Homework1Question1

Exploratory Analysis

Null Hypothesis

- My null hypothesis here is that there is no bias towards poorer or larger percent African American populations for any voting method that is involved in undercounting.

Characterize any undercounting between the equipment machines peopel used for voting in the 2000 election in Georgia

- Initially what I did was plot the difference between number of ballot and number of votes for each county (undercount value) in a box plot categorized by type of voting method used
- The output is below, you can see that while there may be a difference between punch and the other methods, you cannot really tell

```
rawdata = read.csv('georgia2000.csv', header=TRUE)
```

```
boxplot((ballots-votes)~equip,rawdata)
```

- Because of the outliers making the original bar graph hard to understand, I instead created a box plot of the log of the undercounting number per voting method used, that output is below
- You can see clearer now that there is infact a difference not only for punch, but optical voting methods
    - This was unobservable in the original chart

```
boxplot(log(ballots-votes)~equip,rawdata)
```

Consider African American and Poorer populations to evaluate any relationship to voting methods we flagged

Poorer Population

- To evaluate if there was any funny business going on for the poorer population I wanted to see if the poorer populations tended to be using the two voting methods with typcally higher undercounting
    - This would have suggested that their votes may have been affected by the undercounting
- To accomplish this visualization, I plotted a bar graph of the count of poor vs count of non-poor counties and subclassified them by their voting methods

```
boxplot(log(ballots-votes)~equip,rawdata)
```

African American Population

- To evaluate if there was any abnormality with regards to the voting method discrepency in relation to the African American population, I again wanted to see if there appeared to be a bias of higher percent African American populations using the methods flagged earlier
- Because the perAA variable was not categorical like the binary ‘poor’ variable, I had to consider histograms instead
- The below graph shows the frequency of perAA binned into deciles and graphed for each voting method individually

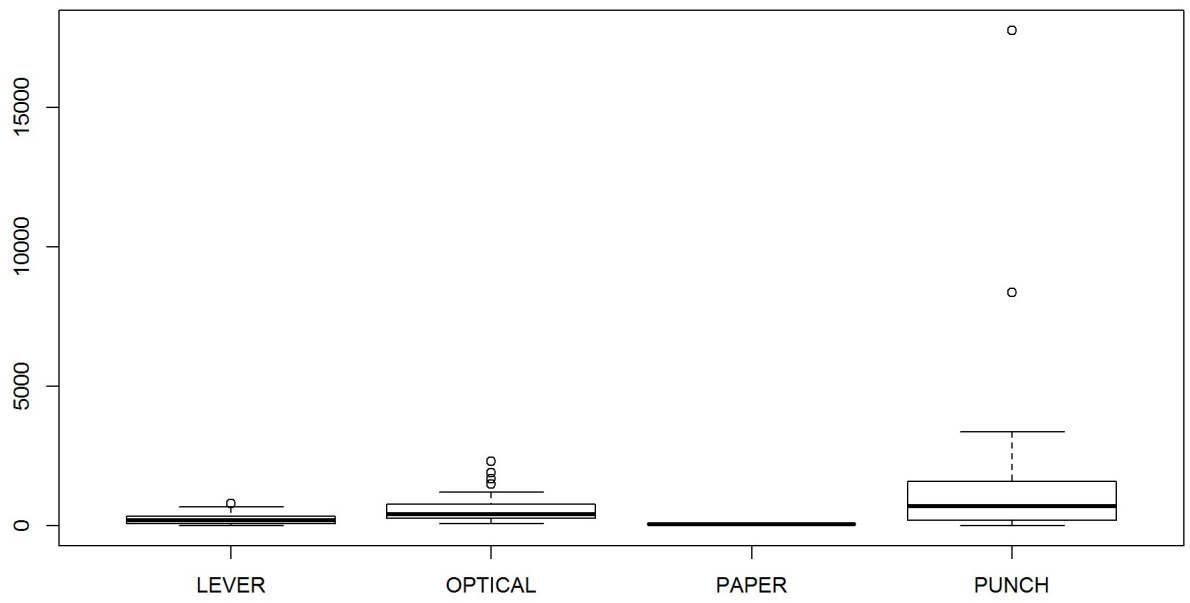
```
boxplot(log(ballots-votes)~equip,rawdata)
```

Conclusion

- To reject my null hypothesis I would have to consider the idea that there was a bias towards Punch or Optical voting methods for either the poor or higher percent African American populations
- From my graphs above I conclude that I cannot reject the null hypothesis
    - The poorer population tends to be biased towards the “lever” voting method and while their next largest frequency method is optical, the non-poor population has a significantly larger share of that method.
    - The higher percent African American population tends to be also biased towards the “lever” voting method. The histogram shows us that sub 40% African American populations predominantly use the optical method. The punch method has very low frequency but it is also mostly concentrated below 50% African American.

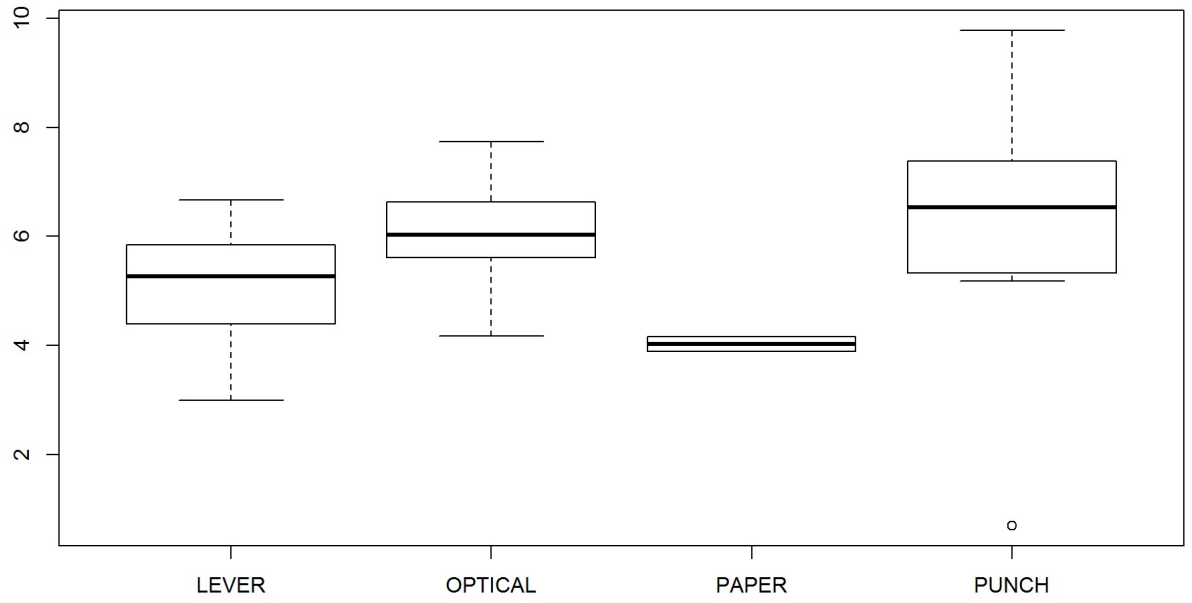
```
library(ggplot2)
set.seed(1)
rawdata = read.csv('georgia2000.csv', header=TRUE)

#To create the box plot for equipment vs undercounted votes
boxplot((ballots-votes)~equip,rawdata)
```




```
#To create the box plot for equipment vs undercounted votes (log form)
boxplot(log(ballots-votes)~equip,rawdata)
```

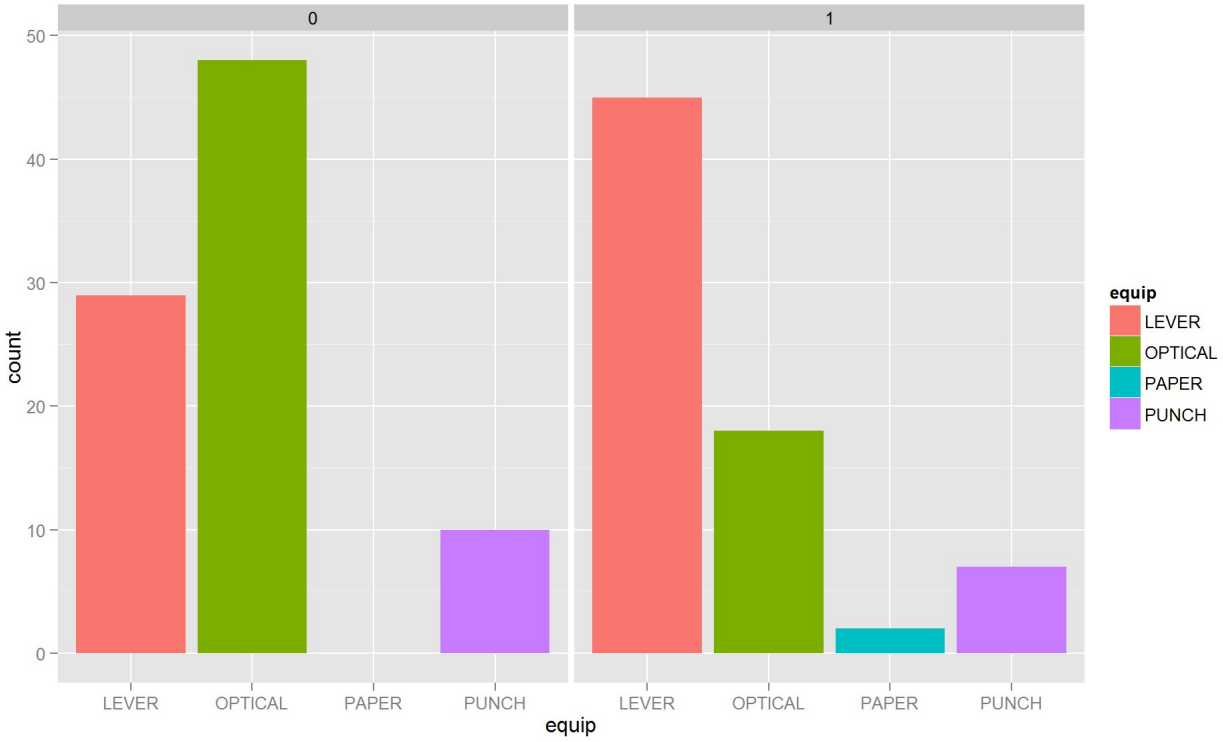
```
## Warning in bplt(at[i], wid = width[i], stats = z$stats[, i], out = z$out[z
## $group == : Outlier (-Inf) in boxplot 1 is not drawn
```




```
#To initially investigate the impact of poor and equip on the number of votes
xtabs(votes~poor+equip, data=rawdata)
```

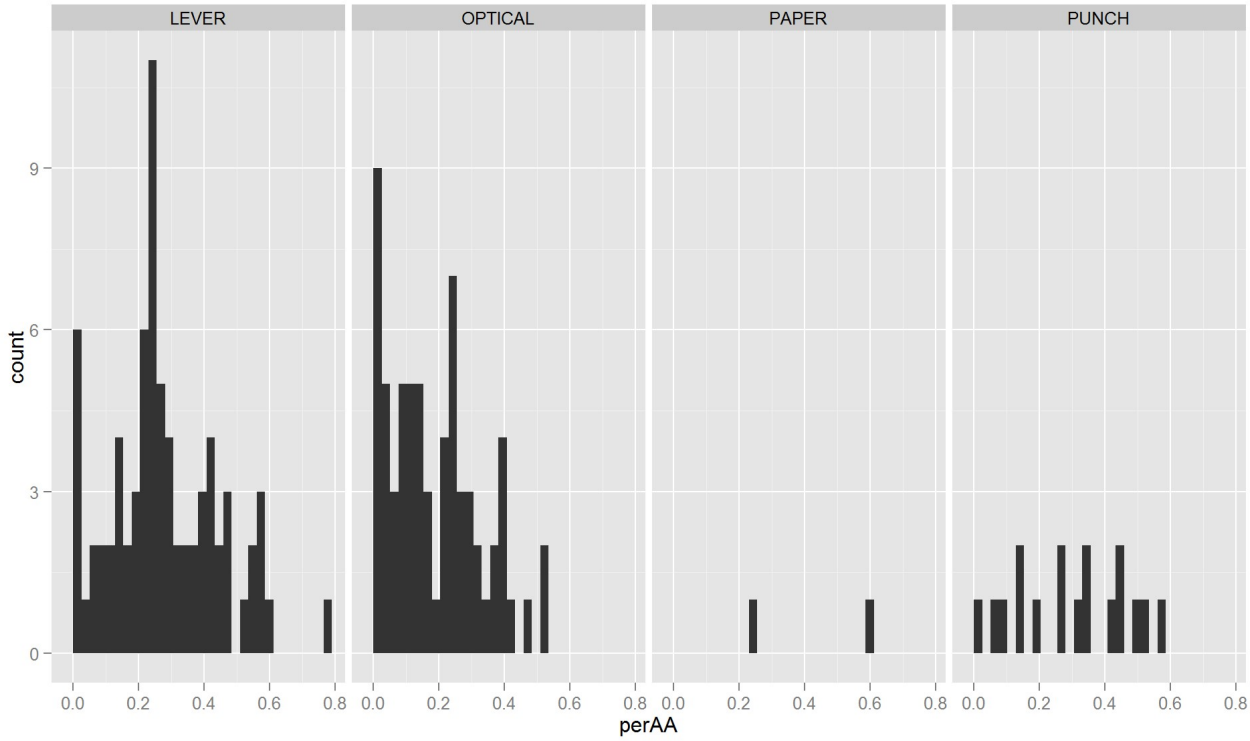
```
##      equip
## poor   LEVER OPTICAL   PAPER   PUNCH
##    0  201710 1290061       0  763276
##    1  209054  107008    3341   22183
```

```
#Bar chart of the different equipment types used for poor and not poor
qplot(equip, data=rawdata, geom="bar", fill=equip) + facet_wrap(~ poor, ncol =
5)
```




```
#histogram of the percent African American to each type of voting equipment
ggplot(data=rawdata, aes(perAA)) + geom_histogram() + facet_wrap(~ equip, ncol
= 5)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```


LEVER
OPTICAL
PAPER
PUNCH
count
9
6
3
0
0.0
0.2
0.4
0.6
0.8
perAA