統計 GIS コース余話
三輪哲久（農業環境技術研究所）

第 1 回（第 0 回）：自由度 $n-1$

– À nous la liberté（自由を我等に） –

筆者からのお願い

大変に冗長なので，暇なときに（あるいは，数回に分けて）お読みください。

1　登場人物

M さん: ある研究所の管理職（統計研究者）。
T 君: 若手の農業研究者。

2　自由度 $n-1$

T 君: 先生。今日は自由度について質問に来ました。
M さん: わしは教育職ではないし，医者でもないので，「先生」と呼ばなくてもいいよ。
T 君: 先生は自分のことを「わし」と呼ぶのですか。
M さん: 広島では，小学生でも自分のことを「わし」と言っている。それに，「先生」はやめなさい。
T 君: それでは，何とお呼びしましょうか。
M さん: まあ，おたがい研究労働者だから，「さん付け」でいいだろう。
T 君: 何ですか，M さん。その「研究労働者」というのは。ずいぶん，しいたげられたような印象を受けますが。
M さん: 君は，R. A. Fisher の有名な “Statistical Methods for Research Workers” というテキストを知っているかい。「研究労働者」というのは “Research Workers” のことだよ。誇りを持っていいんだよ。
T 君: はあ。???
M さん: ところで，「自由度」の何について聞きたいのかね。
T 君: そもそも，自由度とはなんですか。
M さん: まあ，その名のとおり，「自由の程度」，あるいは「自由の度合い」といったところかな。英語では何というか知っているかい。
T 君: 用語集を見ると “degree of freedom” となっていました。まさに，「自由の程度」ですね。
M さん: だいたいはそれでよろしい。しかし，この “degree” というのは「温度」や「角度」の場合の「度」と同様に可算名詞だから，複数の自由度のとき，たとえば自由度 10 は，“ten degrees of freedom” というように複数形にしなければいけない。
T 君: はい（説教がましいなあ）。そうすると自由度が 1 の場合，単数で “one degree of freedom” というのですか。
M さん: たぶんね。ところで，こんなことをやっていると，一向に本題の「自由度 $n-1$」に進まない。実は，わしは非常に忙しいんだよ。
T 君: そうですね。わし，いや，僕も暇ではないんです。今日は，偏差平方和の自由度 $n-1$ について納得したら帰ります。
M さん: ということは，納得しないと帰らんということだな。まあ，いいだろう。つまり，同一の母集団からの n 個の観測値を $Y_1, Y_2, \ldots, Y_n$ としたときの偏差平方和

$$SS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{1}$$

の自由度 $n-1$ について知りたいわけだな。
T 君: あれ，M さん。なんで，データを表わすのに，アルファベットの Y を使うんですか。ほとんどの入門書では X を使ってますよ。
M さん: つまらんことを気にするな。それは，こちらの作戦だよ。君は次回，単回帰分析における平方和の自由度について訊きに来るに決まっている。そのときのために X（あるいは x）は取っておくのじゃ。とにかく Y の方が，なんとなく「確率変数」という気がするだろう。
T 君: まあ，どうでもいいですけど。もう，来ないかもしれませんし。
M さん: この問題は，いわゆる FAQ (Frequently Asked Question) とういうやつで，いろいろなところで説明されている。大学で習わなかったのかね。
T 君: 僕もいくつかのテキストを見てみました。よく見るのは次のような説明です。

> **よく行なわれる説明**
> 偏差平方和 (1) 式において，各項の 2 乗を取る前の n 個の偏差 $Y_i - \bar{Y}$ には，
> $$\sum_{i=1}^{n}(Y_i - \bar{Y}) = \sum_{i=1}^{n} Y_i - n\bar{Y} = 0$$
> のように，加えると必ずゼロになるという制約条件が 1 つある。したがって，偏差平方和 SS の自由度は，n よりも 1 だけ少なくなり $n-1$ となる。分散を計算するときも，SS を n ではなく，$n-1$ で割り $V = SS/(n-1)$ とする。

M さん: うまく説明されているではないか。では，さようなら。

T 君: ちょっ，ちょっと待ってください。なぜ，制約条件が 1 つあると自由度が 1 つ減るのですか。

M さん: $\sum_{i=1}^{n}(Y_i - \bar{Y}) = 0$ という制約があるため，n 個の偏差 $Y_1 - \bar{Y}, Y_2 - \bar{Y}, \ldots, Y_n - \bar{Y}$ のうち，最初の $n-1$ 個の値が決まると，残りの偏差 $Y_n - \bar{Y}$ の値は自動的に決まってしまう。$Y_n - \bar{Y}$ は自由に動くことができないのだよ。したがって，自由度が減るのだ。

T 君: しかし，n 個の偏差のうち 1 つが自由に動けないから自由度が 1 減って $n-1$ になるといわれても，それがどうしたのですか，としか言えないですね。たとえば，$Y_n - \bar{Y}$ が自由に動けないのならば，次のような説明はどうですか。

> **間違った説明**
> n 個の偏差 $Y_i - \bar{Y}$ には，$\sum_{i=1}^{n}(Y_i - \bar{Y}) = 0$ という制約があるため，最後の $Y_n - \bar{Y}$ は自由に動くことができない。したがって，偏差平方和
> $$SS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$
> を計算するときに，n 個の項が自由に動ける場合よりも値が大きくなってしまう。そのため，分散を計算するときは，SS を n ではなく，$n+1$ で割り $V = SS/(n+1)$ とする。

M さん: うっ。よし，それでは本格的に説明しよう。

3 χ^2 分布

M さん: 偏差平方和の自由度は χ^2 分布と密接に関係している。

T 君: t 分布表や F 分布表を参照する場合にも自由度が出てきます。

M さん: そうだね。しかし，t 分布では定義式の分母に χ^2 分布が現われるし，F 分布では定義式の分子と分母に χ^2 分布が現われる。このとき，t 分布や F 分布の自由度は元になった χ^2 分布の自由度に由来しているのだよ。

T 君: そうですか。

M さん: 前回の「正規分布から導かれる分布」で説明したように，χ^2 分布は次のように定義される。

> **χ^2 分布の定義**
> $U_1, U_2, \ldots, U_m$ が互いに独立に標準正規分布 $N(0,1)$ に従うとき
> $$\chi^2 = U_1^2 + U_2^2 + \cdots + U_m^2 = \sum_{i=1}^{m} U_i^2 \quad (2)$$
> は自由度 m の χ^2 分布に従う。自由度 m の χ^2 分布は，$\chi^2(m)$ や χ^2_m のように書かれることもある。

T 君: ちょっと待ってください。「前回」といったって，今日がこのシリーズの第 1 回めですよ。

M さん: その「前回」の分は来年までに準備するつもりだ。自由度 $n-1$ に関する質問が多いので，こちらを先に片付けないといけないのだ。とにかく，χ^2 分布の定義は上のとおりだ。独立な標準正規確率変数の 2 乗を m 個加えたとき，その平方和は χ^2 分布に従い，その自由度が m なのだ。この χ^2 分布の定義は，理解していると思っていいかね。

T 君: まあ，いいことにしましょう。「前回」の分は，来年までに何とかしてもらえるようですから。

M さん: さて，最初の偏差平方和に戻ろう。このとき，次の定理が成り立つ。

> **正規変量の平均と偏差平方和に関する定理**
> $Y_1, \ldots, Y_n$ が互いに独立に，共通の平均 μ，分散 σ^2 をもつ正規分布 $N(\mu, \sigma^2)$ に従うとき，平均 $\bar{Y}$ と偏差平方和 SS
> $$\bar{Y} = \sum_{i=1}^{n} Y_i/n$$
> $$SS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$
> に関して，次の性質が成り立つ。
> (a) $\bar{Y} \sim N(\mu, \sigma^2/n)$ (3)
> (b) $\bar{Y}$ と SS とは独立に分布する。
> (c) $SS/\sigma^2 \sim \chi^2(n-1)$ (4)

M さん: いままでに出てきたギリシャ文字の読み方は，カタカナで近似すると χ（カイ），μ（ミュー），σ（シグマ），Σ（シグマ）だ。学会など，英語で発表しなければならない場合は，発音を辞書で調べておいたほうがいいよ。

T 君: はい。

M さん: また記号に関して，以後は $Y_1, \ldots, Y_n$ のように，適宜 Y_2 を省略して書くことにする。今までは丁寧に $Y_1, Y_2, \ldots, Y_n$ のように書いていた。そうすれば Y_i の添え字 i は，1 ずつ増えることが分かる。高校の教科書ではこのように書いてある。しかし，添え字の動きが分かる場合は，紙数の都合により，$Y_1, \ldots, Y_n$ のように書くことにする。また，これも紙数の都合により，和の記号についても，添え字の範囲を省略して $SS = \sum(Y_i - \bar{Y})^2$ のように書くこともある。

T 君: （独り言。 紙数を気にするよりは，無駄話をやめた方が良いと思いますが。）

M さん: 今日の目的は，自由度 $n-1$ について本質的な部分を理解してもらうことだ。表記上の細かい点については，統一が取れていない可能性がある。気にしないでくれ。たとえば，すでに「正規確率変数」と言ったり，「正規変量」と言ったりしている。もし，意味のわからないことがあったら，いつでも質問してくれ。

T 君: はい。読者に代わって, しっかりと質問します。

M さん: さて本題に戻ろう。上の「正規変量の平均と偏差平方和に関する定理」の中の (4) 式で $n-1$ という値が登場している。

T 君: そうですね。このあたりが, 自由度 $n-1$ の鍵になるのでしょうか。

M さん: そうだ。この定理を理解すれば, 自由度 $n-1$ が分かるはずだ。ところが, 初等的な統計学のテキストには, この定理の証明が与えられていない。

T 君: 残念ですね。それは何故ですか。

M さん: おそらく, 初等的な知識の範囲で説明するのは面倒だからだろう。しかし, 面倒なことを避けていては世の中は前に進まないから, 今日はひとつ, 君に, この定理を理解して（理解したような気になって）帰ってもらおう。

T 君: 初等的なテキストに書かれていないことが分かるのですか。それは, うれしいなぁ。

M さん: さて, この定理を理解するには, 次の 2 つの方法がある。

(1) 多次元正規分布, 変数変換のヤコビアン, 直交行列, 直交変換などの知識を利用する。

(2) 和の記号 $\sum$ だけを使う（+ 根性）。

T 君: (1) の方法は, ほとんど中級以上の数理統計学ではないですか。(2) の方がいいなぁ。

M さん: そうだな。しかし, (1) の方が紙数が省略できるのだけれどね。とにかく (2) の方法でやってみよう。

T 君: お願いします。ただ, 小さい活字の（+ 根性）が気になりますが。

4 平方和の分解

M さん: まず第 1 段階は, (1) 式の偏差平方和を次のように表わすことから始まる。

$$\begin{aligned} SS &= \sum_{i=1}^{n}(Y_i-\bar{Y})^2 \\ &= \left(\frac{Y_1-Y_2}{\sqrt{2}}\right)^2 + \left(\frac{Y_1+Y_2-2Y_3}{\sqrt{6}}\right)^2 + \cdots \\ &+ \left(\frac{Y_1+\cdots+Y_k-kY_{k+1}}{\sqrt{k(k+1)}}\right)^2 + \cdots \\ &+ \left(\frac{Y_1+\cdots+Y_{n-1}-(n-1)Y_n}{\sqrt{(n-1)n}}\right)^2 \end{aligned} \tag{5}$$

T 君: 簡単な平方和の式が, 急に複雑になりましたね。目がくらくらしてきます。

M さん: よく見れば複雑なことはない。まず, 項の数が $n-1$ 個であることに注目してほしい。そして各項は規則的に構成されている。真ん中の添え字 k を使った一般項を見てごらん。Y_1 から Y_k までの和から次の変数 Y_{k+1} の k 倍を引いたものを適当な定数で割り, それを 2 乗している。この添え字 k を 1 から $n-1$ まで変化させて $n-1$ 項の 2 乗和（平方和）を計算している。

T 君: 各項の分母に現われている平方根の中の値はどのようにして決まっているのですか。

M さん: それは, おいおい分かる。わしは, ちょっと茶を飲んでくるから, その間に (5) 式を証明しておいてくれ。証明ができたら第 5 節の「線形結合の平均・分散・共分散」へ進んでいいよ。(読者の皆様も, まずは自分で証明を考えてみてください)。

T 君: なにかヒントをください。

M さん: 最初の k 個の数値 $Y_1,\ldots,Y_k$ だけを使ったときの平均と偏差平方和を次のように定義しよう。

$$\bar{Y}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} Y_i \tag{6}$$

$$SS_{(k)} = \sum_{i=1}^{k}(Y_i-\bar{Y}_{(k)})^2 \tag{7}$$

したがって, いままで単に $\bar{Y}$ や SS と書いていたものは, n 個の数値 $Y_1,\ldots,Y_n$ を使った平均と偏差平方和なので, $\bar{Y}=\bar{Y}_{(n)}$, $SS=SS_{(n)}$ のことだ。このとき $SS_{(k)}$ に関して

$$SS_{(k+1)} = SS_{(k)} + \left(\frac{\bar{Y}_{(k)}-Y_{k+1}}{\sqrt{(k+1)/k}}\right)^2 \tag{8}$$

が成り立つ。この (8) 式は,

$$\begin{aligned} SS_{(k)} &= Y_1^2+\cdots+Y_k^2-k\bar{Y}_{(k)}^2 \\ SS_{(k+1)} &= Y_1^2+\cdots+Y_{k+1}^2-(k+1)\bar{Y}_{(k+1)}^2 \\ \bar{Y}_{(k+1)} &= (k\bar{Y}_{(k)}+Y_{k+1})/(k+1) \end{aligned}$$

の関係から, 容易に根性で導くことができる。何か合いの手を入れてくれないと, わしのしゃべる部分が長くなりすぎてしまうではないか。

T 君: それでは合いの手を。「容易に」と「根性で」とは矛盾しないのですか。

M さん: 先に進もう。(5) 式と (8) 式をよく見てくれ。(5) 式の添え字 k で表わされている一般項と, (8) 式の第 2 項は同じものじゃ。2 乗を計算する前の値を Z_k としよう。

$$\begin{aligned} Z_k &= \frac{\bar{Y}_{(k)}-Y_{k+1}}{\sqrt{(k+1)/k}} \\ &= \frac{Y_1+\cdots+Y_k-kY_{k+1}}{\sqrt{k(k+1)}} \end{aligned} \tag{9}$$

そうすると (8) 式は

$$SS_{(k+1)} = SS_{(k)} + Z_k^2$$

と表わされる。以上より

$$\begin{aligned} SS &= SS_{(n)} = SS_{(n-1)} + Z_{n-1}^2 \\ &= SS_{(n-2)} + Z_{n-2}^2 + Z_{n-1}^2 \\ &= SS_{(1)} + Z_1^2 + \cdots + Z_{n-2}^2 + Z_{n-1}^2 \\ &= Z_1^2 + \cdots + Z_{n-2}^2 + Z_{n-1}^2 \end{aligned} \tag{10}$$

となる。ここで，$SS_{(1)}$ については，$\bar{Y}_{(1)} = Y_1$ より，$SS_{(1)} = 0$ だ。

T 君: もともとの偏差平方和の式 $SS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ では n 個の項の和だったものが，実は $n-1$ 個の項の和 $SS = \sum_{k=1}^{n-1} Z_k^2$ だったのですね。

M さん: そうだよ。このことが，最初に出てきた「よく行なわれる説明」（1 ページ）の「制約条件が 1 つあるから自由度が 1 つ減る」ということに対応している。

T 君: もう少し詳しく説明してください。

M さん: 一般に n 個の変数 $t_1, \ldots, t_n$ の 2 乗和（平方和）

$$SS_t = t_1^2 + \cdots + t_n^2$$

を考えよう。このとき，$c_1, \ldots, c_n$ を定数として，変数 $t_1, \ldots, t_n$ のあいだに

$$c_1 t_1 + \cdots + c_n t_n = 0$$

という制約条件があるとする。そうすると，$t_1, \ldots, t_n$ の線形結合を $n-1$ 個用いて，もとの平方和 SS_t は，この $n-1$ 個の線形結合の 2 乗和として表わすことができるんだ。

T 君: そうすると，制約条件が

$$c_1 t_1 + \cdots + c_n t_n = 0$$

$$d_1 t_1 + \cdots + d_n t_n = 0$$

のように 2 つあると $n-2$ 個の線形結合の 2 乗和で表わすことができるのですか。

M さん: そのとおりだ。

T 君: その証明も $\sum$ の記号を使うだけでできますか。

M さん: グラム・シュミットの直交化というような方法を使えば，できないことはない。しかし，それは，線形代数の直交行列とか逆行列などの概念を $\sum$ の記号で面倒に説明するだけだ。それよりも，線形代数を少しだけ勉強した方がよい。

T 君: そうですか。この先，大変だな。

M さん: そう落ち込まないでいいよ。今回の

$$SS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

という特別な形の平方和（偏差平方和）の場合は，(5) 式，あるいは (10) 式のように，とにかく $n-1$ 個の線形結合の平方和で表わせるということを証明できたわけだ。これだけで，SS の自由度が $n-1$ であることを納得する人もいる。君はどうかね。

T 君: うーん。ある程度納得できますね。

5 線形結合の平均・分散・共分散

M さん: (9) 式で表わされる $Z_1, \ldots, Z_{n-1}$ は，互いに独立な確率変数 $Y_1, \ldots, Y_n$ の線形結合（一次結合）になっている。Z_n が使われずに余っているので，

$$Z_n = \bar{Y} = (Y_1 + \cdots + Y_n)/n \tag{11}$$

と定義しよう。

M さん（独り言）: 数学者ならば，

$$Z_n = \sqrt{n} \cdot \bar{Y} = (Y_1 + \cdots + Y_n)/\sqrt{n}$$

と定義するだろうな。しかし，ここは素直に $Z_n = \bar{Y}$ と定義しておこう。

T 君: 何か言いましたか。

M さん: いや，何でもない。ここで $Z_1, \ldots, Z_n$ は，いずれも，$Y_1, \ldots, Y_n$ の線形結合となっている。一般に，$a_1, \ldots, a_n$ を既知の定数として，$Y_1, \ldots, Y_n$ の線形結合

$$X = a_1 Y_1 + \cdots + a_n Y_n \tag{12}$$

を考えよう。前々々回の「確率分布の性質」で説明したように，その平均と分散は次のように表わされる。

T 君: ちょっと待った。今度は「前々々回」ですか。それでは，「前々回」は何ですか。

M さん: それは，「正規分布の徹底的理解」だよ。

前回「正規分布から導かれる分布」
前々回「正規分布の徹底的理解」
前々々回「確率分布の性質」

T 君: そのうち「前々々々回」とか，あるいは先に進んで「次々々回」とかも出てくるのではないですか。

M さん: そのとおり。分かりにくければ，「前回」は「第 (-1) 回」，「前々回」は「第 (-2) 回」のように通し番号にしようか。

T 君: しかし，それでは今回（第 1 回）と前回（第 (-1) 回）との差が -2 となって，整合性がとれないのではないですか。

M さん: そうだな。それでは今回を「第 0 回」としよう。しかも，ベキ乗に対応させて「前回 $=$ 前^{1}回」，「前々回 $=$ 前^{2}回」，「今回 $=$ 次^{0}回」，「次回 $=$ 次^{1}回」，「次々回 $=$ 次^{2}回」のように表わすことにしよう（ここで「前 $=$ 次$^{-1}$」）。そうすると，後で書き直すときに一斉に書き直すことができて便利だ。

T 君: とにかく，先に進みましょう。

M さん: まず，独立な確率変数 $Y_1, \ldots, Y_n$ の線形結合

$$X = a_1 Y_1 + \cdots + a_n Y_n$$

の平均と分散は

$$\begin{aligned} \mathrm{E}(X) &= a_1 \mathrm{E}(Y_1) + \cdots + a_n \mathrm{E}(Y_n) \\ \mathrm{V}(X) &= a_1^2 \mathrm{V}(Y_1) + \cdots + a_n^2 \mathrm{V}(Y_n) \end{aligned}$$

と表わされる。このことは前[3]回(「確率分布の性質」)で説明した。これは，いいかね。

T 君: はい。

M さん: ここは，素直だね。次に，2 つの線形結合

$$X_1 = a_1Y_1 + \cdots + a_nY_n$$
$$X_2 = b_1Y_1 + \cdots + b_nY_n$$

があるとき，その共分散は

$$\mathrm{Cov}(X_1, X_2) = a_1b_1\mathrm{V}(Y_1) + \cdots + a_nb_n\mathrm{V}(Y_n)$$

で表わされる。これも，いいかね。

T 君: いいことにしましょう。

M さん: そう言ってもらえば，こちらも助かる。これらの平均・分散・共分散の式は $Y_1, \ldots, Y_n$ の分布が正規分布でなくても成り立つ。さて，われわれの Z_k について具体的に見てみよう。$Y_1, \ldots, Y_n$ は互いに独立に同じ分布 $N(\mu, \sigma^2)$ に従っているので，

$$\mathrm{E}(Y_i) = \mu,\ \mathrm{V}(Y_i) = \sigma^2, \quad 1 \le i \le n$$

だ。まず，(9) 式の Z_k について，

$$Z_k = a_1Y_1 + \cdots + a_nY_n$$

と表わしたときの線形結合の係数はどうなっているかね。

T 君: ええと。$1 \le k \le n-1$ の範囲にある k について，Z_k の係数は

$$a_1 = \cdots = a_k = 1/\sqrt{k(k+1)}$$
$$a_{k+1} = -k/\sqrt{k(k+1)}$$
$$a_{k+2} = \cdots = a_n = 0$$

です。係数の和，2 乗の和は

$$a_1 + \cdots + a_n = 0$$
$$a_1^2 + \cdots + a_n^2 = \frac{1 \times k}{k(k+1)} + \frac{k^2}{k(k+1)} = 1$$

となっています。美しいですね。だから，(9) 式の Z_k の定義式で $\sqrt{k(k+1)}$ で割っていたのですね。

M さん: そうだよ。これより，容易に

$$\mathrm{E}(Z_k) = 0,\ \mathrm{V}(Z_k) = \mathrm{E}(Z_k^2) = \sigma^2 \qquad (13)$$
$$1 \le k \le n-1$$

であることが分かる。ここで，$\mathrm{E}(Z_k) = 0$ であるから，$\mathrm{V}(Z_k) = \mathrm{E}(Z_k^2)$ となっている。$Z_n = \bar{Y}$ については，よく知られているように

$$\mathrm{E}(Z_n) = \mu,\ \mathrm{V}(Z_n) = \sigma^2/n$$

だ。次に，2 つの線形結合

$$Z_k = a_1Y_1 + \cdots + a_nY_n$$
$$Z_j = b_1Y_1 + \cdots + b_nY_n \quad (k < j)$$

の共分散を見てみよう。Z_k の係数 a_i については，

$$\sum_{i=1}^{k+1} a_i = 0, \quad a_{k+2} = \cdots = a_n = 0$$

が成り立つ。一方，$k < j$ である Z_j に対しては $k+1$ 番めまでの係数は値が等しく

$$b_1 = \cdots = b_{k+1}$$

なので，

$$\mathrm{Cov}(Z_k, Z_j) = (a_1b_1 + \cdots + a_nb_n)\sigma^2 = 0$$

が成り立つ。これは，$Z_j = Z_n$ の場合も成り立つ。すなわち，$Z_1, \ldots, Z_n$ の間の共分散は，すべてゼロになる。

6 不偏分散

M さん: ところで，(10) 式による SS の表現

$$SS = Z_1^2 + \cdots + Z_{n-1}^2$$

と，(13) 式の $\mathrm{E}(Z_k^2) = \sigma^2$ より，ただちに

$$\mathrm{E}(SS) = \sum_{k=1}^{n-1} \mathrm{E}(Z_k^2) = (n-1)\sigma^2 \qquad (14)$$

であることが分かる。あるいは，

$$\mathrm{E}\left(\frac{SS}{n-1}\right) = \sigma^2 \qquad (15)$$

が成り立つ。したがって，σ^2 の推定には SS を $n-1$ で割った不偏分散 $V = SS/(n-1)$ が使われる。

T 君: なるほど。

M さん: (15) 式が示すように，SS を $n-1$ で割ることによって不偏推定量が得られることから SS の自由度が $n-1$ であると説明することもある。どうかね。

T 君: ここも，ある程度は納得できます。

M さん: なお，(14) 式は，元の $Y_1, \ldots, Y_n$ が正規分布でなくても成立する。また，この式は，今回のような複雑な平方和の分解を行なわなくても簡単に示すことができる。

T 君: 簡単なのであれば，ここで示してくださいよ。

M さん: よく知られた平方和の定義式

$$SS = \sum (Y_i - \bar{Y})^2 = Y_1^2 + \cdots + Y_n^2 - n\bar{Y}^2$$

に，

$$\mathrm{E}(Y_i^2) = \mu^2 + \sigma^2, \quad \mathrm{E}(\bar{Y}^2) = \mu^2 + \sigma^2/n$$

を代入すれば，すぐに求まる。

T 君: 最後のところは，一般に確率変数 X に関して，

$$\mathrm{V}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$$

が成り立つことを使うのですね。

M さん: 補足してくれて，ありがとう。そのとおりだ。ところで，(14) 式は，SS の期待値が，σ^2 の n 倍ではなく，それよりも小さい値になることを示している。このことは，直観的にも説明することができる。

T 君: どういうことですか。

M さん: もし，仮にだ。真の平均値 μ を使うことができたとして，この真の平均値 μ との差の 2 乗和

$$SS_{\mathrm{true}} = \sum_{i=1}^{n} (Y_i - \mu)^2$$

を考えてみよう。各 Y_i は，この真の平均値 μ の周りに分布しているので，この SS_{true} については，その期待値が，σ^2 の n 倍

$$\mathrm{E}(SS_{\mathrm{true}}) = n\sigma^2$$

になることは納得できるだろう。

T 君: それは，そうですね。

M さん: ところが実際には真の平均値 μ は未知であるから, 我々の計算に使うことはできない。そこで代わりに, データから計算した標本平均 $\bar{Y}$ を使って, 偏差平方和 $SS = \sum (Y_i - \bar{Y})^2$ を計算することになる。このとき，観測値 $Y_1, \ldots, Y_n$ が得られたあと計算される標本平均 $\bar{Y}$ は，おおざっぱにいえば，$Y_1, \ldots, Y_n$ の真ん中あたりに来ることになる。本当は固定した値 μ の周りのバラツキを計算しなければならないのに，データから計算した都合のよい（つまり真ん中あたりに位置する）$\bar{Y}$ を使って計算するから，バラツキの程度を過少評価することになるんだ。

T 君: どれくらい過少評価していることになるのですか。σ^2 の n 倍ではなく，$n-2$ 倍とか，$n-3$ 倍になるのですか。

M さん: 君ね。分かっていて，とぼけてはいかんよ。

T 君: すみません。(14) 式の $\mathrm{E}(SS) = (n-1)\sigma^2$ が，その過少評価の程度を示しているのですね。これで，すっきりしました。いままで,「制約条件が 1 つあるから, 偏差平方和を $n-1$ で割る」といわれても, 何となくすっきりしなかったんですよ。今回は, $\mathrm{E}(SS) = (n-1)\sigma^2$ の関係を 2 とおりの方法で証明してもらっていますしね。

M さん: 次は，いよいよ SS/σ^2 が $\bar{Y}$ とは “独立に” χ^2 分布に従うことの説明だ。長くなりそうなので，すこし休憩しよう。

7 休憩

– **Intermission** –

8 正規変量の線形結合の分布

T 君: 独立な正規確率変数 $Y_1, \ldots, Y_n$ の線形結合である $Z_1, \ldots, Z_n$ も正規分布に従うのですか。

M さん: 問題はそこだ。一般に，正規確率変数 $Y_1, \ldots, Y_n$ の線形結合

$$X = a_1 Y_1 + \cdots + a_n Y_n$$

が正規分布に従うことは，仮に $Y_1, \ldots, Y_n$ が互いに独立でなくても成り立つ。それを示すためには，多次元正規分布や変数変換のヤコビアンなどの知識を必要する。しかし幸いなことに，いま我々は独立な正規確率変数を扱っている。そのときは，こつこつと数式を変形すれば，X が正規分布に従うことを示すことができる。これは 前[2] 回（「正規分布の徹底的理解」）で説明した。

T 君: 思い出せないので概略を説明してください（聞いていないのに思い出せるはずがない）。

M さん: 数学的帰納法を使う。まず, 1 つの正規変数 Y について，定数倍した

$$X = aY$$

が正規分布に従うことを示す。これは，1 変数の積分変数の変換で簡単に示すことができる。次に，Y_1 と Y_2 が独立に正規分布に従うとき（Y_1 と Y_2 の平均と分散は異なっていてもよい），和

$$X = Y_1 + Y_2$$

が正規分布に従うことを示す。そのためには，独立な確率変数 Y_1, Y_2 の密度関数を $f_1(y_1), f_2(y_2)$ とすると，和 $X = Y_1 + Y_2$ の密度関数 $g(x)$ が

$$g(x) = \int_{-\infty}^{\infty} f_1(y_1) f_2(x - y_1) dy_1 \qquad (16)$$

と表わされることを使う。$f_1(y_1), f_2(y_2)$ に正規分布の密度関数を代入して計算してみれば，$g(x)$ もまた正規分布の密度関数になることが確かめられる。このとき，任意の p, q^2 に対して，

$$\frac{1}{\sqrt{2\pi q^2}} \int_{-\infty}^{\infty} \exp\left[-\frac{(y-p)^2}{2q^2} \right] dy = 1$$

の関係は使っていいよ。あとは, 数学的帰納法により, (12) 式の線形結合が正規分布に従うことが分かる。

9 $Z_1, \ldots, Z_n$ の独立性

T 君: $Z_1, \ldots, Z_n$ が正規分布に従うことが分かりました。それから，第 5 節の最後のところで $Z_1, \ldots, Z_n$ は互いに共分散がゼロということが示されました。当然，相関係数も全てゼロですね。

M さん: そうだよ。

T 君: そうすると，$Z_1, \ldots, Z_n$ は互いに独立ということになるのですか。

M さん: そうはいかない。それは 前[3]回（「確率分布の性質」）で説明したはずだ。2 つの確率変数が独立であれば，共分散（相関）はゼロになる。しかし，2 つの確率変数の共分散（相関）がゼロであっても，その 2 つの確率変数は必ずしも独立とは限らない。

T 君: そうでしたね（合いの手）。

M さん: しかし，前[2]回 の「正規分布の徹底的理解」で説明したように，正規確率変数であれば，共分散（相関）がゼロであれば独立になる。

T 君: これも思い出せないので説明してください。

M さん: それでは，2 次元正規分布で説明しよう。しかし，困ったな。

T 君: そんなに難しいのですか。

M さん: いや，内容は難しくない。今回の原稿を二段組みで書き始めたため，密度関数の式が一段に収まらなくて，わしの美的感覚に合わないのだよ。特にページや段組みの切り替わり付近では，数式が途中で切れないように気を使うのだ。そのため，数式を次のページ（次の段）に送った方がよいときは，無駄話を長めにしている。

T 君: 無駄話も単純ではないのですね。

M さん: さて，Z_1, Z_2 が 2 次元正規分布

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix} \right)$$

に従うとき，その密度関数は

$$f(z_1, z_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp\left[-\frac{1}{2(1-\rho^2)} \left\{ \frac{(z_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(z_1-\mu_1)(z_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(z_2-\mu_2)^2}{\sigma_2^2} \right\} \right] \quad (17)$$

と書ける（前[2] 回「正規分布の徹底的理解」）。

T 君: 恐ろしい式ですね。n 次元だったら大変ですね。

M さん: いや，n 次元のときは行列を使って表わすから，かえって簡単になる。今回も，直交行列を使えば，こんな長文にはならなかったんだよ。

T 君: すみません。

M さん: 別に謝らなくていいよ。さて，(17) 式で $\rho = 0$ と置くと，あっけなく

$$f(z_1, z_2) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{(z_1-\mu_1)^2}{2\sigma_1^2} \right] \times \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{(z_2-\mu_2)^2}{2\sigma_2^2} \right]$$

となり，Z_1 と Z_2 とが独立であることが分かる。確率変数が独立であることの定義は「同時密度関数が，それぞれの密度関数の積で表わせること」だったことを思い出してくれ。n 次元正規分布についても同様だ。$Z_1, \ldots, Z_n$ の共分散（相関）が全てゼロであれば，$Z_1, \ldots, Z_n$ は互いに独立になる。

10 偏差平方和の分布

M さん: ここまでくれば，最初に示した「正規変量の平均と偏差平方和に関する定理」の証明は簡単だ。

T 君: そんな感じですね。

M さん: では，君が説明してみたまえ。

T 君: そうですか。まず，(10) 式から，偏差平方和は

$$SS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = Z_1^2 + \cdots + Z_{n-1}^2$$

と表わされます。そして，(13) 式から $Z_1, \ldots, Z_{n-1}$ は互いに独立に，平均がゼロ，分散が σ^2 の正規分布に従うのでした。ということは，$Z_1/\sigma, \ldots, Z_{n-1}/\sigma$ は互いに独立に平均がゼロ，分散が 1 の正規分布，すなわち，標準正規分布に従うことになります。そうすると，χ^2 分布の定義から，SS/σ^2 が自由度 $n-1$ の χ^2 分布に従うことはすぐ分かります。

M さん: そのとおりだ。

T 君: また，$Z_1, \ldots, Z_{n-1}$ は，$Z_n = \bar{Y}$ とも独立だったので，SS は，$\bar{Y}$ と独立に分布することになります。

M さん: めでたし，めでたし。

11 まとめ（中締め）

M さん: ここまでの話をまとめておこう。

まとめ

(1) n 個の数値 $Y_1, \ldots, Y_n$ から計算される偏差平方和は，

$$SS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = Z_1^2 + \cdots + Z_{n-1}^2$$
$$Z_k = \frac{Y_1 + \cdots + Y_k - kY_{k+1}}{\sqrt{k(k+1)}}$$

と表わされる。この式は，数式の変形によるだけなので，$Y_1, \ldots, Y_n$ の分布が何であっても成り立つ。

(2) $Y_1, \ldots, Y_n$ が互いに独立に，平均 μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ に従うとき，$Z_1, \ldots, Z_{n-1}$ は互いに独立に，平均ゼロ，分散 σ^2 の正規分布に従う。$Z_1, \ldots, Z_{n-1}$ は $\bar{Y}$ とも独立である。

(3) 以上により，SS/σ^2 は，平均 $\bar{Y}$ とは独立に，自由度 $n-1$ の χ^2 分布に従う。

T 君: まとめると，たったこれだけですか。

M さん: それでは，ここで少し脇道にそれて，数値計算の話をしておこう。

12　数値計算の話

T 君: 「すこし脇道」といっても，上の「まとめ」の短さを見れば分かるように，ほとんどが脇道のような気がしますが。

M さん: このシリーズは，単に統計の知識を身につけるだけでなく，その他の，役に立ったり，役に立たなかったりする広い知識を体で覚えることを目的にしているんだ。

T 君: しかし，この長い冗長な文章は，読んでもらえますかね。

M さん: どうかな。とにかく数値計算の話に移ろう。さて，偏差平方和を計算するには，次の 2 つの式が知られている。もう一度，書いておこう。

$$SS_1 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{18}$$

$$SS_2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \tag{19}$$

T 君: これは，初等的なテキストには必ず出てきますね。

M さん: そうだ。君は，どちらの計算方法が優れていると思う？

T 君: その手には乗りませんよ。昔の手計算の時代はともかく, 今のコンピュータの時代には (19) 式の SS_2 の方法は計算精度が悪くなる可能性があるので使わない方がよいと多くのテキストに書いてあります。

M さん: そのとおり。(19) 式の計算方法では，桁落ちが生じる可能性がある。

T 君: 「桁落ち」って何ですか。そもそも，何と読むのですか。

M さん: 読み方は辞書で調べなさい。

T 君: そんな鰾膠（にべ）もないことを言わないでください。

M さん: 君も難しい漢字を使うではないか。桁落ち（けたおち）については，次[30]回 に説明しよう。

T 君: 今後, 回を改めて説明するつもりはないということですね。

M さん: まあ, 簡単に説明すると次のとおりだ。各 Y_i が大きな値だと，$\sum Y_i^2$ は相当大きな値になる。また，$n\bar{Y}^2$ も大きな値となる。そして，大きい値から似たような大きい値を引くと，有効数字が，ごそっと失われることになるのだ。

T 君: (18) 式 SS_1 の方法だと，$Y_i - \bar{Y}$ という本質的に偏差を表わす部分を先ず計算し，その 2 乗和を計算するので精度が保たれるのですね。

M さん: そのとおりだ。しかし (18) 式は，どうも，わしの美的感覚に合わんのだ。

T 君: また，美的感覚ですか。何がどういけないのですか。

M さん: コンピュータで計算する場合でも，(18) 式のアルゴリズムでは，ループを 2 回まわさなければならない。1 回めのループで $\bar{Y}$ を計算し，2 回めのループで $Y_i - \bar{Y}$ の 2 乗和を計算することになる。さらに悪いことには，1 回めのループで $\bar{Y}$ を計算したあとも，2 回めのループに備えて，各 Y_i の値を保存しておかなければならない。

T 君: (19) 式ではどうなのですか。

M さん: ループは 1 回まわすだけでよい。Y_i と Y_i^2 をどんどんメモリーに加算していき，最後に (19) 式で平方和を計算すればよい。しかもだ。第 i ステップで Y_i を処理しているとき，それまでの $Y_1, \ldots, Y_{i-1}$ の値は捨ててしまってよいのだ。

T 君: その「$Y_1, \ldots, Y_{i-1}$ の値は捨ててしまってよい」ことによって，何が良くなるのですか。

M さん: メモリーのサイズが小さくてすむ。むかし，スピードの遅い CPU に，数十 KB のメモリーを積んで，紙テープやカードリーダからデータを読み込んでいたころには，苦労したもんだ。

T 君: いつ頃の話ですか。

M さん: 30 年以上まえの話だ。さて今回, 自由度 $n-1$ の話をしている際に，平方和の計算方法として第 3 の方法が出てきた。

T 君: 「まとめ」に出てきた

$$SS_3 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = Z_1^2 + \cdots + Z_{n-1}^2 \tag{20}$$
$$Z_k = \frac{Y_1 + \cdots + Y_k - kY_{k+1}}{\sqrt{k(k+1)}}$$

ですね。

M さん: そうだ。そして，第 4 節「平方和の分解」で説明したように，Z_k は第 k 項までの平均

$$\bar{Y}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} Y_i$$

を使って，

$$Z_k = \frac{Y_1 + \cdots + Y_k - kY_{k+1}}{\sqrt{k(k+1)}} = \frac{\bar{Y}_{(k)} - Y_{k+1}}{\sqrt{(k+1)/k}}$$

と書くことができる。ここで，$\bar{Y}_{(k)} - Y_{k+1}$ は第 k 項までの平均と第 $k+1$ 項との差であり，$Y_1, \ldots, Y_n$ のあいだのバラツキに由来している。つまり，Z_k^2 は本質的に $Y_1, \ldots, Y_k$ の変動（の一部）を表わすものになっている。

T 君: そうすると，(20) 式の計算方法は，(18) 式の計算 $\sum(Y_i - \bar{Y})^2$ と同じように，桁落ちが無くなるのですか。

M さん: そのとおり。まあ，(18) 式より多少は精度が悪くなるかもしれない。それと，もうひとつ重要なことは，Z_k は $\bar{Y}_{(k)}$ と Y_{k+1} だけを使って計算できるということだ。

T 君: ということは，$\bar{Y}_{(k)}$ さえメモリーに保存しておけば，$Y_1, \ldots, Y_k$ の値は捨ててしまってよいということですね。

M さん: おっ。なかなか察しがよいではないか。

T 君: つまり，(20) 式の方法は，少ないメモリーで，精度よく，ループを 1 回まわすだけで偏差平方和を計算できるということですか。

M さん: そういうことだ。メモリーの豊富な今の時代には，単純な (18) 式の SS_1 でよいかもしれない。電卓で精度よく計算したい場合などは，(20) 式の SS_3 を使ってくれ。

T 君: データ $Y_1, \ldots, Y_n$ が，時系列的に入ってくる場合にも使えそうですね。

M さん: そうだな。さて，長くなってきたので，そろそろ終わろうか。

T 君: しかし，M さん。9 ページ（奇数ページ）の左欄で原稿を終了するというのは，M さんの美的感覚に照らしてどうですか。

M さん: 極めて良くないな。書物によっては，各章の始まりは奇数ページとする場合も多い。そのとき，前の章が奇数ページで終わっていると，その裏の偶数ページを白紙にし，その次の奇数ページから次の章が始まることになる。出版社はいやがるであろう。奇数ページの左の欄というのが特に良くない。ところで，なんでここが 9 ページめだと分かるのかい。

T 君: LaTeX の文章の中に

```
\label{this-page}\pageref{this-page}
```

と書いておけば，ページが増えたり減ったりしても自動的に計算してくれます。Word では，えっと。

M さん: いや，特にいいよ。それじゃあ，もう少し数値計算の話を続けようか。

T 君: そうしましょう。

13 数値計算の実践

T 君: ところで，(18) 式，(19) 式，(20) 式の 3 つの計算方法で，実際の精度はどれくらい違うのですか。そもそも，(20) 式の SS_3 を使えば，本当に精度が良くなるのですか。

M さん: この統計 GIS コースは実践を旨とするので，数値計算の精度について，C 言語のプログラムで試してみよう。$Y_1 = 1,000,000$ から $Y_{11} = 1,000,001$ まで，0.1 刻みで値を変化させ ($n = 11$)，3 つの方法で計算するプログラムだ。C 言語を知らなければ，とばしても構わないよ。

C プログラム (ss.c)

```
#include <stdio.h>
main(void)
{
  double  y[12], total, avr, ss1, ss2, ss3;
  int     i, k, n=11;

  /* Data: y[1]=1,000,000, ..., y[n]=1,000,001 */
  for(i=1; i <= n; i=i+1)
    y[i] = 1000000.0 + 0.1*(i-1);

  /* Method 1 */
  for(i=1, total=0.0; i <= n; i=i+1)
    total = total + y[i];
  avr = total/n;
  for(i=1, ss1=0.0; i <= n; i=i+1)
    ss1 =  ss1 + (y[i] - avr)*(y[i] - avr);

  /* Method 2 */
  for(i=1, total=0.0, ss2=0.0; i <= n; i=i+1){
    total = total + y[i];
    ss2 = ss2 + y[i]*y[i];
  }
  ss2 = ss2 - total*total/n;

  /* Method 3 */
  for(k=1, avr=y[1], ss3=0.0; k <= n-1; k=k+1){
    ss3 = ss3 + (avr-y[k+1])*(avr-y[k+1])/(k+1)*k;
    avr = (k*avr + y[k+1])/(k+1);
  }

  printf("SS1 =%20.17lf\n"
         "SS2 =%20.17lf\n"
         "SS3 =%20.17lf\n", ss1, ss2, ss3);
}
```

M さん: 上の C プログラムは，本文の記述に合わせるためと，C 言語を知らない人にも感じが分かるように，通常の C プログラムの書き方とは，少し変えてある。たとえば，C 言語では配列の添え字は 0（ゼロ）から始まる。しかし，ここでは本文に合わせて y[1], ..., y[n] の偏差平方和を計算している。また通常は，'i=i+1' とは書かないで 'i++' のように書く。

T 君: なんとなく，やっていることはわかります。

M さん: さっそく，コンパイルして実行してみよう。

コンパイルと実行結果

```
$ gcc ss.c -o ss.exe
$ ./ss.exe
SS1 = 1.10000000004656640
SS2 = 1.10156250000000000
SS3 = 1.09999999955762195
```

M さん: 偏差平方和の正確な値は $SS = 1.1$ だ。

T 君: 実行結果の SS1, SS2, SS3 が，それぞれ (18) 式，(19) 式，(20) 式に対応するのですね。確かに (19) 式の SS_2 はかなり精度が悪いですね。(20) 式の SS_3 は，正確な値 $SS = 1.1$ よりも小さくなっていますが，精度は (18) 式の SS_1 よりも 1 桁悪い程度でしょうか。

M さん: なっ。言ったとおりだろう。

T 君: ところで，M さん。最近は，C 言語を知らなくても，R なら知っている人が多いですよ。R で計算するとどうなりますか。

M さん: R には，最初から `var()` という関数が用意されている。不偏分散 $V = \sum(Y_i - \bar{Y})^2/(n-1)$ を計算してくれるはずだ。しかし，R を使って模擬実験（シミュレーション）をやってみようか。まず，従来の方法の (18) 式 (SS_1) と (19) 式 (SS_2) を計算してみよう。

R での模擬実験 (SS_1, SS_2)

```
> y <- seq(from=1000000.0, to=1000001.0, by=0.1)
> print( n <- length(y) )
[1] 11
> ss1 <- sum( (y-mean(y))^2 )      # Method 1
> ss2 <- sum(y^2) - n*mean(y)^2    # Method 2
> options(digits=22)
> ss1; ss2; (n-1)*var(y)
[1] 1.100000000046566
[1] 1.099609375
[1] 1.100000000046566
```

T 君: 'options(digits=22)' というコマンド（関数）は何ですか。

M さん: 表示桁数を増やすコマンドだ。また R では，代入演算 '<-' を実行するだけでは値が表示されない。`print(n <- length(y))` のように，'print()' 関数を同時に使っておけば代入結果が表示される。

T 君: R で書くと，SS_1 も SS_2 も簡単な式になりますね。精度については，SS_2 は予想どおり精度がよくないですね。また，`(n-1)*var(y)` の値は，SS_1 と完全に同じになっています。R の偏差平方和の計算は信用していいということですね。(20) 式の SS_3 はどうなりますか。

M さん: こんなもんかな。

R での模擬実験 (SS_3)

```
> avr <- y[1]
> ss3 <- 0.0
> for(k in 1:(n-1)){
+   ss3 <- ss3 + (avr - y[k+1])^2/(k+1)*k
+   avr <- (k*avr + y[k+1])/(k+1)
+ }
> ss1; ss2; ss3
[1] 1.100000000046566
[1] 1.099609375
[1] 1.1000000002793966
```

T 君: SS_3 の精度は，C 言語の場合と同様，SS_1 より 1 桁悪くなる程度でしょうか。しかし，SS_1 や SS_2 が 1 行で書けるのに対し，SS_3 を R で計算する式はあまり美しくないですね。

M さん: だから，R ではやりたくなかったんだ。R では，`y[1]` から `y[k]` までの平均は，`mean(y[1:k])` で計算できるので，もう少し簡単にはなる。さらに簡単な記述方法があれば連絡してくれ。

R での模擬実験（SS_3; `mean(y[1:k])` を使う）

```
> ss3 <- 0.0
> for(k in 1:(n-1)){
+   ss3 <- ss3 + (mean(y[1:k]) - y[k+1])^2/(k+1)*k
+ }
> ss1; ss2; ss3
[1] 1.100000000046566
[1] 1.099609375
[1] 1.100000000046566
```

T 君: なんと，精度は SS_1 と全く同じですね。

M さん: 不思議なもんだなあ。

T 君: ところで，M さん。値を 10 倍して，$x[1] = 10,000,000, \ldots, x[11] = 10,000,010$（刻み 1.0）とすると，理論的には，偏差平方和は $10^2 = 100$ 倍の $SS = 110$ となるはずですね。これを (19) 式 (SS_2) の方法で計算するとどうなりますか。

M さん: まあ，疑問に思ったら，とにかく実行してみることだ。

T 君: はい。

10,000,000 ～ 10,000,010 (step=1.0) の偏差平方和

```
> x <- seq(from=10000000.0, to=10000010.0, by=1.0)
> print( n <- length(x) )
[1] 11
> print( ss4 <- sum(x^2) - n*mean(x)^2 )
[1] 110
```

T 君: 最後の "[1] 110" という出力は，結果がぴったり $SS = 110$ ということですか。

M さん: そうだろうね。

T 君: しかし R の内部では，整数でも倍精度の浮動小数点で処理されているのではないですか。よし，C 言語でもやってみよう。

C プログラム (ss_integer.c)

```
#include <stdio.h>
main(void)
{
  double  x[12], total, ss4;
  int     i, n=11;

  /* Data: x[1]=10,000,000, ..., x[n]=10,000,010 */
  for(i=1; i <= n; i=i+1)
    x[i] = 10000000.0 + (i-1);

  /* Method 2 */
  for(i=1, total=0.0, ss4=0.0; i <= n; i=i+1){
    total = total + x[i];
    ss4 = ss4 + x[i]*x[i];
  }
  ss4 = ss4 - total*total/n;

  printf("SS4 =%20.17lf\n", ss4);
}
```

T 君: プログラムはこんなものかな。さっそく，コンパイルして実行してみよう。

コンパイルと実行結果

```
$ gcc ss_integer.c -o ss_integer.exe
$ ./ss_integer.exe
SS4 =110.00000000000000000
```

T 君: ひぇー。C 言語の double 宣言で計算しても，ぴったり $SS = 110$ となりました。double 変数は倍精度浮動小数点のはずなのに。$X_1, \ldots, X_{11}$ が，値が大きくても整数値だからこうなるのかな。

M さん: 浮動小数点の計算ではいろいろ面白いことが起こる。たとえば，R の '==' という演算子は，左辺と右辺とを比較し，両辺の値が等しければ 'TRUE'，異なれば 'FALSE' を返す。そして，次のような変わった結果が出ることもある。

R の '==' 演算子のテスト

```
> print( 0.4*0.4 == 0.16 )
[1] FALSE
> print( 0.5*0.5 == 0.25 )
[1] TRUE
```

T 君: 数学的には，どちらも 'TRUE' のはずなのに，結果は違うのですね。

M さん: 'if(条件文)' などで判定するときには，気をつけないといけない。しかし，話が脇道のさらに枝道にそれていく気配なので，このへんでやめよう。もし可能ならば，浮動小数点の話は 次[31]回 に説明する。

T 君: 分かりました。とにかく今日は，偏差平方和と，その自由度 $n-1$ について訊きに来たのですから。

M さん: よし，これで終わりにしよう。

T 君: そうはいきませんよ。ここは，また奇数ページ（11 ページ）の左欄ですよ。

M さん: 困ったな。それでは，今日の偏差平方和の応用として一元配置分散分析の話をしよう。

T 君: うまく偶数ページの右の欄で原稿を書き終えることができますかね。

M さん: 心配するな。山より大きな獅子は出ない。

T 君: その諺は，もっと深刻な事態に立ち向かうときに使うのかと思っていました。

14 一元配置分散分析

M さん: 今日の話の中心は，$Y_1, \ldots, Y_n$ が互いに独立に同一の正規分布 $N(\mu, \sigma^2)$ に従うとき，偏差平方和 SS と平均 $\bar{Y}$ とは独立であり，SS/σ^2 は自由度 $n-1$ の χ^2 分布に従うということだ。

T 君: そこはよく分かりました。数値計算の部分は，少し難しかったですが。

M さん: なお，SS/σ^2 が自由度 $n-1$ の χ^2 分布に従うということを数式で表わすと

$$SS/\sigma^2 \sim \chi^2(n-1)$$

となる。この式で，左辺に未知パラメータの σ^2 が使われていることが感じが悪いと思う場合は，

$$SS \sim \sigma^2 \chi^2(n-1)$$

と表わすこともある。χ^2 分布に従う確率変数を σ^2 倍した分布に SS が従うという意味だ。$SS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ は，観測値 $Y_1, \ldots, Y_n$ だけから計算される値なので，データの解析手順を考えているときには，こちらの方が都合がよい。

T 君: そうですね。

M さん: 今日の話の中で，偏差平方和 SS と平均 $\bar{Y}$ とが独立に分布するということが重要だ。このことにより，t 検定や F 検定が可能になるのだ。この点は，あまり初等的なテキストには書かれていない。

T 君: そういえば，分散分析のテキストでも，いきなり F 分布が登場したりしますね。

M さん: さて，一元配置実験の話をしよう。因子 A の a とおりの処理（水準）を $A_1, \ldots, A_a$ とする。水準 A_i で n 回の繰り返しがあるとして，そのデータを

$$y_{ij} \quad (1 \le i \le a;\ 1 \le j \le n)$$

としよう。添え字 i は処理の違いを表わし，1 から a までの値をとる。添え字 j は，各処理 A_i での繰り返しを表わし，1 から n までの値をとる。観測値 y_{ij} の総数は $a \times n = an$ 個だ。このとき，実験では an 個の実験単位（試験区）が必要になる。この an 個の実験単位（試験区）を完全にランダムに配置した実験は，「一元配置実験（一元配置法）」，あるいは「一因子完全無作為化実験（一因子完全無作為化法）」とよばれる。よいかね。

T 君: はい。ここまで添え字の動き方を丁寧に説明したテキストは無いと思います。ただですね，いままで確率変数は大文字の Y を使っていたのに，なぜ，ここで小文字の y に変わってしまったのですか。

M さん: そんな質問をするから，脇道の話が多くなってしまうのだ。しかし，この「統計 GIS コース余話」シリーズでは，脇道を恐れずに説明することにしよう。「余話」なのだから。

T 君: はい。お願いします。

M さん: 前[3]回（「確率分布の性質」）において，確率変数を大文字の Y で表わし，その確率変数 Y の実現値（あるいは，確率変数 Y が取りうる値）を，対応する小文字の y で表わした。そうすると，いろいろな確率分布の性質を分かりやすく説明することができる。たとえば，密度関数 $f(y)$ は

$$f(y) = \lim_{\Delta y \to 0} \frac{1}{\Delta y} \mathrm{Prob}\{y \le Y \le y + \Delta y\}$$

と定義できるし，確率変数 Y が区間 $[c,\ d]$ に入る確

率は,

$$\mathrm{Prob}\{c \le Y \le d\} = \int_c^d f(y)dy$$

のような書き方をすることができる。

T 君: 確かに，$\mathrm{Prob}\{y \le Y \le y + \Delta y\}$ などは，大文字 Y と小文字 y を使い分けないと，表現することが難しいですね。

M さん: そうだ。しかし，一旦「確率分布の性質」を理解してしまうと，あまり大文字と小文字とを使い分ける必要がなくなってくる。むしろ，大文字と小文字とを使い分けていると無理が生じてくる。たとえば分散分析で，観測値を表わす確率変数を大文字の Y_{ij} で表わし，その実現値を小文字の y_{ij} で表わすとしよう。そうすると，後で説明する誤差平方和

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2$$

などに対しても，観測値から計算される上記の SS_E のほかに，確率変数としての $\sum\sum(Y_{ij} - \bar{Y}_{i\cdot})^2$ を表わす記号を用意しなければならなくなる。多くの入門書で苦労している。

T 君: そうなんですか（合いの手)。

M さん: そこでだ。「本章以降では，混乱のない限り，確率変数とその実現値を同じ小文字で表わす（場合によっては同じ大文字で表わす)。」と断り書きを入れておいて，小文字だけ（あるいは大文字だけ）を使って表現する方が混乱が生じないのだ。

T 君: さっきは，その断り書きがなかったような気がしますが。まあ，いいか。この方式は，M さんが考えたのですか。

M さん: いや，Rao (1965) の本の第 3 章の始めに書いてあった。

T 君: うまい方法ですね。

M さん: さて，処理 A_i の第 j 番めの観測値が平均 μ_i，分散 σ^2 の正規分布に従うとしよう。

$$y_{ij} \sim N(\mu_i, \sigma^2) \quad (1 \le i \le a; 1 \le j \le n) \quad (21)$$

an 個の y_{ij} は，互いに独立に分布している。平均 μ_i は処理 A_i に固有の平均値だ。ここで，$e_{ij} = y_{ij} - \mu_i$ を考えると，e_{ij} は平均ゼロの正規分布に従うので，(21) 式の代わりに

$$y_{ij} = \mu_i + e_{ij}, \quad e_{ij} \sim N(0, \sigma^2) \quad (22)$$
$$(1 \le i \le a; 1 \le j \le n)$$

と書くこともある。

T 君: こちらの方が，よく見る式ですね。

M さん: ここで特定の処理 A_i について，n 個の観測値 $y_{i1}, \ldots, y_{in}$ に注目しよう。この n 個の観測値は互いに独立に同一の正規分布 $N(\mu_i, \sigma^2)$ に従っている。

T 君: ということは，きょう学んだことが，そのまま使えるということですね。

M さん: そういうことだ。この n 個の観測値の平均と偏差平方和を計算してみよう。

$$\bar{y}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} y_{ij} \quad (23)$$

$$SS_i = \sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot}) \quad (24)$$

T 君: 平均 $\bar{y}_{i\cdot}$ の使われているドット '.' はどういう意味ですか。いままでは，平均は単に $\bar{Y}$ と表わしていましたが。

M さん: 分散分析では，複数の添え字が使われる。今回でも観測値は y_{ij} と表わされている。ドット '.' の記号は，その位置にある添え字に関して平均を計算することを示している。$\bar{y}_{i\cdot}$ では，添え字 j の位置にドット '.' がついているので，(23) 式のように添え字 j に関して平均を計算したことが分かるのだ。合計値を計算するときにも，ドット '.' が使われることがある。この $\bar{y}_{i\cdot}$ を，「処理平均」と呼ぶことにしよう。

T 君: 分かりました。ここで，$y_{i1}, \ldots, y_{in}$ が互いに独立に同一の正規分布 $N(\mu_i, \sigma^2)$ に従うという前提では，今日学んだことをさっそく使うと，

$$\bar{y}_{i\cdot} \sim N(\mu_i, \sigma^2/n)$$
$$SS_i \sim \sigma^2\chi^2(n-1)$$

ということですね。

M さん: そのとおりだ。もう 1 つ重要なことがあったはずだ。

T 君: ええと。そうだ。$\bar{y}_{i\cdot}$ と SS_i は独立に分布するのでした。ちょっと待ってくださいよ。処理が異なれば，当然，もとの y_{ij} は独立に分布していますよね。そうすると，a 通りの処理について処理平均と偏差平方和を計算した

$$\bar{y}_{1\cdot},\ \ldots,\ \bar{y}_{a\cdot},\ SS_1,\ \ldots,\ SS_a$$

は，全てが互いに独立に分布しているのですか。すごいですね。

M さん: このことが，後で F 検定を構成するときに重要になってくるのだよ。

T 君: そうですか。

M さん: まず偏差平方和に注目してみよう。SS_i は，偏差 $y_{ij} - \bar{y}_{i\cdot}$ の平方和であるし，$SS_i \sim \sigma^2\chi^2(n-1)$ から分かるように母平均 μ_i に依存しない。

T 君: 「母平均」とは何ですか。

M さん: 君は分からない点があったらすぐに訊いてくるね。まあ，いいことだがね。(23) 式で，観測値から計算した処理平均 $\bar{y}_{i\cdot}$ が出てきた。一方，今日の

話全体を通じて，正規分布の平均 μ（あるいは μ_i）が使われている。つまり，単に「平均」というと観測値から計算した「標本平均」なのか母集団の平均値としての「母平均」なのかが混乱する可能性がある。それで，特に注意が必要な場合に限り，「母平均」，「標本平均」，「処理平均」と詳しくよぶことにする。大抵の場合は，「平均 $\bar{y}_{i\cdot}$」や「平均 μ の正規分布 $N(\mu, \sigma^2)$」のように，「平均」のあとに対応する記号 $\bar{y}_{i\cdot}$ や μ も一緒に書くので，混乱はないであろう。

T 君: 分かりました。

M さん: とにかく，処理 A_i における偏差平方和 SS_i は，未知の母平均 μ_i に依存しないんだ。そこで，a 個の処理 $A_1, \ldots, A_a$ において計算された偏差平方和を合計した

$$SS_E = \sum_{i=1}^{a} SS_i = \sum_{i=1}^{a}\sum_{j=1}^{n} (y_{ij} - \bar{y}_{i\cdot})^2 \tag{25}$$

を考えよう。SS_E は「誤差平方和」とよばれる。この SS_E の分布はどうなると思うかね。

T 君: 前[1]回「正規分布から導かれる分布」の χ^2 分布の性質を使うのですね。

M さん: そうだ。君も 前[n]回 の使い方に慣れてきたようだね。それで，その χ^2 分布の性質を覚えているかね。

T 君: これは，重要な性質なので覚えています。

> **χ^2 分布の性質 1**
> X が自由度 df の χ^2 分布に従うとき，その期待値は df に等しい。
>
> $$\mathrm{E}[X] = df$$
>
> **χ^2 分布の性質 3**
> X_1, X_2 が，互いに独立に，それぞれ自由度 df_1, df_2 の χ^2 分布に従うとき，その和 $X_1 + X_2$ は自由度 $df_1 + df_2$ の χ^2 分布に従う。
>
> $$X_1 + X_2 \sim \chi^2(df_1 + df_2)$$

SS_E/σ^2 は a 個の独立な χ^2 確率変数の和ですが，数学的帰納法を使えば，SS_E/σ^2 が χ^2 分布に従うことはすぐに分かります。

M さん: その自由度はどうなるかね。

T 君: 各 SS_i の自由度は $n-1$ だったので，a 個の SS_i の和である SS_E の自由度は，自由度も合計して，$df_E = a(n-1)$ となります。つまり SS_E の分布は

$$SS_E \sim \sigma^2 \chi^2(df_E), \quad df_E = a(n-1) \tag{26}$$

ですね。

M さん: 完璧だね。

T 君: この誤差平方和を自由度で割れば，母分散 σ^2 の推定値が得られるのですね。

$$V_E = \frac{SS_E}{df_E}, \quad \mathrm{E}[V_E] = \sigma^2 \tag{27}$$

M さん: 次に，各処理 A_i で計算した $\bar{y}_{i\cdot}$ と SS_i のうち，処理平均 $\bar{y}_{i\cdot}$ について見てみよう。この処理平均は，a 個，つまり処理の数だけある。$\bar{y}_{1\cdot}, \ldots, \bar{y}_{a\cdot}$ だ。この a 個の処理平均を使って，今までと同じように偏差平方和

$$SS'_A = \sum_{i=1}^{a} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

を計算してみよう。

T 君: $\bar{y}_{\cdot\cdot}$ は，添え字 i と j の両方について平均を計算するのですね。

M さん: そうだ。詳しく書けば

$$\bar{y}_{\cdot\cdot} = \frac{1}{a}\sum_{i=1}^{a} \bar{y}_{i\cdot} = \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}$$

となる。通常は，総平均とよばれる。“平均の平均”，英語で言えば，“mean of means” かな。

T 君: むかし，“King of Kings” という映画がありましたね。

M さん: そんな畏れ多いことをいってはいかんよ。この余話と “King of Kings” とを比べるなんて。

T 君: いや，語呂が似ていたものですから。

M さん: 話を元に戻そう。SS'_A は $\bar{y}_{1\cdot}, \ldots, \bar{y}_{a\cdot}$ から計算される値なので，$SS_1, \ldots, SS_a$ から計算される SS_E とは独立に分布することになる。ただし，各 $\bar{y}_{i\cdot}$ は，異なる母平均 μ_i をもつ正規分布 $N(\mu_i, \sigma^2/n)$ に従っている。

T 君: ということは，同一の正規分布からの標本ではないので，今日学んだことが使えないということですね。残念だなぁ。

M さん: そこでだ。a 個の処理 $A_1, \ldots, A_a$ に関して，その効果に違いがないという仮説を想定してみよう。つまり，各処理の母平均 μ_i が全て等しく μ という値をもつという仮説だ。数式で表せば，

$$H_0 : \mu_1 = \cdots = \mu_i = \cdots = \mu_a \equiv \mu \tag{28}$$

となる。この仮説を統計用語で「帰無仮説（きむかせつ）」(null hypothesis) というんだ。

T 君: 「帰無仮説」については，統計の基礎を勉強しているときに何回か出てきました。

M さん: 実際，a とおりの処理 $A_1, \ldots, A_a$ を取り上げて実験を行なうとき，我々が知りたいことは，この a とおりの処理 $A_1, \ldots, A_a$ に関して効果に違いがあるかどうかということだ。つまり，帰無仮説が成

り立っているのか，あるいは帰無仮説は成り立っていなくて，処理のあいだの効果に違いがあるのかということだ。

T 君: 処理が 2 つだけの場合の帰無仮説 $H_0 : \mu_1 = \mu_2$ の拡張ですね。

M さん: そういうことだ。そこで，(28) 式の帰無仮説のもとで，処理平均の平方和を考えてみよう。

T 君: 帰無仮説 (28) のもとでは，$\bar{y}_{1\cdot}, \ldots, \bar{y}_{a\cdot}$ は同一の正規分布に従うことになります。こうなれば，今日の話が使えます。まかしてください。

M さん: ほう。

T 君: 偏差平方和を母分散 σ^2 で割った

$$\frac{SS'_A}{\sigma^2} = \frac{\displaystyle\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{\sigma^2}$$

が自由度 $a-1$ の χ^2 分布に従うのでしょう。

M さん: おしいね。$\bar{y}_{1\cdot}, \ldots, \bar{y}_{a\cdot}$ の分布をもう一度よく見てごらん。

T 君: ええと，$\bar{y}_{1\cdot}, \ldots, \bar{y}_{a\cdot}$ は，互いに独立に，同じ正規分布 $N(\mu, \sigma^2/n)$ に従っています。あっ，そうか。各 $\bar{y}_{i\cdot}$ の分散は，σ^2/n だから，偏差平方和 SS'_A を σ^2/n で割らないといけないのですね。そうすると，

$$\frac{SS'_A}{\sigma^2/n} = \frac{\displaystyle\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{\sigma^2/n} = \frac{\displaystyle n\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{\sigma^2}$$

が自由度 $a-1$ の χ^2 分布に従うことになります。

M さん: そのとおりだ。そこで，単純に $\bar{y}_{i\cdot}$ の平方和を計算した SS'_A ではなく，それを n 倍した

$$SS_A = nSS'_A = n\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \tag{29}$$

を考えてみよう。そうすると，

$$SS_A/\sigma^2 \sim \chi^2(a-1), \quad SS_A \sim \sigma^2\chi^2(a-1)$$

となる。だから，この SS_A の方が都合がいいんだ。この SS_A は「処理平方和」とよばれる。その自由度は $df_A = a-1$ だ。

T 君: そうすると，$SS_A \sim \sigma^2\chi^2(a-1)$ ということは，処理平方和 SS_A をその自由度で割った

$$V_A = SS_A/df_A \tag{30}$$

について，帰無仮説のもとで $\mathrm{E}[V_A] = \sigma^2$ が成り立つのですか。

M さん: そうなるね。しかもだ，上に書いたように SS'_A は SS_E とは独立だったから，SS'_A を n 倍した SS_A も，SS_E とは独立になる。この独立性は，帰無仮説の下でなくても成り立つ。

T 君: そろそろ，F 分布の定義が出てきそうですね。

M さん: そのようだな。

> ***F* 分布の定義**
> X_1, X_2 が互いに独立に，それぞれ自由度 df_1, df_2 の χ^2 分布に従うとき
> $$F = \frac{X_1/df_1}{X_2/df_2} \tag{31}$$
> は自由度 (df_1, df_2) の F 分布に従う。この F 分布は，$F(df_1, df_2)$ と表記される。

T 君: 処理平方和 SS_A と誤差平方和 SS_E に関して，

$$SS_A/\sigma^2 \sim \chi^2(df_A), \quad df_A = a-1$$
$$SS_E/\sigma^2 \sim \chi^2(df_E), \quad df_E = a(n-1)$$

で，SS_A と SS_E とは独立ですから，帰無仮説の下で

$$F = \frac{V_A}{V_E} = \frac{SS_A/df_A}{SS_E/df_E} \sim F(df_A, df_E)$$

が言えるということですね。うまい具合に，分子と分母に出てくる σ^2 がキャンセルされています。

M さん: これで，一元配置分散分析の話の主要部分は終わりだ。最後に平方和の加法性の話をしておかなくちゃならんだろう。

T 君: そうですね。大抵のテキストは，その平方和の加法性の話から始まっていますね。。

M さん: そうなんだ。君もよく見かけているはずだ。ここも，復習の意味で，君が説明してみるかね。

T 君: はい。先ず，an 個のデータ y_{ij} について，総平均 $\bar{y}_{\cdot\cdot}$ からの偏差平方和

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\cdot\cdot})^2$$

を計算するんでしたね。

M さん: そうだ。この平方和 SS_T は，an 個のデータ全体のバラツキを表わしているので，「総平方和」とよばれる。

T 君: そして，これは an 個のデータの偏差平方和なので，第 4 節の方法で $an-1$ 個の項の平方和として表わすことができるわけですね。だから，その自由度は $df_T = an-1$ となります。処理平方和 SS_A と誤差平方和 SS_E は，ここまでの議論と同じですね。

$$SS_A = n\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2, \quad df_A = a-1$$
$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2, \quad df_E = a(n-1)$$

あとは，簡単な計算により，平方和と自由度に関して，加法性

$$SS_T = SS_A + SS_E, \quad df_T = df_A + df_E$$

が成り立つことが確かめられます。

15　本当のまとめ

M さん: n 個の観測値 $Y_1, \ldots, Y_n$ から計算される偏差平方和

$$SS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

に関して, 今日は, いろいろな観点から説明した。ここで，一緒にまとめてみよう。まず，最初に出てきた性質は，偏差平方和の各項の 2 乗を取る前の偏差に関して，その合計がゼロになるという

$$(1) \quad \sum_{i=1}^{n} (Y_i - \bar{Y}) = 0$$

という式だ。つまり，n 個の偏差 $Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y}$ の間に，制約条件が 1 つあるということだ。「この事実によって，自由度が 1 だけ減って $n-1$ になる」といわれても，なかなか納得しがたいであろう。

T 君: そうですね。僕にしても, 2 ページのような「間違った説明」を思いつきますからね。

M さん: 次に，偏差平方和 SS は，

$$(2) \quad SS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = Z_1^2 + \cdots + Z_{n-1}^2$$
$$Z_k = \frac{Y_1 + \cdots + Y_k - kY_{k+1}}{\sqrt{k(k+1)}}$$

のように，$n-1$ 個の項の 2 乗和（平方和）として表わされることを示した（第 4 節)。前にも説明したように，この式は単なる数式の変形から導かれるので，$Y_1, \ldots, Y_n$ がどのような分布に従っていてもよい（互いに異なる分布であってもよい)。

T 君: これは，数式をフォローするのに根性を必要としましたが，面白かったですね。

M さん: この $n-1$ 個の項の 2 乗和（平方和）で表わされることにより，偏差平方和 SS の自由度が $n-1$ であるということもある。

T 君: これも何となく，理解できます。

M さん: ここで，ちょっと注意しておくと，偏差平方和 SS を $n-1$ 項の 2 乗和として表わす方法は，一とおりではない。たとえば，n 個の観測値 $Y_1, \ldots, Y_n$ は互いに平等だから，Y_1 からスタートするのではなく，Y_n からスタートしてもよい。

T 君: そういえば，そうですね。

M さん: 次に，$Y_1, \ldots, Y_n$ が互いに独立に同じ分布に従い，

$$\mathrm{E}(Y_i) = \mu, \quad \mathrm{V}(Y_i) = \sigma^2$$

のように，平均と分散が共通であるとしよう。そうすると，話が大きく進展する。

T 君: Z_k $(1 \le k \le n-1)$ に関して，

$$(3) \quad \mathrm{E}(Z_k) = 0, \quad \mathrm{V}(Z_k) = \sigma^2 \quad (1 \le k \le n-1)$$
$$\mathrm{Cov}(Z_k, Z_j) = 0$$

のように，全て平均がゼロ，分散は等しく σ^2 になるのでしたね。また，互いに共分散（相関）はゼロです。

M さん: この時点で，正規分布を仮定しなくても，偏差平方和の期待値が

$$(4) \quad \mathrm{E}(SS) = (n-1)\sigma^2$$
$$\mathrm{E}\left(\frac{SS}{n-1}\right) = \sigma^2$$

であることが分かる。つまり，偏差平方和 SS をデータの個数 n で割るのではなく，自由度 $n-1$ で割った不偏分散 $V = SS/(n-1)$ を使えば，σ^2 の偏りのない推定量（不偏推定量）が得られることが分かる。このことも，偏差平方和 SS の自由度が $n-1$ であるということの裏付けになる。

T 君: このあたりになると，「偏差平方和 SS の自由度は $n-1$ である」という表現が自然に思えてきますね。

M さん: そうだろう。そして，最後はな ... 。

T 君: そこは，僕にまかせてください。

M さん: そうか。

T 君: $Y_1, \ldots, Y_n$ が互いに独立に同じ正規分布 $N(\mu, \sigma^2)$ に従うとき，SS/σ^2 は，平均 $\bar{Y}$ とは独立に，自由度 $n-1$ の χ^2 分布に従う，つまり

$$(5) \quad SS/\sigma^2 \sim \chi^2(n-1)$$
$$SS \sim \sigma^2 \chi^2(n-1)$$

が成り立つということです。

M さん: これら，(1) ~ (5) のことを総合して，「偏差平方和 SS の自由度は $n-1$ である」というんだ。

T 君: はい，分かりました。

M さん: ところで今日は，自由度 $n-1$ の話題に関連して，$Y_1, \ldots, Y_n$ が独立に同じ正規分布に従うとき，平均 $\bar{Y}$ と偏差平方和 $SS = \sum (Y_i - \bar{Y})^2$ が独立に分布するということを話した。

T 君: そうでしたね。だから，分散分析の F 統計量も，本当に F 分布に従うのですね。今日は話がありませんでしたが，t 検定についても同様ですね。

M さん: ところで，正規分布とは仮定しないで，$Y_1, \ldots, Y_n$ が独立に同じ分布に従うとしよう。このとき，平均 $\bar{Y}$ と偏差平方和 SS が独立になるならば，もとの $Y_1, \ldots, Y_n$ は正規分布であることが言えるんだ。

T 君: えーっ。そんな話があるのですか。正規分布，恐るべしですね。

M さん: 興味があったら，Kagan *et al.* (1973) などを見てくれ。

T 君: たぶん，見ないとおもいますが。とにかく今日は，ありがとうございました。ちょっと自由度の話を訊きに来たのに，なんか，一元配置の分散分析まで分かったような気になりました。

M さん: わしも，こんなに長丁場になるとは思わなかった。まあ，統計で分からないことがあったら，気楽に訊きに来てもいいよ。分かることは答えるし，分からないことは分からないと，はっきり言うから。

T 君: はい，また来ます。ここは左の欄ですが，偶数ページ（16 ページ）で終わってよかったですね。

M さん: 後は，わしがなんとかしておこう。

T 君: 失礼します。

16 参考文献

Rao, C. R. (1965). *Linear Statistical Inference and Its Applications*, Wiley.

Kagan, A. M., Linnik, I. and Rao, C. R (1972). *Characterization Problems in Mathematical Statistics*, Wiley.

17 上級者向け演習問題

筆者: ここまで，お読みいただき，本当にありがとうございました。最後に，上級者の方々（大学院や統計のゼミで数理統計学を学んでいる方々）への演習問題を出しておきます。まずは，自分で考えてみてください。分からなければ，仲間と相談してください。それでも分からなければ，指導教官からヒントをもらって考えてください。

演習問題 1: 今回は，T 君が理解できるような方法で，偏差平方和や，その自由度の説明を行ないました。T 君は，1 変数の積分に関しては，変数変換の公式や，

$$\frac{1}{\sqrt{2\pi q^2}}\int_{-\infty}^{\infty}\exp\left[-\frac{(y-p)^2}{2q^2}\right]dy = 1$$

の関係は理解しているという想定です。また，$\sum$ の記号を使った式変形を根気よくフォローすることはできます。しかし，多重積分における変数変換のヤコビアンは，どうも良く理解していません。そのような T 君に対して，M さんは，うまく説明しているように見えます。しかし，今日の説明の中には一か所，M さんが手抜きをしている箇所があります。間違ったことを言っているわけではありません。数理統計学的にみると不十分なのです。それが，どこであるか指摘してください。また，T 君が理解できるような説明が可能かどうか考えてください。

演習問題 2: 第 8 節「正規変量の線形結合の分布」では，$Y_1, \ldots, Y_n$ が独立であるという前提で，数学的帰納法と，(16) 式の畳込み (convolution) を使って，線形結合 $X = a_1Y_1 + \cdots + a_nY_n$ が正規分布に従うことを説明しました。しかし，2 変数 Y_1, Y_2 が独立でない場合でも，その同時密度関数を $f_{12}(y_1, y_2)$ とすると，和 $X = Y_1 + Y_2$ の密度関数は

$$g(x) = \int_{-\infty}^{\infty} f_{12}(y_1, x - y_1)dy_1$$

と表わされます。そうすると，$Y_1, \ldots, Y_n$ が独立でない場合でも，$f_{12}(y_1, y_2)$ に (17) 式の 2 次元正規分布の同時密度関数を代入して計算すれば，数学的帰納法を使って，線形結合 $X = a_1Y_1 + \cdots + a_nY_n$ が正規分布に従うことを T 君に示すことができるように見えます。この議論は正しいのでしょうか。

演習問題 3: 第 12 節「数値計算の話」で，R における倍精度浮動小数点計算の例がいくつか出てきました。

R での倍精度浮動小数点計算

```
> options(digits=22)
> y <- seq(from=1000000.0, to=1000001.0, by=0.1)
> print( ss2 <- sum(y^2) - n*mean(y)^2 )
[1] 1.099609375
> x <- seq(from=10000000.0, to=10000010.0, by=1.0)
> print( ss4 <- sum(x^2) - n*mean(x)^2 )
[1] 110

> print( 0.4*0.4 == 0.16 )
[1] FALSE
> print( 0.5*0.5 == 0.25 )
[1] TRUE
```

なぜ，このような結果が得られるのかを説明してください。C 言語プログラム（あるいは他言語のプログラム）を書ける人は，計算機の中で何が行なわれているのかを確かめるプログラムを書いてください。

演習問題 4: 繰り返し数が不揃い ($n_i \neq n$) の一元配置分散分析モデル

$$\begin{aligned} y_{ij} &= \mu_i + e_{ij} \quad (1 \le i \le a; 1 \le j \le n_i) \\ & e_{ij} \sim N(0, \sigma^2) \\ & \bar{y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij},\ \bar{y}_{\cdot\cdot} = \frac{1}{\sum n_i}\sum_{i=1}^{a} n_i\bar{y}_{i\cdot} \end{aligned}$$

を考えます。このとき，処理平方和に関して，

$$SS_A = \sum_{i=1}^{a} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \sim \sigma^2\chi^2(a-1)$$

が成り立つことを，T 君にも分かるような方法で説明してください。

演習問題 5: （これは簡単です。）

処理平方和 $SS_A = \sum_{i=1}^{a} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$ について，対立仮説のもとでの期待値 $\mathrm{E}[SS_A]$ を求めてください。