匿名データ作成のための最適な境界値：経済統計データに基づく実証分析

一橋大学経済研究所／独立行政法人統計センター　白川　清美
独立行政法人統計センター　阿部　穂日
一橋大学経済研究所　千葉　亮太

省庁が公的統計調査の匿名データを作成する際には、開示リスクを軽減するため、様々な統計的開示抑制を講じており、その一つとしてトップ（ボトム）コーディングがよく用いられる。これは、各変数に対し上限値（下限値）を設定し、その境界値以上（以下）のデータを一つの階級としてまとめる手法である。これにより、極端に大きい（小さい）値のデータを秘匿することができる。

しかしその一方で、上限値（または下限値）による情報損失が問題となる。例えば実際のデータが複数の分布から成る場合、境界値を分布の境界に設定すると、秘匿された分布を公開されたデータから推測することが出来なくなる（図 1）。

また、省庁は、利用者がどのような分析を行うか仮定していないため、トップコーディングの前に対数変換などのデータ変換を行わない。一方、研究者は、特に経済統計のデータを分析する際、しばしばデータ変換を行う。

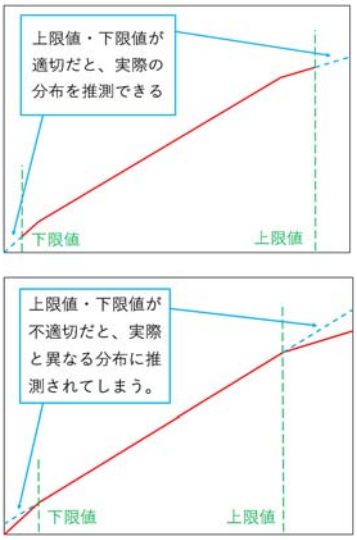

図 1　上限値（下限値）が適切または不適切な場合の推測結果の違い

2017 年度統計関連学会連合大会において、我々はトップコーディングの境界値検証のため、決定木と Chow 検定を組み合せ、適切な境界値を探索する手法について、擬似データによるシミュレーションの結果を報告したが、今回は総務省統計局が実施した平成 27 年科学技術研究調査の調査票情報にこの手法を適用し、トップコーディングにより秘匿された分布と、トップコーディング後のデータから推測される分布の違いを調べた。また、層別に適切なトップコーディングの境界値を推定し、元のデータと秘匿後のデータをデータ変換後に比較することで、トップコーディングが研究者の分析に与える影響について実証分析を行った。

詳細な結果については、報告当日に発表する。

［参考文献］

豊田秀樹．「データマイニング入門：R で学ぶ最新データ解析」，東京図書，2008.
青木繁伸．「Chow 検定」，http://aoki2.si.gunma-u.ac.jp/R/Chow.html，2018 年 6 月 23 日アクセス.
Shirakawa, Kiyomi. "Multi-Stratification for Outlier Detection based on the Graphical Model: Evaluation by Chow Test and AIC.", http://www.nstac.go.jp/services/society_paper/26_05_02.pdf, 2018 年 6 月 23 日アクセス.
阿部穂日、白川清美、千葉亮太．「匿名データの作成のためのリコーディング　―決定木による最適な境界値の分析―」，2016 年度統計関連学会連合大会，2016 年 9 月 6 日.
阿部穂日、白川清美、千葉亮太．「決定木と Chow 検定によるリコーディング境界値の分析」，2017 年度統計関連学会連合大会，2017 年 9 月 6 日.