Data Analysis #2 Version 2 (75 points total)

Gesheva, Mariana

Instructions

R markdown is a plain-text file format for integrating text and R code, and creating transparent, reproducible and interactive reports. An R markdown file (.Rmd) contains metadata, markdown and R code "chunks", and can be "knit" into numerous output types. Answer the test questions by adding R code to the fenced code areas below each item. There are questions that require a written answer that also need to be answered. Enter your comments in the space provided as shown below:

Answer: (Enter your answer here.)

Once completed, you will "knit" and submit the resulting .html document and the .Rmd file. The .html will present the output of your R code and your written answers, but your R code will not appear. Your R code will appear in the .Rmd file. The resulting .html document will be graded and a feedback report returned with comments. Points assigned to each item appear in the template.

Before proceeding, look to the top of the .Rmd for the (YAML) metadata block, where the *title*, *author* and *output* are given. Please change *author* to include your name, with the format 'lastName, firstName.'

If you encounter issues with knitting the .html, please send an email via Canvas to your TA.

Each code chunk is delineated by six (6) backticks; three (3) at the start and three (3) at the end. After the opening ticks, arguments are passed to the code chunk and in curly brackets. **Please do not add or remove backticks, or modify the arguments or values inside the curly brackets**. An example code chunk is included here:

```r
# Comments are included in each code chunk, simply as prompts

#...R code placed here

#...R code placed here
```

R code only needs to be added inside the code chunks for each assignment item. However, there are questions that follow many assignment items. Enter your answers in the space provided. An example showing how to use the template and respond to a question follows.

Example Problem with Solution:

Use *rbinom()* to generate two random samples of size 10,000 from the binomial distribution. For the first sample, use p = 0.45 and n = 10. For the second sample, use p = 0.55 and n = 10. Convert the sample frequencies to sample proportions and compute the mean number of successes for each sample. Present these statistics.

```r
set.seed(123)
sample.one <- table(rbinom(10000, 10, 0.45)) / 10000
sample.two <- table(rbinom(10000, 10, 0.55)) / 10000
```

```r
successes <- seq(0, 10)

round(sum(sample.one*successes), digits = 1) # [1] 4.5
```

```
## [1] 4.5
```

```r
round(sum(sample.two*successes), digits = 1) # [1] 5.5
```

```
## [1] 5.5
```

Question: How do the simulated expectations compare to calculated binomial expectations?

Answer: The calculated binomial expectations are 10(0.45) = 4.5 and 10(0.55) = 5.5. After rounding the simulated results, the same values are obtained.

Submit both the .Rmd and .html files for grading. You may remove the instructions and example problem above, but do not remove the YAML metadata block or the first, "setup" code chunk. Address the steps that appear below and answer all the questions. Be sure to address each question with code and comments as needed. You may use either base R functions or ggplot2 for the visualizations.

##Data Analysis #2

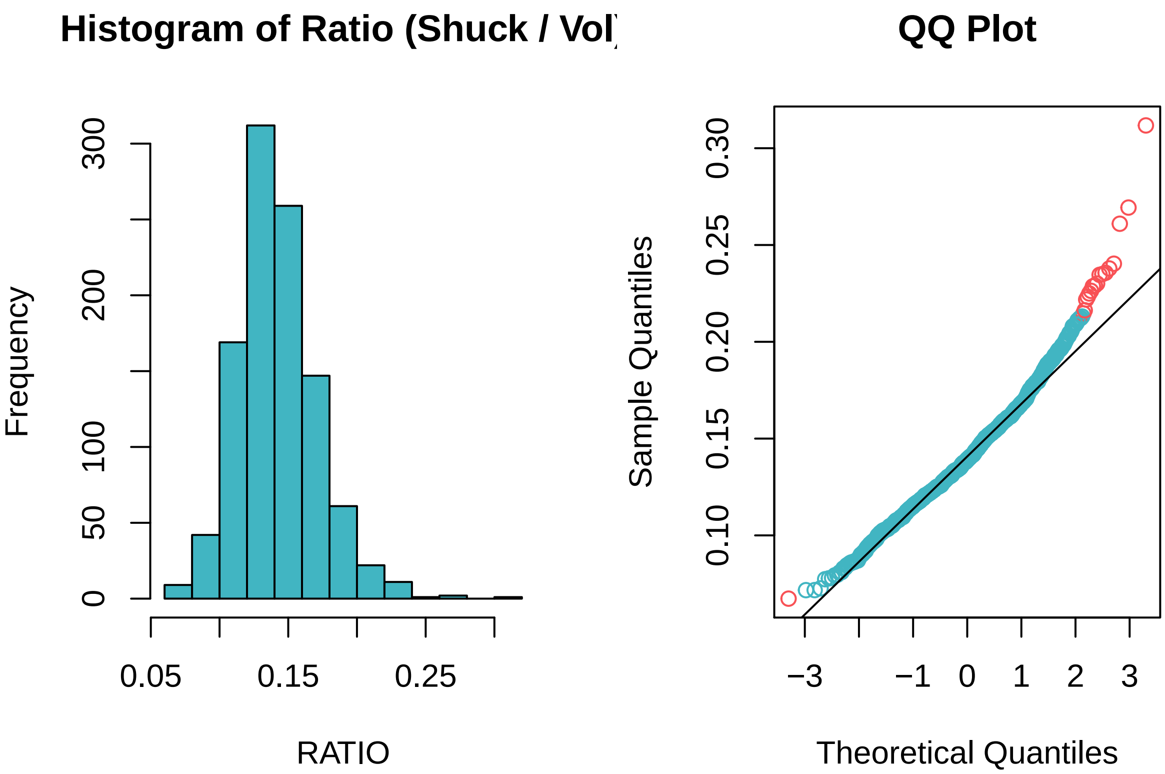
```
## 'data.frame':    1036 obs. of  10 variables:
##  $ SEX   : chr  "I" "I" "I" "I" ...
##  $ LENGTH: num  5.57 3.67 10.08 4.09 6.93 ...
##  $ DIAM  : num  4.09 2.62 7.35 3.15 4.83 ...
##  $ HEIGHT: num  1.26 0.84 2.205 0.945 1.785 ...
##  $ WHOLE : num  11.5 3.5 79.38 4.69 21.19 ...
##  $ SHUCK : num  4.31 1.19 44 2.25 9.88 ...
##  $ RINGS : int  6 4 6 3 6 6 5 6 5 6 ...
##  $ CLASS : chr  "A1" "A1" "A1" "A1" ...
##  $ VOLUME: num  28.7 8.1 163.4 12.2 59.7 ...
##  $ RATIO : num  0.15 0.147 0.269 0.185 0.165 ...
```

Test Items starts from here - There are 10 sections - total of 75 points

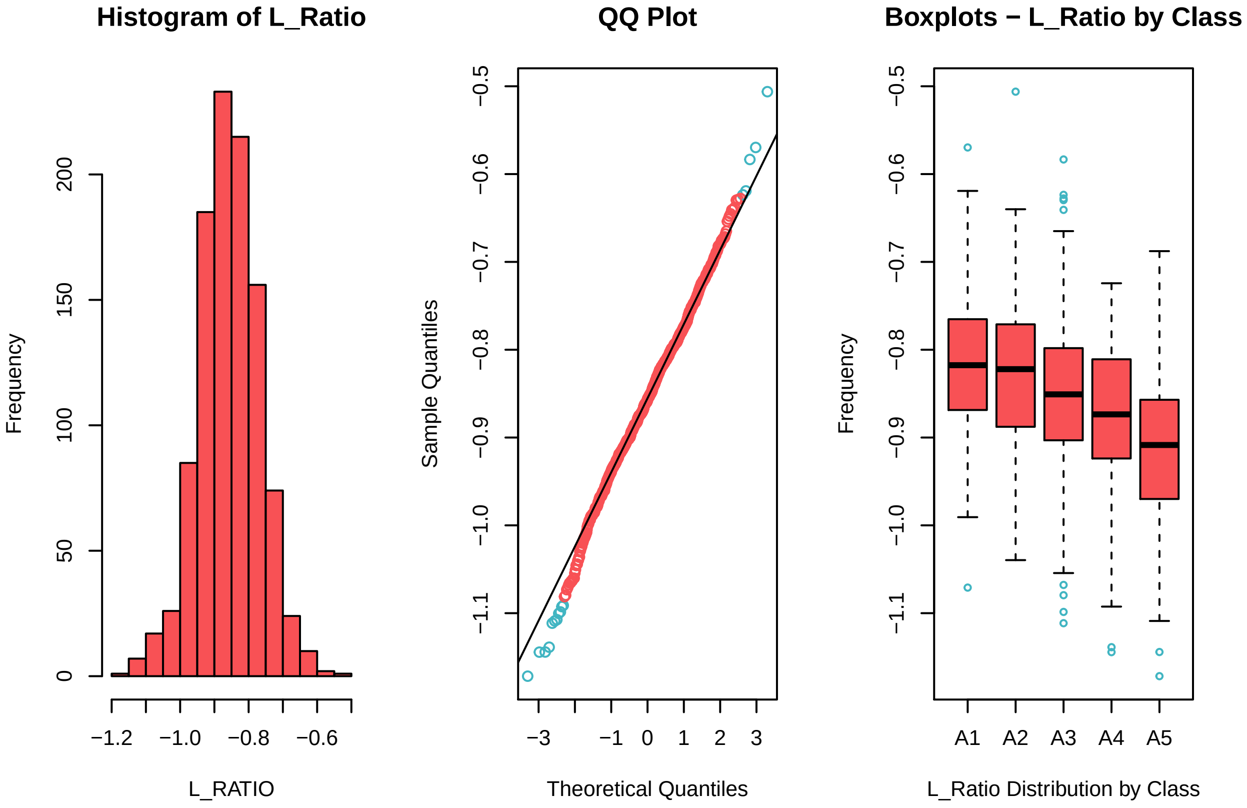
Section 1: (5 points)

(1)(a) Form a histogram and QQ plot using RATIO. Calculate skewness and kurtosis using 'rockchalk.' Be aware that with 'rockchalk', the kurtosis value has 3.0 subtracted from it which differs from the 'moments' package.

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```



(1)(b) Tranform RATIO using *log10()* to create L_RATIO (Kabacoff Section 8.5.2, p. 199-200). Form a histogram and QQ plot using L_RATIO. Calculate the skewness and kurtosis. Create a boxplot of L_RATIO differentiated by CLASS.



(1)(c) Test the homogeneity of variance across classes using *bartlett.test()* (Kabacoff Section 9.2.2, p. 222).

```
##
##  Bartlett test of homogeneity of variances
##
## data:  RATIO by CLASS
## Bartlett's K-squared = 21.49, df = 4, p-value = 0.0002531
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  L_RATIO by CLASS
## Bartlett's K-squared = 3.1891, df = 4, p-value = 0.5267
```

Essay Question: Based on steps 1.a, 1.b and 1.c, which variable RATIO or L_RATIO exhibits better conformance to a normal distribution with homogeneous variances across age classes? Why?

Answer: (The L_Ratio has better conformance to a normal distribution with homogeneous variances across age classes. We see there is less skew in the histogram as well as more evenly distributed outliers in the QQ plot and boxplots. The Bartlett test of homogeneity failed to reject the null hypothesis and this is why we assume that the classes are homogeneous.)

Section 2 (10 points)

(2)(a) Perform an analysis of variance with *aov()* on L_RATIO using CLASS and SEX as the independent variables (Kabacoff chapter 9, p. 212-229). Assume equal variances. Perform two analyses. First, fit a model with the interaction term CLASS:SEX. Then, fit a model without CLASS:SEX. Use *summary()* to obtain the analysis of variance tables (Kabacoff chapter 9, p. 227).

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## CLASS          4  1.055 0.26384  38.370 < 2e-16 ***
## SEX            2  0.091 0.04569   6.644 0.00136 **
## CLASS:SEX      8  0.027 0.00334   0.485 0.86709
## Residuals   1021  7.021 0.00688
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## CLASS          4  1.055 0.26384  38.524 < 2e-16 ***
## SEX            2  0.091 0.04569   6.671 0.00132 **
## Residuals   1029  7.047 0.00685
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Essay Question: Compare the two analyses. What does the non-significant interaction term suggest about the relationship between L_RATIO and the factors CLASS and SEX?

Answer: (The interaction term shows to be insignificant, the p-value is 0.867. On the other hand, the Class and Sex variables show to be significant with the L-Ratio.)

(2)(b) For the model without CLASS:SEX (i.e. an interaction term), obtain multiple comparisons with the *TukeyHSD()* function. Interpret the results at the 95% confidence level (*TukeyHSD()* will adjust for unequal sample sizes).

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = L_RATIO ~ CLASS + SEX, data = mydata)
```

```
##
## $CLASS
##              diff         lwr          upr     p adj
## A2-A1 -0.01248831 -0.03876038  0.013783756 0.6919456
## A3-A1 -0.03426008 -0.05933928 -0.009180867 0.0018630
## A4-A1 -0.05863763 -0.08594237 -0.031332896 0.0000001
## A5-A1 -0.09997200 -0.12764430 -0.072299703 0.0000000
## A3-A2 -0.02177176 -0.04106269 -0.002480831 0.0178413
## A4-A2 -0.04614932 -0.06825638 -0.024042262 0.0000002
## A5-A2 -0.08748369 -0.11004316 -0.064924223 0.0000000
## A4-A3 -0.02437756 -0.04505283 -0.003702280 0.0114638
## A5-A3 -0.06571193 -0.08687025 -0.044553605 0.0000000
## A5-A4 -0.04133437 -0.06508845 -0.017580286 0.0000223
##
## $SEX
##             diff          lwr           upr     p adj
## I-F -0.015890329 -0.031069561 -0.0007110968 0.0376673
## M-F  0.002069057 -0.012585555  0.0167236691 0.9412689
## M-I  0.017959386  0.003340824  0.0325779478 0.0111881
```

Additional Essay Question: first, interpret the trend in coefficients across age classes. What is this indicating about L_RATIO? Second, do these results suggest male and female abalones can be combined into a single category labeled as 'adults?' If not, why not?

Answer: ***(The null hypothesis that the infant ratio vs male/female ration is the same is rejected. The null hypothesis that male and female ratios are significantly different is not rejected. We can combine the male and female in one adult category.)***

Section 3: (10 points)

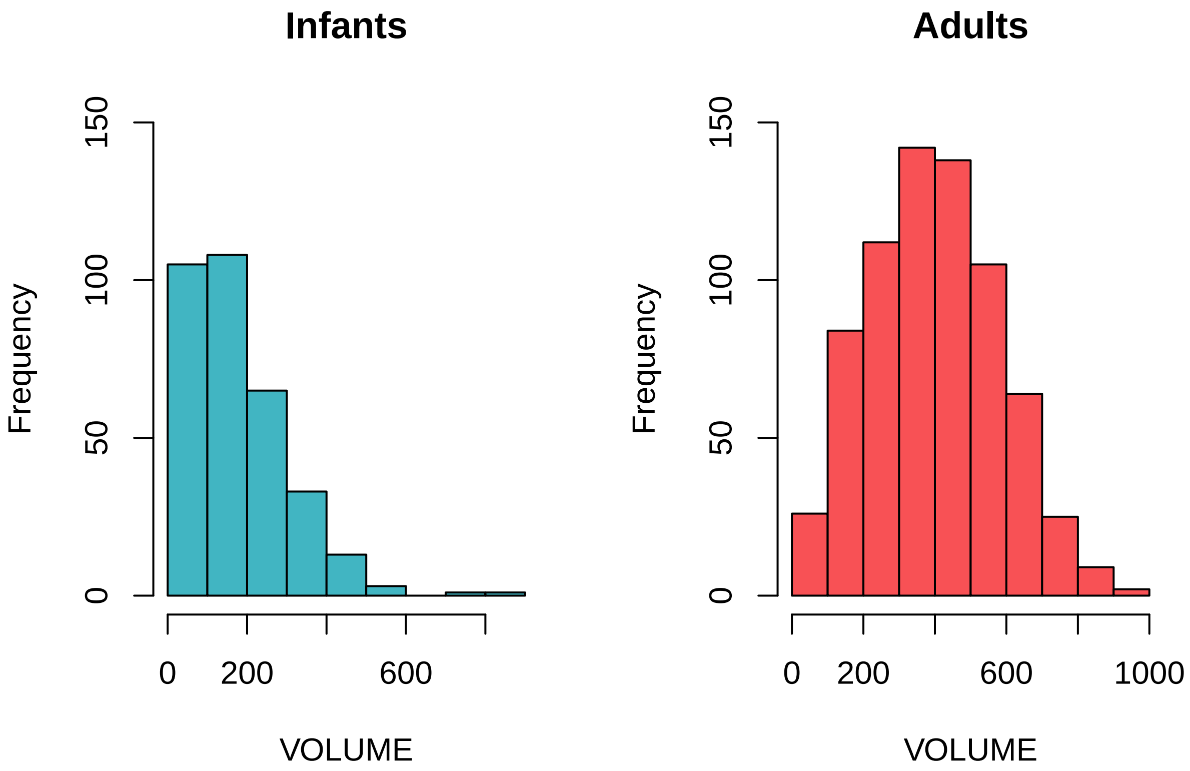
(3)(a1) We combine "M" and "F" into a new level, "ADULT". (While this could be accomplished using *combineLevels()* from the 'rockchalk' package, we use base R code because many students do not have access to the rockchalk package.) This necessitated defining a new variable, TYPE, in mydata which had two levels: "I" and "ADULT".

```
##
## Check on definition of TYPE object (should be an integer):  integer
```

```
##
## mydata$TYPE is treated as a factor:  TRUE
```

```
##
##     ADULT   I
##   F   326   0
##   I     0 329
##   M   381   0
```

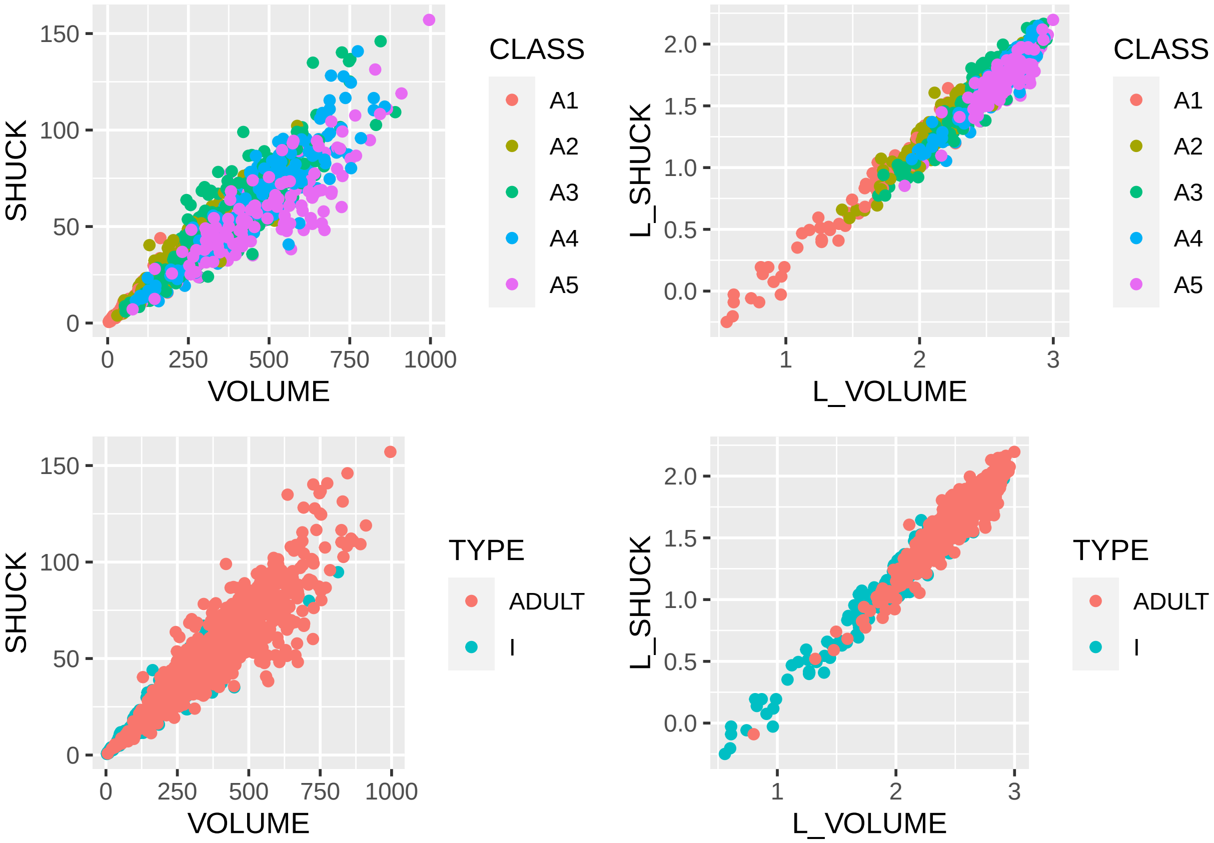
(3)(a2) Present side-by-side histograms of VOLUME. One should display infant volumes and, the other, adult volumes.



Essay Question: Compare the histograms. How do the distributions differ? Are there going to be any difficulties separating infants from adults based on VOLUME?

Answer: (The infant distribution is skewed to the right comparing to the adult distribution. It seems that most of the Infants' volume is 0-300. The majority of the Adults' volume is between 300 and 600. We can easily separate infants from adults judging by their volume.)

(3)(b) Create a scatterplot of SHUCK versus VOLUME and a scatterplot of their base ten logarithms, labeling the variables as L_SHUCK and L_VOLUME. Please be aware the variables, L_SHUCK and L_VOLUME, present the data as orders of magnitude (i.e. VOLUME = 100 = 10ˆ2 becomes L_VOLUME = 2). Use color to differentiate CLASS in the plots. Repeat using color to differentiate by TYPE.



Additional Essay Question: Compare the two scatterplots. What effect(s) does log-transformation appear to have on the variability present in the plot? What are the implications for linear regression analysis? Where do the various CLASS levels appear in the plots? Where do the levels of TYPE appear in the plots?

Answer: (The VOLUME and SHUCK plots overlap so that it is hard to distinguish clear lines between CLASS or TYPE. The log-transformed measures show divisions between the A1 group and the infants from the others. The INFANT abalones have concentrated group at l_volume < 2 and l_shuck <1. The Adult group shows to be right above on the plot.)

Section 4: (5 points)

(4)(a1) Since abalone growth slows after class A3, infants in classes A4 and A5 are considered mature and candidates for harvest. Reclassify the infants in classes A4 and A5 as ADULTS. This reclassification could have been achieved using *combineLevels()*, but only on the abalones in classes A4 and A5. We will do this recoding of the TYPE variable using base R functions. We will use this recoded TYPE variable, in which the infants in A4 and A5 are reclassified as ADULTS, for the remainder of this data analysis assignment.

```
##
## Check on redefinition of TYPE object (should be an integer):  integer
```

```
##
## mydata$TYPE is treated as a factor:  TRUE
```

```
##
## Three-way contingency table for SEX, CLASS, and TYPE:
```

```
## , ,  = ADULT
##
##
```

```
##      A1  A2  A3  A4  A5
##   F   5  41 121  82  77
##   I   0   0   0  21  19
##   M  12  62 143  85  79
##
## , ,  = I
##
##
##      A1  A2  A3  A4  A5
##   F   0   0   0   0   0
##   I  91 133  65   0   0
##   M   0   0   0   0   0
```

(4)(a2) Regress L_SHUCK as the dependent variable on L_VOLUME, CLASS and TYPE (Kabacoff Section 8.2.4, p. 178-186, the Data Analysis Video #2 and Black Section 14.2). Use the multiple regression model: L_SHUCK ~ L_VOLUME + CLASS + TYPE. Apply *summary()* to the model object to produce results.

```
##
## Call:
## lm(formula = L_SHUCK ~ L_VOLUME + CLASS + TYPE, data = mydata)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.270634 -0.054287  0.000159  0.055986  0.309718
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.796418   0.021718 -36.672  < 2e-16 ***
## L_VOLUME     0.999303   0.010262  97.377  < 2e-16 ***
## CLASSA2     -0.018005   0.011005  -1.636 0.102124
## CLASSA3     -0.047310   0.012474  -3.793 0.000158 ***
## CLASSA4     -0.075782   0.014056  -5.391 8.67e-08 ***
## CLASSA5     -0.117119   0.014131  -8.288 3.56e-16 ***
## TYPEI       -0.021093   0.007688  -2.744 0.006180 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08297 on 1029 degrees of freedom
## Multiple R-squared:  0.9504, Adjusted R-squared:  0.9501
## F-statistic:  3287 on 6 and 1029 DF,  p-value: < 2.2e-16
```

Essay Question: Interpret the trend in CLASS levelcoefficient estimates? (Hint: this question is not asking if the estimates are statistically significant. It is asking for an interpretation of the pattern in these coefficients, and how this pattern relates to the earlier displays).

Answer: (TThe coefficients show a stronger negative correlation in L_SHUNK as the classes increase. The previous plots also suggest that the L_SHUCK increases more rapidly in the lower classes. This means that the weight of the shell in older classes is greater than the shack.)

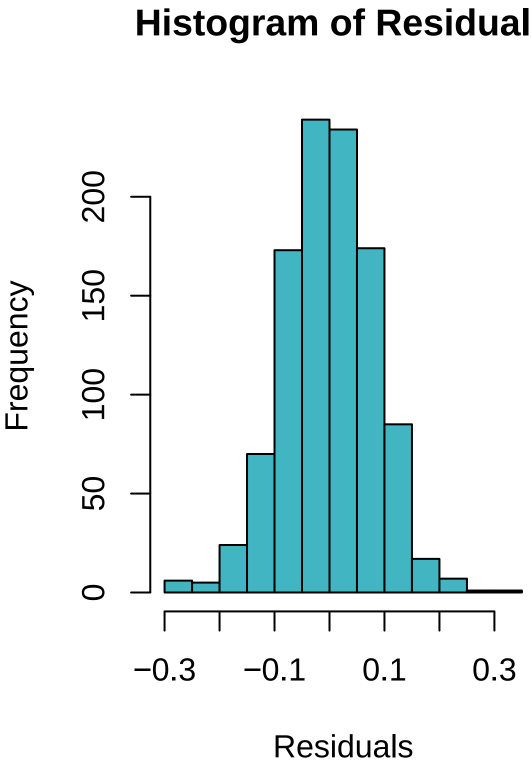
Additional Essay Question: Is TYPE an important predictor in this regression? (Hint: This question is not asking if TYPE is statistically significant, but rather how it compares to the other independent variables in terms of its contribution to predictions of L_SHUCK for harvesting decisions.) Explain your conclusion.

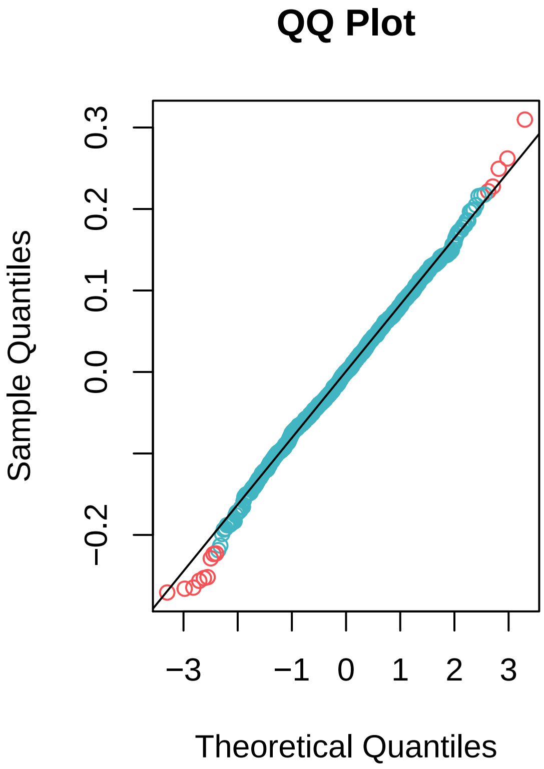
Answer: (Type is statistically significant but does not a strong coefficient to predict L_Shuck. Type would not be a good choice in predicting the L_Shuck compare to the other classes.)

The next two analysis steps involve an analysis of the residuals resulting from the regression model in (4)(a) (Kabacoff Section 8.2.4, p. 178-186, the Data Analysis Video #2).

Section 5: (5 points)

(5)(a) If "model" is the regression object, use model$residuals and construct a histogram and QQ plot. Compute the skewness and kurtosis. Be aware that with 'rockchalk,' the kurtosis value has 3.0 subtracted from it which differs from the 'moments' package.

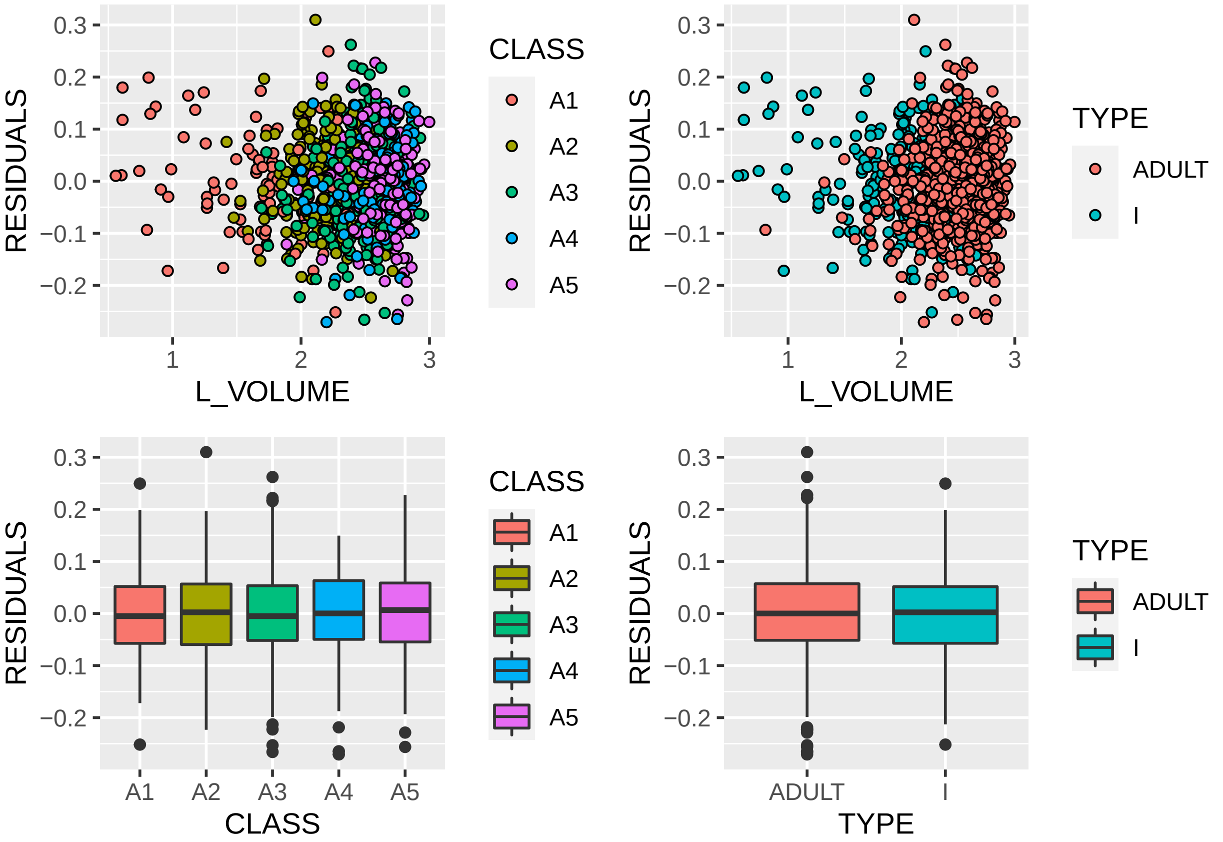




```
## [1] -0.05953853
```

```
## [1] 3.349772
```

(5)(b) Plot the residuals versus L_VOLUME, coloring the data points by CLASS and, a second time, coloring the data points by TYPE. Keep in mind the y-axis and x-axis may be disproportionate which will amplify the variability in the residuals. Present boxplots of the residuals differentiated by CLASS and TYPE (These four plots can be conveniently presented on one page using *par(mfrow..)* or *grid.arrange()*. Test the homogeneity of variance of the residuals across classes using *bartlett.test()* (Kabacoff Section 9.3.2, p. 222).



```
##
##  Bartlett test of homogeneity of variances
##
## data:  RESIDUALS by CLASS
## Bartlett's K-squared = 3.6882, df = 4, p-value = 0.4498
```

Essay Question: What is revealed by the displays and calculations in (5)(a) and (5)(b)? Does the model 'fit'? Does this analysis indicate that L_VOLUME, and ultimately VOLUME, might be useful for harvesting decisions? Discuss.

Answer: (Histogram 5a shows that residuals are evenly distributed and most residual values are 0 or around 0. Similar trends are seen in 5b. This means that most of the variability in the data can fit the model. We can conclude that L_VOLUME and VOLUME can be used for making decisions as they are good predictors for CLASS and TYPE.)

There is a tradeoff faced in managing abalone harvest. The infant population must be protected since it represents future harvests. On the other hand, the harvest should be designed to be efficient with a yield to justify the effort. This assignment will use VOLUME to form binary decision rules to guide harvesting. If VOLUME is below a "cutoff" (i.e. a specified volume), that individual will not be harvested. If above, it will be harvested. Different rules are possible.

The next steps in the assignment will require consideration of the proportions of infants and adults harvested at different cutoffs. For this, similar "for-loops" will be used to compute the harvest proportions. These loops must use the same values for the constants min.v and delta and use the same statement "for(k in 1:10000)." Otherwise, the resulting infant and adult proportions cannot be directly compared and plotted as requested. Note the example code supplied below.

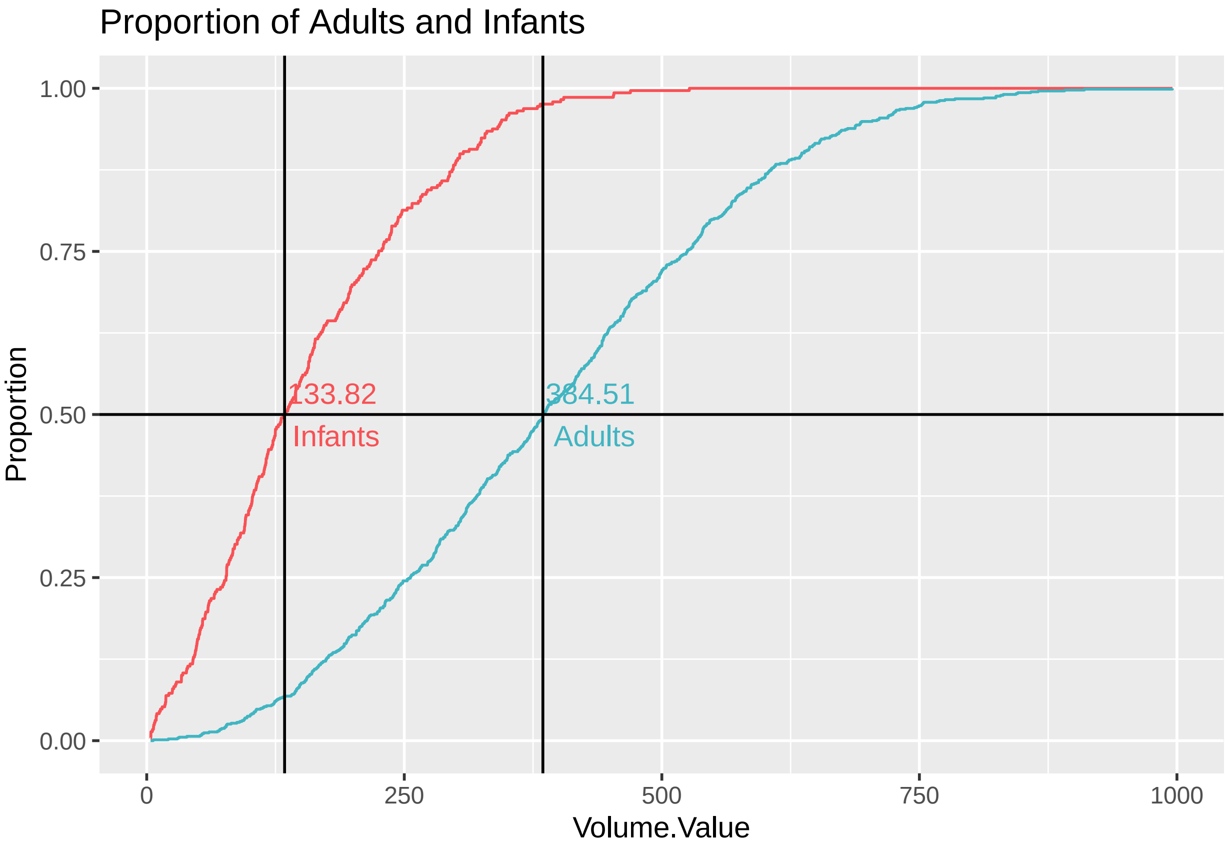
Section 6: (5 points)

(6)(a) A series of volumes covering the range from minimum to maximum abalone volume will be used in a "for loop" to determine how the harvest proportions change as the "cutoff" changes. Code for doing this is provided.

```
## [1] 133.8199
```

```
## [1] 384.5138
```

(6)(b) Present a plot showing the infant proportions and the adult proportions versus volume.value. Compute the 50% "split" volume.value for each and show on the plot.



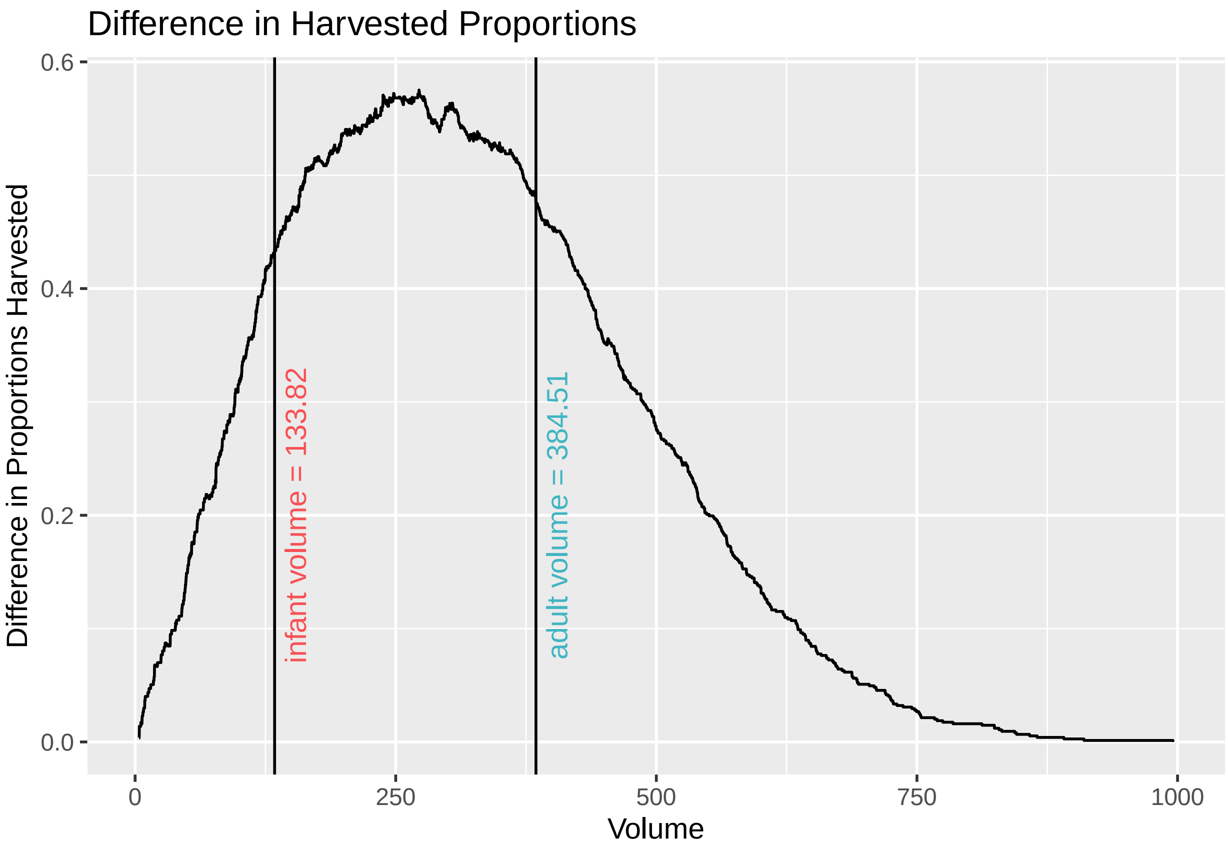
Essay Question: The two 50% "split" values serve a descriptive purpose illustrating the difference between the populations. What do these values suggest regarding possible cutoffs for harvesting?

Answer: (The volume values show a good distinction between adults and infants. The values suggest good cut off points of abalones. More adults should bet harvested than infants.)

This part will address the determination of a volume.value corresponding to the observed maximum difference in harvest percentages of adults and infants. To calculate this result, the vectors of proportions from item (6) must be used. These proportions must be converted from "not harvested" to "harvested" proportions by using (1 - prop.infants) for infants, and (1 - prop.adults) for adults. The reason the proportion for infants drops sooner than adults is that infants are maturing and becoming adults with larger volumes.

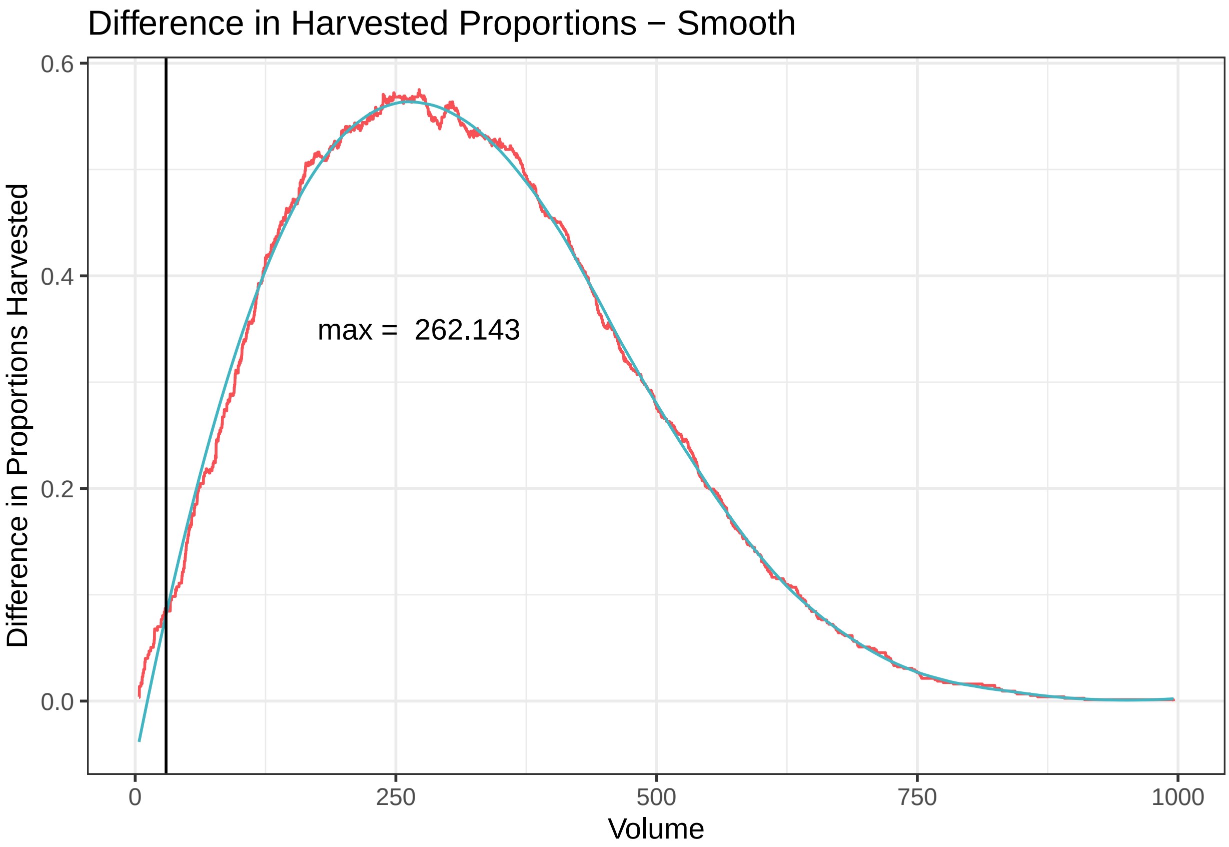
Section 7: (10 points)

(7)(a) Evaluate a plot of the difference ((1 - prop.adults) - (1 - prop.infants)) versus volume.value. Compare to the 50% "split" points determined in (6)(a). There is considerable variability present in the peak area of this plot. The observed "peak" difference may not be the best representation of the data. One solution is to smooth the data to determine a more representative estimate of the maximum difference.



(7)(b) Since curve smoothing is not studied in this course, code is supplied below. Execute the following code to create a smoothed curve to append to the plot in (a). The procedure is to individually smooth (1-prop.adults) and (1-prop.infants) before determining an estimate of the maximum difference.

(7)(c) Present a plot of the difference ((1 - prop.adults) - (1 - prop.infants)) versus volume.value with the variable smooth.difference superimposed. Determine the volume.value corresponding to the maximum smoothed difference (Hint: use *which.max()*). Show the estimated peak location corresponding to the cutoff determined.



(7)(d) What separate harvest proportions for infants and adults would result if this cutoff is used? Show the separate harvest proportions (NOTE: the adult harvest proportion is the “true positive rate” and the infant harvest proportion is the “false positive rate”).

Code for calculating the adult harvest proportion is provided.

```
## [1] 0.7416332
```

```
## [1] 0.1764706
```

There are alternative ways to determine cutoffs. Two such cutoffs are described below.

Section 8: (10 points)

(8)(a) Harvesting of infants in CLASS “A1” must be minimized. The smallest volume.value cutoff that produces a zero harvest of infants from CLASS “A1” may be used as a baseline for comparison with larger cutoffs. Any smaller cutoff would result in harvesting infants from CLASS “A1.”

Compute this cutoff, and the proportions of infants and adults with VOLUME exceeding this cutoff. Code for determining this cutoff is provided. Show these proportions.

```
## [1] 206.786
```

```
## [1] 0.8259705
```

```
## [1] 0.2871972
```

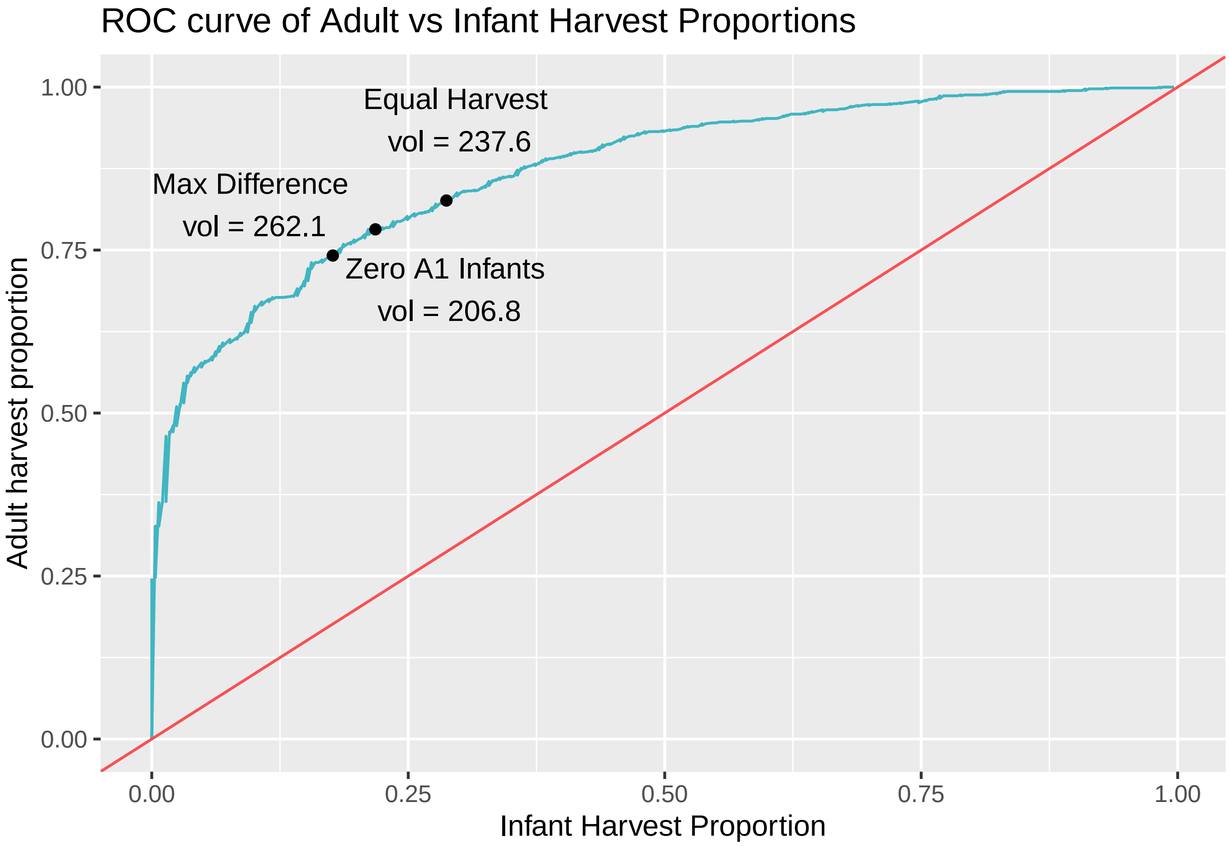
(8)(b) Another cutoff is one for which the proportion of adults not harvested equals the proportion of infants harvested. This cutoff would equate these rates; effectively, our two errors: ‘missed’ adults and wrongly-harvested infants. This leaves for discussion which is the greater loss: a larger proportion of adults not harvested or infants harvested? This cutoff is 237.7383. Calculate the separate harvest proportions for infants and adults using this cutoff. Show these proportions. Code for determining this cutoff is provided.

```
## [1] 0.7817938
```

```
## [1] 0.2179931
```

Section 9: (5 points)

(9)(a) Construct an ROC curve by plotting (1 - prop.adults) versus (1 - prop.infants). Each point which appears corresponds to a particular volume.value. Show the location of the cutoffs determined in (7) and (8) on this plot and label each.



(9)(b) Numerically integrate the area under the ROC curve and report your result. This is most easily done with the *auc()* function from the "flux" package. Areas-under-curve, or AUCs, greater than 0.8 are taken to indicate good discrimination potential.

```
## [1] 0.8666894
```

Section 10: (10 points)

(10)(a) Prepare a table showing each cutoff along with the following: 1) true positive rate (1-prop.adults, 2) false positive rate (1-prop.infants), 3) harvest proportion of the total population

```
##   Strategy         Volume TPR   FPR   Yield
## A Zero A1 Infants 206.8  0.826 0.287 0.689
## B Equal Error     237.6  0.782 0.218 0.638
## C Max Diff        262.1  0.742 0.176 0.598
```

Essay Question: Based on the ROC curve, it is evident a wide range of possible "cutoffs" exist. Compare and discuss the three cutoffs determined in this assignment.

Answer: (In the case of zero infants, we have the highest proportional yield but also a high false-positive rate. This is the least conservative approach. The equal error case has a safe medium rate of false-positive rate and proportional yield, so it is a moderate approach. The best approach, the most conservative, will be the max difference. It shows the lowest proportional yield value and the lowest false-positive rate.)

Final Essay Question: Assume you are expected to make a presentation of your analysis to the investigators How would you do so? Consider the following in your answer:

1. Would you make a specific recommendation or outline various choices and tradeoffs?
2. What qualifications or limitations would you present regarding your analysis?
3. If it is necessary to proceed based on the current analysis, what suggestions would you have for implementation of a cutoff? 4) What suggestions would you have for planning future abalone studies of this type?

Answer: (1. I would not recommend with a high degree of certainty any strategy. The summary above shows that the true-positive rate, false-positive rate, and harvest proportion are actually for all three strategies that are very close in values. 2. I would have to present the visuals od the histograms, box plots, outliers, kurtosis, and skewness. I would point out the difficulty and risks in assessing the classes and the level of quality of the data. 3. If a choice for a harvest strategy needs to be made, I will go with the most conservative - the max difference; it is the safest choice. 4. I would suggest exploring the data collection methods and more in-depth analysis if there are other important variables like environment, seasonality, and other independent variables that might affect.)