Final project

Alessandro Maiuolo, Robert Fulton, Aria Coalson, Maria Croom

December 16, 2018

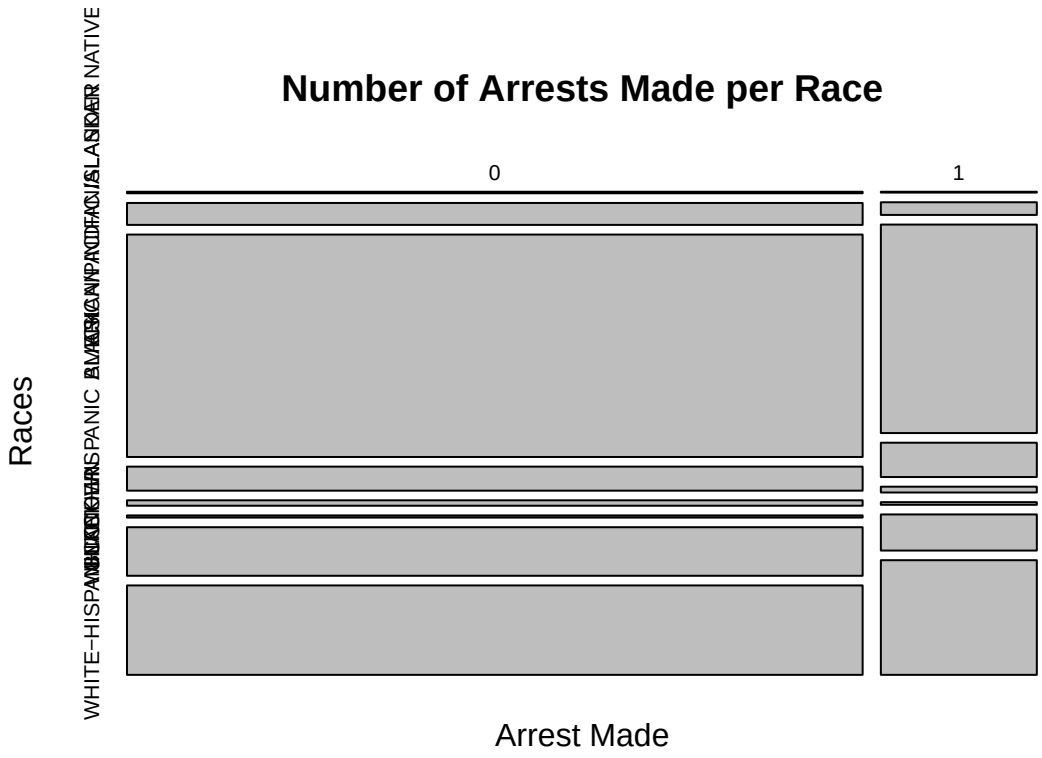
```
download.file('http://math.hmc.edu/m35f/2010_sqf_m35.csv', '2010_sqf_m35.csv')
download.file('http://math.hmc.edu/m35f/2015_sqf_m35.csv', '2015_sqf_m35.csv')

sqf2010 = read.csv('2010_sqf_m35.csv')
sqf2015 = read.csv('2015_sqf_m35.csv')

sqf2015 = sqf2015[!sqf2015$perstop=="**" & !sqf2015$perstop==" ",]
sqf2015$perstop = as.numeric(as.character(sqf2015$perstop))
```

Mosaic plot to visualize number of arrests made per race from the data

```
mosaicplot(table(sqf2015$arstmade, sqf2015$race),
           xlab="Arrest Made", ylab="Races",
           main="Number of Arrests Made per Race")
```



move white on top so when taking the binomial regression we use WHITE as the baseline

```
sqf2015$race <- relevel(sqf2015$race, "WHITE")
```

Takes a binomial regression that looks at how race effects the arrests made using WHITE arrest made as a baseline

```
mod <- glm(arstmade ~ race, data=sqf2015, family="binomial")
summary(mod)
```

```
##
## Call:
## glm(formula = arstmade ~ race, family = "binomial", data = sqf2015)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7274  -0.6021  -0.6021  -0.5411   2.1281
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -1.84721    0.05818 -31.749  < 2e-16
## raceAMERICAN INDIAN/ALASKAN NATIVE -0.30745    0.37797  -0.813 0.415970
## raceASIAN/PACIFIC ISLANDER         -0.25387    0.11286  -2.249 0.024484
## raceBLACK                           0.23131    0.06318   3.661 0.000251
## raceBLACK-HISPANIC                  0.65282    0.08599   7.592 3.15e-14
## raceOTHER                           0.36561    0.16031   2.281 0.022570
## raceUNKNOWN                         0.49138    0.23172   2.121 0.033959
## raceWHITE-HISPANIC                  0.54442    0.06750   8.066 7.29e-16
##
## (Intercept)                       ***
## raceAMERICAN INDIAN/ALASKAN NATIVE
## raceASIAN/PACIFIC ISLANDER        *
## raceBLACK                         ***
## raceBLACK-HISPANIC                ***
## raceOTHER                         *
## raceUNKNOWN                       *
## raceWHITE-HISPANIC                ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 20864  on 22501  degrees of freedom
## Residual deviance: 20708  on 22494  degrees of freedom
## AIC: 20724
##
## Number of Fisher Scoring iterations: 4
```

after fitting the model with regression

```
data <- data.frame(race=as.factor(c("WHITE","BLACK","BLACK-HISPANIC","WHITE-HISPANIC","OTHER","UNKNOWN"
```

predictive probability of the above races getting arrested

```
predict(mod, newdata = data, type = "response")
```

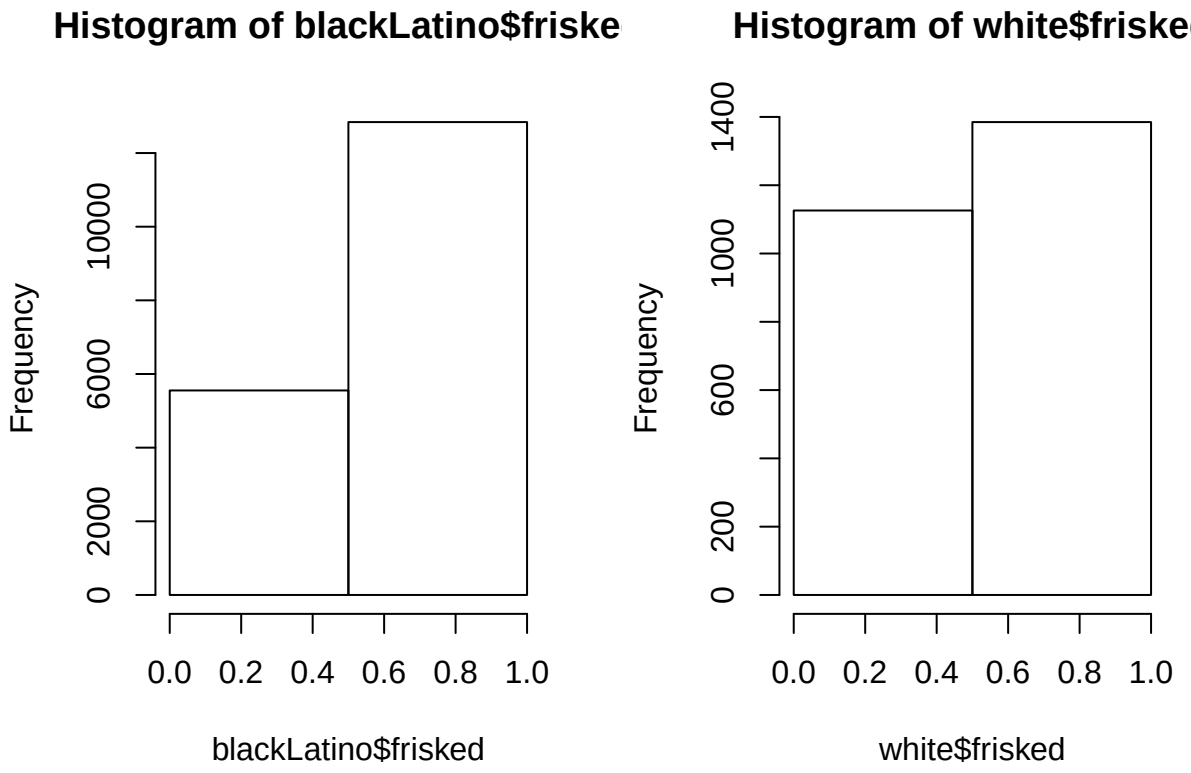
```
##         1         2         3         4         5         6
## 0.1362007 0.1657714 0.2324750 0.2136954 0.1851852 0.2049180
```

Is the proportion of black and latino individuals getting frisked different than the proportion of white individuals getting frisked?

A quick look shows that black and latino people might having more likelihood of being frisked

```r
blackLatino <- subset(sqf2015, sqf2015$race=="BLACK"|sqf2015$race=="WHITE-HISPANIC"|sqf2015$race=="BLACK
white <- subset(sqf2015, sqf2015$race=="WHITE")

par(mfrow = c(1,2))
hist(blackLatino$frisked, breaks = 2)
hist(white$frisked, breaks = 2)
```




```r
x1 = sum(blackLatino$frisked)
n1 = length(blackLatino$frisked)
x2 = sum(white$frisked)
n2 = length(white$frisked)

prop.test(x=c(x1,x2), n=c(n1,n2),  alternative="two.sided", conf.level=0.99)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(x1, x2) out of c(n1, n2)
## X-squared = 217.53, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  0.119298 0.173772
## sample estimates:
##    prop 1    prop 2
## 0.6981081 0.5515731
```

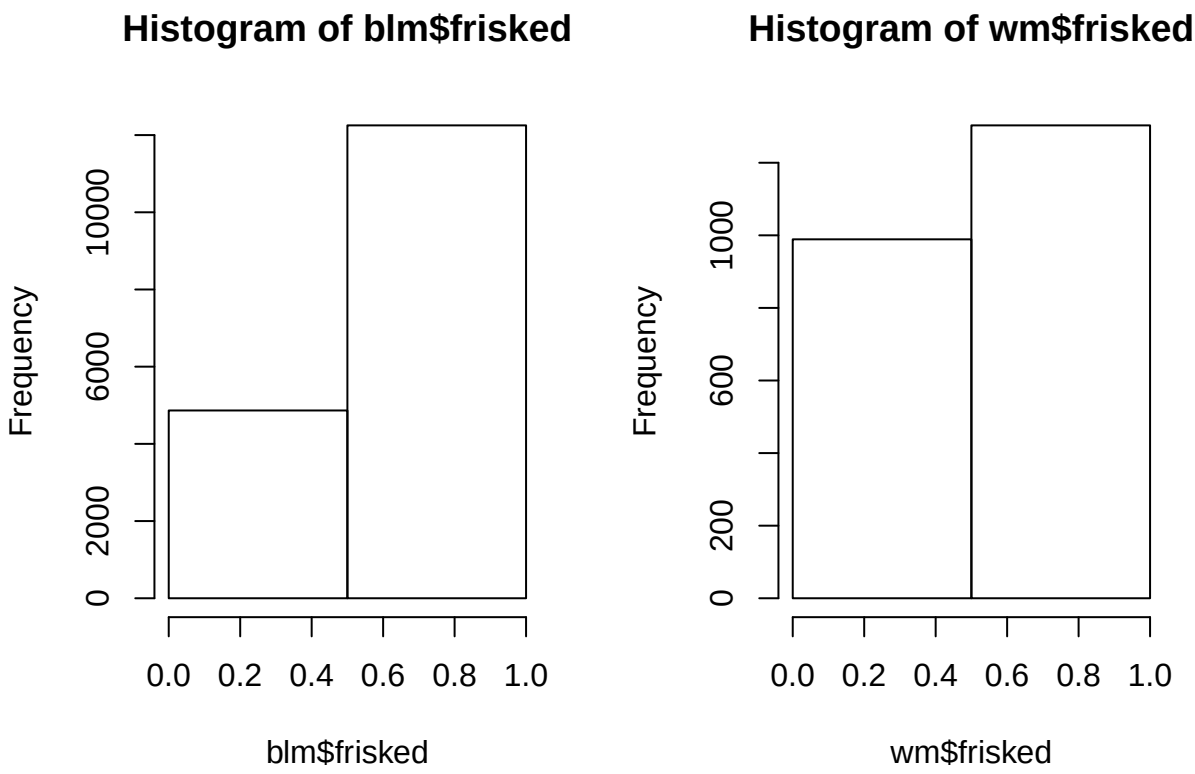
The p-value allows us to reject that the two proportions are the same. (p-value < 2.2e-16) The confidence interval suggests that the blackLatino proportion is higher. (0.119298 0.173772)

Is the proportion of black and latino men getting frisked different than the proportion of white men getting frisked?

```
blackLatino <- subset(sqf2015, sqf2015$race=="BLACK"|sqf2015$race=="WHITE-HISPANIC"|sqf2015$race=="BLACK
white <- subset(sqf2015, sqf2015$race=="WHITE")
blm <- subset(blackLatino, blackLatino$sex=="M") # black and latino men
wm <- subset(white, white$sex=="M") # white men
```

A quick look shows that black and latino men might having more likelihood of being frisked.

```
par(mfrow = c(1,2))
hist(blm$frisked, breaks = 2)
hist(wm$frisked, breaks = 2)
```




```
x1 = sum(blm$frisked)
n1 = length(blm$frisked)
x2 = sum(wm$frisked)
n2 = length(wm$frisked)

prop.test(x=c(x1,x2), n=c(n1,n2),  alternative="two.sided", conf.level=0.99)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(x1, x2) out of c(n1, n2)
```

```
## X-squared = 207.45, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  0.1189444 0.1756166
## sample estimates:
##    prop 1    prop 2
## 0.7157796 0.5684991
```

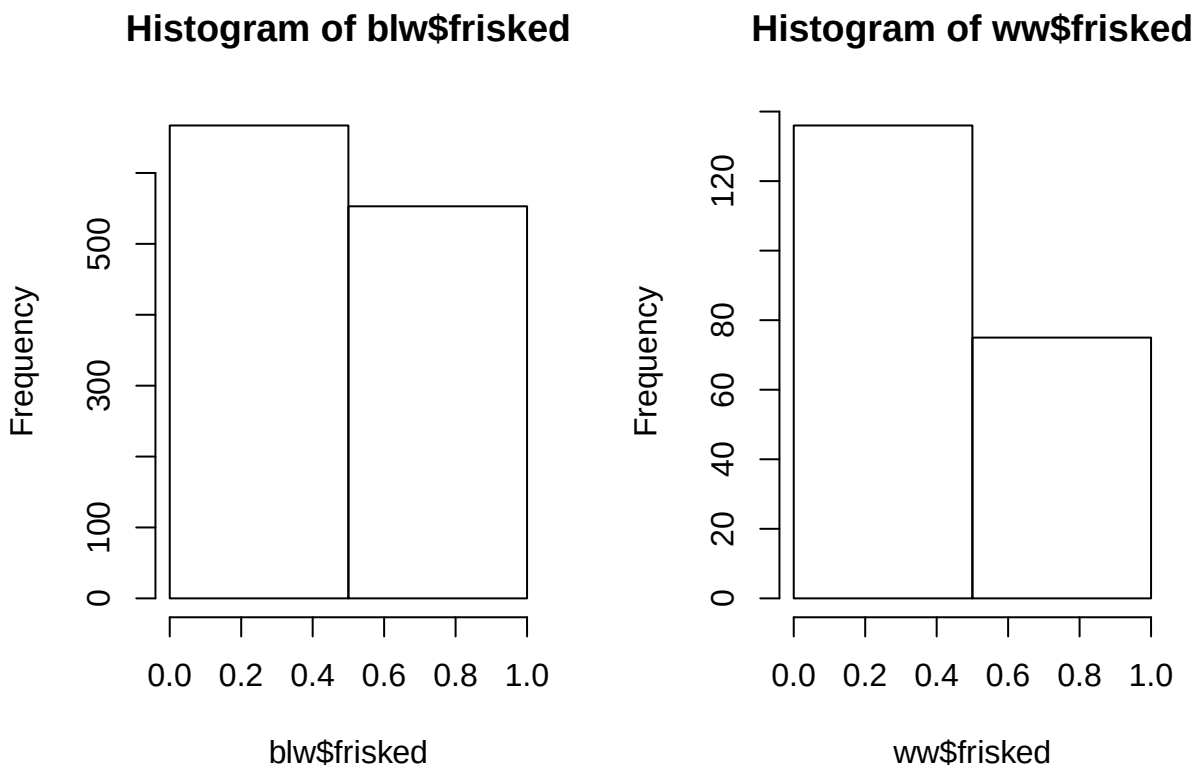
The p-value allows us to reject that the two proportions are the same. (p-value < 2.2e-16) The confidence interval suggests that the blackLatino proportion is higher. (0.1189444 0.1756166)

Is the proportion of black and latino women getting frisked different than the proportion of white women getting frisked?

```
blackLatino <- subset(sqf2015, sqf2015$race=="BLACK"|sqf2015$race=="WHITE-HISPANIC"|sqf2015$race=="BLACK
white <- subset(sqf2015, sqf2015$race=="WHITE")
blw <- subset(blackLatino, blackLatino$sex=="F") # black and latino women
ww <- subset(white, white$sex=="F") # white women
```

A quick look shows that black and latino women might having more likelihood of being frisked

```
par(mfrow = c(1,2))
hist(blw$frisked, breaks = 2)
hist(ww$frisked, breaks = 2)
```




```
x1 = sum(blw$frisked)
n1 = length(blw$frisked)
x2 = sum(ww$frisked)
n2 = length(ww$frisked)
```

```r
prop.test(x=c(x1,x2), n=c(n1,n2),  alternative="two.sided", conf.level=0.98)
```

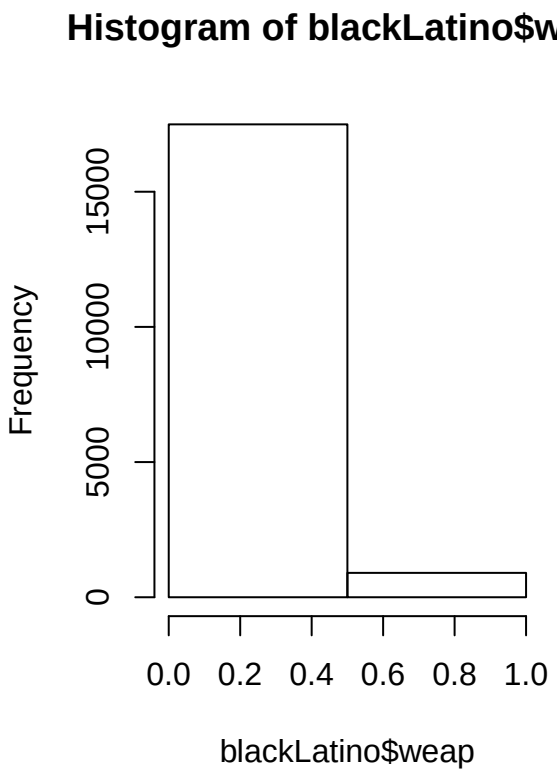
```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(x1, x2) out of c(n1, n2)
## X-squared = 6.5994, df = 1, p-value = 0.0102
## alternative hypothesis: two.sided
## 98 percent confidence interval:
##  0.01152897 0.18412793
## sample estimates:
##    prop 1    prop 2
## 0.4532787 0.3554502
```

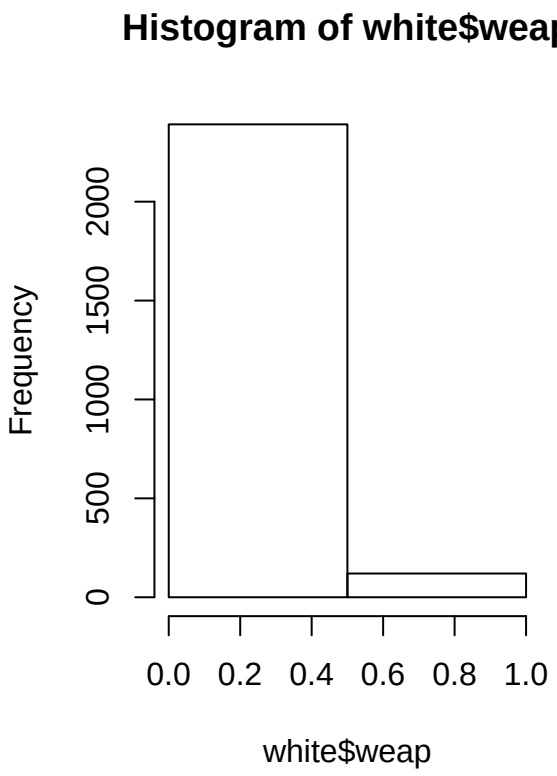
The p-value allows us to reject that the two proportions are the same. (p-value = .0102). The confidence interval suggests that the blackLatino proportion is higher. (0.01152897 0.18412793)

Is the proportion of black and latino individuals having weapons different than the proportion of white individuals having weapons?

A quick look shows that the proportion of blackLatino and white are essentially the same.

```r
par(mfrow = c(1,2))
hist(blackLatino$weap, breaks = 2)
hist(white$weap, breaks = 2)
```






```
blackLatino <- subset(sqf2015, sqf2015$race=="BLACK"|sqf2015$race=="WHITE-HISPANIC"|sqf2015$race=="BLACI
white <- subset(sqf2015, sqf2015$race=="WHITE")

x1 = sum(blackLatino$weap)
n1 = length(blackLatino$weap)
x2 = sum(white$weap)
n2 = length(white$weap)

prop.test(x=c(x1,x2), n=c(n1,n2),  alternative="two.sided", conf.level=0.99)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(x1, x2) out of c(n1, n2)
## X-squared = 0.0496, df = 1, p-value = 0.8238
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  -0.01068568  0.01318169
## sample estimates:
##     prop 1     prop 2
## 0.04903773 0.04778973
```

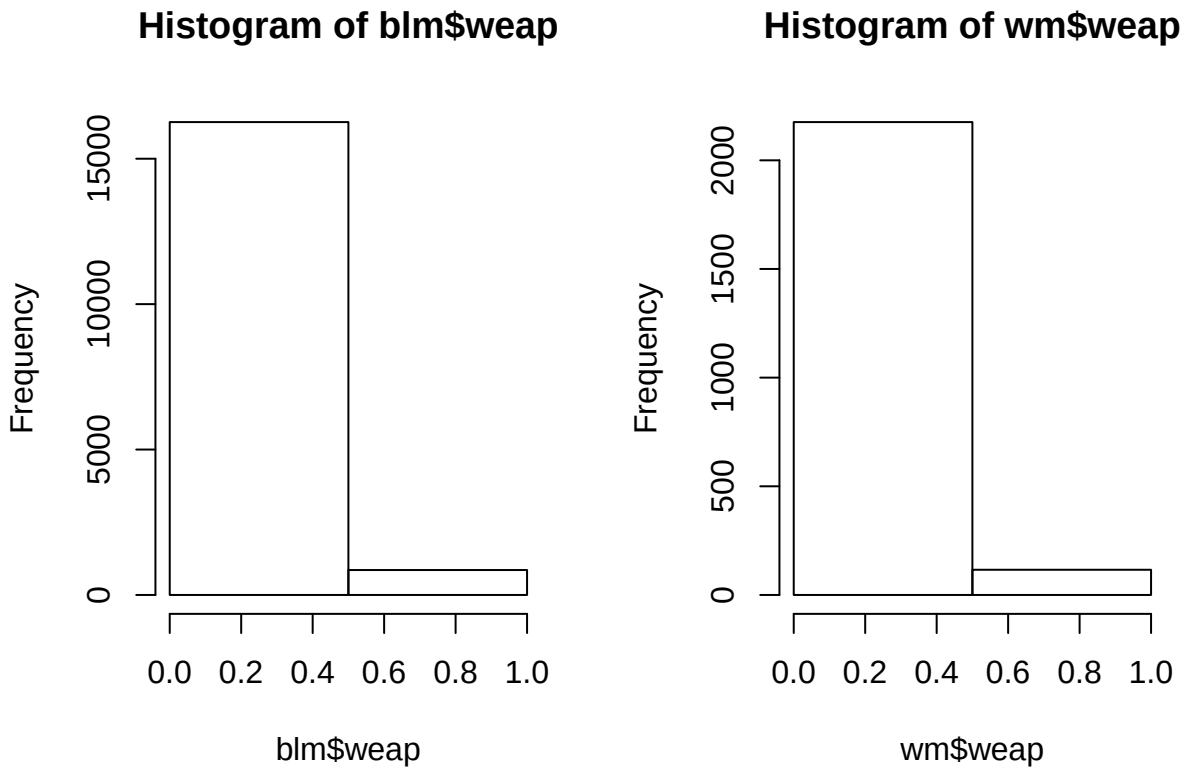
With the p-value we calculated, we fail to reject that the two proportions are the same. (p-value = 0.8238)
The confidence interval reinforces the failed rejection. (-0.01068568 0.01318169)

Is the proportion of black and latino men having a weapon different than the proportion of white men having a weapon?

```
blackLatino <- subset(sqf2015, sqf2015$race=="BLACK"|sqf2015$race=="WHITE-HISPANIC"|sqf2015$race=="BLACK
white <- subset(sqf2015, sqf2015$race=="WHITE")
blm <- subset(blackLatino, blackLatino$sex=="M") # black and latino men
wm <- subset(white, white$sex=="M") # white men
```

A quick look shows that the porportions are essentially the same

```
par(mfrow = c(1,2))
hist(blm$weap, breaks = 2)
hist(wm$weap, breaks = 2)
```




```
x1 = sum(blm$weap)
n1 = length(blm$weap)
x2 = sum(wm$weap)
n2 = length(wm$weap)

prop.test(x=c(x1,x2), n=c(n1,n2),  alternative="two.sided", conf.level=0.99)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(x1, x2) out of c(n1, n2)
```

```
## X-squared = 0.0037261, df = 1, p-value = 0.9513
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  -0.01334205  0.01225478
## sample estimates:
##     prop 1     prop 2
## 0.05006718 0.05061082
```

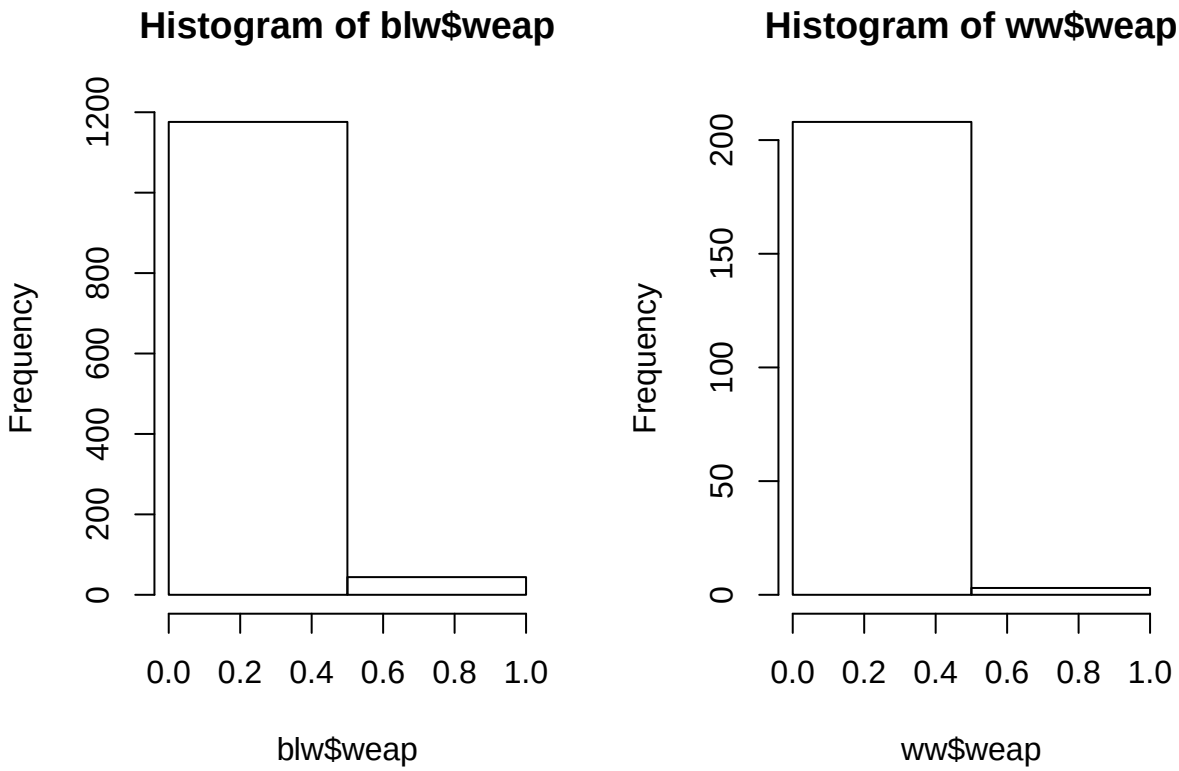
With the p-value we calculated, we fail to reject that the two proportions are the same. (p-value = 0.9513) The confidence interval reinforces the failed rejection. (-0.01334205 0.01225478)

Is the proportion of black and latino women having a weapon different than the proportion of white women having a weapon?

```
blackLatino <- subset(sqf2015, sqf2015$race=="BLACK"|sqf2015$race=="WHITE-HISPANIC"|sqf2015$race=="BLACK
white <- subset(sqf2015, sqf2015$race=="WHITE")
blw <- subset(blackLatino, blackLatino$sex=="F") # black and latino women
ww <- subset(white, white$sex=="F") # white women
```

A quick look shows that black and latino women might have a higher proportion

```
par(mfrow = c(1,2))
hist(blw$weap, breaks = 2)
hist(ww$weap, breaks = 2)
```



```r
x1 = sum(blw$weap)
n1 = length(blw$weap)
x2 = sum(ww$weap)
n2 = length(ww$weap)

prop.test(x=c(x1,x2), n=c(n1,n2),  alternative="two.sided", conf.level=0.99)
```

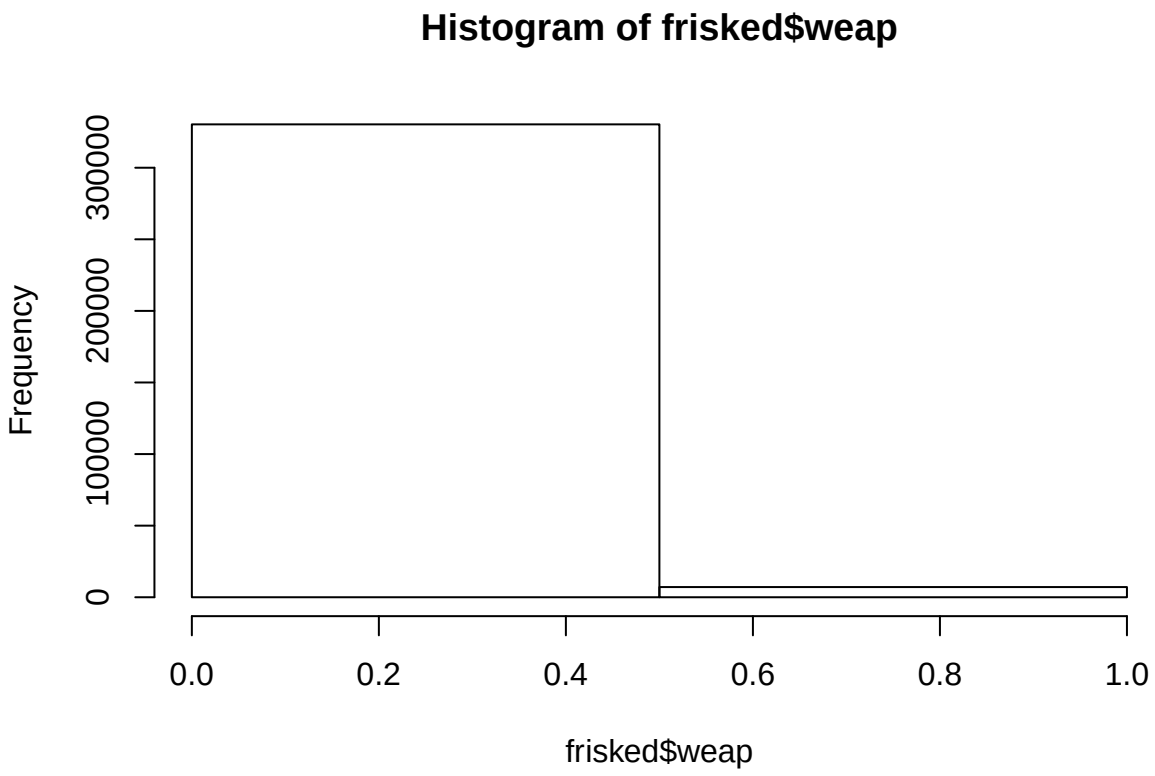
```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(x1, x2) out of c(n1, n2)
## X-squared = 2.059, df = 1, p-value = 0.1513
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  -0.006027662  0.049722791
## sample estimates:
##     prop 1     prop 2
## 0.03606557 0.01421801
```

With the p-value we calculated, we fail to reject that the two proportions are the same. (p-value = .1513) The confidence interval reinforces the failed rejection. (-0.006027662 0.049722791)

Frisk success

```r
frisked <- subset(sqf2010, sqf2010$frisked==1)

par(mfrow = c(1,1))
hist(frisked$weap,breaks = 2) # shows low success rate in frisks
```




```
x1 = sum(frisked$weap)
n1 = length(frisked$weap)
x2 = length(frisked$weap)/50 # article claimed 2%
n2 = length(frisked$weap)

prop.test(x=c(x1,x2), n=c(n1,n2),  alternative="two.sided", conf.level=0.99)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(x1, x2) out of c(n1, n2)
## X-squared = 9.7, df = 1, p-value = 0.001843
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  0.0001860155 0.0019709112
## sample estimates:
##     prop 1     prop 2
## 0.02107846 0.02000000
```

We reject the article's claim (p-value = 0.001843).
Confidence interval suggests that the success rate is higher than 2%. (0.0001860155 0.0019709112)

Compare the proportion of black and latino males stopped to the 41% cited in the report (hypothesis test)

Subsetting the black and "latino" or hispanic populations:

```
black_or_hisp <- subset(sqf2010, sqf2010$race == "BLACK" | sqf2010$race == "BLACK-HISPANIC" | sqf2010$ra
white <- subset(sqf2010, sqf2010$race == "WHITE")
```

Getting male subsets of blacks and hispanics:

```
male_borh <- subset(black_or_hisp, black_or_hisp$sex == "M")
male_white <- subset(white, white$sex == "M")
```

Subsetting by age:

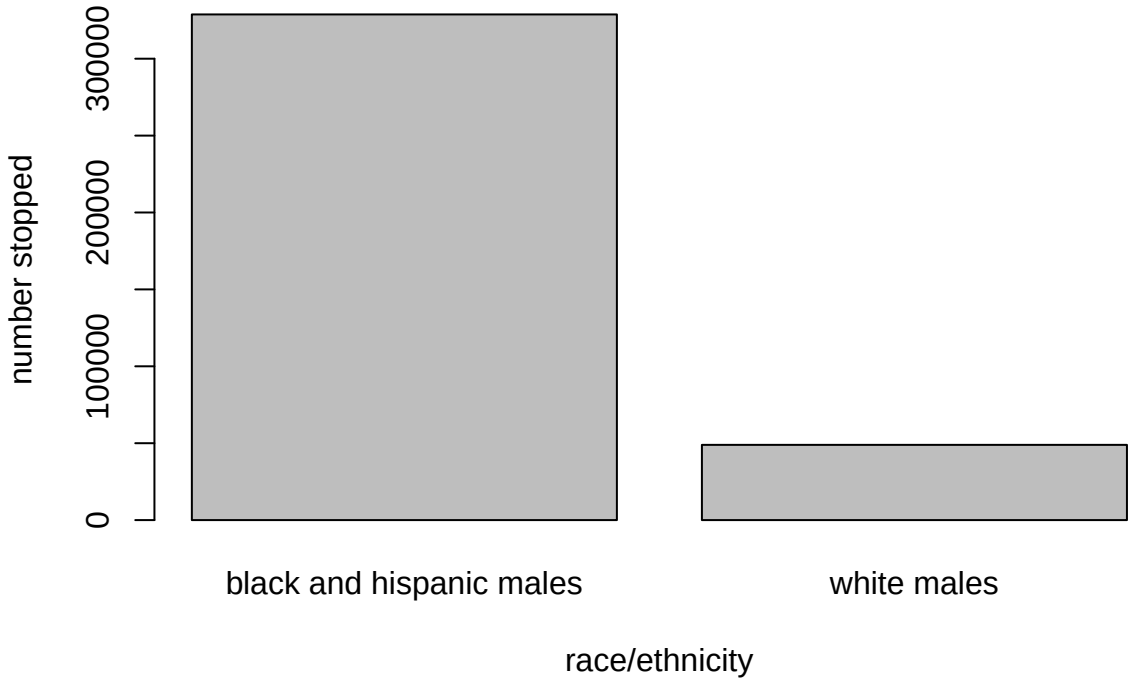
```
m_bh_lt24 = subset(male_borh, male_borh$age < 25)
m_bh_15to24 <- subset(m_bh_lt24, m_bh_lt24$age > 13)
```

Finding proportions:

```
x <- dim(m_bh_15to24)[1]
n = dim(sqf2010)[1]
prop.test(x, n, p = 0.41, alternative = "two.sided", conf.level = 0.95)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  x out of n, null probability 0.41
## X-squared = 46078, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.41
## 95 percent confidence interval:
##  0.2727203 0.2749763
## sample estimates:
##         p
## 0.2738468
```

```
barplot(c(dim(male_borh)[1], dim(male_white)[1]), names.arg = c("black and hispanic males", "white males
```



We reject the null - p-value = 2.2 *10^-16 The 95% percent confidence interval is that 77.83 to 78.04 percent of the people who were stopped were either black males, hispanic males, or both.
This is almost double the proportion (0.41) predicted by the article!