Abalones - Updated Title

Melinda Higgins, PhD - Associate Research Professor

February 11, 2019

## Abalones Dataset from UCI Repository

The abalone dataset is available from the UCI data repository at <https://archive.ics.uci.edu/ml/datasets/abalone>.

## Variables in Abalone Dataset

The variables in the abalone dataset are listed below.

names(abalone)

## [1] "sex" "length" "diameter" "height"   
## [5] "wholeWeight" "shuckedWeight" "visceraWeight" "shellWeight"   
## [9] "rings"

## Summary statistics of variables in abalone

summary(abalone)

## sex length diameter height   
## Length:4177 Min. :0.075 Min. :0.0550 Min. :0.0000   
## Class :character 1st Qu.:0.450 1st Qu.:0.3500 1st Qu.:0.1150   
## Mode :character Median :0.545 Median :0.4250 Median :0.1400   
## Mean :0.524 Mean :0.4079 Mean :0.1395   
## 3rd Qu.:0.615 3rd Qu.:0.4800 3rd Qu.:0.1650   
## Max. :0.815 Max. :0.6500 Max. :1.1300   
## wholeWeight shuckedWeight visceraWeight shellWeight   
## Min. :0.0020 Min. :0.0010 Min. :0.0005 Min. :0.0015   
## 1st Qu.:0.4415 1st Qu.:0.1860 1st Qu.:0.0935 1st Qu.:0.1300   
## Median :0.7995 Median :0.3360 Median :0.1710 Median :0.2340   
## Mean :0.8287 Mean :0.3594 Mean :0.1806 Mean :0.2388   
## 3rd Qu.:1.1530 3rd Qu.:0.5020 3rd Qu.:0.2530 3rd Qu.:0.3290   
## Max. :2.8255 Max. :1.4880 Max. :0.7600 Max. :1.0050   
## rings   
## Min. : 1.000   
## 1st Qu.: 8.000   
## Median : 9.000   
## Mean : 9.934   
## 3rd Qu.:11.000   
## Max. :29.000

## Specific statistics within text

We can use rmarkdown with R to embed R code within text to show the result in the final document instead of the code.

For example, the average height of the abalone is 0.1395164.

[ANSWER KEY] The median height is 0.14, the standard deviation of the heights is 0.0418271 and the min and max are 0, 1.13, respectively.

## Histogram of abalone heights

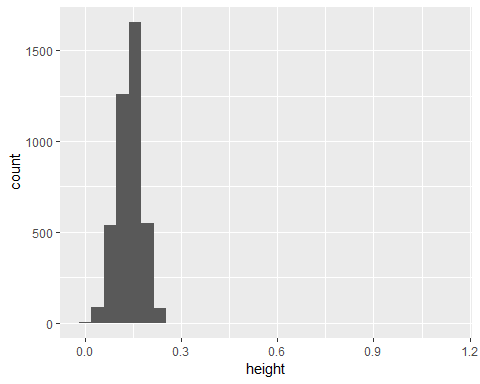
When typing text in rmarkdown we can add formatting like making works **BOLD** or adding other emphasis using *italics*.

We can also add bullets:

* What do you notice about the abalone heights?
* What could we do to investigate this issue further?

# make a histogram of height  
ggplot(data = abalone, aes(x = height)) +  
 geom\_histogram()

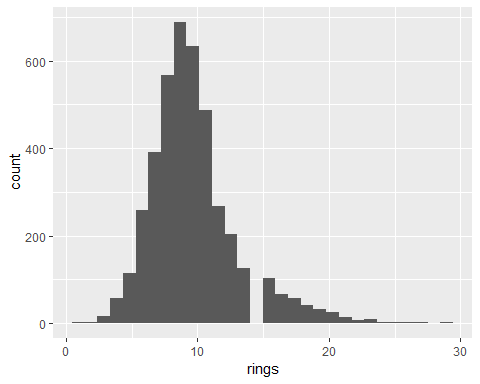
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## [ANSWER KEY] Histogram of abalone rings

# make a histogram of rings  
ggplot(data = abalone, aes(x = rings)) +  
 geom\_histogram()

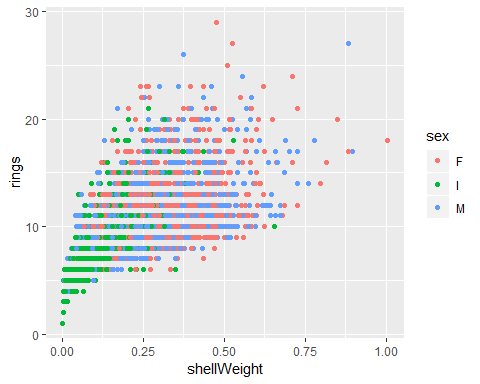
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



[ANSWER KEY] There is some slight skewness to the right (longer tail to the right) for the number of rings of the abalones. The distribution curve is also slightly peaked (positive kurtosis) but this is minor. Given the large sample size (>4000) no transformation is recommended.

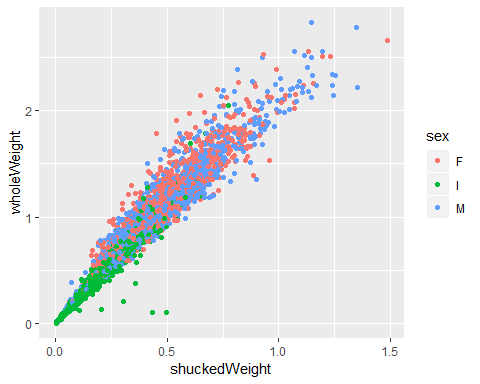
## Scatterplot of abalone rings by shellWeight

# make a scatterplot of rings by shellWeight  
# color points by sex  
ggplot(data = abalone,   
 aes(x = shellWeight, y = rings)) +  
 geom\_point(aes(color = sex))



## [ANSWER KEY] Scatterplot of abalone wholeWeight by shuckedWeight

# make a scatterplot of wholeWeight (y-axis) by   
# shuckedWeight (x-axis)  
# color points by sex  
ggplot(data = abalone,   
 aes(x = shuckedWeight, y = wholeWeight)) +  
 geom\_point(aes(color = sex))



[ANSWER KEY] In general, there is a positive correlation between shucked weight and whole weight which makes sense. however, there are 5 or so points in the lower left corner where the shucked weight is larger than the whole weight which doesn’t seem correct. This should be investigated further to see if these data were entered or recorded correctly.