Lab 3 Writeup

Jiahuan Li

2023-02-06

Introduction

According to Environmental Protection Agency guidelines, a stream segment is considered impaired when more than 10% of water quality measurements exceed numeric criteria (Smith et al., 2001). This is known as a "raw score" assessment method. Water quality measurements were taken from samples collected from a variety of water quality conditions. Thus, it is reasonable to view the assessment process as a statistical decision problem. And in this context, appropriately sampling water bodies is of great importance since it is a critical concern in the statistic field. The objective of this report is to evaluate the impacts of the sample size in the sampling process of water quality evaluation.

Methods

To check the influence of sampling water bodies with different sample sizes, a function simulating the process is applied and the results are further examined at different levels. At first, I created a function with the number of rivers and the observations for each river as inputs. It is designed to directly yield the percentage of impaired river.

After that, I select to simulate certain sampling of 10, 50, 100, and 500 independent and identically distributed rivers with 10, 50, 100, and 500 observations per river and observe their results. Due to the irregularity of their impaired percentage results, I then involved more sampling and illustrate the percentage trend associated with the observation increase. At last, I compared the sampling results with the supposed theoretical results to evaluate the effectiveness.

Results & Conclusion

Certain sampling

Results for the sampling of 10, 50, 100, and 500 rivers with 10, 50, 100, and 500 observations per river are shown below, respectively:

```
## [1] 10.0  0.2
```

```
## [1] 50.00  0.24
```

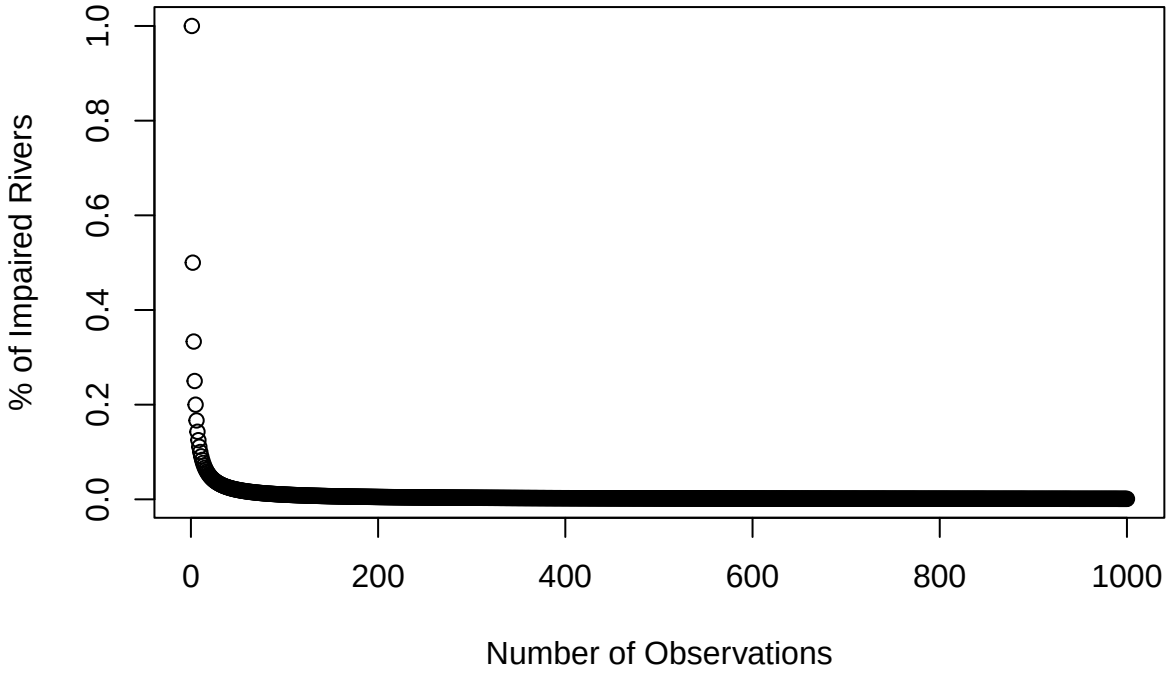
```
## [1] 100.00   0.16
```

```
## [1] 500.000   0.022
```

The first number in each row indicates the number of observation while the second means the percentage of impaired rivers. We can find that although the rivers are identically distributed, the percentage results are not similar, which implies some of the sampling do not reflect the characteristics of the entire population appropriately.

Sampling trend

A graph below illustrate the change of the percentage of impaired rivers with the increase of the number of observations. We can find that the impaired proportion declines rapidly from 0 to 40 obs and then keeps flat. With large sample size, we can detect that these rivers are little likely to be recognized as impaired rivers. But this finding cannot be conclude from a sample with relatively small sample size.



Theoretical examination

Whether the finding from the sample analysis is valid can be examined through theoretical calculation. Unlike the real case, we know the exact population characteristics of the rivers. The true distribution of the log concentration is $\texttt{N}(\mu = 4, \sigma = 1.4)$. Thus, according to the EPA standards, we can calculate the 90 percentile of the rivers' pollutant concentration is $5.794 < 6$. From another perspective, the cumulative percentage of the concentration equals to 6 can be calculated as $0.923 > 0.9$. Therefore, theoretically, these rivers are not impaired rivers, which coincides with the sampling results above.

From the above analysis, we can find that the sample size should be large enough to reflect the real characteristics of the population. Although EPA standards are clear and easy to execute, it requires large sample size to produce convincing results as revealed in this report, which could cost substantial time and money. However, these efforts are unavoidable according to the findings.

References

Smith, E.P., Ye, K., Hughes, C., and Shabman, L. (2001) Statistical assessment of violations of water quality standards under section 303(d) of the Clean Water Act. Environmental Science and Technology, 35(3): 606-612.

Appendix

```
knitr::opts_chunk$set(echo=F, eval=T)
# create the function
h2o <- function(riv, obs) {
  set.seed(1001)
  df <- as.data.frame(matrix(rnorm(riv * obs, mean = 4, sd = 1.4), ncol = obs))
  paste(rep("Riv", nrow(df)), c(1:nrow(df)), sep = "")
  rownames(df) <- paste(rep("Riv", nrow(df)), c(1:nrow(df)), sep = "")
  colnames(df) <- paste(rep("Obs", ncol(df)), c(1:ncol(df)), sep = "")

  df$Test <- rowSums(ifelse(df > 6, 1, 0))   #impair judgment
  impaired <- length(df$Test[df$Test > 0.1 * obs])
  percent_impaired <- impaired / riv  #calculate percentage of impaired rivers
  return (percent_impaired)  # return the percentage of impaired rivers
}
# simulate certain samples using the function above
a = c(10L, h2o (10, 10))
a
b = c(50L, h2o (50, 50))
b
c = c(100L, h2o (100, 100))
c
d = c(500L, h2o (500, 500))
d
percent = h2o (1:1000, 1:1000)
observation = c(1:1000)
plot (observation, percent, xlab = "Number of Observations", ylab = "% of Impaired Rivers")
# theoretical values
e = pnorm (6, 4, 1.4)
f = qnorm (.9, 4, 1.4)
```