IV-3-1. studento の t 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

$$\frac{\mathrm{M}-\mu}{\Phi}=\tau$$

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

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

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、

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

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、

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

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 t の式を書き換えておきます。

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

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。

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

ϕ: *差の母集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。

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

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６.総当たりで和の平均を求める計算

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

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 7. 表 6 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。

$$x_i = M + e_i$$

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 32 の（1）のような構成になります。

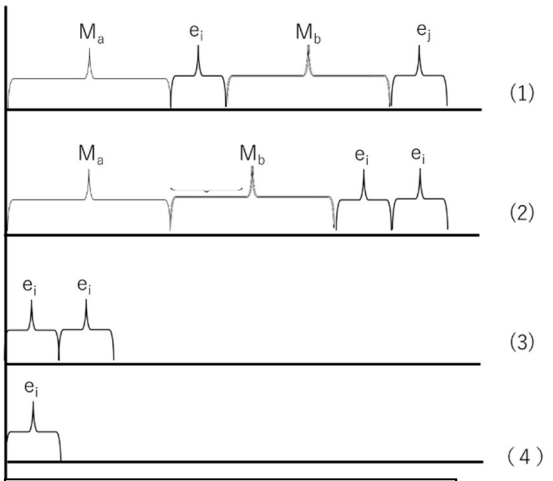

図 32.和のデータの構成要素

図 32 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 6 の総当たりの計算にすると、表 8 のようになります。
表 8.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 9.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。

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

$$SS_{total} = n_b SS_a + n_a SS_b \qquad \text{式 39}$$

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 8 の偏差の和の平均を求める計算に戻ります。

表 8.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 9 に行き、偏差の和の二乗をたし合わせています。表 9 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 8 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 10 の形式で書くと、

表 10-1. 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 10-2. 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、

$$SS_{A+B} = SS_{A+\bar{B}} + SS_{\bar{A}+B} = SS_a + SS_b \qquad 式 40-1$$

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

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

$$\sigma^2{}_{A+B}=\frac{\sigma^2{}_A(n_a-1)+\sigma^2{}_B(n_b-1)}{(n_a+n_b-2)} \qquad 式40-2$$

となります。式 40-2 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 12.表 8 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、

$$\phi^2 = \sigma^2{}_{A-B} = \frac{(n_a - 1)\sigma^2{}_A + (n_b - 1)\sigma^2{}_B}{n_a + n_b - 2} \qquad 式40-3$$

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

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。

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

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。標準誤差に加法性があるのならば、あるいは「総当たり的」に考えれば、

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

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

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

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し、それで原理的にも間違っていないと思います。

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

$$t=\frac{M_a-M_b}{\frac{\sigma_{A-B}}{\sqrt{\nu}}}$$

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

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

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 13. 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。