1. Introduction of experiment

The EPA once required that a water body be declared as “impaired” if 10% of the water quality measurements exceed the limit of a standard. This rule is intended to ensure that the water quality standard is violated at most 10% of the time.

This experiment is about constructing a statistical model (normal distribution), running simulation, collecting data and evaluating EPA’s rule. Eventually, we will prove that EPA’s rule cannot reach its expectation, which is to ensure the violation is at most 10%.

To evaluate the rule, we will repeatedly sample from a water body that is known to be in compliance and determine **how often** the rule will declare the water impaired. Since we already know the distribution of the water quality pollutant, I will write R code and let computer simulate the sampling process. I will change the sampling number from each river and will change the number of rivers to sample from.

2. Method

First, I will build a statistical model of the water quality pollutants which is normal distribution: N(4, 1.4), then I will repeat the sampling process by R’s rnorm() function;

Then I will use a matrix to store the outcomes and compute the number of measures that are above standard (6 in this case);

Then I will compute the ratio of “unhealthy” rivers based on EPA’s rule, and by changing sample size, I will record each ratio for each size.

Finally I will draw a histogram and dot plot to show the test result for each simulation.

3. Results and conclusion

If we take 10 observations from each of 10, 50, 100, 500 rivers, the ratio of rivers declared as “impaired” are:

*0.700 0.540 0.490 0.522*

If we take 50 observations from each of 10, 50, 100, 500 rivers, the ratio of rivers declared as “impaired” are::

*0.300 0.420 0.340 0.358*

If we take 100 observations from each of 10, 50, 100, 500 rivers, the ratio of rivers declared as “impaired” are::

*0.200 0.260 0.230 0.206*

If we take 500 observations from each of 10, 50, 100, 500 rivers, the ratio of rivers declared as “impaired” are::

*0.000 0.000 0.020 0.038*

From above, we can see that only when the sample size gets large (at 500) can the ratio drop under 10%. When the sample size is small, the ratio of wrong result can get significantly high (as high as 70% in this experiment on 10 rivers)

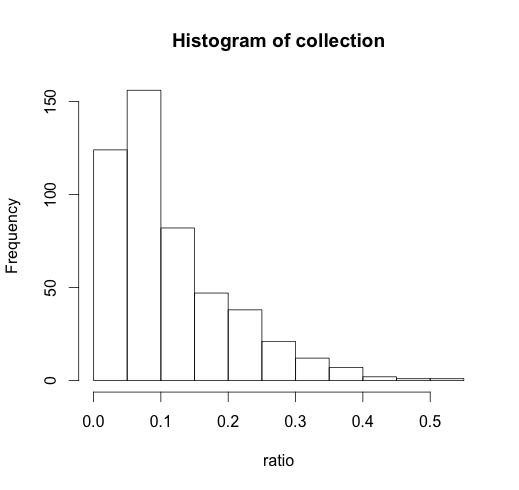
I take the sample size from 1 to 500 with following R code (evaluatefunc is the function I write for calculating the ratio of “impaired” rivers):

*collection <- c()*

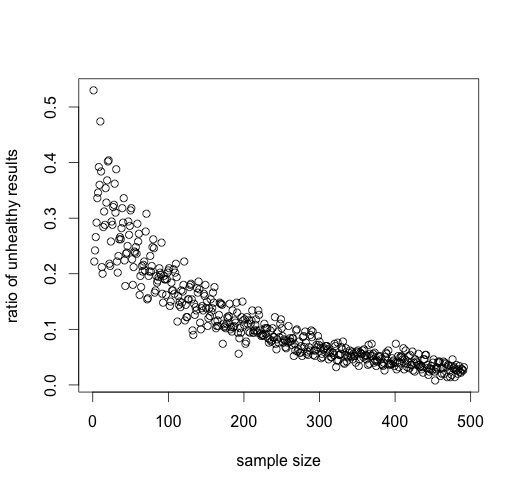
*for (i in 1:500){*

*collection <- c(collection, evaluatefunc(samplesize=i, rivernumber=500) )*

*}*

Then we get the distribution of the “impaired” ratio and we can see close to half of them are above 0.1:  


And from following plot we can clearly see as sample size increases the unhealthy ratio drops, and only after it gets past approximately 200 the ratio gets lower than 10%:



Clearly the rule does not work as EPA expects and in reality **the unhealthy ratio should be stable and not be affected by sample size.**

Appendix:

#import data

dat <- read.csv("AfrPlots.csv", header = T)

attach(dat)

#problem 1

qnorm(p=0.9, mean=4, sd=1.4)

pnorm(q=5, mean=4, sd=1.4)

set.seed(1001)

drawl <- rnorm(n=12, mean=4, sd=1.4)

#problem 2

set.seed(1001)

h2o <- as.data.frame(matrix(rnorm(n=10\*12, mean=4, sd=1.4), ncol=12))

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))

length(h2o$Test[h2o$Test>1])

#problem 3

set.seed(1001)

h2o <- as.data.frame(matrix(rnorm(n=10\*12, mean=4.5, sd=1.4), ncol=12))

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))

length(h2o$Test[h2o$Test>1])

#problem 4

set.seed(1001)

h2o <- as.data.frame(matrix(rnorm(n=10\*36, mean=4, sd=1.4), ncol=36))

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))

#cut-off is 10%\*36 which means 4

length(h2o$Test[h2o$Test>=4])

#problem 5

set.seed(1001)

#create a funtion

evaluatefunc <- function(samplesize, rivernumber){

h2odat <- as.data.frame(matrix(rnorm(n=samplesize\*rivernumber, mean=4, sd=1.4), ncol=samplesize))

rownames(h2odat) <- paste(rep("Riv", nrow(h2odat)), c(1:nrow(h2odat)), sep = "")

colnames(h2odat) <- paste(rep("Obs", ncol(h2odat)), c(1:ncol(h2odat)), sep = "")

h2odat$Test <- rowSums(ifelse(h2odat>6, 1, 0))

cutoff = ceiling(samplesize \* 0.1)

impaired = length(h2odat$Test[h2odat$Test>=cutoff])

#give the ratio of "the impaired"

impaired/rivernumber

}

#test different sample size from 10,50,100 and 500 rivers

sapply(c(10,50,100,500), evaluatefunc, samplesize=10)

sapply(c(10,50,100,500), evaluatefunc, samplesize=50)

sapply(c(10,50,100,500), evaluatefunc, samplesize=100)

sapply(c(10,50,100,500), evaluatefunc, samplesize=500)

#collection is the collection of the unhealthy ratio in 500 rivers

collection <- c()

for (i in 1:500){

collection <- c(collection, evaluatefunc(samplesize=i, rivernumber=500))

}

#percentage of ratio below 10%

sum(collection <= 0.1)/500

hist(collection, xlab="ratio")

plot(collection, ylab="ratio of unhealthy results", xlab="sample size")