An Occupational Data Pipeline that Preserves Uncertainty: From Free-Text to Statistical Inference

Occupational data move through a pipeline: respondents describe their jobs in free text, those descriptions are coded into occupational classifications, classifications are mapped to scale values, and the values enter downstream analyses. Each step introduces uncertainty, and standard practice tends to discard it: a single best-fit code is recorded, a single scale value is assigned, and downstream models treat the result as if it were measured without error. This talk argues that uncertainty should be measured and propagated through the analytical pipeline rather than ignored.

In this talk I focus on two specific steps: automated occupational coding and the subsequent statistical inference. Using HACHICO, a Japanese occupational classifier based on Sentence-LUKE, I treat the coder's output as a full probability distribution over 204 occupational codes rather than a single prediction. These distributions are temperature-calibrated against human-coded reference data, and the uncertainty is carried into downstream analysis via plausible values combined with Rubin's rules. The result is bias- corrected estimates of intergenerational mobility, stratification effects, and other quantities of interest, with proper standard errors.

Applications to Japanese panel and cross-sectional surveys (JLPS, SSP, SSM) illustrate the framework. The same logic applies at the other steps (data collection design and scale construction), though those are beyond the scope of this talk. The approach is general and applies to any classifier that outputs probability distributions, supporting measurement-error correction in survey-based research using occupational data.