Abalones Exploratory Data Analytics Project

Sunny Wu
Predict 401 Section 57
Title: Data Analytics Project Assignment #1

Introduction

Abalones are an economic and recreational resource that is threatened by a variety of factors which include: pollution, disease, loss of habitat, predation, commercial harvesting, sport fishing and illegal harvesting. Environmental variation and the availability of nutrients affect the growth and maturation rate of abalones. Over the last 20+ years it is estimated the commercial catch of abalone worldwide has declined in the neighborhood of 40%. Abalones are easily over harvested because of slow growth rates and variable reproductive success. Being able to quickly determine the age composition of a regional abalone population would be an important capability.

The data are derived from an observational study of abalones. The intent of the investigators was to predict the age of abalone from physical measurements thus avoiding the necessity of counting growth rings for aging. Ideally, a growth ring is produced each year of age. Currently, age is determined by drilling the shell and counting the number of shell rings using a microscope. This is a difficult and time consuming process. Ring clarity can be an issue. At the completion of the breeding season sexing abalone can be difficult. Similar difficulties are experienced when trying to determine the sex of immature abalone. The study was not successful. The investigators concluded additional information would be required such as
weather patterns and location which affect food availability.

Results

R programming language was used to generate meaningful analysis for abalones' living conditions. The observational study of abalones database is used in this study. Additional parameters of VOLUME and RATIO (ratio of shuck weight and total weight) are also calculated based on collected data. The first 6 rows of the studied dataset are presented below in Table 1. In our study, the factors including abalone sex, length, diameter, height, whole weight, shuck weight, rings, class, volume, and ratio are considered.

Table 1. Dataset of studying abalones (First six rows)

SEX	LENGTH	DIAM	HEIGHT	WHOLE	SHUCK	RINGS	CLASS	VOLUME	RATIO
I	5.565	4.095	1.26	11.5	4.3125	6	A1	28.713731	0.1501895
I	3.675	2.625	0.84	3.5	1.1875	4	A1	8.103375	0.1465439
I	10.08	7.35	2.205	79.375	44	6	A1	163.36404	0.2693371
I	4.095	3.15	0.945	4.6875	2.25	3	A1	12.189791	0.1845807
I	6.93	4.83	1.785	21.1875	9.875	6	A1	59.747341	0.1652793
I	7.875	6.09	2.1	27.375	11.5625	6	A1	100.713375	0.114806

A summary of the dataset is presented in Table 2. In total, 1036 groups of 8 variables are collected in the dataset. The statistical analysis in Table 2 presents abalone mean length of 11.08 cm, diameter of 8.62 cm, height of 2.95 cm, and total weight of 105.8 gram. The average volume of abalones is 326,804 cm^3. Average ratio of shucked weight of meat versus whole weight is

0.14. The average length of abalone might suggest the minimum legal length of fishing abalones. The current standard of minimum legal length of catching abalone is 11.5cm. Since the mean length of abalone is 11.08 cm, the current standard has been well established to protect abalones.

Table 2. Summary of Abalone Dataset

SEX	LENGTH	DIAM	HEIGHT	WHOLE
F:326	Min. : 2.73	Min. : 1.995	Min. :0.525	Min. : 1.625
I:329	1st Qu.: 9.45	1st Qu.: 7.350	1st Qu.:2.415	1st Qu.: 56.484
M:381	Median :11.45	Median : 8.925	Median :2.940	Median :101.344
	Mean :11.08	Mean : 8.622	Mean :2.947	Mean :105.832
	3rd Qu.:13.02	3rd Qu.:10.185	3rd Qu.:3.570	3rd Qu.:150.319
	Max. :16.80	Max. :13.230	Max. :4.935	Max. :315.750

SHUCK	RINGS	CLASS	VOLUME	RATIO
Min. : 0.5625	Min. : 3.000	A1:108	Min. : 3.612	Min. :0.06734
1st Qu.: 23.3006	1st Qu.: 8.000	A2:236	1st Qu.:163.545	1st Qu.:0.12241
Median : 42.5700	Median : 9.000	A3:330	Median :307.363	Median :0.13914
Mean : 45.4396	Mean : 9.984	A4:188	Mean :326.804	Mean :0.14205
3rd Qu.: 64.2897	3rd Qu.:11.000	A5: 83	3rd Qu.:463.264	3rd Qu.:0.15911
Max. :157.0800	Max. :25.000	A6: 91	Max. :995.673	Max. :0.31176

The length distribution of abalones is analyzed in Figure 2 to understand whether the environmental regulation has protected abalones well or not. The skewness analysis of length returns a negative skew of -0.67, indicating the mass of the distribution is concentrated on the right of the figure. If the environmental regulation of minimum length is set to be too low, then the length histogram should be have a positive skew, therefore, the regulation of minimum legal length of catching abalones are well established. Kurtosis analysis of length data returns 3.2, indicating a leptokurtic distribution, which has flatter tails.

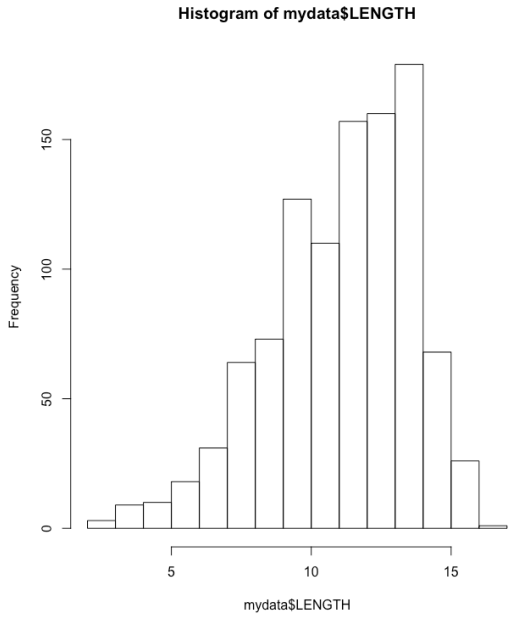

Figure 1. Histogram of abalone length

This research has been further applied to understand the correlation of sex frequency distribution and age class (Table 3 and Figure 2). Sex are classified into 3 groups: infant, male, and female. It could be seen that it is a vague concept when abalone turns adult. In age class A1, the youngest age class, the number of infant is highest, reaching up to 91. However, there is also a sex distribution of 5 female and 12 male in age class A1. In comparison, at the oldest age A6, we see there are still 6 fish classified as infants. Therefore, in the data collection stage, some mistakes were made while identifying sex of fish. It is easy to understand in A1 stage, while fish are still very young, it is harder to identify sex. My suggestion is that an age limit could be established, all fish at age class of A1 and A2 should be considered as infants, and all fish at age class of A3 above should be considered male and female sex.

Table 3. Comparison of Abalone Sex Frequencies

	A1	A2	A3	A4	A5	A6	Sum
F	5	41	121	82	36	41	326
I	91	133	66	21	10	8	329
M	12	62	143	85	37	42	381
Sum	108	236	330	188	83	91	1036

Figure 2 also shows the change of sex distribution between male and female. In age class A2 and A3, more male fish are observed than female. In age group A4, A5, and A6, the number of male and female fish are identical. A possible reason that this observation might be biased is that female abalones are generally smaller than male, therefore, in age group A2 and A3, some female abalone might be classified as infants.

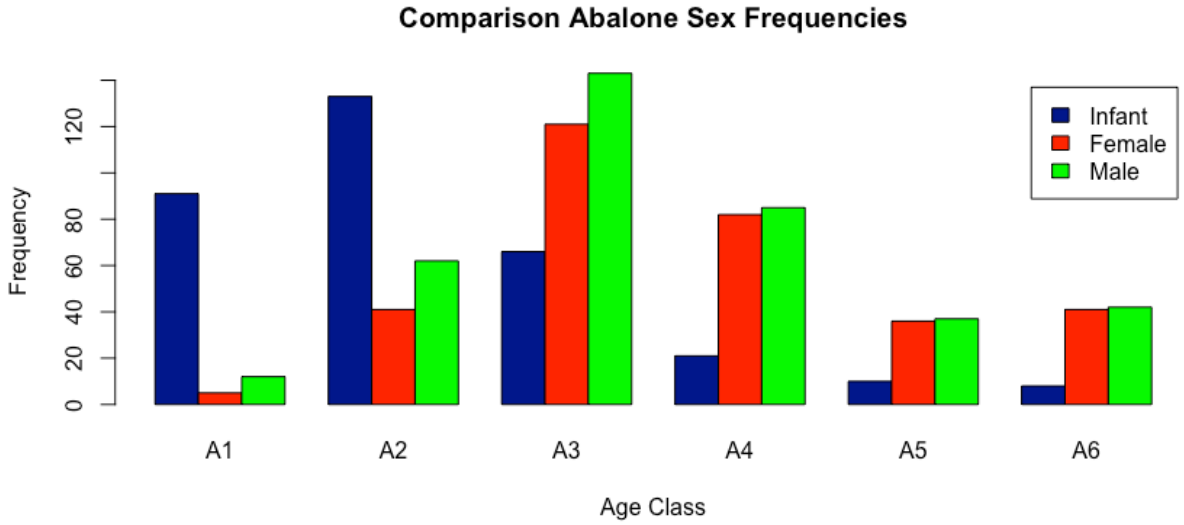

Figure 2. Barplot of Comparison of Abalone Sex Frequencies

The scatterplot of physical measurement including length, diameter, height, whole weight, and shuck weight are plotted in Figure 3. A strong positive linear relationship among length,

diameter, and height are observed. Also, a strong positive linear relationship between whole weight and shuck weight are observed. There is a strong positive curvilinear relationship observed between spatial and weight variables.

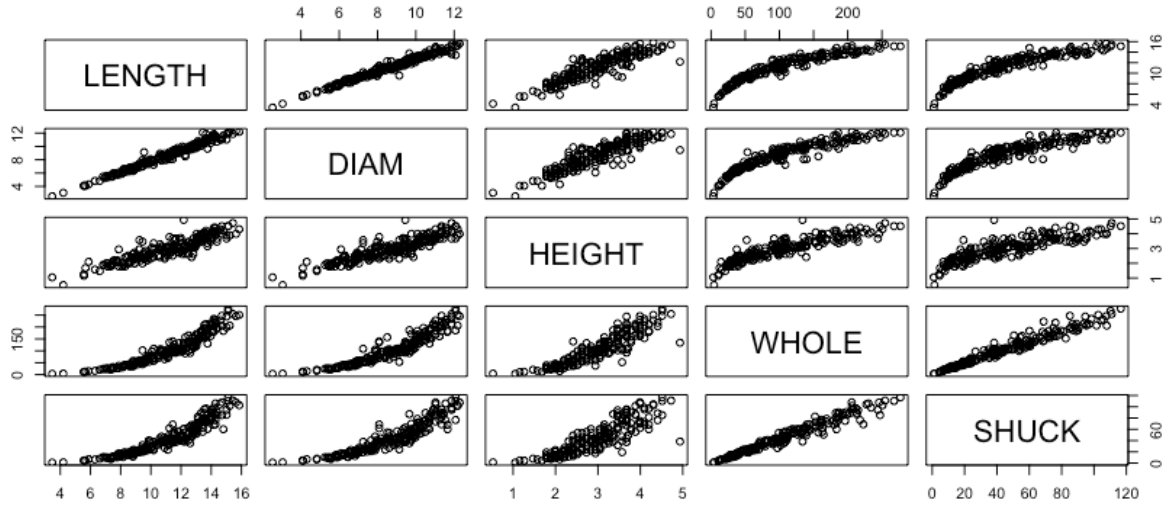

Figure 3. Crossplot of Length, diameter, height, whole, and shuck weight of abalones

Weight increases linearly with volume, rather than any single dimensions such as length, diameter, or height (Figure 4). Also, shuck weight increases with the whole weight with a strong linear relationship, we could observe that when whole weight increases, there is proportionally less shuck weight increase (Figure 5). The difference is that the correlation between whole weight increase with volume is always linear, while the shuck weight increases less with the whole weight increase when the whole weight value gets larger.

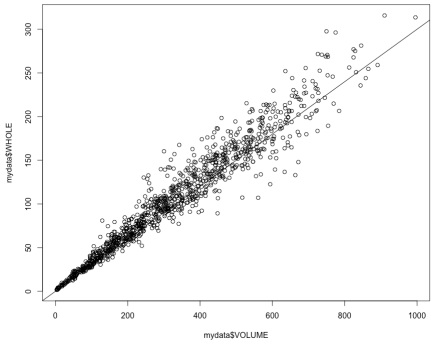

Figure 4. Scatterplot of volume and whole weight

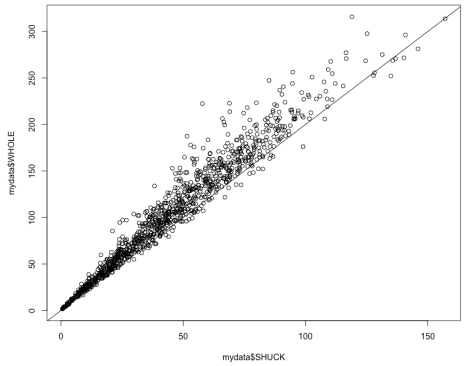

Figure 5. Scatterplot of shuck weight and whole weight

In Figure 6, the histograms, boxplots, and Q-Q plots of VOLUME and WHOLE are displayed. There is greater skewness observed in infant data. In order to further explore our data, regression analysis is recommended. The definition of infant sex means that the fish are still small size and it is harder to identify the sex. Therefore, the skewness is understandable. Comparison of male and female data shows very similar distribution.

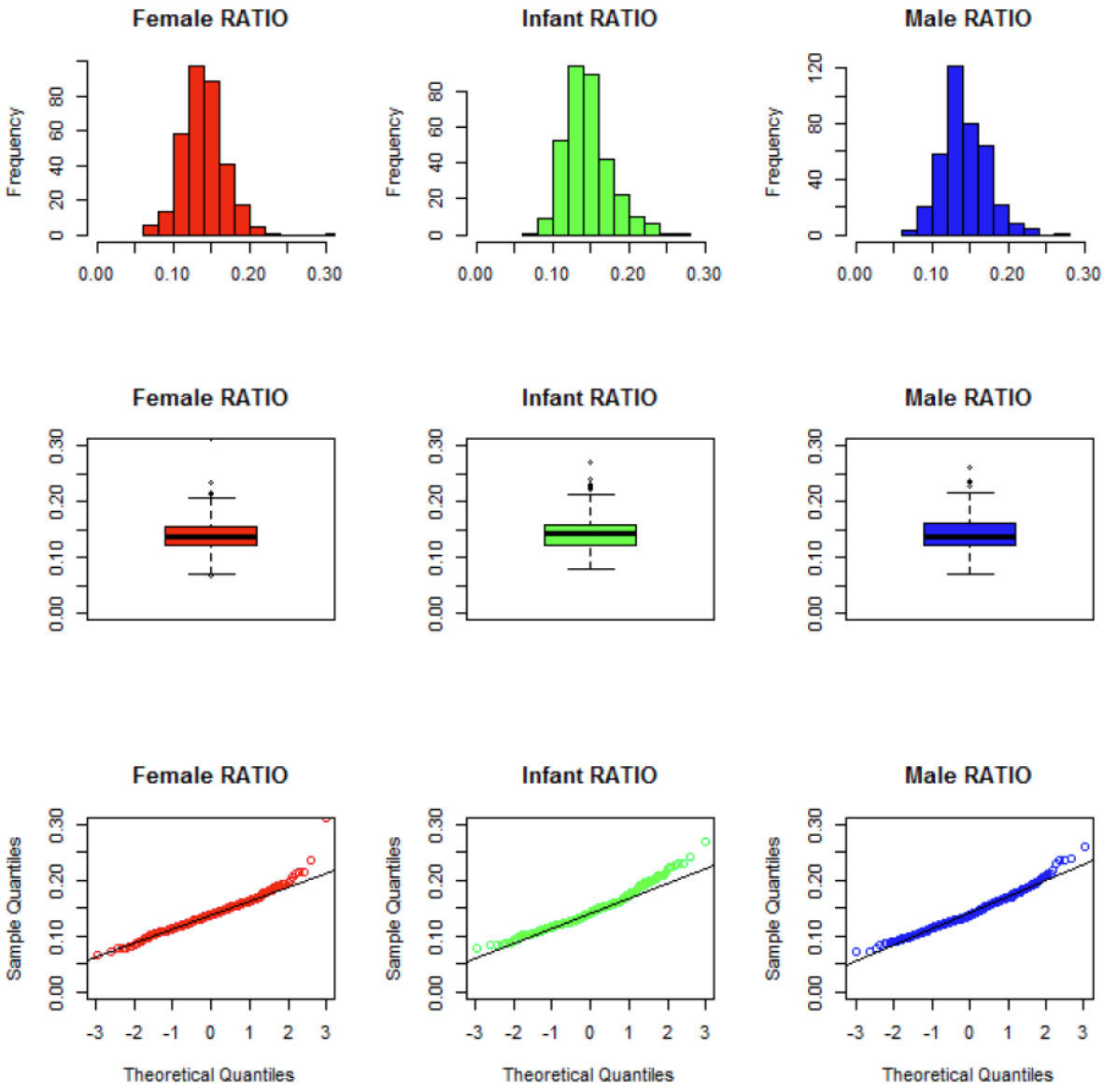

Figure 6. Histogram, boxplots and Q-Q plots of RATIO differentiated by SEX

The boxplots of volume and whole weight in Figure 7 show that the volume and whole weight has similar trend with the age class. In age class A1 and A2, fish are still infants, and the volume and whole weight are smaller than average. In age class A3, A4, A5, and A6, fish are already adults, and therefore, the volume and whole weight maintain steady, except a slight fall in age group A5. The plot of volume and whole weight as a function of RINGS indicates that are no correlations between rings and volume or rings and weight. I would suggest environmental factor data might be collected to facilitate further analysis.

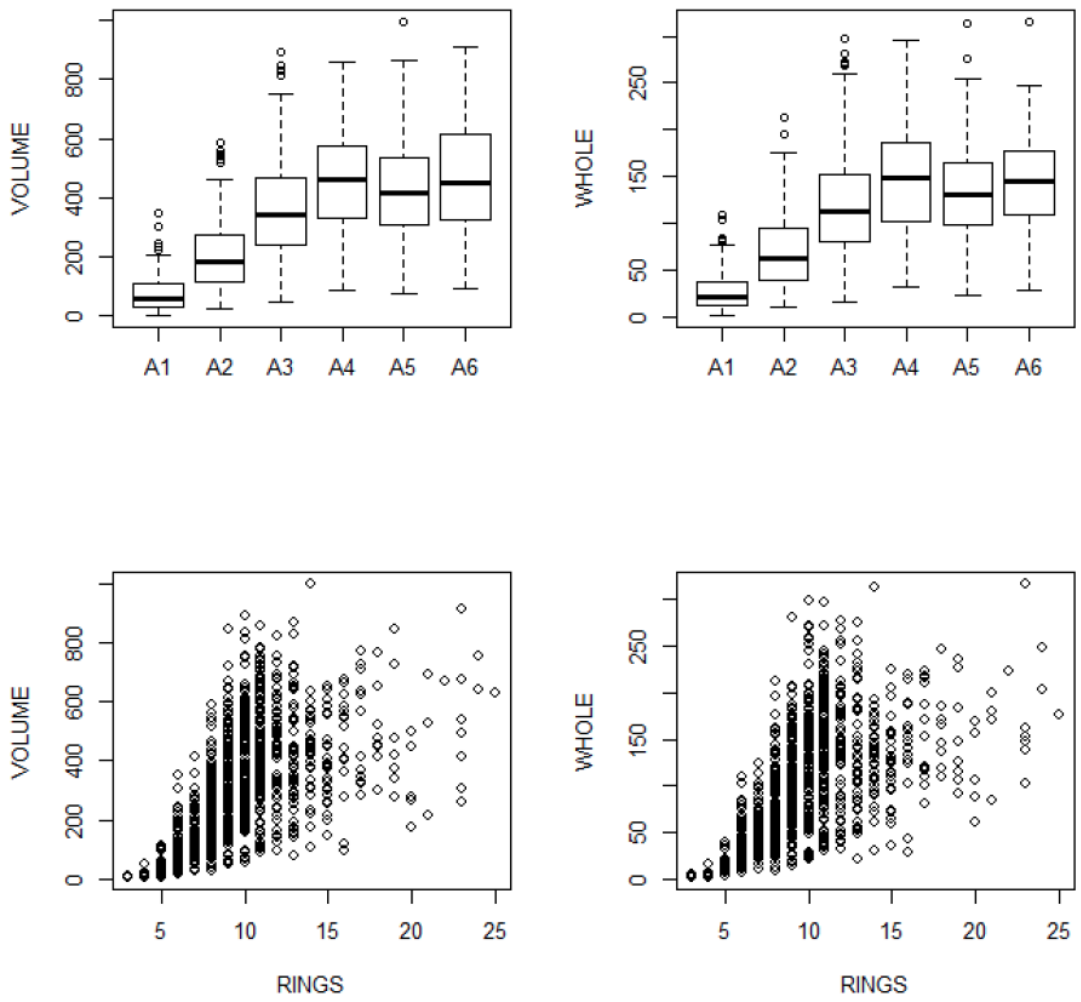

Figure 7. Boxplot of Volume and WHOLE differentiated by class, scatterplot of VOLUME and WHOLE versus RINGS

In Figure 8, mean VOLUME are plotted per CLASS, per SEX. The infant sex has the smallest volume, followed by male, and female has the largest volume. This corresponds to the biological information that female abalones have a smaller size than male. Change of mean volume gradient is high in CLASS A1, and A2, and mean volume stays relatively steady in class A3, A4,

A5, and A6 for male and female abalones. The volume of infant in classes A1 and A2 are more reliable.

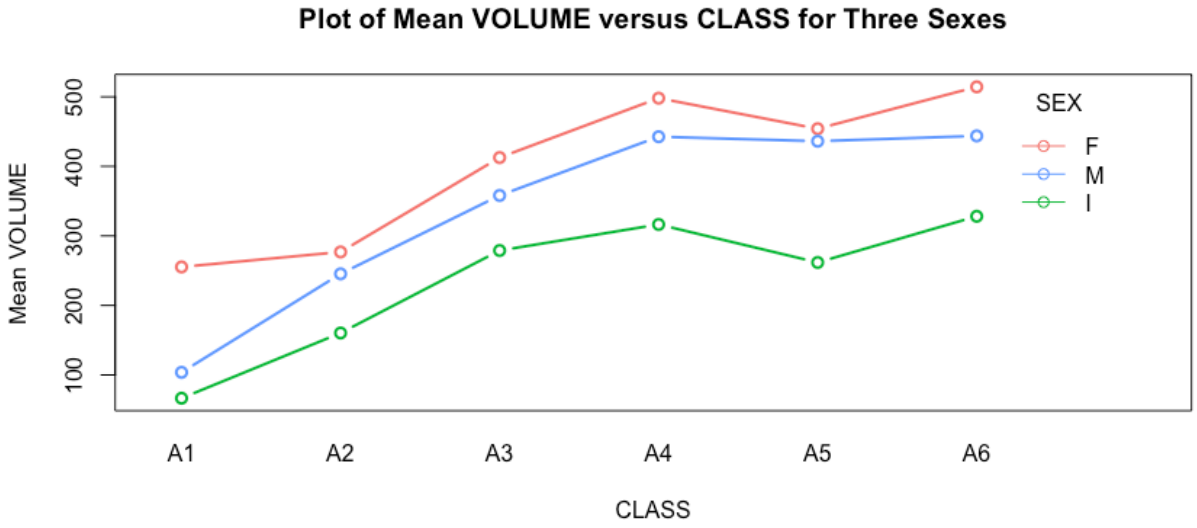

Figure 8. Plot of Mean VOLUME versus CLASS for Three Sexes

In Figure 9, mean RATIO are also plotted per CLASS, per SEX. The mean ratio of infant is significant lower than male and female, while the mean RATIO of Male and female overlaps each other. This indicates that infants' weight are mostly composed of bones, carrying less meat. Again, the dataset that includes infant at age class above A3 should be carefully analyzed because those abalones are no longer infants.

Figure 9. Plot of Mean RATIO versus CLASS for Three Sexes

Conclusion

The physical measurement of abalones should be used for predicting age of abalones. Our study shows that the average length of abalones is 11.1cm, which could be a cut-off line for identifying infant abalone and adult abalones. Bringing that concept into our dataset will significantly improve our data analysis result.

If I am presented with an overall histogram and summary statistics from a sample and no other information, I would ask lots of questions before accepting them as representatives of that population:

1. What is the sample size and the population size?
2. Are there selection bias in my samples? Are there coverage bias? Is the sampling frame established?
3. What sampling technique is used?
4. How well the data collection was done? Are there quality controls during the data collection process? Is the classification clearly defined? In our dataset, the classification of sex: infant, male, and female are poorly defined.
5. What is the source of the sampling?

What do you see as important difficulties with observation studies in general?

Observational studies could be quite problematic. General difficulties include: it could be subjective, time-consuming, and dependent on the role of researcher. In this study, we find the problems with classification of sex, and also the age of abalones. Although we do not consider our analysis as successful, we have got some meaniful results from our analysis. We found that the male and female abalones have similar ratio, and follow the same trend line of volume change with age.

However, the use of observational studies are still quite important since it collects first hand data, and research often needs to be conducted on data with limitations.

Appendix

```
mydata <- read.csv("/Users/sunnywu/Desktop/DataAnalysisAssignment1/abalones.csv", sep =
"")
str(mydata)
#Define VOLUME and RATIO variables
mydata$VOLUME <- mydata$LENGTH * mydata$DIAM * mydata$HEIGHT
mydata$RATIO <- mydata$SHUCK / mydata$VOLUME
summary(mydata)
head(mydata)

summary(mydata$RINGS)
summary(mydata$CLASS)
# ?table(), ?addmargins() to review documentation pages
#barplot()
barplot(table(mydata$SEX,mydata$CLASS)[c(2,1,3),],
      legend.text=c("Infant","Female","Male"),
      main = "Comparison Abalone Sex Frequencies", ylab = "Frequency",
```

```
      xlab = "Age Class", beside = TRUE, col = c("darkblue","red","green"),
      names.arg= c("A1","A2","A3","A4","A5","A6"))

set.seed(123)
work <- mydata[sample(1:nrow(mydata), 200, replace = FALSE),]
plot(work[, 2:6])

plot (x=1:10, y=1:10,type="p",main="Example Scatterplot Using abline()",xlab="X",ylab="Y")
abline(a=0, b=1)
help(abline)
#q5a
with(mydata,
    interaction.plot(CLASS, SEX, VOLUME,type="b",
col=c("#F8766D","#00BA38","#619CFF"),
                lty = 1, ylab = "Mean VOLUME", xlab = "CLASS", lwd = 2,
                trace.label = "SEX", pch = c(1,1,1),
                main = "Plot of Mean VOLUME versus CLASS for Three Sexes"))
out <-aggregate(VOLUME ~ SEX + CLASS, data = mydata, mean)

with(mydata,
    interaction.plot(CLASS, SEX, RATIO,type="b", col=c("#F8766D","#00BA38","#619CFF"),
                lty = 1, ylab = "Mean RATIO", xlab = "CLASS", lwd = 2,
                trace.label = "SEX", pch = c(1,1,1),
                main = "Plot of Mean RATIO versus CLASS for Three Sexes"))
out <-aggregate(RATIO ~ SEX + CLASS, data = mydata, mean)

#Q5b
ggplot(data=out, aes(x= CLASS, y= VOLUME, group = SEX, colour = SEX)) +
  geom_line() + geom_point(size=4) +
  ggtitle("Plot of Mean VOLUME versus CLASS for Three Sexes")

ggplot(data=out, aes(x= CLASS, y = RATIO, group = SEX, colour = SEX)) +
  geom_line() + geom_point(size=4) +
  ggtitle("Plot of Mean RATIO versus CLASS for Three Sexes")

data(mtcars)
par(mfrow=c(3,3))
grid.arrange(ggplot(mtcars,aes(x=factor(cyl),y=mpg, group=cyl)) + geom_boxplot() +
              labs(x="Numbr of Cylinders", y= "Miles per Gallon"),
            ggplot(mtcars,aes(x=factor(cyl),y=hp, group=cyl)) + geom_boxplot() +
              labs(x="Numbr of Cylinders", y= "Horsepower"),
            nrow=1,
            top = "Example Boxplots using ggplot() and grid.arrange()"
            )
skewness(mydata$LENGTH)
skewness(mydata$WHOLE)
```

```
par(mfrow=c(1,1))
hist(mydata$LENGTH)
kurtosis(mydata$LENGTH)
addmargins(table(mydata$SEX,mydata$CLASS))
skewness(mydata$SEX)
plot(mydata$VOLUME,mydata$WHOLE)
abline(a = 0, b = 3/10)

plot(mydata$SHUCK,mydata$WHOLE)
abline(a = 0, b = 30/15)

par(mfrow=c(3,3))
hist(mydata$RATIO, SEX = I)

summary(mydata$LENGTH)
```