハムスターの成長は餌によって変化すると結論できる．

$$S = \sum (x_i - \bar{x})^2 = 65.20$$

$$V_2 = \frac{65.20}{12} = 5.43$$

分散分析では誤差のばらつきより処理のばらつきの方が大きいと判断してよいから，片側検定である．

Ｅ．分散分析の実際（一元配置の場合）

１．分散分析の結果の表現方法

分散分析の結果はふつう下のような表に書いて示す．

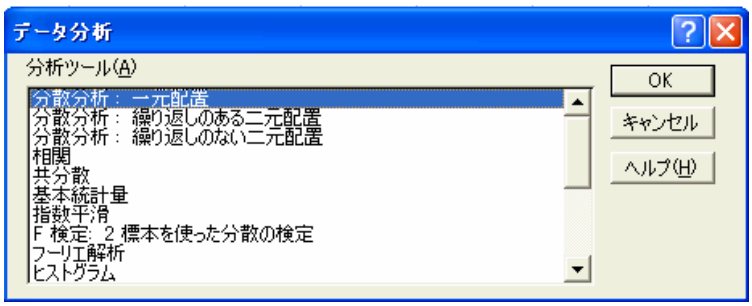
変動因	自由度φ	平方和Ｓ	平均平方（分散）Ｖ	分散比Ｆ
処理間	ϕ_A	S_A	V_A	F_0
誤差	ϕ_E	S_E	V_E	
全体	ϕ_T	S_T		

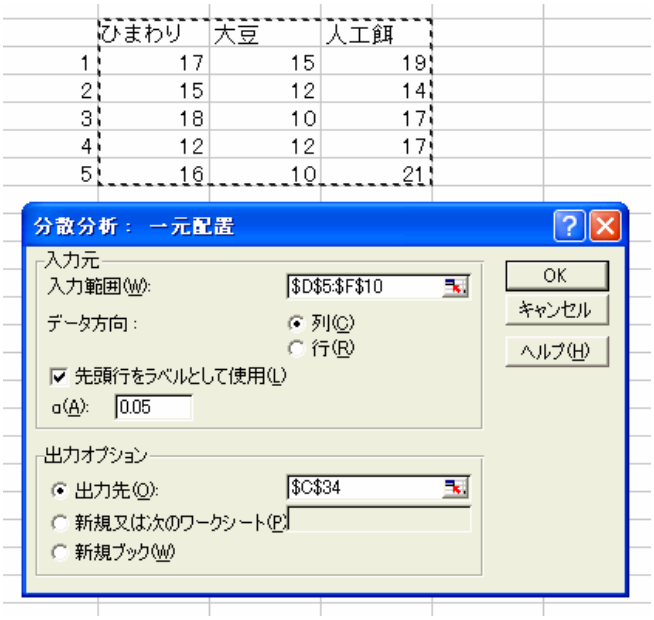
さらに有意水準５%，１%で有意であればそれぞれ，*，**をＦ$_0$の右肩につけるのが慣習となっている．有意差が検出されなかったときは ns をつけることもある．

２．エクセルでの分散分析

例:ハムスターをひまわり，大豆，人工餌の３種類のどれで育てるのが一番よいかを実験した．

	ひまわり	大豆	人工餌
1	17	15	19
2	15	12	14
3	18	10	17
4	12	12	17
5	16	10	21





分散分析：一元配置

概要

グループ	標本数	合計	平均	分散
ひまわり	5	78	15.6	5.3
大豆	5	59	11.8	4.2
人工餌	5	88	17.6	6.8

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	86.8	2	43.4	7.987730061	0.006229	3.88529
グループ内	65.2	12	5.433333			
合計	152	14				

エクセルで分散分析するときはＦ値を調べる必要なない．Ｐ－値をみれば，帰無仮説が成り立つ場合，今回えられたようなデータが出現する確率（Ｐ値）がわかる．Ｐ値が 0.05 以下であれば有意水準 5%で帰無仮説は棄却され，処理間の母平均に差があること，すなわち処理によってハムスターの成長が変わったことが示される．

３．分散分析の結果の意味

分散分析は処理（水準）間の母平均に差がないという帰無仮説を検定している

帰無仮説：$\mu_1 = \mu_2 = \mu_3 = \mu_4$　どの水準でも母平均は同じである

対立仮説：水準（処理）間の母平均のどれか一つは異なる

したがって，分散分析の結果，有意差があることが分かった場合，その意味するところは，処理（水準）の中で一つは母平均の異なるものがある，すなわち処理によって変わるということである．一般的には一番値の小さいものと一番値の大きいものとの間には有意差があるということになる．それ以外の処理（水準）間に差があるかは分散分析ではわからない．これを調べるのは多重比較法である．多重比較法については後の授業で簡単に説明する．

練習１：

例:ヤギの成長がよくなるという薬A，B，C，Dのうち，どれが効果があるかあるいはないか，どれが一番効果があるかを比較したい．したがって，対照区（コントロール，薬を与えない区），A，B，C，Dの５つを比較することになる．処理:薬，水準:５つである．

次のような結果を得た．

処理	対照区	A	B	C	D
	100	105	96	100	115
	102	108	97	97	112
	104	110	100	95	100
	105	106	102	104	105
	103	104	99	103	106
平均	102.8	106.6	98.8	99.8	107.6
標準偏差	1.92	2.41	2.39	3.83	5.94

練習２：右の例で水稲の１穂穎花数には品種間差があるかを分散分析せよ．

	水稲の1穂穎花数の品種間差異を5品種について調査した。				
	コシヒカリ	日本晴	NPT160	IR72	タカナリ
	56	80	188	131	177
	52	84	169	129	207
	64	64	131	106	179
	60	71	233	95	173
	17	81	186	152	109
	98	93	183	152	151
	122	107	167	145	143
	101	104	135	144	162
	128	77	148	153	123
	85	60	126	83	188
	125	96	177	160	96
	60	111	158	115	152
	108	78	127	113	162
	85	80	162	71	111
	83	80	117	76	139
	97	94	144	101	154
	84	85	165	76	132
	108	92	98	68	214
平均	85.16667	85.38889	156.3333	115	154
標準偏差	29.40038	13.9332	31.93007	31.8932	32.76476

帰無仮説：

対立仮説：

検定結果：

４．分散分析をする上での注意

①　なるべく反復数はそろえる

今回学んだ一元配置の分散分析では反復数が個々の処理区で異なっていてもそれほど問題はない．しかし，来週以降に学ぶより複雑な分散分析では，反復数が異なると解析が面倒になるだけでなく，精度も大きく落ちてしまう．実験開始のときは反復数をそろえて実験するのが普通であるが，事故や不注意なので反復数がそろわなくなることもあるかもしれない．しかし，できるだけ反復のそろうように実験することが基本である．なお反復がそろわないからといって，一部のデータを削除するのは間違ったやり方である．

②　複雑な実験はなるべく避ける

前項とも関連するが，分散分析ではデータが複雑になるほど，解析が面倒かつ間違えやすくなる．特にコンピューターで計算させるときは，データの入力の仕方を間違いやすくなり，自分の目的とする分散分析をするにはデータの構造が複雑（あるいはでたらめ）で，解析不能ということもあり得る．そのうえ，反復がそろわなかったときの影響も大きくなる．必要のない複雑な実験は避けるのはもちろんのこと，必要だとしてもできるだけ簡単な実験計画にならないかをよく検討してから実験するべきである．実験計画を立てた時点で，どういう分散分析をするのかを決めておくのが正しい統計解析方法である．

③　正規分布するデータが前提条件である

分散分析では比較する母集団それぞれが正規分布すること，母分散が等しいことが理論的には前提条件になる．しかし，分散分析は多少その前提条件からはずれていても，結果が大きく左右されない頑健性をもっている．

④　すべての水準に対して母分散が等しいことも前提条件である

水準ごとの反復数がみな同じである場合，この前提が多少崩れても影響はあまりない．

⑤　フィッシャーの３原則を満たした実験計画のもとで，分散分析を行う

誤差に系統誤差が入ると解析結果の妥当性が失われる危険性がある．系統誤差を除去したり，分散分析の解析の妨害とならない偶然誤差に転化するのがフィッシャーの３原則に述べられた反復，無作為化，局所管理である．詳細は後の授業で学ぶ予定である．

Ｆ．宿題

１．第５回の宿題で調査したデータを用いて，卵の重さのばらつきが２つの店の間で同じであるかを有意水準 5%でＦ検定（両側検定）せよ．

２．第７回の宿題３．で調べたデータについて分散分析せよ．

３．次回の授業では二元配置の分散分析について学ぶ．２つの処理を組み合わせたときのデータの解析方法である．

①　産地と品種の違いがリンゴの糖度に及ぼす影響を知りたい．
②　施肥量と品種の違いがイネの 1 穂穎花数に及ぼす影響を知りたい．
③　作期と品種の違いがトマトの酸度に及ぼす影響を知りたい．

二元配置で分散分析できそうなデータを研究室の卒業実験などから手に入れる．あるいは自分で実験してもかまわない．なお，データには反復が２つ以上あり，それぞれの処理で反復の数は同じにすること．データの配置は下の図のようになる．

産地と品種の違いがリンゴの糖度に及ぼす影響を各処理３つずつのリンゴで調査した．

産地：青森，長野，山形，　品種：紅玉，ふじ，むつ

	青森	長野	山形
紅玉	（　）（　）（　）	（　）（　）（　）	（　）（　）（　）
ふじ	（　）（　）（　）	（　）（　）（　）	（　）（　）（　）
むつ	（　）（　）（　）	（　）（　）（　）	（　）（　）（　）

他にも次のような例が考えられる

コンビニ弁当の売れ行き　曜日，天気の２つの要因

イネの 1 穂穎花数　品種，施肥の２つの要因

小鳥のさえずる回数　気温，太陽の明るさの２つの要因