STATS 500 HW 1

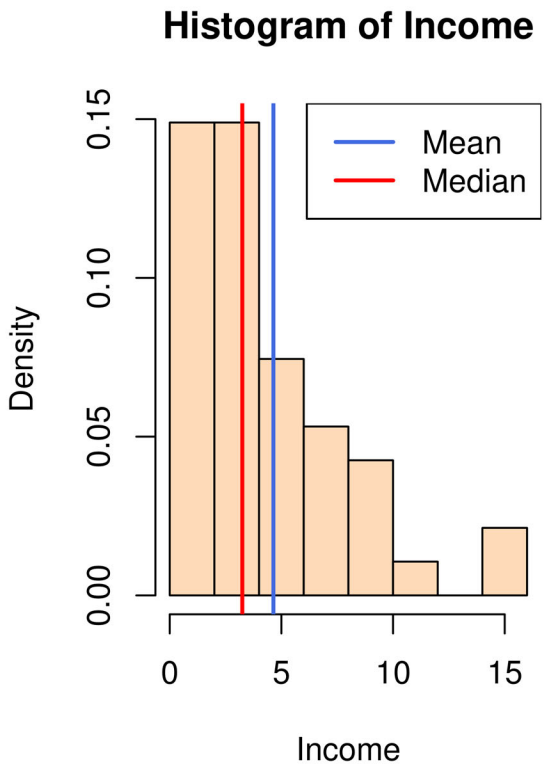
General Numerical

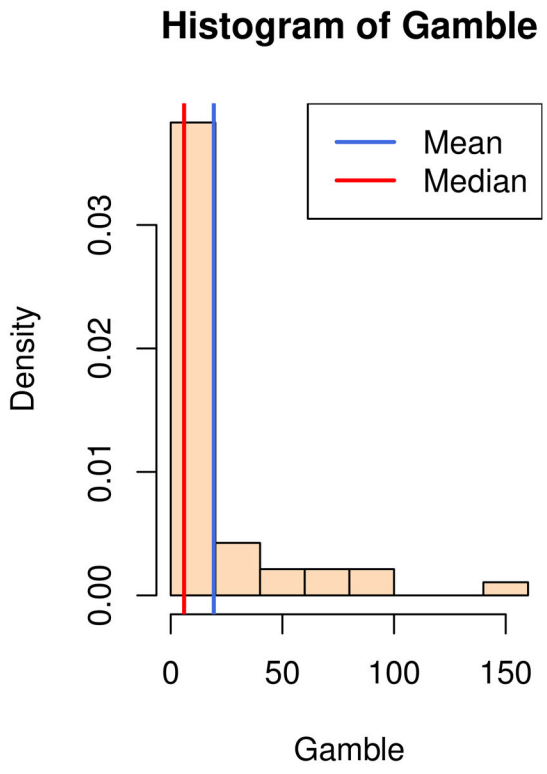
In this general summary we are able to note that the variable **sex** has been converted to a factor and named with the "Male" & "Female" label. There seem to be more males (28) than females (19) in this dataset. At first glance, the varibales **status** , **income**, and **verbal** seem to have no extreme values. Unlike the variable **gamble** in which we are able to note maximum value of 156- this is too far away from the mean & a possible ourlier/error.

```
##      sex         status          income          verbal          gamble
##  FEMALE:19   Min.   :18.00   Min.   : 0.600   Min.   : 1.00   Min.   :  0.0
##  MALE   :28   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00   1st Qu.:  1.1
##               Median :43.00   Median : 3.250   Median : 7.00   Median :  6.0
##               Mean   :45.23   Mean   : 4.642   Mean   : 6.66   Mean   : 19.3
##               3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00   3rd Qu.: 19.4
##               Max.   :75.00   Max.   :15.000   Max.   :10.00   Max.   :156.0
```

Income & Gamble

The **mean** for *income* is: 4.642 pounds per week & the **median** is 3.250 ppounds per week. Whereas, the **mean** for *gamble* is 19.3 pounds per year & the **median** is 6.0 pounds per year. The reason for the means being larger than the medians, in both cases, has to due with the *skewed* distribnution that both of this variables have. Both variables tend to have a few observations that have extreme values. Now, these extreme vaues skew the mean. The key idea is that the mean is easily influeced by extreme values, whereas the median is not. As a result, in both cases we have skewed distributions & in both cases the mean is greater than the median.



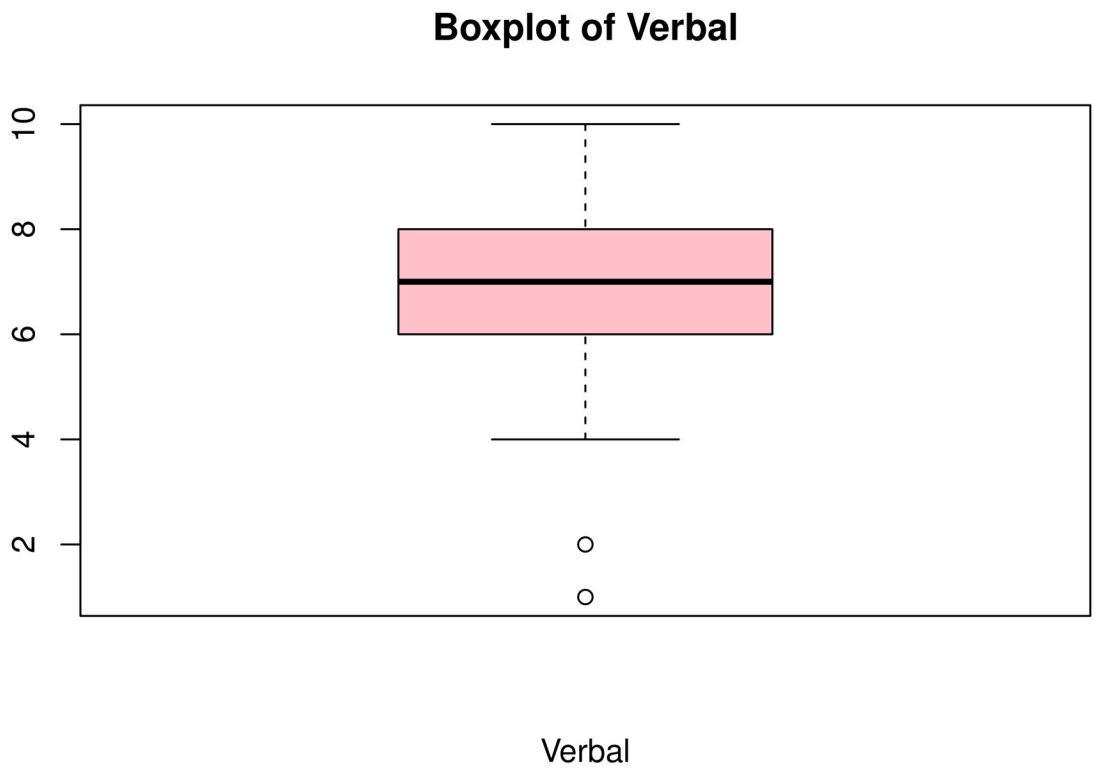


Verbal

There are a **total** of 47 observations in the verbal column, of which only **9** are considered to be uniqe.

```
## [1]  1  2  4  5  6  7  8  9 10
```

In order to find outliers in a data, we can inspect the data graphically through a boxplot, or we can look for the points that are below Q1- 1.5* IQR & above Q3+1.5*IQR. In the boxplot, we are able to see that there are two outliers. These points are the ones that are outside of the whiskers.



We are able to identify these points by checking for the ones that satisfy the aformentioned conditions (1.5* IQR). Such points are **1** and **2** and belong to observation **31** and **35**.

```
teengamb$verbal[teengamb$verbal < 6-1.5*2 ]
```

```
## [1] 2 1
```

```
teengamb[teengamb$verbal == 1, ]
```

```
##     sex status income verbal gamble
## 35 MALE     28    1.5      1   14.1
```

```
teengamb[teengamb$verbal == 2, ]
```

```
##     sex status income verbal gamble
## 31 MALE     18     12      2     88
```

Interesting Findings

Correlation is a quantity that allows us to determine the *stregth* of a relationship among two quantitative variables. After some investigating, I was able to find some strong correlations, where strong is defined as a relationship with a correlation greater than 51%.

```
##             status     income     verbal      gamble
## status  1.00000000 -0.2750340  0.5316102 -0.05042081
## income -0.27503402  1.0000000 -0.1755707  0.62207690
## verbal  0.53161022 -0.1755707  1.0000000 -0.22005619
## gamble -0.05042081  0.6220769 -0.2200562  1.00000000
```

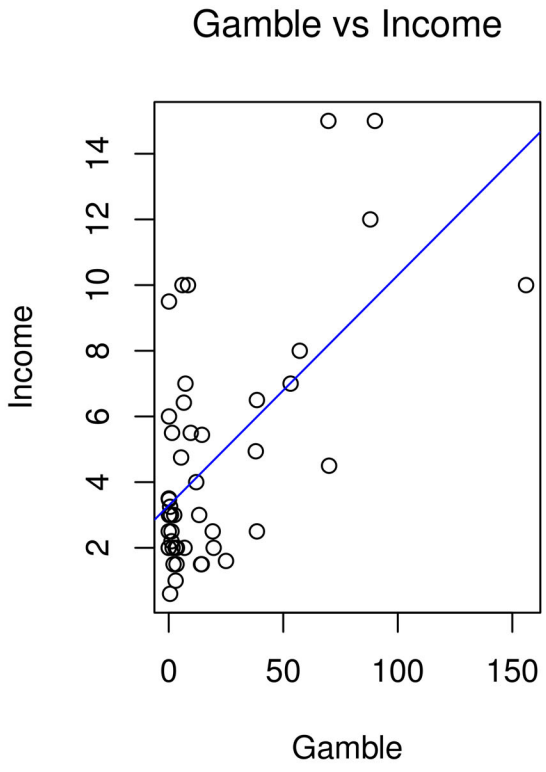
The strongest correlation was found between the variables **gamble-income** and **status-verbal**. Both of these correaltaions are positive. Intuitively, both findings make sense.

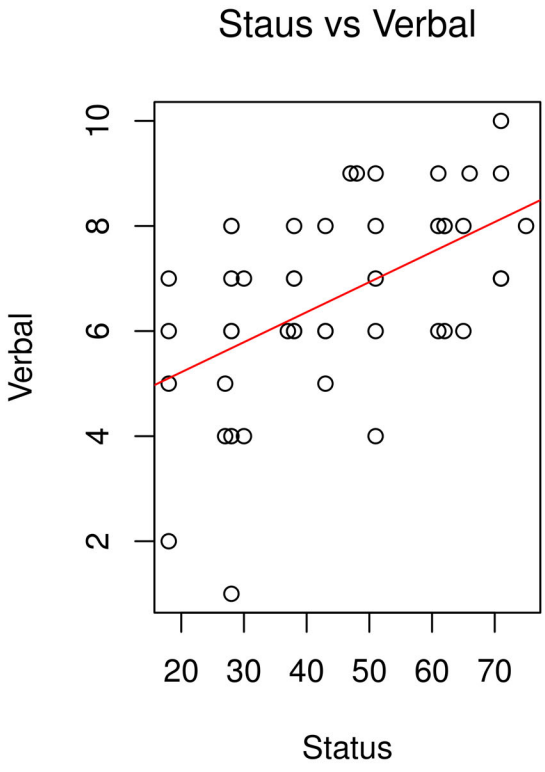
In the case of **gamble-income**, one would expect that the higher the income the more likely one is to gamble. Perhaps, this is not alwatys true, but not necessarily false either.

In the case of **status-verbal**, we are able to see that individuals with a better socioeconomic staus seem to be doing better at the verabal testing. The implications of this finding suggests that rich people are more likely to do good in academia than poor people. Altough a bit shocking, this finding does indeed reflect the general reality.

The **scatterplots** below allow us to get a graphical representation of the strength of these relationship. The blue and red line represent their respecitve regression lines.

Overall, these discoveries are basic in their meaning, but extremely revelatory to a beginner like myself.





References: https://www.r-bloggers.com/adding-measures-of-central-tendency-to-histograms-in-r/

Linear Models with R & Canvas slides