t 検定と F 検定

ここまでの講義で、統計学にある思想というか、信念があるのだと感じた人はいないだろうか。その思想が妥当かどうか私にはわからなが、経験的には大体あっているのだろうという気がする。つまり「その実在を突き詰めることはできないけれど、何かあるべき理想というか観念的な真理のようなものが、点のように質量や体積を持たずに存在し、現実は、その周辺に確率的に分布している。」という考え方だ。この世界観を受け入れて、発展させると次のような応用展開が生まれる。「現実の確率事象の広がり方はそれぞれの事象によって異なるのだが、その事象が起こる確率密度は、真理からの距離に依存し、距離が離れると確率が低くなる。その広がり方を何かの単位で基準化すれば、現実が表れる確率を、標準的な確率分布関数として表現できる。」、こうした考え方から、正規分布、χ^2分布、t分布、F分布、等々、様々な確率分布が確率密度関数として描かれる。こうした、確率密度関数も実は極めて観念的なもので、「観念的な真理（たとえば、「同じ母集団から取り出した 2 つのサンプル集団の平均値の差は 0 であるべきだ。」あるいは、「同じ母集団から取り出した 2 つのサンプル集団の分散の比は 1 であるべきだ。」のようなものを中心にして、これも現実には存在しない広がり方の単位（真の分散）が使われて、確率密度関数が描かれています。

t分布の場合、現実をM、あるべき理想（真理）をμ、広がり方の単位をΦとして

$$\frac{\mathrm{M}-\mu}{\Phi}=\tau \qquad \textit{式 1}$$

τの確率分布を描いたのがt分布です。t検定は、普通、2 つのサンプルグループの平均値の差の有意性の検定に使われます。2 グループの差の有意性の検定以外に、t 分布が使われる例があるかどうか、考えてみたけれど思いつきませんでした。t分布は、「2 つのサンプルグループの平均値の差の有意性の検定：t 検定に使われる確率密度関数」と覚えてしまっても、おそらく、問題ないでしょう。2 つのサンプルグループAとBの間に差があるというための帰無仮説は「本当に差がないのならばこうなるはずだ」だから、仮説となるべき観念的な真理は「同じ母集団から取り出した 2 つのサンプル集団の平均値の差は 0 であるべきだ。」となります。

すると式 1 は

$$\frac{\mathrm{M}-\mu}{\Phi}=\frac{\left(A_{average}-B_{average}\right)-0}{\Phi}=\frac{A_{average}-B_{average}}{\Phi}=\tau \qquad \textit{式 2}$$

となります。ここで、わからないのはΦです。Φも観念的な概念ですから、実際に得られたデータから、Φをどのように推定すればよいのかを考えなくてはなりません。すでに、

私たちは、標準誤差（standard error:*S.E*）という概念を知っています。もし、Mがとるべき、観念的に正しい、標準誤差を知っていれば、

$$\Phi = \overline{S.E}$$

として、

$$\frac{A_{average} - B_{average}}{\overline{S.E}} = \tau \qquad 式3$$

となります。$\overline{S.E}$とわざわざバーを付けたのは、そんなこと知る分けないだろうという、私の気持ちの表現です。何らかの方法で$\overline{S.E}$を外から与えれば、τはt分布にしたがうだろうということです（あくまで観念的に)。

講義で示しておいた標準誤差の式は

$$S.E = \frac{\sigma}{\sqrt{n}}$$

普通、σは母集団の標準偏差で、サンプル集団の平方和 SSと自由度nを使って、

$$\sigma^2 = \frac{SS}{n-1} \qquad 式4$$

と計算して、母集団の分散:普遍分散を求め、その平方根が標準偏差だと習います。この普遍分散は、サンプル集団から求めた普遍分散の推定値で、ここで、議論している$\overline{S.E}$: 2 つのグループの平均値の差とあるべき平均値の差 0 の距離の単位とすべきかどうかわからないと言えばわからないのですが、それでも、このままでは、現実の世界をデータとして、この抽象的な世界観の中に持ち込めないから、現実のサンプル集団から求めた普遍分散σ^2の平方根σを根拠として、S.E を作らざるを得ないでしょう。とりあえず、そういう分散σやサンプルサイズnがあることにして式 3 を書き換えておきます。

$$\frac{A_{average} - B_{average}}{\frac{\sigma}{\sqrt{n}}} = t \qquad 式5$$

この式でtを定義し、$A_{average} - B_{average} = 0$であるべき時に、現実の世界で$t$が観測される確率を議論することになります。これが t 検定でやる作業です。

ですから、サンプル・データの形の応じてσやnをどのように推定するのかが、第一に議論すべき問題です。初歩の統計学の教科書では、多くのページを割いて、この問題を論じています。長い説明で読んでて嫌になるのですが、とりあえず、２つのデータのタイプに分けます。これが、対になったデータの t 検定と対になっていないデータの t 検定という話です。まず、対になったデータの t 検定から話します。対になったデータの t 検定を先に話すのは、σとnの推定がわかりやすく、*式5*をそのまま使って、t 検定が行えるからです。とりあえず、σとnがわからないことにして、*式5*を下の式のように書き換えておきます。

$$\frac{A_{average} - B_{average}}{\frac{\phi}{\sqrt{\nu}}} = t \qquad 式5$$

ϕ: *差の母数団の標準偏差の推定値*、ν: *推定に用いデータのサンプルサイズ*

対になったデータを比べるとは、たとえば、右手と左手とどちらが長いかとか、手と足とどちらが長いかとか、一つの鉢に違う植物をうえて、これを繰り返してどちらの植物の成長が早いかなどをくらべることをイメージすればよいでしょう。先週の雑談のところで話した、養殖場の周辺のサンプリングステーションの、養殖場建設前後のコメツキガニの安定同位体比も対になったデータの例です。対になっているのだから、１番目の人の右手にはそれと対になった左手が必ずあります。このデータをＡ１と B1 と表現する と、以下Ａ２，Ｂ２，Ａ３，Ｂ３のように表現できます。必ず一組になっているのだから、

$$C_k = A_k - B_k \qquad (k = 1, \cdots n)$$

として、n 個の C_kを求めることができます。C_kは標本から得られるデータで、確率的に変動します。ＡとＢの平均に差がないということはＣの平均が 0 ということです。つまり、観念的にあるべきＣの平均を考えると$C_{average} = 0$で、帰無仮説は$C_{average} = 0$です。 この場合は、Ｃデータの数をnデータ数νとして、Ｃから求めた普遍分散σ^2から求めた偏差σを普遍分散ϕとすることに何の問題もないでしょう。また、A,B,C でnが同じだから

$$A_{average} - B_{average} = C_{average}$$

したがって式５を次のように書き換えることが出来ます。

$$\frac{C_{average}}{\frac{\sigma}{\sqrt{n}}} = t \qquad 式6$$

これが、対になったt検定であたえる、t値の式です。

次に、。ϕとνをどのように推定するかという問題を、対になっていない場合について考えます。これは、和の分散、差の分散という問題に帰着します。これについてはブログに書きましたが、あらためてここで説明します。

和の分散

検討するモデル

Aというデータ群（$A_1, A_2, \cdots, A_{n_A}$)と$B$というデータ群（$B_1, B_2, \cdots, B_{n_B}$）があることにします。そこから各データを足し合わせた$A + B$というデータ群とか、$A - B$あるいは$A \times B$というデータ群を作ることを考えます。対になっている場合には足し合わせる相手がわかっているから$A + B$は（$A_1 + B_1, \cdots, A_n + B_n$）のように、$n$個のデータをつくれば良いのですが、対になっていない場合は、AのどのデータとBのどのデータを足し合わせればよいか決まっていません。そこで、考えられる組み合わせのすべてについて$n_A n_B$ 個のデータをつくり、そのデータの分散について考えます。この解説は原理的な理解のためにやっているので、実用的な意味を考えていません。筆者はデータの和という概念を具体的に使う場面を思いつきません。たとえば、鉢に違う種類の植物を植えて、その成長量の和を確率的に論ずることに何か意味があるとは思えないでしょう。ただ、データの和の分散がどのようになる

のかを考えることは、分散をいくつかの要因に取り分けるときるときの基本的な考え方です。私は、この考え方を「総当たり的」な考え方と呼んでいます。「総当たり的」な議論はＦ検定の操作手順を理解するのに役立ちます。

$n_A \times n_B$の総当り表を作ります（表６）。表 7 はその具体例で、A のデータ群として(1,5,6)、B のデータ群として(1,5,6,8)を使って、それらの和のデータを作 っています。

表１.総当たりで和の平均を求める計算

	A_1	$\cdots$	A_i	$\cdots$	A_{n_a}	合計	平均
B_1	$A_1 + B_1$	$\cdots$	$A_i + B_1$	$\cdots$	$A_{n_a} + B_1$	$\sum_{i=1}^{n_a} A_{n_a} + n_a B_1$	$M_A + B_1$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
B_j	$A_1 + B_j$	$\cdots$	$A_i + B_j$	$\cdots$	$A_{n_a} + B_j$	$\sum_{i=1}^{n_a} A_{n_a} + n_a B_j$	$M_A + B_j$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
B_{n_B}	$A_1 + B_{n_B}$	$\cdots$	$A_i + B_{n_B}$	$\cdots$	$A_{n_a} + B_{n_B}$	$\sum_{i=1}^{n_a} A_{n_a} + n_a B_{n_B}$	$M_A + B_{n_B}$
合計	$n_B A_1 + \sum_{j=1}^{n_B} B_j$	$\cdots$	$n_B A_i + \sum_{j=1}^{n_B} B_j$	$\cdots$	$n_B A_{n_a} + \sum_{j=1}^{n_B} B_j$	$n_B \sum_{i=1}^{n_a} A_{n_a} + n_a \sum_{j=1}^{n_B} B_j$	$n_B M_A + n_B M_B = n_B(M_A + M_B)$
平均	$A_1 + n_B M_B$	$\cdots$	$A_j + n_B M_B$	$\cdots$	$A_{n_a} + n_B M_B$	$n_A M_A + n_A M_B = n_A(M_A + M_B)$	$M_A + M_B$

足し合わせた数の平均なのだから、元のグループの平均の和になるのは当然だろうと言えば当然なのですが、いわゆる、重み付きの平均とは違っています。先にそれぞれの平均を求めて、それぞれのデータ数を重みとして、重み付き平均を求める式は $\frac{n_A M_A + n_B M_B}{n_A + n_B}$ です。表１の合計と平均の交差するところを色を付けておきました。黄色のところの合計を求めて、n_aで割って、赤いところの数値を出して、赤いところの数値をn_Bで割って、青の$M_A + M_B$という計算結果を出すというルートでも、n_bで割って紫経由ルートで$M_A + M_B$に行っても良いのですが、黄色をいきなり、$n_a n_B$で割った方が手っ取り早いので、普通はそう計算するでしょう。

このように求めた平均を何と言うのか知りません。総当たり平均とでも言うのでしょうか。それぞれのグループの個々の値が相手のグループの個々の値と対になる回数で重みを付けているので、私には、重み付き平均の一種のような気がします。もちろん、一般に使われてる「重み付き平均」とは大きく意味が異なります。

数式だけを追っているとわかりにくいかもしれないので、具体的な計算例を表２に示しま

した。

表 2.表 1 の計算の具体例

		A				
		1	5	6	合計	平均
B	1	2	6	7	15	5
	5	6	10	11	27	9
	6	7	11	12	30	10
	8	9	13	14	36	12
	合計	24	40	44	108	36
	平均	6	10	11	27	9

合計$M_{A+B}: total = \frac{108}{12} = 9$

$M_A = \frac{1+5+6}{3}$=4

$M_B \frac{1+5+6+8}{4}$=5

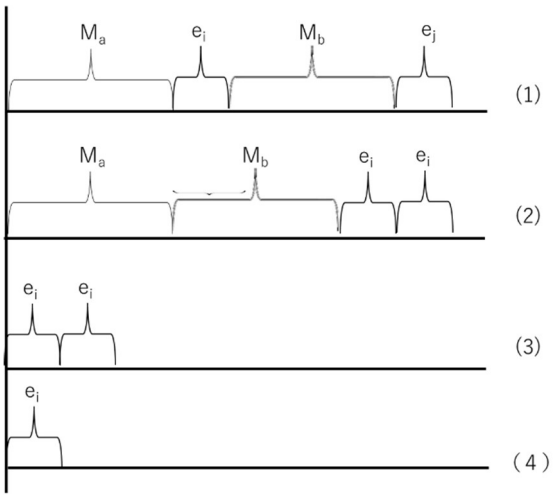
ちなみに、重み付き平均は、$\frac{4\times3+5\times4}{3+4} = \frac{32}{7} = 4{,}571$

確かに、総当たりで行った全体の平均は、$M_A + M_B$になっています。データを足し合わせているのだから、その平均値も平均値の和になるというのは当然ですね。元の 2 つのデータ群の分散を知っていることにして、これらを利用して、足し合わせたデータの母集団の平均値周りの分散を推定することを考えます。第一段階として、個々のデータを構成する要素を分離してSS を計算してみます。このSSをSS_{A+B}、分散を$\sigma^2{}_{A+B}$と表すことにします。個々のデータx_i、平均値をM、e_i を個々のデータの平均値からの隔たりとすると。

$$x_i = M + e_i \qquad 式\ 7$$

となります。これが個々のデータの構成要素です。サンプル集団AのA_iとサンプル集団BのB_jをたし合わせることを考えると、図 1 の（1）のような構成になります。

図 1．和のデータの構成要素



図の横棒は、$A_i + B_j$の値を示す数直線です。直線の左端が 0 です。足し算だから(2)のように順番を変えても値は変わらないでしょう。M_aとM_bはそれぞれの集団の中で共通だから、それを取り除けば、(3)のようになります。eは偏差だから、負の値もあって、直線は 0 の左側に伸びることもあります。

ここで、この足し合わせたものの平均値からの隔たりを$e_{A+B_{ij}}$と表すと（持ってまわった言い方ですが、簡単に言えば、（3）の直線の長さのこと)、$e_{A+B_{ij}} = e_{A_i} + e_{B_j}$となります。

これを表１の総当たりの計算にすると、表３のようになります。

表３.偏差の和の平均を求める計算

	A_1	⋯	A_i	⋯	A_{n_a}	合計	平均
B_1	$e_{A_1}+e_{B_1}$	⋯	$e_{A_i}+e_{B_1}$	⋯	$e_{A_{n_a}}+e_{B_1}$	$0^*+n_a e_{B_1}$	e_{B_1}
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮
B_j	$e_{A_1}+e_{B_j}$	⋯	$e_{A_i}+e_{B_j}$	⋯	$e_{A_{n_a}}+e_{B_j}$	$0^*+n_a e_{B_j}$	e_{B_j}
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮
B_{n_B}	$e_{A_1}+e_{B_{n_b}}$	⋯	$e_{A_i}+e_{B_{n_B}}$	⋯	$e_{A_{n_a}}+e_{B_{n_B}}$	$0^*+n_a e_{B_{n_B}}$	$e_{B_{n_B}}$
合計	$n_b e_{A_1}+0^{**}$	⋯	$n_b e_{A_i}+0^{**}$	⋯	$n_b e_{A_{n_a}}+0^{**}$	$0+0^{**}$	0
平均	e_{A_1}	⋯	e_{A_i}	⋯	$e_{A_{n_a}}$	0	0

＊：$\sum_{i=1}^{n_a} e_{A_1}=0$　偏差の総和は０　＊＊：$\sum_{j=1}^{n_b} e_{B_1}=0$　偏差の総和は０

これも、偏差の平均は０に決まっているだろうと言われれば、その通りです。

同じことを偏差の２乗についてやってみます。

$$\left(e_{A_i}+e_{B_j}\right)^2 = {e_{A_i}}^2+2e_{A_i}e_{B_j}+{e_{B_j}}^2 = {e_{A_i}}^2+{e_{B_j}}^2+2e_{A_i}e_{B_j}$$

と展開した形で書きます。

表４.偏差の和の２乗平均を求める計算

	A_1	⋯	A_i	⋯	A_{n_a}	合計	平均
B_1	${e_{A_1}}^2+{e_{B_1}}^2+2e_{A_1}e_{B_1}$	⋯	${e_{A_i}}^2+{e_{B_1}}^2+2e_{A_i}e_{B_1}$	⋯	${e_{A_{n_a}}}^2+{e_{B_1}}^2+2e_{A_{n_a}}e_{B_1}$	$\sum_{i=1}^{n_a}{e_{A_i}}^2+n_a{e_{B_1}}^2+2\times 0^*\times e_{B_1}$	$\frac{SS_a}{n_a}+{e_{B_1}}^2$
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮
B_j	${e_{A_1}}^2+{e_{B_j}}^2+2e_{A_1}e_{B_j}$	⋯	${e_{A_i}}^2+{e_{B_j}}^2+2e_{A_i}e_{B_j}$	⋯	${e_{A_{n_a}}}^2+{e_{B_j}}^2+2e_{A_{n_a}}e_{B_j}$	$\sum_{i=1}^{n_a}{e_{A_i}}^2+n_a{e_{B_j}}^2$	$\frac{SS_a}{n_a}+{e_{B_j}}^2$
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮
B_{n_B}	${e_{A_1}}^2+{e_{B_{n_b}}}^2+2e_{A_1}e_{B_{n_b}}$	⋯	${e_{A_i}}^2+{e_{B_{n_b}}}^2+2e_{A_i}e_{B_{n_b}}$	⋯	${e_{A_{n_a}}}^2+{e_{B_{n_b}}}^2+2e_{A_{n_a}}e_{B_{n_b}}$	$\sum_{i=1}^{n_a}{e_{A_i}}^2+n_a{e_{B_{n_b}}}^2$	$\frac{SS_a}{n_a}+{e_{B_{n_B}}}^2$
合計	$n_b{e_{A_1}}^2+\sum_{j=1}^{n_b}{e_{B_j}}^2$	⋯	$n_b{e_{A_i}}^2+\sum_{j=1}^{n_b}{e_{B_j}}^2$		$n_b{e_{A_{n_a}}}^2+\sum_{j=1}^{n_b}{e_{B_j}}^2$	$n_b\sum_{i=1}^{n_a}{e_{A_i}}^2+n_a\sum_{j=1}^{n_b}{e_{B_j}}^2$	$\frac{n_b SS_a}{n_a}+SS_b$
平均	${e_{A_1}}^2+\frac{SS_b}{n_b}$	⋯	${e_{A_i}}^2+\frac{SS_b}{n_b}$	⋯	${e_{A_{n_a}}}^2+\frac{SS_b}{n_b}$	$SS_a+\frac{n_a SS_b}{n_b}$	$\frac{\sum_{i=1}^{n_a}{e_{A_i}}^2}{n_a}+\frac{\sum_{j=1}^{n_b}{e_{B_j}}^2}{n_b}$

＊：$\sum_{i=1}^{n_a} e_{A_i}=0$

黄色のマーカーで印を付けたところに注目します。黄色のマーカーで印を付けたところに

注目します。これは、総当たりの和を作った時に全平均からの偏差の 2 乗総和、つまり、$n = n_a n_b$の平方和 SS です。総当たり全ての SS という意味で、SS_{total}という記号にしておきます。SS_{total}を$n_a n_b$で割れば個の総当たりを一つの標本集団としたときの、サンプルの分散です。それを実施した結果が、青のマーカーで印をつけたところです。

$$\frac{SS_{total}}{n_a n_b} = \frac{\sum_{i=1}^{n_a} {e_{A_i}}^2}{n_a} + \frac{\sum_{j=1}^{n_b} {e_{B_j}}^2}{n_b} = \frac{SS_a}{n_a} + \frac{SS_b}{n_b}$$

$$SS_{total} = n_b SS_a + n_a SS_b$$

というのが、ここで試行的に行った計算の一つの結論なのですが、我々が求めているのは、SS_{total}から求めた分散ではありません。この結論は、F 検定をするときに、全体の分散と部分分散を分けるときに意味があるので、知っておいて損はありませんが、一旦それを忘れて、我々が求めている分散が何かを考えます。表 3 の偏差の和の平均を求める計算に戻ります。

表 3 .偏差の和の平均を求める計算（再掲）

	A_1	⋯	A_i	⋯	A_{n_a}	合計	平均
B_1	$e_{A_1} + e_{B_1}$	⋯	$e_{A_i} + e_{B_1}$	⋯	$e_{A_{n_a}} + e_{B_1}$	$0^* + n_a e_{B_1}$	e_{B_1}
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮
B_j	$e_{A_1} + e_{B_j}$	⋯	$e_{A_i} + e_{B_j}$	⋯	$e_{A_{n_a}} + e_{B_j}$	$0^* + n_a e_{B_j}$	e_{B_j}
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮
B_{n_B}	$e_{A_1} + e_{B_{n_b}}$	⋯	$e_{A_i} + e_{B_{n_b}}$	⋯	$e_{A_{n_a}} + e_{B_{n_b}}$	$0^* + n_a e_{B_{n_b}}$	$e_{B_{n_b}}$
合計	$n_b e_{A_1} + 0^{**}$	⋯	$n_b e_{A_i} + 0^{**}$	⋯	$n_b e_{A_{n_a}} + 0^{**}$	$0 + 0^{**}$	0
平均	e_{A_1}	⋯	e_{A_i}	⋯	$e_{A_{n_a}}$	0	0

ここから、いきなり、表 4 に行き、偏差の和の二乗をたし合わせています。表 4 では、A_i列のB_j行の偏差の和の 2 乗を${e_{A_i}}^2 + 2e_{A_i}e_{B_j} + {e_{B_j}}^2$と計算しています。偏差というのは平均値からの距離ですね。表 3 のB_j行を見ていくと、$e_{A_1} + e_{B_j}, \cdots, e_{A_i} + e_{B_j}, \cdots, e_{A_{n_a}} + e_{B_j}$となっています。合計が$n_a e_{B_j}$、平均が$e_{B_j}$だというのも間違いではありません。

しかし。もし、A だけに着目して、列の違いをランダムだと考えれば、偏差の和から行の平均を差し引いたものを偏差とするでしょう。つまり、B_j行については

$$e_{A_i} + e_{B_j} - e_{B_j} = e_{A_i}$$

これが平均値からの距離（偏差）です。つまり図 1 の(4)のようになっているのです。これを表 4 の形式で書くと、

表５．SS_{A+B}を求める計算（行ごとに平均する計算）

	A_1	⋯	A_i	⋯	A_{n_a}	合計	平均
B_1	${e_{A_1}}^2$	⋯	${e_{A_i}}^2$	⋯	${e_{A_{n_a}}}^2$	$\sum_{i=1}^{n_a} {e_{A_i}}^2$	$\frac{SS_a}{n_a}$
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮
B_j	${e_{A_1}}^2$	⋯	${e_{A_i}}^2$	⋯	${e_{A_{n_a}}}^2$	$\sum_{i=1}^{n_a} {e_{A_i}}^2$	$\frac{SS_a}{n_a}$
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮
B_n	${e_{A_1}}^2$	⋯	${e_{A_i}}^2$	⋯	${e_{A_{n_a}}}^2$	$\sum_{i=1}^{n_a} {e_{A_i}}^2$	$\frac{SS_a}{n_a}$
合計	$n_b{e_{A_1}}^2$	⋯	$n_b{e_{A_i}}^2$		$n_b{e_{A_{n_a}}}^2$	$n_b\sum_{i=1}^{n_a} {e_{A_i}}^2$	$n_b\frac{SS_a}{n_a}$
平均	${e_{A_1}}^2$	⋯	${e_{A_i}}^2$	⋯	${e_{A_{n_a}}}^2$	SS_a	$\frac{SS_a}{n_a}$

合計の列の平均値の行SS_aは、SS_{A+B}の期待値を求める計算過程で出てくるものですが、実は、これは計算の途中です。とりあえず、$SS_{A+\bar{B}}$と表して

$$SS_{A+\bar{B}} = SS_a$$

としておきますこれも、総当たりで計算ですが、内容は、B_jを固定して、Aの変動のだけを見てるので、総当たり的に考えれば、A_iを固定して、Bの変動も考えなくてはならないでしょう。列方向に平均をとる計算をしてみます。

表 6. SS_{A+B}を求める計算（列ごとに平均する計算）

	A_1	⋯	A_i	⋯	A_{n_a}	合計	平均
B_1	${e_{B_1}}^2$	⋯	${e_{B_1}}^2$	⋯	${e_{B_1}}^2$	$n_a{e_{B_1}}^2$	${e_{B_1}}^2$
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮
B_j	${e_{B_j}}^2$	⋯	${e_{B_j}}^2$	⋯	${e_{B_j}}^2$	$n_a{e_{B_j}}^2$	${e_{B_j}}^2$
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮
B_n	${e_{B_{n_b}}}^2$	⋯	${e_{B_{n_b}}}^2$	⋯	${e_{B_{n_b}}}^2$	$n_a{e_{B_j}}^2$	${e_{B_j}}^2$
合計	$\sum_{j=1}^{n_b} {e_{B_j}}^2$	⋯	$\sum_{j=1}^{n_b} {e_{B_j}}^2$		$\sum_{j=1}^{n_b} {e_{B_j}}^2$	$n_a\sum_{j=1}^{n_b} {e_{B_j}}^2$	SS_b
平均	$\frac{SS_b}{n_b}$	⋯	$\frac{SS_b}{n_b}$	⋯	$\frac{SS_b}{n_b}$	$n_a\frac{SS_b}{n_b}$	$\frac{SS_b}{n_b}$

結果は同じで、同様の記法を使うと、

$$SS_{\bar{A}+B} = SS_b$$

$SS_{A+\bar{B}}$と$SS_{\bar{A}+B}$を足し合わせれば、網羅的にSS_{A+B}を推測したことになりますから、

$$SS_{A+B} = SS_{A+\bar{B}} + SS_{\bar{A}+B} = SS_a + SS_b \qquad 式\ 8$$

となります。この式をさらに変形すると、期待値としてSS_{A+B}を求めるとき、平均操作を

２回行っているから、$A+B$の自由度はn_a+n_b-2で、

${\sigma^2}_{A+B}=\frac{SS_{A+B}}{n_a+n_b-2}$, ${\sigma^2}_A=\frac{SS_A}{n_a-1}$, ${\sigma^2}_B=\frac{SS_B}{n_b-1}$　だから式式あるいは代入して

$$(n_a+n_b-2){\sigma^2}_{A+B}={\sigma^2}_A(n_a-1)+{\sigma^2}_B(n_b-1)$$

$${\sigma^2}_{A+B}=\frac{{\sigma^2}_A(n_a-1)+{\sigma^2}_B(n_b-1)}{(n_a+n_b-2)} \qquad 式9$$

となります。式９をよく見ると、A の普遍分散と B の普遍分散のそれぞれに自由度
n_a-1、n_b-1で重みを付けた重み付き平均になっていることがわかります。
「等分散性${\sigma^2}_A={\sigma^2}_B$の仮定を受け入れてしまって、そうあるべきだが、実際にデータとして、得られる２つの分散は等しくないから、 データ数の違いを考慮して、自由度で重みを付けて重み付き平均をとる。」と覚えても良さそうです。
結論を要約すると

$$SS_{total}=n_bSS_a+n_aSS_b$$

$$SS_{A+B}=SS_A+SS_B$$

差の分散
差の分散の議論は和の分散とは違って実用的な意味があります。同じ鉢に植えて、同一の環境で育てた、種の違う２つの植物の成長に差があるかという検討は、対になったデータの t 検定で議論出来ますが、同じ植物に異なる肥料を与えるという実験は、一つの鉢ではできないでしょう。異なる鉢で肥料を変えて実験したときに、肥料の違いによって生長に差があるかというのは意味のある議論です。対になっていない t 検定がそれにあたります。
和の分散についての議論のロジックを応用して差の分散について考えます
差の分散
差の分散の議論は和の分散とは違って実用的な意味があります。同じ鉢に植えて、同一の環境で育てた、種の違う２つの植物の成長に差があるかという検討は、対になったデータの t 検定で議論出来ますが、同じ植物に異なる肥料を与えるという実験は、一つの鉢ではできないでしょう。た時に、成長に差があるか意味のある議論です。対になっていない t 検定がそれにあたります。和の分散についての議論のロジックを応用して差の分散について考えます。表 3 の形式で、総当たり的なグループ間のデータの差の平均値を求めてみます。和の場合と異なって、四則演算の交換法則は、差や割り算の場合成立ちませんから、A から B を引くという形で考えます。

表 7.表 3 の形式で偏差の差の平均を求める計算

	A_1	⋯	A_i	⋯	A_{n_a}	合計	平均
B_1	$e_{A_1}-e_{B_1}$	⋯	$e_{A_i}-e_{B_1}$	⋯	$e_{A_{n_a}}-e_{B_1}$	$0^*-n_a e_{B_1}$	$-e_{B_1}$
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮
B_j	$e_{A_1}-e_{B_j}$	⋯	$e_{A_i}-e_{B_j}$	⋯	$e_{A_{n_a}}-e_{B_j}$	$0^*-n_a e_{B_j}$	$-e_{B_j}$
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮
B_{n_B}	$e_{A_1}-e_{B_{n_b}}$	⋯	$e_{A_i}-e_{B_{n_B}}$	⋯	$e_{A_{n_a}}-e_{B_{n_B}}$	$0^*-n_a e_{B_{n_B}}$	$-e_{B_{n_B}}$
合計	$n_b e_{A_1}-0^{**}$	⋯	$n_b e_{A_i}-0^{**}$	⋯	$n_b e_{A_{n_a}}-0^{**}$	$0-0^{**}$	0
平均	e_{A_1}	⋯	e_{A_i}	⋯	$e_{A_{n_a}}$	0	0

この表に示したように、B_j行の平均は、$-e_{B_j}$です。この平均値を$e_{A_i}-e_{B_j}$から差し引くのだから、平均値からの偏差は、$e_{A_i}-e_{B_j}-\left(-e_{B_j}\right)=e_{A_i}$です。これは A＋B の時と同じです。

ですから、

$$SS_{A-\bar{B}}=SS_A$$

となります。これA_i列についても同様でA_i行の平均はe_{A_i}で、偏差は$e_{A_i}-e_{B_j}-e_{A_i}=-e_{B_j}$、その平方は$\left(-e_{B_j}\right)^2={e_{B_j}}^2$となるので、

$$SS_{\bar{A}-B}=SS_B$$

となります。

ちなみに、SS_{total}についても、

$$SS_{total:A+B}=\sum_{i=1}^{n_a}\sum_{j=1}^{n_b}\left(e_{A_i}+e_{B_j}\right)^2=\sum_{i=1}^{n_a}\sum_{j=1}^{n_b}{e_{A_i}}^2+2\sum_{i=1}^{n_a}\sum_{j=1}^{n_b}e_{A_i}e_{B_j}+\sum_{i=1}^{n_a}\sum_{j=1}^{n_b}{e_{B_j}}^2$$

$$=n_bSS_A+n_aSS_A$$

$$\because\sum_{i=1}^{n_a}\sum_{j=1}^{n_b}e_{A_i}e_{B_j}=0$$

$$SS_{total:A-B}=\sum_{i=1}^{n_a}\sum_{j=1}^{n_b}\left(e_{A_i}-e_{B_j}\right)^2=\sum_{i=1}^{n_a}\sum_{j=1}^{n_b}{e_{A_i}}^2-2\sum_{i=1}^{n_a}\sum_{j=1}^{n_b}e_{A_i}e_{B_j}+\sum_{i=1}^{n_a}\sum_{j=1}^{n_b}{e_{B_j}}^2$$

$$=n_bSS_A+n_aSS_A$$

$$\because\sum_{i=1}^{n_a}\sum_{j=1}^{n_b}e_{A_i}e_{B_j}=0$$

となります。ですから

$$SS_{total:A-B}=SS_{total:A+B}=n_bSS_A+n_aSS_A$$

$$SS_{A-B} = SS_A + SS_B$$

$$\sigma^2{}_{A-B} = \sigma^2{}_{A+B} = \frac{(n_a - 1)\sigma^2{}_A + (n_b - 1)\sigma^2{}_B}{n_a + n_b - 2}$$

「差の分散は和の分散に等しい。」と覚えます。
結論を要約すると

$$SS_{total:A-B} = SS_{total:A+B} = n_b SS_A + n_a SS_A$$

$$SS_{A-B} = SS_{A+B} = SS_A + SS_B$$

$$\sigma^2{}_{A-B} = \sigma^2{}_{A+B} = \frac{(n_a - 1)\sigma^2{}_A + (n_b - 1)\sigma^2{}_B}{n_a + n_b - 2}$$

t 検定の話に戻ります。

対になっていないサンプル群間の差の検定で、*差の母集団の標準偏差*ϕと推定の用いたデータのサンプルサイズνをどうするかという問題でした。分散ϕの方は、差の分散の検討結果をそのまま応用して、

$$\phi^2 = \sigma^2{}_{A-B} = \frac{(n_a - 1)\sigma^2{}_A + (n_b - 1)\sigma^2{}_B}{n_a + n_b - 2} \qquad 式10$$

$$\phi = \sigma_{A-B} = \sqrt{\frac{(n_a - 1)\sigma^2{}_A + (n_b - 1)\sigma^2{}_B}{n_a + n_b - 2}}$$

というのがϕに関する結論です。「自由度で重みづけした重み付け平均」というのは言葉としては覚えやすいのですが、式の形は少し複雑で、覚えにくいかもしれません。また、先にAとBの分散を計算しておくのも、ひと手間多い感じです。そこで、

$$SS_{A-B} = SS_A + SS_B$$

つまり、「差の平方和SS_{A-B}は二つのサンプル集団の平方和SS_AとSS_Bの和」だと覚えて、それを「自由度$n_a + n_b - 2$(平均操作が２回は行っているからサンプルサイズ－２)で割って、差の分散を求める。」とした方が、式の形が単純で覚えやすいし、計算するときの手間を２ステップ省略できます。昔、集計用紙を使って表計算して、検定を行っていたころは、計算間違いや、計算誤差を小さくするために、こういうテクニックが必要でしたが、今はそんな必要はないかもしれません。今時、計算はコンピュータがするのだから、式１０だけ覚えておけばよいというのも、一つの考え方でしょう。

用いたデータのサンプルサイズνについて考えます。先回りしますが、

$$\nu = \frac{n_a n_b}{n_a + n_b}$$

これをきちんと説明している教科書を知りません。ネットで探してみましたが、見つかりませんでした。中には、複雑で難しい計算になるので、結果だけ覚えておけば良いという解説もありました。結果だけ見ると大して難しい式ではありません。計算が難しいのではありません。考え方をわかりやすく説明できないのだと思います。私も私なりに考えてみました。

$$式5 \quad \frac{A_{average} - B_{average}}{\frac{\phi}{\sqrt{\nu}}} = t$$

の分母の$\frac{\phi}{\sqrt{\nu}}$というのは、標準誤差のことですね。

$$SE_A = \frac{{\sigma^2}_A}{n_a}$$

$$SE_B = \frac{{\sigma^2}_B}{n_a}$$

$$SE_{A-B} = \frac{{\sigma^2}_{A-B}}{\nu}$$

ところで、標準誤差というのも期待値の周辺の観測値の２次の積率で、一種の分散です。分散は一般に加法的ではありませんが、AとBに共分散（相関）がなければ、加法的で下記のようになります注。

$$\frac{{\sigma^2}_{A-B}}{\nu} = \frac{{\sigma^2}_A}{n_a} + \frac{{\sigma^2}_B}{n_b}$$

ところで、表５，表６の平均の列の平均の行に$\frac{SS_a}{n_a}$と$\frac{SS_b}{n_b}$というのがあります。この計算から何かの母集団の分散を推測しているのではないから

$$\frac{SS_a}{n_a} = {\sigma^2}_A, \quad \frac{SS_b}{n_b} = {\sigma^2}_B$$

なのですが、そもそも、これらは、${\sigma^2}_{A-B}$の期待値を B を固定して行方向に計算した値${\sigma^2}_{A-\bar{B}_j}$と、列方向に A を固定して縦方向に計算した値${\sigma^2}_{\bar{A}-B}$をそれぞれ縦方向、横方向に足して、平均化して、${\sigma^2}_{A-\bar{B}_j}$からσ_{A-B}を、${\sigma^2}_{\bar{A}-B}$からσ_{A-B}求めたもので、どちらも、σ_{A-B}と書けます。ただし、サンプルサイズが違っていますから、それぞれから求めた標準誤差は、$\frac{{\sigma^2}_{A-B}}{n_a}$、$\frac{{\sigma^2}_{A-B}}{n_b}$になります。標準誤差に加法性（加算性？）があるのならば、

$$\frac{{\sigma^2}_{A-B}}{\nu} = \frac{{\sigma^2}_A}{n_a} + \frac{{\sigma^2}_B}{n_b} = \frac{{\sigma^2}_{A-B}}{n_a} + \frac{{\sigma^2}_{A-B}}{n_b} = \left(\frac{1}{n_a} + \frac{1}{n_b}\right){\sigma^2}_{A-B}$$

$$\frac{1}{\nu}=\frac{1}{n_a}+\frac{1}{n_b}$$

$$\nu=\frac{n_a n_b}{n_a+n_b}$$

となります。何か、ごちゃごちゃした説明で、途中に言い訳がたくさん入って、面倒な説明になってしまいました。もっと、原理的に考えると、

$$\nu=\frac{n_a n_b}{n_a+n_b}$$

というのは、加法性があると考えた時のサンプルサイズと「総当たり的」に考えた時のサンプルサイズの比です。もともと、サンプルサイズとは、サンプル集団の大きさをサンプルが１しかない時の、サンプルサイズを１として表した、大きさ（広がり）の比です。加法的に考えたものを「総当たり的」の考えたものの広げているのだから、その比で割り算すべきなのだと考えて、

$$\nu=\frac{n_a n_b}{n_a+n_b}$$

となると理解した方が、シンプルで分かりやすいかもしれません。あるいは、

$$\sigma_{A-B}=\sqrt{\frac{(n_a-1){\sigma^2}_A+(n_b-1){\sigma^2}_B}{n_a+n_b-2}}$$

という、自由度の重み付き平均の式から計算値を与えて。これは、「総当たり的」に求めた分散から来ているので、加法的に求めた分散にするために、加法的なサンプルサイズと「総当たり的」なサンプルサイズの比で、加法的（加算的）な分散を割って、

$$SE=\sigma_{A-B}\sqrt{\frac{n_a+n_b}{n_a n_b}}=\sigma_{A-B}\sqrt{\frac{1}{n_a}+\frac{1}{n_b}}$$

とするのだ、という説明の方が良いかもしれません。つまり、対になっていない場合は総当たり的になっているのだけれど、これを、加法的（加算的にするためにサンプルサイズの比率で割っているという理解です。

最終的な結論としてtの書き方はいろいろありそうですが

$$t=\frac{M_a-M_b}{\frac{\sigma_{A-B}}{\sqrt{\nu}}}\quad 式\ 1\ 2$$

$${\sigma^2}_{A-B}=\frac{(n_a-1){\sigma^2}_A+(n_b-1){\sigma^2}_B}{n_a+n_b-2}$$

$$\nu=\frac{n_a n_b}{n_a+n_b}$$

と書いておくか

$$\sigma_{A-B} = \sqrt{\frac{(n_a-1){\sigma^2}_A + (n_b-1){\sigma^2}_B}{n_a+n_b-2}}$$

$$\frac{1}{\sqrt{\nu}} = \sqrt{\frac{1}{n_a}+\frac{1}{n_b}}$$

と計算しておいて、

$$t = \frac{M_a - M_b}{\sqrt{\frac{(n_a-1){\sigma^2}_A + (n_b-1){\sigma^2}_B}{n_a+n_b-2}}\sqrt{\frac{1}{n_a}+\frac{1}{n_b}}}$$

これを t の定義式とするのもあるかもしれないが、意味が解りません。意味が分かるという点では、式１２の方が良いでしょう。教科書によっては

$$t = \frac{M_a - M_b}{\sigma_{A-B}\sqrt{\frac{1}{n_a}+\frac{1}{n_b}}}$$

と書いてあります。比較的シンプルで意味も分かりやすい記述ですが、$\sqrt{\frac{1}{n_a}+\frac{1}{n_b}}$が何かと問われると、とっさに答えられません。

式を覚えるよりは具体的なデータを使って計算手順を覚えた方が良いかもしれません。表２のデータを使って、実際に計算してみます。

表８．t 検定のためのエクセルシート

	A	A^2	B	B^2
	1	1	1	1
	5	25	5	25
	6	36	6	36
		0	8	64
合計	12	62	20	126
平均	4		5	
データ数	3		4	
SS	14		26	

A：データー数$n_a = 3$、$df_a = 2$ 平均$M_A = 4$ 平方和$SS_A = 14$ 分散$\sigma^2{}_A = \frac{14}{2} = 7$
B：データー数$n_b = 4$、$df_b = 3$ 平均$M_B = 5$ 平方和$SS_B = 26$ 分散$\sigma^2{}_B = \frac{26}{2} = 8.667$
全データ数$n_{total} = n_a + n_a = 7 \quad df_{total} = n_{total} - 1 = 6 \quad df_{A-B} = n_{total} - 2 = 5$

$SS_{A-B} = SS_A + SS_B = 14 + 26 = 40$ 分散$\sigma^2{}_{A-B} = \frac{SS_{A-B}}{df_{A-B}} = \frac{40}{5} = 8$ $\sigma_{A-B} = \sqrt{8} = 2.8284$

サンプルサイズ$\nu_{A-B} = \frac{n_a n_b}{n_a + n_b} = \frac{3\times4}{3+4} = 1.714286$ $\sqrt{\frac{1}{\nu_{A-B}}} = \sqrt{\frac{1}{n_a} + \frac{1}{n_b}} = 0.763763$

$$|M_A - M_B| = 1$$

$$t = \frac{|M_A - M_B|}{\sigma_{A-B}\sqrt{\frac{1}{n_a} + \frac{1}{n_b}}} = \frac{1}{2.8284 \times 0.763763} = 0.46291$$

t 表(両側)によれば$p = 0.95$（5%の危険率）での t の臨海値は 2.571 ですから、A の母集団と B の母集団の平均値が異なっていると結論することはできません。

t 検定には、対になっている t 検定、対になっていない t 検定以外にも、いくつかのタイプがあります。普通、 t 検定は、検定しようとする 2 つのサンプル集団で分散が等しいはずだという前提に立っています。等分散性は分散比＝ 1 を確かめればよいから、F 検定をして確かめれば良いのですが、帰無仮説が否定されれば、等分散性がないと言えますが、帰無仮説が否定されなかったからと言って、等分散性があるとは言い切れないでしょう。むしろ、例えば、割合のデータなどは、原理的に等分散性がありません。だから、原理的に等分散性があるとすべきか、ないとすべきか考えるべきです。等分散性がない場合には、Welch の t 検定というのを使います。また、何らかの情報であらかじめ、母集団の平均と分散がわかっている場合のそれらを与えて、行う検定です。例えば、全国平均がわかっているとき、自分たちの集団がその中でどのくらいに位置付けられるかというようなの検討に使います。偏差値みたいな考え方ですね。私は、 1 度だけ Welch の t 検定をやったこと

がありますが、複雑な式で面倒です。Z 検定やったことがありません。社会学的な分野では使うことがあるかもしれませんが、あまり難しくないから、必要になったら調べて使えばよいでしょう。確かエクセルの分析ツールにも入っていたと思います。

注:　これは和の分散、差の分散のところでやりましたが、もう少し、体系的に理解しておいた方が良いかもしれません。この話は、期待値の加法性という話に帰着します。連続した数値の場合を考えると、積分の形で表した方が良いかもしれませんが、わかりやすさを重視して、離散型で書きます。

$$X = \{x_1, \cdots, x_{n_x}\}$$

$$Y = \{y_1, \cdots, y_{n_y}\}$$

というデータがあったとします。たとえば、サイコロのように 1 ～ 6 の整数がある確率で出てくるデータです。
Xの期待値は

$$E(X) = \sum_{i=1}^{n_a} x_i Pr(x_i)$$

$Pr(x_i)$はx_iとなる確率です。サイコロならば、$Pr(x_i) = \frac{1}{6}$ですが、一般的には確率が全部等しいということはないでしょう。サイコロをいくつか振った合計ということならば、二項分布の確率になるでしょう。Yについても

$$E(Y) = \sum_{i=j}^{n_b} y_j Pr(y_j)$$

となります。$E(X+Y)$について考えると

$$E(X+Y) = \sum_{i=1}^{n_a}\sum_{j=1}^{n_b}(x_i + y_j)\, Pr(x_i, y_j)$$

$Pr(x_i, y_j)$はx_i, y_jの同時確率だから、

$$Pr(x_i, y_j) = Pr(x_i,)Pr(y_j)$$

$$E(X+Y) = \sum_{i=1}^{n_a}\sum_{j=1}^{n_b}(x_i + y_j)Pr(x_i,)Pr(y_j)$$

$$= \sum_{i=1}^{n_a}\sum_{j=1}^{n_b} x_i Pr(x_i,)Pr(y_j) + \sum_{i=1}^{n_a}\sum_{j=1}^{n_b} y_j Pr(x_i,)Pr(y_j)$$

第 1 項についてシグマ記号の順番を入れ替えてx_iを含まない$Pr(y_j)$をその外に出すと、

$$\sum_{i=1}^{n_a}\sum_{j=1}^{n_b} x_i Pr(x_i,)Pr(y_j) = \sum_{j=1}^{n_b}\sum_{i=1}^{n_a} x_i Pr(x_i,)Pr(y_j) = \sum_{j=1}^{n_b} Pr(y_j)\sum_{i=1}^{n_a} x_i Pr(x_i,)$$

$\sum_{i=1}^{n_a} x_i Pr(x_i,) = E(X)$だから

$$= \sum_{j=1}^{n_b} Pr(y_j)E(X) = E(X)\sum_{j=1}^{n_b} Pr(y_j) = E(X)$$

$\because \sum_{j=1}^{n_b} Pr(y_j) = 1$（確率の総和は１）
第二項についても同様で、

$$\sum_{i=1}^{n_a}\sum_{j=1}^{n_b} y_j Pr(x_i,)Pr(y_j) = E(Y)$$

したがって、

$$E(X+Y) = E(X) + E(Y)$$
$$E(X-Y) = E(X) + E(-Y) = E(X) - E(Y)$$

和の平均値は平均値の和、差の平均値は平均値の差というのを面倒くさく言うとこうなります。
分散については、

$$V(X) = \sum_{i=1}^{n_x}(x_i - M_X)^2 Pr(x_i) = \sum_{i=1}^{n_x} {x_i}^2 Pr(x_i) - 2M_X\sum_{i=1}^{n_x} x_i Pr(x_i) + {M_X}^2\sum_{i=1}^{n_x} Pr(x_i)$$
$$= E(X^2) - 2E(X)^2 + E(X)^2 = E(X^2) - E(X)^2$$

$$V(Y) = E(Y^2) - E(Y)^2$$

$$V(X+Y) = \sum_{j=1}^{n_y}\sum_{i=1}^{n_x}\Big(x_i + y_j - \big(E(X)+E(Y)\big)\Big)^2 Pr(x_i, y_j)$$

$$= \sum_{j=1}^{n_y}\sum_{i=1}^{n_x}\big(x_i - E(X)\big)^2 Pr(x_i, y_j) + \sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\Big(y_j - E(Y)\Big)^2 Pr(x_i, y_j)$$

$$+2\sum_{j=1}^{n_y}\sum_{i=1}^{n_x}\big(x_i - E(X)\big)\Big(y_j - E(Y)\Big)Pr(x_i, y_j)$$

第１項

$$\sum_{j=1}^{n_y}\sum_{i=1}^{n_x}\big(x_i - E(X)\big)^2 Pr(x_i, y_j) = \sum_{j=1}^{n_y}\sum_{i=1}^{n_x}\big(x_i - E(X)\big)^2 Pr(x_i)Pr(y_j)$$

$$= \sum_{j=1}^{n_y} Pr(y_j) \sum_{i=1}^{n_x} \left(x_i - E(X)\right)^2 Pr(x_i)$$

$$= \sum_{j=1}^{n_y} Pr(y_j) V(X) = V(X) \sum_{j=1}^{n_y} Pr(y_j) = V(X^2)$$

同様に、第２項についても

$$\sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \left(y_j - E(Y)\right)^2 Pr(x_i, y_j) = \ V(Y)$$

第三項は、積の期待値で共分散（Cov）

$$2 \sum_{j=1}^{n_y} \sum_{i=1}^{n_x} \left(x_i - E(X)\right) \left(y_j - E(Y)\right) Pr(x_i, y_j) = 2Cov(X, Y)$$

だから、

$$V(X + Y) = V(X) + \ V(Y) + 2Cov(X, Y)$$
$$V(X - Y) = V(X) + \ V(Y) - 2Cov(X, Y)$$

したがって、一般には

$$V(X + Y) \neq V(X) + \ V(Y)$$
$$V(X - Y) \neq V(X) + \ V(Y)$$

だが、共分散$Cov(X, Y) = 0$,つまりXとYが独立（相関がない）ならば

$$V(X + Y) = V(X) + \ V(Y)$$
$$V(X - Y) = V(X) + \ V(Y)$$

始めに、これを与えて、

$$\sigma^2{}_{A-B} = \sigma^2{}_A + \sigma^2{}_B$$
$$SE_{A-B} = SE_A + SE_B$$

を前提に話をした方が手っ取り早い。しかし、それだと、作業内容がイメージしにくいでしょう。

F 検定

データの構造と取り扱い

分散の分離

様々な要因が関係して現実が出来上がります。科学では、それらの要因を切り分けて、その影響を論じます。統計学の作業で重要なことは、データの分散（バラつき）をその原因（要因）ごとに切り分けることです。それが出来なければ、因果関係の立証できないでしょう。一つの要因と他の要因が原因と結果のような関係性を持 っているという場合もあって、そんな場合には要因同士の積の分散（共分散）もありますが、それは相関分析になるので後述します。 ここではまず、お互いに関係を持たない、独立した原因がある場合を考えます。

F 検定は、分散の比（F 比）の確率を論ずる分析で、１要因分散分析、２要因分散分析、繰り返しのある２要因分散分析、３要因分散分析等々、一般に「多要因分散分析」と呼ばれる様々な構造のデータ・セットの分析に使われますが、そのためには、データの構造を考えて、データセットから、様々な要因の分散を取り出さなくてはなりません。エクセルの「データ」のところにある「データ分析」には、「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」があり、「F 検定：２標本を使った分散の検定」まであります。さらにＺ検定も含めて、様々なｔ検定もあります。探せば、Ｒにも Pyson にも様々な形で使えるものがあると思います。ですから、今では、計算方法を覚えても意味がないかもしれません。しかし、計算方法がブラックボックスでも、どんな要因を取り出して比較しているのかを理解していることは、正しく分析方法の選択したり、結果を解釈して正しく議論を進めていくうえで重要でしょう。そこで、データがどんな構造をしていて、どのよう要因別に分散を取り出すのかを説明します。

様々な説明の仕方がありますが、ここでは、解析が前提としているモデル（論理構造）を鮮明にするために、いくつかの要因によって出来ているデータセットを、自分たちで作って みることにします。例としては、和の分散で取り上げたデータの組み合わせの例（表２）を使います。要因 A による以下のデータがあるとします。 サンプル集団Aのデータ 1、5、６のデータをA_iと表すことにすると$A_1 = 1, A_2 = 5, A_3 = 6$と表すということです。たとえば、三つの植木鉢に、それぞれ窒素量の違う肥料を入れて、それぞれの鉢に植えた 草の丈が、１ｃｍ、５ｃｍ６ｃｍだったというようなことを考えてください

この情報を整理すると

サンプルサイズ（データ数）$n_a = 3,$

自由度 $df_a = n_a - 1 = 3 - 1 = 2$

平均値 $M_A = \frac{\text{データーの総和}}{\text{サンプルサイズ}} = \frac{Sum_A}{n_a} = \frac{\sum_{i=1}^{n_a} A_i}{n_a} = \frac{1+5+6}{3} = 4$

平方和 $SS_A = \sum_{i=1}^{n_a}(A_i - M_A)^2 = (1-4)^2 + (5-4)^2 + (6-4)^2 = 14$

不偏分散　$\sigma^2{}_A = \frac{SS_A}{df_a} = \frac{14}{2} = 7$

標準偏差 σ_A=$\sqrt{\sigma^2{}_A} = \sqrt{7}$

というデータがあるとき、そのデータが正規分布しているという前提があれば、それらのデー タ群は、平均値と分散で代表させることができます。次に別に要因 B によるデータがあって、サンプル集団Aと同様に、$B_1 = 1, B_2 = 5, B_3 = 6, B_4 = 8$とします。たとえば、4 つの植木鉢があって、そこに入れる肥料のリンの量を 4 段階に変えて、草丈を比較したというようなことを考えてください。

このデータを代表する統計量は

サンプルサイズ（データ数）$n_b = 4$

自由度 $df_b = n_b - 1 = 4 - 1 = 3$

平均値 $M_B = \frac{\text{データーの総和}}{\text{サンプルサイズ}} = \frac{Sum_B}{n_b} = \frac{\sum_{j=1}^{n_b} B_j}{n_b} = \frac{1+5+6+8}{4} = 5$

平方和 $SS_B = \sum_{j=1}^{n_b}\left(B_j - M_B\right)^2 = (1-5)^2 + (5-5)^2 + (6-5)^2 + (8-5)^2 = 26$

不偏分散　$\sigma^2{}_B = \frac{SS_B}{df_b} = \frac{26}{3} = 8.667$

標準偏差 σ_B=$\sqrt{\sigma^2{}_B} = \sqrt{8.667} = 2.944$

この 2 つの要因によるデータの広がり方（バラつき）を比較するために、その分散の比を計算します。 これが F 比です。

$$F = \frac{\sigma^2{}_A}{\sigma^2{}_B} = \frac{7}{8.667} = 0.808$$

この値を見ると、A の要因によるデータの広がりは、B の要因による広がりよりも少し小さいけれど、その違いはあまりないなと、感じるでしょう。多分、人は直感的にこういう比 較をしています。 例えば、データの広がりが、B の 10 分の 1 になっている、次のようなデータ群 C を考えます。

データ

0.1、0.5、0. 6 、0.8

個々のデータをC_jと表すことにすると

$$C_1 = 0.1, C_2 = 0.5, C_3 = 0.6, C_4 = 0.8$$

たとえば、4 つの植木鉢に、異なる人の写真を張り付けて、その人の顔が植木鉢の中の草の 成長に与える影響を比較するというようなことを考えてください。そんなことが成長に影響するはずがないと思うかもしれませんが、決めつけてはいけません。世の中何があるか わかりません。 このデータの統計量は以下の通りです。

サンプルサイズ$n_c = 4$

自由度 $df_C = 4 - 1 = 3$
平均値 $M_C = \frac{0.1+0.5+0.6+0.8}{4}$=0.5
平方和 SS_C=$(0.1 - 0.5)^2 + (0.5 - 0.5)^2 + (0.6 - 0.5)^2 + (0.8 - 0.5)^2 = 0.26$
不偏分散 $= \frac{0.26}{3} = 0.08667$
標準偏差 $\sigma_C = \sqrt{0.08667} = 0.09310$
これも A の統計量と並べて見比べてみましょう。

F 比 $F = \frac{\sigma^2{}_A}{\sigma^2{}_C} = \frac{7}{0.08667}$80.77

A の要因によるデータは、C のデータの１００倍近い広がりもって分布しているのです。A の要因と C の要因の広がりの大きさが、同じだという人はいないでしょう。もちろん、これは 感覚的なものですが、A と C のデータの広がりが同じであるにもかかわらず、たまたま、 母集団から取り出したデータの取り出しかたによって、この様な比になることが、どのくらいの確率で起こるのかを計算しておけば数量的な根拠をもってそのような判断が出来る でしょう。たとえば、5%以下の確率でしか起こらないという境界になる F 比の値を数学的モデルから計算しておき、実際のデータの値と数学的なモデルから計算した値と大きさを比較すれば、実際のデータから計算した値が、データの取り出し方によってたまたま出た値であるかどうか判断することができます。この F 比のことを分散比とよび、こういう分析の 仕方を分散分析（F 検定）といいます。
この議論ができたのは、要因ごとのデータの統計量を私たちが知っていたからです。実際のデータはこういう形では与えられません。様々な要因が加わって出来た結果だけをデータとして与えられます。そこで、２つの要因が重なって出来た（相加的な要因という言い方が良いかもしれません。）データの値がどのようになるかを考えてみます。要因 A と要因 B が相加的に関与してできるデータで、要因同士は独立しています。この場合、A の特定の レベルと B の特定のレベルだけを選び出してたしあわせることはできませんから、すべて のデータを総当りでたしあわせることにします。和の分散のところでやったことと同じです。同じデータを使います。
表９. 二つの要因によるデータの総当たりの和

		A				
		1	5	6	合計	平均
B	1	1+1=2	5+1=6	6+1=7	15	5
	5	1+5=6	5+5=10	6+5=11	27	9
	6	1+6=7	5+6=11	6+6=12	30	10
	8	1+8=9	5+8=13	6+8=14	36	12
	合計	24	40	44	108	36
	平均	6	10	11	27	9

青字がAの要因で決まる部分、赤字が B の要因で決まる部分です。

$$A\text{の平均}M_A = 4,\ \text{平方和}SS_A = 14, \text{分散}{\sigma^2}_A = 7$$

$$B\text{の平均}M_B = 5,\ \text{平方和}SS_B = 26, \text{分散}{\sigma^2}_B = 8.667$$

でした。全体の平均は 9 だから$M_{total} = M_A + M_B$データをたし合わせているのだから、 その平均値も平均値の和になるというのは当然です。

この表で示された、このデータが 2 つのデータを加算的に積み重ねて作ったものだと知らない人は、表中の 2,6,7.6,10,11,7,11,12,9.13.14 というデータから、 すべての合計を

$$Sum_{total} = 2 + 6 + 7 + 6 + 10 + 11 + 7 + 11 + 12 + 9 + 13 + 14 = 108$$

全体の平方和を

$$SS_{total} = (2-9)^2 + (6-9)^2 + (7-9)^2 + (6-9)^2 + (10-9)^2 + (11-9)^2$$
$$+(7-9)^2 + (11-9)^2 + (12-9)^2 + (9-9)^2 + (13-9)^2 + (14-9)^2$$
$$49+9+4+9+1+4+4+4+9+0+16+25 = 134$$

この全体の自由度は、

$$df_{total} = (3\times 4) - 1 = 11$$

だから、 全体の分散は、

$${\sigma^2}_{total} = \frac{SS_{total}}{df_{total}} = \frac{134}{11} = 12.18182$$

と計算するでしょう。

この計算に違和感を感じて、何やっているんだと思う人もいるでしょう。すでに

$$SS_{A+B} = SS_A + SS_B$$

だと知っていますから、何しているんだろうという感じです。それはそれとして、この計算は正直で丁寧な計算ですが少し面倒ですね

個々のデータの平均値からの距離の 2 乗の値を、表にしてみます。

表 10. 全 SS の計算

		A			
		1	5	6	合計
B	1	49	9	4	62
	5	9	1	4	14
	6	4	4	9	17
	8	0	16	25	41
	合計	62	30	42	134

こんな風になっていて、この表の、行ごとに計算したものを、縦にたしあわせるか、列ごとに計算したものを横にたしあわせるかで、全平方和の 134 という値が求まります。 とこ

ろで、緑色で書いた 49 という数値の由来を考えると、$(1+1-9)=49$ という計算です。さらに元をたどれば、$\{(1-4)+(1-5)\}^2$ と言う計算 です。 このように考えると、 1 行目の各列をたしあわせた総和の計算は

$$\{(1-4)+(1-5)\}^2+\{(5-4)+(1-5)\}^2+\{(6-4)+(1-5)\}^2=62$$

と言う計算でもあります(4 は A の平均値で、 5 は B の平均値ですから、それぞれに要来するという意味で、青字、赤字にしました)。この式の外側のカッコを外して展開してみましょう。

$$\{(1-4)+(1-5)\}^2+\{(5-4)+(1-5)\}^2+\{(6-4)+(1-5)\}^2$$
$$=(1-4)^2+2(1-4)(1-5)+(1-5)^2+(5-4)^2+(5-4)(1-5)+(1-5)^2$$
$$+(6-4)^2+2(6-4)(1-5)+(1-5)^2$$
$$=(1-4)^2+(5-4)^2+(6-4)^2+2\{(1-4)+(5-4)+(6-4)\}+3(1-5)^2$$

$=(1-4)^2+(5-4)^2+(6-4)^2+2(1-5)\{(1-4)+(5-4)+(6-4)\}+3(1-5)^2$

です。ところで、

$(1-4)^2+(5-4)^2+(6-4)^2=SS_A=14$

$\{(1-4)+(5-4)+(6-4)\}=0$ (平均値からの距離の総和は 0)

$3(1-5)^2$ はB_1の平均値からの距離の 2 乗のn_a倍: $n_a(B_1-M_B)=3\times16=48$

です。

ですから、列合計は

$$SS_A+n_a(B_1-M_B)^2=62$$

です。

2 列目についても同様で、

$$SS_A+n_a(B_2-M_B)^2=14+3(5-5)^2=14$$

3 列目についても同様で

$$SS_A+n_a(B_3-M_B)^2=14+3(6-5)^2=17$$

4 列目についても同様で

$$SS_A+n_a(B_4-M_B)^2=14+3(8-5)^2=41$$

列を合計すると

$$SS_{total}=n_bSS_A+n_a\sum_{j=1}^{n_b}(B_j-M_B)^2=n_bSS_A+n_aSS_A=62+14+17+41=134$$

となります。これ自体は、和の分散のところで文字式で計算した結果に数値を与えて計算しただけのことです。 2 つの要因で説明されるデータで、それぞれの要因の組み合わせの中に繰り返しがない時の分析を、繰り返しのない 2 要因分散分析言います。何かの要因と何かの要因を組み合わせて実験するような 場合のことです。たとえば、植木鉢を 1 2 個用意して、それを 4 ずつ 3 組に分けて、それ ぞれの組の肥料の窒素濃度を 3 段階、(A1,

A2,A3）に設定する。そのそれぞれの組につい て、４段階のリンの濃度（B1, B2, B3, B4）を設定して、合計１２組の肥料濃度の異なる、 植木鉢を作って、それぞれの植木鉢に草を１本だけ植えて、その成長を比較するというよ うな場合です。実験条件としては、A1B1, A1B2, と言うような組み合わせになります。普通、こういう場合は、草を数本植えて繰り返しを作るで しょうから、あまり現実的な想定ではありませんが、絶対にあり得ないことではないし、 初めはできるだけ単純な方が考えやすいので、この形のデータについて考えて、それから、１要因の分散分析に発展させます。 とにかく次のような表３のようなデータが得られたとします。このデータは、今まで検討してきた合成したデータそのものです。しかし、分析者は合成データだということを知り ません。この場合、それぞれの行ごと列ごとに平均値を求めるでしょう。それらの平均値を使って平均値の平均値を計算し、A、Bの要因による平方和（SS）を計算します。

表 11. 二つの要因によるデータの総当たり（表９を改変）

	A1	A2	A3	合計	平均	偏差*	偏差^2
B1	2	6	7	15	5	-4	16
B2	6	10	11	27	9	0	0
B3	7	11	12	30	10	1	1
B4	9	13	14	36	12	3	9
合計	24	40	44	108	36		26
平均	6	10	11	27	9		
偏差**	3	1	2				
偏差^2	9	1	4	14			

＊：列平均－全平均　＊＊:行平均－全平均

おそらく、列ごと行ごとに平均値を求めて、それを列や業の代表値として、その平均を考えるし、列や行の平均値を使って、列ごと行ごとに平方和を計算するはずです。

列についてみると、

A_i ごとの平均値の平均値

$$\frac{6+10+}{3}=9$$

A_iごとの平均値から求めた平方和

$$(6-9)^2+(10-9)^2+(11-9)^2=14$$

B_i ごとの平均値の平均値

$$\frac{5+9+10+12}{4}=9$$

B_iごとの平均値から求めた平方和

$$(5-9)^2+(9-9)^2+(10-9)^2+(12-9)^2=26$$

A_iごとの平均値から求めた平方和、B_iごとの平均値から求めた平方和はそれぞれ元の A、

B の平方和　$SS_A = 14,\ SS_B = 26$になっています。
どうしてそうなるのかは、多分わかると思いますが、念のために説明します。平均値の方は、分散の和でも説明しましたし、わかると思うので省力します。平方和の方も分散の和で説明しましたが、もう一度説明します。
なります。次に平方和について考えると
１行目について
$\{(M_A + 1) - (M_A + M_B)\}^2$という計算をしているのだから、$(1 - M_B)^2$という計算をしたことになります。
同様に２列目は

$$(5 - M_B)^2$$

3 列目は

$$(6 - M_B)^2$$

4 列目は

$$(8 - M_B)^2$$

ですから　その総和は

$$(1 - M_B)^2 + (5 - M_B)^2 + (6 - M_B)^2 + (8 - M_B)^2 = \sum_{j=1}^{n_b} \left(B_j - M_B\right)^2 = SS_B$$

です。
これは、列についても同様で
　　列の平均値から全平均を引いた値の平方和はSS_Bになります。
このことは、SS_{total}を先に計算しておき、平均値のSSとしてSS_Aを求めれば、$SS_{total} = n_b SS_A + n_a SS_B$の式を使って$SS_B$が求まります。やってみます。

$$SS_B = \frac{SS_{total} - n_b SS_A}{n_a} = \frac{134 - 14 \times 4}{3} = \frac{134 - 56}{3} = \frac{78}{3} = 26$$

反対に

$$SS_A = \frac{SS_{total} - n_a SS_B}{n_b} = \frac{134 - 26 \times 3}{4} = \frac{134 - 78}{4} = \frac{56}{4} = 14$$

昔は計算を簡略化するために、こういう公式も実用的な意味があったのですが、ここでこれを紹介したのは、データから、残渣成分（要因によって説明できない変動）を計算するときにこの考え方が使えるからです。要因によって説明できない変動は、その言葉通り、要因から独立しています。ここで、B がそのような変動だと考えると、SS_Bを$SS_{residual}$と書くことになります。その分散

$$\sigma^2{}_{residual} = \frac{SS_{residual}}{n_b - 1} = \frac{SS_B}{n_b - 1} = \frac{SS_{total} - n_b SS_A}{n_a(n_b - 1)}$$

この残差分散に対して、要因Ａによる分散がどのくらい大きいかということを検討するの

で$F = \frac{{\sigma^2}_A}{{\sigma^2}_{residual}}$です。今使っている例では、$F = \frac{{\sigma^2}_A}{{\sigma^2}_{residual}}$では、$F = \frac{7}{8.667} = 0.808$

で、${\sigma^2}_B$単独で計算した時とF比は同じですが、自由度は違います。Aの自由度は$n_a - 1$で変わりませんが、残渣の自由度は$n_a(n_b - 1)$です。分子の自由度$n_a - 1$，分母の自由度$n_a(n_b - 1)$ F比の確率分布を使って、F比の有意性を検討します。分母の方は自由度$(n_b - 1)$のことをn_a繰り返しているので、それだけ分布が尖っているのです。

全平方和は、部分平方和の総和で、

たとえば

$$SS_{total} = n_b SS_A + n_A SS_B$$

自由度にも同じような、法則があって全自由度は部分自由度の総和です。

今使った例で言えば

$$df_{total} = df_a + df_{residual}$$

$$df_{total} = n_a n_b - 1$$

$$df_a = n_a - 1$$

ですから

$$df_{residual} = n_a n_b - 1 - (n_a - 1) = n_a n_b - n_a = n_a(n_b - 1)$$

です。

表 12. データの合成、A と C

		A				
		1	5	6	合計	平均
C	0.1	1.1	5.1	6.1	12.3	4.1
	0.5	1.5	5.5	6.5	13.5	4.5
	0.6	1.6	5.6	6.6	13.8	4.6
	0.8	1.8	5.8	6.8	14.4	4.8
	合計	6	22	26	54	18
	平均	1.5	5.5	6.5	13.5	4.5

あまり実感がわかないので、もう少し、現実性を持たせます。B の値を 10 分の 1 にした、データについて、総当たり的に和のデータを作ります。

このままだと、合成したデータ間が強いので、元の数値を隠して、データも、各列でランダムに入れ替えます。

表 13. C を実験区A_iの中の繰り返しとしてランダムに順番を変える

	A1	A2	A3
	1.8	5.6	6.5
	1.6	5.5	6.6
	1.1	5.8	6.8
	1.5	5.1	6.1
合計	6	22	26
平均	1.5	5.5	6.5

これならば、繰り返しのある 1 要因分散分析の形になっていますが、もちろん、繰り返しの和や平均は変わりません

表 14 計算手順

	A_1	A_1^2	A_2	A_2^2	A_3	A_3^2	総和	二乗総和
	1.8	3.24	5.6	31.36	6.5	42.25		
	1.6	2.56	5.5	30.25	6.6	43.56		
	1.1	1.21	5.8	33.64	6.8	46.24		
	1.5	2.25	5.1	26.01	6.1	37.21		
合計	6	9.26	22	121.26	26	169.26	54	299.78
平均	1.5		5.5		6.5			
全平方和							57	

実際の計算をどんな手順でやるかは、いろいろなやり方があります。集計用紙で手計算か、エクセルを使うのか。使う計算ソフトによって計算の仕方は違うでしょう。ここでは、あまり一般的ではない計算の手順を例示します。その方が、今までの話の流れにあっていて、モデル的にはわかりやすいからです。次に、一般的に使われている計算手順を示します。

$SS_x = \sum_{i=1}^{n} x_i{}^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2$という、簡便化された SS の計算の仕方を使って、データ入力と同時に、その 2 乗模計算してしまいます。
それらを縦に足して、さらに横に足すと、$\sum_{i=1}^{n} x_i$ =6+22+26=54 と$\sum_{i=1}^{n} x_i{}^2 = 9.26 + 121.26 + 169.26 = 299.78$ が出てきます。
これから、

$$SS_x = \sum_{i=1}^{n} x_i{}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 299.78 - \frac{1}{12} \times 54^2 = 299.78 - 243 = 56.78$$

となって$SS_{total} = 56.78$が出ます。
一方、平均値 1.5, 5.5, 6.5 からは、$\sum_{i=1}^{n} x_i = 1.5 + 5.5 + 6.5 = 13.5$ $\sum_{i=1}^{n} {x_i}^2 = 1.5^2 + 5.5^2 + 6.5^2 = 2.25 + 30.25 + 42.25 = 74.75$が出て、

$$SS_x = \sum_{i=1}^{n} {x_i}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 74.75 - \frac{13.5^2}{3} = 74.75 - \frac{182.25}{3} = 74.75 - 60.75 = 14$$

となって

$$SS_A = 14$$

$${\sigma^2}_A = \frac{ss_A}{df_a} = \frac{14}{3-1} = 7$$

$${\sigma^2}_{residual} = \frac{SS_{residual}}{n_b - 1} = \frac{SS_B}{n_b - 1} = \frac{SS_{total} - n_b SS_A}{n_a(n_b - 1)}$$

の式を使えば、

$${\sigma^2}_{residual} = \frac{SS_{residual}}{n_b - 1} = \frac{SS_B}{n_b - 1} = \frac{SS_{total} - n_b SS_A}{n_a(n_b - 1)}$$

の式を使って

$${\sigma^2}_{residual} = \frac{SS_{total} - n_b SS_A}{n_a(n_b - 1)} = \frac{56.78 - 4 \times 16}{3(4-1)} = \frac{0.78}{9} = 0.08667$$

という計算結果が出て

$$F = \frac{{\sigma^2}_A}{{\sigma^2}_{residual}} = \frac{7}{0.08667} = 80.769$$

F 分布表で、適当な α を選択し、分母の自由度$3(4-1) = 9$、分子の自由度 2 の臨界値を探し、有意性の判定をします。おそらく、この場合は、かなりの確率で有意差と判定されるでしょう。つまり、要因 A による違いは、ランダムな誤差要因に比べてはるかに大きいということです。ちなみに、手元にあった F 表（α=0.05(5%危険率)で、分子の自由度の自由度 2，分母の自由度 3 の臨界値は、9.55 ですが、分子の自由度 2、分母の自由度 9 の臨界値は 4.26 でした。つまり、1 実験の繰り返しを 1 から 4 に増やすことで、臨界値が小さくなり、低い F 比でも有意とされるということで、それだけ検出力が上がるということです。これは、我々の直観的な感覚に一致していますね。(繰り返し起こることは、それだけ再現性が高い。)

F 検定について、考え方は理解できたと思います。F 比を使う検定を ANOVA(analysis of variance)と言います。表 13 のデータの計算例を挙げましたが、原理的な説明の流れを視したために、あえて一般的でない計算手順を使いましたが、要因間の平方和を先に求める方が一般的です。多くの場合、繰り返しの数が実験区ごとに違うからです。魚の実験では、死亡魚が出たりして、繰り返し数がそろわないのが普通です。そういう場合も考えて、計

算手順を作ります。今の時代、実際、手計算する人はいないと思いますが、知っておくとより理解が深まります。
表 15 に、実験区 1 に 6 個、実験区 2 群に 7 個、実験区 3 群に 5 個、実験区 4 に 6 個 のデータセットの例を示しました。この平均値間に差があるかどうかを検討します。

表 15.1 要因分散分析の例

実験区			
1	2	3	4
2	10	8	9
5	2	7	15
3	4	3	8
8	9	4	12
9	13	5	13
4	14		4
	15		

計算手順をは以下の通りです。

1．全平方和(SStotal)を計算する
2．各水準ごとの平均値を計算する
3．残差平方和計算する
4．全平方和から残差平方和を差しいて、水準間の平方和とする
5．全自由度と水準間の自由度の差として、残差自由度を求める
6．残差平方和を残差の自由度で割って、残差分散を求める
7．水準間の平方和を水準間の自由度で割って、水準間の分散をもとめる
8．水準間の分散を残差自由度で割って、これを F 値とする
9．判定のための危険率を定め、水準間の自由度を分子の自由度、残差の自由度を分母の自由度として、F 臨界値の表などを使って、有意性を判定する。

表 16.Excel シート

	A	B	C	D	合計	
	2	10	8	9		
	5	2	7	15		
	3	4	3	8		
	8	9	4	12		
	9	13	5	13		
	4	14		4		
		15				
n_i	6	7	5	6	24	N
T_i	31	67	27	61	186	T
M_i	5.166667	9.571429	5.4	10.16667		
s_i	199	791	163	699	1852	S
T_i^2/n_i	160.1667	641.2857	145.8	620.1667	1567.419	$\Sigma t_i^2/n_i$

表中の記号の説明

一般化して、実験区を$i = 1, \cdots m$とします。

n_i: 実験区i のデーター数

T_i: 実験区i のデーターの合計

M_i: 実験区i の平均値

S_i: 実験区 i のデーターの平方和$\sum x^2$

何回か出てきた SS の簡便な計算式を使います。

$$SS_i = \sum_{j=1}^{n_j} {x_{ij}}^2 - \frac{1}{n}\left(\sum_{j=1}^{n_j} x_{ij}\right)^2$$

ここで

$$S_i = \sum_{j=1}^{n_i} {x_{ij}}^2$$

と表すことにすると

$$SS_i = S_i - \frac{{T_i}^2}{n_i}$$

これを各グループの残差平方和といいます。

この値のすべてのグループについての和は

$$S_1 - \frac{{T_1}^2}{n_1} + S_2 - \frac{{T_2}^2}{n_2} + \cdots + S_m - \frac{{T_m}^2}{n_m}$$

$$= S_1 + S_2 + \cdots + S_m - \left(\frac{{T_1}^2}{n_1} + \frac{{T_2}^2}{n_2} + \cdots + \frac{{T_m}^2}{n_m}\right)$$

これは全体の残差平方和の合計です。

二乗和の総和は

$$S = S_1 + S_2 + \cdots + S_m$$

ですから

全体の残差平方和は

$$SS_{residual} = S - \left(\frac{{T_1}^2}{n_1} + \frac{{T_2}^2}{n_2} + \cdots + \frac{{T_m}^2}{n_m} \right) = S - \sum_{i=1}^{m} \frac{{T_i}^2}{n_i}$$

一方全体の SS はSS_{total}

$$SS_{total} = S - \frac{T^2}{N}$$

です。

$$SS_{total} = SS_{\text{実験区}} + SS_{residual}$$

ですから、

$$SS_{\text{実験区}} = SS_{total} - SS_{residual} = S - \frac{T^2}{N} - \left\{ S - \left(\frac{{T_1}^2}{n_1} + \frac{{T_2}^2}{n_2} + \cdots + \frac{{T_m}^2}{n_m} \right) \right\}$$

$$= \left(\frac{{T_1}^2}{n_1} + \frac{{T_2}^2}{n_2} + \cdots + \frac{{T_m}^2}{n_m} \right) - \frac{T^2}{N}$$

$$= \sum_{i=1}^{m} \frac{{T_i}^2}{n_i} - \frac{T^2}{N}$$

表 16 の計算例では$N = 24, T = 186, S = 1852$ $\sum_{i=1}^{m} \frac{{T_i}^2}{n_i} = 1567.419$

だから

$$SS_{t0tal} = 1852 - \frac{186^2}{24} = 410.5$$

$$SS_{residual} = S - \sum_{i=1}^{m} \frac{{T_i}^2}{n_i} = 1852 - 1567.419 = 284.581$$

$$SS_{\text{実験区}} = 1567.419 - \frac{186^2}{24} = 125.919$$

$${\sigma^2}_{residual} = \frac{SS_{residual}}{df_{residual}} = \frac{SS_{residual}}{df_{total} - df_{\text{実験区}}} = \frac{284.581}{23 - 3} = 14.229$$

$${\sigma^2}_{\text{実験区}} = \frac{SS_{\text{実験区}}}{df_{\text{実験区}}} = \frac{125.919}{3} = 41.973$$

$$F = \frac{41.973}{14.229} = 2.950$$

となります。α＝0.05 の F 分布表で　分子の自由度 3。分母の自由度 20 で臨界値を調べると、3.098 です。わずかですが、臨界値に達していません。

残渣平方和を際に計算するところが、前の説明と違うところです。多くの教科書はこの形で計算しろと書いてあると思います。この方が計算が早いでしょう。ただ、事前の説明なしに

$$SS_{residual} = S - \sum_{i=1}^{m} \frac{T_i^2}{n_i}$$

$$SS_{実験区} = \sum_{i=1}^{m} \frac{T_i^2}{n_i} - \frac{T^2}{N}$$

これらの式の意味を直観的には分かるのは難しいと思います。