R Report 1

Jeffrey Ugochukwu

6/27/2019

I (Introduction)

My name is Jeffrey Ugochukwu and I'm from Vallejo, California. I am currently majoring in Statistical Data Science as my field of study. I am currently a freshman at UC Davis. I'm taking STA 32 because it is a class required for my major. My favorite hobby is being a musician since I'm currently a member of the Cal Aggie Marching Band-Uh. We perfom at many sporting events (football, basketball, etc...), do a lot of field show perfomances, parades, and we even do small gigs requested by specific groups or people whether it would be at the school or just the Davis community in general. Recently, I performed my first ever field show solo during the Picnic Day parade in which all of the band members would kneel at the soloists' prescence while each soloist performs their solo in front of the audience. Coming into this class, I define statistics as an applied mathematical science in which you use quanitative values to measure the success or failure rate of a given problem so that a conclusion can be made on how to approach that problem directly.

II (Data reguarding the Patients)

(a)

The list of columns in the patients.csv is given below:

```
## [1] "age"       "totalchol" "sysBP"     "weight"    "height"    "sedmins"
## [7] "obese"     "marriage"  "gender"
```

(b)

The number of rows in the patients.csv is given below:

```
## [1] 4915    9
```

(c)

This is the summary of the patients.csv. This function treats categorical functions in a given dataset as a measure of the frequency per category and it treats the numeric functions as a calculation that measures the summary statistic of each given category.

```
##       age          totalchol         sysBP         weight
##  Min.   :20.00   Min.   : 92.0   Min.   : 90   Min.   : 33.20
##  1st Qu.:34.00   1st Qu.:166.0   1st Qu.:112   1st Qu.: 67.00
##  Median :48.00   Median :192.0   Median :122   Median : 78.50
##  Mean   :49.27   Mean   :195.7   Mean   :124   Mean   : 81.19
```

```
##  3rd Qu.:63.00   3rd Qu.:221.0   3rd Qu.:134   3rd Qu.: 92.60
##  Max.   :80.00   Max.   :383.0   Max.   :220   Max.   :159.10
##      height         sedmins              obese              marriage
##  Min.   :135.4   Min.   :  0.0   normal     :1305   divorced    : 537
##  1st Qu.:160.1   1st Qu.:180.0   obese      :1840   married     :2569
##  Median :167.2   Median :300.0   overweight :1697   nevermarried: 841
##  Mean   :167.5   Mean   :311.3   underweight:  73   other       : 569
##  3rd Qu.:175.0   3rd Qu.:420.0                      widowed     : 399
##  Max.   :202.7   Max.   :840.0
##  gender
##  F:2489
##  M:2426
##
##
##
##
```

(d)

The mean of the age of the patients is given below:

```
## [1] 49.27243
```

(e)

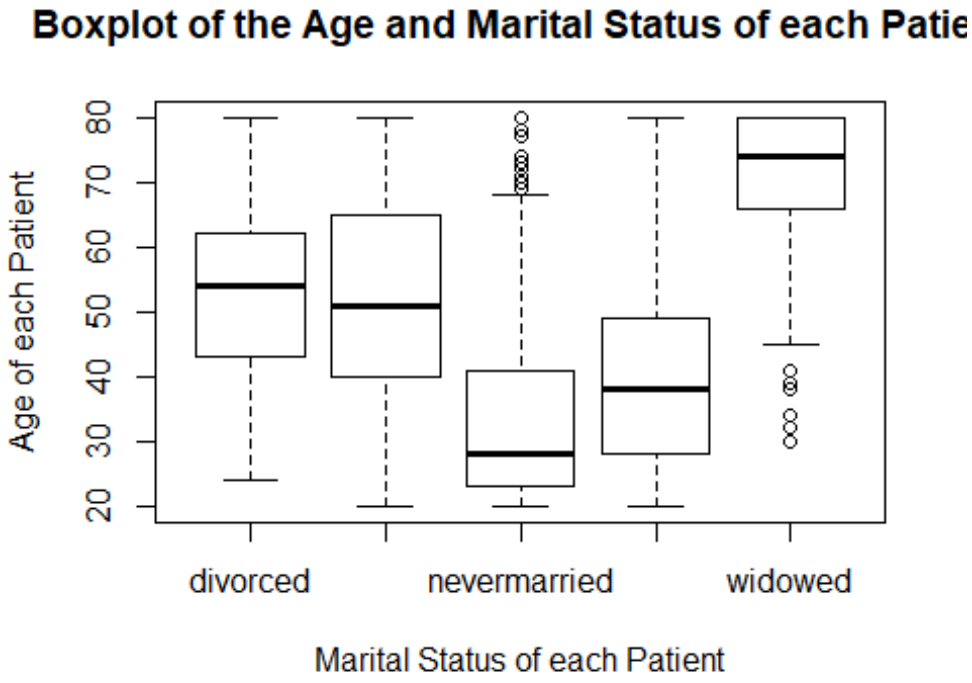
The mean and standard deviation of the age of each gender is given below:

```
##   gender      age
## 1      F 48.83126
## 2      M 49.72506
```

(f)

The boxplot of the age and marital status for each patient is given below. The graph displays the 5 number summaries of the age of patients that are divorced, married, never married, other, and widowed. We can see that the age of the patients that were divorced is roughly symmetric with all of the ages evenly distributed from the age of 24 years old to the age of 80 years old. This category has a 1st quartile of around 43 years old, a median of 54 years old, and a 3rd quartile of around 62 years old. The age of the patients that are married is also roughly symmetric with a even distribution from 20 to 80 years old. It has a 1st quartile of 40 years old, a median of 51 years old, and a 3rd quartile of 65 years old. The patients that never got married is actually skewed to the left with the majority of the values from the 1st quartile to the 3rd quartile ranging from 23 years old to 41 years old. This also has the smallest median out of all of the categories with a median of 30 years old. It also has outliers from that extend outside of the minimum max value of 68 years old. The category where the patients identified as “other” is skewed to the left with the majority of the values from the 1st quartile to the 3rd quartile ranging from 28 years old to 59 years old. It also has the 2nd smallest median with a median of 38 years old. The patients that are currently widowed have a distribution that is skewed to the right with majority of the values from

the 1st quartile to the 3rd quartile ranging from 66 to 80 years old. This also has the largest median out of all of the categories with a median of 74 years old. There's also a noticible trend that the youngest participants in terms of marital status tend to never have previouly married or they have a different situation applying to marriage since the quartile ranges (range from the 1st quartile to the 3rd) are before the age of 60 and the oldest group based on the quartile range tend to be widowed with a quartile range from 66 to 80 years old.

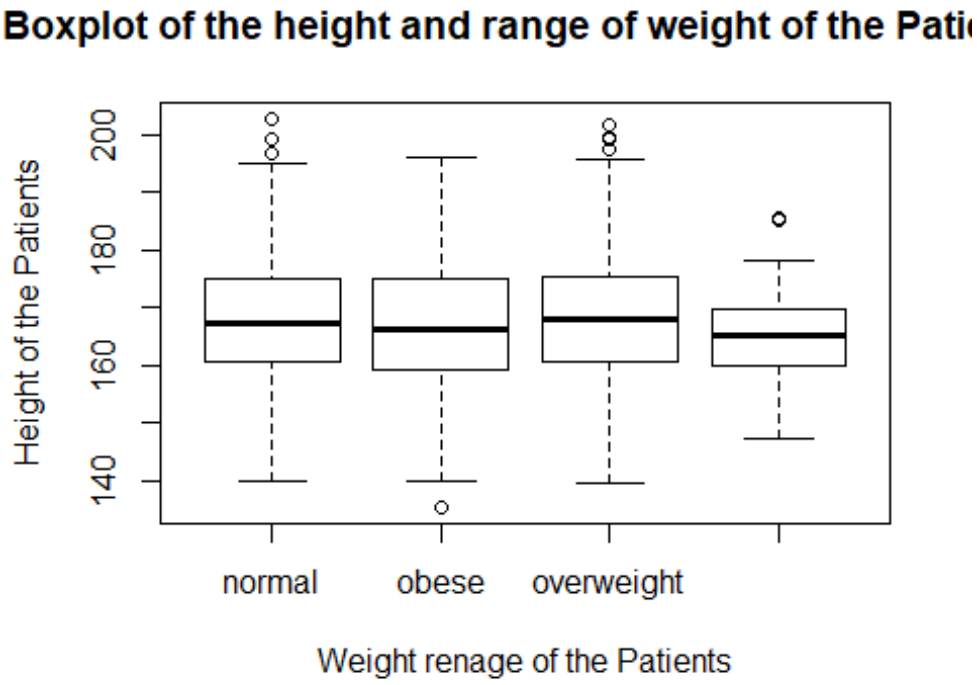



```
## $stats
##      [,1] [,2] [,3] [,4] [,5]
## [1,]   24   20   20   20   45
## [2,]   43   40   23   28   66
## [3,]   54   51   28   38   74
## [4,]   62   65   41   49   80
## [5,]   80   80   68   80   80
## attr(,"class")
##  divorced
## "integer"
##
## $n
## [1]  537 2569  841  569  399
##
## $conf
##          [,1]     [,2]     [,3]     [,4]     [,5]
## [1,] 52.70454 50.22068 27.01931 36.60902 72.89261
## [2,] 55.29546 51.77932 28.98069 39.39098 75.10739
##
```

```
## $out
##  [1] 72 70 70 80 70 71 78 80 70 72 80 80 80 69 73 72 80 77 72 74 70 73 77
## [24] 77 41 30 41 39 34 32 39 38
##
## $group
##  [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 5 5 5 5 5 5 5 5
##
## $names
## [1] "divorced"     "married"      "nevermarried" "other"
## [5] "widowed"
```

(g)

The boxplot represents the height of the patients with correspondence to the weight range of each patient. All graphs that represnt the categories are all roughly symmetric, but we can see how the underweight group has the smallest spread with a range of 147.4 to 178.2 inches in height (excluding outliers). Majority of the categories have the same median of around 168 inches in height with the exception of the group that's obese since they have a median height of 170 inches. Majority of the graphs have outliers that extend past the minimum max height range, but the obese group has an outlier that's below the minimum min height for the range.

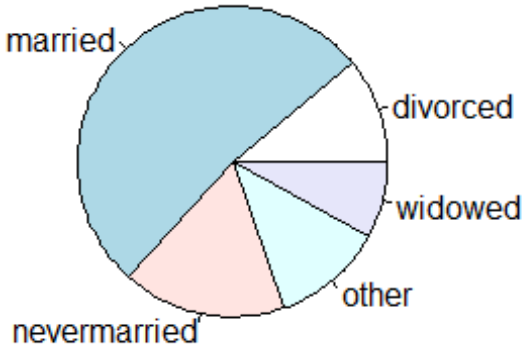



```
## $stats
##       [,1]   [,2]  [,3]  [,4]
## [1,] 139.8 139.80 139.4 147.4
## [2,] 160.5 159.20 160.6 160.0
```

```
## [3,] 167.3 166.35 167.9 165.2
## [4,] 174.9 175.00 175.2 169.6
## [5,] 194.9 195.90 195.5 178.2
##
## $n
## [1] 1305 1840 1697   73
##
## $conf
##          [,1]    [,2]   [,3]     [,4]
## [1,] 166.6702 165.768 167.34 163.4247
## [2,] 167.9298 166.932 168.46 166.9753
##
## $out
##  [1] 199.3 202.7 196.6 135.4 199.2 197.4 197.5 199.6 201.7 185.1 185.5
##
## $group
##  [1] 1 1 1 2 3 3 3 3 3 4 4
##
## $names
## [1] "normal"      "obese"       "overweight"  "underweight"
```

(h)

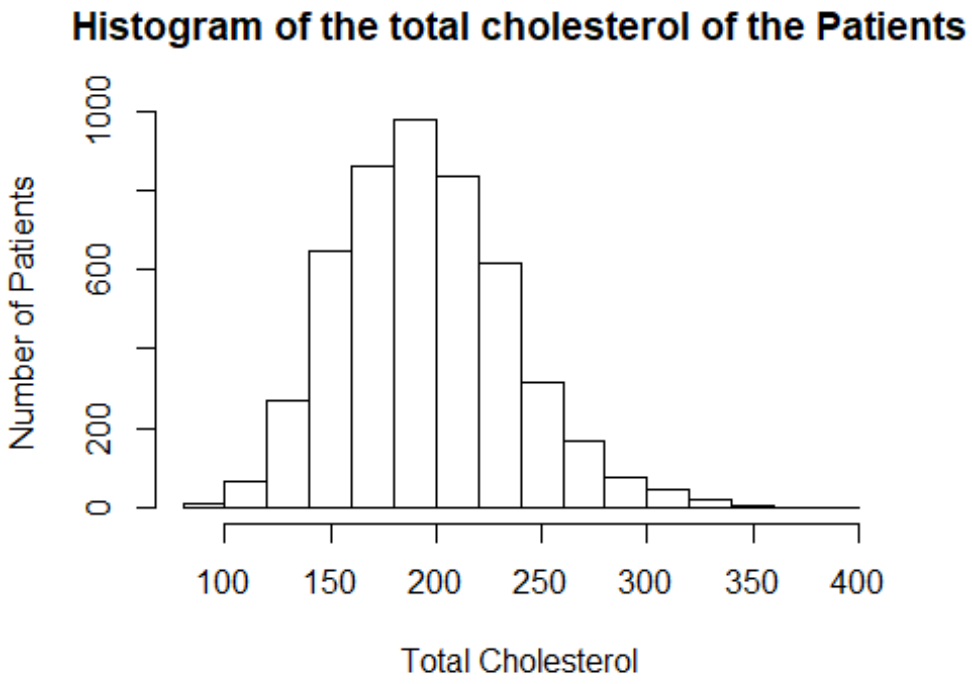
The pie chart below represents the categories of patients based on their marital status. It appears that half of the patients are already maried; the second majority of the patients never had a previous marriage; and the rest of the categories had an even distribution of the frequency of of the amount of patients per group.




```
## NULL
```

(i)

This histogram shows the total amount of cholesterol of all of the patients. We can see here that the distribution is skewed to the right, meaning that the majority of patients have a very high total cholesterol level over the normal total amount, which is 200 mg/dl. This can be concerning because this put them at a much higher risk of heart disease in the future. The distribution is also unimodal with a single peak of the range of 175-200 mg/dl with a frequency of 1000 patients.




```
## $breaks
##  [1]  80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400
##
## $counts
##  [1]   9  66 268 646 859 979 838 614 317 169  77  43  21   6   2   1
##
## $density
##  [1] 9.155646e-05 6.714140e-04 2.726348e-03 6.571719e-03 8.738555e-03
##  [6] 9.959308e-03 8.524924e-03 6.246185e-03 3.224822e-03 1.719227e-03
## [11] 7.833164e-04 4.374364e-04 2.136317e-04 6.103764e-05 2.034588e-05
## [16] 1.017294e-05
##
## $mids
##  [1]  90 110 130 150 170 190 210 230 250 270 290 310 330 350 370 390
##
## $xname
## [1] "patients$totalchol"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

III (Functions in R)

(a)

The fivenum() and quantile() functions both set a purpose on outputing the percentile of a value in the frequency of a dataset. The two differ in which fivenum() only gives the values of the minimum value, the 1st quartile, the median, the 3rd quartile, and the maximum value, whereas the quantile() function is more of an index of the set of data based on where you would imput a specific position on finding a specific value based on what percerntile (in this case it would be a decimal such as 0.25, 0.8, etc…) it is in.

(b)

I have created a function in which I would calculate the standard deviation of the data in an given dataset that is stored in the variable, x. I have created the varibale y in which this would store the equation in which I take the mean of X with the mean() over the square root of the variance of x since the variance is the value of the standard deviation, but it's squared as a conceptual measurement of how far the values of the data are from the mean.

```
R_function = f = function(x)
{
  y=(x-mean(x))/sqrt(var(x))
  return(sqrt(var(y)))
}
x=c(1:100)
f(x)
```

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

Appendex

```
patients <- read.csv("C:/Users/ugoch/Downloads/patients.csv")
patients_list_columns = names(patients)
patients_list_columns
patients_rows = dim(patients)
patients_rows
paitents_sum = summary(patients)
paitents_sum
age_mean = mean(patients$age)
age_mean
age_per_gender_mean = aggregate(age~gender, data = patients, mean)
age_per_gender_mean
bp_of_age_and_marital_staus_per_patient =
boxplot(patients$age~patients$marriage, main = "Boxplot of the Age and
Marital Status of each Patient", xlab = "Marital Status of each Patient",
ylab = "Age of each Patient")
bp_of_age_and_marital_staus_per_patient
bp_of_height_and_weight_range_per_patient =
boxplot(patients$height~patients$obese, main = "Boxplot of the height and
range of weight of the Patients", xlab = "Weight renage of the Patients",
```

```
ylab = "Height of the Patients")
bp_of_height_and_weight_range_per_patient
pie_marital_status =  pie(table(patients$marriage))
pie_marital_status
hist_totalchol =  hist(patients$totalchol, main = "Histogram of the total
cholesterol of the Patients", xlab = "Total Cholesterol", ylab = "Number of
Patients")
hist_totalchol
R_function = f = function(x)
{
  y=(x-mean(x))/sqrt(var(x))
  return(sqrt(var(y)))
}
x=c(1:100)
f(x)
```