

Validation Analysis for GRSC

World Changers
Shaped Here



SMU

Lynne Stokes and Shalima Zalsha

SSC Meeting, Mar 30 – Apr 2, 2021



Planned Sample Design and Analysis

- The planned sample design was a stratified random design
 - 3 Stratification variables
 - Region (FL, ALMS, LA, TX)
 - Habitat (UCB, Natural, Artificial reefs)
 - Depth (Shallow, Mid, Deep)
 - Not all strata were present in all regions, so # of strata varied by region.
- Estimation of total abundance was carried out with a standard stratified sampling estimator:

$$\hat{t} = \sum_{h=1}^H \hat{t}_h$$

- If sampling units are items (e.g., artificial reefs), \hat{t}_h is the mean-per-unit estimator:

$$\hat{t}_{hy,mpu} = N_h \bar{y}_h.$$

- If sampling units are transects of unequal size or two-stage (pyramids), \hat{t}_h is the standard ratio estimator:

$$\hat{t}_{hy,r} = t_{hx} \times \frac{\sum_{i=1}^{n_h} y_{hi}}{\sum_{i=1}^{n_h} x_{hi}} = t_{hx} \times \hat{d}_h,$$

where t_{hx} is the total area of the universe in stratum h and \hat{d}_h is its estimated density.



- The variance of the stratified estimator was estimated by the sum of the estimated variances of the total estimates in each stratum:

$$v(\hat{t}_{hy,mpu}) = N_h^2 \times (s_{hy}^2/n_h) \left(1 - \frac{n_h}{N_h}\right);$$

and

$$v(\hat{t}_{hy,r}) = t_{hx}^2 \times (s_{hd}^2/n_h) \left(1 - \frac{n_h}{\hat{N}_h}\right);$$

where s_{hy}^2 is the sample variance of y_{hi} (# of fish) and s_{hd}^2 is the sample variance of the residuals $d_{hi} = y_{hi} - \hat{d}x_{hi}$ in stratum h (known as Taylor Series variance estimate)

- Note that the fpc in the ratio estimator is a function of the *estimated* number of units in the entire population, $\hat{N}_h = t_{hx}/\bar{x}_h$.
- The estimates and standard errors were computed in SAS PROC SURVEYMEANS



- Sample sizes changed due to complexities of data collection (e.g., malfunctioning gear, technical glitches on video). Sampling was assumed MCAR; no adjustments were made for non-response, except reduced sample size.
- Sample design was adapted when new information was discovered about the habitat (e.g., pyramids in Texas)
 - Stratum was added; design adapted to a cluster design, with first stage = grids. Ratio estimator, where # of pyramids in grid was measure of size, was used.
- Poststrata were added beyond original plan. E.g., Florida was divided into three regions (NW, mid, south). These were treated as strata for estimation purposes.



- Data from LA were mostly unavailable. Data for like habitats from TX were substituted for missing LA data for estimating densities. Then ratio estimates were computed, expanding these densities to LA stratum areas.
- We did not make an estimate of variance (or standard error) for the entire GOM, since the data from TX were reused. This makes the additive formula for variance across strata incorrect.
- Instead, we made two estimates of SE: one for LA (which used substituted data from TX + some LA data), and another for the GOM excluding LA.



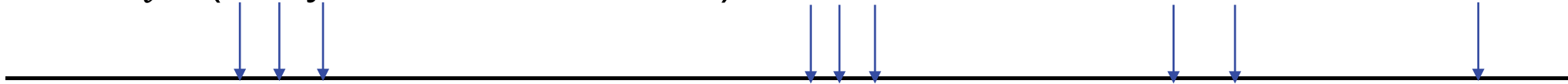
- Several reviewers commented that our variance estimate may be biased low, due to ignoring **measurement error** and **autocorrelation between observations**.
- Under certain simple measurement error models, measurement error actually does not cause underestimation. (Of course if those models don't hold, it will.)
- The reason is similar to the fact that variance for a multi-stage sample design can be estimated almost unbiasedly when first stage fpc is small with only the between-PSU variance (e.g., see Cochran Section 10.4). All survey sampling software (SAS PROC SURVEYMEANS, R *survey*) does it this way.



A visual explanation for effect of measurement error on estimating variance



Suppose $\hat{Y}_i = Y_i + \epsilon_i$, where $Y_i \sim (\mu, \sigma_y^2)$ & $\epsilon_i \sim (0, \sigma_e^2) \rightarrow \text{Var}(\hat{Y}_i) = \sigma_y^2 + \sigma_e^2$

True Y_i 's (but you can't see them)



Measured \hat{Y}_i 's ($\hat{Y}_i = Y_i + \epsilon_i$)



 's are more variable than  's. Thus s^2 already incorporates variability of ϵ_i .



A nerdy explanation for effect of measurement error on estimating variance

- Suppose $\hat{Y}_i = Y_i + \epsilon_i$, where $Y_i \sim (\mu, \sigma_y^2)$ & $\epsilon_i \sim (0, \sigma_e^2) \longrightarrow \text{Var}(\hat{Y}_i) = \sigma_y^2 + \sigma_e^2$
- Let $\bar{\hat{Y}}$ denote the sample mean of $n \hat{Y}_i$'s.
- $\text{Var}(\bar{\hat{Y}}) = \frac{\text{Var}(\hat{Y}_i)}{n} = \frac{\sigma_y^2 + \sigma_e^2}{n}$.
- But the estimated sample variance of $\bar{\hat{Y}}$ if we ignore measurement error is $\frac{\hat{s}^2}{n}$. But $E\left(\frac{\hat{s}^2}{n}\right) = \frac{\sigma_y^2 + \sigma_e^2}{n}$ so it estimates the inflated variability.



Comments on the effect of autocorrelation on estimating variance

- If every transect begins at a random point and the entire transect is taken as a sampling unit, then no underestimation of variance occur with the ratio estimator (i.e., variance of residuals from ratio model). That is, the ratio estimator effectively treats the transect as a cluster, and computes variances appropriately.
- If the transects were broken into pieces and treated as if they start at a random point when they don't, then autocorrelation will cause underestimate variance since the residuals will be correlated.

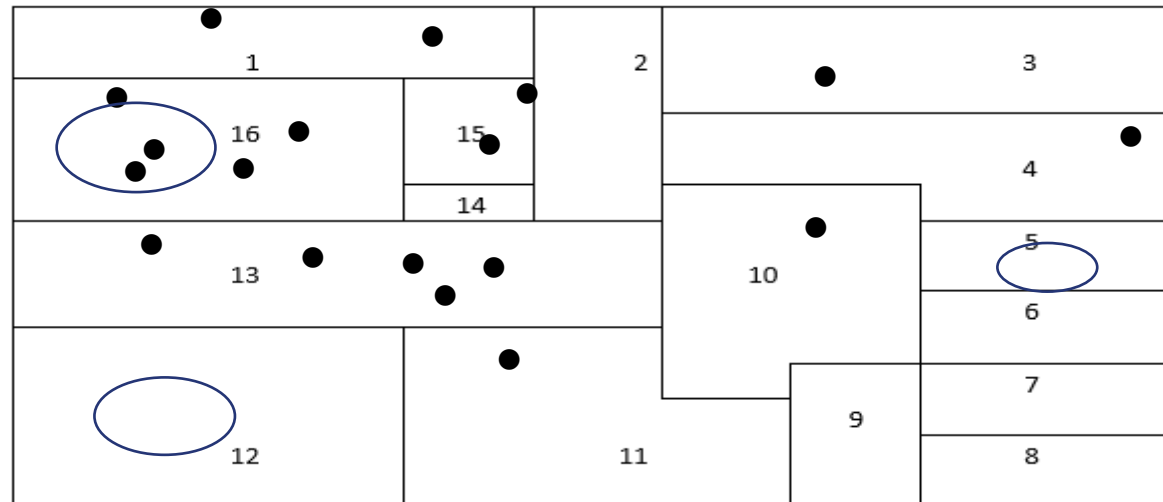


- Treating pipeline data as a random sample may be slightly inaccurate, but not due to autocorrelation, if random starting points are selected.
- Instead it is due to the fact that all transects along the pipeline are treated as having the same probability of selection. Those transects beginning within a half transect of the end of a pipeline have a smaller probability of selection than others.
- This means they should have differing weights, which would affect variance **and** total estimates themselves.
- However, this has the potential of affecting a very small number of sample units.



How does finite sampling point of view relate to estimation approach taken

- We are approximating the sample of transects that could be selected using the “drop a random point and go in a random direction for an arbitrary distance” as defining an idealized population similar to the one below. Here dots are fish and we are interested in estimating the total number of fish from a sample.



The fact that the data are autocorrelated within transects is not relevant, since the randomness comes from the selection.

However, if we split a “transect” in half, we get the same estimate of fish, but a smaller estimate of variance.

My estimate of the number of fish in the stratum would be density in sample * total area

$$= \frac{5+0+0}{12+4+16} * 126 = 20.3$$
 and its SE $v(\hat{t}_{hy,r}) = t_{hx}^2 \times (s_{hd}^2/n_h) \left(1 - \frac{n_h}{\hat{N}_h}\right) = 934$.

