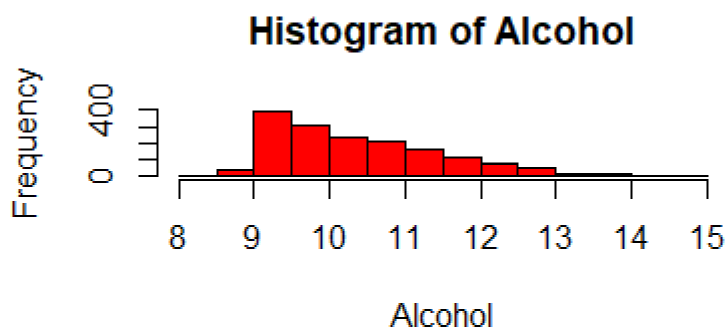
Data Science HW1 Part 2

Matt McCullough

Number 1

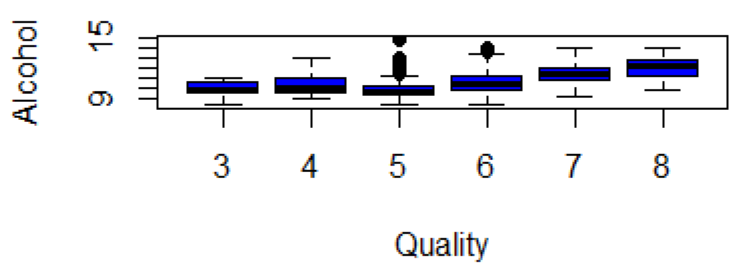
a. Downloaded the red wine file
b. The shape of the distribution of alcohol content in red wine is right skewed



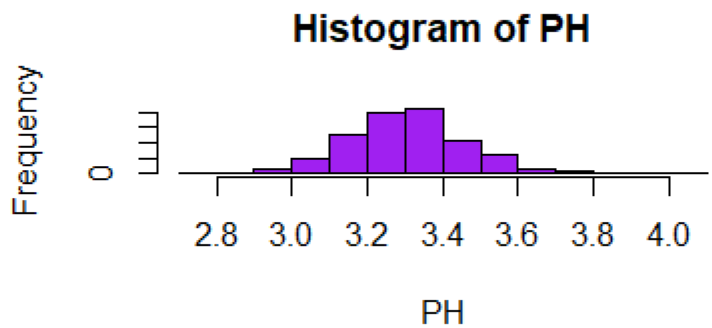
c. The numerical summary that would be most appropriate for this histogram is the five number summary as it highlights the key points of the data, and the mean isn't important given the low range

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
8.40    9.50   10.20   10.42   11.10   14.90
```

d. The quality that tends to have the highest alcohol content is 8



e. I expect the histogram to have the empirical rule hold as it’s a mound shaped distribution



f. The interval that captures 95% of the values is (3.002340 - 3.619886) and the fraction of the values that are inside this interval is 95.30957% of the data

Part 2

2a) $x_1, x_2, \ldots, x_n > 0$ $\quad c > 0$

Show GM of $cx_1, cx_2, cx_3 \ldots, cx_n$ is $|c| \cdot$ GM of $x_1 x_2 \ldots x_n$

old GM $= \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$

The GM of a set of numbers is the n^{th} root of their product

new GM $= \sqrt[n]{cx_1 \cdot cx_2 \cdot \ldots, cx_n}$

Given our new set of numbers all multiplied by c

the GM can be written like this

$\sqrt[n]{c^n(x_1 \cdot x_2 \cdot \ldots \cdot x_n)}$

simplified $\Rightarrow \sqrt[n]{c^n} \cdot \sqrt[n]{(x_1 \cdot x_2 \cdot \ldots \cdot x_n)}$

$\sqrt[n]{c^n}$ can be simplied to $|c|$ and $c > 0$

Since $\sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$ is the original GM,

the GM of $cx_1, cx_2, \ldots, cx_n$ is $c \cdot \overset{(old)}{GM}$

2b) Finite geometric sequence $1, r, r^2, \ldots, r^{n-1}$ $r>0$

n is odd

Show median coincides with GM

median of sequence is $r^{\frac{(n-1)}{2}}$ (median odd - $\frac{n}{2}$ term, even average of middle 2)

GM is $\sqrt[n]{1 \cdot r \cdot r^2 \cdot r^n}$

GM can be written as

$$\sqrt[n]{r^{\sum_{i=1}^{n-1} i}} \longrightarrow r^{\left(\frac{1}{n} \cdot \sum_{i=1}^{n-1} i\right)}$$

Now showing median coincides with GM

Since r is the same base, set exps to eachother

$$\frac{n-1}{2} = \frac{1}{n} \sum_{i=1}^{n-1} i$$

$$\sum_{i=1}^{n-1} i = \frac{(n-1)n}{2}$$

$$\frac{1}{n} \cdot \frac{(n^2-n)}{2} = \frac{n^2-n}{2n} = \frac{n(n-1)}{2n} = \boxed{\frac{n-1}{2}}$$

same as median

Therefore, the GM coincides with the median

3) $x_1, x_2, \ldots, x_n \in \mathbb{R}$ $\quad c \in \mathbb{R}$

Show that standard deviation of $cx_1, cx_2, \ldots, cx_n$ is $|c|$ times the standard deviation of $x_1, x_2, \ldots, x_n$

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$s^2_{(new)} = \frac{\sum_{i=1}^{n} (cx_i - \bar{x}_{(new)})^2}{n-1} \rightarrow \text{sample variance}$$

Finding new mean $\rightarrow$ old mean

$$\bar{x}_{(new)} = \frac{cx_1 + cx_2 + \ldots + cx_n}{n} = c\left(\frac{x_1 + x_2 + \ldots + x_n}{n}\right) = c\bar{x}$$

$$s^2_{(new)} = \frac{\sum_{i=1}^{n} (cx_i - c\bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} c^2(x_i - \bar{x})^2}{n-1}$$

$$\sqrt{c^2} = |c|$$

standard deviation (square root of sample variance) $= \sqrt{\frac{\sum_{i=1}^{n} c^2(x_i - \bar{x})^2}{n-1}} \rightarrow$ effectively multiplied by $\sqrt{c^2} \rightarrow |c|$

Therefore, the standard deviation of $cx_1, cx_2, \ldots, cx_n$ is $|c|$ times the standard deviation of $x_1, x_2, \ldots, x_n$

4a) Let $x_1, \ldots, x_n$ be any real numbers and let $\bar{x} = (x_1 + \ldots + x_n)/n$ be sample mean

(ii) Show that s^2 can also be calculated from the formula

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}^2 n + n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$= \sum_{i=1}^{n} x_i^2 - n\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$$

$$= n \cdot \frac{1}{n^2}\left(\sum_{i=1}^{n} x_i^2\right)$$

$$\sum_{i=1}^{n} x_i^2 - \frac{n}{n^2}\left(\sum_{i=1}^{n} x_i^2\right)$$

$$\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) \quad \text{(two } n\text{'s cancel)}$$

Since $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right)$, s^2 can be calculated with the above formula

4 ii) Show that $\sum_{i=1}^{n} x_i^2 = (n-1)s^2 + n\bar{x}^2$

$(n-1)s^2 = (n-1) \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

n-1 cancels

$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} i\right)$

$(n-1)s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$

$(n-1)s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i^2 - 2\bar{x}x_i + \bar{x}^2)$

$= \sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2$

$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}^2 n + n\bar{x}^2$

$(n-1)s^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$

Since $(n-1)s^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$ and adding $n\bar{x}^2$ would cancel the $n\bar{x}^2$s to just be $\sum_{i=1}^{n} x_i^2$, the above equality is true

5) $b_1, b_2, \ldots, b_n$ is permutation of positive real numbers $a_1, a_2, \ldots, a_n$

Show that $\frac{a_1}{b_1} + \frac{a_2}{b_2} + \ldots + \frac{a_n}{b_n} \geq n$

$\frac{\frac{a_1}{b_1} + \frac{a_2}{b_2} + \ldots + \frac{a_n}{b_n}}{n} \geq \sqrt[n]{\frac{a_1}{b_1} \cdot \frac{a_2}{b_2} \cdots \frac{a_n}{b_n}}$

Since set a and b are the same numbers but b is a permutation, they will multiply to be 1 since the product of set a = product of set b and will cancel

$\frac{\frac{a_1}{b_1} + \frac{a_2}{b_2} + \ldots + \frac{a_n}{b_n}}{n} \geq \sqrt[n]{1}$

$\sqrt[n]{1} = 1$

$\frac{\frac{a_1}{b_1} + \frac{a_2}{b_2} + \ldots + \frac{a_n}{b_n}}{n} \geq 1$

$\frac{a_1}{b_1} + \frac{a_2}{b_2} + \ldots + \frac{a_n}{b_n} \geq n$

Therefore, the above is true

7) For all integers $n>1$, show that $n! < \left(\frac{n+1}{2}\right)^n$

Use AM-GM inequality $AM \geq GM$

Set is $1,2,\dots,n$ (varies based on n)

$$\frac{1+2+\dots+n}{n} \geq \sqrt[n]{1\cdot 2\cdot\dots\cdot n}$$

$$1\cdot 2\cdot\dots\cdot n = n!$$

$$\frac{1+2+\dots+n}{n} \geq \sqrt[n]{n!}$$

$$1+2+\dots+n = \sum_{i=1}^{n} i$$

$$\frac{\sum_{i=1}^{n} i}{n} \geq \sqrt[n]{n!}$$

$$\sum_{i=1}^{n} i = \frac{(n+1)n}{2} \to \frac{n^2+n}{2}$$

$$\frac{\frac{n^2+n}{2}}{n} \geq \sqrt[n]{n!}$$

$$\frac{n^2+n}{2n} \geq \sqrt{n!}$$

$$\frac{\cancel{n}(n+1)}{2\cancel{n}} \geq \sqrt[n]{n!}$$

$$\frac{(n+1)}{2} \geq \sqrt[n]{n!}$$

$$\left(\frac{(n+1)}{2}\right)^n \geq n! \longrightarrow$$ can be written as $n! < \left(\frac{(n+1)}{2}\right)^n$

I couldn't do the other extra credits (6 and 8)

R Code I used to do number 1

```
#Matt McCullough HW1 Part 2

#load in csv file
WineRed=read.csv("winequality-red.csv",header=T,sep=";")

#create histogram of alcohol data
Alcohol = WineRed[["alcohol"]]
```

```
hist(Alcohol,col = "red")

#five number summary
summary(Alcohol)

#pull in quality data
Quality = WineRed[["quality"]]

#box plot
boxplot(Alcohol~Quality, pch = 19, col = "blue")

#histogram of PH data
PH = WineRed[["pH"]]
hist(PH, col = "purple")

#interval of 95% of the data (2 standard deviations away)
interval = c(mean(PH)-2*sd(PH),mean(PH)+2*sd(PH))
interval


#get real fraction of 95% of data
totalInInterval = sum(interval[1] <= PH & PH <= interval[2])
fraction = totalInInterval/length(PH)
fraction * 100
```