Your Favorite Watershed - Stratification Choices

Carl Schwarz

2017-05-14

Contents

1 Introduction

The assignment of a watershed to the FSI category depends on the median CPUE estimated for that watershed and the uncertainty in the estimated median. The uncertainty in the estimated median in turn depends upon the variability in the CPUE values and the number of CPUE values measured in a particular year.

In the case of a simple random sample, the uncertainty in the median can be easily computed. Let Y represent the $log(CPUE)$. Then the $log(median)$ of the CPUE is estimated as the simple sample mean of the $log(CPUE)$, i.e. $\overline{Y}$. The uncertainty of $\overline{Y}$ is denoted as the standard error (SE) and is computed as $SE(\overline{Y}) = s/\sqrt{n}$ where s is the sample standard deviation of Y and n is the sample size.

In some cases, the uncertainty in the log($median$) CPUE can be reduced by stratifying the CPUE. Two potential stratification variables are the stream order or the lower levels of the HUC classification (e.g. the watershed or sub-watershed levels).

In order to compute a mean from a stratified design, on additonal piece of information is needed, namely the relative (population) weights of the strata. For example, the entire sub-basin could composed of 30, 20, and 10 streams of order 2, 3 and 4 respectively. Then the weights for the stream-order stratificaton are $30/60 = 0.5$; $20/50 = 0.333$; and $10/60 = 0.167$ respectively. Notice that the population weights do not have to correspond to the division by sample size, i.e., you could over/under sample in a stratum relative to its population weight.

After stratification, let

- n_i be the observed sample size in stratum i, $i = 1, ...H$ where H is the number of strata.
- $\overline{Y}_i$ be the sample mean for stratum i,
- s_i be the sample standard deviation for stratum i,
- SE_i be the standard error of the mean in stratum i computed as $SE_i = i/\sqrt{n_i}$, and
- W_i be the population weight for stratum i.

The overall mean is found as a weighted average of the stratum means:

$$\overline{Y}_{overall} = W_1\overline{Y}_1 + W_2\overline{Y}_2 + ...$$

The standard error of the overall mean is found as

$$SE(\overline{Y}_{overall}) = \sqrt{W_1^2 SE_1^2 + W_2^2 SE_2^2 + ...}$$

A stratification could lead to reductions in the uncertainty of the overall mean if the values within a stratum have a smaller standard deviation than values across strata, i.e. units within a strataum are homogeneous while units across strata are heterogeneous.

How should units be allocated to strata? There are several possible methods, but the most common are:

- Equal allocation where the total sample size is divided equally among the strata.
- Proportional allocation where the total sample size is divided proportionally among the strata in the same proportion as the population weights. For example, if one stratum had a population weight (W_i) of 0.4, the 40% of the samples should be allocated to this stratum. In other words, larger strata receive more samples.
- Optimal allocation where the total sample size is divided proportionally to the PRODUCT of the population weights and the population standard deviations, i.e .proportion to $W_i S_i$ where S_i is the population standard deviation. Normally, the population standard deviation is unknown, but can be estimated from the respective stratum sample standard deviation (s_i). In other words, large and more variable strata receive more samples.

In most cases, moving from an equal allocation to a proportional allocation provides a large improvement in precision while moving to the optimal allocation only provides a further (modest) improvement.

It is straight forward to compare the efficacy of alternate stratification methods by finding the characteristics of the stratum (mean, standard deviation), dividing a proposed sample by the various allocation methods (equal, proportional or optimal) and then comparing the resulting standard errors of the mean.

2 Extracting the information

CAUTION: The FWIS output format does not yet contain the needed information to examine different stratifications.

Consequently, this template will use a summary data file especially created for this example. This data was extracted from the Nordegg River and contains information on the HUC10 and Stream Order of a number of sampling points selected in 2016.

The data are read in the usual manner. We then find the log(CPUE) and add a small constant to avoid taking log(0):

```
catch.summary <- readxl::read_excel(file.path("Data","Nordegg2016SiteCUESummary_streamorder_CS.xlsx"),
                                    sheet="Sheet1", skip=1)
names(catch.summary) <- make.names(names(catch.summary))
range(catch.summary$BKTR_100m..recomputed.)
```

```
## [1]  0 34
```

```
offset <- 0.5 * min(catch.summary$BKTR_100m..recomputed.[catch.summary$BKTR_100m..recomputed.>0])
offset
```

```
## [1] 0.05
```

```
catch.summary$logCPUE <- log(catch.summary$BKTR_100m..recomputed. + offset)
```

Stratification by HUC10 code and Stream Order will be considered. The stratum statistics are:

Table 1: Stratum statistics for HUC stratification

HUC10	mean	sd
Lower Nordegg River	-2.1	1.8
Middle Nordegg River	1	2.1

HUC10	mean	sd
Upper Nordegg River	0	2.2

Table 2: Stratum statistics for STREAM ORDER stratification

STR_ORDER	mean	sd
2	-1.7	1.9
3	-0.6	2.3
4	1.7	2
5	-0.7	1.5

The FWIS databased likely does NOT include the stratum population weights. You will have to extract this from a GIS analysis of the watershed and then enter the weights here. I've used arbitrary values as an illustration of the methods and results.

Table 3: Population weights for HUC stratification

HUC10	W
Lower Nordegg River	0.23
Middle Nordegg River	0.56
Upper Nordegg River	0.21

Table 4: Population weights for STREAM ORDER stratification

STR_ORDER	W
2	0.05
3	0.54
4	0.31
5	0.1

3 Comparing precision under no stratification and potential stratifications.

Using an arbitary total sample size of 100 for convenience, we allocate this total sample size among the strata using the equal, proportional, or optimal allocation:

Table 5: Allocations for HUC stratification

HUC10	W	mean	sd	n.equal	n.prop	n.opt
Lower Nordegg River	0.23	-2.1	1.8	33.3	23	20.2
Middle Nordegg River	0.56	1	2.1	33.3	56	57.8
Upper Nordegg River	0.21	0	2.2	33.3	21	22

Table 6: Allocations for STREAM ORDER stratification

STR_ORDER	W	mean	sd	n.equal	n.prop	n.opt
2	0.05	-1.7	1.9	25	5	4.4
3	0.54	-0.6	2.3	25	54	59
4	0.31	1.7	2	25	31	29.5
5	0.1	-0.7	1.5	25	10	7.1

Finally we compute the SE of the grand mean using the above formula for each allocation:

Now we can compare the forecasted standard errors under the two potential stratication variables and the three potential qllocation strategies as summarized below:

```
res <-rbind(huc.SE, strorder.SE)
pandoc.table( res,
              caption='Comparing final SE under different stratification and allocations',
              round=c(0,2,2,2,2),keep.trailing.zeros=TRUE,
              justify='lrrrr')
```

Table 7: Comparing final SE under different stratification and allocations

Strat.method	SE.no.strat	SE.equal	SE.prop	SE.opt
HUC	0.24	0.23	0.21	0.21
Stream Order	0.24	0.28	0.22	0.21

For this example, the two stratification methods appears to give similar final standard errors. There is some improvement from stratifying over computing the mean with no stratification.