JILPT Discussion Paper 05-001
2005 年 1 月

層化抽出標本に適用できる1つの検定方法

独立行政法人　労働政策研究･研修機構
情報解析部長　　本川　明

《要旨》

層化抽出され、ウェイトバック集計された統計調査については、通常のカイ2乗検定等が有効でない。本稿では、このような統計調査にも有効なワルド検定を応用した検定方法を紹介する。これは、別の研究のためのツールとして開発されたものだが、一般の統計調査にも適用できるものであり、パソコンの汎用プログラムとして実装された。

（備考）　本稿のとりまとめにあたって、平田周一氏（労働政策研究･研修機構）、石井太氏（厚生労働省）から貴重なコメントをいただいた。もちろん、あり得べき誤謬は執筆者個人に属するものである。なお、本稿は、執筆者個人の責任で発表するものであり、独立行政法人　労働政策研究･研修機構としての見解を示すものではない。

層化抽出標本に適用できる1つの検定方法

目次

第1 章 はじめに

（他の研究の副産物）

本稿では、層化抽出されウェイトバック集計された統計調査について、検定を行うための方法を紹介する。

これは、労働政策研究･研修機構から別途発刊される「インターネット調査は社会調査に利用できるか」（労働政策研究報告書 No.17 2004、以下「インターネット調査研究」と言う）の研究遂行上の必要から生まれたものである。「インターネット調査研究」では、複数の調査で同じ質問をして、その結果に違いがあるかどうかが調べられた。その際に、集計値に表れた調査間の差異が果たして偶然によるものかそうでないのかをどのようにして確かめたらいいのか、ということが問題になったのである。

（普通のカイ2乗検定が使えない）

普通、このような場合には、（ピアソンの）カイ2乗検定が用いられる。例えば、A 調査、B 調査という2種類の調査で「*これからの日本が目指すべき社会のあり方として、あなたのお考えはこのうちど*

れに近いでしょうか」という質問をして、その回答数が次のような分布になったとする。これは、「インターネット調査研究」の実際のデータである。

	計	貧富の差の少ない平等社会	意欲や能力に応じて自由に競争できる社会	どちらともいえない	わからない
計	4,764	884	2,761	912	207
(計の構成比)	(1.000)	(0.186)	(0.580)	(0.191)	(0.044)
A 調査	2,374	432	1,351	467	124
B 調査	2,390	452	1,410	445	83

そこで、「カイ2乗統計量」χを次のように計算する(表の数値は四捨五入されているので結果は手計算と一致しない)。

$$\chi = \frac{(432 - 0.186\times 2374)^2}{0.186\times 2374} + \frac{(1351 - 0.580\times 2374)^2}{0.580\times 2374} + \frac{(467 - 0.191\times 2374)^2}{0.191\times 2374} + \frac{(124 - 0.044\times 2374)^2}{0.044\times 2374}$$
$$+ \frac{(452 - 0.186\times 2390)^2}{0.186\times 2390} + \frac{(1410 - 0.580\times 2390)^2}{0.580\times 2390} + \frac{(445 - 0.191\times 2390)^2}{0.191\times 2390} + \frac{(83 - 0.044\times 2390)^2}{0.044\times 2390}$$
$$= 10.250$$

もし、いずれの調査も単純無作為抽出でサンプリングされており、かつ、単純集計されているならば、このχは漸近的に自由度3のカイ2乗分布に従う。したがって、$\chi = 10.250$を自由度3のカイ2乗分布に当てはめて、有意確率=0.017<0.05 を得ることから、有意水準 5%で調査間の差が有意と「検定」される。

しかし、「インターネット調査研究」で使われた調査は、実際には性・年齢階級による層化抽出によるものであり、また、集計に当たってはウェイトバックが行われた。層化抽出による調査の場合、上のようにして計算されたχは、漸近的にもカイ2乗分布に従わない。したがって、上のような「検定」は無効である(注)。

(注) もし、1より大きな復元倍率を用いてウェイトバックされたならば、サンプルサイズが水増しされたのと同様の効果を生むから、通常のカイ2乗検定が無効であることは容易に分かる。しかし、無効であることの理由はそれだけではない。層化抽出という抽出方法そのものが持つ性質と、ウェイトバックする場合の復元倍率のばらつきとが、より根本的な理由である。

以上のことは、次のように考えると理解しやすい。集計結果の差が有意かどうかは、おおむね、①その差が大きいかどうか、及び②調査の精度が高いかどうか、に左右される。差が大きく精度が高いほど有意になりやすい。一般に、均一の抽出率で層化抽出された調査は、単純抽出された調査より精度が高く、また、ウェイトバックされた集計は、単純集計より精度が低い。通常のカイ2乗検定は、このような状況を反映する仕組みになっていないので、層化抽出されウェイトバックされた調査に対し無効なのである。

(層化抽出に対応した検定方法)

本稿では、層化抽出及びウェイトバックに対応した検定方法が紹介される(後出の命題 2 と命題

3)。「インターネット調査研究」では、これらの手法が使われた。 上の A 調査、B 調査の例を命題 2 の手法で検定すると、有意確率=0.19 であって、調査間の差はほとんど有意でないことが判明する(「3.3 計算例」を参照)。

手法の発想は単純である。通常のカイ2乗統計量は上の例のような平方和を計算するのだが、本稿で示される方法は、平方和を計算する前に、

「漸近的に分散が 1 で互いに無相関になるように各項を線形変換する」

というものである。

(既存の検定手法との関係)

本稿の手法(補題 4)は、ワルド検定(参考 8)の一種である。検定対象とする統計量の分散共分散行列の推計値を用いて、漸近的にカイ2乗分布に従う統計量を構成している。ワルド検定を用いると、分散共分散行列の推計値さえ得られれば単純無作為抽出でなくとも検定を実行できる。ただ、本稿の方法と通常のワルド検定との見かけ上の大きな違いは、ワルド検定が正則な分散共分散行列を前提にしているのに対して、本稿の手法は、これが非正則行列であることを前提にしている点である。計算を遂行する観点からすると、ワルド検定が分散共分散行列の逆行列を計算するのに対して、本稿の手法は、これの対角化を実行する。これは本質的な違いではない。検定対象とするパラメータから従属関係にあるものをあらかじめ省いておけば、正則行列のケースに帰着できるからである。

(インプリケーション)

このように、本稿で紹介する手法は、それ自体目新しいものではなく、統計の専門家の間では周知のことである。ただ、「層化抽出されウェイトバック集計された統計の検定」という特殊ケース(しかし実際の場面で極めて頻繁に遭遇するケース)について、計算式を具体的に示すことには意義があると思われる。

ウェイトバック集計に対しては、その是非も含めて様々な議論がある。しかし、現実には、層化抽出されウェイトバック集計された調査が数多く存在する。層化抽出に対応した検定を行った例としては、「SSM95 調査」を用いた中村[5]がある。しかし、労働の分野では、層化抽出された統計に対応した検定がそれほど広く普及しているように見えない。その原因の一端は、統計の専門家以外にはワルド統計量等に基づく検定が一見面倒そうに見えることにあるように思われる。

実際には、計算はそれほど面倒でない。また、あらかじめパソコンの集計ソフトに組み込んでおけば、通常のカイ2乗検定と大差ない手間暇で実行できるものである。こうした事実が周知されるならば、層化抽出された調査に対して、本稿のような方法が広く使われる可能性もあると思われる。なお、通常のカイ2乗検定が本稿の方法(あるいはワルド検定)と比べてどの程度ずれるかについ

ては、一概に言えない。ただ、一般論として、回答傾向に層間の違いが大きいほど、また、集計倍率に層間の格差が大きいほど、検定結果の食い違いが大きいと考えられる。

冒頭に述べたように、本稿の手法は、2つの調査を比較することを念頭に構成したものである。しかし、この手法は、調査間の比較だけではなく、男女比較など1つの調査の中の異なる属性間の比較に使うこともできる。実際にはこうした使われ方の方が多いかもしれない。

(本稿の構成)

本稿は、4 章に分かれているが、中心になるのは第 3 章である。とくに、「3.3 計算例」だけで検定方法の概要が分かるようになっている。第 2 章は第 3 章のための準備であり、第 4 章は補足である。

第 2 章　モデルの定義

2.1 記号と基本的特性

本稿では、「*次の選択肢から当てはまるものをひとつ選んでください*」というような、選択肢への択一回答を求める調査を想定する。

各選択肢への回答数(後出の S_{il})は、多項分布に従うとする。標本は層化抽出によるものとし、集計は、層ごとに設定された復元ウェイトを乗じて行われるものとする。これらを表現するための記号等を次のように定義する。

k　正の整数(選択肢の個数)

L　正の整数(層の個数)

n_l　$(l=1,...,L)$　正の整数(第 l 層の回答数)

w_l　$(l=1,...,L)$　正数(第 l 層の復元ウェイト)

p_{il}　$(i=1,...,k;l=1,...,L)$　$p_{1l}+...+p_{kl}=1$ なる正数(第 l 層で第 i 選択肢が選ばれる確率)

S_{il}　$(i=1,...,k;l=1,...,L)$　次のような多項分布に従う確率変数(第 l 層の第 i 選択肢への回答数)

$$\sum_{i=1}^{k} S_{il} = n_l \quad (l=1,...,L)$$

n　次により定義される数(回答数)

$$n=\sum_{l=1}^{L} n_l$$

N　次により定義される数(ウェイト付き回答数)

$$N = \sum_{l=1}^{L} w_l n_l$$

p_i　次により定義される数（第 i 選択肢を選ぶ確率）

$$p_i = \frac{\sum_{l=1}^{L} w_l n_l p_{il}}{\sum_{l=1}^{L} w_l n_l} = \frac{1}{N} \sum_{l=1}^{L} w_l n_l p_{il}$$

$S_i \quad (i = 1, \ldots, k)$　次により定義される確率変数（第 i 選択肢へのウェイト付き回答数）

$$S_i = \sum_{l=1}^{L} w_l S_{il}$$

$T_i \quad (i = 1, \ldots, k)$　次により定義される確率変数

$$T_i = \frac{S_i - N p_i}{\sqrt{n}}$$

$\mathbf{T}$　T_i を縦に並べた確率ベクトル

$$\mathbf{T} = \begin{bmatrix} T_1 \\ \vdots \\ T_k \end{bmatrix}$$

2.2　n を大きくする

上で定義した各種の確率変数について、この後、回答数 n を大きくしていった場合の状況を記述していくことになる。その準備として、「n を大きくする」ことと、その層別の内訳である各 n_l の変動との関係を明確にしておく必要がある。

「$n_1 : n_2 : \ldots : n_k$ の比率を一定に保つ形で各 n_l が大きくなる」とできれば簡単だが、各 n_l が整数であることと「比率を一定に保つ」こととは両立しない。そこで、次のように $n_1, \ldots, n_L$ を定めるものとする：

$\alpha_1, \ldots, \alpha_L$ を $\alpha_1 + \ldots + \alpha_L = 1$ なる正数とし、

$n_l = \alpha_l n$ の四捨五入値 $(l = 1, \ldots, L-1)$

$n_L = n - (n_1 + \ldots + n_{L-1})$

明らかに次のことが成立する。

$$\lim_{n \to \infty} \frac{n_l}{n} = \alpha_l \quad (\text{for } l = 1, \ldots L)$$

なお、「四捨五入」というのは厳密な要件ではない。これを「切り上げ」、「切り捨て」、あるいは10や100などの切れの良い値に丸めたとしても、上の極限値は変わらない。また、丸め誤差を上記のように n_L のみで調整するのではなく、$n_1, \ldots, n_L$ に分散させて調整しても同じことである。

n の増大にともなって N が増大、w_l が一定、という設定から分かるように、これは無限母集団を前提としている。N は、母集団サイズではなく、ウェイト調整後の集計値である。また、w_l は、母集団への復元倍率ではなく、単に層間のウェイトを調整するだけの係数である。有限母集団を前提とした場合にどう扱うかは、今後の課題である。

2.3 ウェイトバックカイ2乗統計量(確率が与えられている場合)

この項では、ウェイトバックされた集計値によるカイ2乗統計量を定義する(命題 1)。これは、ワルド統計量の一種である(後出補題4の注を参照)。

命題1 (ウェイトバックカイ2乗検定～確率が与えられている場合～)

k 行 k 列の行列 $\mathbf{C}=(c_{ij})$ を次のように定める。

$$c_{ii}=\frac{1}{n}\sum_{l=1}^{L} w_l{}^2 n_l p_{il}(1-p_{il}) \quad (\text{for } i=1,\ldots,k)$$

$$c_{ij}=-\frac{1}{n}\sum_{l=1}^{L} w_l{}^2 n_l p_{il} p_{jl} \quad (\text{for } i\neq j)$$

このとき、

(1) $\mathbf{BC}{}^t\mathbf{B}=\begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{bmatrix}$ (対角要素以外は0)

となる正則行列 $\mathbf{B}$ が存在する。

(2) $Chi = {}^t\mathbf{T}{}^t\mathbf{BBT}$

と置けば、Chi の分布は、$n\to\infty$ のとき、自由度 $k-1$ のカイ2乗分布に弱収束する。

(3) Chi の値は、$\mathbf{B}$ の選び方によらない。すなわち、別の正則行列 $\mathbf{B}'$ により

$$\mathbf{B}'\mathbf{C}{}^t\mathbf{B}'=\begin{bmatrix} \mathbf{1} & \\ & 0 \end{bmatrix}$$ ($\mathbf{1}$ は $k-1$ 行 $k-1$ 列の単位行列)

となったとすると、

$${}^t\mathbf{T}{}^t\mathbf{B}'\mathbf{B}'\mathbf{T}={}^t\mathbf{T}{}^t\mathbf{BBT}$$

である。

(証明)

(1)と(2)について

後出の補題 4 を適用することとし、① $\mathbf{T}$ の分布が多変量正規分布に弱収束すること、② $\mathbf{C} = \mathrm{V}(\mathbf{T})$ であること、③ $\mathbf{C}$ の階数が $k-1$ であることを示せばよい。

①については、$\mathbf{T}$ が多項分布 S_{il} $(i = 1,...,k; l = 1,...,L)$ の線形結合であることから明か。②については、多項分布 S_{il} の分散共分散が

$$\mathrm{V}(S_{il}) = n_l p_{il}(1 - p_{il})$$
$$\mathrm{Cov}(S_{il}, S_{jl}) = -n_l p_{il} p_{jl}$$

であることを用いて $\mathbf{T}$ の定義式から $\mathrm{V}(\mathbf{T})$ を計算すれば、容易に確かめることができる。③については、次の補題 1 による。

(3)については、後出の補題 2 による。

（命題 1 の証明終わり）

補題 1 （$\mathrm{V}(\mathbf{T})$ の固有値）

$\mathbf{T}$ の分散共分散行列 $\mathrm{V}(\mathbf{T})$ の固有値のうち、1 個は 0 であり、他の $k-1$ 個は正である。

（証明）

$\mathrm{V}(\mathbf{T})$ は分散共分散行列だから、その固有値は 0 又は正である。したがって、$\mathrm{V}(\mathbf{T})$ の階数が $k-1$ であることさえ示せばよい。そのために後出の補題 8 を適用することとして、①各 T_i が確率 0 の値を取らないこと、② $\mathbf{T}$ が張るベクトル空間の次元が $k-1$ であること、を示す。

（T_i が確率 0 の値を取らないこと）

定義により、各 S_{il} は 1 から n_l までの整数値をとり、それぞれの値をとる確率は 0 でない。T_i は、S_{il} の線形関数で表されるので、確率 0 の値をとることはない。

（$\mathbf{T}$ が張るベクトル空間の次元 $\leq k-1$ であること）

$$\sum_{i=1}^{k} T_i = \sum_{i=1}^{k} \frac{S_i - Np_i}{\sqrt{n}} = \frac{1}{\sqrt{n}}\left(\sum_{i=1}^{k} S_i - N\sum_{i=1}^{k} p_i\right) = \frac{N - N \cdot 1}{\sqrt{n}} = 0$$

だから、$T_1,...,T_k$ は一次従属である。したがって、$\mathbf{T}$ が張るベクトル空間の次元 $\leq k-1$ である。

（$\mathbf{T}$ が張るベクトル空間の次元 $\geq k-1$ であること）

事象 B_i を次のように定義する（$1 \leq i \leq k$）。

$$\begin{cases} S_{il} = n_l \\ S_{jl} = 0 \ \ (\text{f or } \ i \neq j) \end{cases}$$

これは「すべての回答が i に集中する」という事象に相当する。

事象 B_i の下で、$\mathbf{T}$ は次の値 $\mathbf{t}_i$ をとる。

$$\mathbf{t}_i = \frac{N}{\sqrt{n}} \begin{bmatrix} -p_1 \\ \vdots \\ -p_{i-1} \\ 1-p_i \\ -p_{i+1} \\ \vdots \\ -p_k \end{bmatrix}$$

後出の補題 7 によりこれらの $\mathbf{t}_i$ を並べた行列の階数が $k-1$ だから、$\mathbf{T}$ が張るベクトル空間の次元 $\geq k-1$ である。

（補題 1 の証明終わり）

補題 2 （${}^t\mathbf{T}\,{}^t\mathbf{BBT}$ の一意性）

命題 1 と同じ記号の下で、${}^t\mathbf{T}\,{}^t\mathbf{B}'\mathbf{B}'\mathbf{T} = {}^t\mathbf{T}\,{}^t\mathbf{BBT}$ である。

（証明）

$$\mathbf{U} = \begin{bmatrix} U_1 \\ \vdots \\ U_k \end{bmatrix} = \mathbf{BT}、\ \mathbf{U}' = \begin{bmatrix} U'_1 \\ \vdots \\ U'_k \end{bmatrix} = \mathbf{B}'\mathbf{T}$$

とする。$\mathrm{E}(\mathbf{U}) = \mathbf{B}\,\mathrm{E}(\mathbf{T}) = 0$ であり、

$$\mathrm{V}(\mathbf{U}) = \mathbf{B}\,\mathrm{V}(\mathbf{T})\,{}^t\mathbf{B} = \mathbf{BC}\,{}^t\mathbf{B} = \begin{bmatrix} \mathbf{1} & \\ & 0 \end{bmatrix}$$

だから、$\mathrm{E}(U_k) = 0, \mathrm{V}(U_k) = 0$ となるが、U_k は $S_{il}\ \ (i = 1,...,k; l = 1,...,L)$ の線形関数だから、実現確率が 0 の値をとることはない。したがって、$U_k = 0$ である。同様に、$U'_k = 0$ である。

$\mathbf{M} = \mathbf{B}'\mathbf{B}^{-1}$ と置けば、$\mathbf{U}' = \mathbf{MU}$ となる。したがって、

$$\mathbf{M}\begin{bmatrix} \mathbf{1} & \\ & 0 \end{bmatrix}{}^t\mathbf{M} = \mathbf{M}\,\mathrm{V}(\mathbf{U})\,{}^t\mathbf{M} = \mathrm{V}(\mathbf{MU}) = \mathrm{V}(\mathbf{U}') = \begin{bmatrix} \mathbf{1} & \\ & 0 \end{bmatrix}$$

となるから、後出の補題 9 により、

$${}^t\mathbf{T}\,{}^t\mathbf{B}'\mathbf{B}'\mathbf{T} = {}^t\mathbf{U}'\mathbf{U}' = {}^t\mathbf{UU} = {}^t\mathbf{T}\,{}^t\mathbf{BBT}\,{}^t$$

となる。

（補題 2 の証明終わり）

第 3 章　2つの調査の比較

3.1　2調査モデル

これまでは、確率 $p_{il}\quad (i=1,...,k;l=1,...,L)$ が分かっているものとして扱ってきたが、現実の場面では、これが与えられていないのが普通である。そこで、この項では、これらを観測値からの推計値で置き換える場合の枠組みを設定する。

ここでは、2 種類の調査が実施されることを想定する。第 1 調査についてはこれまでと同じ記号を用いる。

第 2 調査については、k（選択肢の個数）は第 1 調査と同じとする。

一方、L（層の個数）、p_{il}（第l 層で第 i 選択肢が選ばれる確率）、n_l, w_l, S_{il} 及びこれらから計算される各指標は第 1 調査と第 2 調査で異なるので、第 2 調査では、L'、p'_{il}、p'_i、n'_l、n'、w'_l、N'、S'_{il}、S'_i、T'_i、$\mathbf{T}'$ などと「'」を付けて表示する。S_{il} と S'_{il} は独立であると想定する。

$n_{total}=n+n'$ と置いて、n_{total} が増大するとき n 及び n' は n_{total} にほぼ比例して増大するものとする。すなわち、β を $0<\beta<1$ なる実数として、

$$n=\beta n_{total}\text{の四捨五入値}$$

$$n'=n_{total}-n$$

と想定する。

$$|n-\beta n_{total}|<1$$

である。

p_i 及び p'_i の定義を再掲すると次のようになる。

$$p_i=\frac{1}{N}\sum_{l=1}^{L} w_l n_l p_{il}\ 、\ p'_i=\frac{1}{N'}\sum_{l=1}^{L'} w'_l n'_l p'_{il}$$

これらは、第i選択肢への層合計の回答確率である。これらが第1調査と第2調査で等しいかどうかを検定するのが、この後の目標である。すなわち、帰無仮説を次のように設定する。

帰無仮説　$p_1=p'_1,\ldots,p_k=p'_k$

p_{il}、p'_{il}、及び $p_i=p'_i$ の推計値 π_{il}、π'_{il}、及び π_i を次のように定義する。

$$\pi_{il}=\frac{S_{il}}{n_l}、\ \pi'_{il}=\frac{S'_{il}}{n'_l}、\ \pi_i=\frac{S_i+S'_i}{N+N'}$$

また、T_i、$\mathbf{T}$ の定義において、p_i を π_i に差し替えたものを $\overline{T_i}$、$\overline{\mathbf{T}}$ と置く。

$$\overline{T}_i=\frac{S_i-N\pi_i}{\sqrt{n}}$$

$$\overline{\mathbf{T}} = \begin{bmatrix} \overline{T}_1 \\ \vdots \\ \overline{T}_k \end{bmatrix}$$

3.2 ウェイトバックカイ2乗統計量（確率の推計値を用いる場合）

この項では、確率の推計値を用いる場合について、ウェイトバックされた集計値によるカイ2乗統計量を定義する（命題2）。これも、ワルド統計量の一種である（後出補題4の注を参照）。

命題2 （ウェイトバックカイ2乗検定～確率の推計値を使う場合～）

k 行 k 列の行列 $\overline{\mathbf{C}} = (\overline{c}_{ij})$ を次のように定める。

$$\overline{c}_{ii} = \sum_{l=1}^{L} \omega_l \pi_{il}(1-\pi_{il}) + \sum_{l=1}^{L'} \omega'_l \pi'_{il}(1-\pi'_{il}) \quad (\text{for } i = 1,...,k)$$

$$\overline{c}_{ij} = -\sum_{l=1}^{L} \omega_l \pi_{il} \pi_{jl} - \sum_{l=1}^{L'} \omega'_l \pi'_{il} \pi'_{jl} \quad (\text{for } i \neq j)$$

ただし、$\omega_l = \dfrac{w_l{}^2 n_l N'^2}{n(N+N')^2}$、$\omega'_l = \dfrac{w'_l{}^2 n'_l N^2}{n(N+N')^2}$

このとき、

(1) 確率1で $\lim_{n_{total}\to\infty} \pi_{il} > 0 \ \ (i = 1,...,k; l = 1,...,L)$ となる。

さらに、$\pi_{il} > 0 \ \ (i = 1,...,k; l = 1,...,L)$ のとき、

(2) $$\overline{\mathbf{B}}\overline{\mathbf{C}}{}^{t}\overline{\mathbf{B}} = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{bmatrix}$$ （対角要素以外は0）

となる正則行列 $\overline{\mathbf{B}}$ が存在する。

(3) $\overline{Chi} = {}^{t}\overline{\mathbf{T}}{}^{t}\overline{\mathbf{B}}\overline{\mathbf{B}}\overline{\mathbf{T}}$

と置けば、$\overline{Chi}$ の分布は、$n_{total} \to \infty$ のとき、自由度 $k-1$ のカイ2乗分布に弱収束する。

(4) $\overline{Chi}$ の値は、$\overline{\mathbf{B}}$ の選び方によらない。すなわち、別の正則行列 $\overline{\mathbf{B}}'$ により

$$\overline{\mathbf{B}}'\overline{\mathbf{C}}{}^{t}\overline{\mathbf{B}}' = \begin{bmatrix} \mathbf{1} & \\ & 0 \end{bmatrix}$$ （$\mathbf{1}$ は $k-1$ 行 $k-1$ 列の単位行列）

となったとすると、

$${}^{t}\overline{\mathbf{T}}\,{}^{t}\overline{\mathbf{B}}'\overline{\mathbf{B}}'\overline{\mathbf{T}} = {}^{t}\overline{\mathbf{T}}\,{}^{t}\overline{\mathbf{B}}\overline{\mathbf{B}}\overline{\mathbf{T}}$$

である。

〔証明〕

(1)は、大数の法則〔参考 4〕から得られる。

(2)と(3)について

後出の補題 4 を適用することとし、①$\overline{\mathbf{T}}$の分布が平均$\mathbf{0}$の多変量正規分布に弱収束すること、② $\lim_{n_{total}\to\infty}\overline{\mathbf{C}} = \lim_{n_{total}\to\infty}\mathrm{V}(\overline{\mathbf{T}})$ であること、③$\overline{\mathbf{C}}$の階数が$k-1$であることを示せばよい。$\overline{\mathbf{T}}$及び$\overline{\mathbf{C}}$において、π_{il}、π'_{il}、π_iをp_{il}、p'_{il}、p_iに置き換えたものを$\overline{\mathbf{T}}^*$、$\overline{\mathbf{C}}^*$とする。$\overline{\mathbf{C}}^* = \mathrm{V}(\overline{\mathbf{T}})$である。これは、$\overline{\mathbf{T}}$の定義に従って分散共分散を計算して容易に確かめられる。また、大数の法則により

$$\pi_{il} \to p_{il} \quad \text{(a.e.)}、\pi'_{il} \to p'_{il} \quad \text{(a.e.)}$$

だから、

$$\overline{\mathbf{T}} - \overline{\mathbf{T}}^* \to 0 \quad \text{(a.e.)} \qquad [1]$$

$$\overline{\mathbf{C}} - \overline{\mathbf{C}}^* = \overline{\mathbf{C}} - \mathrm{V}(\overline{\mathbf{T}}) \to 0 \quad \text{(a.e.)} \qquad [2]$$

である。

$\overline{\mathbf{T}}^*$が多項分布S_{il} $(i=1,...,k;l=1,...,L)$の線形結合であって多変量正規分布に弱収束することを考慮すれば、[1]式と参考 3 により①が満たされることが分かる。なお、帰無仮説により、$\overline{\mathbf{T}}$の平均が$\mathbf{0}$であることが保証されている。

②は、[2]式から分かる。

③については、次の補題 3 により$\mathrm{rank}\,\mathrm{V}(\overline{\mathbf{T}}) = k-1$であるが、これは、$p_{1l}+...+p_{kl}=1, p_{il}>0$なる任意の$p_{il}$について成立するから、$\pi_{il}>0$のときは、これを$\pi_{il}$に置き換えても成立する。したがって、$\pi_{il}>0$のときは、$\mathrm{rank}\,\mathrm{V}(\overline{\mathbf{C}}) = k-1$である。

(4)については、確率p_{il}がたまたま推計値π_{il}に一致していた場合を想定して、補題 2 から得られる。

〔命題 2 の証明終わり〕

補題 3 〔$\mathrm{V}(\overline{\mathbf{T}})$の階数〕

$\mathrm{rank}\,\mathrm{V}(\overline{\mathbf{T}}) = k-1$である。

〔証明〕

後出の補題 8 を適用する。$\overline{T_i}$は、S_{il}の1次結合なので確率 0 の値をとらない。したがって、$\overline{\mathbf{T}}$

で張られるベクトル空間の次元が $k-1$ であることを言えばよい。ところが、容易に確かめられるように $\sum_{i=1}^{k} \overline{T}_i = 0$ なので、これは $k-1$ 以下であることが分かる。

このベクトル空間の次元が $k-1$ 以上であることは、次のようにして分かる。

事象 B_i を

$$\begin{cases} S_{il} = n_l \\ S_{jl} = 0 \ \ (\text{f or } j \neq i) \end{cases} \quad \begin{cases} S'_{kl} = n'_l \\ S'_{jl} = 0 \ \ (\text{f or } j \neq k) \end{cases}$$

のように定義する（$1 \leq i \leq k$）。

事象 B_i の下での $\overline{\mathbf{T}}$ の値を $\mathbf{t}_i$ として、$\mathbf{t}_i$ を横に並べた行列を $\mathbf{M}$ とする。

$$\mathbf{M} = (\mathbf{t}_1, ..., \mathbf{t}_k)$$

$\mathbf{M}$ の階数が $k-1$ 以上であることをいえばよい。実際に計算してみると、次のようになる。

$$\mathbf{M} = \frac{NN'}{\sqrt{n}(N+N')} \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \\ 0 & 0 & \cdots & 1 & 0 \\ -1 & -1 & \cdots & -1 & 0 \end{bmatrix}$$

これの階数は、明らかに $k-1$ である。

（補題 3 の証明終わり）

3.3 計算例

命題 2 を使って検定を実行してみよう。「第 1 章 はじめに」で示された例と同じデータを用いる。

（1） A 調査と B 調査のそれぞれについて、層別の回答分布（S_{il}, S'_{il}）と復元ウェイト（w_l, w'_l）を準備する。

A 調査

層	平等社会 S_{1l}	自由競争社会 S_{2l}	どちらとも S_{3l}	わからない S_{4l}	復元ウェイト w_l
男性 20 歳代	7	56	12	4	3.237
男性 30 歳代	16	57	25	1	2.359
男性 40 歳代	18	68	10	3	2.367
男性 50 歳代	20	53	15	6	2.763
男性 60 歳代	13	58	23	0	2.111
女性 20 歳代	23	51	18	6	2.496
女性 30 歳代	23	49	19	9	2.311
女性 40 歳代	22	45	27	7	2.302
女性 50 歳代	17	57	23	8	2.573
女性 60 歳代	20	55	22	6	2.077

B 調査

層	平等社会 S'_{1l}	自由競争社会 S'_{2l}	どちらとも S'_{3l}	わからない S'_{4l}	復元ウェイト w'_l
男性 20 歳代	21	87	23	5	1.904
男性 30 歳代	29	90	21	5	1.643
男性 40 歳代	31	81	25	2	1.686
男性 50 歳代	23	89	28	2	1.842
男性 60 歳代	21	103	12	1	1.449
女性 20 歳代	28	72	34	8	1.758
女性 30 歳代	29	80	23	11	1.621
女性 40 歳代	24	73	41	7	1.603
女性 50 歳代	25	86	29	6	1.851
女性 60 歳代	39	77	26	2	1.490

(2) 上のデータから、$\overline{\mathbf{T}}$ ベクトル及び $\overline{\mathbf{C}}$ 行列を計算する。$\overline{\mathbf{C}}$ 行列は、$\overline{\mathbf{T}}$ ベクトルの分散共分散行列の推計値である。

$\overline{\mathbf{T}}$ ベクトル

平等社会	-0.266
自由競争社会	-0.809
どちらとも	0.414
わからない	0.660

$\overline{\mathbf{C}}$ 行列

	平等社会	自由競争社会	どちらとも	わからない
平等社会	0.378	-0.266	-0.090	-0.022
自由競争社会	-0.266	0.609	-0.278	-0.065
どちらとも	-0.090	-0.278	0.390	-0.023
わからない	-0.022	-0.065	-0.023	0.110

再掲になるが、$\overline{\mathbf{T}}$ ベクトル及び $\overline{\mathbf{C}}$ 行列の計算式は次の通りである。

$$\overline{\mathbf{T}} = \begin{bmatrix} \overline{T}_1 \\ \vdots \\ \overline{T}_k \end{bmatrix} \qquad \overline{T}_i = \frac{S_i - N\pi_i}{\sqrt{n}}$$

$$\overline{\mathbf{C}} = (\overline{c}_{ij})$$

$$S_i = \sum_{l=1}^{L} w_l S_{il} \qquad S'_i = \sum_{l=1}^{L'} w'_l S'_{il}$$

$$n_l = \sum_{i=1}^{k} S_{il} \qquad n'_l = \sum_{i=1}^{k} S'_{il} \qquad n = \sum_{l=1}^{L} n_l$$

$$N = \sum_{l=1}^{L} w_l n_l \qquad N' = \sum_{l=1}^{L'} w'_l n'_l$$

$$\pi_{il} = \frac{S_{il}}{n_l}、\ \pi'_{il} = \frac{S'_{il}}{n'_l}、\ \pi_i = \frac{S_i + S'_i}{N + N'}$$

$$\overline{c}_{ii} = \sum_{l=1}^{L} \omega_l \pi_{il}(1-\pi_{il}) + \sum_{l=1}^{L'} \omega'_l \pi'_{il}(1-\pi'_{il}) \quad (\text{for } i = 1,...,k)$$

$$\overline{c}_{ij} = -\sum_{l=1}^{L} \omega_l \pi_{il}\pi_{jl} - \sum_{l=1}^{L'} \omega'_l \pi'_{il}\pi'_{jl} \quad (\text{for } i \neq j)$$

$$\omega_l = \frac{w_l^2 n_l N'^2}{n(N+N')^2} \qquad \omega'_l = \frac{w'^2_l n'_l N^2}{n(N+N')^2}$$

k：選択肢の個数　L：A 調査（第1調査）の層の個数　L'：B調査（第 2 調査）の層の個数

（3） $\overline{\mathbf{C}}$ 行列の対角化を実行する。

$$\mathbf{Q}\overline{\mathbf{C}}'\mathbf{Q} = \begin{bmatrix} 0.869 & 0 & 0 & 0 \\ 0 & 0.145 & 0 & 0 \\ 0 & 0 & 0.474 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

となる。ただし、$\mathbf{Q}$ は次の行列である。

$\mathbf{Q}$ 行列

-0.373	0.831	-0.410	-0.049
0.313	0.242	0.309	-0.865
0.716	-0.023	-0.698	0.004
-0.500	-0.500	-0.500	-0.500

この $\mathbf{Q}$ 行列は、対称行列の固有値や固有ベクトルを求める通常のテクニックにより計算することができる。主成分分析で用いられるのと同じテクニックである。詳しくは、数値計算法の教科書等を参照のこと。

(4) 次の算式により $\overline{\mathbf{B}}$ 行列を決定する。

$$\overline{\mathbf{B}} = \begin{bmatrix} \frac{1}{\sqrt{0.869}} & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{0.145}} & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{0.474}} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{Q}$$

$\overline{\mathbf{B}}$ 行列

-0.400	0.891	-0.439	-0.052
0.824	0.637	0.812	-2.274
1.041	-0.033	-1.014	0.006
-0.500	-0.500	-0.500	-0.500

(5) $\overline{\mathbf{T}}$ に $\overline{\mathbf{B}}$ を乗じて、$\overline{\mathbf{B}}\overline{\mathbf{T}}$ を得る。「第1章 はじめに」で述べた「漸近的に分散が1 で互いに無相関になるように各項を線形変換する」というのは、この計算を指す。

$\overline{\mathbf{B}}\overline{\mathbf{T}}$ ベクトル

-0.832
-1.900
-0.666
0.000

(6) そこで、

$$\overline{Chi} = {}^{t}\overline{\mathbf{T}}\,{}^{t}\overline{\mathbf{B}}\overline{\mathbf{B}}\overline{\mathbf{T}} = (-0.832)^2 + (-1.900)^2 + (-0.666)^2 + 0.000^2 = 4.75$$

となる。

この4.75を自由度3のカイ2乗分布に当てはめて、有意確率＝0.19を得る。すなわち、層計で「A調査とB調査の結果に差がない」という仮説は、10%の有意水準で棄却されない。

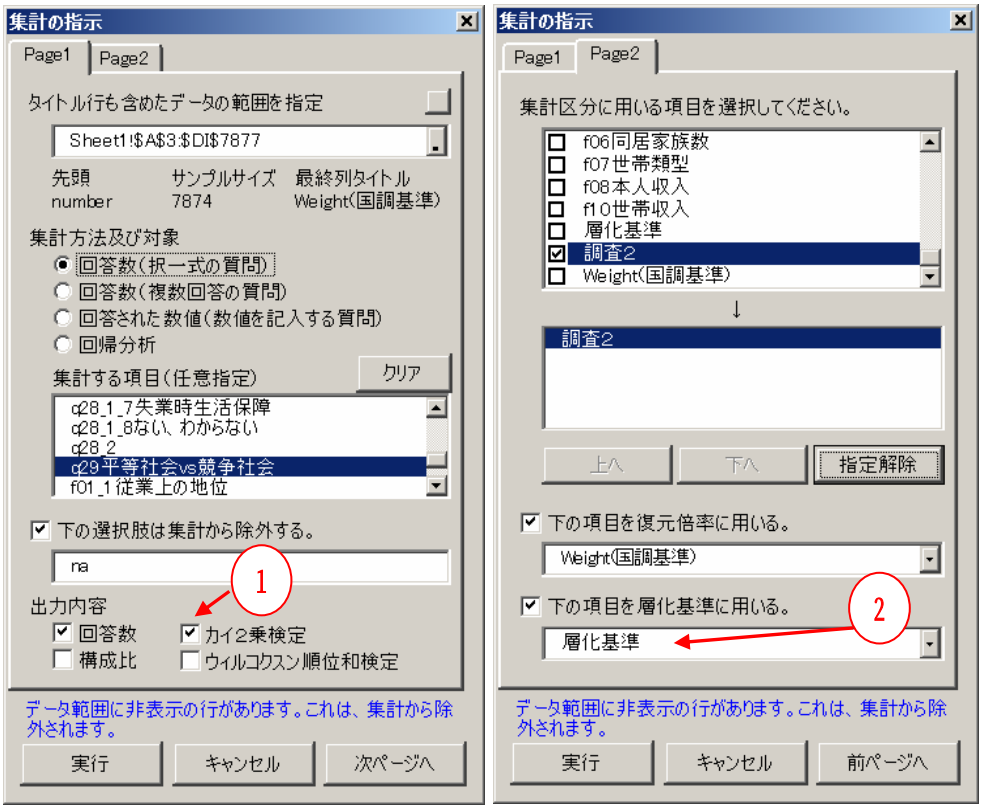
(パソコンのプログラムについて)

上でみたように、本稿の手法による検定を実行するためには、層ごとの集計値と復元ウェイトが必要になる。しかし、これらは、通常の集計作業の中で当然に用いられるものである。そこで、集計プログラムの中に上のような計算をあらかじめ組み込んでおけば、分析者はほとんど負担を感じることなく検定を実行できる。

「インターネット調査研究」では、こういう機能を実際に組み込んだプログラムを新たに開発して用

いた。下の指示フォームをみて分かるように、通常の集計に必要な指示のほか、①カイ2乗検定を行うこと、及び、②層化基準（性・年齢など）が格納されているデータの位置、の2つを指示するだけで検定結果が得られる。

（指示フォーム）



（計算結果）

出力項目=復元回答数						
調査2	q29平等社会vs競争社会					
	計	1	2	3	4	カイ2乗検定
計	4,764	884	2,761	912	207	…
2_A	2,374	432	1,351	467	124	…
3_B	2,390	452	1,410	445	83	0.19139665

第4章 補足

4.1 有限母集団、多段階層化抽出等についてのスケッチ

命題2の方法は、①母集団が近似的に無限母集団とみなせること（抽出率が小さいこと）、②サンプルサイズが大きいこと、③層化一段抽出であること、という前提で適用できる。逆にいえば、母集団サイズまたはサンプルサイズが小さいときや、多段階抽出のときには適用できない。

このような複雑な標本設計の場合でも、命題 2 の $\overline{Chi}$ と類似の算式を構成することは可能と思われる。算式は多少複雑になることが予想されるが、いったんそういうものが明示的に示されたならば、汎用的な手法として活用できる。

また、次のような別のアプローチもある。

(1) 副標本を用いる方法

サンプルをいくつかの小集団(副標本)に分割して、その小集団ごとの集計値から分散共分散行列を推計する方法。

(2) モンテカルロ法(シミュレーション法)

仮想的な母集団をパソコン内に設定し、そこから擬似的なサンプリングを多数回発生させることにより集計値の分布を推計する方法。とくにサンプルサイズが小さいときは、分散共分散行列等を用いる漸近理論が使えないので、この方法が有効と考えられる。

(3) 厳密な計算

母集団サイズが小さいならば、上のモンテカルロ法の特別なケースとして、理論的に可能なサンプリングをすべて尽くすことにより厳密な確率分布が得られる。

これら(1)、(2)、(3)の方法は、調査ごと、場合によっては集計項目ごとにプログラミングが必要になることも考えられ、汎用性に欠ける面もある。また、計算時間も長くなる。しかし標本設計が複雑になっても計算式はあまり複雑にならないという利点がある。パソコンの性能向上が著しいなか、今後、利用可能性が高まることも考えられる。

4.2 ウェイトバック Wilcoxon 順位和検定

「インターネット調査研究」では、Wilcoxon 順位和検定も行ったので、その際に用いた統計量を紹介する。これは、層別の通常の Wilcoxon 順位和統計量を単に平均しただけのものである。

命題 3 (ウェイトバック Wilcoxon 順位和検定)

$$Wcox_l = \sum_{i=1}^{k} S_{il} G_{il} \quad (l = 1, \ldots, L)$$

ただし、$G_{il} = \sum_{j=1}^{i-1} (S_{il} + S'_{il}) + \frac{1}{2}(S_{il} + S'_{il} + 1)$

$$\mu_l = \frac{1}{2} n_l (n_l + n'_l + 1)$$

$$\sigma_l^{\,2} = \frac{n_l n'_l (n_l + n'_l + 1)}{12} - \frac{n_l n'_l \sum_{i=1}^{k} (S_{il}^{\,3} - S_{il})}{12(n_l + n'_l)(n_l + n'_l - 1)}$$

とする。

さらに、$a_1, ..., a_L$ を $a_1 + ... + a_L = 1$ なる実数として、

$$Wcox = \sum_{l=1}^{L} a_l Wcox_l$$

と置く。

このとき、

$$\frac{Wcox - \sum_{l=1}^{L} a_l \mu_l}{\sqrt{\sum_{l=1}^{L} a_l^{\,2} \sigma_l^{\,2}}}$$

の分布は、$n_{total} \to \infty$ のとき標準正規分布に弱収束する。

(証明)

$Wcox_l$ は、第 l 層における通常の Wilcoxon 順位和統計量であって、$\frac{Wcox_l - \mu_l}{\sigma_l}$ の分布は $n_{total} \to \infty$ のとき標準正規分布に弱収束する。このことと、異なる l 同士で $Wcox_l$ が独立なことから、この命題が得られる。

(命題 3 の証明終わり)

「インターネット調査研究」では、命題 3 で $a_l = \frac{w_l n_l + w'_l n'_l}{N + N'}$ と設定した。

4.3 一般的な補題

この項では確率や代数に関する一般的な補題を整理する。

補題 4 (非正則ケースのワルド統計量)

k 次確率ベクトルの列 $\mathbf{Y}_n \quad (n = 1, 2, ..)$ は、平均が $\mathbf{0}$ の多変量正規分布に従う一定の確率ベクトル $\mathbf{Y}$ に法則収束するものとする。また、$\mathbf{V}_1, \mathbf{V}_2, ...$ を、k 次の広義正定値正方確率行列の列とし、次の条件を満たすとする。

① $n \to \infty$ のとき $\mathrm{rank}\, \mathbf{V}_n$ は r に概収束する。

② $n \to \infty$ のとき $\mathbf{V}_n$ は $\mathrm{V}(\mathbf{Y})$ に概収束する。

このとき、

(1)　十分大きな n に対してほとんど至るところ正則な確率行列 $\mathbf{B}_n$ が存在し、ほとんど至るところ

$$\mathbf{B}_n \mathbf{V}_n {}^t\mathbf{B}_n = \begin{bmatrix} \mathbf{1} & \\ & \mathbf{0} \end{bmatrix}. \qquad [3]$$

（$\mathbf{1}$は r 行 r 列の単位行列、$\mathbf{0}$ は $k-r$ 行 $k-r$ 列のゼロ行列）

となる。

(2)　$n \to \infty$ のとき、${}^t\mathbf{Y}_n {}^t\mathbf{B}_n \mathbf{B}_n \mathbf{Y}_n$ の分布は自由度 r のカイ2乗分布に弱収束する。

（証明）

この補題は、ワルド検定の理論（参考 8）を使って証明することもできる。しかし、直接証明してもそれほど難しくないので、直接の証明を掲げる。

（(1)について）

①、②により、十分大きな n に対しほとんど至るところ $\operatorname{rank}\mathbf{V}_n = r$、かつほとんど至るところ $\lim_{n\to\infty}\mathbf{V}_n = \mathrm{V}(\mathbf{Y})$ となる。よって、この条件が満たされる事象及び n に対して補題 5 のように $\mathbf{B}_n$ の値を定めれば、これは、一定の確率行列に概収束する。

（(2)について）

②により $\lim_{n\to\infty}\mathbf{V}_n$ はほとんど至るところ定数行列となるから、補題 5 により $\lim_{n\to\infty}\mathbf{B}_n$ もほとんど至るところ定数行列となるように選ぶことができる。そこで、次のように $\mathbf{B}$、$\mathbf{U}_n$、$\mathbf{U}$ を定義する。

$$\mathbf{B} = \lim_{n\to\infty}\mathbf{B}_n \quad \text{a.e.} \quad \text{（概収束）}$$

$$\mathbf{U}_n = \mathbf{B}_n \mathbf{Y}_n$$

$$\mathbf{U} = \mathbf{B}\mathbf{Y}$$

$\mathbf{U}$ は多変量正規分布であって、参考 3 により、

$$\mathbf{U} = \lim_{n\to\infty}\mathbf{U}_n \quad \text{in law} \quad \text{（法則収束）}$$

となる。

ここで、

$$\mathrm{V}(\mathbf{U}) = \begin{bmatrix} \mathbf{1} & \\ & \mathbf{0} \end{bmatrix} \qquad [4]$$

（$\mathbf{1}$は r 行 r 列の単位行列、$\mathbf{0}$ は $k-r$ 行 $k-r$ 列のゼロ行列）

である。これは、次のようにして示される。

$$\mathrm{V}(\mathbf{U}) = \mathrm{V}(\mathbf{BY}) = \mathbf{B}\,\mathrm{V}(\mathbf{Y})\,{}^{t}\mathbf{B} = \mathbf{B}\,\mathrm{V}(\lim_{n\to\infty}\mathbf{Y}_n \ \text{ in law})\,{}^{t}\mathbf{B} = \mathbf{B}(\lim_{n\to\infty}\mathrm{V}(\mathbf{Y}_n))\,{}^{t}\mathbf{B}$$

$$= \mathbf{B}(\lim_{n\to\infty}\mathbf{V}_n \ \text{a.e.})\,{}^{t}\mathbf{B} = \lim_{n\to\infty}\mathbf{B}\mathbf{V}_n\,{}^{t}\mathbf{B} \ \text{ a.e. } = \begin{bmatrix} \mathbf{1} & \\ & \mathbf{0} \end{bmatrix}$$

したがって、参考 5 によって$\mathbf{U}$の各成分が互いに独立であることが分かるから、参考 6 によって${}^{t}\mathbf{U}\mathbf{U}$が自由度$r$のカイ2乗分布に従うことが分かる。よって、補題 6 によって、${}^{t}\mathbf{Y}_n\,{}^{t}\mathbf{B}_n\mathbf{B}_n\mathbf{Y}_n = {}^{t}\mathbf{U}_n\mathbf{U}_n$の分布が自由度$r$のカイ2乗分布に弱収束することが分かる。

（補題 4 の証明終わり）

（注） 補題 4 において、$r=k$ ならば $\mathbf{B}_n\mathbf{V}_n\,{}^{t}\mathbf{B}_n = \mathbf{1}$ すなわち $\mathbf{V}_n^{-1} = {}^{t}\mathbf{B}_n\mathbf{B}_n$ となるから、${}^{t}\mathbf{Y}_n\,{}^{t}\mathbf{B}_n\mathbf{B}_n\mathbf{Y}_n = {}^{t}\mathbf{Y}_n\mathbf{V}_n^{-1}\mathbf{Y}_n$ である。すなわち、これは普通のワルド統計量（参考 8）となる。

また、$r<k$ のときでも、r 次確率ベクトル$\mathbf{X}_n$とk 次正則行列$\mathbf{A}$によって

$$\mathbf{Y}_n = \mathbf{A}\begin{bmatrix} \mathbf{X}_n \\ \mathbf{0} \end{bmatrix}$$

$$\mathbf{V}_n = \mathbf{A}\begin{bmatrix} \mathbf{W}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}{}^{t}\mathbf{A} = \mathbf{A}\begin{bmatrix} \mathbf{D}_n\,{}^{t}\mathbf{D}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}{}^{t}\mathbf{A} \qquad \mathbf{D}_n \text{ はほとんど至るところ正則な } r \text{ 次確率行列}$$

と書けるならば、

$$\mathbf{B}_n = \begin{bmatrix} \mathbf{D}_n^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix}\mathbf{A}^{-1}$$

とすることができて、${}^{t}\mathbf{Y}_n\,{}^{t}\mathbf{B}_n\mathbf{B}_n\mathbf{Y}_n = {}^{t}\mathbf{X}_n\mathbf{W}_n^{-1}\mathbf{X}_n$ となる。すなわち、これも普通のワルド統計量になる。前述の命題 1 と命題 2 は、このケースに相当する。

補題 5（広義正定値対称行列の収束列）

$\mathbf{C}_1, \mathbf{C}_2, \ldots$ をk 次正方行列の列とし、いずれも広義正定値（i.e. 固有値が 0 以上）であり、階数がrであるものとする。

もし、$\mathbf{C}_1, \mathbf{C}_2, \ldots$ が一定の行列に収束するならば、次のような正則行列の列$\mathbf{B}_1, \mathbf{B}_2, \ldots$が存在する。

(1) $\mathbf{B}_n\mathbf{C}_n\,{}^{t}\mathbf{B}_n = \begin{bmatrix} \mathbf{1} & \\ & \mathbf{0} \end{bmatrix}$ $(n = 1, 2, \ldots)$

$\mathbf{1}$はr 行r 列の単位行列、$\mathbf{0}$ は$k-r$ 行$k-r$ 列のゼロ行列

(2) $n \to \infty$ のとき$\mathbf{B}_n$ は一定の行列に収束

（証明）

$\mathbf{C}_n$ の固有値を$\theta_{n1}, \ldots, \theta_{nr}, 0, \ldots, 0$ とする。適当に並べ替えをして、$n \to \infty$ のとき、$\theta_{n1}, \ldots, \theta_{nr}$ が

それぞれ一定値に収束するようにすることができる（参考 7）。

また、$\mathbf{A}_n$ を、

$$\mathbf{A}_n \mathbf{C}_n {}^{\mathrm{t}}\mathbf{A}_n = \begin{bmatrix} \theta_{n1} & & & \\ & \ddots & & \\ & & \theta_{nr} & \\ & & & \mathbf{0} \end{bmatrix}$$

となるような直交行列とする。

$\mathbf{A}_n$ は、$\mathbf{C}_n$ 及び $\theta_{n1},...,\theta_{nr}$ を係数とする不定一次連立方程式を解いて得ることができるので、これも一定の行列に収束するように選ぶことができる。

そこで、

$$\mathbf{B}_n = \begin{bmatrix} \theta_{n1}^{-\frac{1}{2}} & & & \\ & \ddots & & \\ & & \theta_{nr}^{-\frac{1}{2}} & \\ & & & \mathbf{1} \end{bmatrix} \mathbf{A}_n$$

とすればよい。

（補題 5 の証明終わり）

補題 6（連続関数の収束）

確率ベクトルの列 $\mathbf{X}_n = (X_{1n},..., X_{kn})\ \ (n = 1,2,...)$ が確率ベクトル $\mathbf{X} = (X_1,..., X_k)$ に法則収束するものとする。$\mathrm{f}(x_1,..., x_k)$ を $\mathbf{X}_n\ (n = 1,2,...)$ の値域で定義された連続関数とする。

このとき、確率変数の列 $\mathrm{f}(X_{1n},..., X_{kn})\ \ (n = 1,2,...)$ は、$\mathrm{f}(X_1,..., X_k)$ に法則収束する。

（証明）

法則収束の定義（参考 1）により、任意の有界連続関数 g に対して

$$\lim_{n\to\infty} \mathrm{E}(\mathrm{g}\,\mathrm{f}(X_{1n},..., X_{kn})) = \mathrm{E}(\mathrm{g}\,\mathrm{f}(X_1,..., X_k))$$

であることをいえばよい。しかし、これは、$\mathrm{g}\,\mathrm{f}$ が有界連続関数であることと $\mathbf{X}_n$ が法則収束することから明か。

（補題 6 の証明終わり）

補題 7 （$(\mathbf{a}_1,..., \mathbf{a}_k)$ の固有値）

$a_1,...,a_k$ を $a_1+...+a_k=1$ なる実数として、

$$\mathbf{a}_i=\begin{bmatrix} -a_1 \\ \vdots \\ -a_{i-1} \\ 1-a_i \\ -a_{i+1} \\ \vdots \\ -a_k \end{bmatrix}$$

とするとき、これらを並べた行列 $(\mathbf{a}_1,...,\mathbf{a}_k)$ は、$k-1$ 個の 1 と 1 個の 0 を固有値に持つ。

（証明）

$$\mathbf{M}=\begin{bmatrix} a_1 & \cdots & a_1 \\ \vdots & & \vdots \\ a_k & \cdots & a_k \end{bmatrix}$$

と置くと、明らかに $\mathrm{rank}(\mathbf{M})=1$ だから、$\mathbf{M}$ の固有値のうち $k-1$ 個は 0 である。また、
$\mathrm{tr}(\mathbf{M})=a_1+...+a_k=1$ により、残る 1 個の固有値は 1 であることが知れる。
したがって、$(\mathbf{a}_1,...,\mathbf{a}_k)=1-\mathbf{M}$ により、$(\mathbf{a}_1,...,\mathbf{a}_k)$ の固有値は、$k-1$ 個の 1 と 1 個の 0 であることが分かる。

（補題 7 の証明終わり）

補題 8（分散共分散行列の階数）

$X_1,...,X_m$ をいずれも平均が 0 で分散有限の確率変数であって、かつ実現確率が 0 の値をとることがないものとする。このとき、

$X_1,...,X_m$ の分散共分散行列の階数＝ $X_1,...,X_m$ で張られるベクトル空間の次元

である。

（証明）

$X_1,...,X_m$ で張られるベクトル空間を $\mathbf{S}$ と置く。Y を $\mathbf{S}$ の任意の元とすると、「確率が 0 の値をとることがない」という仮定により、$\mathrm{V}(Y)=0 \Leftrightarrow Y=0$ である。したがって、$\mathbf{S}$ に対して共分散により内積を定義すると、$\mathbf{S}$ は内積空間となる。$\mathbf{S}$ の次元を r として、$Z_1,...,Z_r$ をその正規直交基とする。

$$\begin{bmatrix} X_1 \\ \vdots \\ X_m \end{bmatrix}=\mathbf{A}\begin{bmatrix} Z_1 \\ \vdots \\ Z_r \end{bmatrix}$$

とすれば、$\mathbf{A}$ は階数 r の行列であって、

$$X_1,...,X_m \text{ の分散共分散行列 } = \mathbf{A}(\, Z_1,...,Z_r \text{ の分散共分散行列 })\,{}^t\mathbf{A}$$
$$= \mathbf{A}\begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}{}^t\mathbf{A} = \mathbf{A}\,{}^t\mathbf{A}$$

となる。したがって、$X_1,...,X_m$ の分散共分散行列の階数＝ $\mathrm{rank}(\mathbf{A}\,{}^t\mathbf{A}) = \mathrm{rank}(\mathbf{A}) = r$ を得る。

（補題 8 の証明終わり）

補題 9（ある種のベクトルの平方和の一意性）

$\mathbf{u}$ と $\mathbf{u}'$ を k 個の実数を並べた列ベクトルとし、いずれも末尾の $k-r$ 個の要素は 0 であるとする。

$$\mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_r \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \mathbf{u}' = \begin{bmatrix} u'_1 \\ \vdots \\ u'_r \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

また、$\mathbf{M}$ は k 行 k 列の行列で、

$$\mathbf{M}\begin{bmatrix} \mathbf{1} & \\ & \mathbf{0} \end{bmatrix}{}^t\mathbf{M} = \begin{bmatrix} \mathbf{1} & \\ & \mathbf{0} \end{bmatrix} \qquad \mathbf{1} \text{ は } r \text{ 行 } r \text{ 列の単位行列、} \mathbf{0} \text{ は } k-r \text{ 行 } k-r \text{ 列のゼロ行列}$$

を満たすものとする。

このとき、もし $\mathbf{u}' = \mathbf{M}\mathbf{u}$ ならば、${}^t\mathbf{u}'\mathbf{u}' = {}^t\mathbf{u}\mathbf{u}$ である。

（証明）

$\mathbf{M}$ を上 r 行と下 $k-r$ 行、左 r 列と右 $k-r$ 列に 4 分割して、

$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix}$ とすると、

$$\begin{bmatrix} \mathbf{1} & \\ & \mathbf{0} \end{bmatrix} = \mathbf{M}\begin{bmatrix} \mathbf{1} & \\ & \mathbf{0} \end{bmatrix}{}^t\mathbf{M} = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix}\begin{bmatrix} \mathbf{1} & \\ & \mathbf{0} \end{bmatrix}\begin{bmatrix} {}^t\mathbf{M}_{11} & {}^t\mathbf{M}_{21} \\ {}^t\mathbf{M}_{12} & {}^t\mathbf{M}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{M}_{11}\,{}^t\mathbf{M}_{11} & \mathbf{M}_{11}\,{}^t\mathbf{M}_{21} \\ \mathbf{M}_{21}\,{}^t\mathbf{M}_{11} & \mathbf{M}_{21}\,{}^t\mathbf{M}_{21} \end{bmatrix}$$

より

$$\mathbf{M}_{11}\,{}^{t}\mathbf{M}_{11}=\mathbf{1}$$

となる。したがって、

$${}^{t}\mathbf{u}'\mathbf{u}'={}^{t}\mathbf{u}'\begin{bmatrix}\mathbf{1} & \\ & \mathbf{0}\end{bmatrix}\mathbf{u}'={}^{t}\mathbf{u}\,{}^{t}\mathbf{M}\begin{bmatrix}\mathbf{1} & \\ & \mathbf{0}\end{bmatrix}\mathbf{M}\mathbf{u}={}^{t}\mathbf{u}\begin{bmatrix}\mathbf{1} & \\ & \mathbf{0}\end{bmatrix}{}^{t}\mathbf{M}\begin{bmatrix}\mathbf{1} & \\ & \mathbf{0}\end{bmatrix}\mathbf{M}\begin{bmatrix}\mathbf{1} & \\ & \mathbf{0}\end{bmatrix}\mathbf{u}$$

$$={}^{t}\mathbf{u}\begin{bmatrix}\mathbf{M}_{11}\,{}^{t}\mathbf{M}_{11} & \\ & \mathbf{0}\end{bmatrix}\mathbf{u}={}^{t}\mathbf{u}\begin{bmatrix}\mathbf{1} & \\ & \mathbf{0}\end{bmatrix}\mathbf{u}={}^{t}\mathbf{u}\mathbf{u}$$

となる。

（補題 9 の証明終わり）

4.4 基本的な用語と定理、関数記号

本稿で説明なしに用いた用語や定理をまとめて示す。これらは、主として巻末の参考文献から引用したものである。

参考 1　（定義）確率ベクトルの収束概念

（各点収束）

すべての事象で $\lim_{n\to\infty}\mathbf{X}_n=\mathbf{X}$ となること。

（概収束 converges almost everywhere）

$\Pr(\lim_{n\to\infty}\mathbf{X}_n=\mathbf{X})=1$ となること。

$\lim_{n\to\infty}\mathbf{X}_n=\mathbf{X}$ a.e.　と表記

（確率収束　converges in probability）

任意の正数 ε に対し $\lim_{n\to\infty}(\Pr(|\mathbf{X}_n-\mathbf{X}|>\varepsilon)=0$ となること。

$\operatorname{plim}_{n\to\infty}\mathbf{X}_n=\mathbf{X}$ と表記

（法則収束　convergers in law、　弱収束　converges weakly）

$\mathbf{X}_n,\mathbf{X}$ の分布関数を F_n,F として（$\mathrm{F}_n(\mathbf{x})=\Pr(\mathbf{X}_n\le\mathbf{x}),\mathrm{F}(x)=\Pr(\mathbf{X}\le\mathbf{x})$）、
F の任意の連続点 $\mathbf{x}$ で $\lim_{n\to\infty}\mathrm{F}_n(\mathbf{x})=\mathrm{F}(\mathbf{x})$ となること。
これは、任意の有界連続関数 f に対して $\lim_{n\to\infty}\mathrm{E}(\mathrm{f}(\mathbf{X}_n))=\mathrm{E}(\mathrm{f}(\mathbf{X}))$ となることと同値。

$\lim_{n\to\infty}\mathbf{X}_n = \mathbf{X}$ in law　と表記

$\mathbf{X}_n$ の分布が $\mathbf{X}$ の分布に弱収束する(converges weakly)ともいう。

参考 2　〔定理〕収束の関係

各点収束 ⇒ 概収束 ⇒ 確率収束 ⇒ 法則収束〔弱収束〕

参考 3　〔定理〕一次関数の収束

k を正の整数とし、$\mathbf{X}_n, \mathbf{B}_n$ $(n=1,2,...)$ を k 次元確率ベクトルの列、$\mathbf{A}_n$ $(n=1,2,...)$ を k 行 k 列確率行列の列とする。また、$\mathbf{X}, \mathbf{B}, \mathbf{A}$ をそれぞれ $\mathbf{X}_n, \mathbf{B}_n, \mathbf{A}_n$ と同じサイズの確率ベクトルおよび確率行列とする。

$n \to \infty$ のとき、もし、$\mathbf{X}_n$ が $\mathbf{X}$ に法則収束し、$\mathbf{B}_n, \mathbf{A}_n$ がそれぞれ $\mathbf{B}, \mathbf{A}$ に確率収束するならば、$\mathbf{A}_n\mathbf{X}_n + \mathbf{B}_n$ は $\mathbf{AX} + \mathbf{B}$ に法則収束する。

参考 4　〔定理〕大数の法則

X_n $(n=1,2,...)$ が独立で同一の分布に従い、各 X_n に平均と分散が存在し、$\mathrm{E}(X_n) = \mu, V(X_n) < v$ $(n=1,2,...)$ とするとき、

$\lim_{n\to\infty}\frac{1}{n^c}\sum_{i=1}^{n} X_i = \mu$　a.e.〔概収束〕

となる。ただし、c は 0.5 より大きな数とする。

参考 5　〔定理〕正規分布の独立性判定

$X_1, X_2, \ldots, X_m$ をその同時分布が m 変量正規分布に従う確率変数とする。もし、これらのどの2つも無相関ならば、$X_1, X_2, \ldots, X_m$ は独立である。

参考 6　〔定義〕カイ2乗分布

$X_1, ..., X_k$ を互いに独立で、それぞれ平均 0、分散 1 の正規分布に従う確率変数とするとき、$X_1^2 + ... + X_k^2$ の分布を「自由度 k のカイ2乗分布」という。

参考 7　〔定理〕多項式の根

$\omega_1, ..., \omega_n$ を多項式 $x^n + a_{n-1}x^{n-1} + ... + a_0$ の根とする。$\omega_1, ..., \omega_n$ を適当に並べ替えて $a_{n-1}, ..., a_0$ の関数とみなすとき、これは連続関数となる。

参考８　（定理）一般ワルド検定（Stroudt[4]による）

$\{\boldsymbol{\theta}_n\}$を$p$次元ユークリッド空間$\mathrm{E}^p$の点の列として、$\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\boldsymbol{\delta}_n$と表されるものとする。ここで、$\lim \boldsymbol{\delta}_n = \boldsymbol{\delta}$であって、$\boldsymbol{\theta}_0$と$\boldsymbol{\delta}$は定点である。$\{\mathbf{t}_n\}$を$p$次元確率ベクトルの列として、$n^{\frac{1}{2}}(\mathbf{t}_n - \boldsymbol{\theta}_n\}$の分布は多変量正規分布$N(\mathbf{0}, \boldsymbol{\Sigma}_0)$に弱収束するものとする。$\boldsymbol{\Sigma}_0$は正則とする。$\{\mathbf{S}_n\}$を$p \times p$次正方確率行列の列とし、確率１で正則、かつ、$\mathrm{plim}\,\mathbf{S}_n = \boldsymbol{\Sigma}_0$とする。

$\gamma : \mathrm{E}^p \to \mathrm{E}^r (r \le p)$を、次を満たす関数とする。①$\gamma(\boldsymbol{\theta}_0) = 0$。②$\boldsymbol{\theta}_0$の周り半径$\rho$以内で有界かつ連続な 2 階偏微分が存在する。③$\boldsymbol{\Gamma}_0 \equiv (\partial\gamma_i / \partial\theta_j)\ (1 \le i \le r, 1 \le j \le p)$（$\boldsymbol{\theta}_0$における値）の階数が$r$。

J_nを、$J_n = n\,{}^t\gamma(\mathbf{t}_n)(\mathbf{G}_n \mathbf{S}_n\,{}^t\mathbf{G}_n)^{-1}\gamma(\mathbf{t}_n)$と定義する。ここで、
$\mathbf{G}_n \equiv (\partial\gamma_i / \partial\theta_j)\ (1 \le i \le r, 1 \le j \le p)$（$\mathbf{t}_n$における値）である。

このとき、$n \to \infty$に従って、J_nの分布は非心カイ２乗分布$\chi_r^{\,2}({}^t\boldsymbol{\delta}\,{}^t\boldsymbol{\Gamma}_0(\boldsymbol{\Gamma}_0\boldsymbol{\Sigma}_0\,{}^t\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0\boldsymbol{\delta})$に弱収束する。とくに$\boldsymbol{\delta} = \mathbf{0}$ならば、これはカイ２乗分布に弱収束する。

（注）　参考８で、とくに$\boldsymbol{\delta}_n = \boldsymbol{\delta} = \mathbf{0}, \boldsymbol{\theta}_n = \boldsymbol{\theta}_0 = \mathbf{0}, r = p, \gamma(\boldsymbol{\theta}) \equiv \boldsymbol{\theta}$として、$\mathbf{x}_n = n^{\frac{1}{2}}\mathbf{t}_n$と置けば、

$$J_n = {}^t\mathbf{x}_n \mathbf{S}_n^{-1}\mathbf{x}_n$$

となる。すなわち、分散共分散行列の推計値（$\mathbf{S}_n$）が得られれば、漸近的にカイ２乗分布に従う統計量（J_n）が直ちに構成できる。

関数記号

（以下、X、Y等は確率変数を表す。$\mathbf{X}$は確率ベクトルを表す。Bは事象を表す。$\mathbf{A}$は行列又はベクトルを表す。）

$\mathrm{E}(X)$　Xの期待値

$\mathrm{Cov}(X, Y)$　XとYの共分散（$\mathrm{Cov}(X, Y) = \mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y)))$）

$\mathrm{V}(X)$　Xの分散（$\mathrm{V}(X) = \mathrm{Cov}(X, X) = \mathrm{E}((X - \mathrm{E}(X))^2)$）

$\mathrm{V}(\mathbf{X})$　$\mathbf{X}$の分散共分散行列

$$\mathrm{V}(\mathbf{X}) = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \cdots & \mathrm{Cov}(X_1, X_k) \\ \vdots & & \vdots \\ \mathrm{Cov}(X_k, X_1) & \cdots & \mathrm{Cov}(X_k, X_k) \end{bmatrix} \quad \text{ただし、} \mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix}$$

$\mathrm{Pr}(B)$　Bの確率

${}^t\mathbf{A}$　$\mathbf{A}$の転置行列

$\mathrm{tr}(\mathbf{A})$　$\mathbf{A}$のトレース（対角要素の和）

$\mathrm{rank}(\mathbf{A})$　$\mathbf{A}$の階数

参考文献

[1] 伊藤清「確率論」(岩波書店　初版 1953　第 16 刷)

[2] 竹内啓他「統計学辞典」(東洋経済新報社　初版 1989　第 4 刷)

[3] E.L.レーマン(鍋谷清治他訳)「ノンパラメトリックス　順位にもとづく統計的方法」(森北出版　初版 1978　第 2 刷)

[4] **T. W. F. Stroud "On Obtaining Large-Sample Tests From Asymptotically Normal Estimators" The Annals of Mathematical Statistics 1971, Vol.42, No.4, 1412-1424**

[5] 中村隆1998「1995年SSM調査の標本設計と標本精度－標本抽出法を考慮した分析に向けて」石田浩編『1995 年 SSM 調査シリーズ1社会階層・移動の基礎分析と国際比較』科学研究費補助金特別推進研究(1)「現代日本の社会階層に関する全国調査研究」成果報告書 pp.77-100

[6] Serge Lang "Algebra"(Addison-Wesly Publishing Company, 4th printing 1971)