Final Project Part A

Initialize data

```
if (!file.exists("2015_sqf_m35.csv")) {
  download.file("http://math.hmc.edu/m35f/2015_sqf_m35.csv", "2015_sqf_m35.csv")
}
sqf2015 <- read.csv("2015_sqf_m35.csv")

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

PROBLEM 1

Set helper functions

```
make_ci <- function(xs, alpha) {
  xs_mean <- mean(xs)
  xs_sd <- sd(xs)
  xs_n <- length(xs)
  xs_t <- abs( qt(alpha/2, xs_n - 1) )
  xs_me <- xs_t * xs_sd / sqrt(xs_n)
  return(c(xs_mean - xs_me, xs_mean + xs_me))
}

print_ci <- function(name, ci) {
  print(c(name, ci[1], ci[2]))
}
```

Print confidence intervals

```
# entire sample
print_ci("entire sample", make_ci(sqf2015$perobs, 0.05))
```

```
## [1] "entire sample"    "2.53118181752757" "2.71661837801061"
```

```
# each race
for (r in levels(sqf2015$race)) {
  race_subset <- subset(sqf2015, sqf2015$race == r)
  print_ci(r, make_ci(race_subset$perobs, 0.05))
}
```

```
## [1] "AMERICAN INDIAN/ALASKAN NATIVE" "2.10461868604585"
## [3] "3.60966702823986"
## [1] "ASIAN/PACIFIC ISLANDER" "2.76620021864205"
## [3] "3.30827752886022"
## [1] "BLACK"            "2.40977164311871" "2.69195741513209"
## [1] "BLACK-HISPANIC"   "2.05927253269893" "3.26261587931824"
## [1] "OTHER"            "2.32494116985702" "3.63465478973894"
```

```
## [1] "UNKNOWN"          "1.42794705937611" "2.93270867832881"
## [1] "WHITE"            "2.82094140811401" "3.29853290490869"
## [1] "WHITE-HISPANIC"   "2.3547670598937"  "2.57793660008268"
```

Problem 2

Set variables for the hypothesis tests

```
bhair <- subset(sqf2015, sqf2015$haircolr == "BLACK") # people with black hair
pstop <- bhair$perstop # period of stop for people with black hair
mu_0 <- 7.9 # proposed value for the population mean from the problem
n <- length(pstop)
xbar <- mean(pstop, na.rm = TRUE) # sample mean
sdp <- sd(pstop) # point estimate for standard deviation of population
```

Hypothesis Test: $\mu = \mu_0$

The null hypothesis is $H_0 : \mu = \mu_0$, and the alternate hypothesis is $H_1 : \mu > \mu_0$.

We will first compute the test statistic, z_{test}.

```
z_test <- (xbar-mu_0)/(sdp/sqrt(n)) # z-score for mu being the true average of the population
```

```
## [1] 3.178582
```

The critical value for us to reject the null hypothesis is found by the following code.

```
z_critical <- qnorm(0.95, mean=0, sd = 1)
```

```
## [1] 1.644854
```

Because $z_{test} > z_{critical}$, we can reject the null hypothesis with 95% confidence.

Additionally, we can compute the p-value with the following code.

```
p_value <- pnorm(z_test, mean=0, sd = 1, lower.tail=FALSE)
```

```
## [1] 0.000739986
```

Because this p-value is significantly smaller than 0.05, we get the same result as the hypothesis test: reject the null hypothesis.

Problem 3

Remove extraneous values and subset

```
sqf2015.clean <- subset(sqf2015, 0 < age & age < 100)
sqf2015.male <- subset(sqf2015.clean, sex == "M")
sqf2015.female <- subset(sqf2015.clean, sex == "F")
```

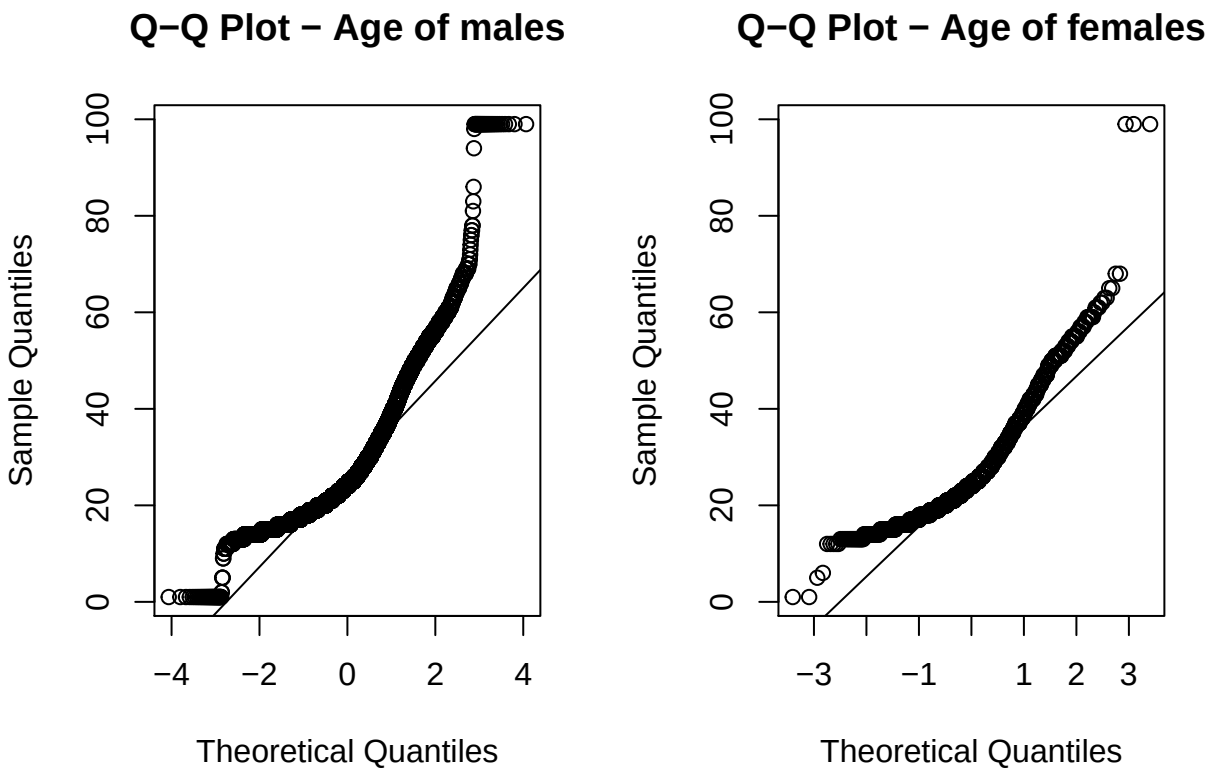
The data was filtered for age between 0 and 100, exclusive. Furthermore, male and female sexes were subsetted.

Plot

```
par(mfrow=c(1, 2))

qqnorm(sqf2015.male$age, main = "Q-Q Plot - Age of males")
qqline(sqf2015.male$age)

qqnorm(sqf2015.female$age, main = "Q-Q Plot - Age of females")
qqline(sqf2015.female$age)
```



We observe that age of males and females are skewed right are heavy tailed. The age of males seems to have more values clustered at the extremes 0 and 100 years.

Problem 4

Confidence Intervals for Proportions

The races are,

```
unique(sqf2015$race)
```

```
## [1] WHITE                          BLACK
## [3] WHITE-HISPANIC                 BLACK-HISPANIC
## [5] OTHER                          ASIAN/PACIFIC ISLANDER
## [7] AMERICAN INDIAN/ALASKAN NATIVE UNKNOWN
## 8 Levels: AMERICAN INDIAN/ALASKAN NATIVE ASIAN/PACIFIC ISLANDER ... WHITE-HISPANIC
```

White

```
data=sqf2015$frisked[sqf2015$race=="WHITE"]
data=data[!is.na(data)]
n=length(data)
p=mean(data)
upperbound=p+1.96*sqrt(p*(1-p)/n)
lowerbound=p-1.96*sqrt(p*(1-p)/n)
c(lowerbound,upperbound)
```

```
## [1] 0.5321204 0.5710258
```

Black

```
data=sqf2015$frisked[sqf2015$race=="BLACK"]
data=data[!is.na(data)]
n=length(data)
p=mean(data)
upperbound=p+1.96*sqrt(p*(1-p)/n)
lowerbound=p-1.96*sqrt(p*(1-p)/n)
c(lowerbound,upperbound)
```

```
## [1] 0.7035497 0.7198178
```

White-Hipanic

```
data=sqf2015$frisked[sqf2015$race=="WHITE-HISPANIC"]
data=data[!is.na(data)]
n=length(data)
p=mean(data)
upperbound=p+1.96*sqrt(p*(1-p)/n)
lowerbound=p-1.96*sqrt(p*(1-p)/n)
c(lowerbound,upperbound)
```

```
## [1] 0.6511231 0.6770941
```

Problem 5

Hypothesis Tests for Proportions part A

We will perform a hypothesis test for the proportions being equal, by testing $H_0 : p = p_0$ and $H_1 : p > p_0$ with a significance level of $\alpha = 0.05$.

```
n <- length(sqf2015$frisked)
p <- sum(sqf2015$frisked) / n # sample proportion of people frisked
po <- 0.675 # proposed population proportion of people frisked

z_test <- (p - po)/(sqrt(p*(1-p)/n)) # calculate test statistic
print(z_test)
```

```
## [1] 0.201498
```

```
z_critical <- 1.645 # critical value for 95% confidence of p > po

reject_H0 <- z_test > z_critical
print(reject_H0)
```

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

Because the test statistic is less than the critical value, we fail to reject the null hypothesis. Hence we cannot conclude with 95% confidence that the population proportion of people being frisked is greater than 0.675.

Hypothesis Tests for Proportions part B

We will perform a t test to determine if the people who refused to show ID were frisked the same amount as people who did not refuse to show ID.

```
sqf2015friskednoid = sqf2015$frisked[sqf2015$typeofid=="REFUSED"]
sqf2015friskedid = sqf2015$frisked[sqf2015$typeofid!="REFUSED"]
t.test(x=sqf2015friskedid,
       y=sqf2015friskednoid,
       alternative = "two.sided",
       mu=0, paired = FALSE,
       var.equal = TRUE,
       conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  sqf2015friskedid and sqf2015friskednoid
## t = 3.5781, df = 22500, p-value = 0.0003469
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03039181 0.10402569
## sample estimates:
## mean of x mean of y
## 0.6775374 0.6103286
```

With a tval of -3.5781 and 22500 degrees of freedom, we generate a 95 percent confidence interval for the differenc of means: $0.03039181 < \mu_1 - \mu_2 < 0.10402569$. Because the confidence interval does not include 0, we conclude with 95% confidence that the two means are not the same. Additionally, the p-value associated with this t value and degrees of freedom is 0.0003469, which is less that 0.05. Hence, we reject H_0 that claims the two means are the same.