Multiwave sampling for two-phase designs

Thomas Lumley

University of Auckland
E-mail: t.lumley@auckland.ac.nz

Abstract: We consider the problem of estimation in two-phase samples, such as analysing EHR data using a validation subsample. Because of concerns about the impact of even nearly-undetectable model misspecification it is of interest to consider weighted estimation using raked weights (rather than a semiparametric-efficient estimator relying on the assumed model). The optimal sampling design depends on the unknown parameters in the model we want to estimate, and also on on the unknown relationship between the phase 1 and phase 2 data. Following McIsaac and co-workers, we consider multi-wave sampling of the validation sample, where a pilot sample is taken to estimate the parameters and allow the design to be optimised. We consider more complex phase-1 information than previous literature, and also the use of prior distributions.