多次元データが単調欠測している場合の平均ベクトルの検定

東京理科大学大学院 理学研究科 応用数学専攻 博士後期課程３年
日本学術振興会 特別研究員DC1
八木（やぎ）　文香（あやか）

はじめに

近年，ビッグデータなどの分析等で統計学が注目されていますが，本研究の内容は，平成29年５月11日に開催された日本計算機統計学会第31回大会の一般セッションにて，「Transformations of Simplified T^2 Statistic with Monotone Missing Data」（単調欠測データにおける簡便なT^2統計量の変換統計量）というタイトルで筆者が講演した内容を中心に紹介するものです。この研究は，東京理科大学理学部応用数学科の瀬尾隆教授との共同研究となっております。

単調欠測データとは

統計データ解析では，大量のデータを取り扱うときばかりでなく，小規模なデータにおいても，データが何らかの理由で欠測（欠損，欠落ともいいます）してしまうことが多く，その際の統計的推定や検定問題を考えることはとても重要です。

例えば，**表１**はある科目について４回の試験結果をまとめたもので，学籍番号や得点に関係なくランダムに並べています。一部省略していますが，試験を受けた学生は50名です。

表１　４回の試験の点数（完全データ）

学生	１回目	２回目	３回目	４回目
1	61	28	57	61
2	40	40	21	61
3	59	59	78	66
4	34	42	54	66
⋮	⋮	⋮	⋮	⋮
50	73	54	73	40

注：次元は４，サンプル数は 50 です。

データは，分かりやすく説明するために実際のデータを少し加工して作成していますが，このようにデータがすべて揃っているデータを完全データといいます。これに対して，データが**表２**のように欠測しているとします。１～30番の学生は４回すべてのテストを受けていますが，31～40番の学生は３回目まで受けて４回目のテストを受けていなく，41～50番の学生は１，２回目は受けて３，４回目のテストを受けていないというデータです。このデータのように一度欠測したら，それ以降の時点で測定されないデータを「単調欠測データ」と呼んでいます。

すなわち，単調欠測データとは，一つの観測ベクトルに対して一度欠測が生じるとそれ以降の成分がすべて欠測となっているデータ

表２　４回の試験の点数（単調欠測データ）

学生	１回目	２回目	３回目	４回目	
1	61	28	57	61	
2	40	40	21	61	
⋮	⋮	⋮	⋮	⋮	30名
30	33	33	64	45	
31	58	49	21	*	
32	61	38	94	*	
⋮	⋮	⋮	⋮	⋮	10名
40	50	67	71	*	
41	36	24	*	*	
42	46	24	*	*	
⋮	⋮	⋮	⋮	⋮	10名
50	73	54	*	*	

注：＊は欠測値を表しています。

からなる観測ベクトルの集まりのことです。一般には，欠測パターンごとに人を並べ替えることによって，データを階段状にすることができます。特に表２のデータの場合，すでにデータが階段状になっており，３段あるので3-step単調欠測データと呼んでいます。もちろん，実際には単調欠測データでない欠測データも数多くあり，その場合の分析は容易ではなく，分析法として数値解析によるものがありますがここでは触れないことにします。また，欠測値の有無はデータに全く無関係であると仮定して議論していきます。

データ解析の基本として，平均ベクトルμ（各次元の母平均を並べたベクトル）や分散共分散行列Σ（各次元の分散と各次元間の共分散を並べた行列）を推定する問題があります。表１の完全データの場合，平均ベクトル（４回の試験の母平均を成分とするベクトル）の推定として，通常各列の50名の標本平均ベクトル（54.24，50.84，65.52，58.90）が用いられますが，これは統計学的に最尤（さいゆう）推定量という，良い性質をもつ推定となっています。それに対して，完全データの場合と同様，単調欠測データの場合も，尤度（ゆうど）関数というものを与えることができるため，それを偏微分することによって最尤推定量が得られます。表２のデータのμの最尤推定値は，$\hat{\mu}$＝(54.24，50.84，65.60，60.82）となり，これが単調欠測データの場合の平均ベクトルμの推定値となります。

一方，表２で列ごとに平均をとった値は，(54.24，50.84，64.95，60.23）であり，先程の最尤推定値と比べると前半の２次元は一致するのですが，後半の２次元は一致しないことが分かります。その理由は，$\hat{\mu}$の後半の２次元はΣの最尤推定量を利用して求められているからです。Σの推定についても重要な問題のひとつとなっていますが，ここでは省略します。

平均ベクトルの検定

次に，例えば，平均ベクトルがある既知のベクトル（μ_0＝(55，60，60，65）とおきます）と等しいかどうかを統計的に検定することを考えてみます。調べたいことについての仮説をたて，データからその仮説が正しいかどうかを統計学を用いて判断することを「統計的仮説検定」といいます。特に，μを平均ベクトルとするとき

$$H_0 : \mu = \mu_0$$

が成り立つかどうかという仮説（H_0を帰無仮説といいます）を統計的に判断することを「平均ベクトルの検定」といいます。

［完全データの場合］

まず，完全データの場合の具体的な手順を紹介していきます。平均ベクトルの検定では，「平均ベクトルがある既知ベクトルと等しいかどうか」を統計的に判断するために，相関を考慮した上で「標本平均ベクトルと既知ベクトル」の距離に基づく量（検定統計量といいます）を用いて検定します。この検定統計量はホテリングのT^2検定統計量と呼ばれ，標本平均ベクトルと標本分散共分散行列の逆行列を用いて表される量になっています。

表１のデータからこのT^2の値を表計算ソフトなどを用いて計算してみると，31.15になります。ここで，気になってくるのが，「今回と同じようなデータセットは，もしも帰無仮説が正しいと仮定した場合にどれくらいの程度で出現するデータセットなのか」ということです。そのためにPr($T^2 \geq 31.15$）を考えてみます。

帰無仮説が正しい下で，このT^2はF分布の定数倍に従うことが知られています。実際に，正規乱数を発生させて100万回のモンテカルロ・シミュレーションを行ったときの，T^2のヒストグラムは**図１**のようになっており，青い曲線とほぼ一致していることが分かります。この曲線がF分布を定数倍した確率密度関数なのです。

このF分布の上側確率を利用すると，$\Pr(T^2 \geq 31.15) = 1.21 \times 10^{-4}$と理論的に求まります。このことは，同じような実験を１万回繰り返した場合に１回くらいしか得られないデータセットだったことを意味します。この場合，帰無仮説が正しくて，めったに起こらないことが偶然に起こったと考えるよりも，帰無仮説は正しくなかったとして，帰無仮説を否定します。すなわち$\mu \neq \mu_0$と判定し，このデータの母平均ベクトルの値は（55，60，60，65）と等しいとはいえないという結論が得られます。

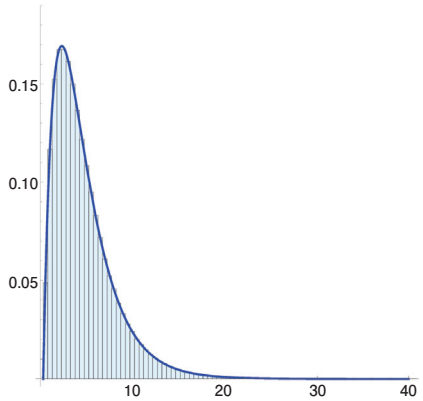

図１　T^2のヒストグラムと理論的に求められた確率密度関数

統計解析の際には，$\Pr(T^2 \geq c) = \alpha$となるようなcが判断基準として使われ（cを上側100α％点といい，$\alpha = 0.05$や0.01などが用いられます），

「データから得られる統計量の値」＞c

ならば「帰無仮説を棄却」

として検定を行います。この問題では$\alpha = 0.01$とすれば，T^2の分布がF分布の定数倍であることから$c = 16.01$となり，$31.15 > 16.01$なので，有意水準１％で帰無仮説を棄却ということになります（**図２**）。

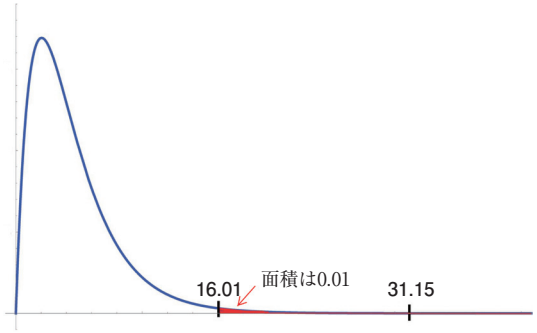

図２　T^2の分布と上側１％点

［単調欠測データの場合］

先ほどのホテリングのT^2検定統計量は使えないので，今回は，新たにT^2型検定統計量を提案します（$\tilde{T}^2$と呼びます）。$\tilde{T}^2$の詳しい定義は割愛しますが，単調欠測データにおけるμとΣの最尤推定量を用いて$\tilde{T}^2$を定義します。そして，データから$\tilde{T}^2$の値を計算すると，その値は，24.17となります。

次に，帰無仮説が真の下で$\Pr(\tilde{T}^2 \geq \tilde{c}) = \alpha$となるような$\tilde{c}$（$\tilde{T}^2$の上側$100\alpha$％点）を導出することが必要になってきますが，これを与えることが完全データのときとは異なり，難しい問題になります。この問題を解決するために，いくつかの方法があるのですが，例えば，カイ二乗分布の上側100α％点を近似として利用する方法（従来の方法と

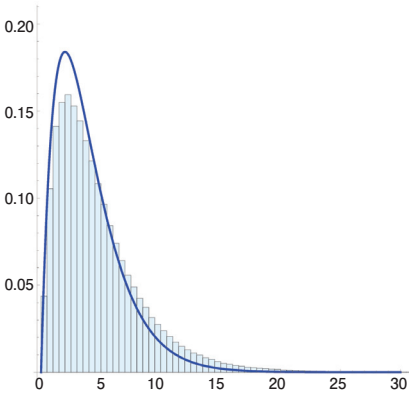

図３　$\tilde{T}^2$型統計量のヒストグラムとカイ二乗分布の確率密度関数

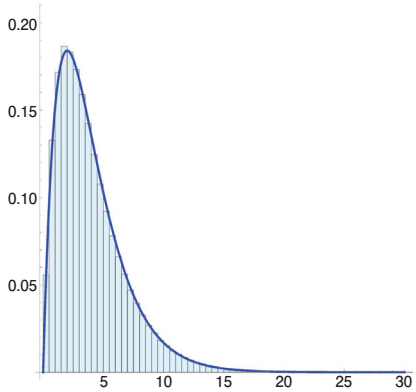

図４　変換統計量のヒストグラムとカイ二乗分布の確率密度関数

呼ぶことにします）があります。ただ，この方法は，サンプル数が少ない場合，近似精度があまり良くないことが知られており（**図３**），良い近似を考える必要が出てきます。

そこで，$\tilde{T}^2$を完全データに類似した形に分解すると，漸近展開と呼ばれる方法が適用できるようになり，近似上側100α％点を求めることに成功しました。表２のデータの場合，このようにして求められた近似上側１％点（＝16.06）を判断基準とすれば，24.17＞16.06であるので帰無仮説は棄却されます。また，他のアイデアとして，$\tilde{T}^2$（検定統計量）を変換し，変換後の分布をカイ二乗分布とみなすことができるような変換統計量を導出しました。このような変換統計量をいくつか提案しているのですが，そのうちのひとつが**図４**のようになっています。表２のデータから得られる変換統計量の値は，19.53であり，判断基準となる値として，カイ二乗分布の上側１％点（＝13.28）を用いれば，19.53＞13.28となるので，帰無仮説を棄却ということになります。これは，先程の完全データの場合と同じ検定結果となっています。

最後に，提案した近似上側100α％点や変換統計量の近似精度を数値的に評価するために，モンテカルロ・シミュレーションを行った結果について紹介します。

今回は「第一種の過誤」が５％に近いかどうかで近似精度を評価します。ここで，第一種の過誤とは，「帰無仮説が真であるのにもかかわらず，帰無仮説を偽として棄却してしまう誤りのこと」をいいます。いくつかのパラメータに対して実験を行っていますが，その結果の一部が**図５**のようになっています。これは，3-step単調欠測データにおいて，１段目の完全データの次元を４，２段目と３段目の次元をそれぞれ３と２とし，２段目と３段目を１段目の半分のサンプル数として，１段目のサンプル数を増やしていった場合の第一種の過誤（％）を表しています。従来の方法よりも，漸近展開近似による方法や変換統計量による方法の方が，第一種の過誤がより５％に近く，近似精度が良くなっていることがみてとれます。

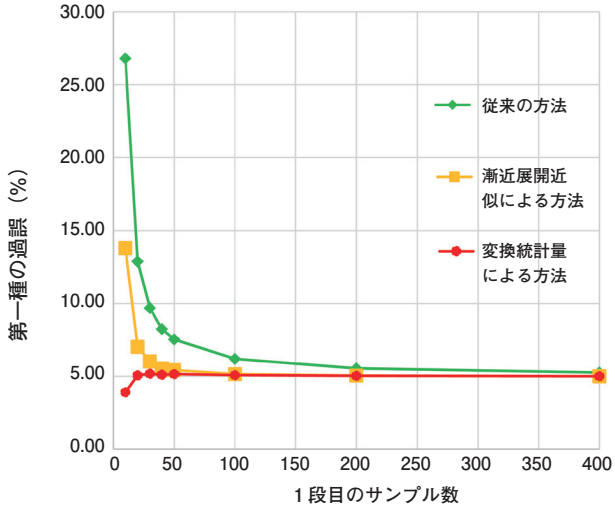

図５　3-step単調欠測データの第一種の過誤

ちなみに，欠測が起きている２，３段目（計20人分）のデータを用いず，１段目（30人分）の完全データのみを用いて検定すると，帰無仮説は棄却されないという結果が得られます。このように，今回提案した方法が有用であることが分かります。

おわりに

多次元データが単調に欠測してしまった場合にも，平均ベクトルの検定ができることを紹介しました。これは母集団が１つの場合でしたが，その拡張として，母集団が２つ以上に増えた場合の検定法について研究を進めています。完全データの場合は，T^2検定は尤度比検定というものになっていますが，欠測データの場合はそうではないことから尤度比検定についても考えています。他にも単調欠測データにおける成長曲線モデルの下での推定問題についても取り組んでいます。