

環境疫学研究における不検出データの問題と取り扱い

疫学研究において、環境因子への曝露を示す生体指標を定量する際、測定値がその分析における検出限界（LOD: limit of detection, detection limit）を下回ることがある。LOD は、測定値がブランク（空試験値）信号と有意に異なるとみなすことが出来る閾値のことを指すが、この場合当該データは不検出（non-detect）とされ、形式的にはその測定値が LOD 未満であるという情報のみ得られることとなる。検出限界は化学物質の分析に伴う値のばらつきに起因するため、微小な量や濃度が含まれる環境科学データには本質的に付随する問題といえ、不検出データへの対処は解析上の課題の一つとなる。例えば、2011 年より環境省が実施している子どもの健康と環境に関する全国調査（エコチル調査）においても、LOD 未満を含むデータに対する標準的な解析手法の確立が求められている。こうした疫学調査において、環境因子曝露の程度を表す変数は、主にそれを規定する要因（食事やその他生活習慣、人口動態特性など）を探索・検証するための結果変数として扱われる場合と、ある健康関連アウトカムリスク要因とみなして説明変数とされる場合がある。特に後者における解析では単純な値の補完あるいは変数のカテゴリ化（検出 vs. 不検出 など）を用いる事例が多いが、解析の目的に応じてより advanced な補完法や既存の生存時間解析あるいは打ち切りデータの解析手法を応用した方法も近年提案され始めている。一方で、それら手法を実際に適用した研究例は現時点ではほとんど見当たらない。課題研究では、各手法の比較検討ならびに改良を実データに基づいて行うことを検討している。

本抄読会では、環境疫学研究における検出限界の問題について、その背景の概説、当分野の研究で頻繁に引用される米国国民健康栄養調査（NHANES: National Health and Nutrition Examination Survey）のデータ例の紹介、提案されている統計手法のレビューを行う。

主要文献

- Helsel DR. *Statistics for censored environmental data using Minitab and R, 2nd ed.* New Jersey: John Wiley & Sons; 2011.
- Arunajadai SG, Rauh VA. Handling covariates subject to limits of detection in regression. *Environ Ecol Stat.* 2012; **19**: 369-91.
- Dinse GE, Jusko TA, Ho LA, et al. Accommodating measurements below a limit of detection: a novel application of Cox regression. *Am J Epidemiol.* 2014; **179**: 1018-24.