

# International Journal of Population Data Science

Journal Website: [www.ijpds.org](http://www.ijpds.org)



## Pilot Samples for Estimation of Linkage Error

Gavin Thomson<sup>1</sup> and Matt Wray<sup>1</sup>

<sup>1</sup>Office for National Statistics, Newport, United Kingdom

Efficient estimation of linkage error in large linked datasets requires careful allocation of limited clerical review resources. We investigate the use of small stratified “pilot” samples as an initial, low-cost strategy for informing sample size calculation and allocation across strata when estimating binomial error proportions. Stratified candidate links and unlinked pairs exhibit heterogeneous error rates and variances; however, without prior variance information, optimal allocation – such as Neyman allocation – cannot be applied. We show that pilot samples of approximately  $n_h = 30$  per stratum provide sufficiently informative variance estimates at minimal cost, enabling efficient allocation of subsequent sampling effort. Using binomial absolute margin-of-error behaviour as a rough guide we calculate uncertainty for error proportions up to a maximum of  $p = 0.5$ , and use simulations to show that beyond  $n_h \approx 30$  the marginal reduction in uncertainty grows negligible.

Simulations illustrate also that these small pilots successfully differentiate high- and low-variance strata, supporting targeted sampling in the final round. When pilots are combined with Neyman allocation, the resulting stratum-level confidence intervals were shown to become more homogeneous, reducing overall population-level variance relative to proportional allocation. Empirical comparisons therefore show that, for equal total clerical review effort, Neyman allocation informed by pilots consistently yields narrower population-level confidence intervals than proportional allocation (0.85% vs 0.93% MoE). This confirms that even very small pilot samples can materially improve efficiency by enabling variance-driven sample size calculation and allocation. The method offers a practical, scalable approach for organisations conducting clerical review of linkage errors in large datasets.

