なぜ，「コレスポンデンス分析・Dual Scaling・数量化理論 再考」か

多摩大学　今泉　忠

1. はじめに

“big data”という用語が用いられるようになり，2015年の現在では「ビッグデータ」に関する研究が様々な分野で進められており，統計学に関連する諸分野でも「ビッグデータ」と言われる多種多様なデータを扱うことが多くなったきた．その場合に，扱うデータの変数の型，質的変数や量的変数などによって手法が選択されて分析を行なわれる場合がある．広義に言えば定量分析と定質分析とでも言えよう．従来のデータ分析を行うために手法についても同様に分類して扱う変数の型から整理すると，主として量的変数を扱う重回帰分析や主成分分析やクラスター分析や判別分析などの多変量解析（分析）の手法が挙げられる．一方，主として質的変数を扱う数量化の手法として，コレスポンデンス分析・Dual Scaling・数量化理論や一般化線形モデルの手法が挙げることができる．これらをまとめて多次元データ分析とも呼ぶ場合もある．

2. 多次元データ分析

多変量解析では，相関係数などの関連度をもとにして分析する．これらの手法で質的変数は扱うのは難しい.「ビッグデータ」でのテキストデータの分析では語句の共起度などをもとにして，コレスポンデンス分析（対応分析）などを適用する場合がある．しかし，この場合のコレスポンデンス分析は，独立性からの偏差をもとに χ^2 距離モデルを用いて分析する．この意味においては，関連性と独立性の対応について何らかの検討が必要と考えられる．一方，Dual Scalingや数量化理論では，より多様なデータの形式を扱うために相関比 η^2 などの量を用いて場合もある．これにより，数量化I類や数量化II類と呼ばれている手法として提案され，コレスポンデンス分析は数量化III類と呼ばれている手法と対応する．

3. 「データから考える」ための再考の必要性

質的変数を分析する場合のデータにおいて，回答者は，ある項目のカテゴリーの 1 つに必ず反応しなければないないのか，しなくともよいのかなども重要な課題である．さらに，従来は事前に設定された2~10個位の変数からなるデータであったが，最近では事後的に100~300個となったデータである．このような二種類のデータが表層的に異なるのか，または，その背後的な反応でも異なるのか検討をする必要があろう．

このように考えると，多次元データ分析の手法として扱う変数の型により手法を整理することは重要であるが，それ以前に「データ分析」を「データを通じて分析を行う」と考えれば，データを通して何を明らかにしたいのか，どのようなデータの形式として収集されたのか，を明確にすることが重要であることがわかる．この問は従前から繰り返された問ではあるが，ここで，コレスポンデンス分析・Dual Scaling・数量化理論に関する再考を行うことで「データ分析」としてより広い適用可能性について検討する．