Environ 710 Lab 3: Sampling

Natalia Neal-Walthall

9.14.18

Introduction

The Clean Water Act, which is enforced by The Environmental Protection Agency (EPA), requires that all US states assess water quality parameters such as dissolved oxygen levels, pollutant loading, and fecal coliform contamination on an on-going basis. When violations are present in a waterway, it is considered "impaired" and the state must take action to bring water quality back into compliance. EPA guidelines define a stream segment as impaired when greater than 10% of the measurements taken for a given water quality parameter exceed some defined criteria. However, obtaining enough values to ensure confidence in the results presents something of a challenge. Sampling and analysis are often costly and time consuming, so it is desirable to utilize a method which provides reliable results from a limited number of observations.

Statistical methods, such as the binomial method employed herein are important for understanding how well a given sampling event predicts reliable results. However, care must be taken when using these methods to apply appropriate parameters to limit error inherent to the process. An overabundance of type 1 errors (false positive) will result in unnecessary costs to monitor the site and implement control technologies. On the other hand, an overabundance of type 2 error (false negatives) may pose a serious risk to human and environmental health. The type one error rate is chosen by the analyst; in this case, it is 10%. Once this value has been determined, the type II error rate for sample size n can then be calculated. If the type II error rate is unnacceptably large, it may be reduced by choosing a greater type I error rate, increasing sample size, and/ or by decreasing measurement uncertainty.

This report explores the results obtained in similar analyses when the number of simulated rivers and theoretical samples (observations) is varied over a wide scale, while keeping parameters resulting from sample analyses (mean and standard deviation) constant.

Methods

To conduct this simulation, I completed the following steps:

1. Construction of the "rivers" function which simulates random sampling:

Takes input (number of rivers, observations).

Randomly generates the required number of values from a binomial distribution based on the true distribution input.

Counts the number of impaired rivers for each simulation, where impairment is pollutant concentration >6 in 10% of more of observations as per EPA guidelines.

Returns a value for percent of rivers considered impaired for each simulation.

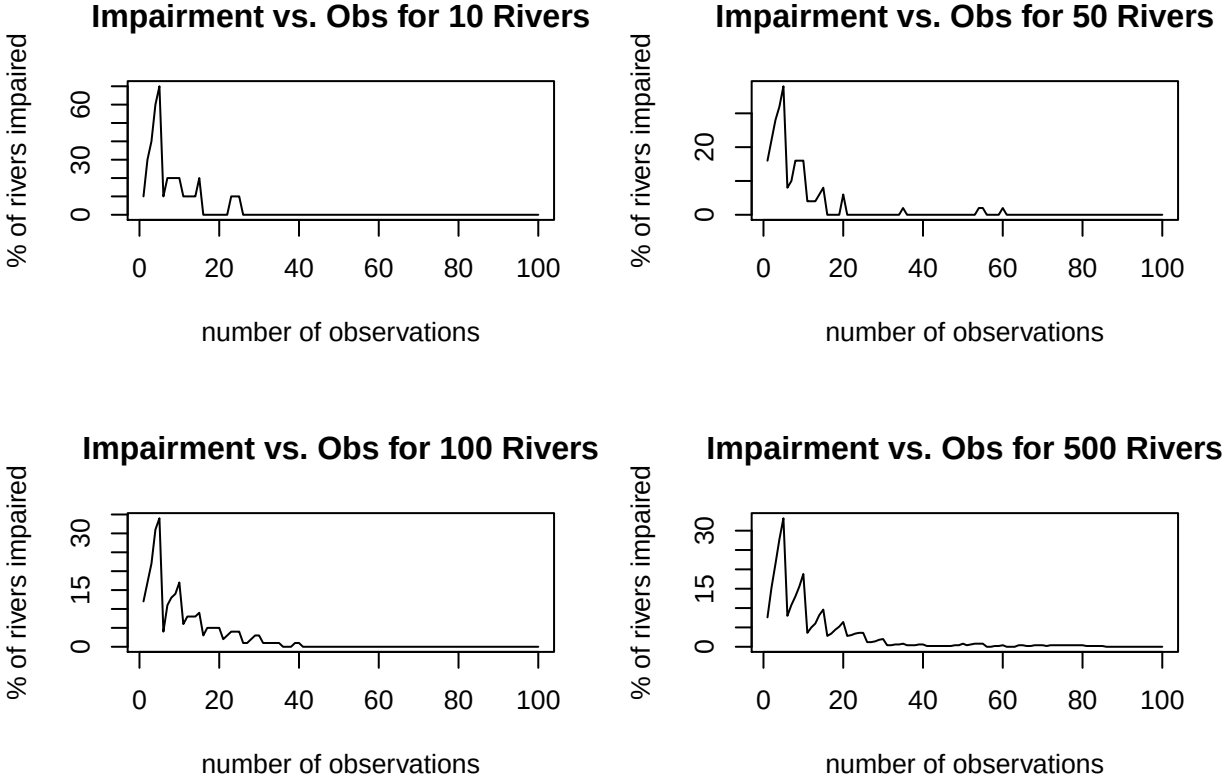
2. Construction of for loop which utilizes the "rivers" function to return percent of rivers impaired for 10, 50, 100, and 500 rivers, with observations ranging from 1-500.

3. Graphed the percent impairment vs. obserations for each river simulation (10, 50, 100 & 500 rivers) to gain a clearer picture of the direct effect number of observations has on the percent of rivers declared impaired. However, based on the initial trial it was immediately apparent that for all cases, number of observations greater than 100 was superfluous and including in the graph made details in the more dynamic portion of the graph indecipherable. Therefore, the loop was ammended to include observations from 1-100. These graphs are shown below.

4. Examined the effects of sample size on standard error.

Results and Conclusions

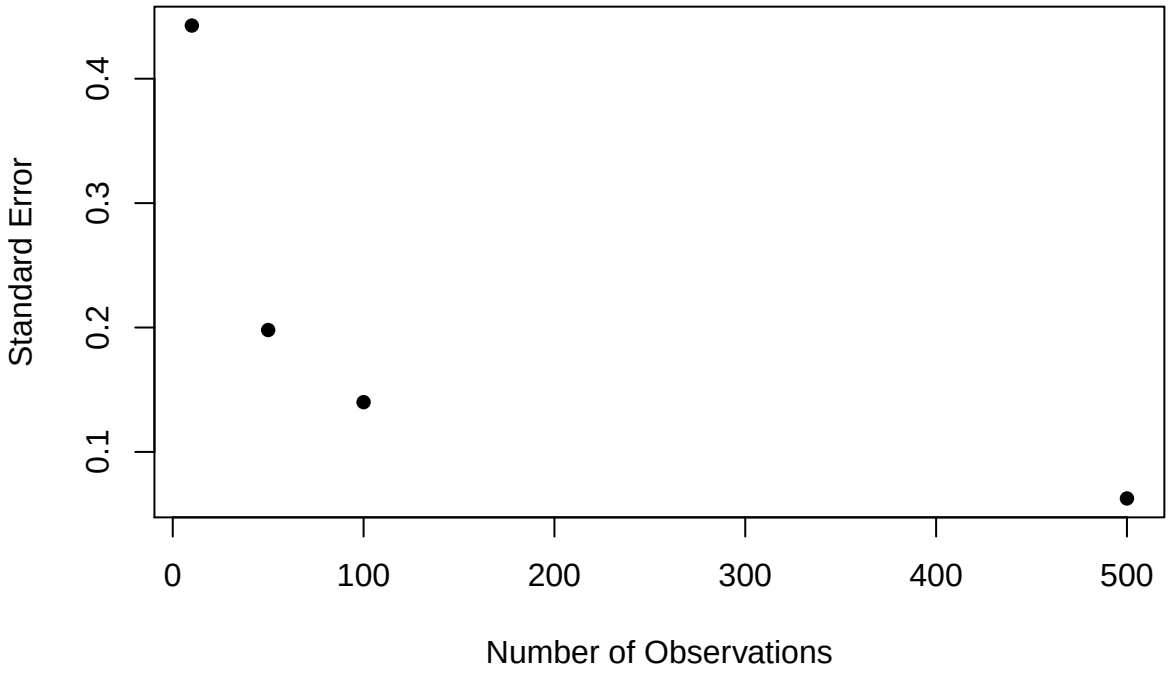
As shown in figure 1, the simulations for varying numbers of rivers showed similar results in that all tended to have significant spikes in percent of rivers impaired when fewer observations were taken and trended to zero within 100 observations. Not shown are the values from 100-500 obsevations, but it was investigated and % impairment remained at zero thrughout.



Though general patterns are similar, there were some interesting differences among the results obtained for 10, 50, 100, and 500 rivers. In particular, the "spike" effect for a low number of observations was of greatest importance for 10 rivers. In this case, roughly 70% of rivers were shown to be impaired, whereas other simuations showed around 40% maximum.

Based on the results of this simulation, it appears that the EPA method will likely result in an overabundance of false positive results unless greater than 40 measurements are taken. This may be feasible for some parameters, but may not be for others and this should be taken into account

when designing a sampling protocol. As can be seen below, the effects of sample size on standard error, which is often used to indicate the validity of the mean, will have a huge impact based on a standard error of 1.4. Having a small sample size also leaves results particularly vulnerable to effects from unforseen heterogenaity or seasonal variation. It is therefore crucial that analysts carefully consider wheather the samples obtained to provide the mean and SD used were true replicates and/or increase the number of observations taken wherever possible.



‘‘‘

```
#Function takes input for (rivers, observations), returns a value for number of impaired river
rivers <- function (x, y)
{
set.seed(1001)
h2o <- as.data.frame(matrix(rnorm(x*y, mean=4, sd=1.4), nrow = x, ncol= y))
rownames(h2o) <- paste(rep("Riv", nrow(h2o)), c(1:nrow(h2o)), sep = "")
colnames(h2o) <- paste(rep("Obs", ncol(h2o)), c(1:ncol(h2o)), sep = "")
h2oTest <- ifelse(h2o>=6, 1, 0)
imp <- (rowSums(h2oTest))/y
fin <- ifelse(imp>=.2, 1, 0)
(sum(fin)/x)*100
}
#vectors to store data from for loop
r10 <- c()
r50 <- c()
r100 <- c()
r500 <- c()
#for loops for 10, 50, 100, 500 rivers with observations ranging from 10-500
for(i in 1:100) {
r10[i] <- rivers(10,i)
```

```
r50[i] <- rivers(50,i)
r100[i] <- rivers(100, i)
r500[i] <- rivers (500, i)
}
# 4 figures arranged in 2 rows and 2 columns
par(mfrow=c(2,2))
plot(x = 1:100, y = r10, xlab = "number of observations", ylab = "% of rivers impaired", type =
plot(x = 1:100, y = r50, xlab = "number of observations", ylab = "% of rivers impaired", type =
plot(x = 1:100, y = r100, xlab = "number of observations", ylab = "% of rivers impaired",type =
plot(x = 1:100, y = r500, xlab = "number of observations", ylab = "% of rivers impaired",type =
#Calculating and plotting Standard Errors
SE10 <- 1.4/(sqrt(10))
SE50 <- 1.4/(sqrt(50))
SE100 <- 1.4/(sqrt(100))
SE500 <- 1.4/(sqrt(500))
SE <- c(SE10, SE50, SE100, SE500)
OBS <- c(10, 50, 100, 500)
plot(OBS, SE, xlab = "Number of Observations", ylab = "Standard Error", pch = 16)
```