Final Project Part B

Keizo Morgan, Max Holloway, Bowen Jiang, Jorge Canedo

Initialize Data

```
download.file('http://math.hmc.edu/m35f/2010_sqf_m35.csv',
              '2010_sqf_m35.csv')
sqf2010 <- read.csv("2010_sqf_m35.csv")

sqf2010 <- subset(sqf2010, sqf2010$weight > 50 & sqf2010$weight < 400)
sqf2010 <- subset(sqf2010, sqf2010$age < 100)
```

Problem 1

```
v <- t.test(sqf2010$perobs[sqf2010$haircolr == "BLACK"],
       sqf2010$perobs[sqf2010$haircolr == "BLOND"],
       alternative = "two.sided",
       mu = 0,
       paired = FALSE,
       var.equal = TRUE,
       conf.level = 0.95)
print(v)
```

```
##
##  Two Sample t-test
##
## data:  sqf2010$perobs[sqf2010$haircolr == "BLACK"] and sqf2010$perobs[sqf2010$haircolr == "BLOND"]
## t = -3.3831, df = 457950, p-value = 0.0007169
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.33803440 -0.09003401
## sample estimates:
## mean of x mean of y
##  2.451975  2.666010
```

The null hypothesis is that there is no difference between the period of observation in those with black hair and those with blond hair ($\mu_{black} - \mu_{blond} = 0$). Therefore the null value is zero.

The alternative hypotheis is that there is a difference between the period of observation in those with black hair and those with blond hair ($\mu_{black} - \mu_{blond} \neq 0$).

The alpha value is 1 minus the confidence level of 95%, so it is 0.05. We reject the null hypothesis if the p-value is less than or equal to alpha, and we fail to reject the null hypothesis if the p-value is greater than alpha.

From the t-test we have a p-value of 7.611143210^{-4}, which is less than the alpha of 0.05, therefore we reject the null hypothesis and conclude that there is a difference between the period of observation in those with black hair and those with blond hair.

Furthermore, we are 95% confident that the true difference of period of observation in those with black hair and those with blond hair is between -0.3380344 and -0.090034.

Problem 2

Code for running the t-test

```
t.test(sqf2010$perobs[sqf2010$sex=="M"],
       sqf2010$perobs[sqf2010$sex=="F"],
       alternative="two.sided",
       mu=0, paired=FALSE, var.equal=TRUE,
       conf.level=0.95)
```

```
##
##  Two Sample t-test
##
## data:  sqf2010$perobs[sqf2010$sex == "M"] and sqf2010$perobs[sqf2010$sex == "F"]
## t = -5.3342, df = 588970, p-value = 9.6e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.19925487 -0.09217391
## sample estimates:
## mean of x mean of y
##  2.466840  2.612554
```

Analysis

The null hypothesis is $H_0 : \mu_1 - \mu_2 = 0$, and the alternate hypothesis is $H_1 : \mu_1 - \mu_2 \neq 0$; the null value is 0; the significance level is 95%; we reject the null hypothesis with 95% confidence, because the 95% confidence interval does not contain 0; the p-value is $9.6 * 10^-8$, which is much less than 0.05, so we are very confident that the population means are not the same; the confidence interval is $-0.19925487 < \mu_1 - \mu_2 < -0.09217391$; this means that we can say with 95% confidence that the difference of means is $\neq 0$, meaning that one population has a larger population than another population with 95% confidence.

Problem 3

Code for the proportion test

```
sqf2010.c <- sqf2010[sqf2010$contrabn==1, ]
x1 <- sum(sqf2010.c$arstmade)
n1 <- length(sqf2010.c$arstmade)

sqf2010.nc <- sqf2010[sqf2010$contrabn==0, ]
x2 <- sum(sqf2010.nc$arstmade)
n2 <- length(sqf2010.nc$arstmade)

prop.test(x=c(x1, x2), n=c(n1, n2), alternative="two.sided",
          conf.level=0.95)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(x1, x2) out of c(n1, n2)
## X-squared = 111580, df = 1, p-value < 2.2e-16
```

```
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.7813279 0.7947289
## sample estimates:
##     prop 1     prop 2
## 0.84104145 0.05301302
```

Analysis

The proportion of people carrying contraban that were arrested is 0.84104145; the proportion of people who were not carrying contraban that were arrested is 0.05301302; $H_0 : p_1 = p_2$; $H_1 : p_1 \neq p_2$; the p-value is $2.2 * 10^{-16}$, showing that is extremely unlikely that the proportions are the same (we reject the null hypothesis that claims the proportions are the same).

Problem 4

Frisked

```
sqf2010.male <- sqf2010[sqf2010$sex == "M",]
sqf2010.female <- sqf2010[sqf2010$sex == "F",]
x_male_frisked <- sum(sqf2010.male$frisked)
n_male_frisked <- length(sqf2010.male$frisked)
x_female_frisked <- sum(sqf2010.female$frisked)
n_female_frisked <- length(sqf2010.female$frisked)

v_frisked <- prop.test(x = c(x_male_frisked, x_female_frisked),
                       n = c(n_male_frisked, n_female_frisked),
                       alternative = "two.sided",
                       conf.level = 0.95)

print(v_frisked)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(x_male_frisked, x_female_frisked) out of c(n_male_frisked, n_female_frisked)
## X-squared = 14419, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.2991538 0.3082176
## sample estimates:
##    prop 1    prop 2
## 0.5838085 0.2801227
```

The null hypothesis is that there is no difference in the proportion of males that were frisked and females that were frisked ($p_{male,frisked} - p_{female,frisked} = 0$).

The alternative hypothesis is that there is a difference in the proportion of males that were frisked and females that were frisked ($p_{male,frisked} - p_{female,frisked} \neq 0$).

From the two-sample proportion test for frisked, we see that 58.3808452% of males were frisked and 28.0122738% of females were frisked.

Furthermore, the p-value of this difference is close to zero, and is less than any alpha value given. Therefore we reject the null hypothesis and conclude that there is a difference in the proportion of males that were frisked and females that were frisked.

Arrested

```
x_male_arrested <- sum(sqf2010.male$arstmade)
n_male_arrested <- length(sqf2010.male$arstmade)
x_female_arrested <- sum(sqf2010.female$arstmade)
n_female_arrested <- length(sqf2010.female$arstmade)

v_arrested <- prop.test(x = c(x_male_arrested, x_female_arrested),
                        n = c(n_male_arrested, n_female_arrested),
                        alternative = "two.sided",
                        conf.level = 0.95)

print(v_arrested)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(x_male_arrested, x_female_arrested) out of c(n_male_arrested, n_female_arrested)
## X-squared = 530.59, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.0326244 -0.0267654
## sample estimates:
##     prop 1     prop 2
## 0.06651424 0.09620914
```

The null hypothesis is that there is no difference in the proportion of males that were arrested and females that were arrested ($p_{male,arrested} - p_{female,arrested} = 0$).

The alternative hypothesis is that there is a difference in the proportion of males that were arrested and females that were arrested ($p_{male,arrested} - p_{female,arrested} \neq 0$).

From the two-sample proportion test for arrested, we see that 6.6514239% of males were arrested and 9.6209138% of females were arrested.

Furthermore, the p-value of this difference is close to zero, and is less than any alpha value given. Therefore we reject the null hypothesis and conclude that there is a difference in the proportion of males that were arrested and females that were arrested.

Problem 5

T Test Analysis

Here we are given a t test of the difference of two binary variables (arstmade and contrabn). We use a two sided test with confidence level of 95%.

```
t.test(sqf2010$arstmade,
       sqf2010$contrabn,
       alternative = "two.sided",
```

```
       mu = 0, paired = TRUE,
       var.equal = TRUE,
       conf.level = 0.95)
```

```
##
##  Paired t-test
##
## data:  sqf2010$arstmade and sqf2010$contrabn
## t = 164.76, df = 598640, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.04829724 0.04946019
## sample estimates:
## mean of the differences
##              0.04887872
```

We then achieve a t value of 165.04 and a p value of 2.2e-16. This correlates to a 95 percent confidence interval of: 0.04826533 to 0.04942545. The outputted sample mean of the differences is 0.04884539. The implications of this sample mean of differences is that for the sample calculated the majority of stops resulted in no contraban or arrest with a slight skew towards an arrest made without contraban. A pval that small also implies that the null hypothesis of both binary variables being equal (hence correlated) is rejected.