COMP4441_final project

Abigail Ward, Simran Kota

8/2/2021

This data comes from The Washington Post (https://bit.io/bitdotio/police_shooting/). The records were collected via local news reports, law enforcement websites, social media, and independent databases such as Killed by Police and Fatal Encounters. This list is continually updated with additional records and new information of prrevious cases. The FBI and the Centers for Disease Control and Prevention also logged fatal shootings by police, but had an incomplete list. Each record represents a civilian in the United stated who was shot and killed by a police officer in the line of duty from 2015 to 2021. Not included in this data set are deaths of people in police custody, fatal shootings by off-duty officers, and non-shooting deaths. The records include details about each individual such as race, gender, age, location, signs of mental illness, etc.

Initial exploration of the data:

```
dat<-read.csv("fatal-police-shootings-data.csv")
dat$date <- as.Date(dat$date)
```

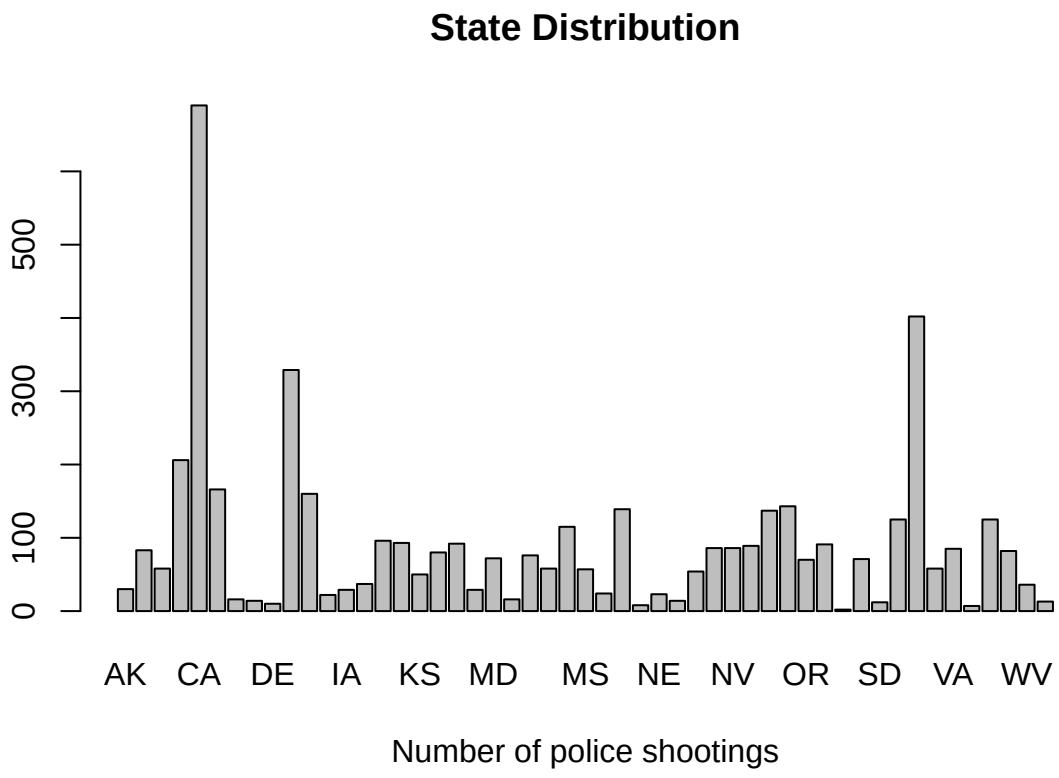
The initial data set has `6471` observations of `17` features. However, there are a number of null values present in the data.

```
dat <- dat %>% na_if("") %>% na.omit
dat <- dat[dat$date < as.Date('2021-01-01'),]
```

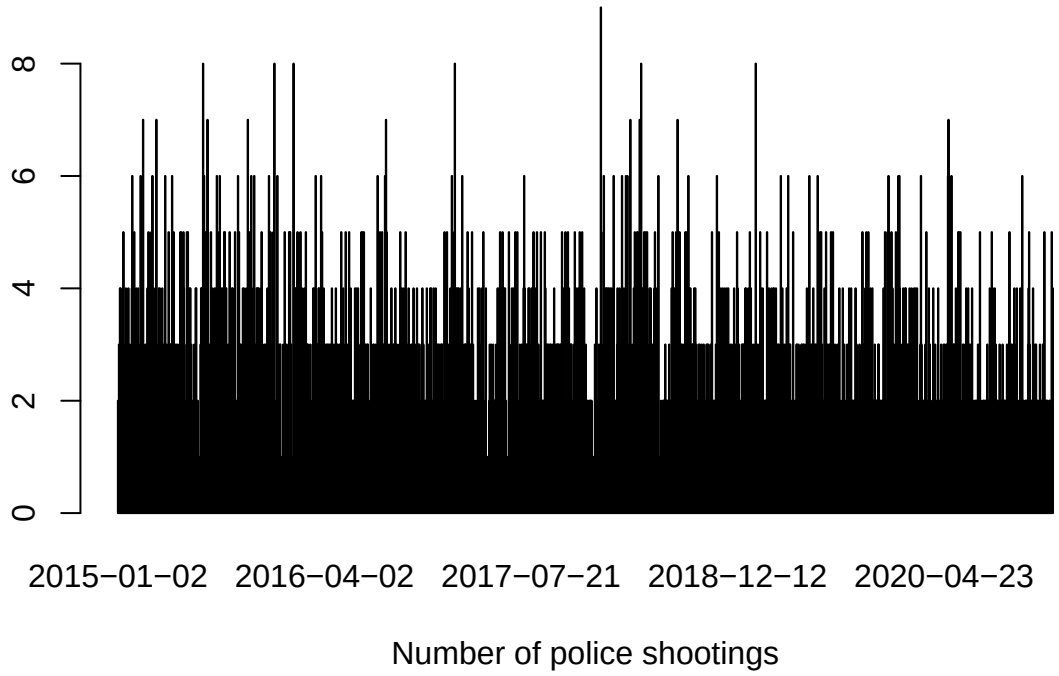
After removing the nulls and empty strings, we have `4666` observations of `17` features. We also drop the data for 2021 because it is incomplete and skews our results.

Exploratory Data Analysis

```
countState <- table(dat$state)
barplot(countState, main="State Distribution",
        xlab="Number of police shootings")
```

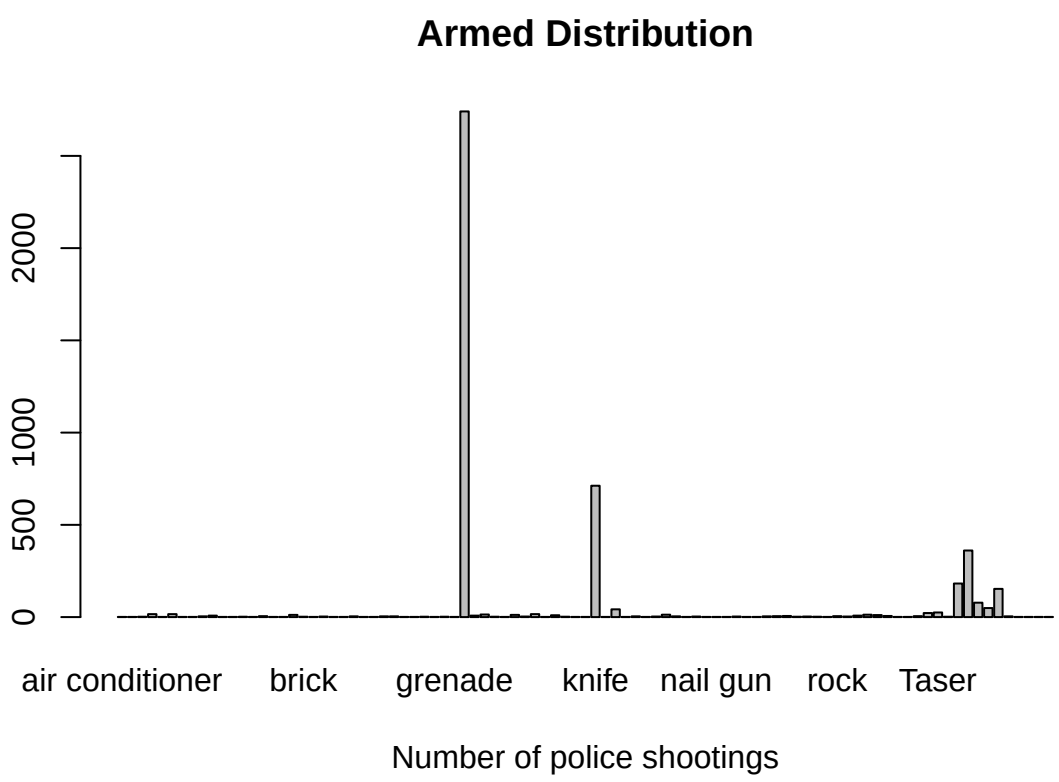



```
countDate <- table(dat$date)
barplot(countDate, main="Date Distribution",
        xlab="Number of police shootings")
```

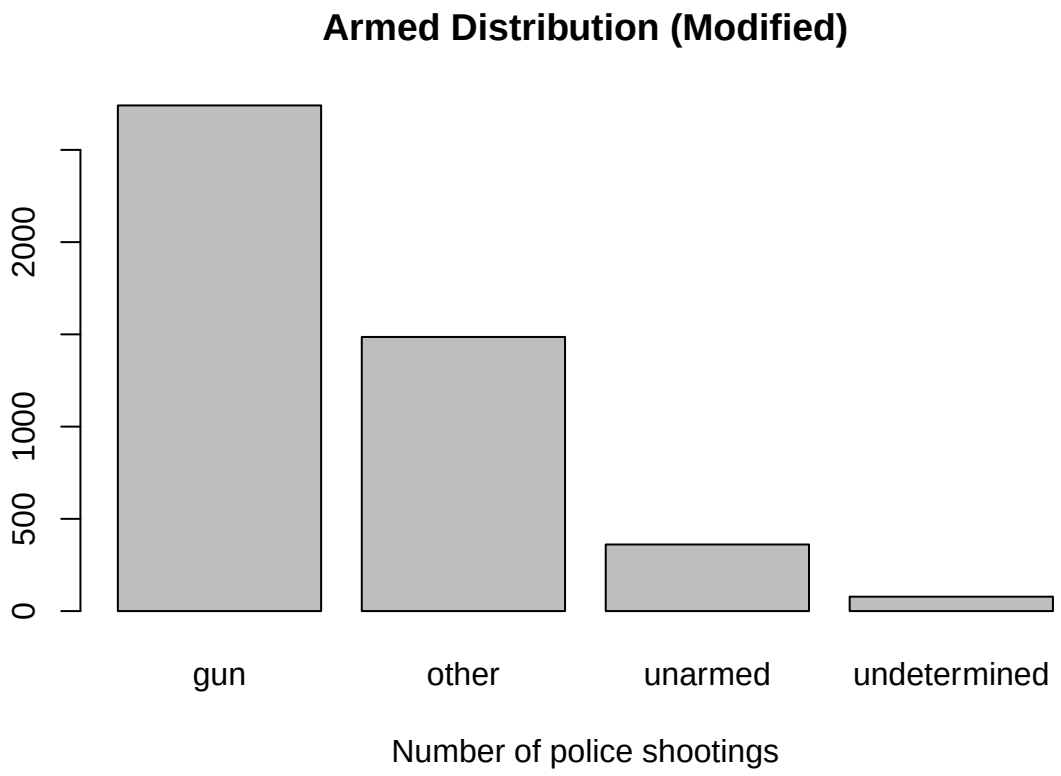



```
countAge <- table(dat$age)
barplot(countAge, main="Age Distribution",
        xlab="Number of police shootings")
```

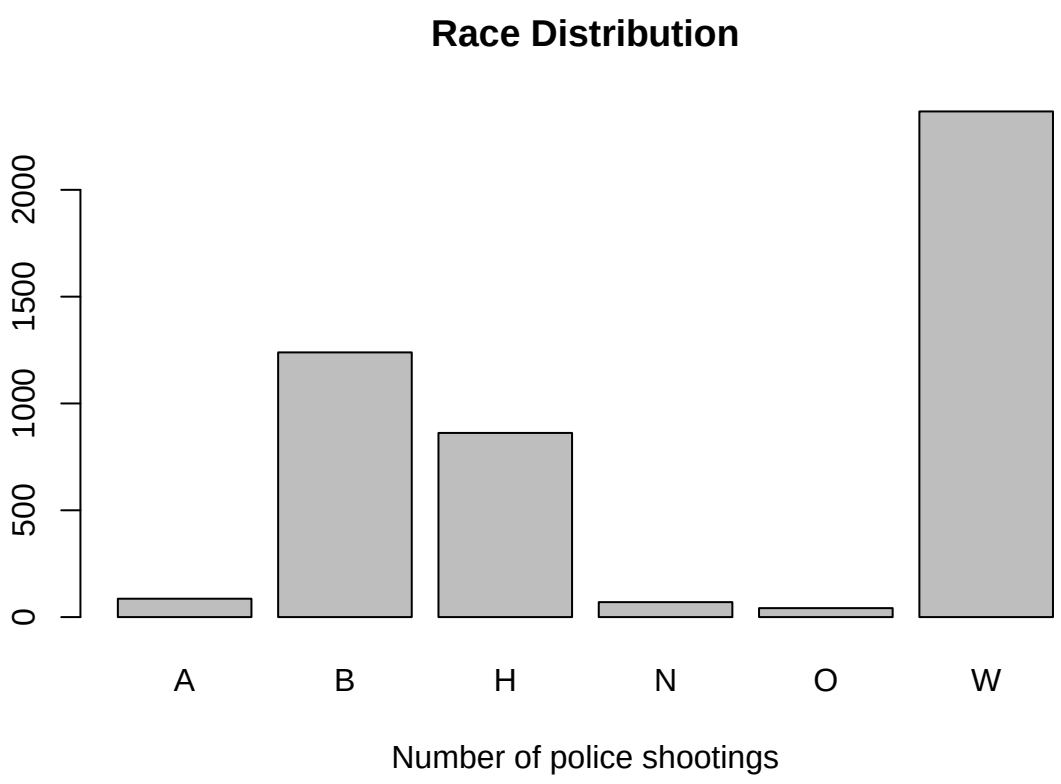

```
countArmed <- table(dat$armed)
barplot(countArmed, main="Armed Distribution",
        xlab="Number of police shootings")
```




```
gun<-("gun")
unarmed<-("unarmed")
dat$group <- with(dat, ifelse(armed %in% gun, "gun",
                              ifelse(armed %in% unarmed, "unarmed",
                                     ifelse(armed %in% ("undetermined"),
                                            "undetermined", "other" ))))
countArmed.grouped <- table(dat$group)
barplot(countArmed.grouped, main="Armed Distribution (Modified)",
        xlab="Number of police shootings")
```




```
countRace <- table(dat$race)
barplot(countRace, main="Race Distribution",
        xlab="Number of police shootings")
```



Findings

- The age variable appears to be approximately Normally distributed, but skewed to the left.
- There are too many individual values in the armed variable to conduct a meaningful analysis. To make the values more meaningful, we constructed four categories to sort the values into: gun, other, unarmed, and undetermined.
- It appears that there are significantly more Caucasian people represented in the data than any other race.

Goal

The goal of this analysis is to assess the significance various factors such as age, race, mental illness and gender have on the population of civilians in the United States who have been fatally shot by the police.

Time Series

The one letter abbreviations for race are as follows:

A Asian,

B Black or African American,

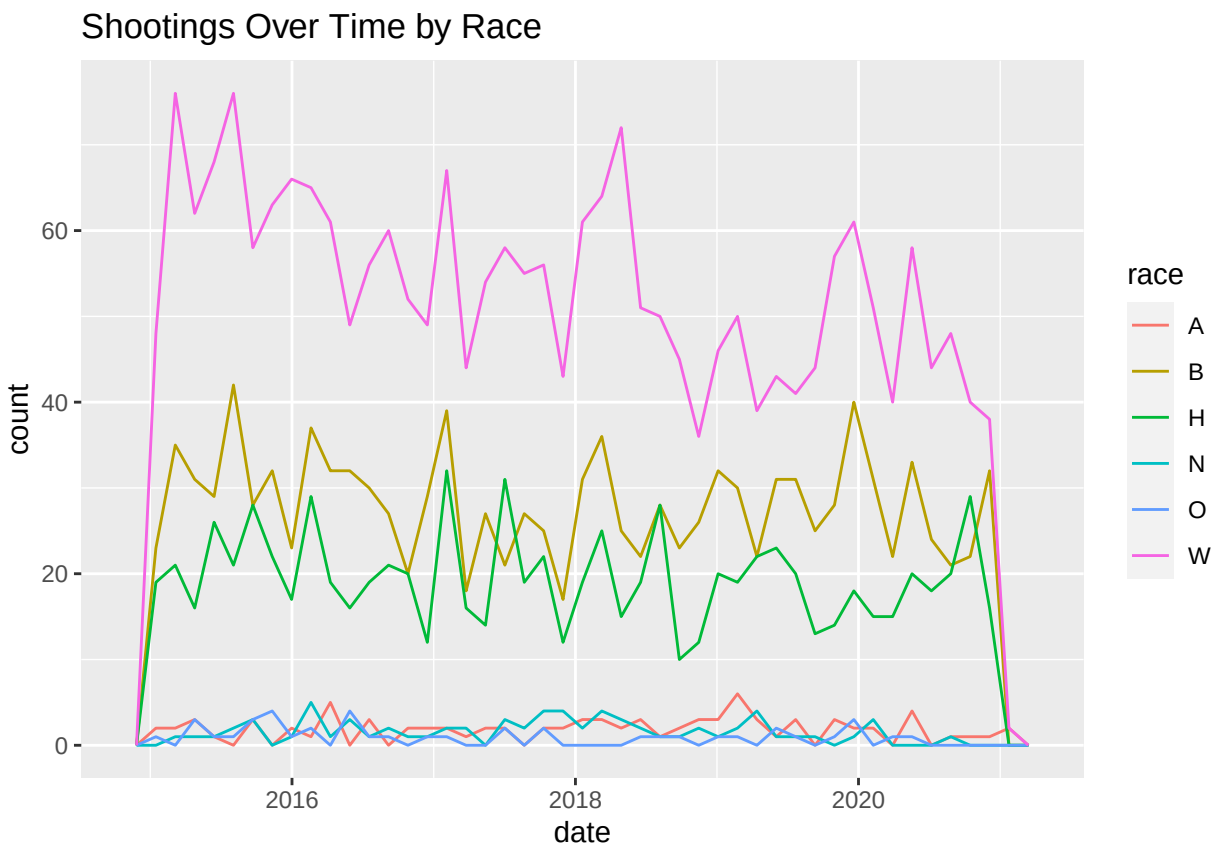
H Hispanic,

N American Native or Alaskan Native,

O Unknown/Other/Two ore more,

W White

```
ggplot(dat, aes(x=date, color=race)) + geom_freqpoly(binwidth=50) +
  labs(title="Shootings Over Time by Race")
```



There does not appear to be any specific trend over time in the number of people who are fatally shot.

Two-Sample Tests

We will now attempt to examine whether there is a significant difference between the observed number of shootings for various values of the features.

Here we construct a data frame based on the original data that shows the total number of shootings per year by race.

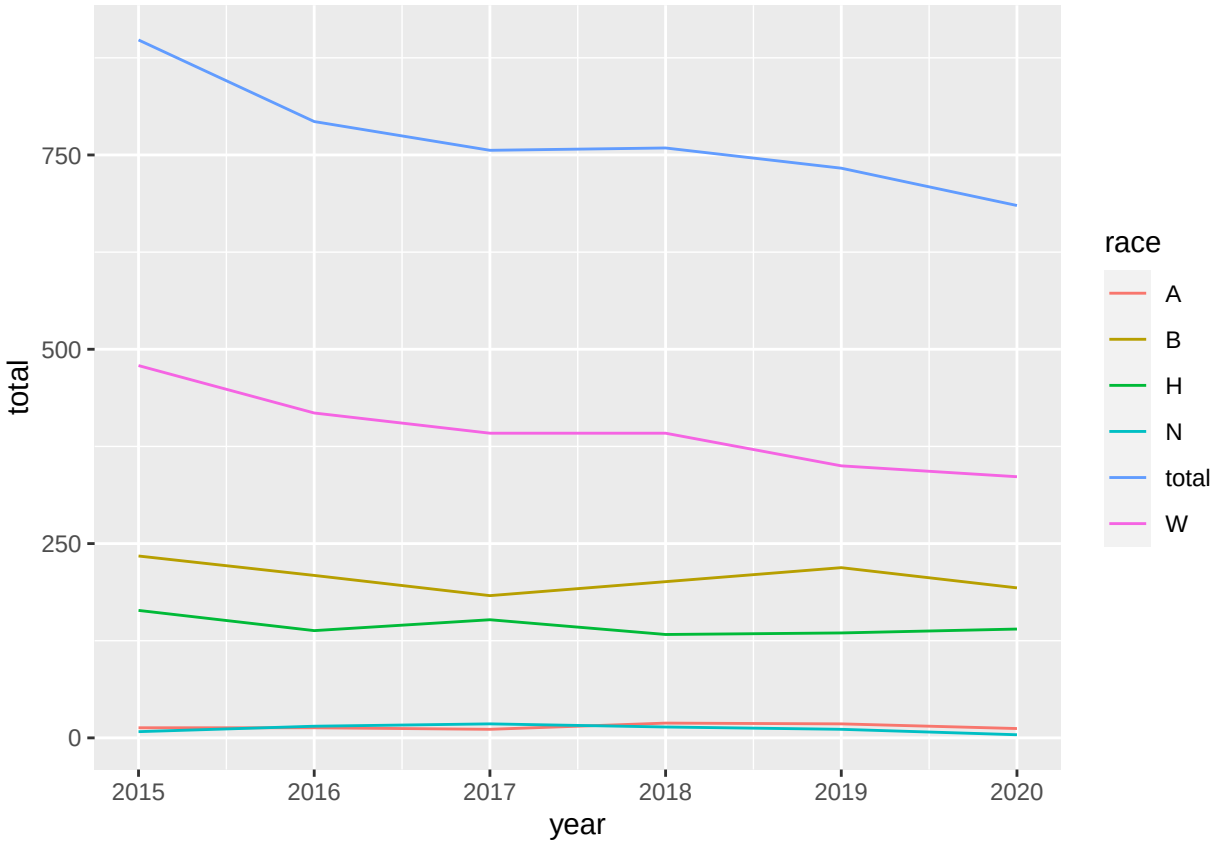
```
by_year <- data.frame(year=2015:2020)
by_year$A <- (dat[dat$race == "A",] %>% mutate(year=format(date, "%Y")) %>%
                group_by(year) %>% tally())$n
by_year$B <- (dat[dat$race == "B",] %>% mutate(year=format(date, "%Y")) %>%
                group_by(year) %>% tally())$n
by_year$W <- (dat[dat$race == "W",] %>% mutate(year=format(date, "%Y")) %>%
                group_by(year) %>% tally())$n
by_year$H <- (dat[dat$race == "H",] %>% mutate(year=format(date, "%Y")) %>%
                group_by(year) %>% tally())$n
```

```
by_year$N <- (dat[dat$race == "N",] %>% mutate(year=format(date, "%Y")) %>%
                group_by(year) %>% tally())$n
by_year$total <- apply(by_year[,2:6], 1, sum)
by_year
```

```
##   year  A   B   W   H  N total
## 1 2015 13 234 479 164  8   898
## 2 2016 13 209 418 138 15   793
## 3 2017 11 183 392 152 18   756
## 4 2018 19 201 392 133 14   759
## 5 2019 18 219 350 135 11   733
## 6 2020 12 193 336 140  4   685
```

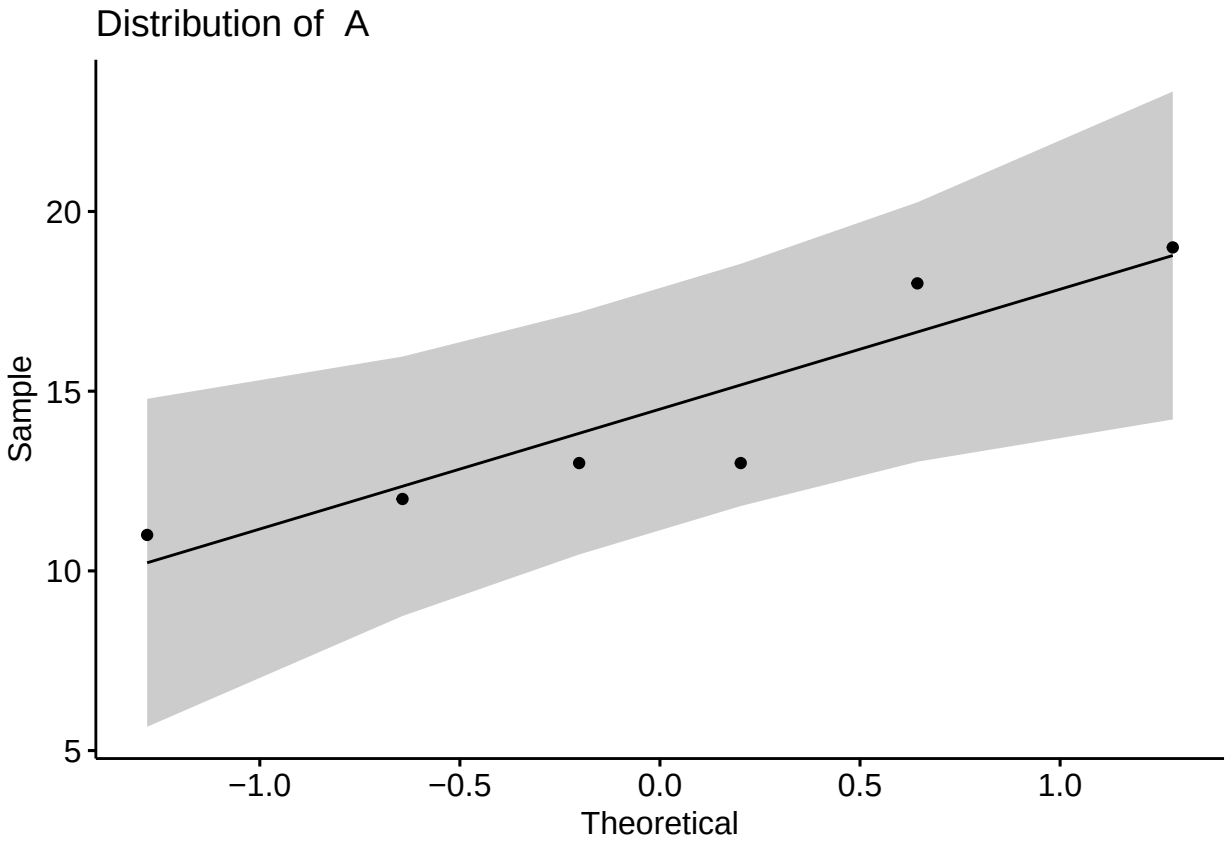
We will reformat the data frame wherein each row will represent a unique year-race combination. Examining the trend over time by race, we see no significant differences between the trends.

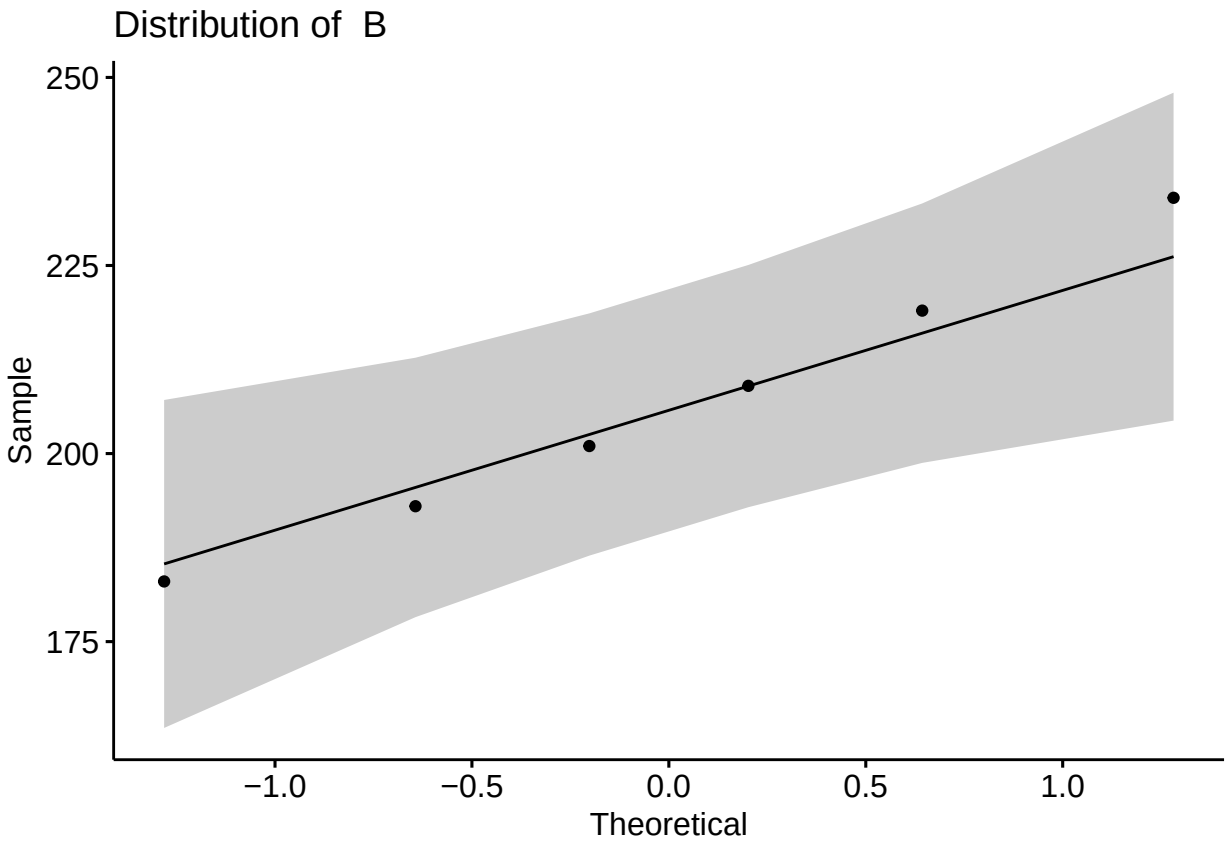
```
by_long <- by_year %>%
  gather(race, total, -year)
ggplot(data=by_long, aes(x=year, y=total, color=race)) + geom_line()
```

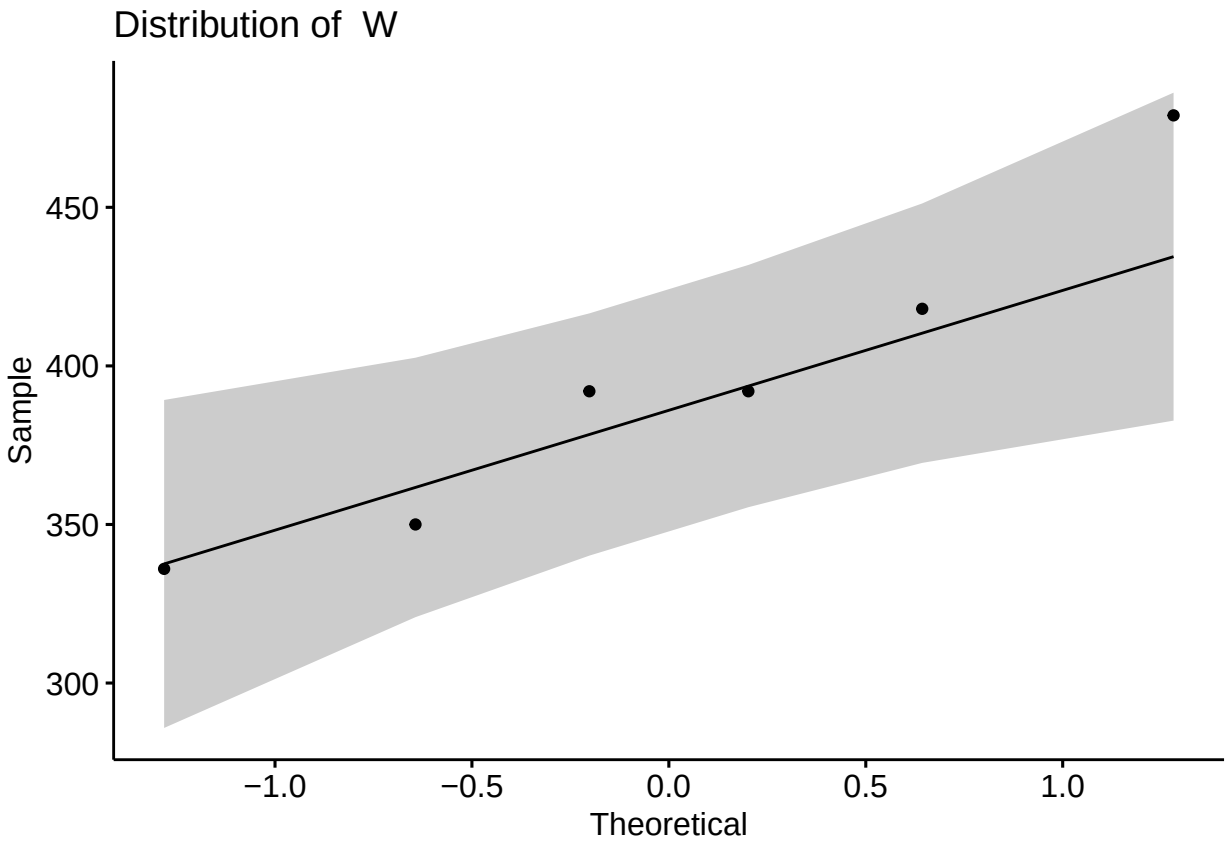


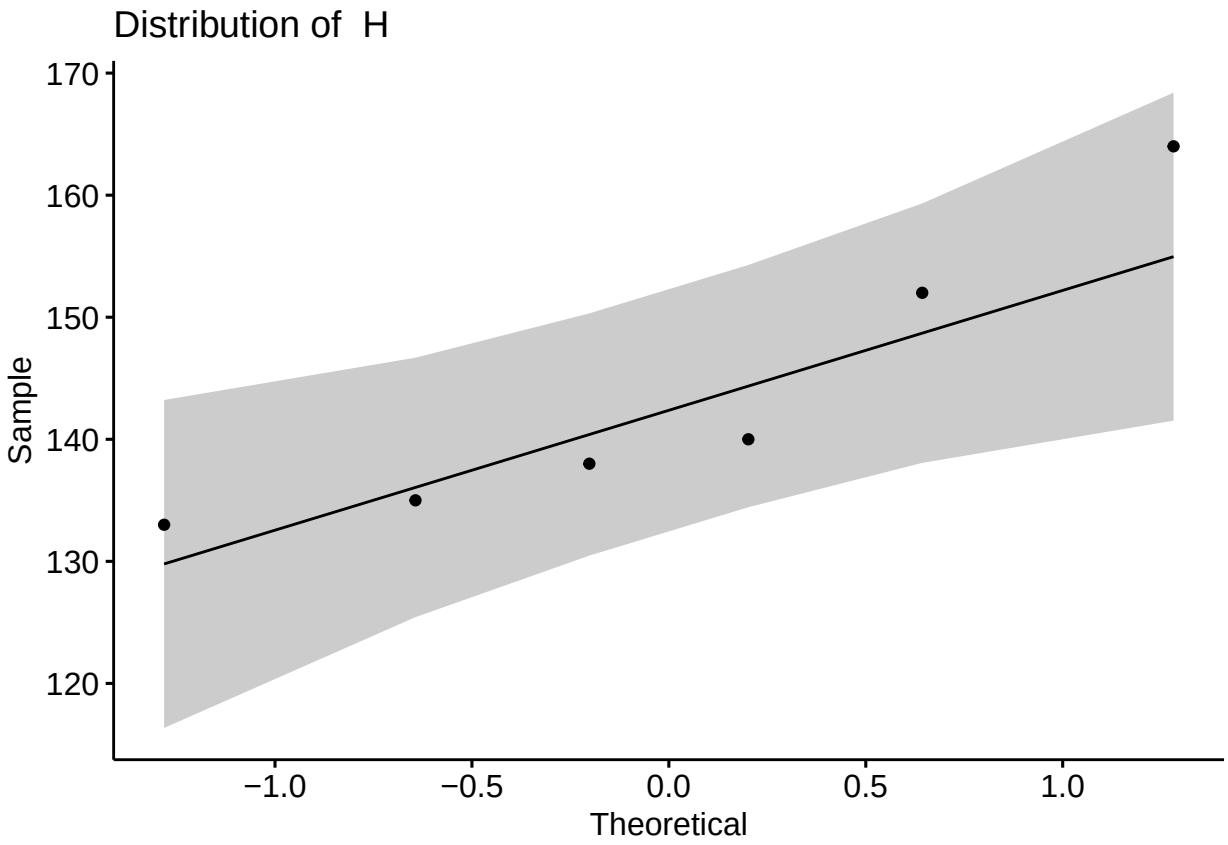
We will now verify that the data for each race is normally distributed in order to conduct a two-sample T-test between various races.

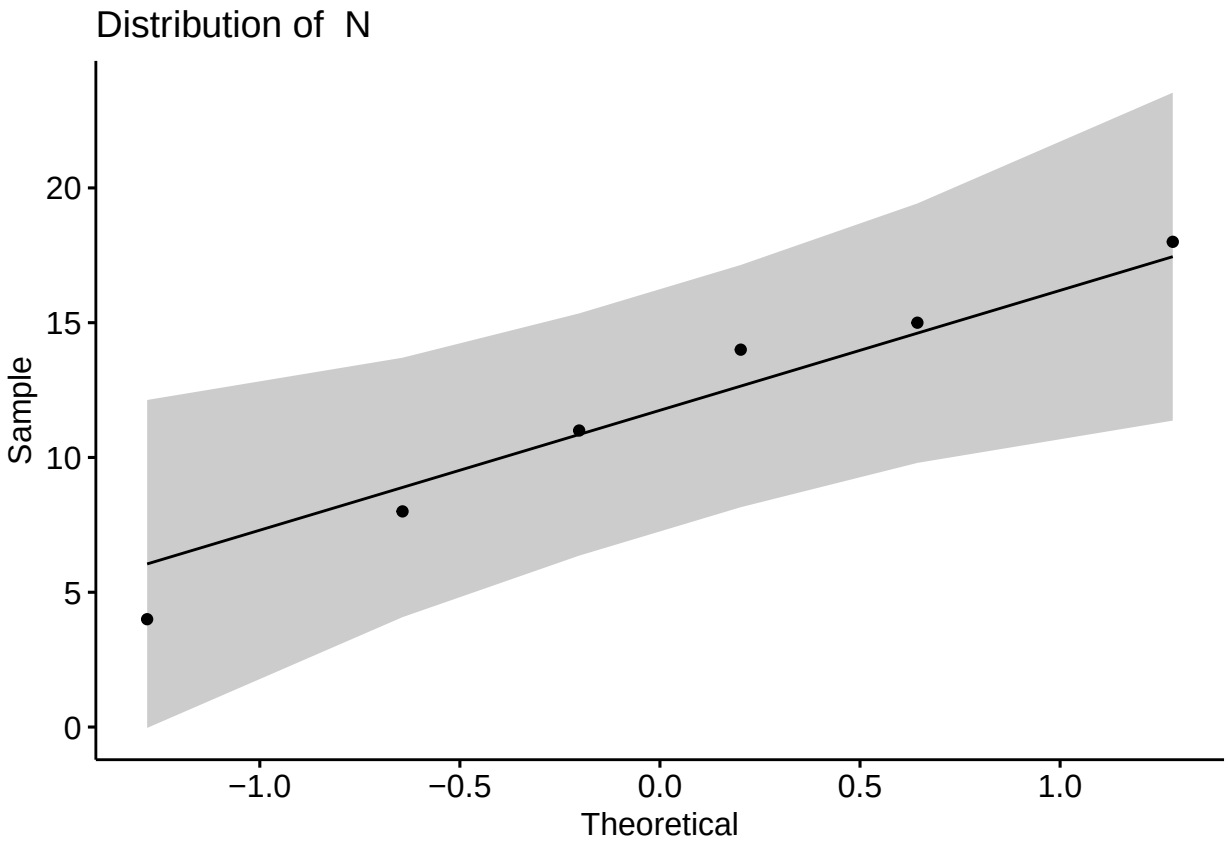
```r
for (i in 2:6) {
  print(ggqqplot(by_year[,i], title=paste('Distribution of ', colnames(by_year)[i])))
}
```




Distribution of B
250
225
200
175
Sample
−1.0
−0.5
0.0
0.5
1.0
Theoretical


Distribution of W
Sample
450
400
350
300
−1.0
−0.5
0.0
0.5
1.0
Theoretical


Distribution of H
170
160
150
140
130
120
Sample
−1.0
−0.5
0.0
0.5
1.0
Theoretical



Since the data is normally distributed, we may proceed.

```
t.test(by_year$W - by_year$B)
```

```
##
##  One Sample t-test
##
## data:  by_year$W - by_year$B
## t = 10.613, df = 5, p-value = 0.0001284
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  142.4637 233.5363
## sample estimates:
## mean of x
##       188
```

```
t.test(by_year$total - by_year$W)
```

```
##
##  One Sample t-test
##
## data:  by_year$total - by_year$W
## t = 38.575, df = 5, p-value = 2.206e-07
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
```

```
##  351.0993 401.2341
## sample estimates:
## mean of x
##  376.1667
```

```
t.test(by_year$B - by_year$H)
```

```
##
##  One Sample t-test
##
## data:  by_year$B - by_year$H
## t = 8.34, df = 5, p-value = 0.0004053
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  43.46658 82.20009
## sample estimates:
## mean of x
##  62.83333
```

The t-tests indicate that the data for each race cannot be considered to be iid samples, and that there is a significant difference between the distributions of each race.

```
illness <- dat %>% mutate(year=format(date, "%Y")) %>%
  group_by(year, signs_of_mental_illness) %>% tally()
t.test(data=illness, n ~ signs_of_mental_illness)
```

```
##
##  Welch Two Sample t-test
##
## data:  n by signs_of_mental_illness
## t = 16.526, df = 9.8888, p-value = 1.576e-08
## alternative hypothesis: true difference in means between group False and group True is not equal to 0
## 95 percent confidence interval:
##  339.0671 444.9329
## sample estimates:
## mean in group False  mean in group True
##            584.8333            192.8333
```

In a two sample t-test where we look at the differences between the group with Mental Illness and the group without, we also see a statistically significant difference. However, 24.8% of the data population had signs of mental illness, while only ~20.6% of adults in the US experienced mental illness in 2019 (according to the National Alliance on Mental Illness). We can conduct a t-test to examine whether this difference is statistically significant.

```
props <- data.frame(year=2015:2020,
                    n=illness[illness$signs_of_mental_illness == "True", 3])
illness$year <- as.integer(illness$year)
tot.pop <- aggregate(x = illness$n, by=list(illness$year), FUN=sum)
colnames(tot.pop) <- c("year", "n")
tot.pop$prop <- props$n / tot.pop$n
tot.pop
```

```
##   year   n      prop
## 1 2015 912 0.2697368
## 2 2016 802 0.2768080
## 3 2017 761 0.2733246
## 4 2018 762 0.2257218
## 5 2019 741 0.2240216
## 6 2020 688 0.2078488
```

```
t.test(tot.pop$prop, mu=0.206)
```

```
##
##  One Sample t-test
##
## data:  tot.pop$prop
## t = 3.2469, df = 5, p-value = 0.02277
## alternative hypothesis: true mean is not equal to 0.206
## 95 percent confidence interval:
##  0.2143829 0.2781043
## sample estimates:
## mean of x
## 0.2462436
```

By conducting a one sample t-test between the observed proportions of mental illness in the data and the actual proportion of mental illness in US adults, we see that the difference is statistically significant. We have strong evidence to reject the null hypothesis, showing that although there are more people who are fatally shot that don't have signs of mental illness, there is a higher chance of someone being fatally shot if they do show signs of mental illness.

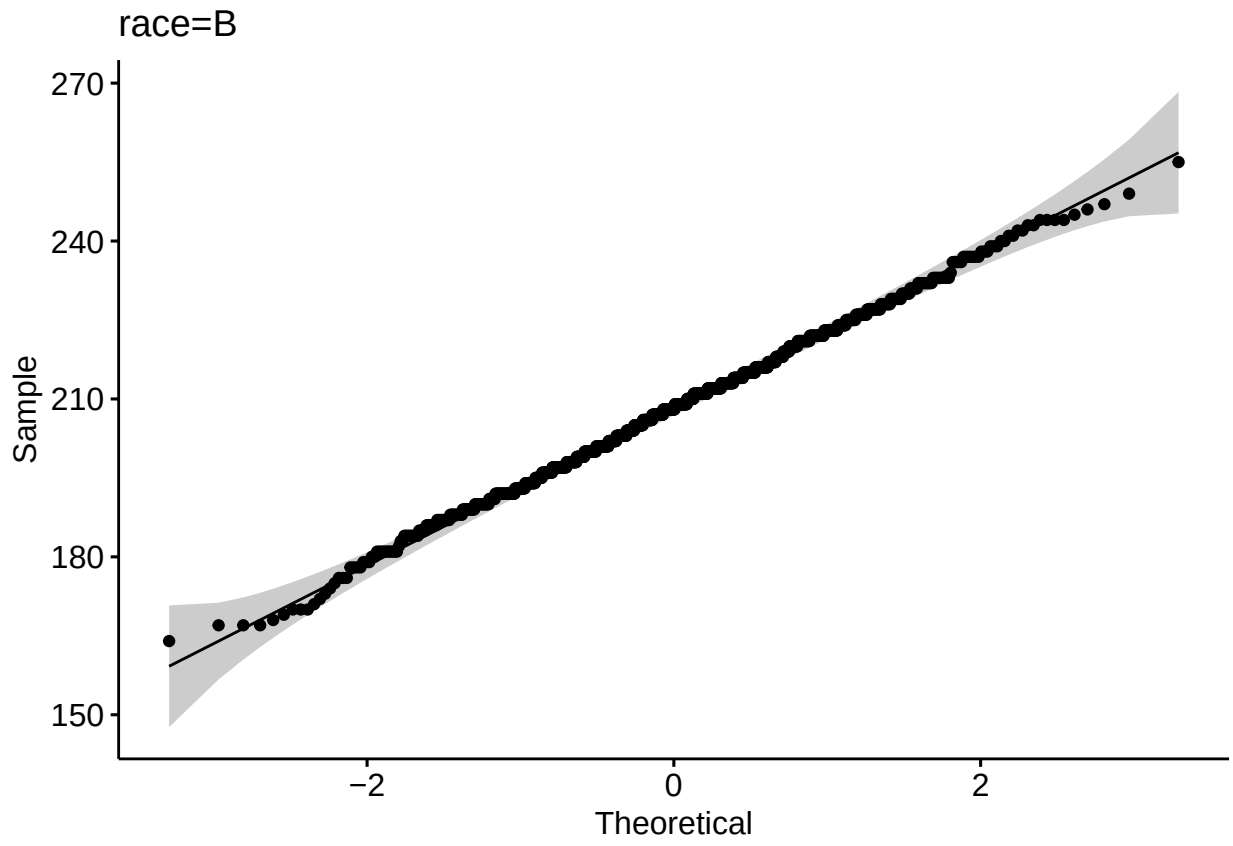
Statistical Model

We will investigate whether the number of people that are fatally shot for a given race in our data is consistent with a population in which each person, regardless of race, is equally likely to be fatally shot. To do this, we will create 1000 samples of a population of 500000 people in the US based on 2019 Census data regarding racial proportions. The probability of getting fatally shot regardless of race is 1/315, according to Business Insider.

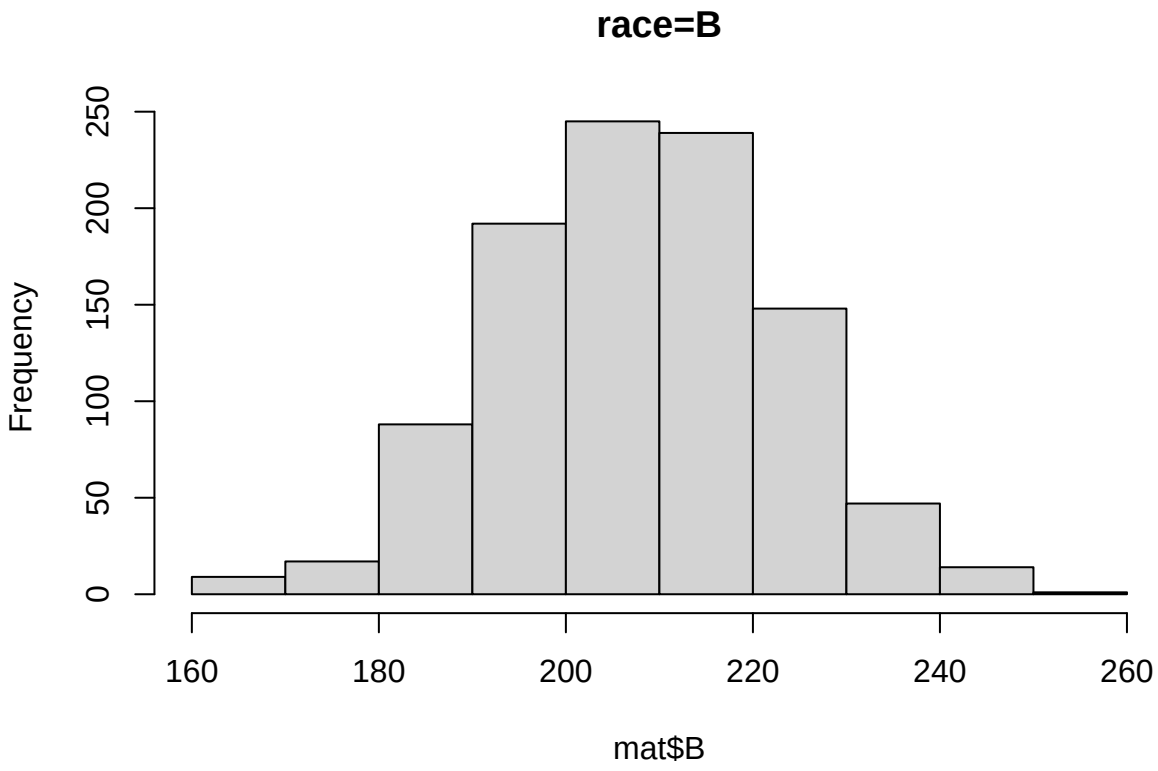
The racial proportions are: A Asian 5.9%, B Black or African American 13.4%, H Hispanic 18.5%, N American Native or Alaskan Native 1.3%, O Unknown/Other/Two or more races 2.8%, W White 60.1%

```
set.seed(1234)
gen_samp <- function() {
  gen_pop <- data.frame(id=1:5000, race=sample(unique(dat$race)[1:6],500000,
                                               replace=TRUE,
                                               prob = c(0.059, 0.601, 0.185,
                                                        0.134, 0.028, 0.013)))
  gen_pop$shot <- sample(c(0, 1), 500000, replace=TRUE,
                         prob = c(314/315, 1/315))
  gen_pop <- gen_pop[gen_pop$shot == 1,]
  gen_pop <- gen_pop %>% group_by(race) %>% tally()
  return (gen_pop$n)
}
mat <- matrix(rep(NA, 6000), ncol=6)
```

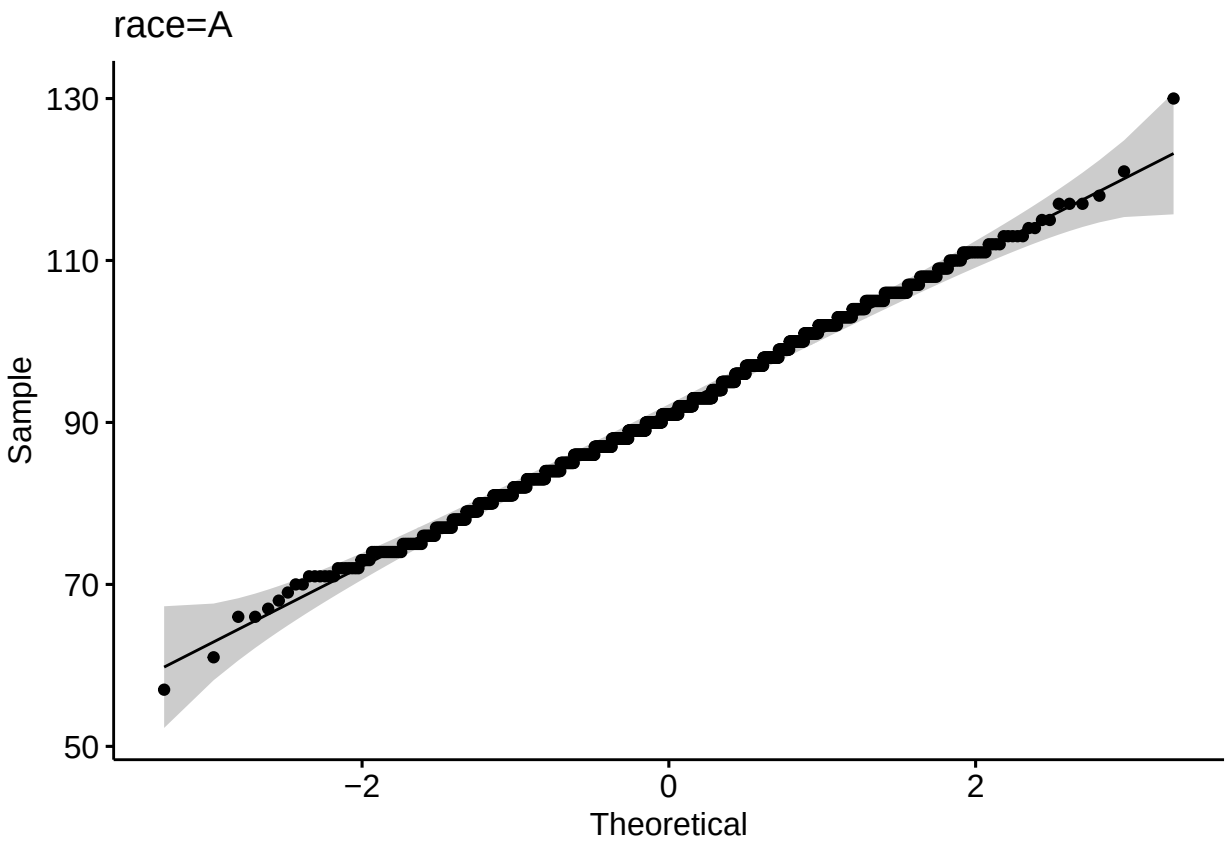
```r
for (i in 1:1000) {
  mat[i,] = gen_samp()
}
mat <- data.frame(mat)
colnames(mat) <- c("A", "B", "H", "N",  "O", "W")
ggqqplot(mat$B, title='race=B')
```



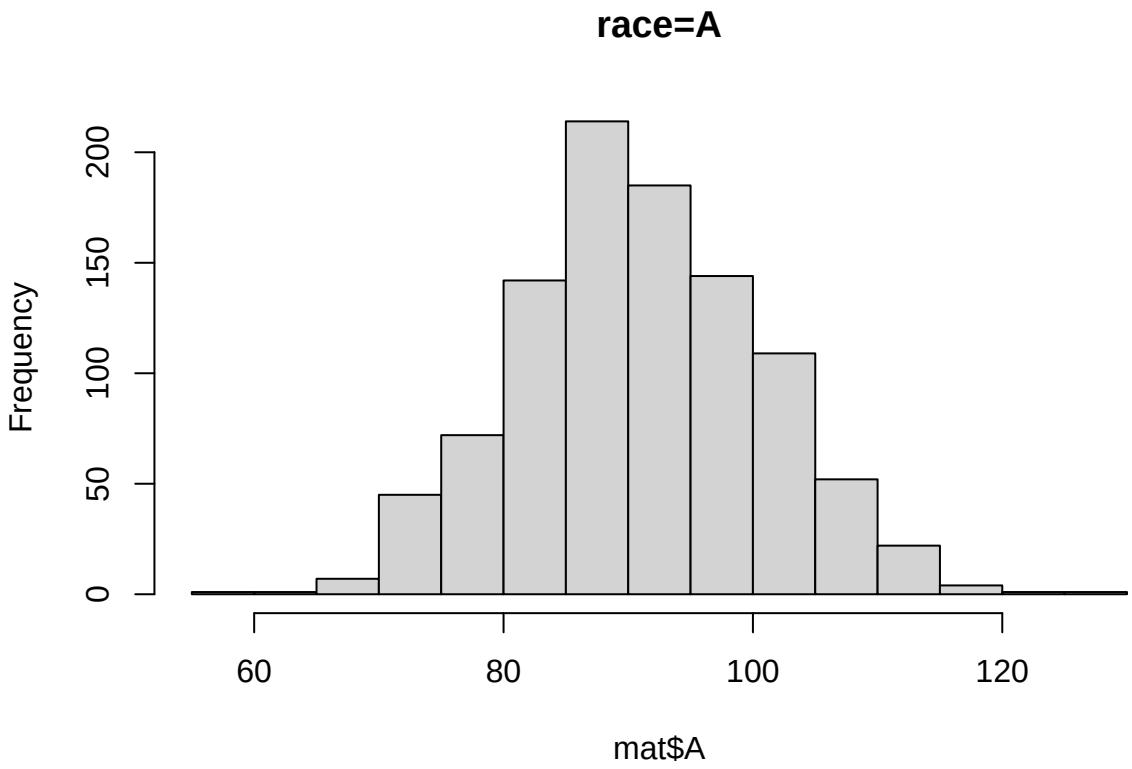

```r
hist(mat$B, main='race=B')
```



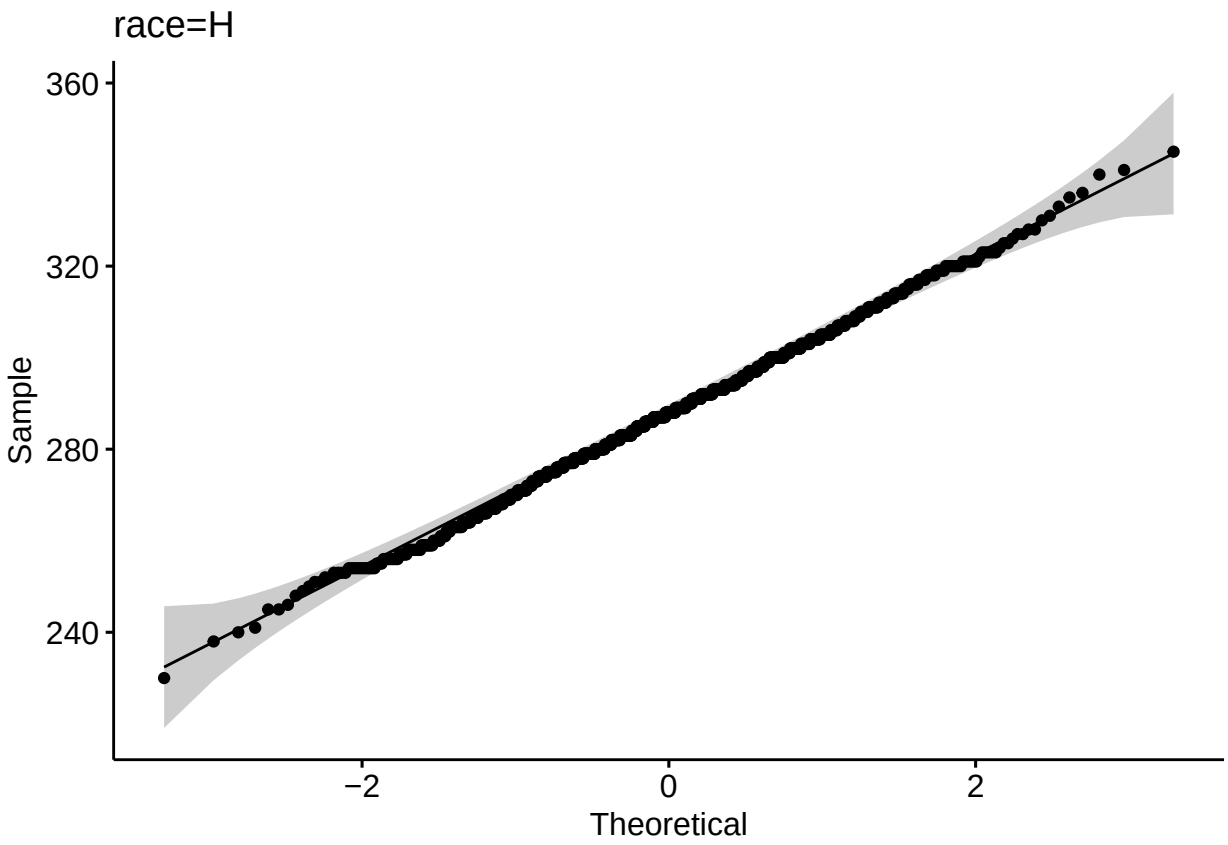

```
ggqqplot(mat$A, title='race=A')
```



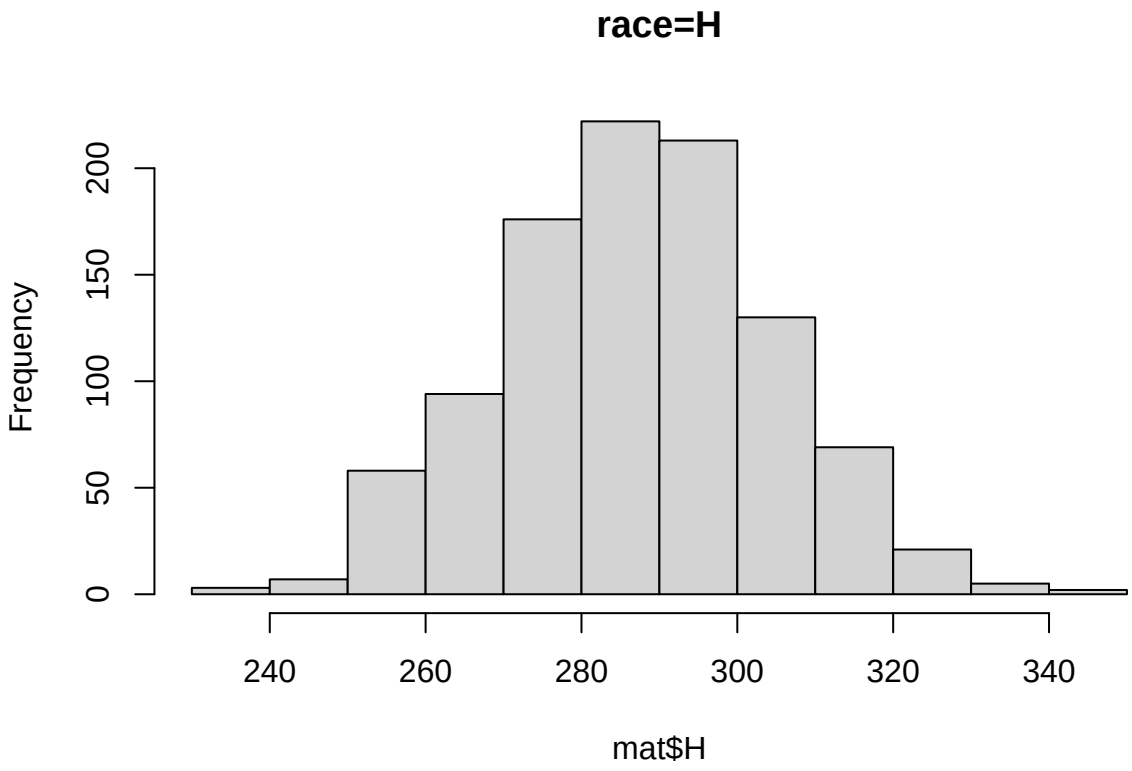

```
hist(mat$A, main='race=A')
```



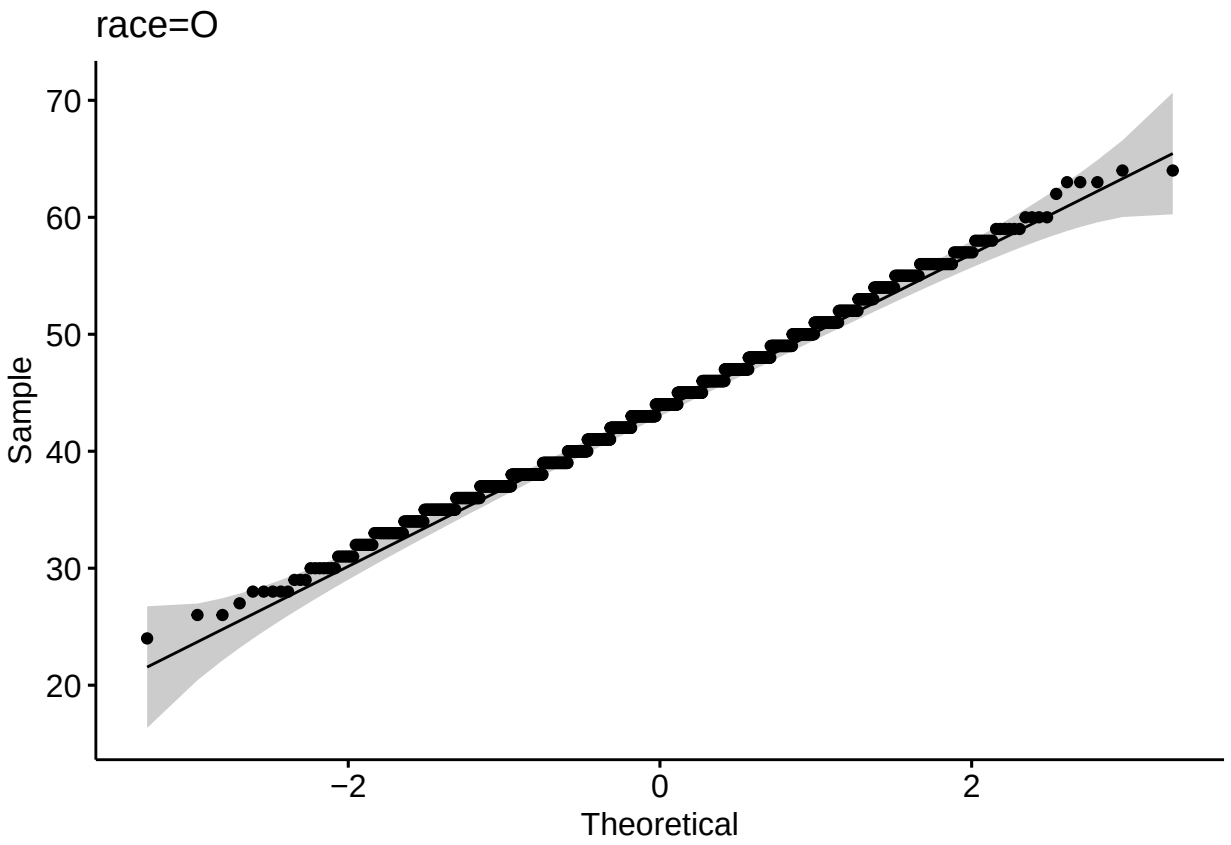

```
ggqqplot(mat$H, title='race=H')
```



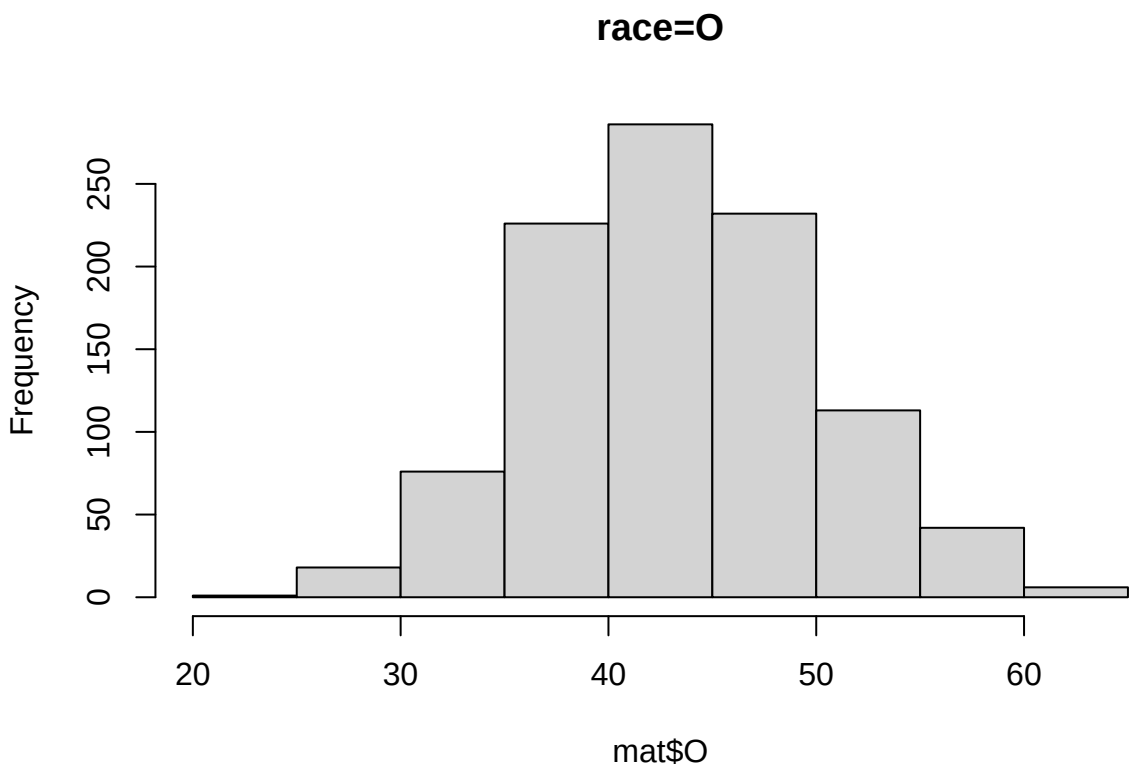

```
hist(mat$H, main='race=H')
```



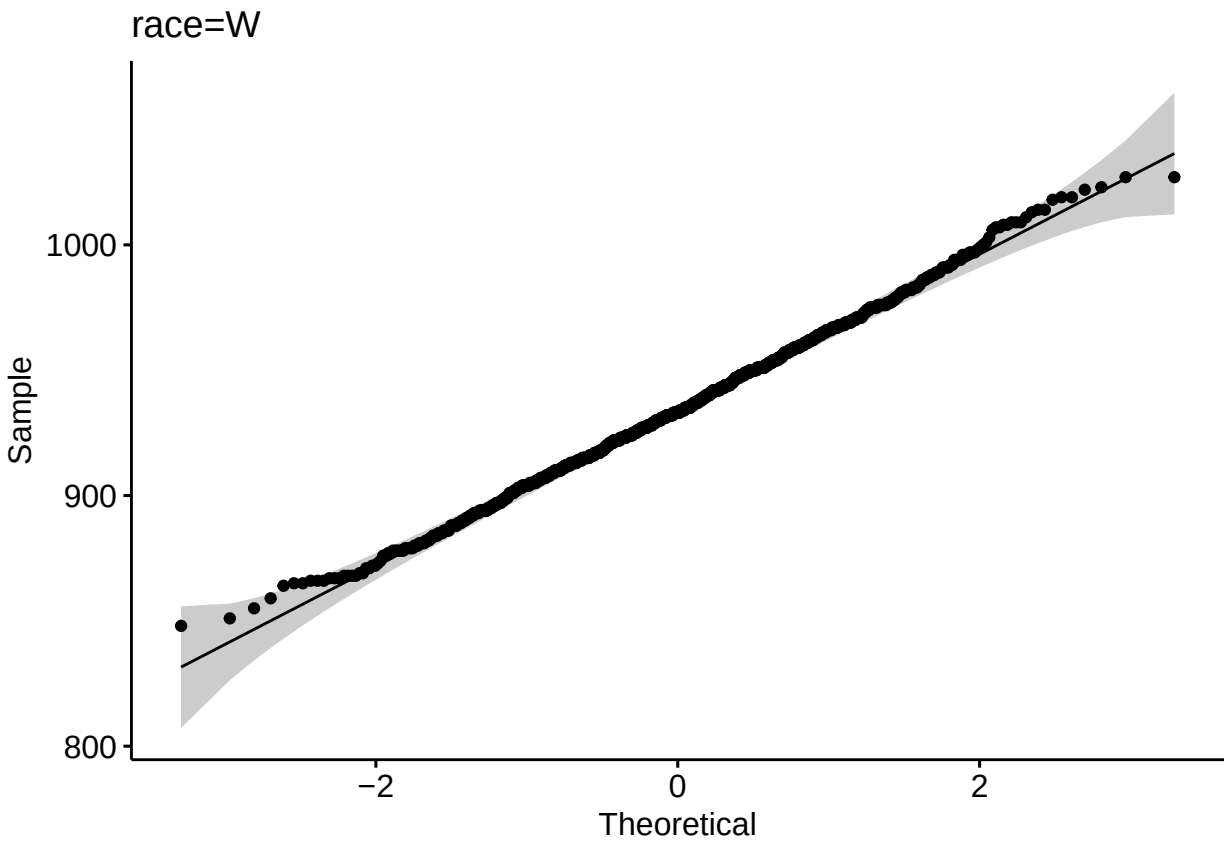

```
ggqqplot(mat$O, title='race=O')
```



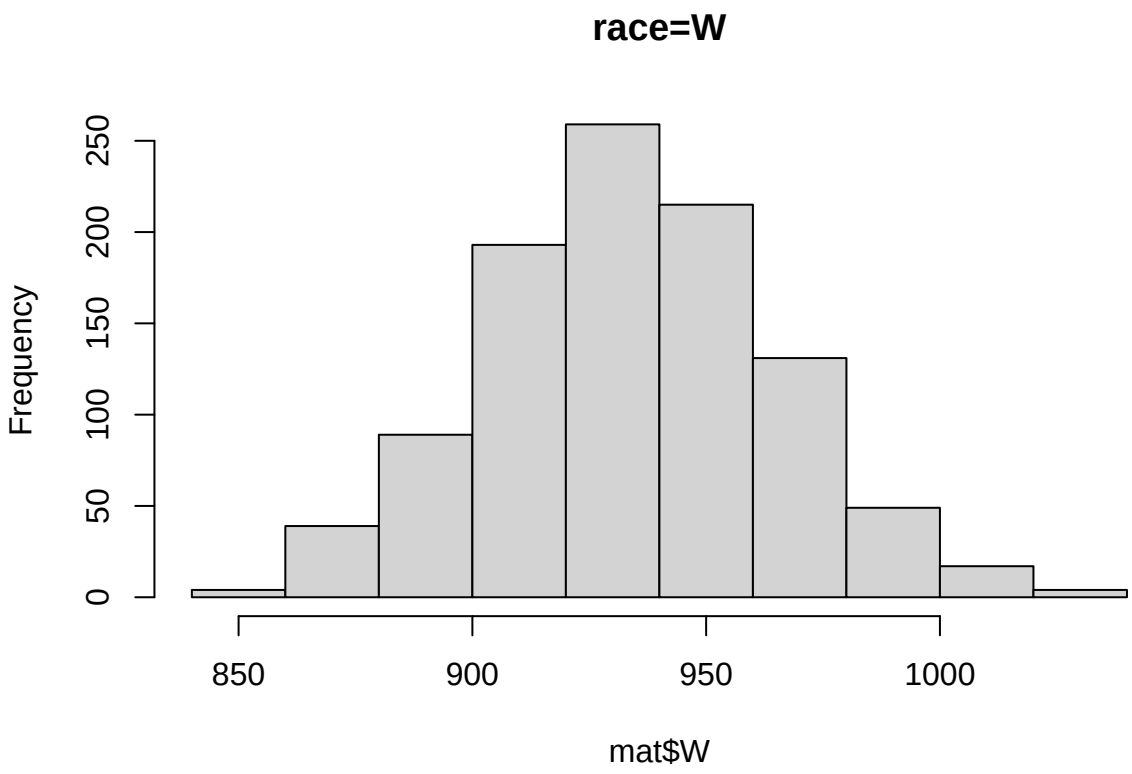

```
hist(mat$O, main='race=O')
```



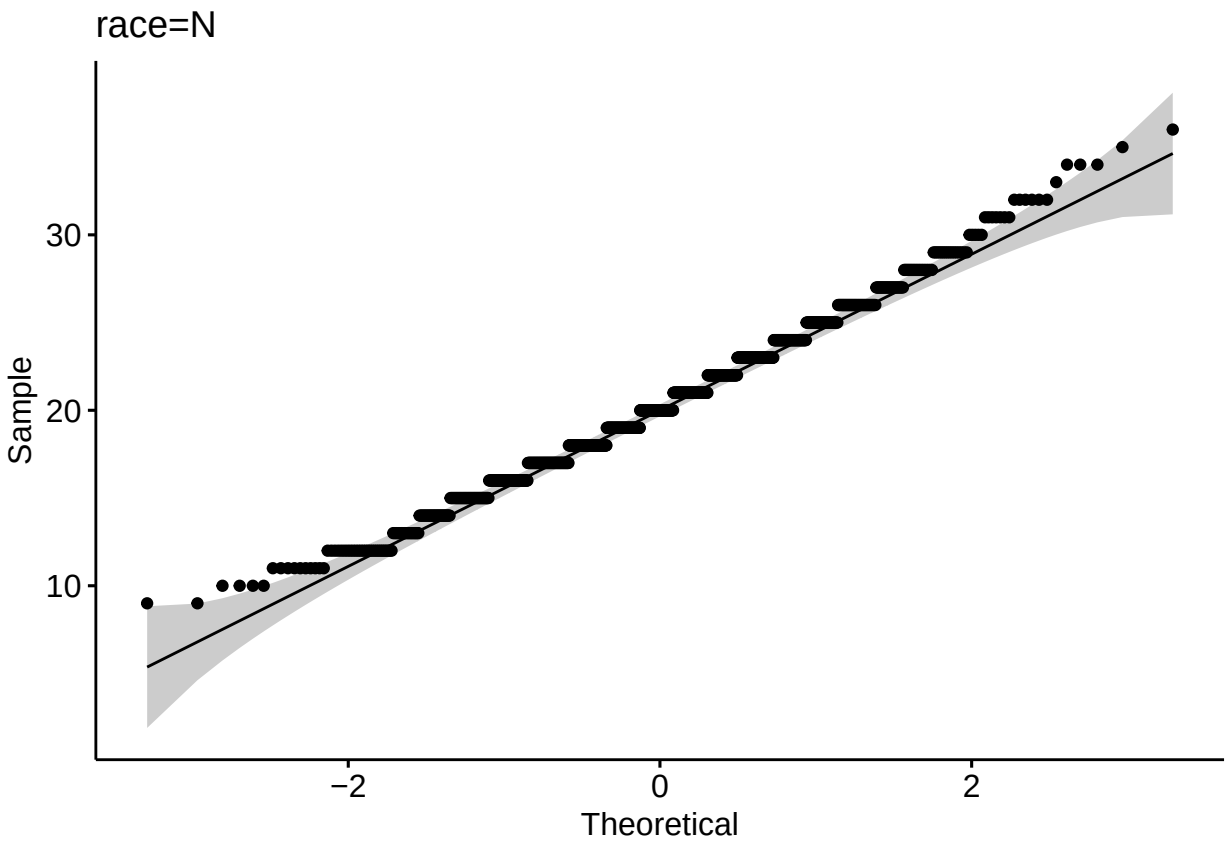

```
ggqqplot(mat$W, title='race=W')
```



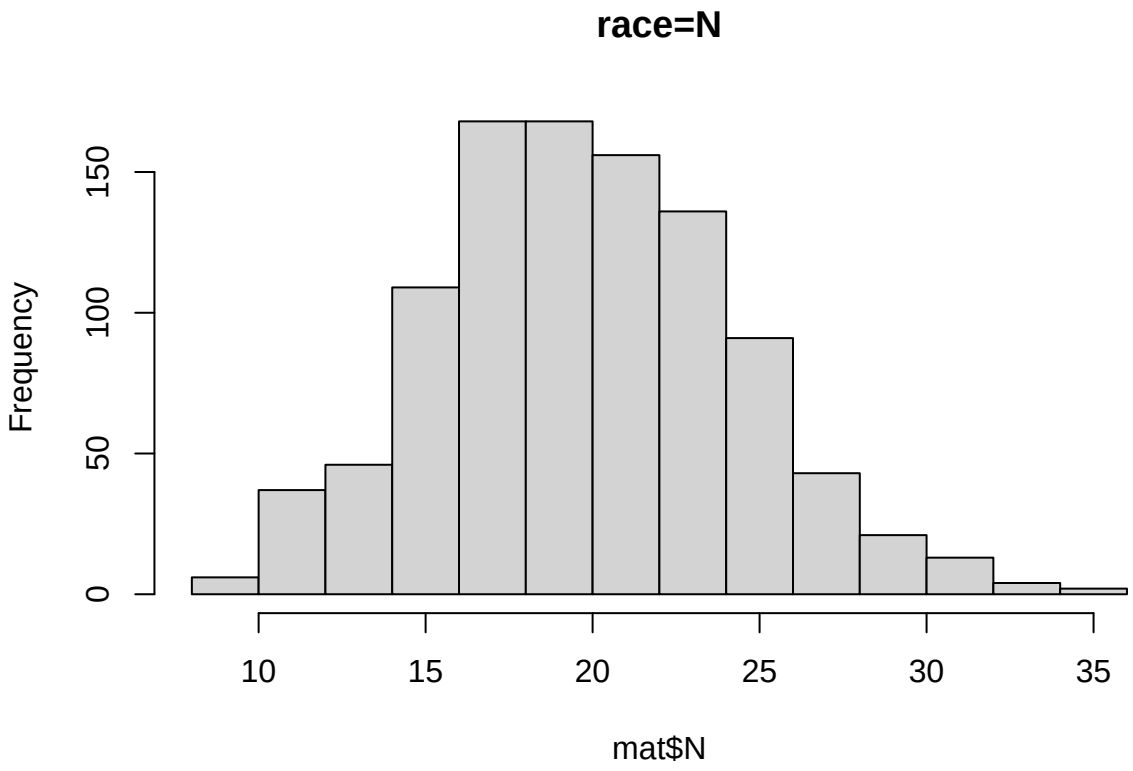

```
hist(mat$W, main='race=W')
```




```
ggqqplot(mat$N, title='race=N')
```




```
hist(mat$N, main='race=N')
```




```
mat$total <- apply(mat, 1, sum)
```

The data appears to be normally distributed, indicating that we may use a Student's T-test. We will test the hypothesis that the proportion of fatal shootings for a given race, on average, is equal to the observed proportion of fatal shootings under the assumption that all races are equally likely to be a victim of a fatal shooting.

```
obs.val <- dat %>% group_by(race) %>% tally()
race <- 'B'
mean <- as.integer(obs.val[obs.val$race == race,2])
t.test(mat$B / mat$total, mu=mean / sum(obs.val$n))
```

```
##
##  One Sample t-test
##
## data:  mat$B/mat$total
## t = -486.89, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0.2655379
## 95 percent confidence interval:
##  0.1307091 0.1317915
## sample estimates:
## mean of x
## 0.1312503
```

Here we conducted the T-test for the African American population. The mean of our simulated populations was approximately 13%, while the observed mean was 26%. The 95% confidence interval does not include

the observed mean, and we get a p-value of 2.2e-16, providing strong evidence to reject the null hypothesis. This also shows us that the percentage of African-Americans that are fatally shot is significantly higher than it would be in the case where all races are equally likely to be fatally shot.

```
race <- 'W'
mean <- as.integer(obs.val[obs.val$race == race,2])
t.test(mat$W / mat$total, mu=mean / sum(obs.val$n))
```

```
##
##  One Sample t-test
##
## data:  mat$W/mat$total
## t = 210.88, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0.5072868
## 95 percent confidence interval:
##  0.5884468 0.5899715
## sample estimates:
## mean of x
## 0.5892091
```

Here we conducted the T-test for the Caucasian population. The mean of our simulated populations was approximately 59%, while the observed mean was 50%. The 95% confidence interval does not include the observed mean, and we get a p-value of 2.2e-16, providing strong evidence to reject the null hypothesis. This also shows us that the percentage of Caucasians that are fatally shot is lower than it would be in the case where all races are equally likely to be fatally shot.

```
race <- 'A'
mean <- as.integer(obs.val[obs.val$race == race,2])
t.test(mat$A / mat$total, mu=mean / sum(obs.val$n))
```

```
##
##  One Sample t-test
##
## data:  mat$A/mat$total
## t = 206.27, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0.0184312
## 95 percent confidence interval:
##  0.05726296 0.05800891
## sample estimates:
##  mean of x
## 0.05763593
```

Here we conducted the T-test for the Asian population. The mean of our simulated populations was approximately 5%, while the observed mean was 1.8%. The 95% confidence interval does not include the observed mean, and we get a p-value of 2.2e-16, providing strong evidence to reject the null hypothesis. This also shows us that the percentage of Asians that are fatally shot is lower than it would be in the case where all races are equally likely to be fatally shot.

```
race <- 'H'
mean <- as.integer(obs.val[obs.val$race == race,2])
t.test(mat$H / mat$total, mu=mean / sum(obs.val$n))
```

```
##
##  One Sample t-test
##
## data:  mat$H/mat$total
## t = -10.432, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0.1847407
## 95 percent confidence interval:
##  0.1808243 0.1820645
## sample estimates:
## mean of x
## 0.1814444
```

Here we conducted the T-test for the Hispanic population. The mean of our simulated populations was approximately 18%, while the observed mean was 18.4%. Although we get a p-value of 2.2e-16 which is strong evidence to reject the null hypothesis, the observed mean is very close to the simulated mean, and just outside the confidence interval. This shows us that the percentage of Hispanics that are fatally shot is almost consistent with the case where all races are equally likely to be fatally shot.

```
race <- 'N'
mean <- as.integer(obs.val[obs.val$race == race,2])
t.test(mat$N / mat$total, mu=mean / sum(obs.val$n))
```

```
##
##  One Sample t-test
##
## data:  mat$N/mat$total
## t = -24.703, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0.01500214
## 95 percent confidence interval:
##  0.01262326 0.01297340
## sample estimates:
##  mean of x
## 0.01279833
```

Here we conducted the T-test for the Native-American population. The mean of our simulated populations was approximately 1%, while the observed mean was 1.5%. Although we get a p-value of 2.2e-16 which is strong evidence to reject the null hypothesis, the observed mean is very close to the simulated mean, and just outside the confidence interval. This shows us that the percentage of Native Americans that are fatally shot is almost consistent with the case where all races are equally likely to be fatally shot.

After analyzing all the observed values for all races against the case where all races are equally likely to be fatally shot, we see that the only race with a statistically significantly higher rate of shootings than the simulated case is African Americans.

Tests of Independence

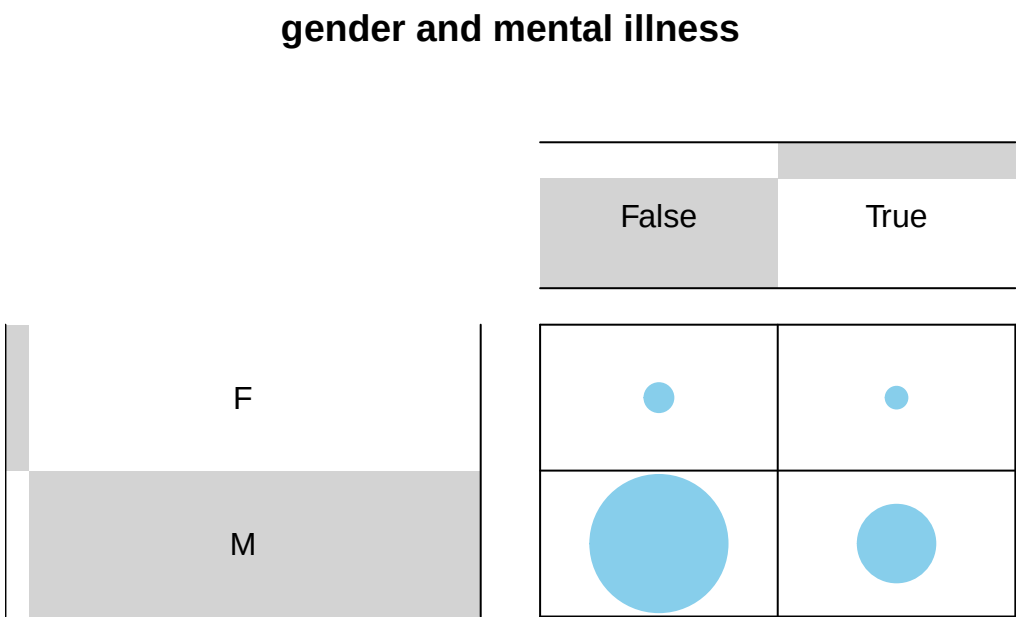
Chi squared analysis or the fisher test will be implemented to assess if there is a significant association between two characteristics of people fatally shot by the police. The null hypothesis for all assessments is that the two characteristics are independent. In places where the chi squared test was not appropriate because the category values were less than 5 or close to 5, the fisher test was implemented. This was an appropriately sized data set to use the fisher test as the data set size was not too large as to cause extensive computation times.

Gender and Mental Illness: Here we assessed the independence of the gender and signs of mental illness of people who were fatally shot by the police using the chi squared test. As shown in the plot, the largest proportion of the sample set is male and with no mental illness, and the smallest proportion is females with signs of mental illness. With $\chi^2 = 13.641$, $df = 1$, $p - value = 0.0002213$, we reject the null hypothesis that gender and signs of mental illness are independent. There is a significant relationship between the gender and the signs of mental illness in people who were fatally shot by the police.

```
gender_mental <- table(dat$gender, dat$signs_of_mental_illness)
gender_mental
```

```
##
##     False True
##   F   149   84
##   M  3360 1073
```

```
balloonplot(t(gender_mental), main ="gender and mental illness", xlab ="",
            ylab="", label = FALSE, show.margins = FALSE)
```




```
chisq.test(dat$gender, dat$signs_of_mental_illness)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  dat$gender and dat$signs_of_mental_illness
## X-squared = 16.031, df = 1, p-value = 6.232e-05
```

Gender and Race: This fisher tests was used to assess the independence of gender and race of people who were fatally shot by the police. With a $p-value = 0.0009995$, we reject the null hypothesis that gender and race are independent. There is a relationship between race and gender resulting in a higher proportion of males being shot and within that, there is a high proportion of those males being white followed by being black and then Hispanic.

```
gender_race<-table( dat$race,dat$gender)
gender_race
```

```
##
##        F    M
##   A    4   82
##   B   46 1193
##   H   29  833
##   N    5   65
##   O    3   39
##   W  146 2221
```

```
balloonplot(t(gender_race), main ="gender and race", xlab ="", ylab="",
            label = FALSE, show.margins = FALSE)
```


```
fisher.test(gender_race, simulate.p.value=TRUE)
```

```
##
```

```
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  gender_race
## p-value = 0.002499
## alternative hypothesis: two.sided
```

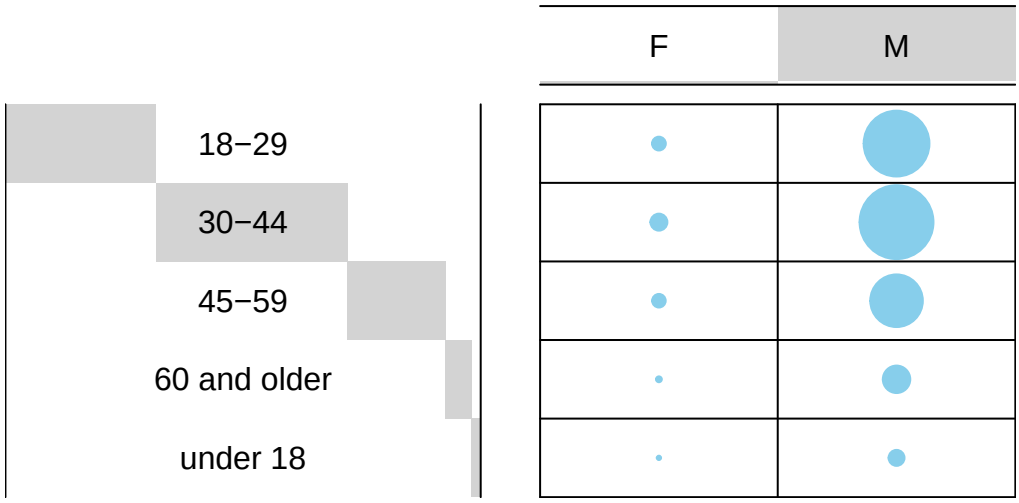
Gender and age: We used a chi squared test to assess the relationship between gender and age in our data set. The age ranges were created to reduce the possible choices for age and were set based on the age commonly considered for minors, young adults, middle age, adulthood, and older adults. Note that the "under 18" category appears at the bottom of the table. We can see that there is a higher presence of males than females and we have also seen in the previous tests. We also see that there is are significantly less people that fall into the "under 18" category and the "60 and older category. This indicated that there is a relationship between the age and gender, however the fisher test resulted in a p-value = 0.05697. We fail to reject the null hypothesis and cannot conclude that there is a significant relationship between gender and the age categories we have selected.

```
under_18<-(0:17)
a18_29<-(18:29)
a30_44<-(30:44)
a45_59<-(45:59)
a60_100<-(60:100)
dat$ageGroup <- with(dat, ifelse(age %in% under_18, "under 18",
                           ifelse(age %in% a18_29, "18-29",
                            ifelse(age %in% a30_44, "30-44",
                             ifelse(age %in% a45_59, "45-59",
                                 "60 and older" )))))

gender_age<-table( dat$ageGroup,dat$gender)
gender_age
```

```
##
##                   F    M
##   18-29          62 1416
##   30-44          94 1784
##   45-59          60  904
##   60 and older   11  246
##   under 18        6   83
```

```
balloonplot(t(gender_age), main ="gender and age", xlab ="", ylab="",
            label = FALSE, show.margins = FALSE)
```




```
fisher.test(gender_age, simulate.p.value=TRUE)
```

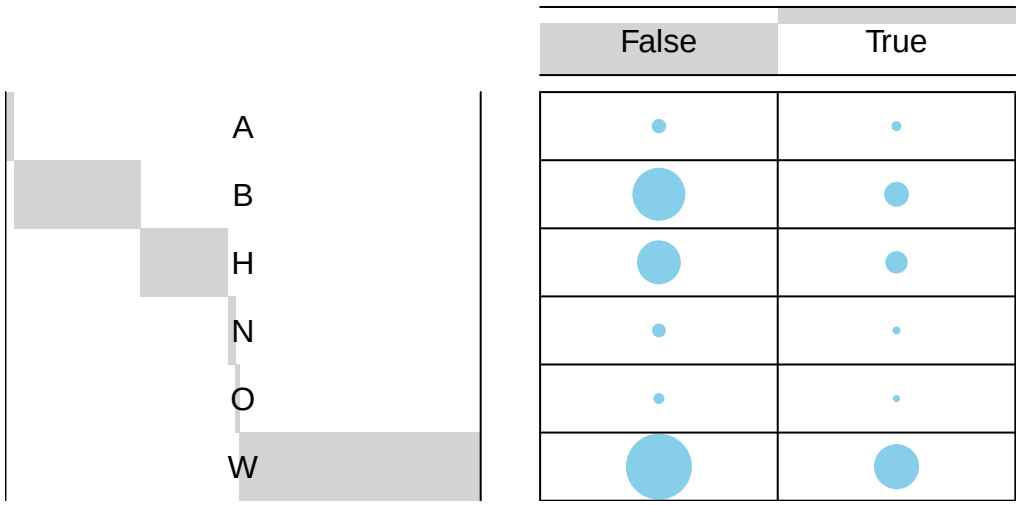
```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  gender_age
## p-value = 0.2009
## alternative hypothesis: two.sided
```

Race and Mental Illness: As shown previously, race and gender have a statistically significant relationship, we also wanted to see if there is a relationship between race and signs of mental illness. The fisher test was implemented and resulted in a p-value = 0.0004998. We reject the null hypothesis that the two factors are independent.

```
race_mental<-table( dat$race,dat$signs_of_mental_illness)
race_mental
```

```
##
##     False True
##   A    61   25
##   B  1033  206
##   H   697  165
##   N    56   14
##   O    32   10
##   W  1630  737
```

```
balloonplot(t(race_mental), main ="race and signs of mental illness", xlab ="",
            ylab="", label = FALSE, show.margins = FALSE)
```




```
fisher.test(race_mental, simulate.p.value=TRUE)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  race_mental
## p-value = 0.0004998
## alternative hypothesis: two.sided
```

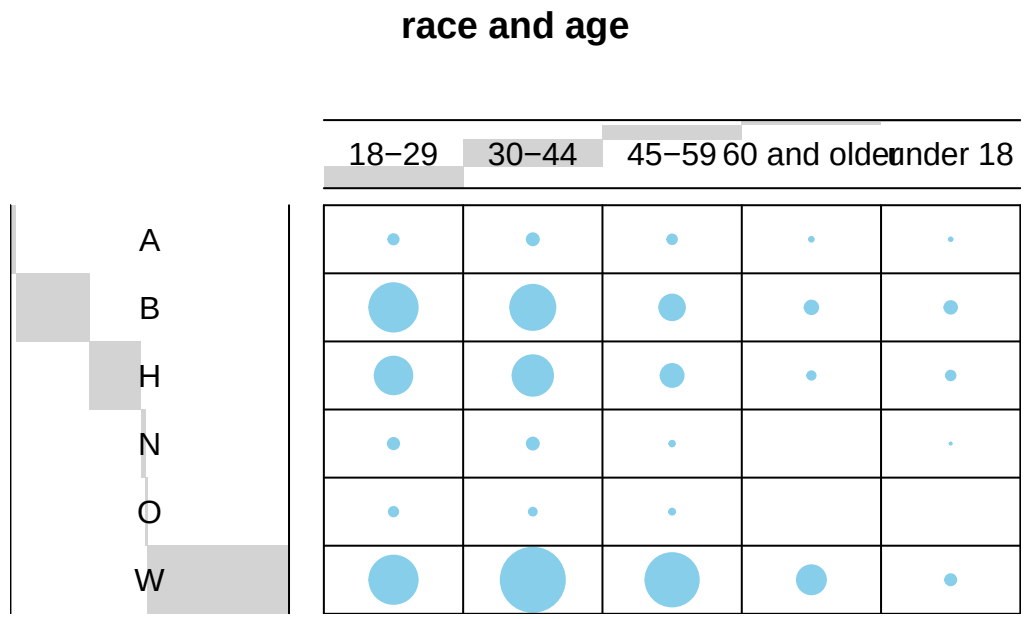
Race and Age: Using the fisher test to assess the significance of the relationship between race and age, we found a p-value of 0.0004998 meaning the two factors are not independent. This indicates that the age a civilian in shot fatally by the police is dependent on their race.

```
race_age<-table( dat$race,dat$ageGroup)
race_age
```

```
##
##     18-29 30-44 45-59 60 and older under 18
##   A    24    33    21            5        3
##   B   540   470   151           42       36
```

```
##    H   326   378   121          16        21
##    N    29    33     7           0         1
##    O    20    14     8           0         0
##    W   539   950   656         194        28
```

```
balloonplot(t(race_age), main ="race and age", xlab ="", ylab="",
            label = FALSE, show.margins = FALSE)
```




```
fisher.test(race_age, simulate.p.value=TRUE)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  race_age
## p-value = 0.0004998
## alternative hypothesis: two.sided
```

##Logistic regression We used multinomial logistic regression to predict the race based on 4 variables. The predicted variable race is both categorical and with more than two possible outcomes. The independent variables are gender (M/F), signs_of_mental_illness (True/False),body_camera (True/False), and side of the United States (West/East) all of which are dichotomous.

We set the reference level to race="W" and trained the model with 70% of the original data set. The accuracy of this model is 52.1% which the number of classifications a model correctly predicts divided by

the total number of predictions made. The final negative log-likelihood is 5394.922383 which is fairly high indicating the model is performing badly. The model performance indicates that there are likely additional or different factors in a situation which better predict the race of the person shot and killed by the police.

```
west<-c('WA','OR','CA', 'MT', 'ID', 'WY', 'NV', 'UT', 'CO', 'AZ', 'NM', 'AK',
        'HI')
dat$side <- with(dat, ifelse(state %in% west, "west", "east" ))
with(dat, table(dat$side, dat$race))
```

```
##
##           A    B    H    N    O    W
##   east   32 1019  303   26   19 1654
##   west   54  220  559   44   23  713
```

```
race_side<-table(dat$side, dat$race)
fisher.test(race_side, simulate.p.value=TRUE)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  race_side
## p-value = 0.0004998
## alternative hypothesis: two.sided
```

```
with(dat, table(dat$body_camera, dat$race))
```

```
##
##             A    B    H    N    O    W
##   False    69 1016  744   57   38 2109
##   True     17  223  118   13    4  258
```

```
race_camera<-table(dat$body_camera, dat$race)
fisher.test(race_camera, simulate.p.value=TRUE)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  race_camera
## p-value = 0.0004998
## alternative hypothesis: two.sided
```

```
#Race and other factors
dat_model<-with(dat, data.frame(race, gender, signs_of_mental_illness,
                                body_camera, side))
index <- createDataPartition(dat_model$race, p = .70, list = FALSE)
train <- dat_model[index,]
test <- dat_model[-index,]
train <- within(train, race <- as.factor(race))
train <- within(train, race <- relevel(race, ref = "W"))
#Training
multinom_model <- multinom(race ~ ., data = dat_model)
```

```
## # weights:  36 (25 variable)
## initial  value 8360.349683
## iter  10 value 5916.348111
## iter  20 value 5197.704964
## iter  30 value 5160.165811
## final  value 5159.764536
## converged
```

```
summary(multinom_model)
```

```
## Call:
## multinom(formula = race ~ ., data = dat_model)
##
## Coefficients:
##   (Intercept)    genderM signs_of_mental_illnessTrue body_cameraTrue
## B   3.4711366  0.1277618                  -0.7200182      0.17724306
## H   2.1491646  0.2823254                  -0.5143032     -0.40935838
## N   0.4116071 -0.5206642                  -0.5122184     -0.02739151
## O   0.0829082 -0.4831582                  -0.2396454     -0.79093603
## W   4.2663335 -0.3093207                   0.1283346     -0.55044412
##      sidewest
## B -2.06715798
## H  0.11813174
## N  0.00529077
## O -0.28807902
## W -1.33143247
##
## Std. Errors:
##   (Intercept)   genderM signs_of_mental_illnessTrue body_cameraTrue  sidewest
## B   0.5561397 0.5398289                  0.2533520       0.2868586 0.2364592
## H   0.5655344 0.5475702                  0.2543308       0.2908927 0.2354283
## N   0.7184491 0.6930825                  0.3840655       0.4130243 0.3346075
## O   0.8136057 0.7901991                  0.4355399       0.5945450 0.3832952
## W   0.5401928 0.5228115                  0.2445086       0.2825649 0.2286360
##
## Residual Deviance: 10319.53
## AIC: 10369.53
```

```
#exponent of the coefficients from our model to see these risk ratios.
exp(coef(multinom_model))
```

```
##   (Intercept)   genderM signs_of_mental_illnessTrue body_cameraTrue  sidewest
## B   32.173290 1.1362823                  0.4867434       1.1939213 0.1265449
## H    8.577689 1.3262101                  0.5979171       0.6640762 1.1253924
## N    1.509241 0.5941258                  0.5991649       0.9729802 1.0053048
## O    1.086442 0.6168322                  0.7869069       0.4534202 0.7497024
## W   71.259883 0.7339454                  1.1369333       0.5766936 0.2640987
```

```
#predicted probabilities for each of our outcome levels
head(round(fitted(multinom_model), 2))
```

```
##      A    B    H    N    O    W
```

```
## 1 0.04 0.08 0.28 0.02 0.01 0.57
## 2 0.03 0.14 0.38 0.03 0.01 0.41
## 3 0.01 0.36 0.11 0.01 0.01 0.51
## 4 0.04 0.08 0.28 0.02 0.01 0.57
## 5 0.03 0.14 0.38 0.03 0.01 0.41
## 6 0.01 0.36 0.11 0.01 0.01 0.51
```

```
#Validating
train$RacePredicted <- predict(multinom_model, newdata = train, "class")
train <- within(train, RacePredicted <- as.factor(RacePredicted))
train <- within(train, RacePredicted <- relevel(RacePredicted, ref = "W"))
# Building classification table
tab <- table(train$race, train$RacePredicted)
# Calculating accuracy - sum of diagonal elements divided by total obs
round((sum(diag(tab))/sum(tab))*100,2)
```

```
## [1] 52.06
```

```
#Predicting
test$RacePredicted <- predict(multinom_model, newdata = test, "class")
# Building classification table
tab1 <- table(test$Race, test$RacePredicted)
tab1
```

```
##
##        A    B    H    N    O    W
##   A    0    0    0    0    0    0
##   B    0   77    0    0    0    0
##   H    0    0   50    0    0    0
##   N    0    0    0    0    0    0
##   O    0    0    0    0    0    0
##   W    0    0    0    0    0 1270
```

Conclusion

The time series analysis shows that there are not any any specific trend over time in the number of people of each race who are fatally shot by the police.

Using two-sample tests, we found that the data for each race cannot be considered iid and the distribution for each race is significantly different. We also found that there is a statistically significant difference between the group with mental illness and the group without mental illness. There is a statistically significant difference in the observed proportions of mental illness in the data and the actual proportion of mental illness in US adults.

By creating a statistical model based on census data and the probability of being fatally shot, we found that the percentage of African-Americans who are fatally shot is significantly higher than it would if all races were equally likely. We also found that the percentage of Caucasians and the percentage of Asians who are fatally shot is significantly lower than expected if all races were equally likely. For the Hispanic population and the Native American population, the percentage is almost consistent with the case where all races are equally likely to be fatally shot.

We used the chi squared test and the fisher test to assess the independence of a few of the factors in the data set. We found that gender and mental illness, gender and race, race and mental illness, and race and

age are all independent. We also found that gender and age are not independent and there is a significant association between two characteristics.

Using a multinomial logistic regression, we attempted to predict the race of a person based on the gender, signs of mental illness, body camera presence, and side of the United Stated the shooting occurred. The model was inaccurate and performed poorly indicating that there are additional factors or the factors we selected were not as important as we had hypothesized.