**Post-stratification notes for three potential applications**

Post-stratification is the assignment of sample units to a stratum after sampling has taken place, often according to a variable that is not known at the time of sample selection. As a result, 1) the number of observations in the group (post-stratum) is a variable and 2) the proportion of observations in any given group might not be the same as the proportion of units in the same group for the population (imbalance due to randomization).

We currently have three scenarios where this topic has come up.

Proximity and CWB allocation, and Interspersion metrics:

When allocating for the fixed gear (combined HAL and POT gears) stratum, we are separating HAL and POT by only allowing each gear type to be neighbors with the same gear type; HAL trips can only be neighbors to other HAL trips. In this case, the data are not actually post-stratified since the proximity indices are computed for all trips in the stratum (across gears) and the neighbors are defined by gear type as well as the spatial and temporal extent of the neighborhood. Similarly, the interspersion metric is computing the index across all trips in the design according to the comparison of interest (e.g. EM to observed trips) within neighborhoods, where neighbors are defined by gear type and the spatial and temporal extent of the neighborhood.

In both cases, although the analysis is done “by gear type”, we are not post-stratifying we are using gear type to define neighbors.

Post-stratification of trips from simple random sample

In this case, either within a stratum or in the not-stratified case, sample units are assigned to post-strata (*h*) after the sample is taken usually based on another covariate. This method is also used to balance samples where by simple random chance, the sample is not proportional to the population. To estimate the overall (post-stratified) mean, the true weighting for each post-stratum (*Np/N*) is used to weight the post-stratum-specific means (, where is the mean for the post-stratum, *p = 1, 2, …P* indexes the post-strata, *N* is the number of sample units in the population, *Np* is the number of units in post-stratum *p,* and is the estimated variance for post-stratum *p*.

The estimated variance for the post-stratified mean consists of two terms, the first is the variance for a stratified estimate of the mean under proportional sampling while the second term is the added variance associated with the random sample size in each post-stratum (variance from the sample (*n*) not being distributed proportionally to the population, i.e. some post-strata have more sample relative to the population as a whole).

Note that this formulation of the estimated variance is from Thompson, 2012 and differs from the estimated variance presented in Cochran 1963 or Cochran 1977 where a different derivation of the variance is used.

These equations apply regardless of the parameter being estimated, hence would be appropriate for post-stratified estimates of observer-effect metrics. Not that in the case that there are post-strata with sampling strata, the post-strata-specific estimated mean and variance would be computed for each stratum and then combined to estimate a total for the population.

For reference, the estimator of the stratified mean is

Where is the overall mean from a stratified sample, *h = 1, 2, 3, … L*, indexes strata, other terms as previously defined. The estimated variance given by

[Craig note – in cases where I am using data from the past, there is no little n, so the variance formula reduces to this:

]

where is the variance of the mean for stratum *h* and other terms are as defined previously. Note that sampling must be independent for each stratum.

Redesign of sampling strata

In this situation, we have a stratified population with simple random selection of sample units taken independently in each stratum. However, when reassigning sampled trips under new stratification schemes, the sampling probability for trips within the new stratum vary (since the original strata had different sample probabilities).

We would start by defining post-strata within each stratum using the new stratification covariates. Hence, within each stratum, there would be a post-strata defining each of the proposed new strata and we’d compute the post-stratified estimates for the metric of interest. Finally, we would combine the post-stratified estimates across sampling strata to estimate the metric for the new stratification.

As above, to estimate the overall (post-stratified) mean, the true weighting for each post-stratum (*Nph/Nh*) is used to weight the post-stratum-specific means (, where is the mean for the post-stratum, *p = 1, 2, …Ph* indexes the post-strata, *Nh* is the number of sample units in the sampling stratum, *Nph* is the number of units in post-stratum *p* and stratum *h,* and is the estimated variance for post-stratum *p* and stratum *h*.

This is the equivalent of computing the post-stratified mean for each stratum, where post-strata are the new proposed strata definitions. The estimated variance for the post-stratified mean consists of two terms, the first is the variance for a stratified estimate of the mean under proportional sampling while the second term is the added variance associated with the random sample size in each post-stratum (variance from the n not being distributed proportionally to the population, i.e. some post-strata have more sample relative to the population as a whole).

Once we have the estimates of the metric for each post-stratum and stratum, we combine across strata to get the overall estimated mean for the new strata definitions.

The estimator of the stratified mean is

where N is the number of sample units in the population, Nh is the number of units in the stratum, and is the mean value of the metric within the stratum, in this case the post-stratified mean. The estimated variance given by

where, as previously, *h = 1, 2, …L* indexes the stratum and is the variance of the mean for stratum *h*.

Another alternative approach would be to use a Horwitz-Thompson estimator of the mean where each individual observation is weighted by its sampling probability. This is a more complex approach that would require knowing the selection probabilities (got that) and creating small-area estimates for the metrics under each new sampling stratification, and then, as before, combining across strata. At this time, I do not recommend this second approach.