January 19, 2022

The results below are generated from an R script.

```
---
title: "Abalone Dataset Summary"
output: pdf_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE, out.width="50%", fig.align = 'center')
library(tidyverse)
```

a) Abalone Data Set Summary

The Abalone data set is provided by UCI. It predicts the age of abalone from some physical
measurements such as length, weight, and diameter and etc. The data originally comes from
a study conducted by Warwick, Tracy, Simon, Andrew and Wes in 1994. There are a total of
4177 observations (rows) and 9 variables (columns) with the "Sex" variable being character,
"Rings" integer, and the rest double. The following table outlines the detail of all
attributes:

+----------------+------------+---------------------+-----------------------------+
| Name | Data Type | Meas. | Description |
+:===============:+:==========:+:====================:+:============================:+
| Sex | nominal | M, F, and I (infant) | |
+----------------+------------+---------------------+-----------------------------+
| Length | Continuous | mm | Longest shell measurement |
+----------------+------------+---------------------+-----------------------------+
| Diameter | Continuous | mm | perpendicular to length |
+----------------+------------+---------------------+-----------------------------+
| Height | Continuous | mm | with meat in shell |
+----------------+------------+---------------------+-----------------------------+
| Whole weight | Continuous | grams | whole abalone |
+----------------+------------+---------------------+-----------------------------+
| Shucked weight | Continuous | grams | weight of meat |
+----------------+------------+---------------------+-----------------------------+
| Viscera weight | Continuous | grams | gut weight (after bleeding) |
+----------------+------------+---------------------+-----------------------------+
| Shell weight | Continuous | grams | after being dried |
+----------------+------------+---------------------+-----------------------------+
| Rings | Integer | | +1.5 gives the age in year |
+----------------+------------+---------------------+-----------------------------+

```{r echo=FALSE}
abalone_col_names <- c("Sex",
```
```

```
                        "Length",
                        "Diameter",
                        "Height",
                        "Whole Weight",
                        "Shucked Weight",
                        "Viscera Weight",
                        "Shell Weight",
                        "Rings")
abalone <- read_csv('abalone.data',col_names = abalone_col_names,
                    col_types = cols(
                      Rings = col_integer()
                    )
)
```

b) Histogram of Length

Now we will select the Length variable to be studied. In the following, we plot a density histogram with a plot of the pdf of a normal distribution superimposed:

```{r echo=FALSE}
hist(abalone$Length,
     main = "Length of Abalone",
     prob = TRUE,
     xlab = "Length (mm)",
     ylab= "Density")

mu = mean(abalone$Length)

v = var(abalone$Length)
s = sqrt(v)

# s = sd(abalone$Length)

xs = sort(abalone$Length)
ys = dnorm(xs, mean = mu, sd = s)
lines(xs,ys,cex=0.1)
```

*** Additionally, if we must display the y-axis in proportion instead of density:

The following graph, created with *ggplot*, scales the y-axis into percentage.

```{r echo=FALSE}


p <- ggplot(abalone, aes(x=Length)) +
         geom_histogram(aes(y = (..count..)/sum(..count..)), binwidth = 0.05, bins = 20,
         colour="lightgrey", fill="darkgrey") +
  stat_bin(aes(y=(..count..)/sum(..count..),
               label=paste0(round((..count..)/sum(..count..)*100,1),"%")),
           geom="text", size=3, binwidth = 0.05, vjust=-0.5) +
  labs(x="Length (mm)", y="Proportion")
p
```

```

And if we superimpose a plot of the pdf of a normal distribution on the above graph, it scales down the y-axis and looks like this:

```{r echo=FALSE}
p + stat_function(fun = dnorm, args = list(mean = mean(abalone$Length),
                            sd = sd(abalone$Length)))
```
```