Lab 3 problems

Grace Randall

2024-01-31

Introduction

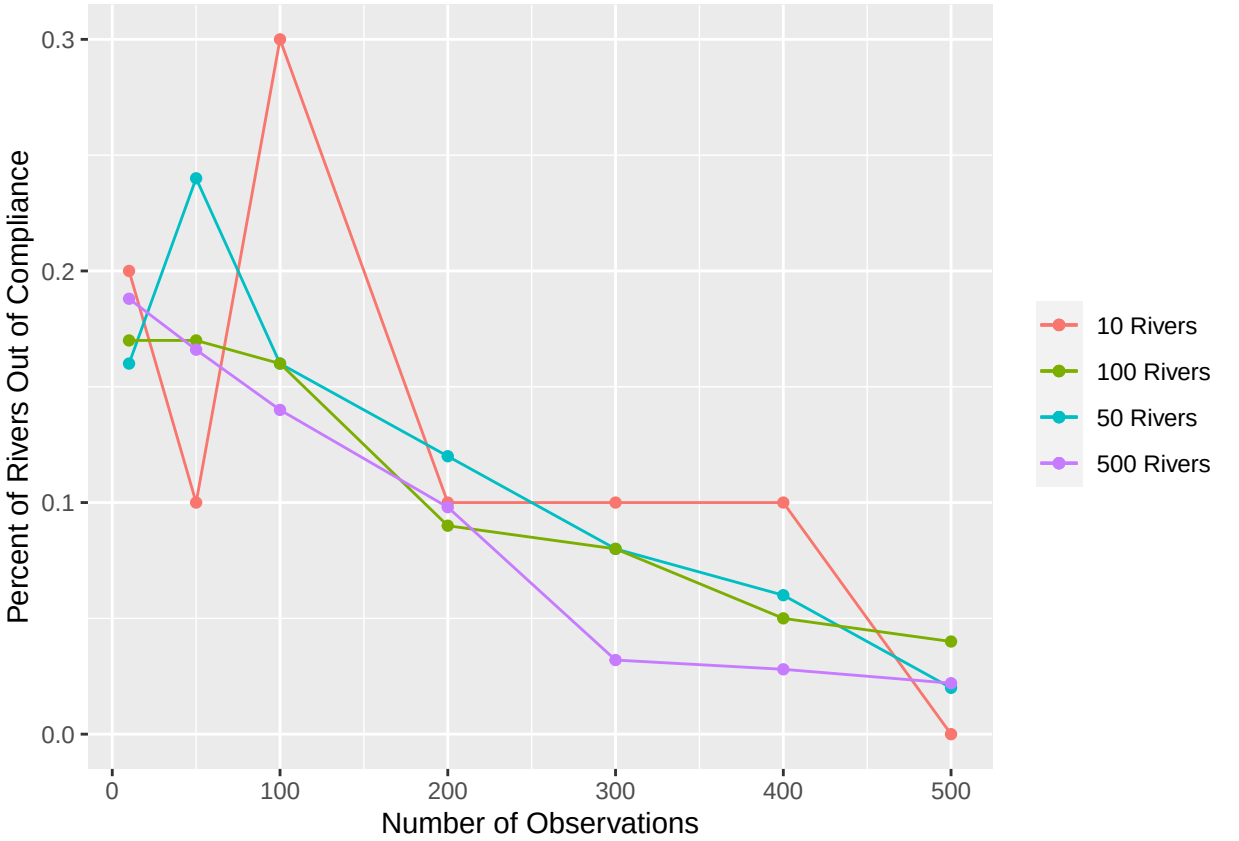
For this report I examine the effect of sampling probabilities on the determination of the compliance of rivers to the regulations of water quality set out by the EPA. The EPA requires that the concentration of a contaminant in the river does not exceed the concentration in the regulation in more than 10% of the measurements taken. While this is intended to require that the concentration of the contaminant does not exceed the limit more than 10% of the time, the actual regulation is dependent on the sample that is taken. This sample may or may not accurately represent the conditions in the river, however it is impossible to measure the concentration everywhere in the river. This leads to the question of how measurements should be taken to assess if the rivers are out of compliance. If the assessment is done on an insufficient sample, then this could lead to compliant rivers being classified as uncompliant and vice versa, which could led to a waste of resources or danger to the public respectively.

Methods

To gain more insight into this issue I attempted to look into the sampling distribution of the percentage of the rivers that would be inaccurately considered to not be in compliance. I simulated the sampling of rivers with different numbers of observations per river and replicated it with both a small and large number of rivers. I used a hypothetical situation where the true distribution of concentrations of pollutant concentration in the river followed a log normal distribution with a mean of 4 and a standard deviation of 1.4. In this hypothetical the log of the standard was 6, so this river should be considered to be in compliance with the regulations. Through repeated simulated sampling I was able to assess how often this river would be inaccurately assessed with different numbers of measurements taken.

Results and Conclusion

Multiple simulations made trends in the percent of rivers misclassified evident. As the number of measurements per river increased the percentage of rivers that were misclassified decreased. This is to be expected because as the number of measurements increases the distribution of the measurements should become closer to the true distribution of values in the river, so there is less likely to be a set of measurements that, as a whole, lies far from the center of the true distribution. With increasing numbers of rivers being tested this pattern became more clear. When there were only 10 rivers, the percent of rivers that were misrepresented did not show a clear relationship with the number of observations, whereas when there were 500 rivers this relationship was clear. The figure below shows an example of how the percent misidentified changed with number of observations and number of rivers.



The decision of how many measurements should be taken when determining the compliance of a river is a trade off between the certainty with the results and the effort needed to make the determination. While the simulations with 500 measurements had the lowest misidentification rate, taking this many measurements of every river that needs to be tested is not feasible in most cases. Taking a lower number of measurements is more achievable, but it requires one to risk the dangers of misidentifying the status of a river.

I would recommend that the EPA require at least 100 measurements of each river. This is an achievable number of measurements but it resulted in a misidentification rate of about 15% when considered over many rivers, which is fairly good. The high misidentification rate in the simulation with 10 rivers shows that this percentage of misidentification can still lead to a relatively high number of misidentifications. Because of this continued level of error, I would also recommend that the EPA look for opportunities where the number of observations could be easily increase. For example if it is possible to implement a device to take periodic concentration measurements without hands on effort. This would could greatly improve the number of observations without significantly increasing effort. I would also suggest increasing the number of observations in situations where it is especially essential to accurately gauge the concentrations in the rivers, such as with a contaminant that can easily be deadly. This may be essential because increasing the number of observations should not only decrease the chance of misidentifying a compliant river as noncompliant, but also of misidentifying a noncompliant river as compliant.

Appendix

```
Compliance_test <- function(num_obs,num_riv){

  set.seed(1001)
  too.many <- round(0.10 * num_obs, digits = 0)

  h2o <- as.data.frame(matrix(rnorm(num_riv*num_obs, mean=4, sd=1.4),
  ncol= num_obs))

  rownames(h2o) <- paste(rep("Riv", nrow(h2o)), c(1:nrow(h2o)), sep = "")
  colnames(h2o) <- paste(rep("Obs", ncol(h2o)), c(1:ncol(h2o)), sep = "")
  h2o$Test <- rowSums(ifelse(h2o>6, 1, 0))
  return(length(h2o$Test[h2o$Test>too.many])/num_riv)

}
test_small <- Compliance_test(10,500)

count_r <- c(1, 2, 3, 4)
River_nums <- c(10, 50, 100, 500)

count_o <- c(1, 2, 3, 4, 5, 6, 7)
obs_nums<- c(10, 50, 100, 200, 300, 400, 500) #i wanted more info

Tests <- data.frame(matrix(ncol = 4, nrow = 4))

rownames(Tests) <-
  c("10 Observations","50 Observations","100 Observations","500 Observations")

colnames(Tests) <-
  c("10 Rivers","50 Rivers","100 Rivers","500 Rivers")

for (i in count_o) {
  for (j in count_r) {
  Tests[i,j] <- Compliance_test(obs_nums[i],River_nums[j])
}
}

library(ggplot2)
Plot <- ggplot(data=Tests)+
  geom_point(aes(x=obs_nums, y=Tests$`10 Rivers`,color="10 Rivers"))+
  geom_line(aes(x=obs_nums, y=Tests$`10 Rivers`,color="10 Rivers"))+

  geom_point(aes(x=obs_nums, y=Tests$`50 Rivers`,color="50 Rivers"))+
  geom_line(aes(x=obs_nums, y=Tests$`50 Rivers`,color="50 Rivers"))+

  geom_point(aes(x=obs_nums, y=Tests$`100 Rivers`,color="100 Rivers"))+
  geom_line(aes(x=obs_nums, y=Tests$`100 Rivers`,color="100 Rivers"))+

  geom_point(aes(x=obs_nums, y=Tests$`500 Rivers`,color="500 Rivers"))+
  geom_line(aes(x=obs_nums, y=Tests$`500 Rivers`,color="500 Rivers"))+

  ylab("Percent of Rivers Out of Compliance")+
  xlab("Number of Observations")+
```

```
  theme(legend.title=element_blank())
Plot
```