様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 ２６ 日現在

機関番号：３５３０２
研究種目：基盤研究(C)（一般）
研究期間：2018～2020
課題番号：１８Ｋ１１２０６

研究課題名（和文）高次元データに対する正規性の検定

研究課題名（英文）Normality test for high dimensional data

研究代表者
中川　重和（Nakagawa, Shigekazu）

岡山理科大学・総合情報学部・教授

研究者番号：９０２４８２０３

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：本研究では，高次元データの特徴を多変量正規分布に従うかどうかで捉えることとし，その判定法を構築する．そのための基礎理論をより強固にするために，標本歪度および標本尖度分布の正確表現を与える（課題1，2）．最終課題を「標本歪度・尖度分布の正確表現に基づいた高次元データにおける正規性の検定」（課題3）と設定する．
課題1，2に関し，超幾何関数を駆使して基礎理論を構築しようとしたが，思うように捗らず，そこで，標本歪度分布の確率密度関数のフーリエ展開による近似表現にたどり着いた．これにより，標本歪度分布の確率密度関数のフーリエ展開表現を得ることができ，arXivへ投稿した．

研究成果の学術的意義や社会的意義
本研究では，標本歪度分布の精確表現を得た．従来法では簡潔に表現できていなかったことが，今回はより精確にそして簡潔に表現できるようになった．この視点において学術的意義深いと思われる．また，フーリエ展開表現が標本尖度分布へも応用が期待できる点も大事である．
高次元データの特徴を多変量正規分布の正規性の観点から捉えるという視点では，本研究は不完全であるが，データ解析の基礎部分の再構築という視点で評価できる．従って，データサイエンス教育の観点で社会的意義があると思われる．

研究成果の概要（英文）：In this research, we consider the characterization of high-dimensional data based on whether or not they are drawn from a multivariate normal population, and testing for normality. We give concrete representations of the pdfs of sample skewness or kurtosis (Tasks 1 and 2). The final task is set as "normality test in high-dimensional data based on those sampling distributions".
Regarding tasks 1 and 2, we tried to construct a basic theory by using hypergeometric functions, but it did not progress as expected. So, we arrived at an approximate expression by Fourier series of the pdf of the sample skewness. As a result, our manuscript related to a Fourier series representation is submitted to arXiv.

研究分野：統計科学

キーワード：高次元データ　正規性の検定　歪度　尖度

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

遺伝子データに代表される高次元データが容易に取得できるようになった現在，高次元データ解析手法の開発は急務である．データの次元dと標本数nの関係により，いくつかの高次元の設定が考えられる．高次元小標本の場合や，高次元大標本の場合などである．いずれにせよ，$d > n$の状況では，分散共分散行列が特異になってしまうなどの不都合が起こり，従来，$n > d$ の状況で考えられてきた手法をそのまま適用することは不可能である．

青嶋他(2013)は，高次元大標本，小標本のいずれの状況でも，高次元空間に現れる幾何学的特徴を正確に把握した上で，既知の統計量と関連づけることにより，様々な高次元データ解析手法の開発に成功した．また，同じ枠組みで，　Glombek は高次元大標本のときの正規母集団における母共分散行列の sphericity 検定統計量として，Jarque--Bera 統計量の漸近分布を導いた．幾何学的特徴以外に高次元データを捉える術はないだろうか．これは，先行研究を顧みた際に抱いた素朴な疑問である．もしも得られた高次元データの従う分布が正規分布であるとの特徴付けが可能ならば，先の疑問へのひとつの解答となり得る．

そこで，高次元の設定として n を固定し$d \to \infty$ とする場合を考え，さらには，多変量正規分布の性質「確率ベクトルが多変量正規分布に従うならば，その任意の部分集合も多変量正規分布に従う（特に，成分は一変量正規分布に従う）」の対偶も考慮に入れると，高次元データの従う分布が正規分布であるかどうかの判定問題は，一変量の正規性の検定問題に帰着できる．従って，本研究では高次元データに対する正規性の検定の開発をこのような立場から行う．

２．研究の目的

本研究では，正規母集団からの標本歪度および標本尖度の正確分布（近似分布ではなく，解析的表現）を与えることを第 1 の目的とする．第 2 の目的は，これらの結果を用い，n を固定し $d \to \infty$とするときの高次元データに対する正規性検定の構築である．nを固定させるためには，標本歪度・尖度の正確分布導出が不可欠になる．そのために以下の課題を設定する．

・標本歪度および標本尖度分布の正確表現を与える（課題 1，2）．
・標本歪度・尖度分布の正確表現に基づいた高次元データにおける正規性の検定（課題 3）

本研究の独自性は，高次元データの特徴を正規性を有するか否かで表現することにある．また，本研究の創造性は，小標本での標本歪度・尖度の分布の解析的表現に再度着目し，2017 年の研究業績 (主成分に基づく多変量正規性の視覚的検定)ならびに 2016 年の研究業績 (標本歪度・尖度の同時密度関数および同時モーメントの標本の大きさに関する漸化式表現)を積極的に活用し，与えることにある．

課題 1，2 の解決は，正確な棄却域の構成へと直結する．例えば，有意水準 5% の両側棄却域は $\{s_3 | \ |s_3| > 0.7071\}$ となる．

課題 3 の解決は，例えばある種の遺伝子データは多変量正規分布に従うけれども，別種のそれはそうではないといった，高次元データの分類の目安になり得る．以上のように，筆者は，本研究が高次元データ解析の発展へ寄与するだけでなく，正規性検定の精度を向上する点においても，大変重要であると確信している．

３．研究の方法

課題１.正規標本からの標本歪度の密度関数の解析的表現

大きさnの正規標本からの標本歪度とその密度関数をそれぞれ$s_n, f_n(s_n)$とする．また，ガウスの超幾何関数を$F(a;b;c;x)$とする．$f_3(s_3), f_4(s_4)$については，ガウスの超幾何関数による表現が知られているが，しかしながら，$n \geq 5$の場合についての解析的表現は知られていない．そこで，Mulholland（1977）による nに関する漸化式を用いると，密度関数 $f_n(s_n)$ はガウスの超幾何関

数 $F(a;b;c;x)$で表現できると予想できる.

以下に具体的手順を示す.
1.計算機実験により超幾何関数のパラメータ a,b,cを予想.
2.既知の結果 $f_3(s_3),f_4(s_4)$の関係から，漸化式の積分の果たす役割を解明.
3.一般の n の場合の分布形を特定.
4.数値実験による検証.

課題 2.正規標本からの標本尖度の密度関数の解析的表現
標本尖度t_nの分布導出は標本歪度$g_n(t_n)$のそれより困難であることが予想される．McKay の結果は，$g_4(t_4)$のガウスの超幾何関数表示を与えているが,$n \geq 5$，$g_n(t_n)$の解析的表現は知られていない．2016 年の研究業績の結果(s_n,t_n)の同時密度関数$h(s_n,t_n)$の漸化式を利用する．課題 2 の解決においても積分評価が最大のポイントになる.

課題 3．標本歪度・尖度分布の正確表現に基づいた高次元データにおける正規性の検定
2017 年の研究業績の多変量正規性の検定手法を高次元データに拡張することが基本的アイデアである．各変量の一次結合であるd本の主成分を求め，その標本歪度・尖度の値を，それらの同時確率密度等高線が描かれた平面上にプロットする．このとき，各主成分の標本歪度・尖度の値が有意水準に対応する確率密度等高線で囲まれた領域に全て入るならば，正規母集団からの標本とみなしてよいと結論できる．高次元データへ拡張するには，通常の主成分分析を「ノイズ掃き出し法」もしくは「クロスデータ行列法」による高次元主成分分析に置き換えることで同じ枠組みで高次元データへ拡張できる.

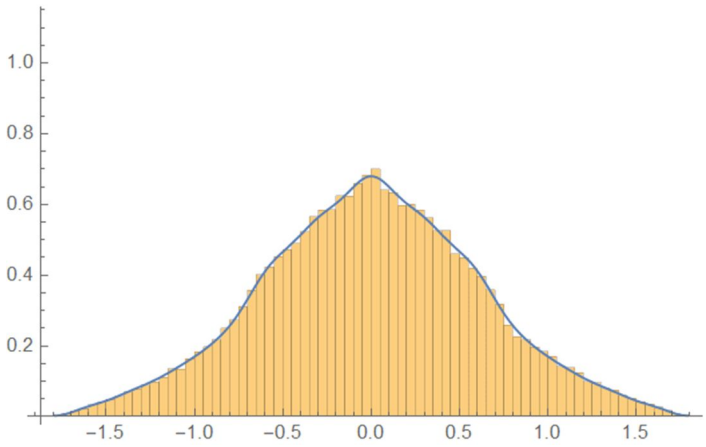
４．研究成果

(1) 「正規性の検定」（共立出版，2019）の執筆
本書は，正規性の検定への準備（第 1 章），正規性の検定の基本的枠組（第 2 章），様々な正規性の検定（第 3 章），標本歪度・尖度に基づく検定（第 4 章），検出力比較（第 5 章），標本歪度・尖度分布の再考（第 6 章），および補遺（付録）からなる．正規性の検定に関する基本事項から書き進め，本研究へ至る着想（第 6 章）についてもまとめている．また，研究成果を社会・国民へできるだけわかりやすく説明するため観点で寄与している．さらに，本研究課題へ至る着想などを整理整頓し，新しいアイデア創出に役立った.

(2) 標本歪度分布のフーリエ級数展開近似

Approximation to probability density functions in sampling distributions based on Fourier cosine series,
Shigekazu Nakagawa, Hiroki Hashiguchi, Yoko Ono,
ArXiv:2103.117.12,　22, Mar., 2021

大きさnの正規標本からの標本歪度をs_nとし，その密度関数を $f_n(s_n)$とする．研究当初は，ガウス超幾何関数での表現を求めていたが，うまくいかなかった．そこで，$f_n(s_n)$が対称分布であること，および歪度が有限な値しか取らないことから，フーリエ級数展開近似が考えられるが，$f_n(s_n)$は未知であるから，直接的にはフーリエ係数を求めることができない．この点を克服する鍵となるのが，s_nのモーメントである．これを用いてフーリエ係数を構成することが可能となり，フーリエ級数展開による$f_n(s_n)$の近似が可能となった．本提案手法は，特に，小標本で有効であることも分かった．既存結果より精度がよくなっていることも分かった．図は，提案手法で得た$f_6(s_6)$のグラフとモンテカルロ法（繰り返し数は10^6回）によるヒストグラムである.


1.0
0.8
0.6
0.4
0.2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Mura, H., Hashiguchi, H., Nakagawa, S. and Ono, Y.	４．巻 55
２．論文標題 Holonomic properties and recurrence formula for the distribution of sample correlation coefficient	５．発行年 2019年
３．雑誌名 SUT journal of mathematics	６．最初と最後の頁 39--52
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Hiroki Watanabe, Masashi Hyodo, Shigekazu Nakagawa	４．巻 -
２．論文標題 Two-way MANOVA with unequal cell sizes and unequal cell covariance matrices in high-dimensional settings	５．発行年 2020年
３．雑誌名 Journal of Multivariate Analysis	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.jmva.2020.104625	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Shigekazu Nakagawa, Hiroki Hashiguchi, Yoko Ono	４．巻 2103.11712
２．論文標題 Approximation to probability density functions in sampling distributions based on Fourier cosine series	５．発行年 2021年
３．雑誌名 ArXiv	６．最初と最後の頁 2103.11712
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 Shigekazu Nakagawa, Hiroki Hashiguchi and Yoko Ono
２．発表標題 New Omnibus test for normality based on a moment ratio
３．学会等名 Statistical Computing:Challenges and Opportunities in Data Science（国際学会）
４．発表年 2018年

〔図書〕　計1件

１．著者名 中川 重和	４．発行年 2019年
２．出版社 共立出版	５．総ページ数 148
３．書名 正規性の検定	

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	橋口　博樹 (Hiroki Hashiguchi) (50266920)	東京理科大学・理学部第一部応用数学科・教授 (32660)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関