Lab 3: Sampling

ENVIRON 710: Applied Statistical Modeling*

The goal of this lab is to learn about sampling. The first part of the lab provides more useful functions in R that should increase your ability to manipulate data. The second part of the lab aims to increase your understanding of sampling through a simulation of water quality standards.

For this lab, write a short discussion on the effectiveness and unintended consequence(s) of the EPA's water quality rule based on the simulation results. Your work will be graded on the clarity of the writing and the accuracy of the interpretation of the simulation results. In the write-up, you should focus on what you did and why, not how you did it. Use the following format:

- *Introduction*: describe the objective of the study
- *Methods*: describe the simulation procedure. See Smith et al (2001) for more discussion.
- *Discussion and Conclusion*: discuss your findings. Propose recommendations to the EPA concerning its impairment regulation and water quality sampling procedures.

Write your report in R Markdown, including your answers and your R-code, and submit it to the class Sakai site under the folder Assignments. See below for more details on the lab write-up.

More functions in R

In this lab, we introduce a few new R commands.
- `sort()` - sorts a vector into ascending or descending order
- `order()` - orders more than one variable, best for sorting data frames
- `qnorm()` - returns the quantile for a given probability from a normal distribution
- `paste()`- concatenates (puts together) vectors after they have been converted to characters
- `rep()` - repeats the values in a vector, whether they are individual values or a vector of values
- `rownames()` - sets row names for a data frame or matrix
- `colnames()`- sets column names for a data frame or matrix
- `matrix()` - defines an object as a matrix
- `data.frame()` - defines an object as a data frame
- `ncol()` - returns the number of columns in a matrix or data frame
- `nrow()` - returns the number of rows in a matrix or data frame
- `rowSums()` - returns column sums across rows
- `ifelse{}` - returns the appropriate given value depending on whether a logical test is TRUE or FALSE

Let's practice functions for indexing data once it is on your workspace. Download the `AfrPlots.csv` file from Sakai. Remember that this database consists of tree plot data from 30 1-ha tree plots in Africa. There were two census periods, the initial census and then a census 4 years later, in which all the trees were measured.

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

Each of the below lines should give you the `MeanGr` data for the second tree census. The first two lines select the data for the 2nd census.

```
MeanGr <- dat$MeanGr[dat$CensusNo == 2]
 dat$MeanGr[31:60]
```

*Created by John Poulsen with edits from TAs.

The next two lines bank on the fact that there could not be growth during the first tree census, and remove all the NA's from the data leaving just the values of tree growth for the second census.

```
MeanGr[!is.na(MeanGr)]
 MeanGr[is.na(MeanGr)==F]
```

To index the dataframe on multiple criteria, we write slightly more complicated logical arguments. For example, let's extract the data for the plots with more than 400 trees in the second census.

```
dat[dat$Trees > 400 & dat$CensusNo == 2,]
```

To Do

Extract the data for the plots where basal area `BasalArea` is greater than 30 cm, the number of trees `Trees` is greater than the mean number of trees for the database, and the trees are found in the second census.

To sort a single vector or dataframe column, use the `sort()` function. What does the `decreasing` argument do?

```
sort(dat$ChaveMoist, decreasing = T)
sort(dat$ChaveMoist, decreasing = F)
```

In R you can also create your own data frame by assigning vectors of data to variables within the `data.frame()` function, such as:

```
new_dat <- data.frame(a = factor(c("Hi", "Med", "Low", "Hi")),
                      x = c("C","B","C","A"),
                      y = c(24, 15, 22, 4))
```

Note that the length of each vector must be the same within `data.frame()`.

Rows or columns can be named using `rownames()` and `colnames()`, such as:

```
colnames(new_dat) <- c("A", "X", "Y")
```

We can sort dataframes as well by single or multiple variables, using `order()`. What do each of these lines of code do? Note the use of `decreasing` and the minus sign in the 2nd and 5th lines of code.

```
new_dat[order(new_dat$X), ]
 new_dat[order(new_dat$X, decreasing = T), ]
  new_dat[order(new_dat$A, new_dat$Y), ]

dat[order(dat$ChaveMoist),]
 dat[order(-dat$CensusNo, dat$BasalArea),]
```

Remember from the last lab, that `set.seed()` is used to set the starting point of the pseudo random number generator in R. If the starting point is the same, then you will end up with the same set of random numbers from the same function.

Sampling with R – EPA Water Quality Standards

The objectives of this assignment are:

- To learn how to translate a government rule on water quality management into a statistical problem.

- To interpret statistical results and to communicate these results clearly in the context of an environmental management problem.

The Clean Water Act requires that states regularly report the health of water bodies and submit a list of waters that do not meet established water quality standards. The United States Environmental Protection Agency (EPA) is responsible for developing rules for water quality assessment. 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. Many water quality experts believe that this approach to impairment designation is flawed. To learn why the rule may be flawed, we will conduct simulations to determine what would happen if this rule were used in practice. In this exercise we will examine a generic water quality parameter. Many water quality pollutants have a lognormal distribution: the logarithm of the pollutant concentration variable is approximately normal. As a result, we will use the logarithm of a pollutant concentration variable and assume a normal distribution of the log-transformed variable. For the sake of this assignment, we will assume that the natural log of the water quality standard is 6.

A simulation is a statistical tool for evaluating the behavior of a random variable. Because water quality measurements are random, results from any sampling study of the water quality of a lake or river are subject to sampling error. By using simulation, we can see how often the EPA's rule will make mistakes, that is, declare a water body to be impaired when it is not, and vice versa. 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 to be impaired. It is impossible to do so in practice, but if we know the distribution of the water quality pollutant, we can let the computer simulate the sampling process. We will simulate taking a water sample and measuring its concentration by drawing a random observation from the known distribution.

Work through the following five problems. *Only turn in Problem 5 as the lab write-up.*

Problem 1

The water quality standard (in natural log) is 6, and we know that the true distribution of the log concentration is $N(\mu = 4, \sigma = 1.4)$. This distribution has a 90th percentile value of 5.8, which means that if we repeatedly sample from this distribution, 90% of the time we will have values below 5.8 (in compliance with the water quality standard of 6).

How do we know the value of the 90th percentile for this distribution? We use `qnorm` to find the quantile (log concentration value) for a given probability of a normal distribution with a mean of 4 and a standard deviation of 1.4. This line of code calculates the quantile of a given probability for a normal distribution with our parameters.

```
qnorm(p = 0.90, mean = 4, sd = 1.4)
```

If the water quality standard was lowered to 5, what proportion of samples would have water quality values lower than this standard? This is calculated by finding the cumulative probability from $-\infty$ to 5 given $X \sim N(4, 1.4)$, which tells us the amount of probability under the curve to 5.

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

If we lowered the standard to 5, then 76.2% of the samples would be 5 or lower.

Now let's sample from this distribution, using `set.seed()` to get reproducible results. Pull 12 random values from our distribution of pollutant concentration.

```
set.seed(1001)
 draw1 <- rnorm(12, mean = 4, sd = 1.4)
  sort(draw1)
```

Based on EPA's rule, if two or more measurements (10% of 12 = 1.2) exceed 6, the water is impaired. Is the water impaired based on this rule?

To Do

Write a logical statement in R that tells you how many of the simulated water samples exceed the cut-off of 6.

Problem 2

Let's imagine that there are 10 rivers or lakes that have the same distribution of pollutants. Below we simulate taking 12 observations from each water body (i.e., `rnorm(10*12, mean = 4, sd = 1.4)`. The `h2o` line of code creates a 10 x 12 matrix and fills it with simulated pollutant values. It also converts the matrix to a dataframe using `as.data.frame()` so that we can add non-numeric columns to it later on. To make the matrix easy to understand, we change the row and column names so that they are meaningful. The result of the simulation indicates how many rivers would be wrongly declared to be impaired.

```
set.seed(1001)

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

## Create row labels, using paste and nrow
 paste(rep("Riv", nrow(h2o)), c(1:nrow(h2o)), sep = "")
##  [1] "Riv1"  "Riv2"  "Riv3"  "Riv4"  "Riv5"  "Riv6"  "Riv7"  "Riv8"  "Riv9"
## [10] "Riv10"

## Now assign it to the matrix using rownames
 rownames(h2o) <- paste(rep("Riv", nrow(h2o)), c(1:nrow(h2o)), sep = "")

 colnames(h2o) <- paste(rep("Obs", ncol(h2o)), c(1:ncol(h2o)), sep = "")
```

Based on EPA's rule, if greater than 10% of observations exceed 6, the water is considered impaired. With a sample size of 12, a water body is impaired when the number of violations is 2 or more. How many rivers would be wrongly declared as "impaired''?

There are many approaches to automatically count up the number of observations that exceed 6 for each river (row). The line of code below changes all values of h2o to 1 or 0, if they are greater than or less than 6, respectively: `ifelse(h2o>6, 1, 0)`. Then `rowSums()` adds up all the 1's across each row of the data frame, indicating the number of observations that surpassed the water quality standard. (This could be broken into two steps, if you find that embedding it in one line of code is difficult to understand.)

```
h2o$Test <- rowSums(ifelse(h2o>6, 1, 0))
```

We can then figure out how many rivers have more than 1 sample that violates the standard by counting the number of rivers with more than 1 impairment.

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

In this example, how many impaired rivers are there?

Problem 3

Now suppose the distribution of the pollutant variable is $N(\mu = 4.5, \sigma = 1.4)$. If we repeatedly sample from this population distribution, we expect to violate the standard, 6, more than 10% of the time. What proportion of observations would violate the standard from this new distribution?

Use `pnorm` to determine the cumulative probability of getting a pollutant value less than or equal to 6 under this new distribution, $N(\mu = 4.5, \sigma = 1.4)$.

```
pnorm(6, mean = 4.5, sd = 1.4)
```

In a population with a mean pollutant concentration of 4.5, 85.8% of values are less than or equal to 6. Therefore, we expect approximately 14.2% of observations to be greater than 6.

Assuming we are still using a sample size of 12, what proportion of water samples would be in compliance under the EPA's rule? Use `set.seed(1001)` whenever you use the `rnorm` function here and throughout the lab.

```
    set.seed(1001)
    sort(rnorm(12, mean = 4.5, sd = 1.4))
```

In this example, samples are below the standard 75% of the time (9 out of 12).

Let's make a graph of the two distributions to picture the change from a mean of 4 to 4.5.

```
ggplot(data = data.frame(x = c(-2, 10)), aes(x)) +
       stat_function(fun = dnorm, args = list(mean = 4, sd = 1.4), col = "darkviolet") +
       stat_function(fun = dnorm, args = list(mean = 4.5, sd = 1.4), col = "darkorange") +
        geom_segment(x=4, y=dnorm(x=4,mean=4,sd=1.4),xend=4,yend=0, col = "darkviolet",
                     linetype = "dashed") +
        geom_segment(x=4.5, y=dnorm(x=4.5,mean=4.5,sd=1.4),xend=4.5,yend=0,
                     col = "darkorange", linetype = "dashed") +
        ylab("Probability density") + xlab("X") + theme_bw()
```

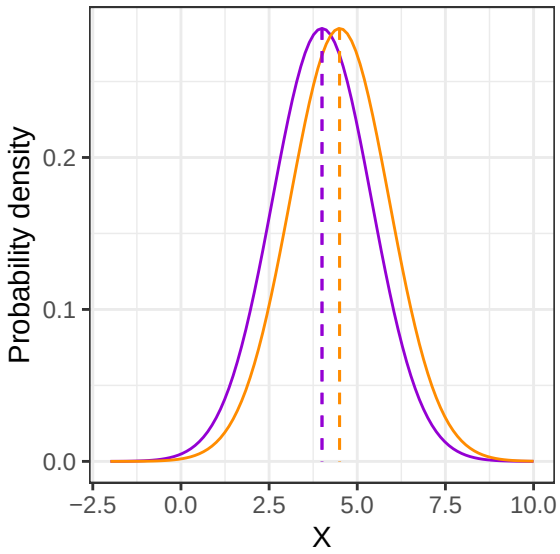

Figure 1: Two probability distributions for the concentration of water pollutants.

Problem 4

Using the original distribution of $N(\mu = 4, \sigma = 1.4)$, what happens if the sample size is 36 (3 years of observations)? What is the cut-off for the number of samples that can exceed the standard of 6 (e.g., 10% of 36)?

```
set.seed(1001)
 nobs <- 36
  too_many <- round(0.10 * nobs, digits = 0)
```

```
h2o <- as.data.frame(matrix(rnorm(10*nobs, mean=4, sd=1.4),
       ncol= nobs))
 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$Test>too_many])
  length(h2o$Test[h2o$Test>too_many])/10
```

The cut-off for a sample of 36 is calculated by `too_many`. In this example, 2 out of 10 rivers violate the standard. For 20% of the rivers, more than 10% of the 36 samples are polluted.

Problem 5 - Lab Write-up

For the lab write-up, evaluate the number of impaired rivers using a true distribution of pollutant concentration of $N(\mu = 4, \sigma = 1.4)$ and a water standard of 6. Simulate the sampling of 10, 50, 100, and 500 independent and identically distributed rivers with 10, 50, 100, and 500 observations per river. What can you conclude about the EPA water standards? What recommendations would you make to the EPA? (Again, use `set.seed(1001)`.)

Try to implement the simulations within a function that can be called with different numbers of rivers and observations.

The lab will be scored on the basis of 40 points. It should be 2 pages long (with at least 1 figure and the R code as an Appendix) and include the following elements:

- Introduction (5 points). Write a brief Introduction to your report that articulates the problem – appropriately sampling water bodies – and its importance. End the Introduction with the objectives of your report or questions to which you are going to provide answers. (Smith et al. 2001 is on Sakai under Lab 3, and it might be of help.) (1-2 paragraphs)
- Methods (10 points). Clearly explain the process of your analysis - what did you do from the simulation of water samples to figuring out whether the EPA's standards are appropriate or not.(1-2 paragraphs)
- Results and Conclusion (15 points). Discuss your results and any patterns that you detected in the relationship between numbers of observations or rivers sampled and the proportion of impaired water bodies. Include a graph(s) that clearly presents your results. The graph(s) should consist of the '% of Impaired Rivers' on the y-axis and the 'Number of Observations' on the x-axis to demonstrate potential changes in results with sampling for a given number of rivers. The graph could be replicated for different numbers of rivers. Then, discuss what your results mean for decisions being taken about whether a water body is impaired or not. What are the strengths or weaknesses of the EPA standards? How would you suggest that the EPA sample rivers in the future? (3-4 paragraphs)
- R-code (10 points). The R-code should be provided as an Appendix and should work. The Appendix is not counted in the paper length. Remember, the R-code can be sent to an appendix by including the following in the header of a new R-code chunk at the end of your R Markdown document: `{r, ref.label=knitr::all_labels(),echo=TRUE,eval=FALSE}`.

All references should be cited using the format of the reference below.

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.