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)

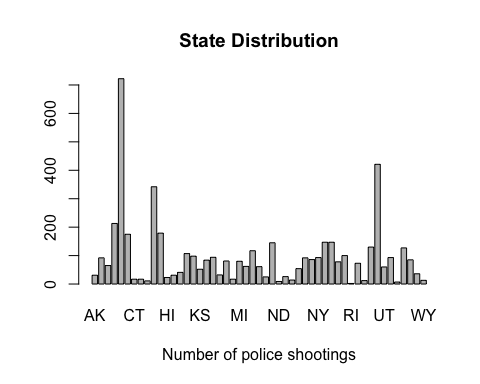
The initial data set has nrow(dat) observations of ncol(dat) features. However, there are a number of null values present in the data.

dat <- dat %>% na\_if("") %>% na.omit

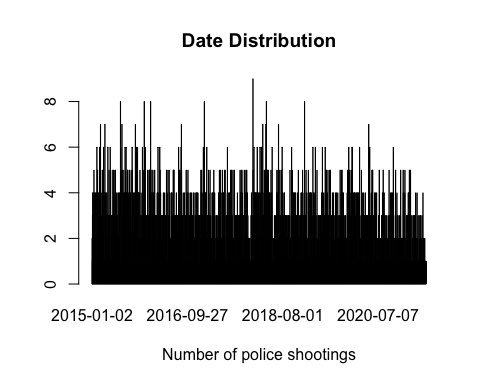
After removing the nulls and empty strings, we have nrow(dat) observations of ncol(dat) features.

### Exploratory Data Analysis

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



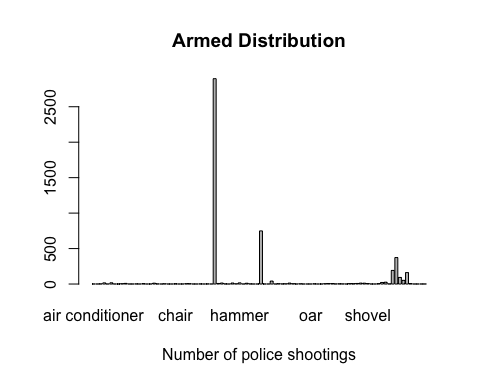
countDate <- table(dat$date)  
barplot(countDate, main="Date Distribution",  
 xlab="Number of police shootings")



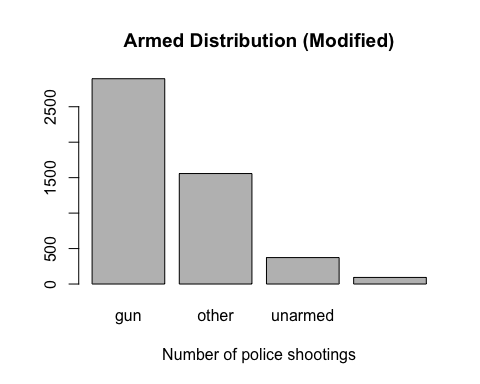
countAge <- table(dat$age)  
barplot(countAge, main="Age Distribution",  
 xlab="Number of police shootings")



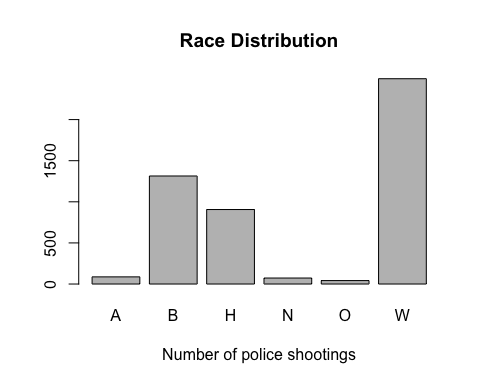
countArmed <- table(dat$armed)  
barplot(countArmed, main="Armed Distribution",  
 xlab="Number of police shootings")



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")

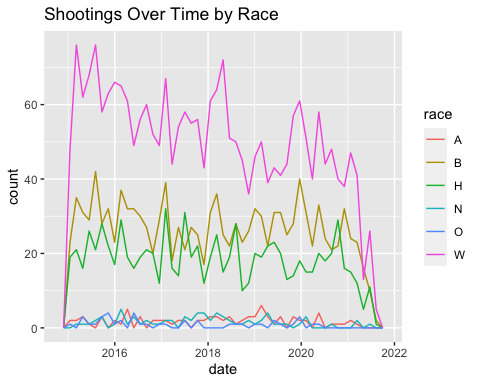


countRace <- table(dat$race)  
barplot(countRace, main="Race Distribution",  
 xlab="Number of police shootings")

 #### 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.

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 W White

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

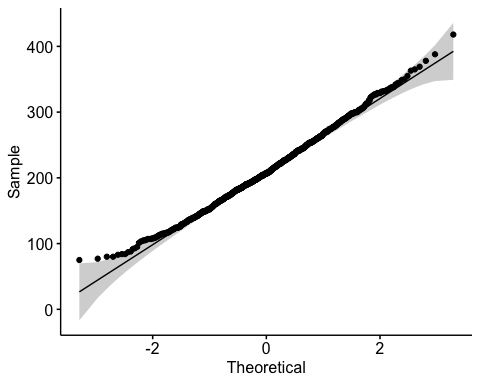
 There does not appear to be any specific trend over time in the number of people who are fatally shot.

## Statistical Model

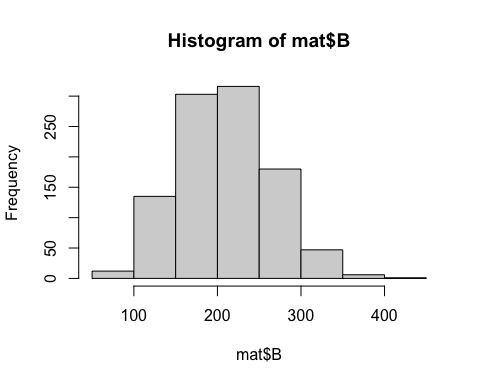
We will investigate whether number of people that are fatally shot for a given attribute in our data is consistent with a population in which each person, regardless of said attribute, is equally likely to be fatally shot. To do this, we will create 1000 samples of the general population in the US based on 2019 Census data regarding on racial proportions from <https://www.census.gov/quickfacts/fact/table/US/PST045219>. The probability of getting fatally shot regardless of race is 1/315.

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%

gen\_samp <- function() {  
 gen\_pop <- data.frame(id=1:5000, race=sample(unique(dat$race)[0:5],500000, replace=TRUE, prob = c(0.059, 0.601, 0.185, 0.134, 0.028)))  
 gen\_pop$shot <- sample(c(0, 1), 5000, replace=TRUE, prob = c(314/315, 1/315))  
 gen\_pop <- gen\_pop[gen\_pop$shot == 1,]  
 gen\_pop <- gen\_pop %>% count(race)  
 return (gen\_pop$n)  
}  
  
mat <- matrix(rep(NA, 5000), ncol=5)  
for (i in 1:1000) {  
 mat[i,] = gen\_samp()  
}  
mat <- data.frame(mat)  
colnames(mat) <- c("A", "B", "H", "O", "W")  
ggqqplot(mat$B)



hist(mat$B)



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 %>% count(race)  
race <- 'B'  
mean <- 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 = -476.98, df = 999, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0.2671275  
## 95 percent confidence interval:  
## 0.1326139 0.1337162  
## sample estimates:  
## mean of x   
## 0.133165

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 26.7%. 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 <- 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 = 223.15, df = 999, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0.5076235  
## 95 percent confidence interval:  
## 0.5954882 0.5970473  
## sample estimates:  
## mean of x   
## 0.5962677

Here we conducted the T-test for the Caucasian population. The mean of our simulated populations was approximately 59.7%, while the observed mean was 50.8%. 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 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 <- 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 = 217.35, df = 999, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0.01768652  
## 95 percent confidence interval:  
## 0.05821034 0.05894875  
## sample estimates:  
## mean of x   
## 0.05857954

Here we conducted the T-test for the Asian population. The mean of our simulated populations was approximately 5.7%, 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 <- 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 = 0.1583, df = 999, p-value = 0.8742  
## alternative hypothesis: true mean is not equal to 0.1841838  
## 95 percent confidence interval:  
## 0.1836009 0.1848690  
## sample estimates:  
## mean of x   
## 0.1842349

Here we conducted the T-test for the Hispanic population. The mean of our simulated populations was approximately 18.4%, while the observed mean was also 18.4%. The 95% confidence interval does include the observed mean, and we get a p-value of 0.7749, which is not strong evidence to reject the null hypothesis. This shows us that the percentage of Hispanics that are fatally shot is 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

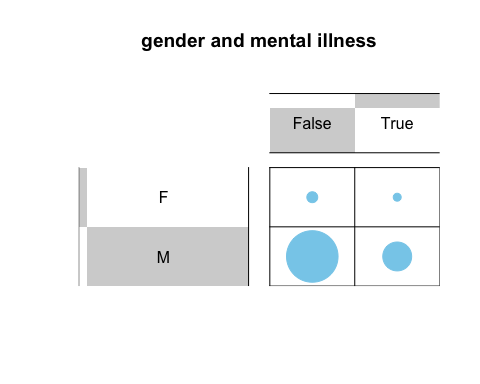
Chi squared analysis or the fisher test will be implemented to assess if there is a significant association between two characteristics of people fatally shot by the police. The null hypothesis for all assessments is that the two characteristics are independent. In places where the chi squared test was 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 , , , 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 159 84  
## M 3558 1118

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)

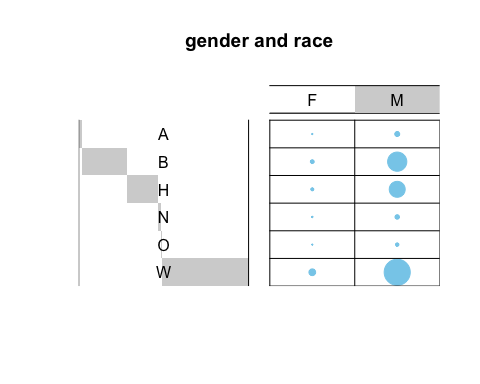
##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: dat$gender and dat$signs\_of\_mental\_illness  
## X-squared = 13.641, df = 1, p-value = 0.0002213

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 , 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 83  
## B 48 1266  
## H 29 877  
## N 5 68  
## O 3 39  
## W 154 2343

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



fisher.test(gender\_race, simulate.p.value=TRUE)

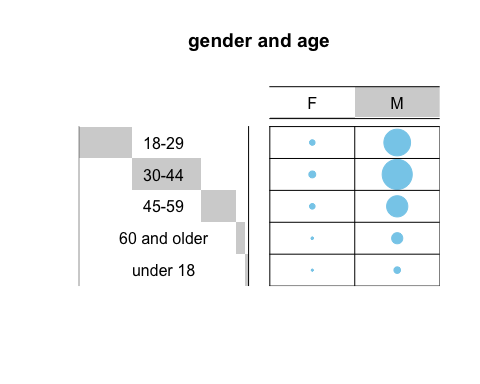
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: gender\_race  
## p-value = 0.001499  
## alternative hypothesis: two.sided

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 1491  
## 30-44 98 1896  
## 45-59 63 941  
## 60 and older 12 259  
## under 18 8 89

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 = 0.06597  
## 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 62 25  
## B 1094 220  
## H 737 169  
## N 59 14  
## O 32 10  
## W 1733 764

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)

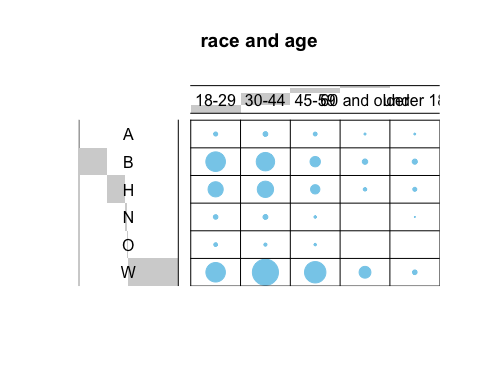
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: race\_mental  
## p-value = 0.0004998  
## alternative hypothesis: two.sided

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 4  
## B 571 504 159 42 38  
## H 342 396 128 17 23  
## N 30 34 8 0 1  
## O 20 14 8 0 0  
## W 566 1013 680 207 31

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

## 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.

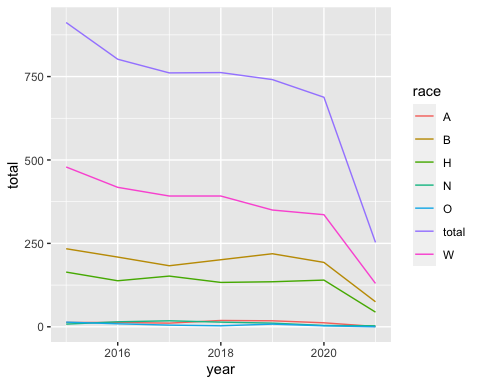
Here we construct a data frame based on the original data that shows the total number of shootings per year by race.

by\_year <- data.frame(year=2015:2021)  
by\_year$A <- (dat[dat$race == "A",] %>% mutate(year=format(date, "%Y")) %>% group\_by(year) %>% count())$n  
by\_year$B <- (dat[dat$race == "B",] %>% mutate(year=format(date, "%Y")) %>% group\_by(year) %>% count())$n  
by\_year$W <- (dat[dat$race == "W",] %>% mutate(year=format(date, "%Y")) %>% group\_by(year) %>% count())$n  
by\_year$H <- (dat[dat$race == "H",] %>% mutate(year=format(date, "%Y")) %>% group\_by(year) %>% count())$n  
O <- (dat[dat$race == "O",] %>% mutate(year=format(date, "%Y")) %>% group\_by(year) %>% count())$n  
O[7] <- 0  
by\_year$O <- O  
by\_year$N <- (dat[dat$race == "N",] %>% mutate(year=format(date, "%Y")) %>% group\_by(year) %>% count())$n  
by\_year$total <- apply(by\_year[,2:7], 1, sum)  
by\_year

## year A B W H O N total  
## 1 2015 13 234 479 164 14 8 912  
## 2 2016 13 209 418 138 9 15 802  
## 3 2017 11 183 392 152 5 18 761  
## 4 2018 19 201 392 133 3 14 762  
## 5 2019 18 219 350 135 8 11 741  
## 6 2020 12 193 336 140 3 4 688  
## 7 2021 1 75 130 44 0 3 253

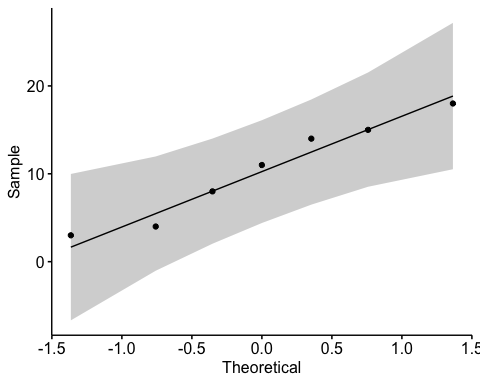
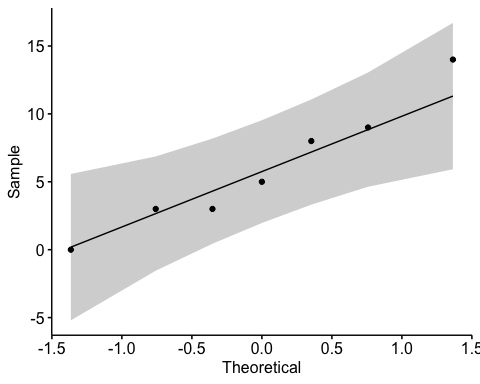
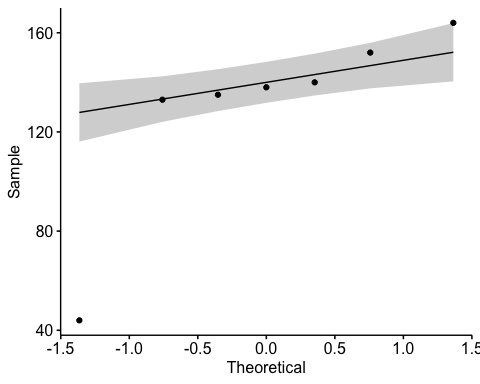
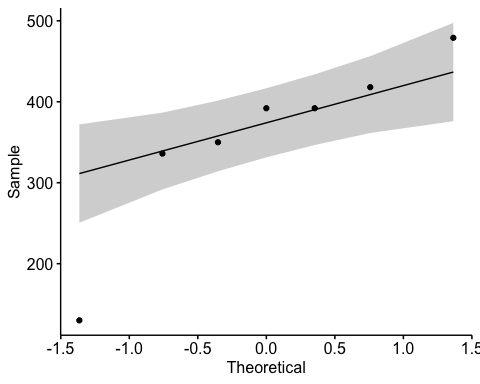
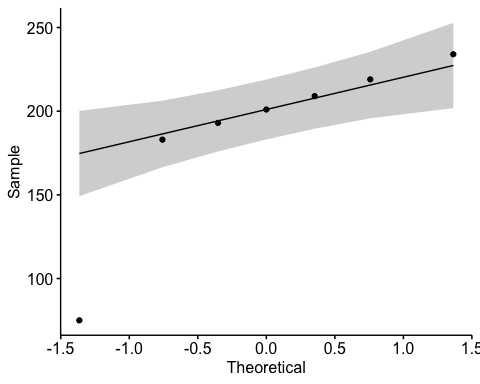
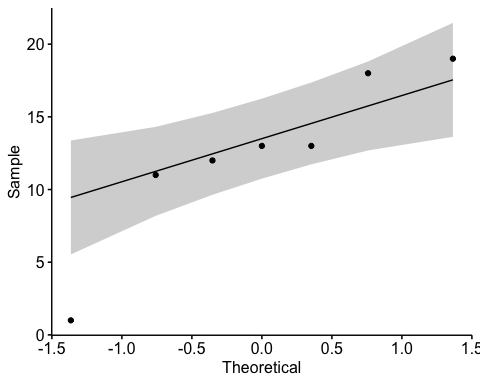
We will reformat the data frame wherein each row will represent a unique year-race combination. Examining the trend over time by race, we see no significant differences between the trends.

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.

for (i in 2:7) {  
 print(ggqqplot(by\_year[,i]))  
}



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 = 6.9864, df = 6, p-value = 0.000428  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 109.8099 228.1901  
## sample estimates:  
## mean of x   
## 169

t.test(by\_year$total - by\_year$W)

##   
## One Sample t-test  
##   
## data: by\_year$total - by\_year$W  
## t = 9.0121, df = 6, p-value = 0.0001045  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 252.0564 439.9436  
## sample estimates:  
## mean of x   
## 346

t.test(by\_year$B - by\_year$H)

##   
## One Sample t-test  
##   
## data: by\_year$B - by\_year$H  
## t = 7.449, df = 6, p-value = 0.0003016  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 39.13959 77.43184  
## sample estimates:  
## mean of x   
## 58.28571

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) %>% count()  
t.test(data=illness, n ~ signs\_of\_mental\_illness)

##   
## Welch Two Sample t-test  
##   
## data: n by signs\_of\_mental\_illness  
## t = 5.8703, df = 8.3131, p-value = 0.000323  
## alternative hypothesis: true difference in means between group False and group True is not equal to 0  
## 95 percent confidence interval:  
## 219.0689 499.5025  
## sample estimates:  
## mean in group False mean in group True   
## 531.0000 171.7143

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, it is worth noting that the mean of the group with Mental Illness is far lower than the mean of the group without, indicating that signs of mental illness do not necessarily make a person more likely to be fatally shot.