統計学雑記

秋田大（北海道大）

平成 26 年 11 月 4 日

目次

1 拘束条件により減る自由度

推定および検定では「これこれは自由度何々の云々分布に従うから $\cdots$」という論法が繰り広げられる．この自由度と言うのがなかなかとっつきにくいもので，基本的には標本数から幾らか自然数を引いた値になる．例えば文献 [9] の χ^2 分布についての記述を見ると（式番号は筆者），

> χ^2 分布を用いると，正規母集団からの標本 $X_1, X_2, ..., X_n$ に基づく標本分散 s^2 の標本分布は，次のようにまとめられる．
>
> 標本分散を
>
> $$s^2 = \frac{1}{n-1}\left\{(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right\} \tag{1}$$
>
> とするとき，統計量
>
> $$\chi^2 \equiv (n-1)s^2/\sigma^2 \tag{2}$$
>
> は自由度 $n-1$ の χ^2 分布 $\chi^2(n-1)$ に従う．

と書かれている．ここで，χ^2 分布とはどのような分布であるかを書いておくと，$Z_1, Z_2, ..., Z_k \sim \mathrm{N}(0,1)$ が独立である時に，

$$\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_k^2 \tag{3}$$

が従う分布のことを**自由度 k の χ^2 分布**という．式 (1) と式 (3) は非常に似ているのであるが，なぜか標本分散を用いると自由度が 1 つ小さくなる．文献 [9] での大雑把な説明を引用すると（同じく式番号は筆者），

> μ が $\bar{X}$ におきかえられていることから
>
> $$(X_1 - \bar{X}) + (X_2 - \bar{X}) + \cdots + (X_n - \bar{X}) = 0 \tag{4}$$

という制限が効く．n 個の変数の和が常に 0（定数）でなければならないなら，第 n 変数は，第 1 変数から第 $(n-1$ 変数の $(n-1)$ 個の変数により，完全に 1 つの値に固定されざるを得ない（和=0 となるように）．したがって，自由に動ける変数（実質的な意味での変数）の個数は $n-1$ である．

と説明されている．また，別の頁には

> $\bar{X}$ は未知の母平均 μ の代わりに用いたもの（推定量ともいう．第 11 章参照）であるが，通常，推定をするとその個数 (1) だけ，自由度が減る．

とある．他の教科書やインターネット上の解説でも同様の説明が見られる[*1]．

しかしこの説明はなんともわかったような気には当初はさせてくれるが，よくよく考えるといささか腑に落ちないものがある．式 (4) で $\bar{X}$ を定数として扱っているが $\bar{X}$ は $X_1, X_2, ..., X_n$ から計算されるものではないのか？実際 $\bar{X} = \sum X_i/n$ を代入すると式 (4) はただの恒等式であり制限として有効であるようには思えない．そもそも標本を取ってくるときには $\bar{X}$ という値は存在しておらず，それこそ n 個の標本 $X_1, X_2, ..., X_n$ を自由に取ってきたのではないのか？大多数の人が上のような説明で納得できたとしても，やはり筆者のように疑問を感じる人は少なからずいるのは事実である．[5] では，母分散の推定量として不偏分散を計算する際に偏差平方和を $n-1$ で割る理由について，単に拘束条件で自由度が減るという説明では，同様の論法で偏差平方和を $n+1$ で割るという説明もできてしまうことが指摘されている．

では結局のところこの自由度が減少する理由はいかにして説明されるのであろうか．[5] ではその原因が偏差平方和と χ^2 分布の関係にあるとし，偏差平方和を母分散で割ったものが自由度 $n-1$ の χ^2 分布に従うことを数学的に示している．[5] と同様の理屈で，一般に拘束条件が k 個ある時に自由度が $n-k$ になることが示せる．方針としては，式 (3) が使用できるように $X_1, X_2, ..., X_n$ から $n-k$ 個の新しい独立な統計量を作り出すということになる．統計量 Θ は

$$\Theta = \begin{pmatrix} X_1 & X_2 & \cdots & X_n \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \mathbf{X^T X} \tag{5}$$

と書き表されているとする[*2]．ここで $X_1, ..., X_n$ に対して

$$C\mathbf{X} = 0 \tag{6}$$

と表現される $\mathrm{rank}\ C = k < n$ の拘束条件があるとする[*3]．つまり実際に考えなくてはならない空間は $\mathrm{Ker}\ C$ である．$\mathrm{Ker}\ C$ の基底を $\mathbf{a}_1, ..., \mathbf{a}_{n-k}$ として，

$$\begin{aligned} \mathbf{X} &= (\mathbf{a}_1, ..., \mathbf{a}_{n-k}) \begin{pmatrix} Y_1 \\ \vdots \\ Y_{n-k} \end{pmatrix} \\ &= A\mathbf{Y} \end{aligned} \tag{7}$$

とすると，Θ は

$$\Theta = \mathbf{Y}^T A^T A \mathbf{Y} = \mathbf{Y}^T B \mathbf{Y} \tag{8}$$

[*1] なぜ自由度が小さくなるかについて教科書で何も説明されていない場合も多々ある

[*2] 偏差平方和なら $X_1, ..., X_n$ は $\bar{X}$ で引かれていると考える

[*3] $\sum X_i = 0$ なら $C = (1\ 1\ \cdots\ 1)$ である．

と書ける．B は実対称行列なので，正規直交行列 U による対角化ができて

$$\begin{aligned}\Theta &= \mathbf{Y}^T U \Lambda U^T \mathbf{Y} = \left(U^T\mathbf{Y}\right)^T \Lambda \left(U^T\mathbf{Y}\right)^T \\ &= \left(\mathbf{u}_1\cdot\mathbf{Y},...,\mathbf{u}_{n-k}\cdot\mathbf{Y}\right)^T \Lambda \left(\mathbf{u}_1\cdot\mathbf{Y},...,\mathbf{u}_{n-k}\cdot\mathbf{Y}\right) \\ &= \left(\sqrt{\lambda_1}\mathbf{u}_1\cdot\mathbf{Y}\right)^2 + \cdots + \left(\sqrt{\lambda_{n-k}}\mathbf{u}_{n-k}\cdot\mathbf{Y}\right)^2 \\ &= \mathbf{W}_1^2 + \cdots + \mathbf{W}_{n-k}^2 \end{aligned} \tag{9}$$

となる．ここで，$\mathbf{u}_1,...,\mathbf{u}_{n-k}$ は U をつくる B の固有ベクトルであり，各固有値を $\lambda_1,...,\lambda_{n-k}$ としている．こうして新しくできた統計量 $\mathbf{W}_1,...,\mathbf{W}_{n-k}$ は，$i\neq j$ について

$$\begin{aligned}\mathrm{Cov}(\mathbf{W}_i,\mathbf{W}_j) &= \sqrt{\lambda_i\lambda_j}\mathrm{Cov}(\mathbf{u}_i\cdot\mathbf{Y},\mathbf{u}_j\cdot\mathbf{Y}) \\ &= \sqrt{\lambda_i\lambda_j}V(\mathbf{Y})\mathbf{u}_i\cdot\mathbf{u}_j = 0\end{aligned} \tag{10}$$

となるので，各々独立である．以上より，統計量 Θ が独立な $n-k$ 個の確率変数で表されていることが証明された．

2　標本調査の標本数公式

世論調査では通常人口の一部に対してアンケートを行い，その結果から母集団である調査対象全体について推測する．このように母集団の一部から観測値を得ることを**標本調査**という．一方，国勢調査など母集団全体に対する調査は**全体調査**あるいは**悉皆調査**と呼ばれる．全体調査で得られた平均や分散などの統計量はまさにその母集団の統計量となるが，標本調査では推測を行わなくてはならず，それゆえ誤差が生じる．そこで調査結果として所望の誤差を実現するためには標本数がどれだけ必要かを標本調査の前には考えなくてはならない．

この問題を整理すると，まず確率変数 X として表されるデータを母集団 Ω から n 個抽出し，X の分布の統計量 μ を推定した値が $\hat{\mu}$ であるとする．このとき，その $100-\alpha\%$ 信頼区間 $[\hat{\mu}_-,\hat{\mu}_+]$ が，設定された誤差を $\varepsilon>0$ として

$$\hat{\mu}-\varepsilon\leq\hat{\mu}_-\leq\hat{\mu}\leq\hat{\mu}_+\leq\hat{\mu}+\varepsilon \tag{11}$$

となるような n を求めるということになる．よって $\hat{\mu}-\hat{\mu}_-$ および $\hat{\mu}_+-\hat{\mu}$ を数式で表し，その式を n で解けば問題は解決される．この問題は抽出方法が復元抽出であるか非復元抽出であるかで難しさが大きく変わる．非復元抽出の場合には抽出したデータ $X_1,X_2,...,X_n$ が独立ではなく，$\mathrm{Cor}(X_i,X_j)\neq 0$ であるため数式の計算が複雑になるのである．[4] では，正規分布 $\mathrm{N}(\mu,\sigma^2)$ に従うデータを大きさ N の有限母集団から n 個非復元抽出するときの標本平均の分散の期待値 $V(\bar{X})$ の公式：

$$V(\bar{X}) = \frac{N-n}{N-1}\frac{\sigma^2}{n} \tag{12}$$

がとてもややこしい式変形の結果として導かれている．よって $\alpha\%$ 信頼区間に誤差 ε を含めるためには，標準正規分布の上側 $\alpha/2\%$ 点を $Z_{\alpha/2}$ として

$$\begin{aligned}\varepsilon &\geq Z_{\alpha/2}\sqrt{V(\bar{X})} \\ \varepsilon^2 &\geq Z_{\alpha/2}^2 V(\bar{X}) \\ &= Z_{\alpha/2}^2\frac{N-n}{N-1}\frac{\sigma^2}{n}\end{aligned} \tag{13}$$

となればよい．これを n について解くと

$$n \geq \frac{N}{\left(\frac{\varepsilon}{Z_{\alpha/2}\sigma}\right)^2 (N-1)+1} \tag{14}$$

となる．式中に母分散 σ^2 が含まれているため，何らかの方法で母分散を知っておくか，予備調査でおおよその目安をつけておくことが必要である．

報道機関がよく調査している政党支持率などの Yes/No で回答するタイプのデータについては，調査数を n，真の政党支持率を p とした二項分布 $\mathrm{Bi}(n,p)$ に従う．必要な調査数 n は，二項分布を正規分布として近似し，二項分布の分散 $\sqrt{p(1-p)}$ を式 (14) の σ^2 に代入して得られる

$$n \geq \frac{N}{\left(\frac{\varepsilon}{Z_{\alpha/2}}\right)^2 \frac{N-1}{p(1-p)}+1} \tag{15}$$

という公式で算出される．この公式でも母集団のパラメータ p が含まれているため，何か目安を用いるか，$p(1-p)$ が最大になる $p=0.5$ で n を計算しておくという方法がとられる．

式 (15) は医学，社会科学等で頻繁に使われているが，このように各項がどのように数学的に導かれているかまで理解している人はほぼいないと思われる．そしてついには各項の意味が理解できずこの公式自体を疑い出す人も中にはいる [8][3]．特に [8] では，式 (15) の分母の $+1$ について，

> この 1 がなければ，k（信頼度係数）より誤差範囲を小さな値にすると「サンプリング数が母集団より大きくなる」と言うマヌケな状況になってしまうので，それを避けるために 1 を足しているんじゃないだろうか（違うかもしれないけど・・・）。

と書いている．さらに，$N-1/p(1-p)$ の $N-1$ に至っては，

> N の大きさと無関係に 1 だけ引くなんて，いかにも「この数式には裏付けがあるんです」的な何だか姑息なものを感じちゃうのである。

と述べている．おそらく医学系での数式が経験則によるものが多いことから，この公式についても経験的に当てはまっているから使われている式であると勘違いしたのではないだろうか．一応二項分布を正規分布に近似するという微妙な部分はあるが，標本数を母集団より小さくするために $+1$ をしているのではなく，標本数が母集団より小さいことを前提にしているから自然とそうなっているのであって，またこの公式は人の意思と関係なく式変形で得られた純粋な裏付けがあるのである．

3 ノンパラメトリックな回帰

2 変数間の直線関係を見出す際にはたいてい最小二乗法による線形回帰分析が行われる．すなわち，観測された変数 $x_i, y_i (i=1,2,...,N)$ について

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{16}$$

というモデルを仮定し，誤差 ε_i の二乗和 $\sum \varepsilon_i^2$ が最小になるように α, β を求めることになる．この最小二乗法によるパラメータの推定において，実際には次のような誤差 ε_i の仮定がある：

1. 不偏性

$$E(\varepsilon_1) = E(\varepsilon_2) = \cdots = E(\varepsilon_N) = 0 \tag{17}$$

2. 等分散性

$$V(\varepsilon_1) = V(\varepsilon_2) = \cdots = V(\varepsilon_N) = \sigma^2 \tag{18}$$

3. 無相関性

$$i \neq j \Rightarrow \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \tag{19}$$

この仮定の下では，y_i の線型結合で表される推定量として，最小二乗法で推定されるパラメータは最良のもの（最良線型不偏推定量）となる [10]．さらに

(4) 正規性[*4]

$$\varepsilon_i \sim \mathrm{N}(\mu, \sigma^2) \tag{20}$$

が仮定できると，y_i の線型結合で表される推定量に限らず，任意の推定量の中で最良な推定量であることが示される．もちろん最小二乗法による計算自体はこの仮定がなくてもすることはできる．ただその場合には他の推定量の方がより良い推定量になることがあるのである．例えば，誤差 ε がコーシー分布

$$f(\varepsilon) = \frac{1}{\pi}\frac{1}{1+\varepsilon^2} \tag{21}$$

に従う場合では，そもそも $E(\varepsilon), V(\varepsilon)$ が存在せず，仮定 (1),(2) が満たされなくなる．誤差 ε_i が $\mu = 0$ の正規分布に従うとしても，その分散が i によって変化する場合には (2) の仮定は妥当ではなくなってしまう．分散が i に依存する場合というのは例えば，x の値が大きくなると y のバラつきが大きくなる場合などが挙げられ，実際にはそんなに目にしないわけではない．x のスケールが何桁にもなるような場合にはそれこそ y のバラつきが一定であるというようなことは少ないであろう．

分布に関する仮定が当てはまらない時には最小二乗法や平均値によるパラメータの推定よりも中央値など別の推定量がより良い推定量となることがある．この代替として使われる推定量は分布に関する仮定が少ない，あるいは仮定がないという性質がある．このように分布関する仮定をあまり置かない統計的手法のことを**ノンパラメトリック**な手法という．一方最小二乗法のように一般的に分布に仮定を置く手法のことを**パラメトリック**な手法という．

回帰直線を求める場合にもノンパラメトリックな手法が提案されており，その一つが **Passing-Bablok 法（PB 法）**である．この手法は本来 2 種類の計測手法の相同性を調べる場合を想定している [19]．つまり，2 種類の計測手法により，真値 x_i°, y_i° が誤差 ξ_i, η_i を伴って

$$x_i = x_i^\circ + \xi_i \tag{22}$$
$$y_i = y_i^\circ + \eta_i \tag{23}$$

としてそれぞれの計測結果 x_i, y_i が得られたとする．この時，全ての i に共通なあるパラメータ α, β により

$$y_i^\circ = \alpha + \beta x_i^\circ \tag{24}$$

という真値の関係があったとして，$\alpha = 0, \beta = 1$ であるかどうかを調べるのが Passing-Bablok 法である．Passing-Bablok 法では誤差 ξ_i, η_i の分布に関する仮定が少なく，誤差が正規分布でなくてもよいし，x や y に依存する誤差であってもよい．とはいえ，この手法の発表論文 [19] では

$$\forall i \quad \frac{V(\xi_i)}{V(\eta_i)} = \beta^2 \tag{25}$$

[*4] (1) が仮定されていれば $\mu = 0$．また (2) が仮定されていればこの σ^2 は (2) のものと同じ

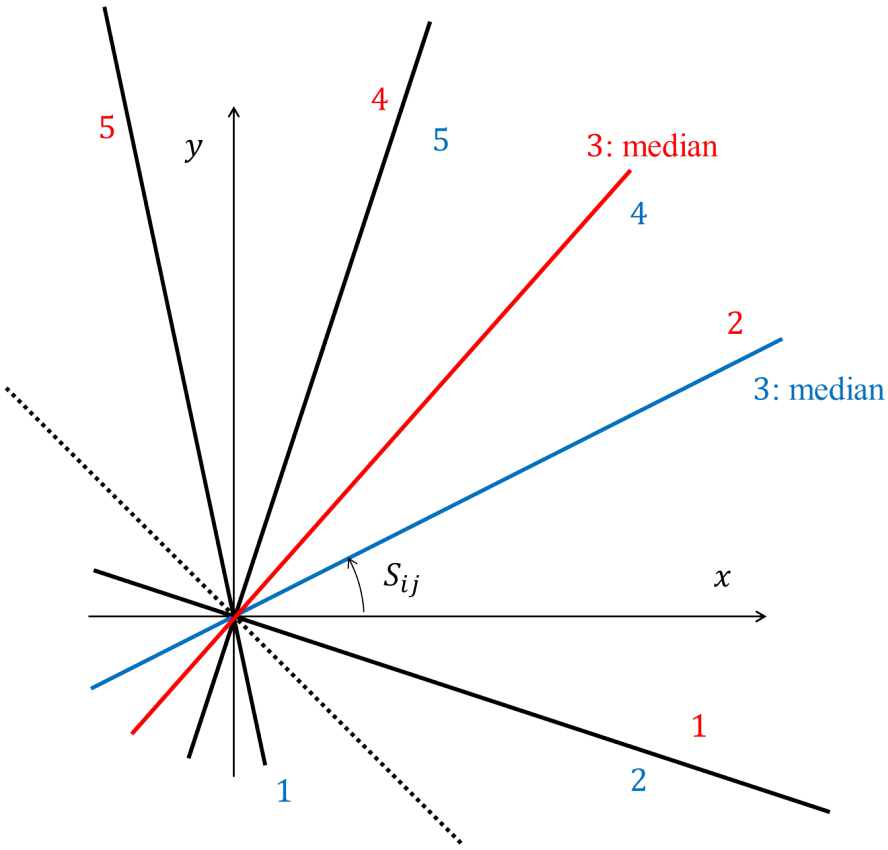

図 1　Passing-Bablok 法における傾きの並び替え：$-90°$ から $90°$ で並び替えると青文字のように中央値の傾きが 1 より小さいものになってしまうが，$-45°$ から $135°$ に並び替えた赤文字の系列では，中央値が傾き 1 のものになっている．

という条件が仮定されており，本当に任意の分布でよいわけではないようである．

Passing-Bablok 法がノンパラメトリックである所以は β を中央値として推定する所にある．まずあらゆる $i \neq j$ の組に対して $(x_i, y_i), (x_j, y_j)$ の傾き S_{ij} を求める．S_{ij} は $-\infty$ から ∞ までの値をとり得るが，これは角度に直すと $-90°$ から $90°$ となる．よってもし $\beta = 1$ であり S_{ij} が 1 の前後に対称に分布していたとすると，このまま S_{ij} の中央値を取ったのでは図 1 のようにその値が 1 より小さくなってしまう恐れがある．そこで，S_{ij} を角度にして $-45°$ から $135°$ の間に並べ替え，この中央値 b により β の推定値とする．この並び替えは正しく傾きを見積もるだけでなく，2 種類の計測法のどちらを x_i にしても同じ推定値を出すことにもつながっている．α の推定には $y_i - bx_i$ の中央値 a を用いる．パラメータの信頼区間の計算には，Passing と Bablok による解析的な手法 [19] の他，ジャックナイフ法やブートストラップ法が使われているようで，R のライブラリ `mcr` で Passing-Bablok 法により回帰を行う場合には，どの手法で信頼区間を出すかオプションで指定することができる．α の信頼区間が 0 を含み，β の信頼区間が 1 を含むならば，$y_i^\circ = x_i^\circ$，つまり 2 種類の手法が同じ物理量の計測になっていると言ってよいだろうと結論付けられる．

このように，Passing-Bablok 法は傾き 1 が想定される場合の回帰である．そのため線形モデルのように係数の値そのものが重要な場合にこの方法を適用するのはおそらく誤りであろう．

参考文献

[1] C. ラダクリシュナ ラオ, “統計的推測とその応用”, 東京図書 (1992).
統計の数学理論が基礎から応用まで書かれている． 難しい.

[2] C. Clopper and E. S. Pearson, “The use of confidence or fiducial limits illustrated in the case of the binomial”, Biometrika, pp. 404–413 (1934).
二項分布のパラメータ推定における信頼区間の話.

[3] “統計学のサンプル数 2000 の根拠は？”, http://oshiete.goo.ne.jp/qa/1552827.html.

[4] 小杉考司, “なぜ不偏分散は $n-1$ で割るのか”, http://www.kosugitti.sakura.ne.jp/wp/wp-content/uploads/2013/08/est.pdf.
不偏分散の式と有限母集団からのサンプリングについて詳細な式展開で説明されている．不偏分散の式の導出は探せばわりと見つかるが，有限母集団から非復元抽出を行う時のサンプル平均および分散の計算について事細かに計算されているのは珍しい.

[5] 三輪哲久, “自由度 n-1”, http://cse.niaes.affrc.go.jp/miwa/ja/niaes-stat-gis/miscellaneous/StatGIS_df_n-1.pdf.
統計の自由度の概念のとても納得できる解説.

[6] 竹内啓, “数理統計学の方法的基礎”, 東洋経済新報社 (1973).
推測や検定など統計を実際に使う際における細かい問題を厳密に考察した本.

[7] 竹内啓, 藤野和建, “2 項分布とポアソン分布”, UP 応用数学選書 2, 東京大学出版会 (1981).

[8] 夏井睦, “世論調査に見る統計処理”, http://www.wound-treatment.jp/next/wound225.htm.

[9] 東京大学教養学部統計学教室編, “統計学入門”, 基礎統計学 I, 東京大学出版会 (1991).
統計学の非常にいい教科書． 他の統計入門書に比較するとやや分厚いが，基礎から詳しく書いてあるためおすすめである.

[10] 東京大学教養学部統計学教室編, “自然科学の統計学”, 基礎統計学 III, 東京大学出版会 (1992).
[9] よりも高度な内容になっており，社会調査や心理実験データが中心の [11] に対して工業製品のデータや物理実験のデータなどが対象になっている.

[11] 東京大学教養学部統計学教室編, “人文・社会科学の統計学”, 基礎統計学 II, 東京大学出版会 (1994).
書名の通り人文・社会科学での実践的な内容になっている．しかし数式は免れない．自然科学の人間にはおそらく不要.

[12] 永田靖, 吉田道弘, “統計的多重比較法の基礎”, サイエンティスト社 (1997).
多重比較法だけに集中してその概念と 様々な手法を解説している.

[13] 稲垣宣生, 山根芳知, 吉田光雄, “統計学入門”, 裳華房 (1992).

[14] 稲毛敏宏, “Passing-bablok 法のアルゴリズム”, http://www.kms.ac.jp/~clinilab/person/ing/lib/method.html.
Passing-Bablok 法の日本語の解説.

[15] 竹内啓, “数理統計学 - データ解析の方法”, 東洋経済新報社 (1963).

[16] 竹村彰通, “現代数理統計学”, 創文社現代経済学選書 8, 創文社 (1991).

[17] 菅民郎, “らくらく図解 アンケート分析教室”, オーム社 (2007).

[18] 食品総合研究所, “サンプル数の理論的決め方”, http://www.naro.affrc.go.jp/org/nfri/

yakudachi/sampling/pdf/logical-sample-number.pdf.
サンプル数決定のための公式がある.

[19] H. Passing and W. Bablok, “A new biometrical procedure for testing the equality of measurements from two different analytical methods. application of linear regression procedures for method comparison studies in clinical chemistry, part i”, Clinical Chemistry and Laboratory Medicine, **21**, 11, pp. 709–720 (1983).
2 種類の計測手法の相同性を調べるノンパラメトリックな手法である Passing-Bablok 法の論文.

[20] H. Passing and W. Bablok, “Comparison of several regression procedures for method comparison studies and determination of sample sizes application of linear regression procedures for method comparison studies in clinical chemistry, part ii”, Clinical Chemistry and Laboratory Medicine, **22**, 6, pp. 431–445 (1984).
[19] の続き. 2 種類の計測手法の相同性を調べる手法として, 変動係数の違いや外れ値の存在でどの手法が良いかが論じられている.