言語（教育）研究における多変量解析の初歩と注意点：各手法の特徴、注意事項、ソフトウェアに関して

The First Step for Multivariate Analyses by Novice Researchers: Statistical Features, Notes and Software for Multivariate Analyses

仁科　恭徳

NISHINA　Yasunori

投稿日：2020 年 2 月 8 日
受理日：2020 年 10 月 15 日

グローバル・コミュニケーション学部

（要約）

本稿では、統計初心者には取り扱いが難しいとされる多変量解析に関して、各手法の特徴と各々の違いを簡潔にまとめ、筆者の研究の経験から得たいくつかの注意事項も紹介した上で、最終的に現在使用可能な統計ソフトを紹介する。特に、多変量解析においては、変数の選択や予測と要約の二つの手法に関する理解が必要不可欠である。また、統計ソフトから自動的に生成された各多変量解析手法の結果の解釈においても、留意すべき点もある。このような点に関して、統計初心者にも分かりやすく多変量解析を紹介し、統計初心者の心の障壁をなくすことを本稿の目的とする。

キーワード：多変量解析、統計、言語研究、言語教育研究

1．はじめに

統計を用いた分析を苦手とする文系研究者は少なくない。しかしながら、言語（教育）研究に限定すれば、収集したデータを客観的に分析しエビデンスを提示する上で、統計の知識は必要不可欠である。中でも本稿で取り上げる多変量解析は、有意差検定（χ^2検定やt検定）よりもハードルが高いため、使用することに躊躇している文系研究者も多いのではないだろうか。また、安易に使用することで、算出結果の解釈に孕んでいる危険性に気づかないものも多い。

最近では、学際的見地から異なる専門分野の研究者が共に特定の研究テーマに取り組むような共同研究も増えている。この場合、データ収集から分析手法、結果の解釈に至るまで、互いに把握・理解している必要がある。しかし、統計に長けた研究者とそうではない研究者の双方が、研究で使用する統計分析の目的や結果の解釈などに関してある程度の共通理解がなければ、共同研究自体が頓挫する可能性もある。各統計手法のアルゴリズムを正確に理解できなかったとしても、各手法で取り扱う変数の違いや目的、生成された結果の解釈の仕方に関しては最低限把握している方がよい。可能であれば、簡単な統計分析は自分で実施できるとなおよい。特に、近年の言語（教育）研究では、一般的な有意差検定に加えて多変量解析を用いた論文の数は増える一方である。これは言語や教育活動に潜む複雑な事象をシンプルに説明したり予測することを研究目的としているためである。

そこで本稿では、統計手法の中でも、このように今や文系の研究においても需要が高くなった多変量解析に注目し、各手法の根本的な目的や特徴、その異なり、生成結果に関して気をつけるべき点、統計ソフトの紹介などを独自の視点からまとめる。

2．多変量解析とは

コンピュータが到来してからはビッグデータの時代へと移行し、言語学の諸分野においても、客観的な解析データをもって言語事象を解明するコーパス言語学が隆盛を迎えている。膨大な言語データを多変量解析にかけることで、今まで母語話者の直感では気づかなかったような新たな言語事実が明らかになりつつある。奥野他（1971）では、「どんな対象についても、その特徴を把握するには多種類のデータを求めるのが普通で、これを多変量・多変数のデータ」（p.1）と呼び、多変量解析法とは「多変量のデータを的確に評価する手法」（p.2）であり、「互いに相関のある多変量（多種類の特性値）のデータのもつ特徴を要約し、かつ所与の目的に応じて総合するための手法である」（p.2-3）と説明している。複雑な情報をシンプルに解釈し表示する分析手法であることから、マーケティング調査や言語の量的分析等と相性がよい。ただし、統計ソフトを使えばデータを挿入しいくつかの選択項目をクリックするだけで何かしらの結果が算出されるため、事前に正しくデータ処理を施し、正しい知識をもって結果を解析しなければ、誤った結論を導く可能性もある（特に第３節において、（筆者の経験から）多変量解析の実行において気をつけるべきいくつかの点をまとめたので、そちらも参照されたい）。

2．1．変数の違い

多変量解析には様々な手法がありその選択肢は幅広い。朝野（2000）、Larson-Hall（2015）、竹内・

水本（2012）などの良書や田畑（2004a）には言語学や言語教育で活用できる統計手法に関して詳説されている。また､フリーソフトのＲを用いて多変量解析を詳説している藤井(2010)や中村(2009)においては、数式も多く少々テクニカルな解説だが参考になる。これらの良書でまとめられているように、多変量解析の各手法の決定的な違いの一つは、扱う変数の異なりである。変数には量的変数と質的変数があり、扱う変数がどちらに類別されるかは、そのデータの尺度によって決まる。

表１が示すように、データの尺度は大きく四つに分かれている。名義尺度は、他と区別・分類するための名称や記号、番号などを指す。順序尺度は、順序や大小のみに意味があるが、その間隔には意味がないものを指す。間隔尺度は、目盛りが等間隔になっており、その間隔自体に意味があるものを指す。比例尺度は、０を原点として間隔と比率に意味があるものを指す。名義尺度と順序尺度は質的変数として、間隔尺度と比例尺度は量的変数と見なされる。各変数の意味をしっかり理解していなければ、誤った手法を選択する可能性もあるため、多変量解析を実施する上では、ここが出発点であり重要でもある。

表１．変数の尺度のまとめ

種類	尺度の説明	例	質・量の区別
名義尺度	他と区別・分類するためのもの	男女、血液型、郵便番号、学籍番号など	質的
順序尺度	順序に意味があるが間隔には意味がないもの	１位／２位／３位のような順位、１．好き／２．ふつう／３．嫌いなどの好み、１級／２級／３級など検定試験のレベルなど	質的
間隔尺度	目盛りが等間隔になっているもの	温度、（テストなどの）点数、知能指数など	量的
比例尺度	原点があり、間隔や比率に意味があるもの	身長、時間、速度、長さ、重さなど	量的

多変量解析はマーケティングの分野でも重宝される解析手法である。多変量解析に関する有益な解説はWeb上でも容易に手に入る。表１をはじめ、本稿の内容も上掲の良書や資料に加え、例えば、マーケティングリサーチのマクロミル（https://www.macromill.com/service/data_analysis/multivariate-analysis.html)、データ分析基礎知識 Albert（https://www.albert2005.co.jp/knowledge/statistics_analysis/multivariate_analysis/multivariate_basis)、intage の多変量解析とは（https://www.intage.co.jp/glossary/056/)、株式会社アイスタット統計分析研究所（https://istat.co.jp/ta_commentary)、日経リサーチの多変量解析（https://www.nikkei-r.co.jp/glossary/id=1643）なども参考にしている。言うまでもなく、これらのウェブサイトで詳説されている内容は重複している部分も多いが、いずれもターゲットとする読者が一般人向けのため、解説が親切なものが多くて文系研究者にも参考になろう。

２．２．多変量解析における二つの手法１：予測の手法

既に紹介した統計書や統計を解説したウェブサイトなどにもまとめられているように、多変量解析は大きく分けて「予測」と「要約」の二つの手法に分けることができる。まず､「予測」の手法は、

文字通り原因から結果を予測する手法であり、重回帰分析、判別分析、ロジスティック回帰分析、数量化Ⅰ類、数量化Ⅱ類などがある。これらの手法の根本的な違いは、表２に挙げるように、説明変数（つまり原因）と目的変数（つまり結果）における変数の選択にある。なお、数量化の基本的な概念とは、質的変数を擬似的に数値化したダミー変数を用いることで多変量解析を実行する手法と考えてよい。

表２．予測の手法の例

	説明変数（原因）		目的変数（結果）	
	量的変数	質的変数	量的変数	質的変数
重回帰分析	○		○	
判別分析	○			○
ロジスティック回帰分析	○			○
決定木	○	○	○	○
数量化Ⅰ類		○	○	
数量化Ⅱ類		○		○

まず､重回帰分析は､一つの目的変数（量的変数）を二つ以上から成る複数の説明変数（量的変数）で予測する分析手法である（単回帰分析は、一つの目的変数（量的変数）を一つの説明変数で予測する分析手法である)。例えば、５教科のテストの合計点を数学と英語の点数から予測するような場合に用いる。どの要因がどの程度影響しているかを算出し、結果の予測のみならず、その精度を知ることも可能である。重回帰分析は第二言語習得研究などの言語教育分野で頻繁に用いられている手法であり、例えば、Dethorne *et al.*（2005)、Zareva（2005)、Paavola *et al.*（2005）などがある。また、同手法を用いた言語研究の例として、李（2016）では、日本語の読解の授業で参考となる文章難易度を予測するリーダビリティ公式を、重回帰分析を用いて作成している。

次に、判別分析では、説明変数に量的変数、目的変数に質的変数を選択する必要があり、例えば、英語を学習した合計時間から英検２級に合格するかどうかといった判別を予測する場合などに使える。石川・前田・山崎（2010）では、判別分析に関して手際よく詳説されているので是非参考にされたい。判別分析を用いたコーパス言語学研究に、日本人英語学習者書き言葉コーパスを用いてN -gram 分析を実施し、CEFR レベルの基準となる特性を解明した能登原（2015）がある。

ロジスティック回帰分析も説明変数では量的変数、目的変数では質的変数を選択する必要があるが、判別分析で紹介した例で例えると、どの程度の確率で合格するかどうかを予測することができる。見目（2004）では、英語の軽動詞句と意味的に等価であろうと仮定される１語動詞との分別において（この研究では *decide* と *make a decision*)、時制や極性、レジスターなど13の因子群のどれが判別要因となるかを、ロジスティック回帰分析を用いて分析している。

決定木（decision tree）は、樹木構造でデータを分類していく手法で、説明変数・目的変数共に量的変数から質的変数まで利用することができる。視覚的にも結果を解釈しやすく、外れ値に対しても頑健であるという長所もある[1]。決定木にはいくつかのアルゴリズムがあり、代表的なものにGini 係数を用いて分岐を行う CART がある[2]。今道（2013）では、ドイツ語の語尾選択に関わるとされる規則がどの程度の精度で反映されているのか、そして、反映されていない場合に名詞にはど

のような特徴が見られるのか、という二つの問いに取り組み、後者に関してCARTを用いて分類モデルを構築し分析している。なお、決定木とアンサンブル学習を組み合わせた機械学習法であるランダムフォレストという手法も近年のコーパス言語学の研究では使われることが多くなった[3]。小林・金丸（2012）では、英語学習者が作成した英作文の様々な量的データ（例えば、平均の文長や代名詞の頻度など）を説明変数とし、コンピュータによる英語力レベル判断を目的変数としてランダムフォレストを実行し、総語数と異語数が目的変数に最も影響していることを明らかにしている。他にもサポートベクターマシン（SVM）やニューラルネットワークといった手法があるが、Rを用いた決定木、SVMなどに関しては金森・竹之内・村田（2009）を参照されたい。

数量化Ⅰ類は、重回帰分析と類似した手法である。しかしながら、説明変数として質的変数を選択する必要があり、男性と女性、大人と子供、好きと嫌い、TOEICの点数が400以下、401～700、701以上などを１と０、１と２と３などダミー変数としての識別コードに変換することで量的変数を予測することが可能となる。例えば、ある教育手法を用いて教えた学生とそうでない学生を１と０で表し、英語力が上昇したかどうかの確率を予測することが可能である。富田･須田･五十嵐･佐々木（2005）では、高校１年生1409名（有効回答数1,240名）を対象に、数量化Ⅰ類を用いてライフスタイルが疲労感に与える要因を分析し、その関与の大きさを調査している。

最後に、数量化II類は、判別分析と類似しているが、説明変数も目的変数も質的変数である場合に用いることができる。判別分析と同じく、既に分かっている傾向や結果のデータを活用しモデルとなる数式を作成することで、新しいサンプルがどのように分類されるかを予測する分析手法である。服部・野々上・門田（2009）では、小学生445名を対象に、ライフスタイルに関する23要因を説明変数とし、普段の自覚症状の訴え数を目的変数として、数量化II類を用いて分析している。

２．３．多変量解析における二つの手法２：要約の手法

多変量解析のもう一つの手法は「要約」の手法である。この手法には主成分分析（PCA: Principle Component Analysis)、因子分析（FA: Factor Analysis)、クラスター分析（Cluster Analysis)、対応分析（Correspondence Analysis)、（林の）数量化III類、多次元尺度法（MDS: Multi Dimensional Scaling）などがあり、いずれも複雑な情報を統計的に精査しシンプルに表示するための次元縮小法であることから、根本的に類似した解析法と捉えてよい。予測の手法とは異なり、説明変数や目的変数という概念がそもそもない。英語教育やコーパス言語学で好まれて使用される解析手法でもある[4]。なお、コーパス言語学で多用される多変量解析の解説は、Oakes（1998）が詳しい。

表３．要約の手法の例

	量的変数	質的変数
主成分分析	○	
因子分析	○	
クラスター分析	○	
対応分析［コレスポンデンス分析］		○
数量化III類		○
多次元尺度（構成）法[MDS]		○

* 対応分析はコレスポンデンス分析、多次元尺度法はMDSとも呼ぶ。

まず、主成分分析は、多数の変数を少数の指標に要約して表現する方法である。主成分分析では、少数の変数に合成して要約し、全体的な要約になりやすい第一主成分得点と、より変数間の差異が明らかとなる第二主成分得点を用いて二次元プロットで可視化している研究が多い。実際の分析の手続きとしては、一般的に主成分分析では、（データをどの程度説明しているかを表す）固有値が１以上あるいは累積寄与率 90% 以上などに適合した有効な主成分のみを分析に採用するのが慣例となっており（石川・前田・山崎 , 2010 参照）、該当する主成分が三つ以上採用される場合はそれぞれを組み合わせた各散布図からサンプルがうまく分類されているかを判定するのが一般的である。主成分分析は、Nishina（2007）、水本（2009）、水本・野口（2009）、田畑（2004b）などのコーパス言語学研究において、語彙頻度に基づき、多数のテクストを分類する手法として用いられている。また、石川（2008）では、265 名の大学生から獲得した英作文データを用いて、延べ語数や異なり語数、語長、文数などの 10 種類の指標から主成分を抽出し、主成分得点に基づいてエッセイのタイプを診断することで、エッセイの自動評価手法を模索している。

次に、因子分析は、多数の変数の背後に潜んでいる少数の概念を抽出することで複雑な現象の原因を探る手法であり、主成分分析とは対極に位置している。石川（2009）では、３変数１因子モデルの探索的因子分析を用いて、1990 年代の 100 万語のアメリカ英語コーパスである FROWN コーパス中に生起した３種の頻度副詞 *often*、*sometimes*、*rarely* の頻度データを分析し、その共通性（共通因子）と独自性（独自因子）を調査している。

クラスター分析は、共通の特性をもつ変数あるいはサンプルをグループ化する手法で、階層的手法と非階層的手法がある。非階層クラスターは事前にグループ数を決定しておく必要があるが、階層クラスターはその限りではない。水本（2008）はクラスター分析を用いて、大学生 40 名が書いた自由英作文から抽出した７種の語彙指標と TOEIC 模擬テスト、評定者によるエッセイの総合的評価の計９種の変数の関係性を可視化した（詳しくは、水本（2008）を参照）。吉村他（2019）では、英語ドラマ制作活動時に学生に課したアンケート調査結果をクラスター分析した結果、学習者は成功したグループ、部分的にうまくいったグループ、相対的に失敗したグループに分類され、各グループにおいて協同学習を実施する上で重要とされる八つの原理の影響度合いが異なることが確認された。

対応分析（コレスポンデンス分析とも言う）と数量化 III 類は、主成分分析と同じ目的で用いる手法で、量的データの場合は対応分析、質的データの場合は数量化 III 類を用いる。次節で詳細に述べるが、主成分分析よりも対応分析の方が二次元プロットにマッピングした結果に変数間の違いが現れやすい。対応分析はコーパス言語学の研究で多用される統計手法であり、Nishina（2007）における高頻出語を用いたジャンル分析や、小林（2008）における高頻出語を用いた学習者コーパスの分類分析などがある。数量化 III 類に関しては、1960 年代の 100 万語のイギリス英語コーパスである LOB コーパスを分析した古橋・高橋（1995）や、1990 年代の１億語のイギリス英語コーパスである British National Corpus を属性指標に基づきサブコーパスに細分化し分析した高橋 (2019) がある。特に後者では、「元来、性差のみの二値尺度については、性差を区分する尺度が現れにくい」(p.221) とされていたが、５段階の年齢グループと性差による 10 個のサブコーパスの言語を解析することで、男女を明確に分ける尺度を断定することに成功している。

多次元尺度（構成）法（以下、MDS）は、二つの対象間の違い（距離）をもとに、その距離を保つようにしながら、可能な限り低次元（通例、二次元）で各点の座標を求め、データの構造を考察

する手法である。この手法では、個体間の類似性はその距離に反映されている。計量多次元尺度法と非計量多次元尺度法に分かれる。MDS はコーパス言語学の分野において、特に 23 種の話し言葉と書き言葉のジャンルの言語的特徴を分析した Biber（1988）やその語の Biber の一連の研究は有名である。他にも、Nesi（2009）では、イギリスの学生の書き言葉を集めたコーパスである BAWE（The British Academic Written English）を用いて、Biber（1988）で特定されたコミュニケーション上の機能と関連している「次元」（Biber はディメンションと呼んでいる）と、4 段階の学生レベル、13 種のジャンル、4 種の学問分野との関連性を調査している。

3．各手法で気をつけるべき点

前節では、多変量解析の各手法の目的と違い、言語（教育）研究などにおける先行研究を簡単に紹介した。全体的には、研究者の手元にあるデータから何かしらの予測をしたいのか、あるいは分かりやすく要約したいのか、そして、その場合の変数の取り扱いの理解こそが、多変量解析を実施する上で必要となる前提知識である。使用する手法が決まれば、さらにその手法の実行に関していくつかの選択が迫られるが、それは既に紹介した統計の良書を参考されたい。

本節では、著者が現在まで自身の個人・共同研究で実施してきた多変量解析の中でも、特に重回帰分析、対応分析、主成分分析に注目し、統計書などには実際に触れられていなかったり、一般的に軽視されているような留意事項に関して簡単にまとめる。以下、少し専門的な内容にはなるが、これから多変量解析を使う予定の文系研究者、あるいは既に使っている文系研究者にも参考になれば幸いである。

3．1．重回帰分析における留意すべき点

多変量解析の手法によっては、分析上、留意すべき項目がある。例えば、重回帰分析においては、予測値と観測値の差である残差がある程度正規分布に従っているか、あるいは正規分布に近いものになっているか、均一に残差が分散されているか、残差間で相関がないか、などをチェックする必要がある。これは、他に比べて極端に逸脱したデータが混ざっていた場合、あるいは残差の分布が正規分布からかけ離れていた場合などでは、算出された回帰モデルに何かしらの問題が生じていると解釈されるのが一般的である。このような場合には、対数変換して正規分布に近い分布にしておくなどの対応を講じることで、より意味のある回帰式を得ることが可能となる。他の多変量解析の手法に関する諸条件に関しては、本稿で紹介したような概論書を参考にして頂きたい。

また、重回帰分析や判別分析などでは変数の選択法がいくつか存在する。例えば、変数を強制的に指定する方法や、全ての変数の組み合わせを計算するような総当たり法、一定のルールに従って変数を出したり入れたりして選択していくステップワイズ法がある。重回帰分析で最もポピュラーなのは、ステップワイズ法である。ステップワイズ法では、変数の選択規則によって、変数増加法、変数減少法、変数増減法、変数減増法の 4 種に分かれる。どの方法が最善であるかは一概には言えないが、JUSE StatWorks（表 4 参照）で選択可能な「逐次選択 4 方法」では、上記四つの方法から「寄与率が最も高い変数の組み合わせを自動的に求める方法」（https://www.i-juse.co.jp/statistics/support/faq/70015.html）であり、悩むようであれば、このような自動化されたアルゴリ

ズムを使うのも一つの手であろう。以上のような変数選択の諸手法に関しては奥野他（1971）が詳しい。

なお、重回帰分析においては、目的変数間で強い相関が認められる場合に生じる多重共線性についても気をつける必要がある（多重共線性に関しては統計概論書を参考にされたい）。SPSS などを用いた場合、一般的には VIF（Variance Inflation Factor）という値が 10 を超えているかどうかをチェックすることで、多重共線性の発生を確認することができる。多重共線性が発生している場合は、一般的に相関が高い説明変数の片方を削除することで回避することができる。しかしながら、筆者も従事した吉村他（2019）の分析中に遭遇したケースとして、VIF の値は問題ないにも関わらず（つまり多重共線性が認められなくても）、回帰係数が負を示している場合がある。以下、吉村他（2019）の共同研究者の方にご提供頂いた Dewaele & Dewaele（2018）の分析データを参照されたい。このような例外に遭遇した場合、ステップワイズ法を用いて説明変数を減らすことで対応可能である[5]。このあたりは統計の入門テキストで触れられていないこともあるので、統計初心者は留意する必要があろう。

Table 1: Pearson correlation analyses between independent variables and WTC.

Independent variable	Pearson *r*	*p*
Private FLE	.46	.0001
Attitude FL	.45	.0001
Language Level	.45	.0001
FLCA	−.41	.0001
Teacher's FL use	.33	.0001
Social FLE	.30	.0001
Attitude teacher	.29	.0001
Relative standing	.25	.0001
Test results	.24	.001
Age	.24	.001

Table 3: Multiple regression analysis with WTC a dependent variable (sorted according to Beta value).

	B	*SE*	*Beta*	*t*	*P*
FLCA	−.21	.05	−.26	−4.0	.0001
Language level	.17	.05	.23	3.4	.001
Teacher's FL use	.12	.04	.18	2.9	.005
Attitude FL	.11	.05	.18	2.4	.019
Social FLE	.14	.07	.14	1.9	.060
Age	.06	.03	.13	2.4	.018
Private FLE	.07	.08	.08	.86	.390
Attitude teacher	−.01	.04	−.01	−.17	.862
Relative standing	.00	.06	−.01	−.08	.890
Test results	−.01	.01	.00	.00	.990

図 1

Dewaele & Dewaele（2018, p.30）より引用（Pearson *r* = ピアソンの積率相関係数（２変数間の相関を示す指標で - 1 から 1 の値をとる）; *p* = 有意確率（解析結果を解釈する際の目安）; *B* = 標準化していない偏回帰係数; *SE* = 偏回帰係数の標準誤差 ; *Beta* = 標準偏回帰係数 ; *t* = *t* 値）
* 詳しくは当該論文や統計概論書を参照のこと

３．２．対応分析の解釈に関して

対応分析によって生成される行変数と列変数の対称マップの解釈は注意しなければいけない点がある。まず、吉村他（2019）では、対応分析を用いることにより学生が執筆したジャーナルから協同学習を成功させる上で必要とされる８原理がどのように反映されているかを可視化した。しかしながら、水本（2009）も指摘するように、対応分析の注意すべき点は「外れ値に影響される」（p.63）ことにある。実際、吉村他（2019）における対応分析の解析結果においても、外れ値が認められたが差異を表すことができた点はよかった。しかしながら、外れ値に引っ張られて解釈が難しかった

のも事実である。

また、藤本（2017）も指摘するように、「行変数内、または列変数内のカテゴリーポイント間の距離は、数理的に定義されているが、行変数と列変数の間の距離は定義されていない」（p.141）ことから、対称マップに同時布置された「相異なる集合に属する二つの点の親近性を解釈することは非常に危険である」（大隈他 , 1994, p.75）と言えよう。次節で挙げるような統計処理を簡易に実行できるソフトに添付された説明書等の解説だと、このことが「省かれてしまい、「わかりやすいグラフ」が一人歩きすることになる」（p.148）と警鐘を鳴らしている。

対称プロットの解釈において唯一可能なことは、藤本（2017）が Greenacre（2007）の考えをまとめている中で、「一方のベクトルと他方のベクトルが同じ方向を向いている場合には、Symmetric Map での表示での問題にはならず、そうでない場合は、問題になるという」（p.149）と指摘するように、同時布置の場合はベクトル方向のみを勘案事項として、生成されたマップを解釈するのが無難なようである。残念ながら、この点を軽視して対応分析の結果を解釈している言語（教育）研究も少なからずある。

3．3．主成分分析の留意点に関して

主成分分析は、できる限り収集したデータの情報を圧縮した形で項目間（変数間とケース間）の関係性を可視化する目的で使うことができる（金 , 2007）。主成分分析と対応分析は類似している点が多々あるが、根本的な違いとして、量的データを扱う場合は主成分分析、質的データを扱う場合は対応分析という点が挙げられる。「コーパスにおける語やコロケーションの頻度のようなデータは、尺度が等間隔になっているような量的データではなく、質的データ」（水本 , 2009, p.54）と解釈されることから、一般的には対応分析を用いる方が妥当である。しかしながら、出村他（2004）、Tabachnick & Fidell（2006）を参考に水本（2009）でまとめられているように、データの頻度が適当なレンジにまたがっている場合は相関係数の計算が可能であることから、「記述目的として主成分分析を適応することも可能」である（水本 , 2009, p.54）。実際には、Burrows（1989）をはじめ、田畑（2004b）、Nishina（2007）、水本（2009）、水本・野口（2009）などのコーパス分析においても主成分分析は用いられている。

また、主成分分析では、第一主成分に最も説明力の高い成分を抽出し主成分得点が作られる。つまり、元データの総合的な情報を反映した第一主成分、および第二主成分でプロット化した場合、対応分析と比べて項目間の差異が分かりにくいことが多い。これを避けるために、あえて第二主成分と第三主成分を用いて二次元プロットを生成する方法もあり、こちらが対応分析の結果と酷似しているという報告がある（君山 , 2002; 水本 , 2009; 水本・野口 , 2009）。いずれにしても、第 2.3 節で述べたように、固有値 1.0 以上の主成分が複数ある場合には、第二主成分 + 第三主成分などの組み合わせであっても分析上の価値があると言えよう。

なお、主成分分析では、転置行列を用いる（元の行列データのケースと変数を転置して行う）手法がある（中村純作 , personal communication, July 15, 2007）。水本（2009）でもその特徴がまとめられているように、Burrows（1989）以降、コーパス言語学などで使用されている手法である。対応分析では行列を入れ替えても結果は同じであるが、主成分分析の場合は異なる結果が得られるので注意が必要である。水本の研究なども参考に、どの成分を用いて二次元プロットを生成するの

か、転置行列を用いて分析するかなど、必要に応じて探索的に分析するのが望ましいのではないだろうか。

いずれにしても、主成分分析と対応分析のこのような違いは詳細に把握しておく必要があろう。

４．文系研究者にも優しい統計ソフト

最後に、本節では主に文系研究者にも優しい統計ソフトを紹介したい。統計ソフトは大きく分けて、まず、GUIベースとCUIベースに分かれる。一般的に文系研究者にとってコマンド入力してやり取りを行うCUIはハードルが高く、パソコンの指示に従って数回クリックするだけで実行できるGUIの方がはるかにユーザー・フレンドリーである。また、有料か無料かという違いも重要である。表４では、これらの違いやOS、搭載している多変量解析手法などを統計ソフトごとに簡単にまとめている。なお､括弧中のMとWは､Macと Windowsを表す（Linuxに関しては､ユーザーが限られることが予想されるのでここでは省略する)。

表４．各統計ソフトの情報

	有料・無料	無料	搭載している多変量解析の一例
GUI	無料	HAD（W・M）	重回帰分析、判別分析、因子分析、クラスター分析､数量化分析､構造方程式モデリング（SEM）など
	無料	PSPP（W・M）	ロジスティック回帰分析、クラスター分析、因子分析など
	無料	JAMOVI（M）	主成分分析、因子分析など
	無料	JASP（M）	主成分分析、因子分析、構造方程式モデリング（SEM）など
	有料	IBM SPSS Statistics Base(W･M)	主成分分析、クラスター分析など
	有料	JMP（W・M）	主成分分析、線形判別分析、PLS回帰、階層型クラスター分析、K Meansクラスター分析、潜在クラス分析など
	有料	Stata（W・M）	因子分析、主成分分析、判別分析、多変量線形回帰､構造方程式モデリング（SEM)､クラスター分析、多次元尺度構成法（MDS)、対応分析など
	有料	XLSTAT BASIC+（W・M）	主成分分析、対応分析、因子分析、判別分析、多次元尺度法（MDS）など
	有料	S-Plus（W）	判別分析、因子分析、主成分分析、多次元尺度法（MDS）など
	有料	エクセル統計（W）	重回帰分析､ロジスティック回帰分析､判別分析、主成分分析、因子分析、クラスター分析、多次元尺度法（MDS)、数量化分析、対応分析など
	有料	EXCEL多変量解析（W・M）	重回帰分析､ロジスティック回帰分析､判別分析、主成分分析、因子分析、クラスター分析、数量化IV類など[6]

	有料・無料	無料	搭載している多変量解析の一例
GUI	有料	JUSE-StatWorks（W）	主成分分析、数量化 III 類、因子分析、判別分析、数量化 II 類、クラスター分析など
	有料	Minitab（W・M）	主成分分析、因子分析、判別分析、クラスター分析、コレスポンデンス分析など
CUI	無料	R（W・M）	主成分分析、因子分析、対応分析、多次元尺度法（MDS）、判別分析など[7]
		PSPP（W・M） * コマンドでも実行可能	ロジスティック回帰、クラスター分析、因子分析など
	有料	S-Plus（W） * S -Plus は S 言語を使用	重回帰分析、主成分分析、因子分析、多次元尺度法（MDS）、判別分析、クラスター分析など
		Stata（W・M） * コマンドでも実行可能	因子分析、主成分分析、判別分析、多変量線形回帰、構造方程式モデリング（SEM）、クラスター分析、多次元尺度構成法（MDS）、対応分析など

*M = Mac、W = Windows * GUI = Graphical User Interface、CUI = Character User Interface

昨今、何十万円もする高額な有料の統計ソフトの機能や使い勝手を踏襲した無料の統計ソフトが公開されている。個人的に無料ソフトの中でも使いやすいインターフェースを搭載していると感じるのは、JASP と JAMOVI である。また、Excel でまとめたデータからそのまま統計解析が実行できる XLSTAT、エクセル統計、EXCEL 統計（エクセル統計と EXCEL 統計は別のソフト）はいずれも初心者には使い勝手がよいが有料である。本格的に統計処理の学習を始めるのであれば、やはり R か高額ではあるが IBM の SPSS が無難であろう。コーパス言語学では R、英語教育では SPSS がスタンダードであったが、近年、後者でも R を使った分析が多くなっている。

5. まとめ：統計処理の現在とこれから

以上、本稿では、文系研究者が知っておくべき多変量解析の初歩と注意すべき点、現在利用可能である統計ソフトを簡潔にまとめた。このような頻度主義統計は文系の諸分野を専門とする研究者にとっては、未だ取っ付き難いかもしれない。ただし、自身の研究の妥当性や有効性を検証し、そのエビデンスを客観的に提示するためには必要不可欠な研究の術でもある。まずは表 4 で示した無料統計ソフトに触れてみることから始めてみてもよいのではないだろうか。本稿で紹介した文献以外にも、優れた統計書やウェブサイト、論文はオンライン上でも多数見つかるので是非参考にして欲しい。

なお、草薙（2017）が指摘するように、2010 年代からの統計改革によって効果量や検定力の提示が言語教育研究においても推奨されている[8]。水本・竹内（2010）の検定・分析の種類別の代表的な効果量の指標と大きさの目安（p.51）や、各手法における具体的なサンプルサイズの例（pp.59-70）は大変参考になる。コーパス言語学においても効果量の重要性は指摘されており（Gries, 2010）、石川・長谷部・住吉（2020）でも紹介されているようにコンコーダンサー（コーパス分析ソフト）によっては、「実質的な頻度差の目安となる効果量」（p.22）の表示が可能となっている。しかしながら、効果量などの指標はあくまで説明できる一つのごくわずかな指標にすぎず、その値にばかり依存しすぎないことが重要である。なぜならば、草薙（2014）は *t* 検定を例に説明しているが、検

定統計量 *t* は効果の大きさと標本の大きさを掛け合わせたものであるため、統計的に有意であったとしても、効果が小さい場合もあるからである（南風原 , 2002; 水本・竹内 , 2008）。

また、従来の頻度主義統計よりもベイズ統計を実施することで「統計的帰無仮説検定の一部の問題とは無縁」になり、「自然、柔軟、そして人間らしい意思決定ができる」（草薙 , 2017, p.2）といった利点も指摘されている。今後、その動向にますます注目が集まるであろう。

謝辞

匿名の２名の査読者の方々から貴重なご助言をいただいた。ここに記して謝意を述べたい。

注

1　決定木（decision tree）は分類木（classification tree）とも呼ばれ、結果変数が量的な値をとる場合には回帰木（regression tree）とも呼ばれる（藤井 , 2010, p.142）。

2　Gini 係数とは不純度や不平等さを測る指標であり、雑多な情報を分類する上で、どれほどの不純物が混ざっているかを０～１の数値を用いて表す（完全に純粋な場合は０）。「社会における経済格差を表す指標」とも呼ばれる（村山・若宮・荒牧 , 2018, p.698）。決定木の場合、子ノードの Gini 係数の平均は親ノードの値よりもが小さくなるように分岐していく。

3　アンサンブル学習とは、複数の学習器を組み合わせることで「複雑な処理を行うことができる機械を構成する」手法である（金森・竹之内・村田 , 2009, p.154）。

4　主成分分析や因子分析では相関行列の固有値分解を、対応分析や数量化 III 類では頻度行列の特異値分解を、判別分析では分散行列の固有値分解を、多次元尺度法では距離行列の固有値分解を実施しており、入力するデータが異なる。

5　明治大学の廣森友人教授には特に吉村他（2019）分析時に、本稿で紹介した資料の提供を含め本件に関する貴重なご意見を頂き大変勉強になった（私信：2019 年３月 11 日）。ここに感謝の意を示す。

6　数量化 IV 類は、順序尺度からなるデータを用いた多次元尺度法と言える（エクセル統計数量化 IV 類の概要より https://bellcurve.jp/ex/function/quant4.html）。

7　R の基本パッケージの中にある多変量解析に関しては、RjpWiki 中の「R の基本パッケージ中の多変量解析関数一覧」を参照されたい（http://www.okadajp.org/RWiki/? R の基本パッケージ中の多変量解析関数一覧）。

8　「十分な検出力をもった研究を行うために行われるサンプルサイズの計算」（岡田 , 2007, p.1）のことをパワーアナリシスと呼ぶ。R の base パッケージには *t* 検定、カイ２乗検定、分散分析の三つの検定手法に対応したパワーアナリシスが組み込まれている。GUI ソフトの G *Power（http://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html）を用いても、サンプルサイズを求めることができる。

参考文献

［１］ 朝野熙彦（2000）『入門多変量解析の実際（第２版）』東京：講談社サイエンティフィック.

［２］ Biber, D.（1988）. *Variation across speech and writing*. Cambridge, UK: Cambridge University

Press.

［３］ Burrows, J. F.（1989）.‘A Vision’as a revision? *Eighteenth-Century Studies, 22*, 551-565.

［４］ 出村慎一・西嶋尚彦・長澤吉則・佐藤進（編）.（2004）.『健康・スポーツ科学のための SPSS による多変量解析入門』東京：杏林書院 .

［５］ Dethorne, L. S., Johnson, B. W., & Loeb, J. W.（2005）. A closer look at MLU: What does it really measure? *Clinical Linguistics and Phonetics, 19*（８）, 635-648.

［６］ Dewaele, J. M., & Dewaele, L.（2018）. Leaner-internal and learner-external predictors of willingness to communicate in the FL classroom. *Journal of the European Second Language Association, 2*（１）, 24-37.

［７］ 藤井良宜（2010）『Ｒで学ぶデータサイエンス１ カテゴリカルデータ解析』東京：共立出版 .

［８］ 藤本一男（2017）「対応分析のグラフを適切に解釈する条件」『津田塾大学紀要』*49*, 141-153.

［９］ 古橋聰・高橋薫（1993）「タグ付きコーパスの句レベルの解析について」『中京大学教養論叢』*33*（２）, 117-146.

［10］ 古橋聰・高橋薫（1995）「LOB-Corpus におけるカテゴリーの特徴について－多変量統計解析法による分析－」『中京大学教養論叢』*35*（３）, 727-747.

［11］ Greenacre. M.（2007）. *Correspondence analysis in practice*（*second edition*）. London, UK: Chapman & Hall/CRC Interdisciplinary Statistics.

［12］ Gries, S. T.（2010）. Useful statistics for corpus linguistics. In A. Sánchez & M. Almela（Eds.）, A mosaic of corpus linguistics: Selected approaches, 269-291. Frankfurt am Main: Peter Lang.

［13］ 南風原朝和（2002）『心理統計学の基礎―統合的理解のために』東京：有斐閣アルマ .

［14］ 服部伸一・野々上敬子・門田新一郎（2009）「小学生の自覚症状の訴え数とライフスタイル要因との関連について－数量化 II 類を用いた検討－」『小児保健研究』*68*（６）, 643-653.

［15］ 今道晴彦（2013）「ドイツ語属格の語尾選択規則とその有用性：決定木による分析の試み」『統計数理研究所共同研究レポート』*290*, 73-87.

［16］ 石川慎一郎（2008）「主成分分析を用いた英文エッセイ自動診断システムの構築の可能性」『学習者コーパスの解析に基づく客観的作文評価指標の検討 統計数理研究所共同研究リポート』*215*, 29-42.

［17］ 石川慎一郎（2009）「因子分析（３変数１因子モデル）を用いた FROWN コーパスにおける頻度副詞の共通性と独自性の検討」『コーパス言語研究における量的データ処理のための統計手法の概観 統計数理研究所共同研究リポート』*232*, 119-127.

［18］ 石川慎一郎・前田忠彦・山崎誠（2010）『言語研究のための統計入門』東京：くろしお出版 .

［19］ 石川慎一郎・長谷部陽一郎・住吉誠（2020）『コーパス研究の展望』東京：開拓社 .

［20］ 金森敬文・竹之内高志・村田昇（2009）『Ｒで学ぶデータサイエンス５ パターン認識』東京：共立出版 .

［21］ 金明哲（2007）『R によるデータサイエンス』東京：森北出版株式会社 .

［22］ 君山由良（2002）『コレスポンデンス分析と因子分析によるイメージの測定法』東京：データ分析研究所 .

［23］ 草薙邦広（2014）「教育実践のなかで集団に対する処遇の結果を適切に解釈するための定量的方法－効果量の利用とその限界点－」『外国語教育メディア学会（LET）関西支部メソドロジー研究部会報告論集』*6*, 46-84.

［24］ 草薙邦広（2017）「外国語教育研究者のためのベイズ統計入門」第 57 回外国語教育メディア学会全国研究大会ワークショップ資料（pp.1-15）. 名古屋学院大学 . Retrieved from https://drive.google.com/file/ d /0BwTqO-zjLUJnZjJzTnZfSS03VlU/view

［25］ 見目卓之（2004）「Why do you 'make a decision' instead of 'decide'? －ロジスティック回帰分析に基づく英語軽動詞に関するケーススタディ－」英語コーパス学会第 23 回大会発表資料 .

［26］ 小林雄一郎（2008）「高頻出語を用いた学習者コーパスの分類」『学習者コーパスの解析に基づく客観的作文評価指標の検討 統計数理研究所共同研究リポート』*215*, 69-82.

［27］ 小林雄一郎・金丸敏幸（2012）「パターン認識を用いた課題英作文の自動評価の試み」『電子情報通信学会技術研究報告』, *112*（103）, 37-42.

［28］ Larson-Hall, J.（2015）. *A guide to doing statistics in second language research using SPSS and R*（*The 2nd ed*）. London: Routledge.

［29］ 水本篤（2008）「自由英作文における語彙の統計指標と評定者の総合的評価の関係」『学習者コーパスの解析に基づく客観的作文評価指標の検討』統計数理研究所共同研究リポート 215, pp.15-28.

［30］ 水本篤（2009）「コーパス言語学研究における多変量解析手法の比較：主成分分析 vs. コレスポンデンス分析」『統計数理研究所共同研究リポート』*232*, 53-64.

［31］ 水本篤・野口ジュディー（2009）「多変量解析を用いた PERC コーパスの領域分類」『統計数理研究所共同レポート』*232*, 85-106.

［32］ 水本篤・竹内理（2008）「研究論文における効果量の報告のために―基礎的概念と注意点 ―」『関西英語教育学会紀要 英語教育研究』*31*, 57-66.

［33］ 水本篤・竹内理（2010）「効果量と検定力分析入門－統計的検定を正しく使うために－」『より良い外国語教育研究のための方法 外国語教育メディア学会（LET）関西支部メソドロジー研究部会 2010 年度報告論集』47-73.

［34］ 村山太一・若宮翔子・荒牧英治（2008）「WORD GINI: 語の使用の偏りを捉える指標の提案とその応用」『言語処理学会第 24 回年次大会発表論文集』698-701.

［35］ 中村永友（2009）『R で学ぶデータサイエンス 2 多次元データ解析法』東京：共立出版 .

［36］ Nakamura, J., & Sinclair, J.（1995）. The world of woman in the Bank of English: Internal criteria for the classification of corpora. *Literacy and Linguistic Computing, 10*, 99-110.

［37］ Nesi, H.（2009）. A multidimensional analysis of student writing across levels and disciplines. In M. Edwardes（Ed.）, Taking the measure of applied linguistics: Proceedings of the BAAL Annual Conference, University of Swansea, 11-13. September 2008. London: BAAL/ Scitsiugnil Press.

［38］ Nishina, Y.（2007）. A corpus-driven approach to genre analysis: The reinvestigation of academic newspaper and literary texts. *ELR Journal, 1*（２）. [online journal]

［39］ 能登原祥之（2015）「N -gram 分析を通した CEFR レベル基準特性の特定：動詞共起フレー

ムに焦点をあてて」『言語処理学会第 21 回年次大会発表論文集』876-879.

［40］ Oakes, M. P.（1998）. *Statistics for corpus linguistics*. Edinburgh, UK: Edinburgh University Press.

［41］ 岡田昌史（2007）「R Commander を用いた統計解析の基礎（３）」Retrieved from http://www.okadajp.org/RWiki/?plugin=attach&refer=100 万ヒット記念コーナー &openfile=RcmdrPractice3.pdf

［42］ 奥野忠一・久米均・芳賀敏郎・吉澤正（1971）『多変量解析法』東京：日科技連 .

［43］ 大隅昇・アランモリノウ・馬場康維・ルドヴィックルバール・ケネスワーウィック（1994）『記述的多変量解析法』東京：日科技連出版社 .

［44］ Paavola, L., Kunnari, S., & Moilanen, I.（2005）. Maternal responsiveness and infant intentional communication: Implications for the early communicative and linguistic development. *Child: Care, Health & Development, 31*（６）, 727-735.

［45］ 李在鎬（2016）「日本語教育のための文章難易度に関する研究」『早稲田日本語教育学』*21*, 1-16.

［46］ Tabachnick, B. G., & Fidell, L. S.（2006）. *Using multivariate statistics*（*5th international ed.*）. Boston, MA: Pearson/Allyn & Bacon.

［47］ 田畑智司（2004a）『コーパス言語学のための多変量解析入門』英語コーパス学会第 24 回 大会ワークショップ配布資料 . Retrieved February 20, 2019, from http://www.lang.osaka-u.ac.jp/~tabata/JAECS2004/JAECS2004hand.pdf

［48］ 田畑智司（2004b）「-ly 副詞の生起頻度解析による文体識別―コレスポンデンス分析と主成分分析による比較研究―」『電子化言語資料分析研究』97-114.

［49］ 高橋薫（2019）「LOB Corpus から BNC へと移行した語彙分析の成果について」『コーパスと英語研究』（pp. 213-224）. 東京：ひつじ書房 .

［50］ 竹内理・水本篤（2012）『外国語教育ハンドブック』東京：松柏社 .

［51］ 富田勤・須田美由紀・五十嵐直子・佐々木胤則（2005）「高校生におけるライフスタイルと疲労自覚症状との関連－主に数量化Ⅰ類を用いた分析から－」『北海道教育大学紀要自然科学編』*65*（１）, 29-37.

［52］ 吉村征洋・廣森友人・桐村亮・仁科恭徳（2019）「英語ドラマ制作によるプロジェクト型協同学習が学習者の心理的側面に与える影響」『JACET Kansai Journal』*21*, 23-44.

［53］ Zareva, A.（2005）. Models of lexical knowledge assessment of second language learners of English at higher levels of language proficiency. *System, 33*, 547-562.