Lab 2: Comparing Means

w203 Statistics for Data Science

Eddie / Girija / Naveen / Sonal

We will read the csv file into variable 'A' and then filter out the nonserious and dishonest records (only the absolute ones, that is who answer as 'Never' honest and 'Never' serious in answering the questions on the survey) out before we start processing the data:

[RESPONSE QUALITY]

Question 1: [nonserious] We sometimes find people don't always take surveys seriously, instead providing funny or insincere answers. How often would you say that you were not serious in answering questions on this survey?
Question 2: [honest] How often would you say you answered the questions honestly on this survey?

Possible Answers:
Never [1]
Some of the time [2]
About half the time [3]
Most of the time [4]
Always [5]

In [1]:
```
A = read.csv("anes_pilot_2018.csv")
anes_pilot_2018_df <- A[A$nonserious != 5 & A$honest != 1, ]
```

Below we install and import two libraries that we use to test practical significance in our study

install.packages("effsize")
install.packages("lsr")

In [2]:
```
library(effsize)
library(lsr)
```

Research Questions

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

Introduce your topic briefly. (5 points)

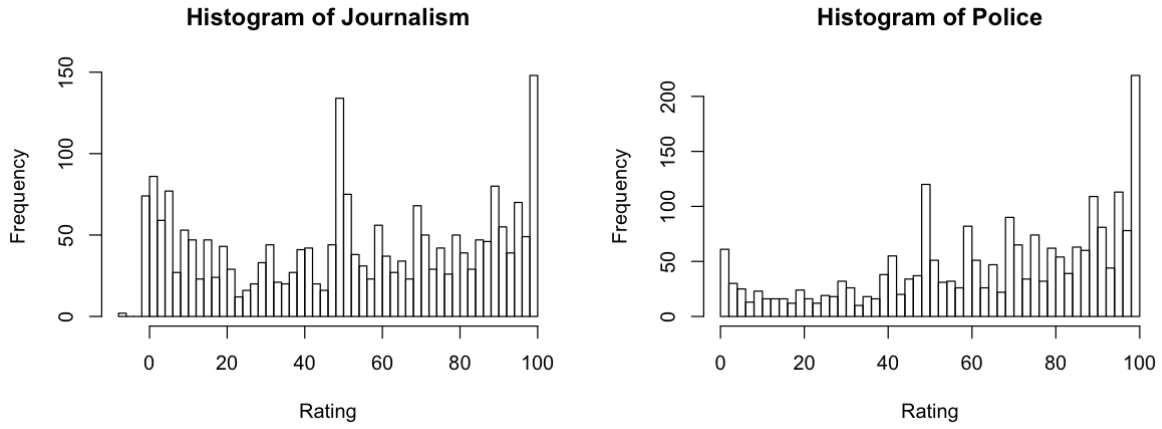
Explain how your variables are operationalized. Comment on any gaps that you can identify between your operational definitions and the concepts you are trying to study.

> There are two independent variables in this question (respect for police and respect for journalists.) The two fields in the data are ftjournal and ftpolice which are the ratings respondents gave to journalism and police respectively. We are trying to conduct a two independent samples t test which compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different.The potential gap is that these two variables might not be truely independent which we assumed for our test.
>
> Another assumption here is that the sample size is adequate and the samples were randomly selected and there is no clustering in the sample, which means the sample is i.i.d.

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 [3]:
```
x <-(anes_pilot_2018_df$ftjournal)
y <-(anes_pilot_2018_df$ftpolice)
options(repr.plot.width=10, repr.plot.height=4)
par(mfrow=c(1,2))
hist(x,breaks=50,main='Histogram of Journalism',xlab='Rating')
hist(y,breaks=50,main='Histogram of Police',xlab='Rating')
```



In [4]:
```
summary(x)
summary(y)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  -7.00   21.00   52.00   52.36   81.25  100.00 
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00   48.00   70.00   64.92   90.00  100.00 
```

In [5]:
```
#checking the two records that has -7 in
x <- A[A$ftjournal<0, ]
x
```

	version	caseid	weight	weight_spss	form	follow	addtime	reg	whenreg	howreg	··
51	ANES 2018 Pilot Study main version 20190129	55	0.3204701	0.1860009	2	1	3	1	4	-1	··
597	ANES 2018 Pilot Study main version 20190129	636	0.4936004	0.2864856	1	1	3	1	4	-1	··

-7 means no answer. That means those records should be excluded from the mean comparision.

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

Explain why your test is the most appropriate choice, focusing on its statistical assumptions.

> Since these are cardinal variables, we will conduct a independent sample t-test with hypothesis that the means of "respect police" and "respect jouralism" are the same. If we can reject the NULL hypothesis i.e. the means are not the same, we will be able to compare the means to detrmine which one is more repected by the voters.

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 [6]:
```
x <-anes_pilot_2018_df$ftjournal[anes_pilot_2018_df$ftjournal>=0]
y <-anes_pilot_2018_df$ftpolice[anes_pilot_2018_df$ftpolice>=0]

t.test(x,y)
```

```
        Welch Two Sample t-test

data:  x and y
t = -13.879, df = 4499.5, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.27919 -10.74453
sample estimates:
mean of x mean of y
 52.40917  64.92103
```

In this test, the p-value is smaller than the significant level therefore, we can reject the hyphothesis. We can conclude that the means of "respect for police" vs "respect for Journalism" are diffrent. Comparing the means and we can conclude that the respect for police is higher amoung voters.

In [7]:
```
cohen.d(y,x)
```

```
Cohen's d

d estimate: 0.4101066 (small)
95 percent confidence interval:
    lower     upper
0.3515759 0.4686372
```

In this test, we found that the cohen d estimate was only 0.41 which is small, This means that the two groups' mean values differ by no more than 0.41 standard deviation or more.
We also looked at the correlation between the two groups and it appeared to be very small as well. So, we can conclude that the practical significance is small, even though the statistical significance is high.

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.

Age:

-- The independent variable that can be used for this analysis is - "birthyr"

-- this is the participant's year of birth , to calculate the age substract from 2018.

-- I am considering 2018 - the year of this study instead of current year, beacuse we are interested in Survey findings.

Republican / Democrat:

-- There are multiple variables that can be used to identify the voter as Dem / Rep: pidstr -- Would you call yourself a strong (Democrat/Republican) or a not very strong pidlean -- Do you think of yourself as closer to the Republican Party or to the Democratic pid1d -- Generally speaking, do you usually think of yourself as a Democrat, a Republican pid1r -- Generally speaking, do you usually think of yourself as a Republican, a Democrat pid2d -- pid1d: Something else pid2r -- pid1r: Something else

-- We can ignore pid2d , pid2r because these are participants who did not identify as Dem/Rep. -- We can ignore pidlean - this is only asked if IF pid1d=3 OR 4 OR NO ANSWER OR pid1r = 3 OR 4 OR NO ANSWER

-- According to the code book : " pid2d RESPONSE CODE VALUES MATCH pid1d BUT ORDER (2,1,3,4) DIFFERS" , based on this, I would consider responses to either of this questions to consider Rep/Dem.

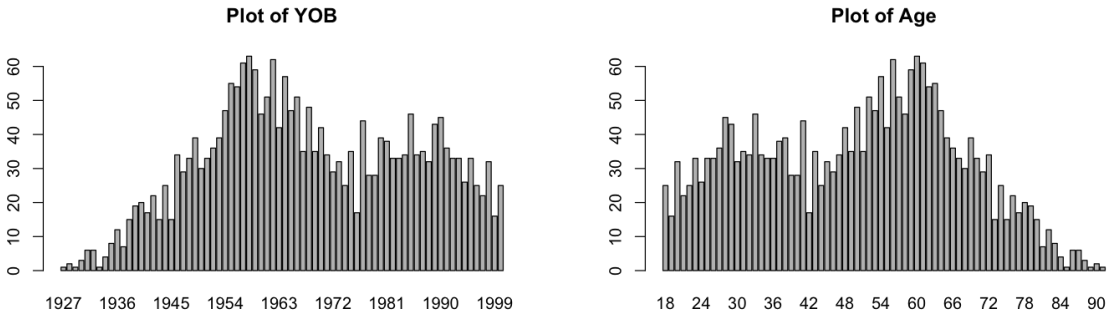
Perform an exploratory data analysis (EDA) of the relevant variables. (5 points)

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.

Age

In [8]:
```
options(repr.plot.width=12, repr.plot.height=4)
par(mfrow=c(1,2))
YOB <- anes_pilot_2018_df$birthyr
YOB_count <- rle(sort(YOB))
YOB_plot <- data.frame(number=YOB_count$values, n=YOB_count$lengths)
barplot(YOB_plot$n,names.arg=YOB_plot$number,main = 'Plot of YOB')

age <- (2018 - YOB)
age_count <- rle(sort(age))
age_plot <- data.frame(number=age_count$values, n=age_count$lengths)
barplot(age_plot$n,names.arg=age_plot$number, main = 'Plot of Age')
```



Republican / Democrat:

In [9]:
```
summary(anes_pilot_2018_df$pid1d)
summary(anes_pilot_2018_df$pid1r)

anes_pilot_2018_df$age <- 2018 - anes_pilot_2018_df$birthyr

A_Rep <- anes_pilot_2018_df[anes_pilot_2018_df$pid1d == 2 | anes_pilot_2
018_df$pid1r == 2,]
A_Dem <- anes_pilot_2018_df[anes_pilot_2018_df$pid1d == 1 | anes_pilot_2
018_df$pid1r == 1,]

paste('Total number of participants who identidy themselves as Republica
ns: ' , length(A_Rep$age))
paste('Total number of participants who identidy themselves as Democrat
s: ' , length(A_Dem$age))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-7.0000 -1.0000 -1.0000  0.4136  2.0000  4.0000 
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  -7.00   -1.00   -1.00    0.49    2.00    4.00 
```

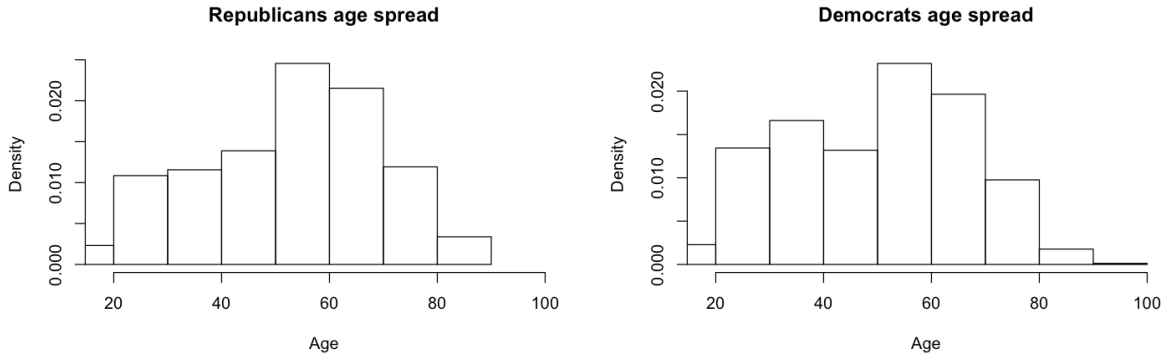
'Total number of participants who identidy themselves as Republicans: 562'

'Total number of participants who identidy themselves as Democrats: 789'

In [10]:
```
v_rep <- (anes_pilot_2018_df[anes_pilot_2018_df$pid1d == 2 | anes_pilot_
2018_df$pid1r == 2,]$age)
v_dem <- (anes_pilot_2018_df[anes_pilot_2018_df$pid1d == 1 | anes_pilot_
2018_df$pid1r == 1,]$age)

par(mfrow=c(1,2))
hist(v_rep, breaks = 10, main = "Republicans age spread"
                  ,xlim = c(18,100) , freq = FALSE, xlab = "Age")

hist(v_dem, breaks = 10, main = "Democrats age spread"
                  ,xlim = c(18,100) , freq = FALSE, xlab = "Age")
```



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

Explain why your test is the most appropriate choice, focusing on its statistical assumptions.

-- Since Age is a cardinal variable, we would consider T-Test -- Since we are trying to answer if Rep. Voters are Older or Younger than Dem Voters, We will perform 2-tail T-test comparing means.

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 [11]:
```
t.test(v_rep,v_dem,alternative = c("two.sided"))
```

```
        Welch Two Sample t-test

data:  v_rep and v_dem
t = 2.9506, df = 1202, p-value = 0.003233
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9166479 4.5546590
sample estimates:
mean of x mean of y
 53.19573  50.46008
```

-- P-value is less than 0.05 , so we reject the Null hypthesis which is. 'True difference between the means is equal to zero' -- since the mean(x) > mean(y), we can conclude that the 'Republican voters are older than Democratic voters'

In [12]:
```
cohen.d(v_rep,v_dem)
```

```
Cohen's d

d estimate: 0.1631035 (negligible)
95 percent confidence interval:
     lower      upper
0.05464599 0.27156095
```

cohen test concludes that these two groups don't differ by more than 0.22, therefore, there is very small practical significance.

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.

To address this question the data set provides 3 variables that relate to beliefs of voters on Russian election interference.
The 3 variable are **'russia16'**, **'muellerinv'** and **'coord16'**

To determine which test statistics to apply to answer the above question we look at the 3 related variables to see which of them are to be observed and measured:
'russia16' and **'coord16'** are ordinal variable with 2 levels, while **'muellerinv'** is ordinal variable with 7 levels

For our study we choose **'muellerinv'** (mueller investigation) variable. We specifically choose this variable because the above question is directly related to it.

The above question is asking if the voters believe that investigation is baseless or not - If the voters think its baseless then they disapprove the investigation or on the other side, if they think it does have some basis they approve the investigation - this is exactly what the **'muellerinv'** variable addresses

We do not consider **'russia16'** and **'coord16'** for this study for the following reasons:

1. The beliefs of voters for **'russia16'**, i.e. if russia interfered or not does not directly apply to their belief in conducting, approving or disapproving federal investigations
2. The beliefs of voters for **'coord16'**, i.e. if Donald Trump's 2016 campaign probably coordinated with the Russians does not apply directly to their belief in conducting, approving or disapproving federal investigations

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.

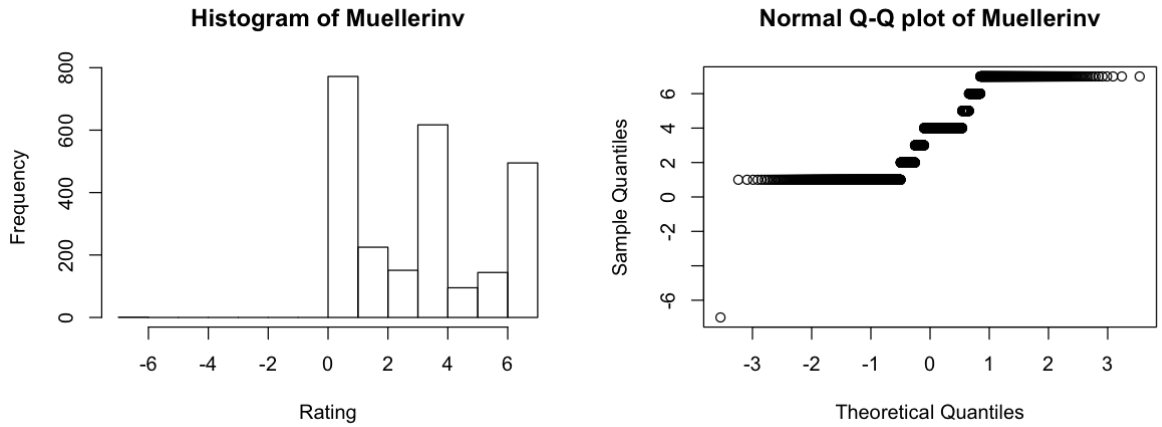
In [13]:
```
head(anes_pilot_2018_df[c('version', 'caseid', 'weight', 'russia16', 'muellerinv', 'coord16')])
summary(anes_pilot_2018_df[c('russia16', 'muellerinv', 'coord16')])
```

	version	caseid	weight	russia16	muellerinv	coord16
	ANES 2018 Pilot Study main version 20190129	1	1.2555080	2	4	2
	ANES 2018 Pilot Study main version 20190129	2	0.5694911	1	1	2
	ANES 2018 Pilot Study main version 20190129	3	0.9720616	1	5	1
	ANES 2018 Pilot Study main version 20190129	4	2.4500732	2	4	2
	ANES 2018 Pilot Study main version 20190129	5	1.6348200	1	2	2
	ANES 2018 Pilot Study main version 20190129	6	1.8184054	2	4	1

```
    russia16        muellerinv        coord16      
 Min.   :-7.000   Min.   :-7.000   Min.   :-7.000  
 1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.000  
 Median : 1.000   Median : 4.000   Median : 1.000  
 Mean   : 1.415   Mean   : 3.555   Mean   : 1.468  
 3rd Qu.: 2.000   3rd Qu.: 6.000   3rd Qu.: 2.000  
 Max.   : 2.000   Max.   : 7.000   Max.   : 2.000  
```

Above we show a snippet of the three related variables we had considered initially and their summaries.

In [14]:
```
options(repr.plot.width=10, repr.plot.height=4)
par(mfrow=c(1,2))
hist(A$muellerinv,breaks=10,main='Histogram of Muellerinv',xlab='Rating')
qqnorm(A$muellerinv, main='Normal Q-Q plot of Muellerinv')
```



We can see some evidence of a non-normal distribution. This can be confirmed by looking at the qq plot

In [15]:
```
C <-anes_pilot_2018_df[anes_pilot_2018_df$muellerinv>=0, ]
miss = dim(anes_pilot_2018_df)-dim(C)
paste('Total responses in muellerinv: ', dim(anes_pilot_2018_df)[1])
paste('Skipped/Non-responses in muellerinv: ', miss[1])
```

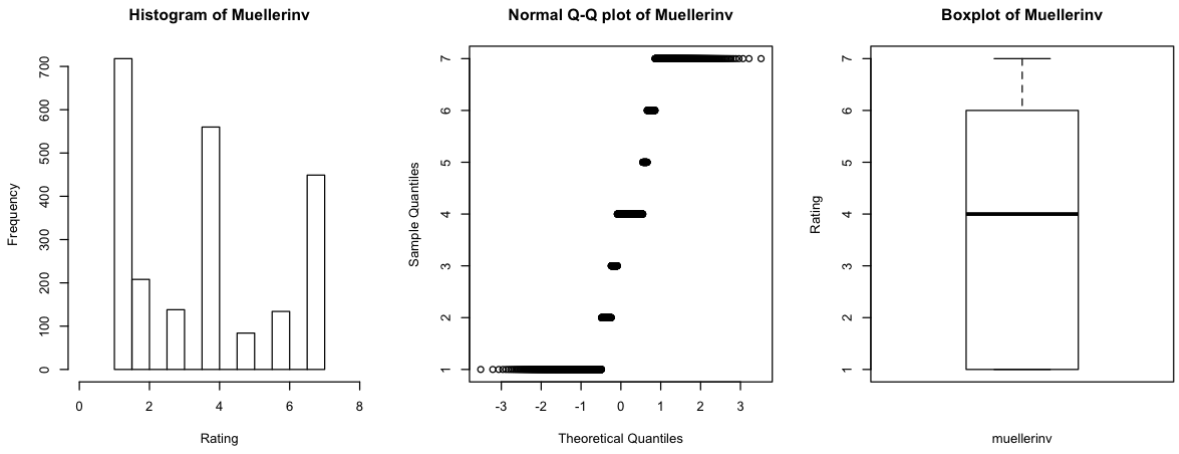
'Total responses in muellerinv: 2292'

'Skipped/Non-responses in muellerinv: 1'

- **Additionally, there is a outlier at -7 due to non-response code**

In this case, we could be empirical and try do two models - one with non-responses taken as missing values (e.g. exclude them from analysis) and second with non-responses recoded as a middle position on scales. Then just compare results.
However, since the proportion of nonresponses is 1/2292 = 0.00043 which is very low, we will exclude them from our study sample

In [16]:
```
par(mfrow=c(1,3))
options(repr.plot.width=10, repr.plot.height=4)
hist(C$muellerinv,breaks=10,main='Histogram of Muellerinv',xlab='Rating'
,xlim=c(0,8))
qqnorm(C$muellerinv, main='Normal Q-Q plot of Muellerinv')
boxplot(C$muellerinv,  ylab = "Rating", main='Boxplot of Muellerinv',
        xlab = "muellerinv")
```



We can see some evidence of a non-normal distribution. This can be confirmed by looking at the qq plot
We also show the box-plot showing the median and the quartiles.

In [17]:
```
summary(C$muellerinv)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00    1.00    4.00    3.56    6.00    7.00 
```

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

Explain why your test is the most appropriate choice, focusing on its statistical assumptions.

Based on the EDA, We have chosen to perform a Wilcoxon test on the muellerinv variable
Since our sample is Non-parametric, we use a one-sample Wilcoxon signed rank test. We think it is a good alternative to one-sample t-test when the data cannot be assumed to be normally distributed. It can be used to determine whether the median of the sample is equal to/less than/greater than a known standard value (i.e. theoretical value).

The main reasons for our choice are:

```
1. The variable is ordinal and uses likert scale
2. Looking at the histogram for the varible the data does not sho
w normality,
   one could argue that since the sample is large we could apply
 CLT here but since the data itself is ordinal
   and distribution seems very skewed from normal we stick to app
lying non-parametric tests on the data
```

Additonally, our Null Hypothesis is as follow:

- We hypothesize that the voters think that investigations are baseless i.e they see no basis at all for the investigation.
- Since the direction from 'no basis' is on the lower numeric side i.e. more basis based on the scale is towards value '1', we choose a one-tailed test i.e our alternate hypothesis is that the median voter belief is less than atleast 6 (Disapprove moderately strongly)

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 [18]:
```
# One-sample wilcoxon test
result <- wilcox.test(C$muellerinv, mu = 6, alternative = "less")
result
```

```
        Wilcoxon signed rank test with continuity correction

data:  C$muellerinv
V = 119880, p-value < 2.2e-16
alternative hypothesis: true location is less than 6
```

In [19]:
```
#### Statistical significance -  p-value
result$p.value
```

1.80748100382539e-290

The p-value of the test is ~0, p-value < 2.2e-16, which is less than the significance level alpha = 0.05.

We can reject the null hypothesis that the voters think that investigations are baseless i.e they see no basis at all for the investigation.
We conclude that voters do believe that there is some basis to federal investigations of Russian election interference

Once we determine p-value for statistical significance. We look at what is the overall size of the effect? we can see practical significance by performing cohen.d test

In terms of interpreting the test values > .8 are large effects, < 0.2 are small effects and values in between are moderate effects

In [20]:
```
cohensD(C$muellerinv, mu=6)
```

1.08204409043491

We ran the cohen test between the mullerinv and the theoritical mean of 6, because the hypothesis is that the voters believed that the investigation was baseless.
The cohen D test showed that the cohen D number is greater than 0.8 which indicated that there is large practical significance.

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

Introduce your topic briefly. (5 points)

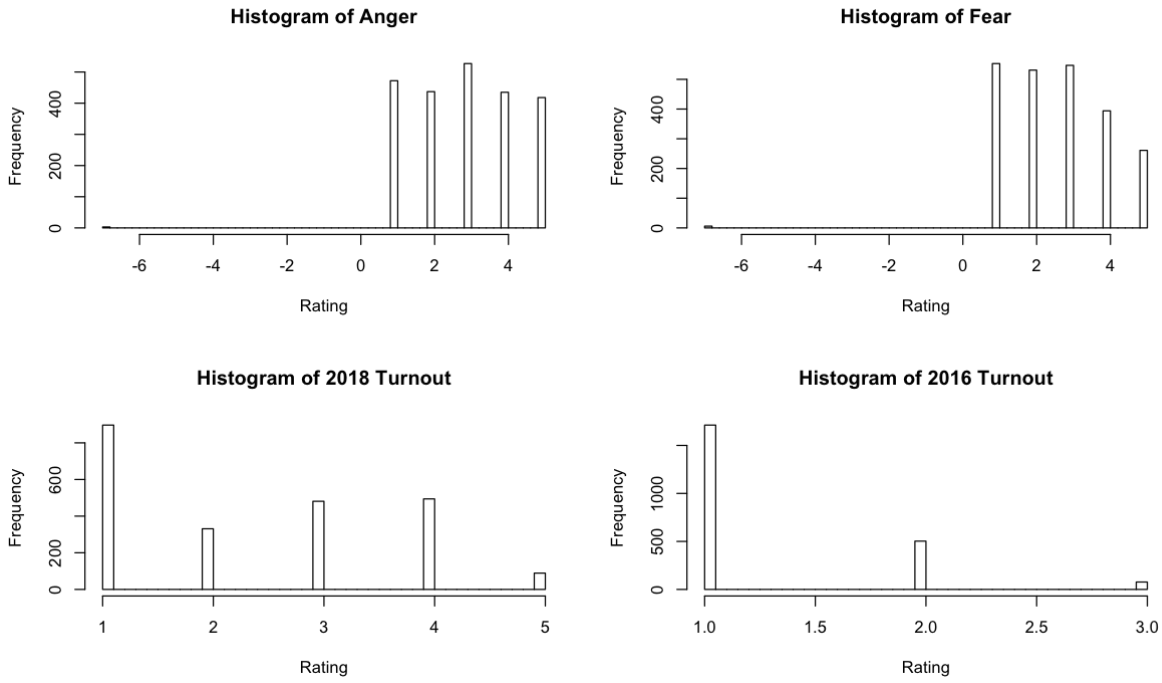
Explain how your variables are operationalized. Comment on any gaps that you can identify between your operational definitions and the concepts you are trying to study.

The independent variables here are the level of anger and and level of fear. The dependent variable here is the increase in voter turnout from 2016 to 2018. In the dataset, there are three different measurements of anger and fear. We feel that the general election anger and fear are most revelent and more broad based measurement, therefore, we chose this for our test. The 2016 turnout is on a 3 level scale while the 2018 turnout is on a 5 level scale. We want to compare the means of anger and fear within the subset of voters who didn't vote in 2016 but voted in 2018, which might indicate if the voters changed from not voting to voting driven by more fear or anger. The gap maybe that anger and fear might be actually dependent to each other. If that was the case, then our test would be misleading since our assumption is that they are independent variables.

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 [21]:
```
anger <-(anes_pilot_2018_df$geangry)
fear <-(anes_pilot_2018_df$geafraid)
t18 <-(anes_pilot_2018_df$turnout18)
t16 <-(anes_pilot_2018_df$turnout16)
options(repr.plot.height=6)
par(mfrow=c(2,2))
hist(anger,breaks=50,main='Histogram of Anger',xlab='Rating')
hist(fear,breaks=50,main='Histogram of Fear',xlab='Rating')
hist(t18,breaks=50,main='Histogram of 2018 Turnout',xlab='Rating')
hist(t16,breaks=50,main='Histogram of 2016 Turnout',xlab='Rating')
```



Exploring the histograms of different datasets. In the fear and anger histograms, we can see that there are a few voters who didn't answer these two questions, which should be excluded from the test since they don't bear any pratical meaning to the test. The histograms of 2018 and 2016 turnouts showed that they are on different scales. However, after reading the descriptions of the different ratings, we concluded that rating 1,2,3 in 2018 turnout are the voters who voted in 2018, rating 1,3 are the voters who voted in 2016. Therefore, we will create the subset of the voters who didn't vote in 2016 (2) but voted in 2018 (1,2,3) and we will exclude the voters who didn't answer the fear or anger questions (-7).

In [22]:
```
#checking the two records that has -7 in
a <- anes_pilot_2018_df[anes_pilot_2018_df$geangry<0,]
f <- anes_pilot_2018_df[anes_pilot_2018_df$geafraid<0,]
```

In [23]:
```
summary(anger)
summary(fear)
summary(t18)
summary(t16)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -7.000   2.000   3.000   2.939   4.000   5.000 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -7.000   2.000   3.000   2.659   4.000   5.000 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   1.000   2.000   2.366   4.000   5.000 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   1.000   1.000   1.287   2.000   3.000 
```

In [24]:
```
x <- anes_pilot_2018_df[anes_pilot_2018_df$turnout16 != 2,]$geangry #$turnout16
#paste(head(x))
x_2 <- anes_pilot_2018_df[anes_pilot_2018_df$turnout16 == 1,]$geafraid
#paste(head(x_2))
y <- anes_pilot_2018_df[anes_pilot_2018_df$turnout18 == 1 | anes_pilot_2018_df$turnout18 == 2 | anes_pilot_2018_df$turnout18 == 3,]$turnout18
#paste(length(y))

didnotvote10voted18_geafraid <- anes_pilot_2018_df[(anes_pilot_2018_df$turnout18 == 1 | anes_pilot_2018_df$turnout18 == 2 | anes_pilot_2018_df$turnout18 == 3) & anes_pilot_2018_df$turnout16 == 2 & anes_pilot_2018_df$geafraid > 0, ]$geafraid
# head(didnotvote10voted18_geafraid)

didnotvote10voted18_geangry <- anes_pilot_2018_df[(anes_pilot_2018_df$turnout18 == 1 | anes_pilot_2018_df$turnout18 == 2 | anes_pilot_2018_df$turnout18 == 3) & anes_pilot_2018_df$turnout16 == 2 & anes_pilot_2018_df$geangry > 0, ]$geangry
# head(didnotvote10voted18_geangry)
```

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

Explain why your test is the most appropriate choice, focusing on its statistical assumptions.

- Based on the fact that the two variables are both ordinal variables, we will conduct a wilcoxon test.

> The main reasons for our choice are:
>
> 1. The variables are ordinal and uses likert scale
> 2. Looking at the histogram for the varibles the data does not show normality, one could argue that since the sample is large we could apply CLT here but since the data itself is ordinal and distribution seems very skewed from normal we stick to applying non-parametric tests on the data

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 [25]:
```
wilcox.test(didnotvote10voted18_geafraid,didnotvote10voted18_geangry)
```

```
        Wilcoxon rank sum test with continuity correction

data:  didnotvote10voted18_geafraid and didnotvote10voted18_geangry
W = 3821.5, p-value = 0.8802
alternative hypothesis: true location shift is not equal to 0
```

In [26]:
```
cohen.d(didnotvote10voted18_geafraid,didnotvote10voted18_geangry)
```

```
Cohen's d

d estimate: -0.01862199 (negligible)
95 percent confidence interval:
     lower      upper
-0.3161928  0.2789489
```

The hypothesis is that the mean of fear and anger for voters who didn't vote in 2016 but voted in 2018 would be the same. Our test result showed that P value is very high, therefore, it is not statistically significant enough to reject the hypothesis. The cohen test resulted a very small d estimates of 0.027 which is negligible. That indicates that there was no practical significance in the data.

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.

For this section we will look at how Trump Approval rating affected voter turnout in 2018 for the voters who identify themselves as republicans and those that identify themselves as non-republicans.

The variable we will use for approval rating is: 'apppres' which is described as "Do you approve, disapprove, or neither approve nor disapprove of the way Donald Trump is handling his job as president?" and ranges from 1 to 7.

Another variable we will use for this section is the 'pid1d' which is described as "Generally speaking, do you usually think of yourself as a Democrat, a Republican, an independent, or what?" and ranges from 1 to 4 as: **Democrat [1]** Republican [2] **independent [3]** something else [4] And simultaneously we will use 'pid1r' which is described as: Generally speaking, do you usually think of yourself as a Republican, a Democrat, an independent, or what? **Republican [2]** Democrat [1] **independent [3]** something else [4]

The final variable we will use for this analysis is 'turnout18' which is described as "In the election held on November 6, did you definitely vote in person on election day, vote in person before Nov 6, vote by mail, did you definitely not vote, or are you not completely sure whether you voted in that election?

Definitely voted in person on Nov 6 [1] Definitely voted in person, before Nov 6 [2] **Definitely voted by mail [3]** Definitely did not vote [4] __ Not completely sure [5]

The dataset we will use has already filtered out the nonserious and dishonest records (only the absolute ones, that is who answer as 'Never' honest and 'Never' serious in answering the questions on the survey) out before we start processing the data. We are assuming this data if left in would be contaminating the dataset.

We will filter out non-voters using condition turnout18 !=4

In order to analyze how Trump Approval rating manifested itself in the voter pattern for people who identify themselves as Republicans we will: Filter all the data from variable 'apppres' only for the voters who identify themselves as Republicans to ensure the dataset remains consistent - that is we will check the approval ratings against the turnout in 2018 election against the same dataset.

In order to analyze how Trump Approval rating manifested itself in the voter pattern for people who do not identify themselves as Republicans we will: Filter all the data from variable 'apppres' only for the voters who do not identify themselves as Republicans to ensure the dataset remains consistent - that is we will check the turnout in 2018 election against the same dataset.

The potential gap is that these two variables Trump job approval apppres and the voter turnout for 2018 election might not be truely dependent which we assumed for our test.

Another assumption here is that the sample size is adequate and the samples were randomly selected and there is no clustering in the sample, which means the sample is i.i.d.

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, focusing on its statistical assumptions.

Let us first check the smmary of all the 4 variables to ensure there are no outliers:

In [27]:
```
paste("Summary of Trump presidency approval rating apppres:") # ordinal
summary(anes_pilot_2018_df$apppres)
paste("Mean of Trump presidency approval rating: ", mean(anes_pilot_2018
_df$apppres))

paste("Summary of Voter Turnout for 2018 election turnout18:") # ordinal
summary(anes_pilot_2018_df$turnout18)
paste("Mean of Voter Turnout for 2018 election: ", mean(anes_pilot_2018_
df$turnout18))

paste("Summary of People who identify themselves as as a Democrat, a Rep
ublican, an independent, or something else: pid1d:")
summary(anes_pilot_2018_df$pid1d)
paste("Mean of pid1d: ", mean(anes_pilot_2018_df$pid1d))

paste("Summary of People who identify themselves as as a Democrat, a Rep
ublican, an independent, or what pid1r:")
summary(anes_pilot_2018_df$pid1r)
paste("Mean of pid1r: ", mean(anes_pilot_2018_df$pid1r))

paste("Summary of Voter Turnout for 2018 election turnout18 who identify
themselves as Republicans and voted:")
v_repubturnout <- anes_pilot_2018_df[(anes_pilot_2018_df$pid1d == 2 | an
es_pilot_2018_df$pid1r == 2) & anes_pilot_2018_df$turnout18 !=4 ,]$turno
ut18
summary(v_repubturnout)

paste("Summary of Voter Turnout for 2018 election turnout18 who do not i
dentify themselves as Republicans and voted:")
v_repubturnout <- anes_pilot_2018_df[anes_pilot_2018_df$pid1d != 2 & ane
s_pilot_2018_df$pid1r != 2 & (anes_pilot_2018_df$pid1d > 0 | anes_pilot_
2018_df$pid1r > 0) & anes_pilot_2018_df$turnout18 !=4 ,]$turnout18
summary(v_repubturnout)
```

'Summary of Trump presidency approval rating apppres:'

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00    2.00    5.00    4.56    7.00    7.00 
```

'Mean of Trump presidency approval rating: 4.56020942408377'

'Summary of Voter Turnout for 2018 election turnout18:'

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   1.000   2.000   2.366   4.000   5.000 
```

'Mean of Voter Turnout for 2018 election: 2.36605584642234'

'Summary of People who identify themselves as as a Democrat, a Republican, an independent, or something else: pid1d:'

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-7.0000 -1.0000 -1.0000  0.4136  2.0000  4.0000 
```

'Mean of pid1d: 0.413612565445026'

'Summary of People who identify themselves as as a Democrat, a Republican, an independent, or what pid1r:'

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  -7.00   -1.00   -1.00    0.49    2.00    4.00 
```

'Mean of pid1r: 0.489965095986038'

'Summary of Voter Turnout for 2018 election turnout18 who identify themselves as Republicans and voted:'

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   1.000   1.000   1.728   2.000   5.000 
```

'Summary of Voter Turnout for 2018 election turnout18 who do not identify themselves as Republicans and voted:'

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   1.000   2.000   1.915   3.000   5.000 
```

The above data shows that the variables pid1d and pid1r shows some unexpected data. Since we will be basing our datasets on this variable we need to check if there is a subset that identifies themselves as Republicans meaning that variable pid1d == 2 or pid1r == 2

In [28]:
```
#(anes_pilot_2018_df[anes_pilot_2018_df$pid1d == 2,]$pid1r)
#(anes_pilot_2018_df[anes_pilot_2018_df$pid1r == 2,]$pid1d)
rep <- anes_pilot_2018_df[anes_pilot_2018_df$pid1d == 2 | anes_pilot_201
8_df$pid1r == 2,]
# Republicans identified themselves through variable pid1d
rep_v <- rep[rep$pid1d == 2,]$pid1d
# Republicans identified themselves through variable pid1r
rep_v2 <- rep[rep$pid1r == 2,]$pid1r
paste("There are ", length(rep_v) + length(rep_v) , " people who identif
y themselves as Republicans")
```

'There are 598 people who identify themselves as Republicans'

In [29]:
```
non_rep <- anes_pilot_2018_df[anes_pilot_2018_df$pid1d != 2 & anes_pilot
_2018_df$pid1r != 2 & (anes_pilot_2018_df$pid1d > 0 | anes_pilot_2018_df
$pid1r > 0),]
nonrep_v <- non_rep[non_rep$pid1d != 2 & non_rep$pid1d > 0,]$pid1d
nonrep_v2 <- non_rep[non_rep$pid1r != 2 & non_rep$pid1r > 0,]$pid1r
paste("There are ", length(nonrep_v) , " people who do not identify them
selves as Republicans")
```

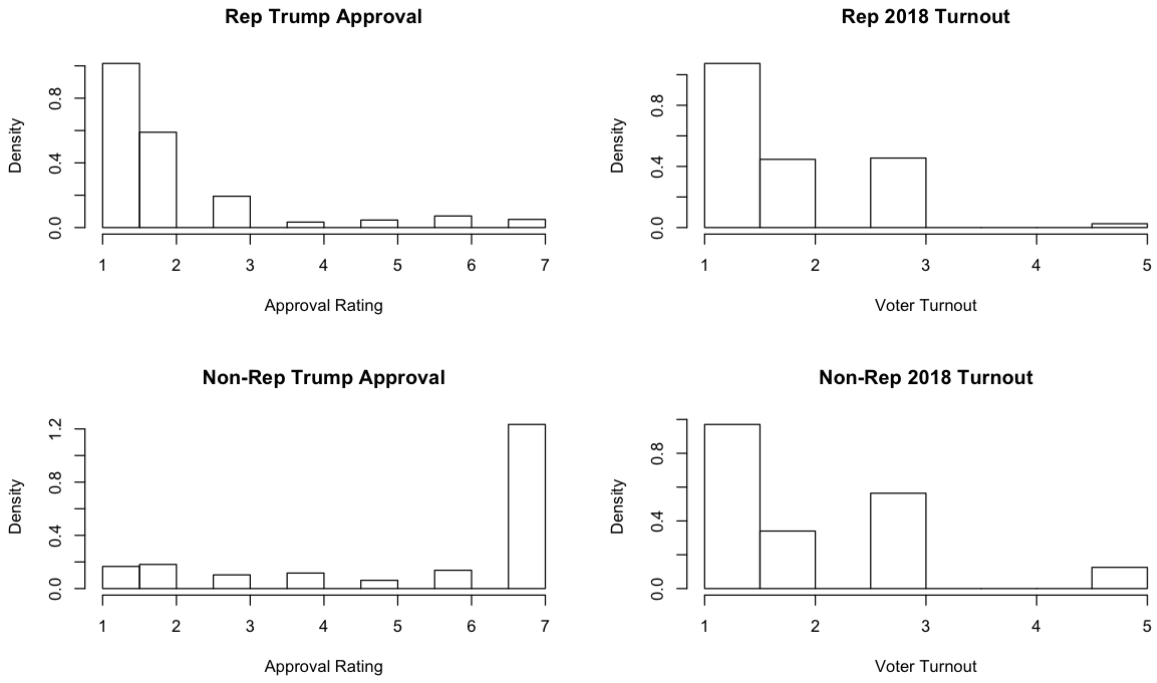
'There are 772 people who do not identify themselves as Republicans'

In [30]:
```
# Approval rating for the Trump presidency for the dataset identifying t
hemselves as Republicans
par(mfrow=c(2,2))
options(repr.plot.width=10, repr.plot.height=6)
v_trumpapproval <-(anes_pilot_2018_df[(anes_pilot_2018_df$pid1d == 2 | a
nes_pilot_2018_df$pid1r == 2) & anes_pilot_2018_df$turnout18 !=4,]$apppr
es)
#paste("Approval rating for the Trump presidency against voter turnout f
or the dataset identifying themselves as Republicans")
hist(v_trumpapproval, breaks = 10, main = "Rep Trump Approval"
     ,xlim = c(1,7), freq = FALSE, xlab = "Approval Rating")

# Turnout for 2018 election for voters identifying themselves as Republi
cans
v_repubturnout <- (anes_pilot_2018_df[(anes_pilot_2018_df$pid1d == 2 | a
nes_pilot_2018_df$pid1r == 2) & anes_pilot_2018_df$turnout18 !=4 ,]$turn
out18)
hist(v_repubturnout, breaks = 10, main = "Rep 2018 Turnout"
     ,xlim = c(1,5) , freq = FALSE, xlab = "Voter Turnout")

#paste("Approval rating for the Trump presidency against voter turnout f
or the dataset not identifying themselves as Republicans")
# Approval rating for the Trump presidency for the dataset not identifyi
ng themselves as Republicans
v_trumpapproval2 <-(anes_pilot_2018_df[anes_pilot_2018_df$pid1d != 2 & a
nes_pilot_2018_df$pid1r != 2 & anes_pilot_2018_df$turnout18 !=4,]$apppre
s)
hist(v_trumpapproval2, breaks = 10, main = "Non-Rep Trump Approval"
     ,xlim = c(1,7) , freq = FALSE, xlab = "Approval Rating")

# Turnout for 2018 election for voters not identifying themselves as Rep
ublicans
v_voterturnout <- (anes_pilot_2018_df[anes_pilot_2018_df$pid1d != 2 & an
es_pilot_2018_df$pid1r != 2 & anes_pilot_2018_df$turnout18 !=4,]$turnout
18)
hist(v_voterturnout, breaks = 10, main = "Non-Rep 2018 Turnout"
     , xlim = c(1,5) , freq = FALSE, xlab = "Voter Turnout")
```



We could also pull up a qq-plot to see how normal this variable looks.

A perfectly normal variable would show up as a nice diagonal line on the qq-lot.

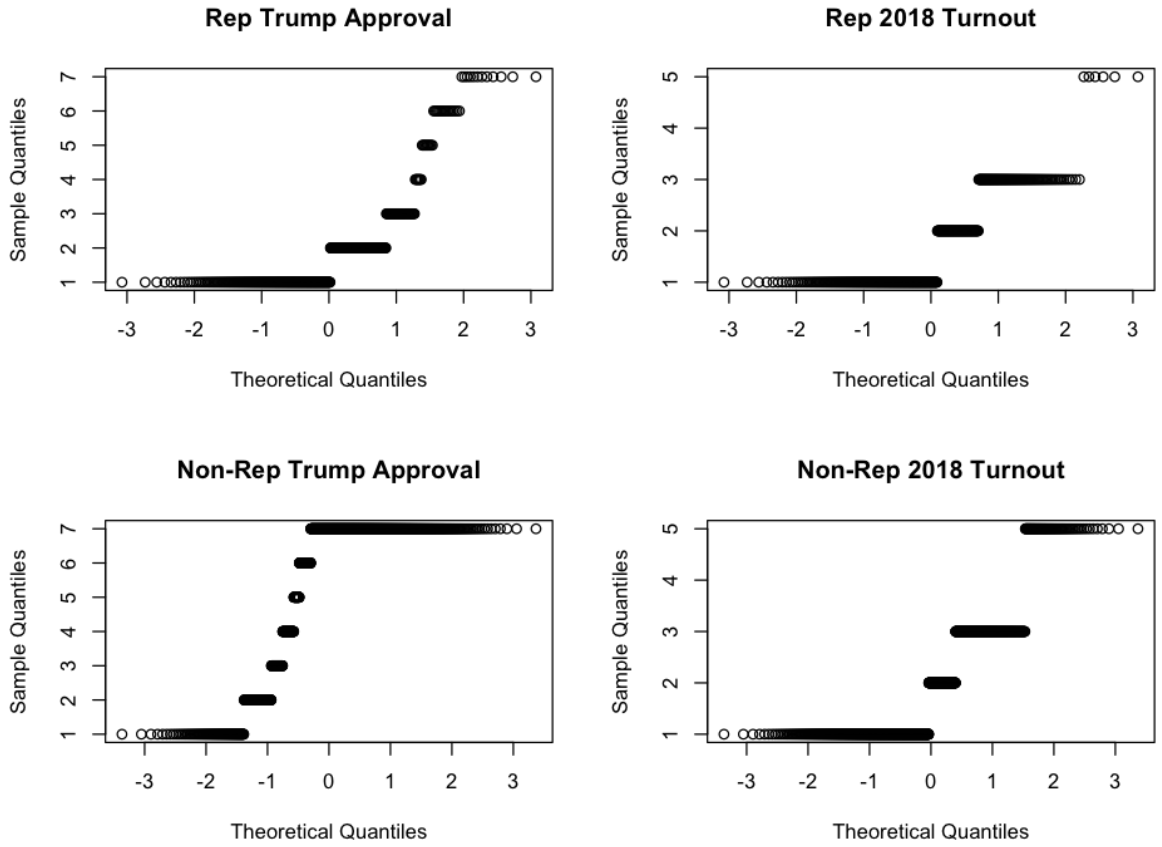
We can see some evidence of a non-normal distribution. This can be confirmed by looking at the qq plot

In [31]:
```
# Approval rating for the Trump presidency for the dataset identifying t
hemselves as Republicans
par(mfrow=c(2,2))
options(repr.plot.width=8, repr.plot.height=6)
v_trumpapproval <-(anes_pilot_2018_df[(anes_pilot_2018_df$pid1d == 2 | a
nes_pilot_2018_df$pid1r == 2) & anes_pilot_2018_df$turnout18 !=4,]$apppr
es)
qqnorm(v_trumpapproval, main = "Rep Trump Approval")

# Turnout for 2018 election for voters identifying themselves as Republi
cans
v_repubturnout <- (anes_pilot_2018_df[(anes_pilot_2018_df$pid1d == 2 | a
nes_pilot_2018_df$pid1r == 2) & anes_pilot_2018_df$turnout18 !=4 ,]$turn
out18)
qqnorm(v_repubturnout, main = "Rep 2018 Turnout")

#paste("Approval rating for the Trump presidency against voter turnout f
or the dataset not identifying themselves as Republicans")
# Approval rating for the Trump presidency for the dataset not identifyi
ng themselves as Republicans
v_trumpapproval2 <-(anes_pilot_2018_df[anes_pilot_2018_df$pid1d != 2 & a
nes_pilot_2018_df$pid1r != 2 & anes_pilot_2018_df$turnout18 !=4,]$apppre
s)
qqnorm(v_trumpapproval2, main = "Non-Rep Trump Approval")

# Turnout for 2018 election for voters not identifying themselves as Rep
ublicans
v_voterturnout <- (anes_pilot_2018_df[anes_pilot_2018_df$pid1d != 2 & an
es_pilot_2018_df$pid1r != 2 & anes_pilot_2018_df$turnout18 !=4,]$turnout
18)
qqnorm(v_voterturnout, main = "Non-Rep 2018 Turnout")
```



Since the Trump approval rating and voter turnout in 2018 are both ordinal variables we will be using the Wilcoxon Signed-Rank Test. The main reasons for our choice are:

```
1. The variables are ordinal and uses likert scale
2. Looking at the histogram for the varibles the data does not sh
ow normality,
   one could argue that since the sample is large we could apply
 CLT here but since the data itself is ordinal
   and distribution seems very skewed from normal we stick to app
lying non-parametric tests on the data
```

The Null Hypothesis in this case will be that the Voter pattern will follow the Trump approval rating for Republicans, that is for the sub-section approving the Trump job. For the Non-Republicans the Trump approval would be inversed - that is for people approving of the Trump job this sub-section would vote to ensure Republicans would stay in power.

Conduct your test. (2 points)

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

In [32]:
```
paste("Wilcoxon rank sum test for Republican voters")
#t.test(v_trumpapproval , v_repubturnout, paired=T)
republican_voters <- wilcox.test(v_trumpapproval , v_repubturnout, paire
d = TRUE)
republican_voters
```

'Wilcoxon rank sum test for Republican voters'

```
        Wilcoxon signed rank test with continuity correction

data:  v_trumpapproval and v_repubturnout
V = 26566, p-value = 0.1356
alternative hypothesis: true location shift is not equal to 0
```

The wilcoxon test showed a P value > 0.05, which indicates that we fail to reject our NULL hypothesis.
So we cannot make a conclusion on our alternate hypothesis that states that republican voters who a high Trump job approval rating will vote less.

In [33]:
```
#### Statistical significance -  p-value
paste("p-value for Republican voter turnout for 2018:", republican_voter
s$p.value)
cohen.d(v_trumpapproval , v_repubturnout)
paste("Correlation for Republican voter turnout for 2018:", cor(v_trumpa
pproval , v_repubturnout))
```

'p-value for Republican voter turnout for 2018: 0.135605544568047'

```
Cohen's d

d estimate: 0.1942987 (negligible)
95 percent confidence interval:
     lower      upper
0.06665671 0.32194060
```

'Correlation for Republican voter turnout for 2018: -0.00795221549853379'

The cohen test shows a small value for the d estimate, therefore, there is no practical significance.

In [34]:
```
paste("Wilcoxon rank sum test for Non-Republican voters")
#t.test(v_trumpapproval2 , v_voterturnout, paired=T)
non_republican_voters <- wilcox.test(v_trumpapproval2 , v_voterturnout,
paired = TRUE)
non_republican_voters
```

'Wilcoxon rank sum test for Non-Republican voters'

```
        Wilcoxon signed rank test with continuity correction

data:  v_trumpapproval2 and v_voterturnout
V = 734050, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

The wilcoxon test showed a small P value which indicated statistical significance in the data, so we can reject the hypothesis that Trump approval rating for voters who identified themselves as non-republican also voted in 2018.

In [35]:
```
#### Statistical significance -  p-value
paste("p-value for Non-Republican voter turnout for 2018:", non_republic
an_voters$p.value)
cohen.d(v_trumpapproval2 , v_voterturnout)
paste("Correlation for Non-Republican voter turnout for 2018:", cor(v_tr
umpapproval2 , v_voterturnout))
```

'p-value for Non-Republican voter turnout for 2018: 1.07254816160257e-185'

```
Cohen's d

d estimate: 2.051567 (large)
95 percent confidence interval:
   lower    upper
1.957383 2.145751
```

'Correlation for Non-Republican voter turnout for 2018: 0.00325737465852943'

The cohen test is showing large value for the d estimate, therefore, there is practical significance.

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.

The first wilcoxon test showed a P value > 0.05 , which indicates that we fail to reject our NULL hypothesis.
So we cannot make a conclusion on our alternate hypothesis that states that republican voters who a high Trump job approval rating will vote less.
The cohen test shows a small value for the d estimate, therefore, there is no practical significance.

The second wilcoxon test showed a small P value which indicated statistical significance in the data, so we can reject the hypothesis that Trump approval rating for voters who identified themselves as non-republicans also voted in 2018. The cohen test showing large value for the d estimate, therefore, there is practical significance.

It is reasonable to conclude that we cannot draw any conclusions based on republican voter pattern based on Trump approval rating.
However, from our second wilcoxon test we can conclude that the Non-Republican voters who approved Trump's job also voted in 2018. Based on our study the conclusions seems counter-intuitive to the data observed in Trump approval, However we factored in the voter turnout and independent candidate supporters.
It might help us design the strategy to encourage more republican voters to keep voting in the future elections.