W203 Statistics for Data Science
Lab 2: Comparing Means

Monday, 4 PM

Summer 2020

July 5, 2020

Mackenzie Lee, Aditya Mengani

Table of Contents

Lab 2: Comparing Means

w203 Statistics for Data Science

The Data

The American National Election Studies (ANES) conducts surveys of voters in the United States. While its flagship survey occurs every four years at the time of each presidential election, ANES also conducts pilot studies midway between these elections. You are provided with data from the 2018 ANES Pilot Study.

For a glimpse into some of the intricacies that go into the design of this study, take a look at the introduction to the ANES User's Guide and Codebook (https://electionstudies.org/wp-content/uploads/2019/02/anes_pilot_2018_userguidecodebook.pdf).

It is important to consider the way that the ANES sample was created. Survery participants are taken from the YouGov panel, which is an online system in which users earn rewards for completing questionnaires. This feature limits the extent to which results generalize to the U.S. population.

To partially account for differences between the YouGov panel and the U.S. Population, ANES assigns a survey weight to each observation. This weight estimates the degree to which a citizen with certain observed characteristics is over- or under-represented in the sample. For the purposes of this assignment, however, you are not asked to use the survey weights. (For groups with a strong interest in survey analysis, we recommend that you read about R's survey package (http://r-survey.r-forge.r-project.org/survey/). We will assign a very small number of bonus points (up to 3) to any group that correctly applies the survey weights and includes a clear explanation of how these work).

In [1]:
```
A = read.csv("anes_pilot_2018.csv")
#Add how this dataset is a sample of the overall US Population of voters
```

In [2]:
```
#check if there are any duplicate value rows in the data
which(duplicated(A) | duplicated(A[nrow(A):1, ])[nrow(A):1])
```

There are no duplicate rows in the data.

Following is an example of a question asked on the ANES survey:

> How difficult was it for you to vote in this last election?

The variable `votehard` records answers to this question, with the following encoding:

- -1 inapplicable, legitimate skip
- 1 Not difficult at all
- 2 A little difficult
- 3 Moderately difficult
- 4 Very difficult
- 5 Extremely difficult

To see the precise form of each question, take a look at the Questionnaire Specifications (https://electionstudies.org/wp-content/uploads/2018/12/anes_pilot_2018_questionnaire.pdf).

Assignment

You will use the ANES dataset to address five research questions. For each question, you will need to operationalize the concepts (selecting appropriate variables and possibly transforming them), conduct exploratory analysis, deal with non-response and other special codes, perform sanity checks, select an appropriate hypothesis test, conduct the test, and interpret your results. When selecting a hypothesis test, you may choose from the tests covered in the async videos and readings. These include both paired and unpaired t-tests, Wilcoxon rank-sum test, Wilcoxon signed-rank test, and sign test. You may select a one-tailed or two-tailed test.

Please organize your response according to the prompts in this notebook.

In [19]:
```
# ## Submission Guidelines
# - Submit _one_ report per group.
# - Submit *both* your pdf report as well as your source file.
# - **Only analyses and comments included in your PDF report will be considered
for grading.**
# - Include names of group members on the front page of the submitted report.
# - Naming structure of submitted files:
#     - PDF report: [student_surname_1]\_[student_surname_2][\_*]\_lab\_2.pdf
#     - Jupyter Notebook: [student_surname_1]\_[student_surname_2][\_*]\_lab\_
2.ipynb
```

Research Questions

In [3]:
```
#Dataframe shape
paste('# columns:',ncol(A))
paste('# rows:',nrow(A))
```

'# columns: 767'

'# rows: 2500'

Question 1: Do US voters have more respect for the police or for journalists?

Introduce your topic briefly. (5 points)

Explain how your variables are operationalized. Comment on any gaps that you can identify between your operational definitions and the concepts you are trying to study.

2020 has been rife with political distrust in all levels of life. The president has called news agencies and journalists a source of fake new, while citizens protest against police brutality. Both positions are generally acknowledged to be necessary for society, but with abuses in both fields, it would be interesting to compare how US voters view both positions.

To answer this question, we must first find the relevant variables from the data set relating to police and journalists. The relevant variables are listed below:

Police variables*:

- ftpolice
- ftpolice_page_timing
- ftpolice_skp
- ord_ftpolice

Journalist variables*:

- ftjournal
- ftjournal_page_timing
- ftjournal_skp
- ord_ftjournal

Based on the available variables in the data set, we will use *ftpolice* to gauge respect towards police and *ftjournal* to gauge respect toward journalists. We do not need to worry about the other variables as they play no role in relating to respect and relate only to the survey design. By selecting these variables, we must examine the rating scale and meaning that the survey assigned to the responses. The survey used a thermometer to visually represent ratings and respondent would click on the thermometer which had numbers ranging from 0 to 100 and descriptions next to chosen increments in the following manner:

It is important to note that ratings are not equally spaced. Ratings increase by intervals of 15 from 0 to 30, but then increase by 10 from 40 to 70 and back to intervals of 15 from 70 to 100. These ratings suggest that ratings in the range from 30 to 70 carry greater weight than those that fall outside this range. The descriptions for the ranges below 30 and above 70 all have the same description in terms of favor and can be seen as affecting responses. It is necessary to translate this scale in order to answer the question, so we assume that favorable feelings translate to respect.

Using this scale, we assume that higher ratings correspond to greater respect. A rating between 0 and 50 means people view police or journalists negatively and disrepectfully while a rating between 50 and 100 means people view them respectfully. A rating of 50 means a person doesn't care or does not lean in either direction of respect.**

**see below for definitions*

***see note below*

In [ ]:
```
#Scale for choosing ratings:
#ftpolice (how would you rate the police?)
#ftjournal (how would you rate journalists?)
# 100; Very warm or favorable feeling
# 85; Quite warm or favorable feeling
# 70; Fairly warm or favorable feling
# 60; A bit more warm or favorable feeling than cold
# 50; No feeling at all
# 40; A bit more cold or unfavorable than warm
# 30; Fairly cold or unfavorable feeling
# 15; Quite cold or unfavorable feeling
# 0; Very cold or unfavorable feeling

# The questionairre states that ratings between 50 and 100 correspond to favora
ble feelings
# while ratings between 0 and 50 mean that you feel unfavorable toward that per
son.
# A rating of 50 would mean you do not feel particularly warm or cold toward th
e person.


#Related variables we do not use:
# ftpolice_page_timing (how much time (seconds) someone spent on this question)
# ftpolice_skp (indicates if respondent skipped the rating question)
# ord_ftpolice (indicates the order in which this question appeared out of all
the rating questions)

# ftjournal_page_timing*
# ftjournal_skp*
# ord_ftjournal*

#* same format as police variables

#Note:
# An important idea to explore with this scale is how to interpret and compare
results and
# how to account for individual interpretation of the response. For example, ca
n we assume a
# rating of 78 is more respectful than a 79? The original scale and interpretat
ion using favor had
# 9 buckets (i.e. descriptions), and we choose to keep this format because if n
ot, we would have to
# transform the ratings with our own buckets and group values. A possible trans
form is:

# 1. 0 - 9 (no respect)
# 2. 10 - 19
# 3. 20 - 29
# 4. 30 - 39
# 5. 40 - 49
# 6. 50 (neutral)
# 7. 51 - 59
# 8. 60 - 69
# 9. 70 - 79
# 10. 80 - 89
# 11. 90 - 100 (utmost respect)

# Even with these groupings, we still see a loss of data in the sense that diff
erent ratings
# in the same bucket will be treated equally. Personal interpretation of the de
scriptions would
# affect responses and for some people the same score may mean different thing
```

```
s, so this approach
# may seem appropriate. However, the differences in buckets could misconstrue r
esults. For example,
# a rating of 11 and 29 seem to have a significant difference from face value.
Using the above transform,
# they would be placed in the 2nd and 3rd buckets and this situation would be a
nalagous to values
# of 19 and 20 in the same 2nd and 3rd buckets; these values show a smaller dif
ference but are treated
# equally. To account for this, we can say that smaller differences in value co
rrespond to smaller
# differences in respect. We note that a value of -7 means the respondent had n
o answer and we will
# drop this data.
```

Perform an exploratory data analysis (EDA) of the relevant variables. (5 points)

This should include a treatment of non-response and other special codes, basic sanity checks, and a justification for any values that are removed. Use visual tools to assess the relationship among your variables and comment on any features you find.

In [4]:
```
#Checking for missing values
(which(is.na(A$ftpolice)))
(which(is.na(A$ftjournal)))
```

In [1]:
```
#There are no missing values in the data.
```

In [5]:
```
paste("Summary of ftpolice showing respect for police:")
summary(A$ftpolice)
paste("Summary of ftjournal showing respect for journalists:")
summary(A$ftjournal)
```

'Summary of ftpolice showing respect for police:'

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00   47.00   70.00   64.68   90.00  100.00 
```

'Summary of ftjournal showing respect for journalists:'

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  -7.00   21.00   52.00   52.26   82.00  100.00 
```

In [104]:
```
#Both variables are numerical and real-valued ranging from 0 to 100 (integers).
#-7 appears only in *ftjournalists* showing some respondent(s) did not provide
a rating.
#The descriptive statistics for police are generally higher for all values,
#except the max since the rating scale is the same, suggesting that police were
viewed more
#respectfully than police. The mean and median support this idea as they are bo
th higher than those
#of the ratings toward journalists which may mean there are more higher ratings
for police than for journalists.
```

In [6]:
```
#Removing negative values
neg_journal = A$ftjournal[A$ftjournal <0]
neg_police = A$ftpolice[A$ftpolice <0]

paste("# negative ratings for journalists: ", length(neg_journal))
paste("# negative ratings for police: ", length(neg_police))

pos_journal = A$ftjournal[A$ftjournal >=0]
pos_police = A$ftpolice[A$ftpolice >=0]

paste("# total ratings for journalists: ", length(pos_journal))
paste("# total ratings for police: ", length(pos_police))
```

'# negative ratings for journalists: 2'

'# negative ratings for police: 0'

'# total ratings for journalists: 2498'

'# total ratings for police: 2500'

In [105]:
```
#Only *ftjournal* had negative values and we cannot impute these negative value
s
#since that would be adding false data. Replacing a negative value with any val
ue
#in the range from 0 to 100 would be saying that respondent had that specific a
ttitude of respect
#toward journalists and this would be a false representation. Adding another va
lue outside this
#range would not make sense since it would be unintellible according to the rat
ing scale.
#Consequently, removing them is the most appropriate case. Because we are deali
ng with one sample and
#two different opinions from the same people, we cannot remove only the values
from journalists as we must
#compare that single person's views of journalists and the police. Our sample s
ize is now 2,498.
```

In [7]:
```
#Removing values
A[A$ftjournal < 0 , c('ftjournal', 'ftpolice')]

#Removing the rows
journal = A[-c(51, 597),]$ftjournal
police = A[-c(51, 597),]$ftpolice
```

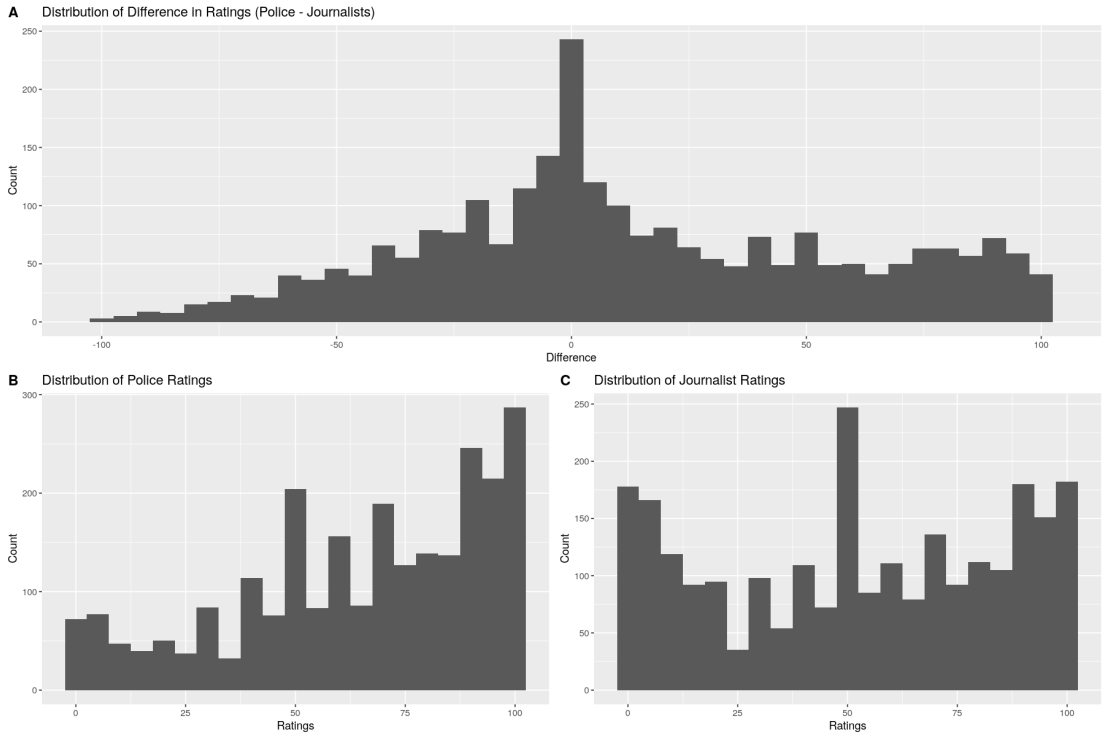
A data.frame: 2 × 2

	ftjournal	ftpolice
	<int>	<int>
51	-7	84
597	-7	91

In [20]:
```
#The two displayed rows are the two rows in the data in which there is no
#journalist rating. As discussed, we will be removing them, but it is important
#to note the police rating values. Both values are high on the rating scale lea
ning
#to utmost respect. As we continue the analysis, it will be important to rememb
er that two high
#rating police values were removed.

#We cannot impute any values since that would add false data (i.e. imputing 0 f
or -7 means
#we add a rating of Viewed with no respect which could skew data).
```

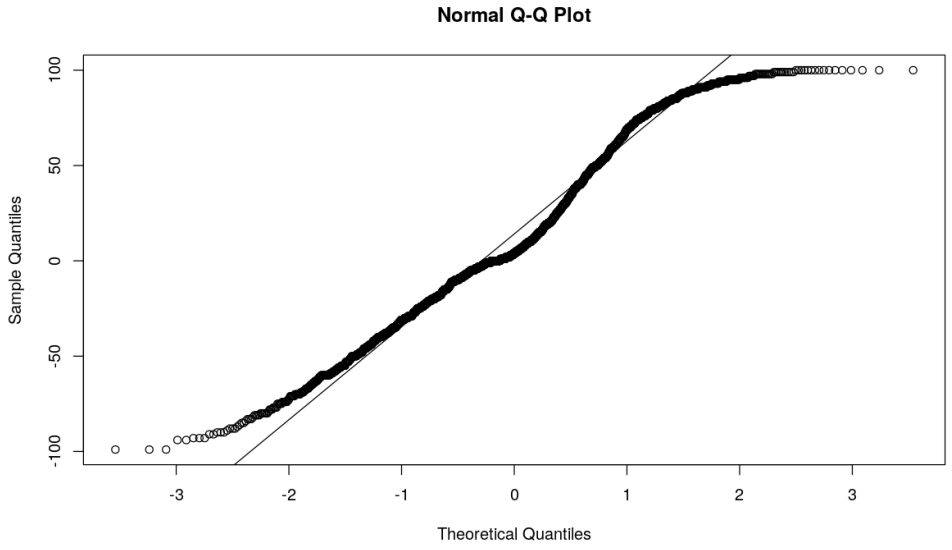
In [8]:
```
#Controls plot size
#necessary libraries
library(ggplot2)
library(gridExtra)
library(repr)

p1 = qplot(police, geom="histogram", binwidth = 5,xlab = 'Ratings',
           ylab = 'Count', main = 'Distribution of Police Ratings')
p2 = qplot(journal, geom="histogram", binwidth = 5,xlab = 'Ratings',
           ylab = 'Count', main = 'Distribution of Journalist Ratings')
p3 = qplot(police-journal, geom="histogram", binwidth = 5,xlab = 'Difference',
           ylab = 'Count', main = 'Distribution of Difference in Ratings (Polic
e - Journalists)')
```

In [9]:
```
library(ggpubr)
options(repr.plot.width=15, repr.plot.height=10)
#Arranging layout of the plots
ggarrange(p3,
          ggarrange(p1, p2, ncol = 2, labels = c("B", "C")),
          nrow = 2,
          labels = "A"
          )
```



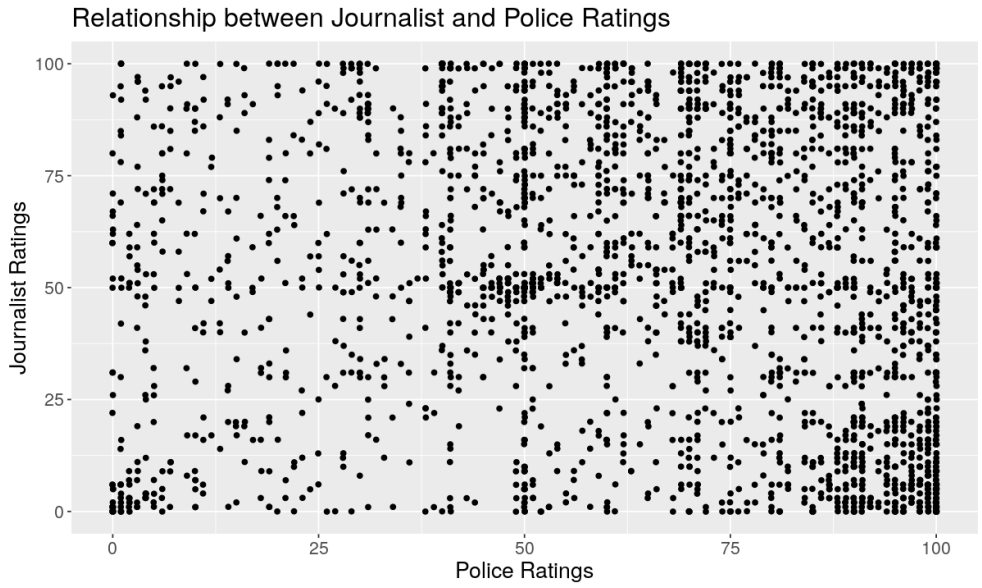
In [19]:
```
# Beginning with plots B and C in the bottom of the above plots, we see that ne
ither variable
# for police or journalist ratings follows a normal distribution. Plot B showin
g police ratings
# is left skewed, but shows a slight peak at 50 indicating many people felt no
particular way toward police.
# The two removed values would have fallen on the right side of the plot where
the majority of the data lies,
# so we can assume that the loss of data should not have a significant impact s
ince the removed values were not
#outliers. In plot C, the distribution of journalists seems trimodal with peaks
around 0, 50, and 100
# suggesting a wide range of stronger feelings compared to those toward police.

# Plot A showing the differences in the ratings more closely follows a normal d
istribution.
# Since we are subtracting journalist ratings from police ratings, a positive v
alue in this
# plot indicates that the person gave a higher rating to police than journalist
s and a negative
# value indicates the person gave a higher rating to journalists than to polic
e.
# A difference of 0 indicates the person gave the same rating to both police an
d journalists.
# The distribution is not perfectly normal and we can see more positive values
indicating that
# more people rated police better than journalists, but there is a central peak
at 0.
```

In [10]:
```
#normality of differences
diff = police - journal
options(repr.plot.width=10, repr.plot.height=6)

qqnorm(diff)
qqline(diff)
```



In [20]:
```
#The above Q-Q plot is a way to visualize how normal a distribution is.
#A perfectly normal distribution would fall on the diagonal line and deviations
from
#the line indicate non-normality. We can see from this plot that there are devi
ations
#primarly at the lower and higher values of theoretical quantiles, but most of
the values still
#fall on the line. Thus, we can assume that the differences between police and
journalist
#ratings are approximately normality.
```

In [11]:
```
#Scatterplot
rating = do.call(rbind, Map(data.frame, A=police, B=journal))
colnames(rating) = c('police', 'journal')

options(repr.plot.width=10, repr.plot.height=6)
ggplot(rating, aes(x=police, y=journal)) +
    geom_point(position = position_jitter(w = 0.05, h = 0.05))+ #Adding jitter
    labs(x='Police Ratings', y='Journalist Ratings')+
    theme(text = element_text(size=16))+
    ggtitle('Relationship between Journalist and Police Ratings')
```



In [22]:
```
#The above scatterplot shows the relationship between journalist and police rat
ings.
#We see from the plot that there is no relationship between the two variables.
#There does seem to be some heavy spread of values that are above 50 in that ma
ny ratings
#for police that are over 50 also have journalist ratings that are also over 5
0,
#but these values still do not show any meaningful relationship
```

Based on your EDA, select an appropriate hypothesis test. (5 points)

Explain why your test is the most appropriate choice. List and evaluate all assumptions for your test.

We will be conducting a dependent samples t-test to answer the question.*

The important assumption of the dependent t-test is that the difference between ratings should be normal. The Q-Q plot and histogram of the differences between police and journalist ratings show that the differences are not perfectly normal, but they are approximately normal. Accounting for the fact that the t-test is robust to deviations from normality, we can use the dependent t-test.

We setup our test as such where μ_p is the mean ratings for police and μ_j is the mean ratings for journalists:

$H_0 : \mu_p = \mu_j$

$H_A : \mu_p \neq \mu_j$

We use a two-sided test at the 5% significance level since we have no reason to believe that US voters would respect police more or less than journalists. Additionally, using a one-sided test would increase our rejection region and chance of rejecting the null hypothesis.

**see below for justification*

In [ ]:
```
#Test justification
# We are comparing respect for two different groups (i.e. police and journalist
s),
# but we are comparing the opinions of the same people (i.e. US voters). Thus,
we cannot
# treat the two groups (i.e. two sets of ratings) as independent since they are
two different
# opinions of the same people. Additionally, we are interested in how police ra
tings compare to
# journalist ratings (both of which have the same units), so there is a natural
pairing between
# the data points. We do not use a non-parametric test, since those tests are m
ore appropriate for
# ordinal rather than numerical variables. While there were ordered description
s for the ratings,
# we are using the numerical scale from 0 to 100 to interpret the results as di
scussed in the
# introduction for the topic. Consequently, the ratings are numerical rather th
an ordinal.
```

Conduct your test. (5 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result. Make sure you relate your findings to the original research question.

In [12]:
```
#Dependent t-test using police ratings as the first group and journalist rating
s as the second group
test = t.test(police, journal, paired = T)
test
```

```
        Paired t-test

data:  police and journal
t = 13.711, df = 2497, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 10.58776 14.12160
sample estimates:
mean of the differences
               12.35468
```

In [13]:
```
#Effect size correlation calculation
t = test$statistic[[1]]
df = test$parameter[[1]]
r = sqrt((t^2)/(t^2+df))
round(r,3)
```

0.265

In [69]:
```
#Looking at the results, we see that our t-value is large at 13.711 and  there
is a small p value <2.2e-16
#suggesting that we can reject the null hypothesis in favor of the alternative
hypothesis
#that the means of the two groups are not equal at the 5% significance level.
#This result is highly statistically significant since our p-value is very smal
l.
#This idea is supported by the 95% confidence interval of (10.588, 14.122),
#which does not contain 0 and suggest that a difference of 0 does not fall in t
he confidence interval.
#The mean of the differences is positive suggesting that the police ratings ten
ded to higher than
#journalist ratings.

#The correlation value is small 0.265 suggesting a small effect size.
#While the result is statistically significant, the effect size is small,
#so not practically significant suggesting that we cannot take the result to me
an that
#US voters definitely view police more respectably than journalists.
#This sample may have reflected those results, but the small effect size sugges
ts this result
#should not be seen as fact. We can look at the mean of the differences to supp
ort
#the idea that police have a higher rating than journalists since the differenc
e is positive 12.355
#saying the mean rating difference between police and journalist ratings is abo
ut 12 points.
#Since we are interpreting our results numerically, we can say that police
#have more respectful views, but the difference is not noticeably large keeping
in mind the scale is out
#of 100. Perhaps US voters may view police more respectably than journalists,
#but the difference may be quite small practically speaking.
```

Question 2: Are Republican voters older or younger than Democratic voters?

Introduce your topic briefly. (5 points)

Explain how your variables are operationalized. Comment on any gaps that you can identify between your operational definitions and the concepts you are trying to study.

Younger generations tend to be viewed as more liberal than their older counterparts. This means younger voters may lean Democrat and this question will help provide insight into whether this statement holds true.

In order to solve this question, we determined that the question speaks about "political affiliation of voter" and "relative age of the voter groups" belonging to Republican and Democrats.

a) Political Affiliation(Democrat vs Republican):

variables that determine the political party affiliation from the dataset. Thus we identified below variable identifies the party affiliation of the voter.

- pid7x*

Assumptions and Reasoning for choosing pid7x:

We had two choices for determining political affiliation of voter:

1. Use only pid7x and give own definition for Republic/Democratic voter
2. Use pid7x + other columns based on own definition

We decided to use the option 1.** Won't use second since that is like the safe approach.

***see note below*

b) Age of voter group(Young or Old):

Similarly, as the question speaks about the voter being younger and older, we investigated for the variables that determine age.

We did not find any direct variable that determines the age of the voter, except "birthyr" that determines the year born.

- birthyr*

This new variable "age" = 2018 - "birthyr" can be accurately used to determine the relative age of voter groups.

**see below for definitions*

In [30]:
```
#Scale for:
#pid7x (Measure voter self-identification toward party)
# -7 - no answer
# 1 - Strong dem
# 2 - Not very strong dem
# 3 - Ind, closer to dem
# 4 - Independent
# 5 - Ind, closer to rep
# 6 - Not very strong rep
# 7 - Strong rep

#Definition
#birthyr (profile variable determining the year of birth of voter)
# Using this we designed a rule to determine age based on the below observatio
n:
# The ANES study was implemented in Dec 21, 2018. So  use 2018 as the base year
which
# can be used to determine the age of the voter when combined with the "birthy
r".


#Note:
# Option 1 : For our study we identified a person as a Republican/Democrat if t
hey self-identify
# themselves as belonging to those parties. Then we can exclude voters who have
choosen either -7,4
# as they have not responded or are an independent voter who can have a alterin
g party affiliation.
# So, using this approach we can only keep options 1,2,3,5,6,7 who have some in
clination
# for Democrat/Republic.

# Option 2: Instead, If we assume that the Republic/Democrat groups refers to p
eople who would
# vote for those parties rather than people who self-identified as belonging to
those parties then
# we can keep all categories except 4, -7 but then would need to find what they
actually voted for
#     to check. Using another variable could drop more data.
# Further, for this approach we could use other variables like house18p","senat
e18p","gov18p"
# but decided not to as this is dropping a lot of data, which we felt is a safe
approach.
```

Perform an exploratory data analysis (EDA) of the relevant variables. (5 points)

This should include a treatment of non-response and other special codes, basic sanity checks, and a justification for any values that are removed. Use visual tools to assess the relationship among your variables and comment on any features you find.

In [14]:
```
#datatype of variable birthyr is integer
paste("The data type and sample distribution for birthyr is real-valued : ")
str(A$birthyr)
```

'The data type and sample distribution for birthyr is real-valued : '

```
 int [1:2500] 1986 1972 1999 1975 1989 1992 1960 1962 1978 1957 ...
```

In [15]:
```
#datatype of variable birthyr
paste("The data type and sample distribution for pid7x is real valued: ")
str(A$pid7x)
```

'The data type and sample distribution for pid7x is real valued: '

```
 int [1:2500] 6 6 3 4 6 2 3 7 7 2 ...
```

In [16]:
```
#Created a new variable called "age"
A$age = 2018 - A$birthyr

paste("Summary of age variable : ")
summary(A$age)
```

'Summary of age variable : '

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.00   35.00   52.00   49.48   62.00   91.00 
```

In [17]:
```
#datatype for A$age variable
paste("The data type and sample distribution for age is real valued : ")
str(A$age)
```

'The data type and sample distribution for age is real valued : '

```
 num [1:2500] 32 46 19 43 29 26 58 56 40 61 ...
```

In [18]:
```
# The datatype for age is a numeric variable.

# The age variable represents a true age of voter and do not have any bad value
s like minor age (<18), >100,
# or negative  age. Hence this can be used for measurement without any further
treatment as there is no bad
# data that needs to be imputed
```

In [19]:
```
#Checking for missing values in age
paste("Checking the null functions in age : ")
(which(is.na(A$age)))
```

'Checking the null functions in age : '

There are no missing values in age

In [20]:
```
# Republican voter group
# We have choosen the voters who have choosen 5,6,7 as a republican voter
paste("Summary distribution of republican group : ")
republican <- A$age[(A$pid7x == 5) | (A$pid7x == 6) | (A$pid7x == 7)]
summary(republican)
paste("# total republican count : ", length(republican))
```

'Summary distribution of republican group : '

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.00   41.00   56.00   53.33   66.00   90.00 
```

'# total republican count : 849'

In [21]:
```
# Democrat voter group
# we have choosen the voters who have choosen 1,2,3 as a democrat voter
paste("Summary distribution of democrat group : ")
democrat <- A$age[(A$pid7x == 1) | (A$pid7x == 2) | (A$pid7x == 3)]
summary(democrat)
paste("# total democrat count: ", length(democrat))
```

'Summary distribution of democrat group : '

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.00   35.00   52.00   49.64   63.00   91.00 
```

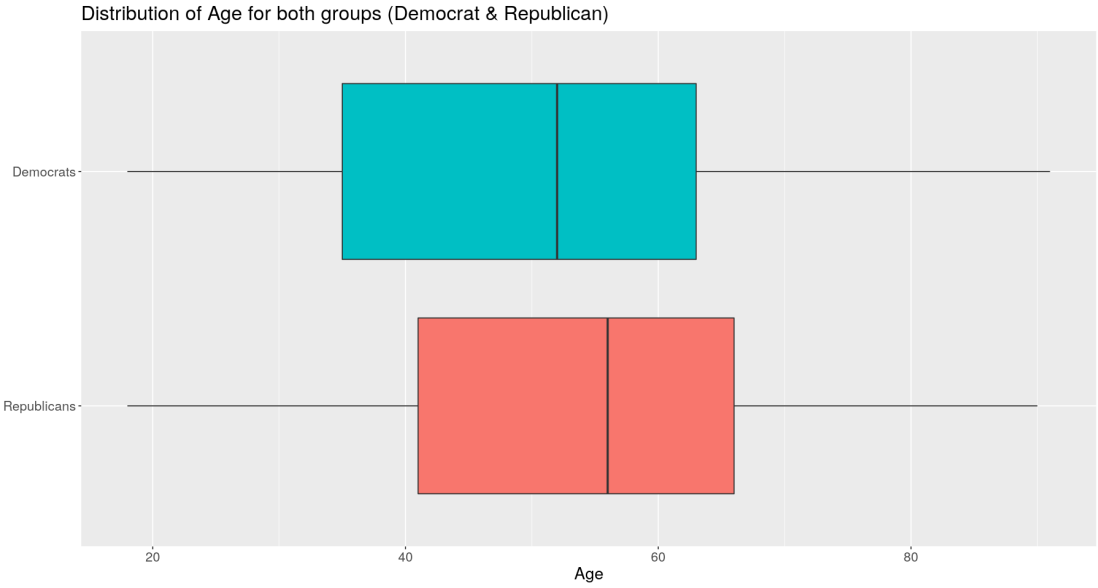
'# total democrat count: 1136'

In [24]:
```
#Checking for missing values
paste("Checking the missing values for democrat and republic groups : ")
(which(is.na(democrat)))
(which(is.na(republican)))
```

'Checking the missing values for democrat and republic groups : '

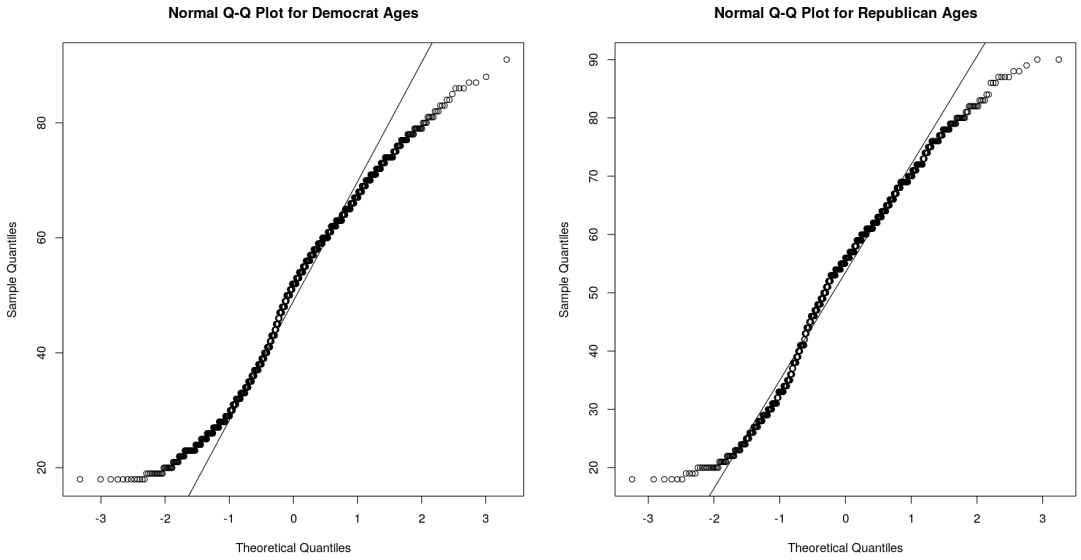
In [2]:
```
# There are no missing values for democrat and republican variables.
# For variable pid7x, we exclude voters who choose option 4(independent) and -7
(no-answer) from our analysis.
```

In [25]:
```
library(ggplot2)
options(repr.plot.width=15, repr.plot.height=8)

repdem = suppressWarnings(do.call(rbind, Map(data.frame, A=republican, B=democr
at)))
colnames(repdem) = c('rep', 'dem')
library(reshape)
repdem = melt(repdem, measure.vars = c('rep', 'dem'))

ggplot(aes(y = value, x = variable, fill = variable), data = repdem) +
    geom_boxplot(show.legend = FALSE)+
    scale_x_discrete(labels= c('Republicans', 'Democrats'))+
    ggtitle('Distribution of Age for both groups (Democrat & Republican)')+
    theme(text = element_text(size=16), axis.title.y= element_blank())+
    labs(y='Age')+
    coord_flip()
```



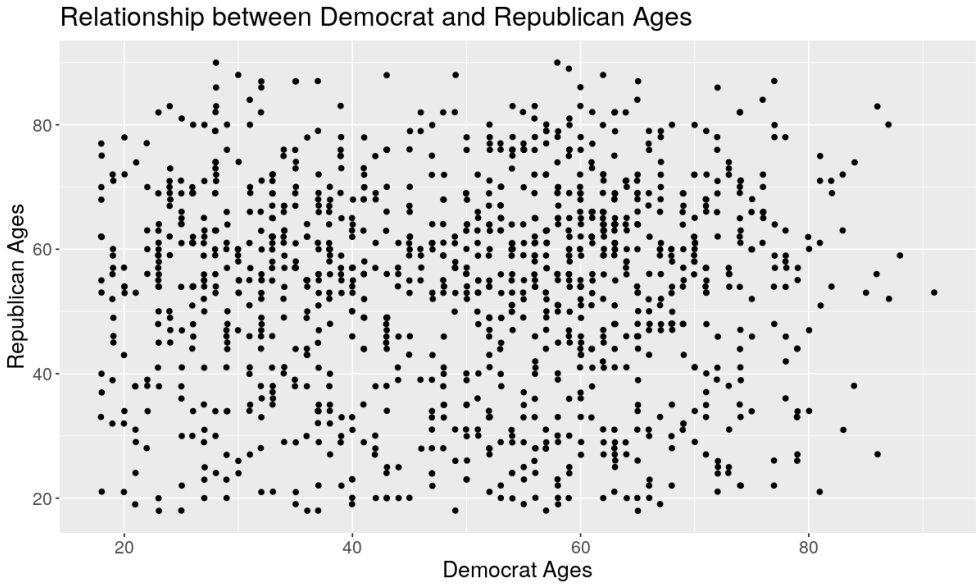
In [3]:
```
# The above Q-Q plots for two datasets show how normal their distributions are.
# A perfectly normal distribution will fall on diagonal line and deviations ind
icate non-normality.
# From above plots there are deviations primarily at lower and higher values of
theoretical quantiles,
# but most of the values still fall on the line. Thus, we can assume that the d
ifferences between Democrat
# and Republican age groups are approximately normal.
```

In [26]:
```
par(mfrow=c(1,2))

options(repr.plot.width=15, repr.plot.height=8)

qqnorm(democrat, main ='Normal Q-Q Plot for Democrat Ages')
qqline(democrat)

qqnorm(republican, main ='Normal Q-Q Plot for Republican Ages')
qqline(republican)
```



In [4]:
```
# We can see from the scatterplot that there is no relationship between the two ages of both groups.
# There is no trend nor relationship we can see.
```

In [27]:
```
#Scatterplot
demrep = suppressWarnings(do.call(rbind, Map(data.frame, A=democrat, B=republic
an)))
colnames(demrep) = c('dem', 'rep')
library(ggplot2)
options(repr.plot.width=10, repr.plot.height=6)
ggplot(demrep, aes(x=dem, y=rep)) +
    geom_point(position = position_jitter(w = 0.05, h = 0.05))+ #Adding jitter
    labs(x='Democrat Ages', y='Republican Ages')+
    theme(text = element_text(size=16))+
    ggtitle('Relationship between Democrat and Republican Ages')
```



We can see from the scatterplot that there is no relationship between the two ages of both groups. There is no trend nor relationship we can see.

Based on your EDA, select an appropriate hypothesis test. (5 points)

Explain why your test is the most appropriate choice. List and evaluate all assumptions for your test.

Based on the exploratory data analysis, we choose to do the independent 2-samples, 2-tailed t-test to answer the question.*

The important assumptions of the independent t-test are:

- T-test allows us to test mean differences while accounting for variablity.
- Assumption of i.i.d.: Democrats and Republicans are independent as they represent different sets of people. According to the survey design, we can assume respondents are drawn from the same distribution and are unrelated, thus independent of each other.
- Assumption of Normality: Mean distributions of age groups for Democrats and Republican voters as per the QQ-plot show that the differences are not perfectly normal, but they are approximately normal. The independent t-test is robust to deviations from normality. Also with CLT for large sample sizes (>30), our data can be approximately normal with the t-distribution.

Thus, we use a parametric independent samples t-test for our analysis, we are dealing with a numeric variable age to compare between the two group samples(Democrat and Republic). Moreover t-test measures mean differences while accounting for variability between the two groups. Let μ_R represent the mean age of Republicans and μ_D represent the mean age of Democrats.

Hence our null Hypotheses would be: $H_0 : \mu_R = \mu_D$

Alternate Hypotheses would be: $H_A : \mu_R \neq \mu_D$

We choose to use a two-sided independent t-test as the question speaks about Republican voters younger or older than Democratic voters. That would mean we need to consider the effects on both the sides.

see below for justification

In [ ]:
```
#Test justification
# Even though we are measuring the same variable(age),
# we are actually trying to compare the data between two sets of samples for
# different groups(Democrats and Republicans).
```

Conduct your test. (5 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result. Make sure you relate your findings to the original research question.

In [28]:
```
# perform independent t-test
paste("The t-test result is : ")
tt_dem_rep <- t.test(republican,democrat)
tt_dem_rep
```

'The t-test result is : '

```
        Welch Two Sample t-test

data:  republican and democrat
t = 4.805, df = 1823.3, p-value = 1.674e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.182347 5.192611
sample estimates:
mean of x mean of y
 53.32744  49.63996
```

In [29]:
```
#different size of republicans and democrat
# correlation effect
paste("The correlation effect measured is ")
t_dem_rep = tt_dem_rep$statistic[[1]]
df_dem_rep = tt_dem_rep$parameter[[1]]
r_dem_rep = sqrt((t_dem_rep^2)/(t_dem_rep^2+df_dem_rep))
round(r_dem_rep,3)
```

'The correlation effect measured is '

0.112

In [16]:
```
#Statistical Significance:
# At a  5% significance level, we can reject the null hypothesis since our p-va
lue is
#very small  < (1.674e-06) and the result is highly statistically significant.

# #Practical Significance:
# On assessing the correlation, we see that the r estimate is 0.112 (which is s
ignificantly low effect size).
#This would mean that even though our result is statistically significant it is
not practically siginificant,
#due to the large overlap in the sample data between the two groups.

# Considering CI range mentioned below has positive values and mean of x is hig
her than that of y,
# 95 percent confidence interval: (2.182, 5.193)
# sample estimates:
# mean of Democrat ages =  53.327 and mean of Republican ages= 49.640


# w.r.t question, based on the practical and statistical signifiance measured,
#we can conclude that the Republican voters are older than the democrat voters
statistically,
#as the mean age of Republican voter group is greater than that of the Democrat
with the CI for 95%
#range being positive. But the practical significance of this finding cannot co
nfirmed due to the small
#effect size.
```

Question 3: Do a majority of independent voters believe that the federal investigations of Russian election interference are baseless?

Introduce your topic briefly. (5 points)

Explain how your variables are operationalized. Comment on any gaps that you can identify between your operational definitions and the concepts you are trying to study.

The federal investigations of Russian election interference have been a polarizing issue. In general terms, Republicans believe the investigations are baseless, while Democrats support the investigation. However, the lines are blurred regarding independent voters. Since independent voters do not align to either party, their views do not follow party lines as closely as Republican or Democratic voters. Understanding independent voters' views on the federal investigations provides insights into their voting behaviors as well as revealing attitudes they have toward the investigations.

The relevant variables are*:

- pid7x
- russia16
- coord16
- muellerinv

There are a few key terms in the question to operationalize beginning with the definition of an independent voter. In keeping consistent with question 2's idea of defining voters based on their most likely voting behavior rather than self-identification, we would only use people who declared themselves as independents (*pid7x* == 4) and exclude those who may state they are independents, but lean republic or democrat. The next part of the question to define is what constitutes federal investigations of Russian election interference.

We must first define "baseless" as mentioned in the question. Baseless in this context can be interpreted to mean people believing there lacks evidence to warrant a federal investigation. All variables are relevant and it is worth noting that the Mueller Investigation was an official investigation of Russian interference in the 2016 US election and thus people who approve of the investigation must also believe that there is some base in the investigation. Consequently, we can split the sample into two groups: those who believe investigations are baseless and those who do not. Those who believe investigations are baseless are defined by:

1. *russia16* = 1
2. *coord16* = 1
3. *muellerinv* = 1, 2, or 3

If these three conditions are not met, then the the respondent believes the investigations are baseless.**

The last part to operationalize is how we can measure the result. By splitting the groups into those who believe the investigations are baseless or not, we have two distinct groups and we need a value to use as the measurement variable. The Mueller Report was a divisive report that created two groups: those who supported it or not. Although we use three variables to define the groups, we only use *muellerinv* and use the scores for this variable to measure if a majority of independent voters believe the investigation was baseless or not.

**see below for definitions*

***see note below*

In [ ]:
```
#Scale for:
#russia16 (Measures if people think the Russian government probably interfered
#     in the 2016 presidential election):
# 1. Russia probably interfered
# 2. This probably did not happen

#Scale for:
#coord16 (Measures if people think Donald Trump's 2016 campaign probably coordi
nated with the Russians)
# 1. Probably coordinated with the Russians
# 2. Probably did not

#Scale for:
# muellerinv (Measures if people approve of Robert Mueller's investigation)
# 1. Approve extremely strongly
# 2. Approve moderately strongly
# 3. Approve slightly
# 4. Neither approve nor disapprove
# 5. Disapprove slightly
# 6. Disapprove moderately strongly
# 7. Disapprove extremely strongly

#Scale for:
#pid7x (Measure voter self-identification toward party)
# -7 - no answer
# 1 - Strong dem
# 2 - Not very strong dem
# 3 - Ind, closer to dem
# 4 - Independent
# 5 - Ind, closer to rep
# 6 - Not very strong rep
# 7 - Strong rep

#Note:
# We go up to "approve slightly" as an answer for *muellerinv* because slight a
pproval
# is still an indication that the person believes there is some base warranting
the investigation.
# For *russia16* and *coord16*, we only use those who indicate probably interef
ered or coordinated
# and treat the reponse in a binary sense.
```

Perform an exploratory data analysis (EDA) of the relevant variables. (5 points)

This should include a treatment of non-response and other special codes, basic sanity checks, and a justification for any values that are removed. Use visual tools to assess the relationship among your variables and comment on any features you find.

In [30]:
```
#Create the two sets of independent voters
indep = A[A$pid7x == 4,]

#Those who believe the investigation is not baseless
base.df = indep[ (indep$russia16 == 1) & (indep$coord16 == 1) & ((indep$mueller
inv == 1)| (indep$muellerinv == 2)
                | (indep$muellerinv == 3)),]
#Those who believe the investigation is baseless
no.base.df = indep[!rownames(indep)%in% rownames(base.df),]

base = base.df$muellerinv
no.base = no.base.df$muellerinv
```

In [31]:
```
#Check for missing data
(which(is.na(base)))
(which(is.na(no.base)))
```

In [32]:
```
#There are no missing values.
```

In [33]:
```
paste("Number of independent voters: ", nrow(indep))
paste("Number of those who believe the investigation was not baseless: ", nrow
(base.df))
paste("Number of believe the investigation was baseless: ", nrow(no.base.df))
```

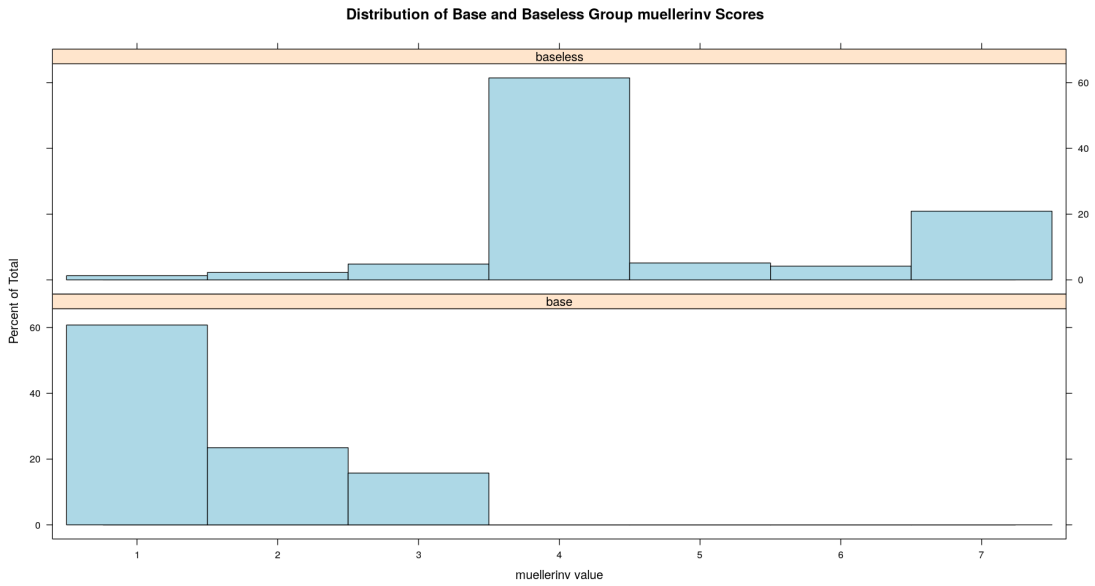
'Number of independent voters: 417'

'Number of those who believe the investigation was not baseless: 106'

'Number of believe the investigation was baseless: 311'

In [29]:
```
#Our sample size is 417 after applying the operationalized restrictions.
#Slightly over one quarter of the sample believes the investigation is warrante
d.
#We notice that the lengths of our two groups (base and no base) are different,
#but this is okay unless we conduct a dependent test.
```

In [34]:
```
options(repr.plot.width=15, repr.plot.height=8)

inv = suppressWarnings(do.call(rbind, Map(data.frame, A=base,
                                          B=no.base)))
colnames(inv) = c('base', 'baseless')
library(reshape)
inv = melt(inv, measure.vars = c('base', 'baseless'))
inv$Likert.f = factor(inv$value,ordered = TRUE)
library(lattice)
histogram(~Likert.f | variable,
          data = inv,
          layout = c(1,2),
          col = 'lightblue',
          xlab = 'muellerinv value',
         main = 'Distribution of Base and Baseless Group muellerinv Scores')
```



In [46]:
```
#The above plot shows the counts of the *muellerinv* variable for the two group
s.
#Remember that muellerinv had the following answer choices:

#1. Approve extremely strongly
#2. Approve moderately strongly
#3. Approve slightly
#4. Neither approve nor disapprove
#5. Disapprove slightly
#6. Disapprove moderately strongly
#7. Disapprove extremely strongly

#Based on the scale, we see that people in the base group
#(independent voters who believe that the investigation has a base) only voted
1, 2, or 3,
#which all fall into the approval range. On the other hand, the baseless is mor
e spread with
#responses in all 7 categories. Most responses are 4 indicating a neutral feeli
ng toward the
#investigation, but we see several responses in greater than 4 indicating disap
proval of the investigation.
#There are a small amount of responses in this group that do approve the invest
igation.
```

Based on your EDA, select an appropriate hypothesis test. (5 points)

Explain why your test is the most appropriate choice. List and evaluate all assumptions for your test.

We use the Wilcoxon rank-sum test using the null hypothesis of comparisons.*

Let X be a random variable representing the base group and let Y be the random variable representing the baseless group. The hypotheses can be set up as such:

$H_0 : P(X < Y) = P(X > Y)$

$H_A : P(X < Y) > P(X > Y)$

The null hypothesis can be interpreted in the following way: the probability that X is less than Y equals the probability that X is greater than Y. The alternative hypothesis can be interpreted as whether the probability of obtaining a voter who believes in the investigation is baseless is higher than the probability of obtaining a voter who believes the investigation is not baseless. We use a one-tailed test with a 5% significance level since we are interested in investigating whether a majority of independent voters believe the investigations are baseless. We know that using a one-sided test would increase our rejection region and chance of rejecting the null hypothesis, and we want to lessen the probability we falsely reject the null hypothesis in favor of the alternative because we care if the majority really does believe baseless investigation. Consequently, we want to lessen our type 1 error rate so we use a 1% significance level instead. While making this change increases our type 2 error rate (probably of failing to reject the null hypothesis when the null hypothesis is false), we care more about falsely rejecting the null hypothesis to address the question. We are only looking for one side of the test (majority) and we do not care about the other, only if that majority idea is false.

The two assumptions for this test are that the variable we are measuring is an on ordinal scale and that each pair (X_i, Y_i) is drawn independently of other pairs from the same distribution. The first assumption is met since *muellerinv* is an ordinal variable.

The second assumption is also met since each sample in the data is independent and unrelated to the others. We are assuming that independent voters are not related and do not influence each other's opinions. This claim is supported by the survey design, which states that respondents are drawn from the same distribution and unrelated to each other. With these assumptions met, we continue with the test. The sampling procedure from the survey design states that respondents are independent of one another so this assumption is met.

**see below for justification*

In [ ]:
```
#Test justification
# We first note that the variable we are measuring is an ordinal variable and
# the only valid operation on levels is comparing. We do not use a parametric t
est
# because we have an ordinal variable. We use an independent test, since we are
dealing
# with two distinct groups. Although the sample is based on independent voters,
we are dividing
# independent voters into a base and baseless group and we are interested in co
mparing their
# approval of the investigation. The groups are not dependent since the same in
dependent voter
# cannot both approve or disapprove of the investigation at the same time.
```

Conduct your test. (5 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result. Make sure you relate your findings to the original research question.

In [35]:
```
#One-tailed test
one.test = wilcox.test(base, no.base, alternative = 'less', conf.level = 0.99)
one.test
```

```
        Wilcoxon rank sum test with continuity correction

data:  base and no.base
W = 630, p-value < 2.2e-16
alternative hypothesis: true location shift is less than 0
```

In [36]:
```
library(rcompanion)
wilcoxonR(no.base, base)
#this library uses the equation for correlation: (r = Z/(√Nobs))
# https://rcompanion.org/handbook/F_04.html
```

r: -0.116

In [70]:
```
#We see the same result with the one-tailed test in that we have an extremely s
mall p-value
#and we can reject the null hypothesis with a highly statistically significant
result.
#This helps give direction to our interpretation and we can say with this test
that the
#probability of drawing a voter who believes the investigation is baseless is h
igher than
#the probability of drawing a voter who believes the investigation has a base.

#We calculate the correlation to measure the effect size, which is almost 0 at
a value
#of -0.116 meaning an extremely small effect size.
#This suggests that although we obtained a highly statistically significant res
ult,
#the result is not practically significant and we cannot definitively say that
a majority
#of independent voters believe the investigations were baseless.
```

Question 4: Was anger or fear more effective at driving increases in voter turnout from 2016 to 2018?

Introduce your topic briefly. (5 points)

Explain how your variables are operationalized. Comment on any gaps that you can identify between your operational definitions and the concepts you are trying to study.

After the 2016 presidential elections, results were polarizing causing emotions to run high. Of these emotions, fear and anger were some of the most common and motivating factors influencing politics. Examining which of these emotions can cause higher voter turnout would be key information for policy makers to shape how to run elections how to appeal to voters.

To answer the question, we need to find the relavant variables from the data set related to voter emotions and voter turnouts.

The relevant variables are listed as below:

Voter emotions(related to anger and fear)*:

- geangry
- geafraid
- dtangry
- dtafraid
- imangry
- imafraid

Voter turnouts*:

- turnout18
- turnout16

We will only use definite answers *turnout18* and *turnout16* to measure voter turnout since if people are unsure if they voted, this could provide inaccuracies in the results.

To answer the question, We will use *geangry* and *geafraid* as variables to measure anger and fear, respectively, and we will exclude *dtangry*; *dtafraid*; *imangry*; and *imafraid* since these 4 variables relate to specific issues and people.

New variable diff_turnout: Based on the above variables, we decided the best way to analyze the issue would be to use a interval scale variable *diff_turnout* that measures the voter's voting pattern (increase,neutral,decrease) between 2016 to 2018. We gave a score of +1 for increase, 0 for neutral and -1 for decrease in the voting pattern. So, there will be a same numerical distance between each value with an arbritary zero value in it, with possible values of -1,0,1

Below is the rule chart for assigning the score for the voter based on his response for 2016 and 2018 election for variables Turnout18 and Turnout16. For example, if a voter chooses either 1,2 or 3(definitely voted), meaning for Turnout18 and 2(definitely did not vote) for Turnout 16, we consider it as 1. Similarly, if he did not vote in 2018, but voted in 2016, it is -1. If he did not voted in 2018 and did not vote in 2016 then it is 0, no change. The entire possible rules are listed below:

Turnout18	Turnout16	Score
1,2,3	2	1
1,2,3	1	0
1,2,3	3	1
4	1	-1
4	2	0
4	3	-1
5	1	0
5	2	1
5	3	0

A positive score would indicate an increase in voter turnout, a negative score would indicate a decrease, and a score of 0

means there was no change.

Now we will create two samples for the variable *diff_turnout* as follows:

- Sample_Angry_only: Sample of angry people who are not fearful
- Sample_Fear_only : Sample of fearful people who are not angry These two datasets has the variable *diff_turnout*. We will now compare this variable between two samples(Sample_Angry_only and Sample_Fear_only).

*see scale below

In [ ]:
```
#Scale for:
#geafraid (How angry do you feel about the way things are going in the countr
y?)
#geangry (How afraid do you  feel about the way things are going in the countr
y?)
# 1 Not at all
# 2 A little
# 3 Somewhat
# 4 Very
# 5 Extremely
# -7 no answer

#Scale for:
#dtangry (Measures anger caused by Donald Trump)
#dtafraid (Measures fear caused by Donald Trump)
#imangry (Measures anger caused by immigration)
#imafraid(Measures fear caused by immigration)
# 1 Not at all
# 2 A little
# 3 Somewhat
# 4 Very
# 5 Extremely
# -1 no answer

#Scale for:
#turnout18 (Tracks how and if people voted in the November 6, 2018 midterm elec
tions)
# 1 Definitely voted in person on Nov 6
# 2 Definitely voted in person, before Nov 6
# 3 Definitely voted by mail
# 4 Definitely did not vote
# 5 Not completely sure

#Scale for:
#turnout16 (Tracks if people voted in the November 6, 2016 presidential electio
n):
# 1 Definitely voted
# 2 Definitely did not vote
# 3 Not completely sure

#turnout18ns (For those that were unsure if they voted in the November 6, 2018
election,
#    measures how unsure people were about voting)
#turnout16b (For those that were unsure if they voted in the November 6, 2016 e
lection,
#    measures how unsure people were about voting)
```

Perform an exploratory data analysis (EDA) of the relevant variables. (5 points)

This should include a treatment of non-response and other special codes, basic sanity checks, and a justification for any values that are removed. Use visual tools to assess the relationship among your variables and comment on any features you find.

In [37]:
```
paste("Summary turnout16 responses : ")
summary(A$turnout16)

paste("Distribution of turnout16 responses : ")
table(A$turnout16)

paste("Total # turnout16 responses : ")
length(A$turnout16)

paste("Datatype for turnout16 responses is categorical : ")
str(A$turnout16)
```

'Summary turnout16 responses : '

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   1.000   1.000   1.306   2.000   3.000 
```

'Distribution of turnout16 responses : '

```
   1    2    3 
1841  552  107 
```

'Total # turnout16 responses : '

2500

'Datatype for turnout16 responses is categorical : '

```
 int [1:2500] 1 2 2 3 1 1 1 1 2 1 ...
```

In [38]:
```
#checking for missing values
paste("Checking for missing values for turnout16 : ")
(which(is.na(A$turnout16)))
```

'Checking for missing values for turnout16 : '

There are no missing values for turnout16

In [39]:
```
paste("Summary turnout18 responses : ")
summary(A$turnout18)

paste("Distribution of turnout18 responses:")
table(A$turnout18)

paste("Total # turnout18 responses : ")
length(A$turnout18)

paste("Datatype for turnout18 responses is categorical: ")
str(A$turnout18)
```

'Summary turnout18 responses : '

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   1.000   2.000   2.392   4.000   5.000 
```

'Distribution of turnout18 responses:'

```
  1   2   3   4   5 
968 357 517 544 114 
```

'Total # turnout18 responses : '

2500

'Datatype for turnout18 responses is categorical: '

```
 int [1:2500] 1 4 4 5 1 1 3 3 4 3 ...
```

In [40]:
```
#checking for missing values
paste("Checking for missing values for turnout18 : ")
(which(is.na(A$turnout18)))
```

'Checking for missing values for turnout18 : '

There are no missing values for turnout18

In [41]:
```
# code to derive the variable diff_turnout based on the rule chart explained in
question (a). It uses
#an if/else method of R to assign a value based on the options choosen for turn
out16 and turnout18 variables
#diff_turnout represents the increase, decrease, or no change in voter turnout
from 2016 to 2018
A$diff_turnout <- ifelse(A$turnout18 %in% c(4) & A$turnout16 %in% c(1,3) ,-1,
            ifelse(A$turnout18 %in% c(4) & A$turnout16 %in% c(2),0,
            ifelse(A$turnout18==5 & A$turnout16 %in% c(2),1,
            ifelse(A$turnout18==5 & A$turnout16 %in% c(1,3),0,
            ifelse(A$turnout18 %in% c(1,2,3) & A$turnout16 %in% c(2,3),1,0)))))

paste("Distribution of  diff_turnout variable showing change in voter behaviour
from 2016 to 2018 : ")
table(A$diff_turnout)

paste("summary of diff_turnout variable showing change in voter behaviour from
2016 to 2018 : ")
summary(A$diff_turnout)

paste("Datatype for diff_turnout responses : ")
str(A$diff_turnout)
```

'Distribution of diff_turnout variable showing change in voter behaviour from 2016 to 2018 : '

```
  -1    0    1
 145 2185  170
```

'summary of diff_turnout variable showing change in voter behaviour from 2016 to 2018 : '

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -1.00    0.00    0.00    0.01    0.00    1.00
```

'Datatype for diff_turnout responses : '

```
 num [1:2500] 0 0 0 0 0 0 0 0 0 0 ...
```

In [42]:
```
#checking for missing values
paste("Checking for missing values for diff_turnout : ")
(which(is.na(A$diff_turnout)))
```

'Checking for missing values for diff_turnout : '

There are no missing values for diff_turnout variable

In [43]:
```
# Create one sample of angry people who are not fearful(Sample_Angry_only)
angry_only_voter = A[A$geangry %in% c(2,3,4,5) & A$geafraid == 1,]

paste("Summary of angry_only_voter :")
paste(nrow(angry_only_voter))
```

'Summary of angry_only_voter :'

'269'

In [44]:
```
# Create another sample of fearful people who are not angry(Sample_Fear_only)
afraid_Only_voter = A[A$geafraid %in% c(2,3,4,5) & A$geangry == 1,]

paste("Summary of afraid_Only_voter :")
paste(nrow(afraid_Only_voter))
```

'Summary of afraid_Only_voter :'

'189'

In [45]:
```
#Note : For our analysis, we decided to make sure we match the sample size for
both the datasets.
#Hence we created a new sample(angry_only_voter_sample) mentioned below for the
angry_only_voter with a
#length equal to that of afraid_Only_voter so that both samples have the same l
ength for analysis.
```

In [46]:
```
set.seed(10)
angry_only_voter_sample = angry_only_voter[sample(nrow(angry_only_voter), nrow
(afraid_Only_voter)), ]
paste("Length of angry_only_voter_sample :")
paste(nrow(angry_only_voter_sample))
paste("Length of afraid_Only_voter :")
paste(nrow(afraid_Only_voter))
```

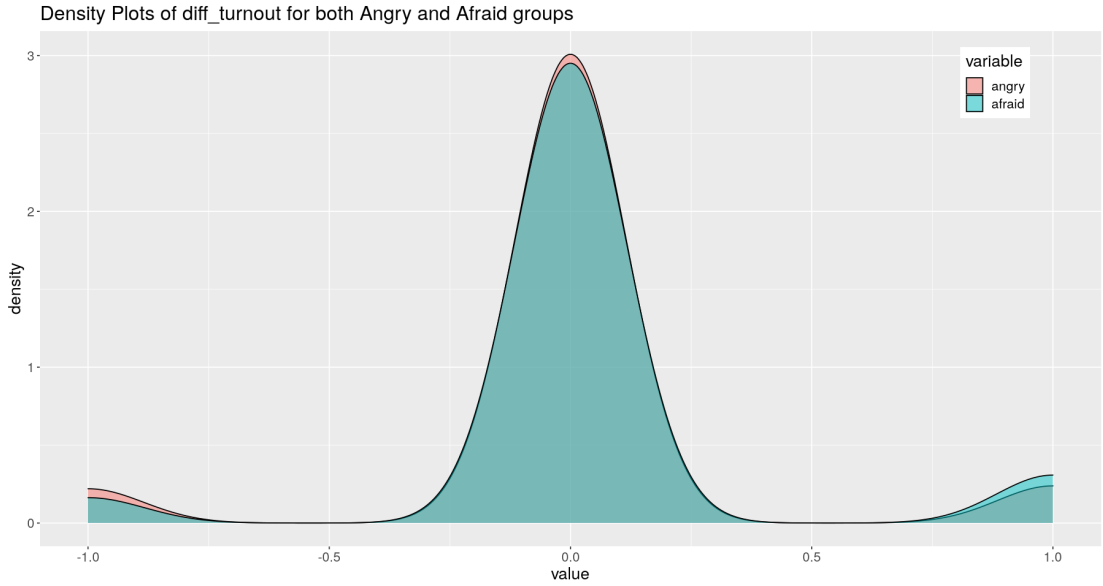
'Length of angry_only_voter_sample :'

'189'

'Length of afraid_Only_voter :'

'189'

In [47]:
```
library(ggplot2)
options(repr.plot.width=15, repr.plot.height=8)

emot = suppressWarnings(do.call(rbind, Map(data.frame, A=angry_only_voter_sampl
e$diff_turnout,
                                           B=afraid_Only_voter$diff_turnout)))
colnames(emot) = c('angry', 'afraid')
library(reshape)
emot = melt(emot, measure.vars = c('angry', 'afraid'))
options(repr.plot.width=15, repr.plot.height=8)
# emot$Likert.f = factor(emot$value,ordered = TRUE)

ggplot(emot, aes(x = value, fill = variable)) +
    geom_density(alpha = 0.5) +
    ggtitle('Density Plots of diff_turnout for both Angry and Afraid groups') +
    theme(text = element_text(size=16), legend.position = c(0.9, 0.9))
```



In [6]:
```
#The above density plot shows the distribution of values for both groups' *diff
_turnout* score.
#The distributions are nearly identical and the large majority of values are 0
in both groups with
#slightly more 1 values in the afraid group indicating a slightly larger increa
se in
#voter turnout in that group.
```

Based on your EDA, select an appropriate hypothesis test. (5 points)

Explain why your test is the most appropriate choice. List and evaluate all assumptions for your test.

We will be selecting a Wilcox rank sum test for answering the question. The dataset samples that measure the variable diff_turnout between 2016 and 2018 measure for two distinct groups(angry and fear).*

Here are our assumptions:

- Metric scale : The diff_turnout variable measures the difference in turnouts in a metric scale, because the signs of the values matter to indicate increase, decrease, or no change in voter turnout.
- Each (X_i, Y_i) pair is drawn independently where X represents the diff_turnout value of the angry group and Y represents the diff_turnout value of the fearful group, and we know this since the data is identically distributed based on the survey design and obtaining all respondents in the same way from the same YouGov panel. Independent because survey design mentions independence.
- X has the same distribution as $Y - \Delta$ where Δ is the constant defining the differnce in distribution and we know this by looking at the density plots and how similar they are and in this case it looks like Δ may even be 0.

Hence we can conclude the below: Null hypothesis is the difference in distributions is equal to 0 $H_0 : \Delta = 0$

Alternative hypothesis is difference in distributions is not equal zero (two-tailed) $H_A : \Delta \neq 0$

**see below for justification*

In [ ]:
```
#Test justifications:
# We defined the two groups as being mutually exclusive in the sense that
# a person cannot be both angry and fearful, only one. In this sense, the group
s must
# be treated independently. We would use a nonparametric test, since even thoug
h anger
# and fear variables are ordinal in nature, the diff_turnout variable is used t
o measure the
# difference in emotions on an interval scale. We will use the hypothesis of me
ans in this case
# as well since we are interested in the value of the diff_turnout variable.
```

Conduct your test. (5 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result. Make sure you relate your findings to the original research question.

In [48]:
```
# computing the wilcox.test function in R for Rank Sum test
paste("Comuputing the wilcox.test result :")
angry_voter <- angry_only_voter_sample$diff_turnout
afraid_voter <- afraid_Only_voter$diff_turnout
w.test = suppressWarnings(wilcox.test(angry_voter, afraid_voter, paired = T))
w.test
```

'Comuputing the wilcox.test result :'

```
        Wilcoxon signed rank test with continuity correction

data:  angry_voter and afraid_voter
V = 440.5, p-value = 0.331
alternative hypothesis: true location shift is not equal to 0
```

In [53]:
```
# effect size R correlation
library(rstatix)
library(reshape)

paste("Effect size correlation is : ")
pols = do.call(rbind, Map(data.frame, A=angry_voter, B=afraid_voter))
colnames(pols) = c('angry_voter', 'afraid_voter')
gek = melt(pols, measure.vars = c('angry_voter', 'afraid_voter'))
round(wilcox_effsize(gek, value~variable, paired=T)$effsize,3)
```

'Effect size correlation is : '

Effect size (r): 0.075

In [17]:
```
#Considering the high p-value(0.331), we fail to reject the null Hypothesis.
#Thus we conclude that there is not a significant difference in fear or anger i
n increasing
#the voter turnout between 2016 and 2018, since, fear and anger have identical
populations.

#The correlation coefficient r test result(Effect size (r): 0.075) backs the st
atistical test(p-value)
#finding. As effect size is also significantly low and close to 0, showing that
that there is a lot
#of overlap in the samples of anger and fear, which also indicate that the dist
ributions are very
#similar and that the difference (delta) between the two distributions is very
trivial or small.

#Based on these findings,to answer the question, we can conclude that neither a
nger or
#fear were more effective in increasing the voter turnout between 2016 to 2018
and they
#use aplayed equal role in increasing the voter turnouts between 2016 and 2018.
```

Question 5: Select a fifth question that you believe is important for understanding the behavior of voters

Clearly argue for the relevance of this question. (10 points)

In words, clearly state your research question and argue why it is important for understanding the recent voting behavior. Explain it as if you were presenting to an audience that includes technical and non technical members.

Explain how your variables are operationalized. Comment on any gaps that you can identify between your operational definitions and the concepts you are trying to study.

Do Republican voters have more respect for Barack Obama or Hillary Clinton?

The two party system in the United States has influenced elections from local to national levels and it is common for people to vote more along party lines rather than treating issues independently. This leads to many people who lean Republican to vote for Republicans and many people who lean Democrat to vote for Democrats and rarely for the opposite party. President Barack Obama and 2008 and 2016 Democratic frontrunner Hillary Clinton have seen low support from Republicans, but we are interested in which of these two Democrats Republicans view more favorably. Investigating this question would lead to interesting insights into Republic voters' views. There are confounding factors like race and gender that we know affect people's views on candidates, but we will capture this information with the overall rating people gave for the two politicians.

Why else is this important?

The relevant variables in the data are:

- ftobama (Measures voters' rating of Obama)
- fthrc (Measures voters' rating of Clinton)
- pid7x (Measures voters' self-identified party affiliation)

The *ftobama* and *fthrc* variables use the same measuring system as discussed in question 1 about the thermometer and a rating scale from 0 to 100, 0 being an unfavorable view and 100 being a favorable view. We will use the same numeric interpretation as we did in question 1 with the scale and assume that differences in rating correlate to the same differences in views. For example, a voter who rated Obama a 50 would view Obama more favorably than a voter who rated Obama a 48. Additionally, we will translate favor over to respect similarly as we did in question 1.

We will define Republican voters as we have done in question 2 with the *pid7x* values of 5, 6, 7.*

We define Republicans by their voter leanings so we include values 5, 6, 7.

**refer to scale below*

In [22]:
```
#Scale for pid7x:
#-7 - no answer
#1 - Strong dem
#2 - Not very strong dem
#3 - Ind, closer to dem
#4 - Independent
#5 - Ind, closer to rep
#6 - Not very strong rep
#7 - Strong rep
```

Perform EDA and select your hypothesis test (5 points)

Perform an exploratory data analysis (EDA) of the relevant variables.

This should include a treatment of non-response and other special codes, basic sanity checks, and a justification for any values that are removed. Use visual tools to assess the relationship among your variables and comment on any features you find.

Based on your EDA, select an appropriate hypothesis test. Explain why your test is the most appropriate choice. List and evaluate all assumptions for your test.

In [54]:
```
#Define Republicans
rep = A[(A$pid7x == 5)|(A$pid7x == 6)|(A$pid7x == 7),]

paste('# Republicans:', nrow(rep))
```

'# Republicans: 849'

There are 849 Republicans

In [55]:
```
paste("Summary of ftpolice showing respect for Obama:")
summary(rep$ftobama)
paste("Summary of ftjournal showing respect for Clinton:")
summary(rep$fthrc)
```

'Summary of ftpolice showing respect for Obama:'

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    1.00    6.00   21.25   35.00  100.00 
```

'Summary of ftjournal showing respect for Clinton:'

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  -7.00    1.00    3.00   11.25   12.00  100.00 
```

In [7]:
```
# Both variables are numerical and real-valued ranging from 0 to 100 (integer
s). -7 appears as the minimum value
# in Clinton's rating indicating that some voters did not provide a rating for
Clinton.
# Every descriptive statistic except the max is higher for Obama's ratings than
for Clinton's rating indicating
# that Republicans may tend to view Obama more positively than Clinton.
#However, all values except the max are less
# than 50 indicating that Republicans still may view both politicians negativel
y.
```

In [56]:
```
#Removing negative values
neg_obama = rep$ftobama[rep$fobama <0]
neg_clinton = rep$fthrc[rep$fthrc <0]

paste("# negative ratings for Obama: ", length(neg_obama))
paste("# negative ratings for Clinton: ", length(neg_clinton))

pos_obama = rep$ftobama[rep$ftobama >=0]
pos_clinton = rep$fthrc[rep$fthrc >=0]

paste("# total ratings for Obama: ", length(pos_obama))
paste("# total ratings for Clinton: ", length(pos_clinton))
```

'# negative ratings for Obama: 0'

'# negative ratings for Clinton: 1'

'# total ratings for Obama: 849'

'# total ratings for Clinton: 848'

In [8]:
```
# We have one negative value for Clinton's rating that we must remove.
# We cannot impute these negative values since that would be adding false data.
# Replacing a negative value with any value in the range from 0 to 100 would be
saying that Republican
# had that specific attitude of respect toward Clinton and this would be a fals
e representation.
# Adding another value outside this range would not make sense since it would b
e unintellible according
# to the rating scale. Consequently, removing them is the most appropriate cas
e. Because we are dealing
# with one sample and two different opinions from the same people, we cannot re
move only the values from
# journalists as we must compare that single person's views of Obama and Clinto
n. Our sample size is now 848.
```

In [57]:
```
#Removing values
rep[rep$fthrc < 0 , c('ftobama', 'fthrc')]

#Removing the rows
obama = rep[rep$fthrc >= 0,]$ftobama
clinton = rep[rep$fthrc >= 0,]$fthrc
```

A data.frame: 1 × 2

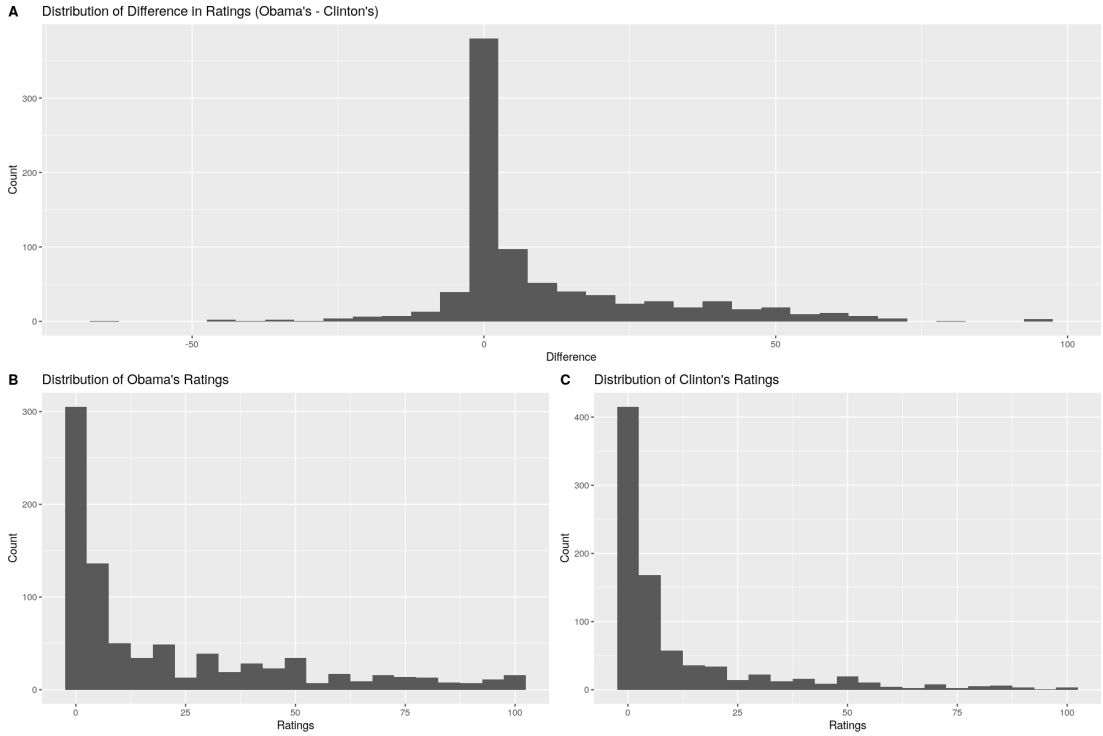
	ftobama	fthrc
	<int>	<int>
1397	5	-7

In [9]:
```
# The one displayed row is the row in the data which contains the missing value
# for Clinton's rating. We see the corresponding rating for this voter is 5 ind
icating a very negative of Obama.
# From the descriptive statistics, we know that the majority of values seem to
be negative for Obama, so by
# removing this row, we are not removing an outlier.
```

In [58]:
```
#Checking for missing values
(which(is.na(obama)))
(which(is.na(clinton)))
```

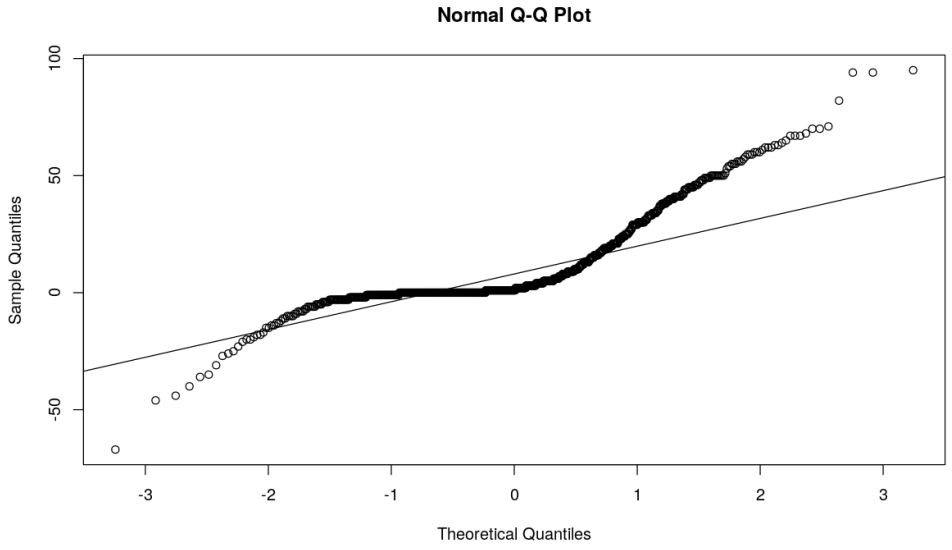
There are no missing values in the data.

In [59]:
```
library(ggpubr)
p1 = qplot(obama, geom="histogram", binwidth = 5,xlab = 'Ratings',
           ylab = 'Count', main = 'Distribution of Obama\'s Ratings')
p2 = qplot(clinton, geom="histogram", binwidth = 5,xlab = 'Ratings',
           ylab = 'Count', main = 'Distribution of Clinton\'s Ratings')
p3 = qplot(obama-clinton, geom="histogram", binwidth = 5,xlab = 'Difference',
           ylab = 'Count', main = 'Distribution of Difference in Ratings (Obam
a\'s - Clinton\'s)')

options(repr.plot.width=15, repr.plot.height=10)
ggarrange(p3,                                                  # First row with
scatter plot
          ggarrange(p1, p2, ncol = 2, labels = c("B", "C")), # Second row with
box and dot plots
          nrow = 2,
          labels = "A"                                        # Labels of the s
catter plot
          )
```



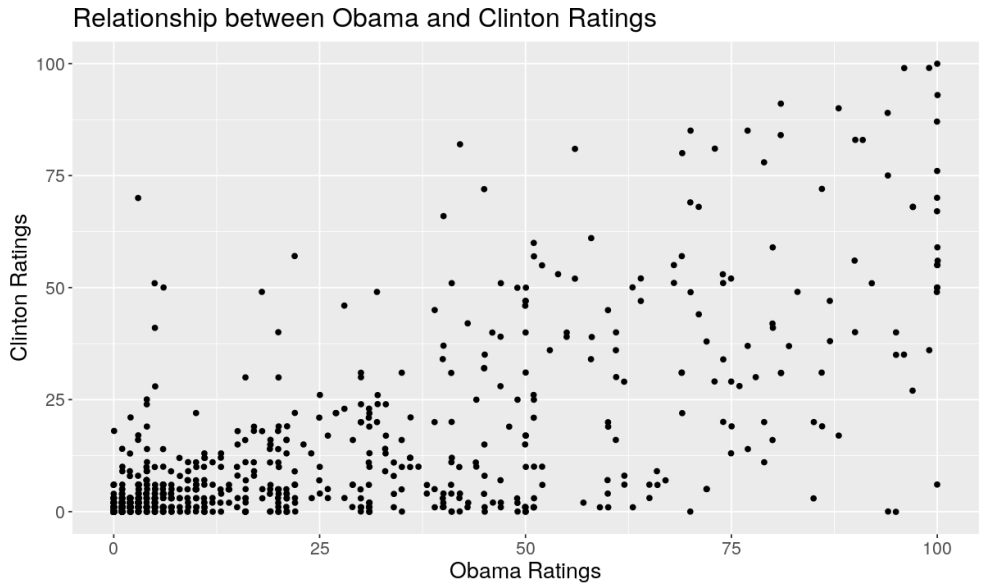
In [10]:
```
# Beginning with plots B and C in the bottom of the above plots, we see that ne
ither variable for Obama
# or Clinton's ratings follows a normal distribution. Plot B showing Obama's ra
tings is right skewed,with a
# majority of values being less than 25 indicating many Republicans felt unfavo
rable toward Obama.
# The one removed values would have fallen where this peak is so we can assume
that the loss of data does not
# have a significant impact. In plot C, the distribution of Clinton's rating sh
ows similar results, but the plot
# is even more right skewed. Obama's ratings had small peaks across the higher
ratings and even several high
# ratings, but Clinton's ratings are consistently low with less higher ratings
than Obama's.

# Plot A showing the differences in the ratings more closely follows a seemingl
y normal distribution.
# Since we are subtracting Clinton's ratings from Obama's ratings, a positive v
alue in this plot indicates
# that the person gave a higher rating to Obama than Clinton and a negative val
ue indicates the person gave a
# higher rating to Clinton than to Obama. A difference of 0 indicates the perso
n gave the same rating to both
# Obama and Clinton. The distribution is not perfectly normal and we can see mo
re positive values indicating
# that more people rated Obama better than Clinton, but there is a tall central
peak at 0 indicating that many
# people viewed both politicians equally.
```

In [60]:
```
#normality of differences
diff = obama - clinton
options(repr.plot.width=10, repr.plot.height=6)

qqnorm(diff)
qqline(diff)
```



In [12]:
```
# We can see from this Q-Q plot that there are deviations from the line at the
lower tail, but especially
# at the higher tail where points deviate frequently from the line. There is al
so a dip in the middle of the plot.
# We can interpret these results to mean that the distribution is heavy-tailed
based on the flatness in
#the middle part of the plot and the upward trend of the points toward the righ
t.
#Looking at the prior histograms, we know that several of Obama's ratings were
in the very
# positive range (>80) compared to the number of same ratings for Clinton.
# The two tails of the Q-Q plot both deviate from the line and we see that the
right tail points
#jump toward the endindicating a slight right skew, which we can see from the h
istogram of the differences.

#https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot
```

In [61]:
```
#Scatterplot
rating = do.call(rbind, Map(data.frame, A=obama, B=clinton))
colnames(rating) = c('obama', 'clinton')

ggplot(rating, aes(x=obama, y=clinton)) +
    geom_point(position = position_jitter(w = 0.05, h = 0.05))+ #Adding jitter
    labs(x='Obama Ratings', y='Clinton Ratings')+
    theme(text = element_text(size=16))+
    ggtitle('Relationship between Obama and Clinton Ratings')
```



In [13]:
```
# Looking at the scatterplot reveals that there may be a weak relationship betw
een Obama's and Clinton's ratings.
# The graph has a slight upward trend indicating that those Republicans who vie
wed Obama more positively and
# respectfully felt similarly toward Clinton. We also see that the majority of
ratings for both candidates are
# clustered below 25.
```

Conduct your test. (2 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result.

We opt to use the Wilcoxon Signed-Rank test.*

The important assumption of the dependent t-test is that the difference between ratings should be normal. The Q-Q plot and histogram of the differences between ratings iss non-normal, so it would be a smarter choice to use a non-parametric test that does not rely on the normality assumption.

The next assumption for the signed-rank test is that data are paired and drawn i.i.d. We know are data is paired due to the fact we are taking two different measurements from the same people and we can assume that based on the survey design that used sample matching that the respondents are independent from one another. It is also stated in the survey design that respondents were selected from a panel called the YouGov panel that consists of a large and diverse set of over a million respondents so we can assume the distribution is identical for each respondent from this organization they were drawn. Additionally, the independence assumption is met through the survey design and how each respondent is independent of each other.

The last assumption of this test is that the difference between pairs follows a symmetric distribution around the mean of the distribution. We can assume this assumption is met by looking at the histogram of the differences. The mean visually seems to be at 0 with a somewhat symmetric distribution of values above and below the mean. While not perfect, we know that the data is just a sample that will have intrinsic randomness to it.

We setup our test as such where μ is the mean of the difference between Obama's and Clinton's ratings:

$H_0 : \mu = 0$

$H_A : \mu \neq 0$

We opt for a two-sided test at a 5% significance level over a one-sided test since we are more interested in if Republicans view Obama and Clinton differently. Moreover, we don't have any strong reason to believe they would prefer one over the other and a one-tailed test increases the rejection region as well.

**see below for justification*

In [ ]:
```
#Test justification
# We first note that we are dealing with a dependent test,
# because we are comparing how the same Republicans view two different people.
# There is a link between the two data sets, so we cannot use an independent te
st.
# Secondly, we are dealing with a numeric variable.
```

In [62]:
```
#We use 5% significance
test = wilcox.test(obama, clinton, paired = T, conf.int = T)
test
```

```
        Wilcoxon signed rank test with continuity correction

data:  obama and clinton
V = 179522, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
  8.500011 12.000040
sample estimates:
(pseudo)median
      10.49998
```

In [14]:
```
# We see that our p-value is extremely small and we can reject the null hypothe
sis at a 5%
#significance level. This test suggests that the mean of the differences is not
0 and that
#Republicans view Obama and Clinton differently.
```

In [63]:
```
pols = do.call(rbind, Map(data.frame, A=obama, B=clinton))
colnames(pols) = c('obama', 'clinton')
library(reshape)
pols = melt(pols, measure.vars = c('obama', 'clinton'))

library(rstatix)
round(wilcox_effsize(pols, value~variable, paired=T)$effsize,3)
#Uses the formula r = Z/sqrt(N) where Z is the Z-statistic divided by the squar
e root of the sample size (N)
```

Effect size (r): 0.527

In [15]:
```
# We calculated a correlation with a large effect size 0.527 suggesting that we
can interpret the difference in
# ratings to be meaningful and practically significant.
```

Conclusion (3 points)

Clearly state the conclusion of your hypothesis test and how it relates to your research question.

Finally, briefly present your conclusion in words as if you were presenting to an audience that includes technical and non technical members.

In [18]:
```
#We reject the null hypothesis in favor of the alternative hypothesis that the
mean difference between
#Obama's and Clinton's ratings is not 0. We have a large effect size indicating
that
#this result can be practically significant. We can look at the confidence inte
rval to
#help answer the question as well. The confidence interval is (8.500, 12.000) w
ith both end points
#being positive suggesting that there is a positive difference in values and th
e mean ratings for
#Obama are higher. As such, Republicans, having little respect for both candida
tes, have more
#respect for Obama than for Clinton.
```