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:

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

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

Planned Sample Design and Analysis

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

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

and

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

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, $\widehat{N}_h = t_{hx}/\bar{x}_h$.
- The estimates and standard errors were computed in SAS PROC SURVEYMEANS

Changes to sample design and analysis plan

- 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.

Changes to sample design and analysis plan

- 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.

Variance estimation

- Several reviewers commented that our variance estimate may be biased low, due to ignoring **measurement error** and **autocorrelation between observations**.
- These will be discussed here. A preview of conclusions are:
- Under certain simple measurement error models, measurement error actually does not cause underestimation of variance. (Of course if those models don’t hold, it will.)
- Autocorrelation within a cluster (transect) does not bias variance estimate in a cluster design when variances are based on transect-to-transect variability.

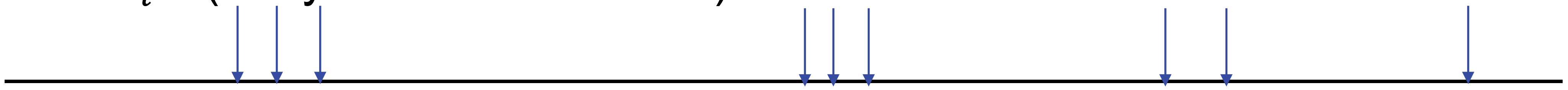
Measurement error

- The reason that measurement error variance is captured in the (measurement error-afflicted) observed values is similar to the reason that variance for a multi-stage sample design can be estimated almost unbiasedly with only the between-PSU variance (e.g., see Cochran Section 10.4). (This is true if first stage fpc is small).
- This is why survey sampling software (SAS PROC SURVEYMEANS, R *survey*) computes variance for multi-stage designs by measuring variability among PSU's.

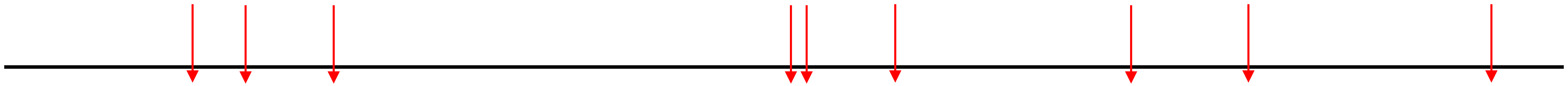
A visual explanation for effect of measurement error on estimating variance

Suppose $\hat{Y}_i = Y_i + \in_i$, where $Y_i \sim (\mu, \sigma_y^2)$ & $\in_i \sim (0, \sigma_e^2)$ ➡ 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 ($\hat{Y}_i = Y_i + \in_i$)



↓ 's are more variable than ↓ 's. Thus s^2 already incorporates variability of $\in_i$.

A nerdy explanation for effect of measurement error on estimating variance

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

Is there neglected autocorrelation in variance estimates from transects?

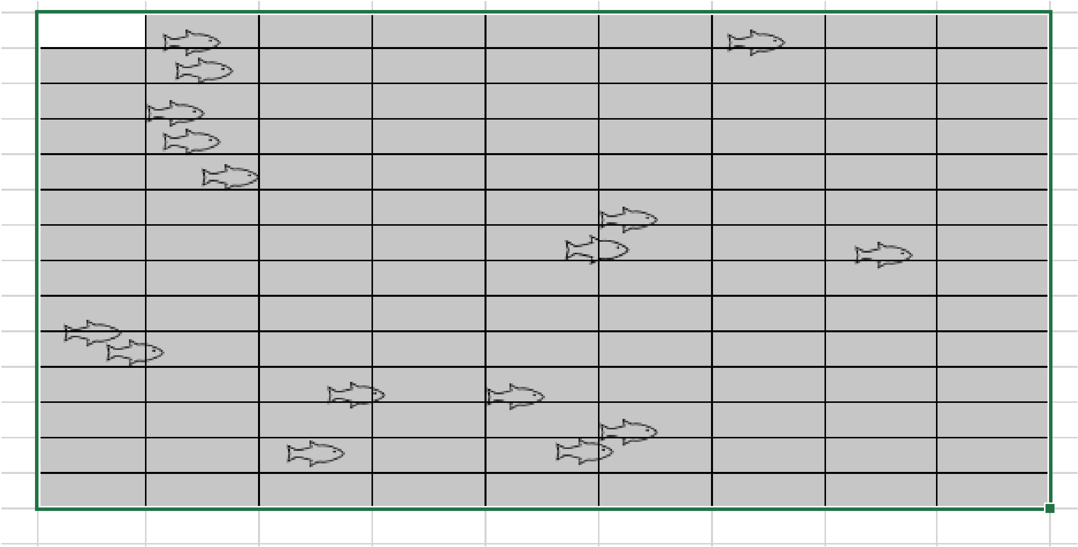
The answer, under general conditions, is NO.

Here is why:
Variances were estimated using a “design-based” approach; i.e., assuming that stochasticity arises ONLY from the randomness of the sample selection procedure.

That means that if sampling units are selected as A SRS, they are by definition, independent.

If two units are NOT selected at random (such as happens when multiple grid units are in a transect but only the transect location is selected randomly), then the transect to transect variance still properly captures the variability, since then the transect Total is the observed value of the sampling unit.

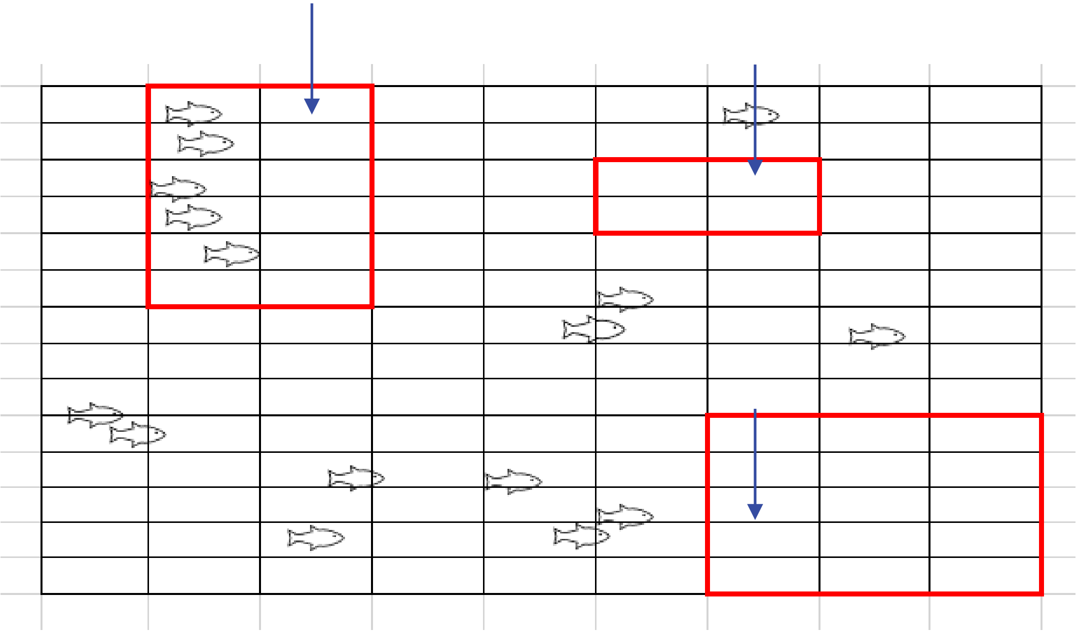
Total area t_x = 126 units



Is there neglected autocorrelation in variance estimates from transects?

Suppose that a sample of 3 transects are planned, each to be started at a randomly chosen point (grid cell). Then a transect is run from that grid in a random direction. Transects may differ in area due for practical reasons.

3 sample grids are indicated by arrows, and the transects run in red.



The estimate of the number of fish in the universe would be density in sample * total area

$= \frac{5+0+0}{12+4+15} * 126 =$ 20.3 and its SE $\sqrt{v(\hat{t}_{hy,r})} = \sqrt{t_{hx}^2 \times (s_{hd}^2/n_h)\left(1-\frac{n_h}{\hat{N}_h}\right)} = \sqrt{126^2 \times (s_{hd}^2/n_h)\left(1-\frac{31}{126}\right)} =$ 175

The fact that the data are autocorrelated within transects is not relevant, since the observations are transect totals,and transects are independently selected.

When would the autocorrelation bias variance estimates?

However, if we split a “transect” In half, we would get the same estimate of fish from the ratio estimator but a smaller estimate of variance for two reasons: (1) a larger sample size and a smaller unit-to-unit variability. This would be a incorrect variance estimate.

Also, if the size of the transect was correlated with the number of fish in the transect, then the estimator itself would be biased. The variability of the transect sizes are due to issues unrelated to number of fish, we believe.

If there is a two-stage sample, in which transect grids are subsampled, then SE of final estimate would be larger. However, as noted earlier, a nearly unbiased ESTIMATE of this larger variance can be made from transect to transect estimated totals capture variance from both stages.

Comments on the other effects on estimated variance (pipelines)

- Treating pipeline data as a SRS may be slightly inaccurate, but not due to autocorrelation.
- 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 could affect variance **and** abundance estimates themselves.
- However, this has the potential of affecting a very small number of sample units.
- If pipeline sites are not selected independently from sites in other strata, that can produce a covariance between estimates between the two strata. This was not taken into account and could bias variance downward. However, the contribution to total abundance is relatively quite a small fraction of the total, so will likely be a small contributor to variance. I believe the impact could actually be estimated if detailed knowledge of the pairing were considered.

A question about bias from low sampling rate

- A question was raised about whether a small sampling rate causes an underestimate of abundance estimate.
- Bias means that if there was repeated sampling, the average over all estimates would be smaller (or larger) than the true abundance.
- The estimators we used were either unbiased (mpu estimator) or asymptotically unbiased (ratio estimator).That means that regardless of sample size (if it is large enough) there is no or negligible bias.
- What is true, however, is that the variance (actually CV) of abundance estimate can be large. This means that the estimate can be far from the true value, but the direction is not predictable.
- Poor allocation to strata in a design likewise does not affect bias, but only raised the variance. Allocation is poor if sample is not allocated proportionally to N_hS_h