Sampling design CI

Bill Kristan

10/18/2021

## Confidence intervals for estimates of greenness and wetness

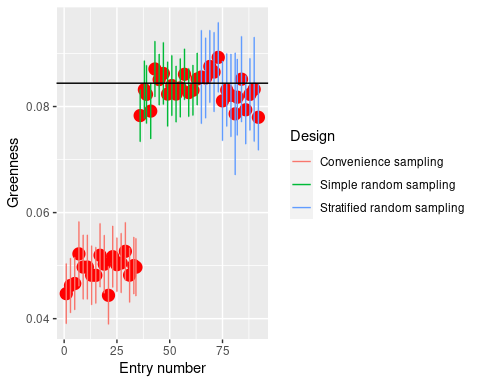
Confidence intervals were submitted by 21 students, but several had to be discarded because of data entry errors. The data with usable confidence intervals were imported:

Import the data:

read.csv("student\_results\_f21.csv") -> ci

Plot the greenness estimates and intervals, with the true value 0f 0.0844 shown as a horizontal line:

library(ggplot2)  
  
ggplot(subset(ci, ci$Variable == 'Greenness'), aes(x = Int.number, y = Estimate, color = Design, group = Design)) + geom\_point(size = 4, color = "red") + geom\_errorbar(aes(ymin = LL, ymax = UL), width = 0.1) + geom\_hline(yintercept = 0.0844) + labs(x = "Entry number", y = "Greenness")

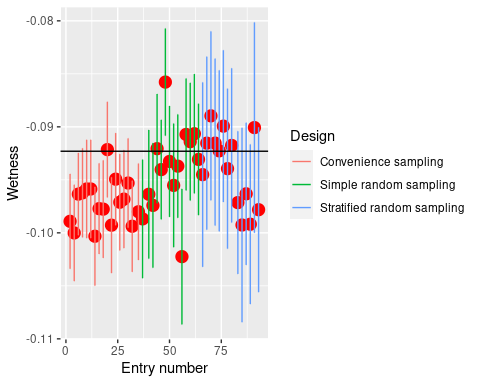


There are 18 convenience sampling intervals, 15 SRS intervals, and 15 StRS intervals. We would only expect 5% of these not to contain the population mean of 0.0844, 5% of 18 or 15 is less than 1, so we would expect 0 or 1 interval to fail to include the true value with each of the designs. There is one interval that does not include 0.0844 from simple random sampling, and none from stratified random sampling. Convenience sampling never included the true value, and the estimates were consistently lower than the true value by a large margin.

The confidence interval widths do not differ markedly between the methods - convenience sampling is thus just as precise as the other two methods, but is inaccurate. Stratifying seems not to have helped with precision, which could be an indication that the greenness was not very different between the strata, such that calculating separate means for each did not improve precision.

Plot the wetness results:

ggplot(subset(ci, ci$Variable == 'Wetness'), aes(x = Int.number, y = Estimate, color = Design, group = Design)) + geom\_point(size = 4, color = "red") + geom\_errorbar(aes(ymin = LL, ymax = UL), width = 0.1) + geom\_hline(yintercept = -0.0923) + labs(x = "Entry number", y = "Wetness")



The bias in wetness along roads is less pronounced than it was for greenness - the estimates are generally low, and there are both more intervals that do not contain the true mean for convenience sampling than expected, and more than for the other two sampling designs (7 of the 17 intervals do not contain the true mean for convenience sampling, but only two of 15 for SRS, and none for StRS).

Again, none of the methods had drastically different confidence interval sizes - the primary problem was the bias caused by sampling along roads, but all of the estimates were approximately equally precise.

Given these results, convenience sampling is a poor choice compared to either SRS or StRS. The choice between SRS and StRS is less clear, and would probably be decided based on whether having estimates of means for each strata is useful in ways other than estimating the mean for the watershed.