

Valid inference from new data sources: what can we learn from the experience of handling complex sampling designs?

Maria Giovanna Ranalli

University of Perugia, Italy

Statistics is definitely based on data. From a classical statistical perspective, data are often viewed as a faithful snapshot of a target population. In practice, however, reality is much more complicated and this picture is frequently distorted as when using filters by the processes through which data are generated, selected, and observed. This is particularly true in recent years in which statistical agencies and researchers increasingly rely on self-selected surveys or on administrative records, sensor data, social media, and other non-survey sources, often collected for non-statistical purposes. These data are typically non-IID, subject to selection mechanisms that are only partially observed or poorly understood, and may correspond to populations that differ from the inferential target. Valid inference from these new and emerging data sources critically depends on explicitly accounting for these processes. The talk draws lessons from decades of experience with inference under complex sampling designs, where the sampling mechanism is an integral part of the inferential framework. In such settings, ignoring the sampling design can lead to biased likelihoods and invalid inference, while incorporating it—through design-based reasoning, weighting, or joint models for outcomes and inclusion indicators—restores validity. By framing inference in terms of a superpopulation model coupled with an explicit selection process, it is possible to highlight how departures between population and observed-data distributions can be both systematic and informative. Revisiting these ideas provides a principled foundation for modern multi-source data integration and more reliable statistical inference.

DIPARTIMENTO
DI SCIENZE STATISTICHE



SAPIENZA
UNIVERSITÀ DI ROMA

