Lab 2

w203 Section 2 Penner Mon (6:30pm)

Team: Adam Sohn, Alvin Lim, Khyati Tripathi

In [2]:
```
A = read.csv("anes_pilot_2018.csv")
```

In [78]:
```
#install.packages("repr",repos = "http://cran.us.r-project.org")
#install.packages("effsize",repos = "http://cran.us.r-project.org")
```

In [3]:
```
#Global

#Filtering A to only include participants who answered both the below question
s in the manner showed:
    #[nonserious] Only include participants who were 'Never' (1) not serious i
n answers.
    #[honest] Only include participants who 'Always' (5) answered honestly.
A_quality <- A[A$nonserious == 1 & A$honest == 5,]

#importing libraries
library(repr)
library(effsize)

#standard plot size.
options(repr.plot.width=4, repr.plot.height=3)

#Note that all plots should contain below arguments
    #cex.lab=.8, cex.axis=.8, cex.main=.8, cex.sub=0.8
```

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.

Response

Respect for police is operationalized as [ftpolice], 101 integer inputs between 0,100, where 0(100) is coldest(warmest).

- [ftpolice] "How would you rate the police (with regards to warm/cold feelings towards)?"

Respect for journalists is operationalized as [ftjournal], 101 integer inputs between 0,100, where 0(100) is coldest(warmest).

- [ftjournal] "How would you rate journalists (with regards to warm/cold feelings towards)?"

[ftpolice] and [ftjournal] are the best candidates for variable operationalization. The only difference in the respective questions is the word police/journalist. That much said, warmth of personal feeling is not an ideal analogue for respect. Whereas respect may create warm feelings, a lack of warm feeling does not necessarily connote a lack of respect. For example, some people may fear the police yet still respect them.

[trustmedia] was another candidate considered for variable operationalization. [trustmedia] is explicitly seeking a measure of trust, which is the concept we are trying to study. However, not all respondants may treat 'media' as a direct analogue for 'journalist'. For instance, a responder may distrust the institution of 'media', yet have specific journalists they trust.

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.

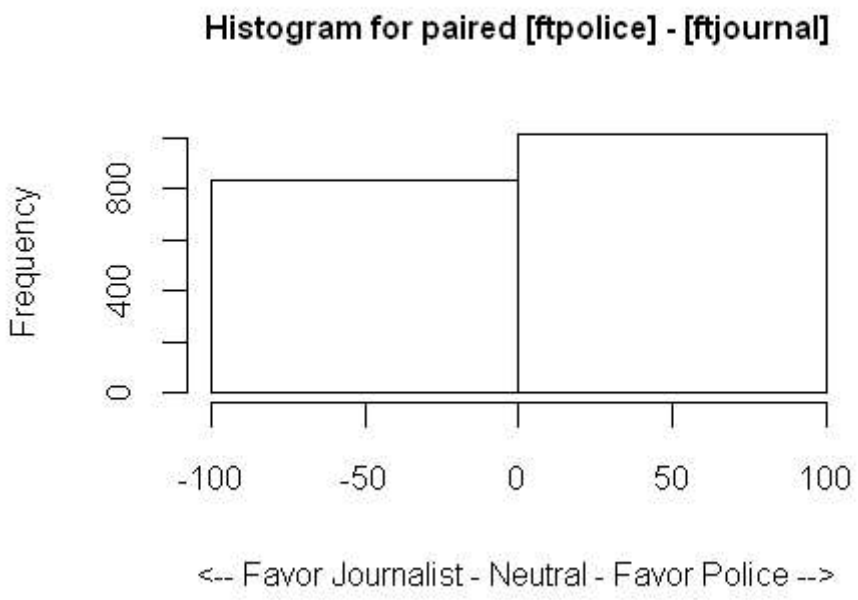
In [4]:
```
#Filtering out any '-7' values which would indicate a non-response on either
 [ftpolice] or [ftjournal]
mod_A_quality = A_quality[A_quality$ftpolice >=0 & A_quality$ftjournal >=0,]

#Creation of a paired (by respondant) delta of sentiment towards police and jo
urnalists.
#Positive delta indicates warmer feelings towards police than journalists.
mod_A_quality$p_j_delta = mod_A_quality$ftpolice - mod_A_quality$ftjournal
```

In [5]:
```
#[ftpolice]
cat('EDA for [ftpolice]: "How would you rate the police?"')
cat("\n\nMin, Max Value (0-100): ",min(A_quality$ftpolice[A_quality$ftpolice>=
0]),",", max(A_quality$ftpolice), "\nUnpaired Mean Value: ", round(mean(A_qual
ity$ftpolice)), "\nUnpaired Median Value: ", round(median(A_quality$ftpolice
)),"\nResponse Count: ", length(A_quality$ftpolice[A_quality$ftpolice >= 0]),
"\nNon-Response Count: ", length(A_quality$ftpolice[A_quality$ftpolice <0]))
strrep('-',50)

#[ftjournal]
cat('EDA for [ftjournal]: "How would you rate the journalists?"')
cat("\n\nMin, Max Value (0-100): ",min(A_quality$ftjournal[A_quality$ftjournal
>=0]),",", max(A_quality$ftjournal), "\nUnpaired Mean Value: ", round(mean(A_q
uality$ftjournal)), "\nUnpaired Median Value: ", round(median(A_quality$ftjour
nal)), "\nResponse Count: ", length(A_quality$ftjournal[A_quality$ftjournal >=
0]),"\nNon-Response Count: ", length(A_quality$ftjournal[A_quality$ftjournal <
0]))

strrep('-',50)
cat('Median delta of respondant-paired feelings towards police and journalist
s:', round(median(mod_A_quality$p_j_delta)),'\nNote1: Positive delta indicates
preference for police. \nNote2: Any non-responses for either [ftpolice] or [ft
journal] are filtered out.')
```

```
EDA for [ftpolice]: "How would you rate the police?"

Min, Max Value (0-100):  0 , 100
Unpaired Mean Value:  67
Unpaired Median Value:  72
Response Count:  1851
Non-Response Count:  0

'--------------------------------------------------'

EDA for [ftjournal]: "How would you rate the journalists?"

Min, Max Value (0-100):  0 , 100
Unpaired Mean Value:  53
Unpaired Median Value:  54
Response Count:  1849
Non-Response Count:  2

'--------------------------------------------------'

Median delta of respondant-paired feelings towards police and journalists: 5
Note1: Positive delta indicates preference for police.
Note2: Any non-responses for either [ftpolice] or [ftjournal] are filtered ou
t.
```

In [6]:
```
cat('Note that histogram is used for ordinal values only in this case that his
togram is merely counting the favor of police or journalists.')
hist(mod_A_quality$p_j_delta, main = "Histogram for paired [ftpolice] - [ftjou
rnal]", breaks = 2,xlab = "<-- Favor Journalist - Neutral - Favor Police -->",
cex.lab=.8, cex.axis=.8, cex.main=.8, cex.sub=0.8)
```

```
Note that histogram is used for ordinal values only in this case that histogr
am is merely counting the favor of police or journalists.
```



Note the created data structures foster reproducability of this study on same basis or others, such as favoring 'capitalists' rather than 'police'.

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.

Response

An appropriate test for establishing a significant delta in 'respect' among the US population between the police and journalists is **Wilcoxon, 2-sided, Signed Rank Test**.

Although [ftpolice] and [ftjournal] data appear to be continuous and Cardinal, the data is Ordinal. The choices are sentiment-based, so whereas we can expect a consistent scale interpretation within a respondant's answers, we can not expect consistency between respondants.

Statistical assumptions for Wilcoxon Signed Rank Test are:

- Paired differences - Each respondant has given their respective input on police and journalists on the same scale. The paired test shows mangitude of individual preference, which is necessary per the research question.
- 2-sided - Although EDA appears to indicate preference of police over journalists, 2-sided is appropriate (as opposed to 1-sided) since we are still open to full statistical study revealing the opposite case.
- Rank-based (Ordinal) numbers - The 0-100 scale operates as an ordinal scale, conveniently with all integers in between as possible responses so re-assignment of scale to integers is not required.

For the test, the following are the hypotheses:

- $H_0: \theta_{police-journalists} = 0$
- $H_a: \theta_{police-journalists} \neq 0$
  where $\theta_{police-journalists}$ is the median paired difference of sentiments towards police and journalists

Another candidate under consideration was binomial test. The binomial test was not chosen as the research question seeks magnitude of sentiment, not a examination of preference. A preference examination would over-represent minor preferences and under-represent extreme preferences.

Also considered, but not chosen, was to treat values as Cardinal and use a Paired, 2-sided t-test. Interestingly, this would yield identical result, but is innapropriate for Ordinal values.

For practical significance we can not use Cohen's d as the data are not continuous. For the Wilcoxon signed rank test, **we will obtain effect size by dividing Z test statistic by square root of the number of observation pairs**.

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 [7]:
```
# Wilcox signed rank test for paired data
wilcoxModel <- wilcox.test(mod_A_quality$ftpolice, mod_A_quality$ftjournal, pa
ired = T,)
wilcoxModel

# Effect size
# Cohen's d is not appropriate as the data are not continuous.
# Some authors (e.g. Pallant, 2007) suggest dividing the Z test statistic by t
he square root of the number of observation pairs
Zstat <- qnorm(wilcoxModel$p.value/2) # Calculate the standardised z statistic
effsize <- abs(Zstat)/sqrt(length(mod_A_quality$ftpolice))
cat("Effect size: ", effsize)
```

```
	Wilcoxon signed rank test with continuity correction

data:  mod_A_quality$ftpolice and mod_A_quality$ftjournal
V = 1033500, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

Effect size:  0.2455704
```

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

Statistical significance

With a statistical significance observed p < 2.2e-16 compared to p_{crit} of 0.05, null hypothesis $\theta_{police-journalists} = 0$ (ie. US voters have equivalent respect for the police and journalists) is rejected. Given the median of paired differences between [ftpolice] and [ftjournal] is positive (+5), the null hypothesis is rejected in the direction favoring the case of $\theta_{police} > \theta_{journalists}$ (US voters have more respect for police than journalists).

Practical significance

The calculated effect size of 0.25 shows that although the statistical significance of the delta in respect for police than journalists is sizeable, the practical significance of the delta is small.

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.

Response

There are several ways to treat political affiliation of respondant.

Respondants could be identified based on their actual voting patterns [house18p]/[senate18p]/[gov18p], stated political leanings [pidlean], personal ideology [pid1d]/[pid1r], and reported party identification [pid7x].

Because voting a party-line is not a given, how one votes is arguably less indicative of one's political leanings than one's self-professed stance.

And while it might be illuminating to juxtapose expressed political identity against actual voting patterns, that is not the purpose of this exercise.

It might also be fruitful including Independents with Democrat/Republican leanings. However, considering the research question, it is most appropriate to define Democrat or Republican based on a [pid7x] response of 1,2 or 6,7 respectively.

Our EDA noted [pid7x] is not a combination of values from other variables. However, as [pid7x] appears to be a direct record of political affiliation, it is likely the best single variable for our purposes.

[pid7x] Party ID summary

- -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

Another consideration here is the word "voters" in the research question. Are all respondants voters? Or is a voter someone who has actually cast a vote in the last elections? If the latter, should we allow into the category people who voted once, twice, or thrice in the previous elections?

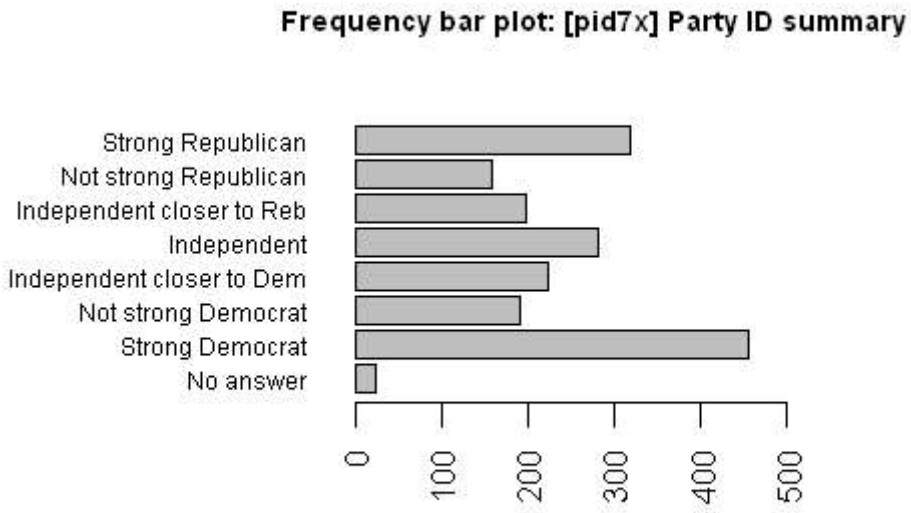
Since we are attempting to compare the ages of Democrats and Republicans, a definition of "voter" that only includes those who had previously voted would remove the youngest slice of the sample. To preserve that slice, and also because in common discussion a "voter" would generally be understood as one able to vote, we shall interpret a "voter" as any person in this dataset.

To operationalize age, [birthyr] tells the respondant's birth year, and has no missing values.

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 [9]:
```
# Quick look at [pid7x]
counts <- table(A_quality$pid7x)
par(las=2) # make label text perpendicular to axis
par(mar=c(5,8,4,2)) # increase y-axis margin.
barlabels <- c("No answer", "Strong Democrat", "Not strong Democrat", "Indepen
dent closer to Dem","Independent","Independent closer to Reb","Not strong Repu
blican","Strong Republican")
barplot(counts,main ='Frequency bar plot: [pid7x] Party ID summary',horiz=TRUE
,names.arg=barlabels,cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8, cex.na
mes=.6, xlim = c(0, max(counts)+100), xpd = FALSE)
```

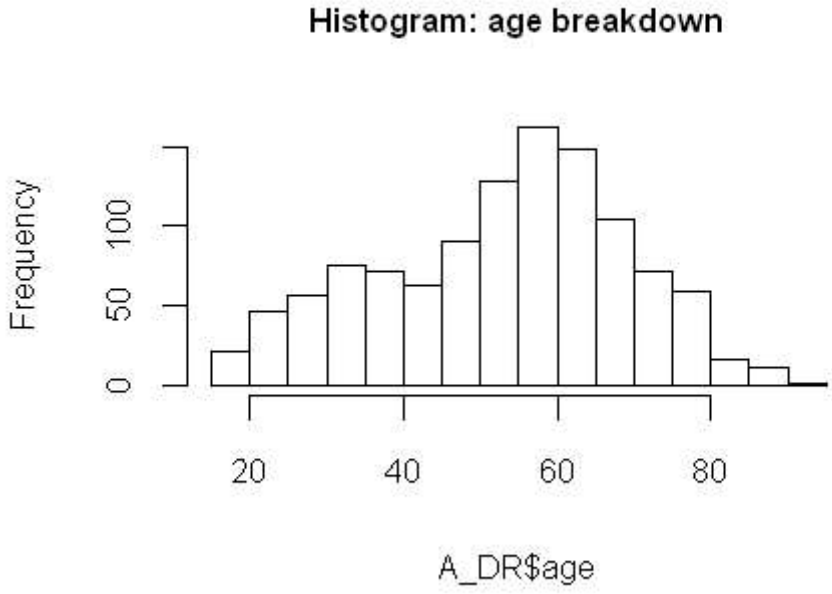


In [10]:
```
cat("Let's now subset a dataframe containing only voters of interest\n")
cat("and a new column [party] to represent all partisans thusly defined.\n\n")
A_DR <- A_quality[A_quality$pid7x == 1 | A_quality$pid7x == 2 | A_quality$pid7
x == 6 | A_quality$pid7x == 7,]
party <- c("D", "D", "3", "4", "5", "R", "R")
A_DR['party'] <- party[A_DR$pid7x]
cat("... done.\n\n")

cat("Sanity check - do the numbers add up? (1+2 = D, 6+7 = R)")
table(A_DR['pid7x'])
table(A_DR['party'])
cat('yes')
```

```
Let's now subset a dataframe containing only voters of interest
and a new column [party] to represent all partisans thusly defined.

... done.

Sanity check - do the numbers add up? (1+2 = D, 6+7 = R)

  1   2   6   7
456 192 158 319

  D   R
648 477

yes
```

In [11]:
```
cat("Age is 2019 - birth year.\n")
A_DR['age'] <- 2018 - A_DR['birthyr']
cat("... done.\n")

cat("Let's also do a quick range check to see if we have some unexpected value
s:", range(A_DR['age']),"\n")
cat("Looks fine.")

hist(A_DR$age, main = "Histogram: age breakdown", breaks = 20, cex.lab=.8, ce
x.axis=.8, cex.main=.8, cex.sub=0.8)
```

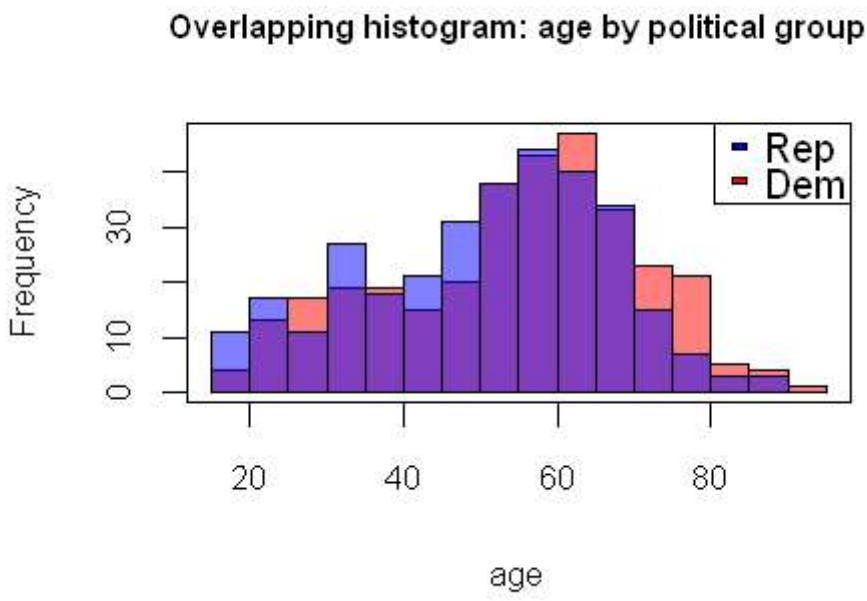
```
Age is 2019 - birth year.
... done.
Let's also do a quick range check to see if we have some unexpected values: 1
8 91
Looks fine.
```



In [12]:
```
cat("Now let's split age by political affiliation to give us a better overview
of our subset.\n\n")

h1 <- A_DR[party=="D",]
h2 <- A_DR[party=="R",]

cat("Note: In the histogram below, purple represents an overlap between Republ
icans and Democrats.")
hist(h1$age, main = "Overlapping histogram: age by political group", breaks =
20, cex.lab=.8, cex.axis=.8, cex.main=.8, cex.sub=0.8, xlab = "age", col=rgb(1
,0,0,0.5))
hist(h2$age, col=rgb(0,0,1,0.5), add=T)
box()
par(cex = 1) #set legend font to 0.7
legend("topright", c("Rep", "Dem"), fill=c("blue", "red"), bty="o", text.width
=10, y.intersp=3)
```

Now let's split age by political affiliation to give us a better overview of our subset.

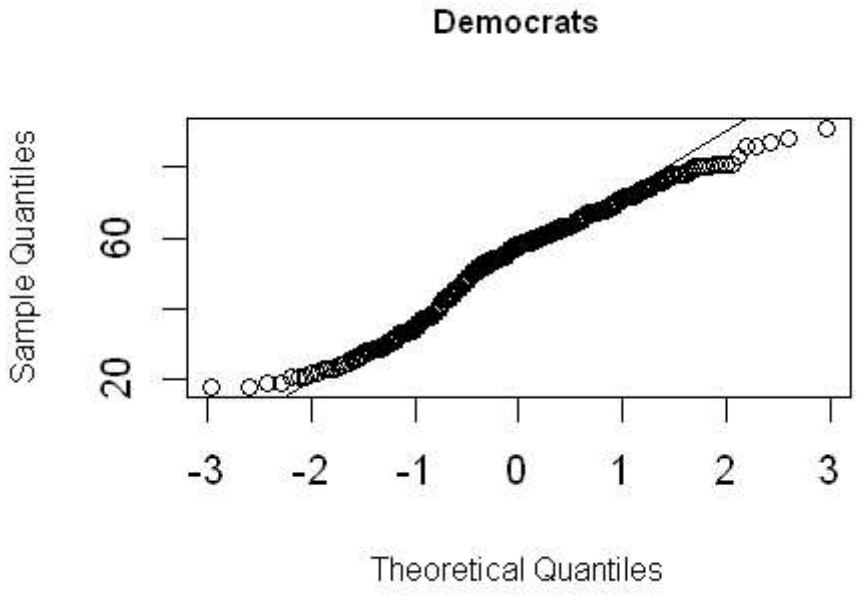
Note: In the histogram below, purple represents an overlap between Republicans and Democrats.



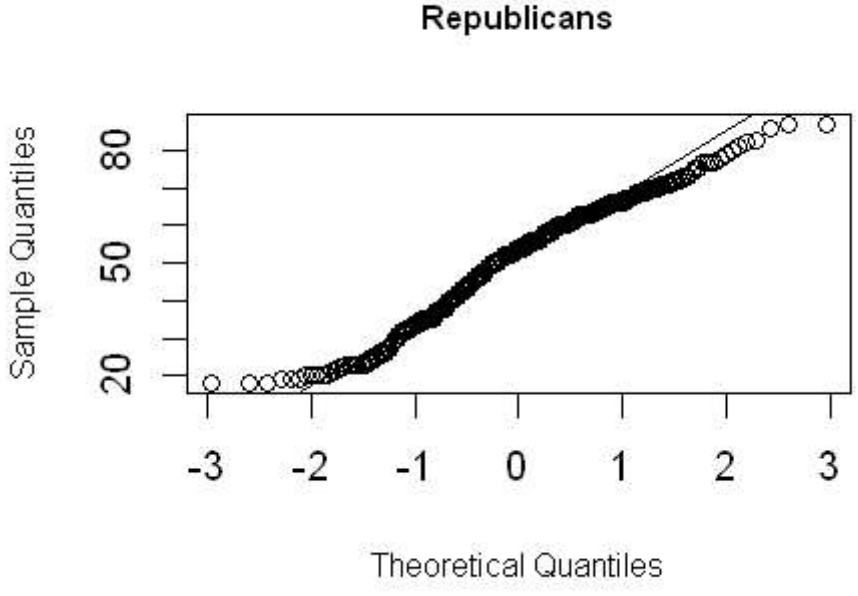
In [13]:
```
cat("We now examine distribution normality using Shapiro-Wilk tests and normal
probability plots.")
qqnorm(h1$age, main = "Democrats", cex.lab=.8, cex.main=.8)
qqline(h1$age)
shapiro.test(h1$age)

qqnorm(h2$age, main = "Republicans", cex.lab=.8, cex.main=.8)
qqline(h2$age)
shapiro.test(h2$age)
```

We now examine distribution normality using Shapiro-Wilk tests and normal pro
bability plots.

```
        Shapiro-Wilk normality test

data:  h1$age
W = 0.97078, p-value = 4.2e-06
```




```
        Shapiro-Wilk normality test

data:  h2$age
W = 0.97295, p-value = 1.026e-05
```



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.

Response

We have two independent samples of ratio data with roughly similar sizes and approximately normal distributions here. An independent samples t-test is therefore the most appropriate test to run if its assumptions can be supported.

- Dependent variable is approximately normal within each group.
- Homogeneity of variances.

Though the Shaprio-Wilk test results are significant, the test is well known to be overly sensitive. Eyeballing the normal probability plots and histograms, we see that the data are sufficiently normal for a parametric t-test especially when considering our large sample sizes.

Unlike the student's t, the Welch's t accounts for differences in group variances and is the better option in many cases (except when sample sizes are very small). We have very large samples in this case and will **proceed with the Welch's t**.

The research question is a two-sided one ("younger or older"), and so a two-tailed test is called for. There is no reason to set alpha at a value other than .05.

Our hypotheses are:

- Null Hypothesis: H_0: $\mu_{Democrats} = \mu_{Republicans}$
- Alternative Hypothesis: H_a: $\mu_{Democrats} \neq \mu_{Republicans}$
  where μ_X is mean age of X

For a measure of practical significance, **we will use Cohen's d**.

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 [89]:
```
t.test(h1$age, h2$age)
cohen.d(h1$age, h2$age)
```

```
        Welch Two Sample t-test

data:  h1$age and h2$age
t = 2.7935, df = 639.68, p-value = 0.00537
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.053827 6.041204
sample estimates:
mean of x mean of y
 54.89752  51.35000

Cohen's d

d estimate: 0.2204848 (small)
95 percent confidence interval:
    lower     upper
0.0650135 0.3759561
```

Results

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

Statistical significance

A Welch's test on voter age by political affiliation, t(639.68) = 2.7935, p = .00537, shows a difference between Democrats and Republicans significant at the p < .01 level. We therefore reject the null hypothesis that $\mu_{Democrats} = \mu_{Republicans}$ and conclude that Democrats (mean age 54.90) are older than Republicans (mean age 51.35) on the aggregate.

Practical significance

The Cohen's d value of 0.22 represents a small practical difference between the groups.

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.

Response

Keying in on the term 'baseless', for the investigation to have basis (merit), at least one of the below statements must be true:

- A: 'The Russians at least interfered with the election.'

  or

- B: 'The Trump campaign colluded w/ the Russians.'

Statement A above is an analogue to the [russia16] question.

- [russia16] 'Do you think the Russian government probably interfered in the 2016 presidential election to try to help Donald Trump win, or do you think this probably did not happen?'

Statement B above is an analogue to the [coord16] question.

- [coord16] 'Do you think Donald Trump's 2016 campaign probably coordinated with the Russians, or do you think his campaign probably did not do this?'

Interestingly the [muellerinv] question more directly is inquiring about the investigation than the above survey questions, yet the crux of the question appears to be regarding the administration of the investigation as opposed to a belief regarding the underlying truth. Whether or not the investigation is administered well has no bearing on the validity of the investigation's basis. Therefore, [muellerinv] will not be considered by our study.

- [muellerinv] 'Do you approve, disapprove, or neither approve nor disapprove of Robert Mueller's investigation of Russian interference in the 2016 election?'

Therefore,

> If and only if a respondant replied negatively to both [russia16] and [coord16] will they be counted as believing that the investigation is baseless.

Regarding the identification of respondants as 'Independent' voters, the survey has a profile element [pid7x] for which any of the below responses are to be regarded as 'Independent':

- [p1d7x] = 'Ind, closer to Dem' (3) or 'Independent' (4) or 'Ind, closer to Rep' (5)

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 [90]:
```
#Modify A_quality dataset for filtering down to Independents only
A_quality_russia <- A_quality[A_quality$pid7x == 3 | A_quality$pid7x == 4 | A_
quality$pid7x == 5,]

#Modify A_quality_russia dataset to eliminate non-responses.
A_quality_russia_respon <- subset(A_quality_russia, A_quality_russia$caseid !=
1683)

cat('The code executed creates 2 datasets, "A_quality_russia" which filters to
only Independents and "A_quality_russia_respon" which filters to only those of
"A_quality_russia" who have responded')
```

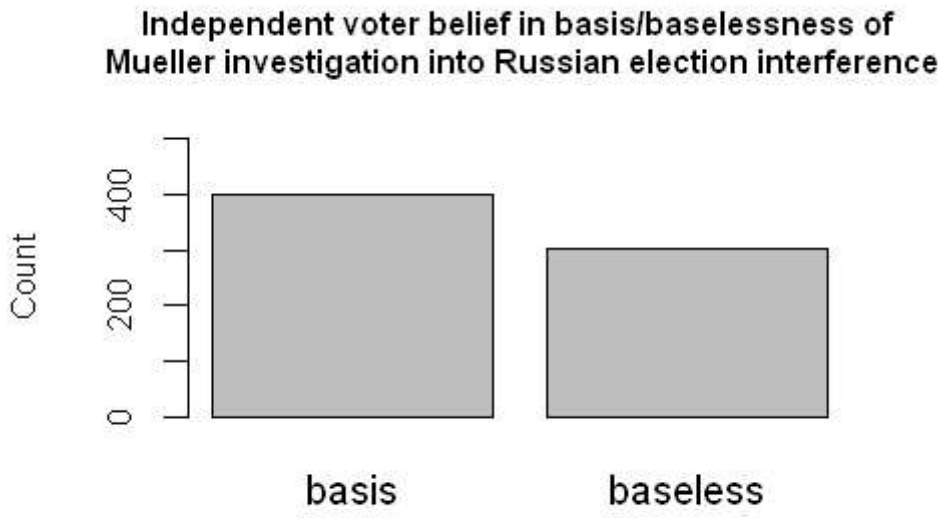
```
The code executed creates 2 datasets, "A_quality_russia" which filters to onl
y Independents and "A_quality_russia_respon" which filters to only those of
"A_quality_russia" who have responded
```

In [91]:
```
#[russia16]
cat('EDA for [russia16] using A_quality_russia: "Do you think the Russian gove
rnment probably\ninterfered in the 2016 presidential election to try to help D
onald Trump win, \nor do you think this probably did not happen?"')
cat("\n\nRussia probably interfered (1): ",length(A_quality_russia$russia16[A_
quality_russia$russia16 == 1]),"\nThis probably did not happen (2): ", length(
A_quality_russia$russia16[A_quality_russia$russia16 == 2]), "\nTotal Response
s: ", length(A_quality_russia$russia16[A_quality_russia$russia16 == 1]) + leng
th(A_quality_russia$russia16[A_quality_russia$russia16 == 2]),"\nNon-Responses
(-7): ", length(A_quality_russia$russia16[A_quality_russia$russia16 == -7]))
cat('\ncaseid\'s for non-respondant: ', (A_quality_russia$caseid[A_quality_rus
sia$russia16 == -7]))

strrep('-',50)
#[coord16]
cat('EDA for [coord16] using A_quality_russia: "Do you think Donald Trump's 20
16 campaign probably\n coordinated with the Russians, or do you think his camp
aign probably did not do this?"')
cat("\n\nProbably coordinated with the Russians (1): ",length(A_quality_russia
$coord16[A_quality_russia$coord16 == 1]),"\nProbably did not (2): ", length(A_
quality_russia$coord16[A_quality_russia$coord16 == 2]), "\nTotal Responses: ",
length(A_quality_russia$coord16[A_quality_russia$coord16 == 1]) + length(A_qua
lity_russia$coord16[A_quality_russia$coord16 == 2]),"\nNon-Responses (-7): ",
length(A_quality_russia$coord16[A_quality_russia$coord16 == -7]))
cat('\ncaseid\'s for non-respondant: ', (A_quality_russia$caseid[A_quality_rus
sia$coord16 == -7]))

strrep('-',50)
#[russia16] & [coord16]
cat('EDA for synthesis of [russia16] & [coord16] using A_quality_russia_respo
n')
count_basis <- (length(A_quality_russia_respon$coord16[(A_quality_russia$coord
16 == 1 | A_quality_russia$russia16 == 1)]))
cat('\n\nMueller investigation has a basis ([russia16] == 1 | [coord16 ==1]):
 ',count_basis)
count_baseless <- length(A_quality_russia_respon$caseid) - length(A_quality_ru
ssia$coord16[A_quality_russia$coord16 == 1 | A_quality_russia$russia16 == 1])
cat('\nMueller investigation is baseless (Complement of basis): ', count_basel
ess)
```

```
EDA for [russia16] using A_quality_russia: "Do you think the Russian governme
nt probably
interfered in the 2016 presidential election to try to help Donald Trump win,
or do you think this probably did not happen?"

Russia probably interfered (1):  381
This probably did not happen (2):  320
Total Responses:  701
Non-Responses (-7):  1
caseid's for non-respondant:  1683



'-----------------------------------------------'

EDA for [coord16] using A_quality_russia: "Do you think Donald Trump’s 2016 c
ampaign probably
 coordinated with the Russians, or do you think his campaign probably did not
do this?"

Probably coordinated with the Russians (1):  354
Probably did not (2):  348
Total Responses:  702
Non-Responses (-7):  0
caseid's for non-respondant:

'-----------------------------------------------'

EDA for synthesis of [russia16] & [coord16] using A_quality_russia_respon

Mueller investigation has a basis ([russia16] == 1 | [coord16 ==1]):  400
Mueller investigation is baseless (Complement of basis):  301
```

In [92]:
```
#Creating a vector for respondant belief in basis/baselessness of Mueller inve
stigation
A_quality_russia_respon$baseless <- ifelse((A_quality_russia_respon$coord16 ==
1 | A_quality_russia_respon$russia16 == 1), 0, 1)

#barplot for basis vs baseless counts
barplot(c(count_basis, count_baseless), ylim = c(0,500), ylab = 'Count', main
= 'Independent voter belief in basis/baselessness of\n Mueller investigation i
nto Russian election interference',names.arg = c('basis','baseless'), cex.lab=
.8, cex.axis=.8, cex.main=.8, cex.sub=0.8)
axis(2, at = seq(0,500, by = 100),labels=FALSE)

cat('The code executed creates a vector for respondant belief in basis/baseles
sness of\nthe Mueller investigation. In this vector, "Mueller investigation ha
s a basis"\nis represented by "1" and "Mueller investigation is baseless"\nis
 represented by "0"')
```

```
The code executed creates a vector for respondant belief in basis/baselessnes
s of
the Mueller investigation. In this vector, "Mueller investigation has a basi
s"
is represented by "1" and "Mueller investigation is baseless"
is represented by "0"
```



Note that the data structures created create ease of reproducability of this study on same basis or others, such as 'Republicans' rather than 'Independants'.

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.

Response

To determine statistically whether the majority opinion of survey response for Independent voters is in favor of either the statement that the Mueller investigation has a basis, or that the Mueller investigation is baseless, we will run a binomial test. For this data set, the binomial test satisfies the requirement that the outcome options are dichotomous (ie. two), mutually exclusive, mutually exhaustive, and of type Nominal. Furthermore, the requirement that each test is independant from other tests is satisfied.

For the binomial test, the following are the hypotheses:

- Null Hypothesis: H_0: prob. of success = 0.5, representing no majority opinion among Independent voters that 'Mueller investigation has a basis' or 'Mueller investigation is baseless'
- Alternative Hypothesis: H_a: prob. of success $\neq$ 0.5, representing a majority opinion among Independent voters that 'Mueller investigation has a basis' or 'Mueller investigation is baseless'

For evaluating between 2 choices, we could consider applying the Central Limit Theorem and a 2-sided paired t-test to determine probability of a single choice. However, the data is Nominal, violating the statistical requirement of t-test (Cardinal).

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 [93]:
```
binom.test(sum(A_quality_russia_respon$baseless), length(A_quality_russia_resp
on$baseless), 0.5)
```

```
        Exact binomial test

data:  sum(A_quality_russia_respon$baseless) and length(A_quality_russia_resp
on$baseless)
number of successes = 301, number of trials = 701, p-value = 0.0002097
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.3923947 0.4669752
sample estimates:
probability of success
             0.4293866
```

Results

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

Statistical significance

With a statistical significance observed p = 0.0002, compared to a p_{crit} of 0.05, the null hypothesis that there is no majority opinion among Independant voters for either 'Mueller investigation has a basis' or 'Mueller investigation is baseless' is rejected. The null hypothesis is rejected in the direction that the majority opinion among Independent voters is that the Mueller investigation has a basis.

Practical significance

This conclusion has the practical significance of ~ $\frac{1}{3}^{rd}$ more survey findings mapping to 'Mueller investigation has a basis'(400) from Independent voters than mapping to 'Mueller investigation is baseless'(301).

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.

Response

To answer this question, we consider:

- How angry/fearful are voters
- How to define increase in voter turnout

The questionnaire contains the following anger- and fear-related items:

- geangry ([geafraid]) How do you feel about the way things are going in the country: angry(afraid)
- dtangry ([dtafraid]) Think about Donald Trump. How often would you say you've felt each of the following ways because of the kind of person Donald Trump is or because of something he has done?: angry(afraid)
- imangry ([imafraid]) Think about immigrants coming from other countries to live in the United States. How often would you say you've felt each of the following ways because of immigrants coming from other countries to live in the United States? angry(afraid).

The im and dt items were each only delivered to half of the sample. In addition, these are targeted at a non-exhaustive issue set (immigrants and Trump).

The ge- items elicted emotional feedback on a general level and are more suitable as measures of fear and anger. Also, ge- items applied to all respondants. ge- items were coded:

- -7 No Answer
- 1 Not at all
- 2 A little
- 3 Somewhat
- 4 Very
- 5 Extremely

To decide if someone is more angry or more fearful, we subtract geafraid from geangry (taking care to remove rows with -7 first). The components and resultant are Ordinal.

[geangry] - [geafraid]

- -4 to -1 more afraid
- 0 neither more afraid nor more angry
- 1 to 4 more angry

It is necessary at this stage to convert these Ordinal outcomes into a binary in order to map our data to the binary form of the research question. We recode this value into a new variable 'Angrier', with "more angry" voters being given a value 1 and "more afraid" voters being given a value 0; those who are "neither more afraid nor more angry" are removed from analysis per the research question.

For voter turnout, we have the following items to work with:

- [turnout16] In 2016 (presidential election)... did you definitely vote, definitely not vote, or are you not completely sure whether you voted?
- [turnout18] In (2018 presidential election) ... 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?

[turnout16] is coded thus:

- 1 Definitely voted
- 2 Definitely did not vote
- 3 Not completely sure

[turnout18] is coded thus:

- 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

As we are studying increase (not change) in voter turnout, we are interested in the voters who did not vote in 2016 but did in 2018. In both groups, respondents who were "not completely sure" regarding voting were removed from analysis. For [turnout18], responses 1,2,3 will be aggregated to generate a single "definitely voted" answer.

Following this, we create a new [VoteUp] binary variable, with a 1 for individuals who did not vote in 2016 but did in 2018, and a 0 for everyone else.

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 [15]:
```
cat("First, let's remove respondents who did not answer either ge item or were
not completely sure about voting either year.\n")

A_Vinc <-  subset(A_quality, geangry > 0 & geangry < 6 & geafraid > 0 & geafra
id < 6 & turnout16 > 0 & turnout16 < 3 & turnout18 > 0 & turnout18 < 5)
cat("We have", nrow(A_Vinc), "of", nrow(A_quality), "rows left.")
```

```
First, let's remove respondents who did not answer either ge item or were not
completely sure about voting either year.
We have 1807 of 1851 rows left.
```

In [16]:
```
cat("Next, we calculate the anger - fear differences and remove those who scor
e the same.\n")

A_Vinc["Ang_Afr"] <- A_Vinc$geangry - A_Vinc$geafraid
A_Vinc <-  subset(A_Vinc, Ang_Afr != 0)
cat("We have", nrow(A_Vinc), "of", nrow(A_quality), "rows left.")
```

```
Next, we calculate the anger - fear differences and remove those who score th
e same.
We have 1012 of 1851 rows left.
```

In [17]:
```
cat("And then we recode the ordinal anger - fear difference into a binary valu
e.\n")
A_Vinc$Angrier <- ifelse((A_Vinc$Ang_Afr > 0), 1, 0)

cat("We do the same for turnout increase.\n")
A_Vinc$VoteUp <- ifelse((A_Vinc$turnout16 == 2 & A_Vinc$turnout18 > 0 &  A_Vin
c$turnout18 < 4), 1, 0)
```

```
And then we recode the ordinal anger - fear difference into a binary value.
We do the same for turnout increase.
```

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.

Response

Since we have a two-way table of frequencies, **we run McNemar's test**. This is an exact chi-square test for paired binomial data that tells us if row and column marginal frequencies are equal.

The assumptions of the test are all met:

- You must have one Nominal variable with two categories and one independent variable with two connected groups.
- The two groups in the dependent variable must be mutually exclusive.
- Sample must be a random sample.

The null hypothesis of marginal homogeneity states that the two marginal probabilities for each outcome are the same.

Therefore, our hypotheses are:

- Null Hypothesis: H_0: the anger-fear difference has no effect on voting turnout increase
- Alternative Hypothesis: H_a: the anger-fear difference has an effect on voting turnout increase

There is no universally agreed-upon effect size measure associated with McNemar's test. Instead, we will describe the exact number of cases different between the groups.

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]:
```
cat("How do our 2 new categorical variables interact? Let's create a contingen
cy table.\n")
cat("[Angrier]: 0 == more fearful, 1 == more angry\n")
cat("[VoteUp]: 0 == no voting behaviour increase, 1 == did not vote in 2016 bu
t did in 2018")
AV_table <- table(A_Vinc$Angrier, A_Vinc$VoteUp, dnn=c("Angrier","VoteUp"))
AV_table
```

```
How do our 2 new categorical variables interact? Let's create a contingency t
able.
[Angrier]: 0 == more fearful, 1 == more angry
[VoteUp]: 0 == no voting behaviour increase, 1 == did not vote in 2016 but di
d in 2018
```

Angrier	VoteUp 0	1
0	342	14
1	641	15

In [19]:
```
cat("The voting increase values of 14 and 15 in the table above are small but
 unsurprising:\n")
cat("voter turnout across the country was notably smaller in 2018 than in 201
6.\n\n")

cat("Just to be sure that we don't have a recoding problem here though, let us
briefly examine the raw turnout data.\n")
cat("[2016]: 1 == voted, 2 == did not vote\n")
cat("[2018]: 1, 2, 3 == voted, 4 == did not vote")
table(A_Vinc$turnout16, A_Vinc$turnout18, dnn=c("2016","2018"))

cat("As we see here, most respondents voted in 2016.\n")
cat("Also, the marginal values here accord with what we see in the earlier Ang
rier by VoteUp table.")
```

```
The voting increase values of 14 and 15 in the table above are small but unsu
rprising:
voter turnout across the country was notably smaller in 2018 than in 2016.

Just to be sure that we don't have a recoding problem here though, let us bri
efly examine the raw turnout data.
[2016]: 1 == voted, 2 == did not vote
[2018]: 1, 2, 3 == voted, 4 == did not vote

    2018
2016   1   2   3   4
   1 411 144 255  31
   2  16   6   7 142

As we see here, most respondents voted in 2016.
Also, the marginal values here accord with what we see in the earlier Angrier
by VoteUp table.
```

In [20]:
```
cat("As the values in the secondary diagonal of the contingency table add up t
o a large number (>25),\n")
cat("we do not need to run the exact variant of McNemar's and can use the regu
lar version.")
mcnemar.test(AV_table)
```

```
As the values in the secondary diagonal of the contingency table add up to a
large number (>25),
we do not need to run the exact variant of McNemar's and can use the regular
version.

        McNemar's Chi-squared test with continuity correction

data:  AV_table
McNemar's chi-squared = 598.28, df = 1, p-value < 2.2e-16
```

In [21]:
```
cat("Let us now derive the proportions of turnout increase within each emotion
group.\n")
cat("More Angry:", 15/(641+15),"\n")
cat("More Afraid:", 14/(342+14))
```

```
Let us now derive the proportions of turnout increase within each emotion gro
up.
More Angry: 0.02286585
More Afraid: 0.03932584
```

Results

Question 4: Was anger or fear more effective at driving increases in voter turnout from 2016 to 2018?
Answer: As determined through statistical testing, fear was more effective at driving increases in voter turnout from 2016 to 2018.

Statistical significance A McNemar's chi-square test was performed on salient emotion (fear vs. anger) and increase in voting turnout from 2016 to 2018 (yes/no). A total of 1,012 voters were assessed and the test result ($p < 2.23e-16$, which is smaller than the standard alpha value of 0.05) suggests that emotion and turnout increase are related.

Practical significance Though the increased voting numbers are similar for the two emotional groups (More Angry= 15, More Afraid = 14), the groups are very different in size (More Angry = 641, More Afraid = 342). In terms of group proportions, the More Angry group saw a 2.29% increase in voter turnout while the More Afraid group saw a 3.93% increase.

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.

Research Background:

We chose the comprehend if education level affects political affiliation.

There are conflicting, commonly held beliefs, regarding the education levels of Democrats and Republicans.

- People in manufacturing (especially, union members) are traditionally of low educational level and considered primarily Democrat.
- President Trump (Republican) received support in the Rust Belt that has significant manufacturing jobs.
- Higher income individuals tend to vote liberal, according to Andrew Gelman, author of "Economic Divisions and Political Polarization in Red and Blue America". It stands to reason that income is a positive function of education.

In summary, there is no clarity on which party's members are more educated.

Research Question:

"Are Republicans more or less educated than Democratics?"

We will use available data on education level of Republicans/Democrats. We will then test for statistical evidence of one party being more educated than the other.

Use of Data Analysis:

The campaign strategists of both parties can use such data analysis to determine focus demographics to maximize the return on investment.

Operationalization:

For operationalization, we investigated different options to determine the political party of a respondant. For example, we considered using actual votes. We settled on "Party ID" as the political affiliation of the respondant.

The following questionaire variables were relevant to the research question.

[educ]: Education level of respondant.

- 1 No High School
- 2 HS Graduate
- 3 Some College
- 4 2-year
- 5 4-year
- 6 post-grad

[pid7x]: Political party of respondant.

- 1 Strong Democrat
- 2 Not very strong Democrat
- 3 Independent closer to Democrat
- 4 Independent
- 5 Independent closer to Republican
- 6 Not very strong Republican
- 7 Very strong Republican

We classified respondants as follows:

- Democrat [pid7x] values 1,2
- Republican [pid7x] values 6,7

Next, a vector containing Democratic voter education level was created by considering the education levels of participants who had pid7x values 1,2. Similarly, a Republican vector was created using pid7x values 6,7.

The variables [educ] and [pid7x] are relevant and adequate for our research question. We did not find gaps between the information needed for the research question and these variables.

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.

Exploratory Data Analysis (EDA):

We performed EDA by noting education levels of Democrat and Republican respondants. 'No response' for either [educ] or [pid7x] are skipped.

For sanity check, we observed histograms of the education levels of all affiliations. The histograms do not follow a Normal distribution, but exhibit similar trends.

We also compared median (5 for Democrats and 3 for Republicans) of [educ].

In [23]:
```
#Renaming A_quality to A_explr
A_explr <- A_quality

# All voters with responses
A_explr_AllVoters <- A_explr[A_explr$pid7x != -7,]

#Modify dataset for filtering down to Democrats and Republicans only
A_explr_Democrats <- A_explr[A_explr$pid7x == 1 | A_explr$pid7x == 2,]
A_explr_Republicans <- A_explr[A_explr$pid7x == 6 | A_explr$pid7x == 7,]
```

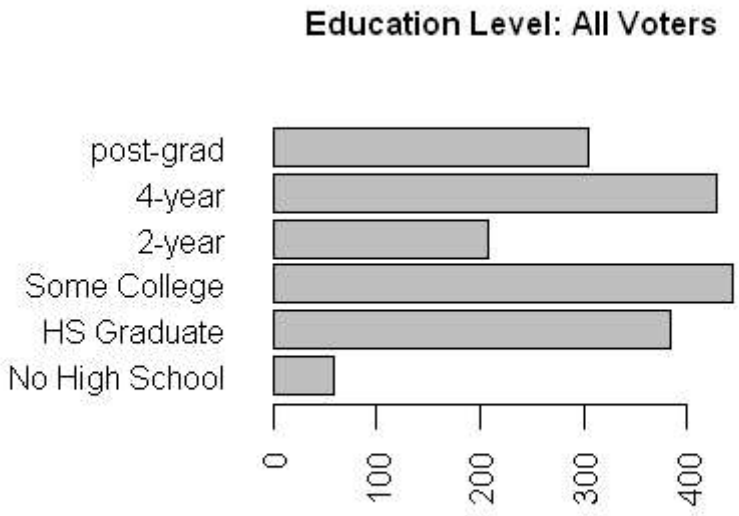
In [24]:
```
educOfAllVoters <- A_explr_AllVoters$educ
length_educOfAllVoters <- length(educOfAllVoters)
paste('Number of Samples of Education Level Of All Voters:', length_educOfAllV
oters)
#hist(educOfAllVoters, xlab="Education Level of All Voters", breaks=20)
#legend("topright", c("1: No High School", "2: HS Graduate", "3: Some Colleg
e", "4: 2-year", "5: 4-year", "6: post-grad"),xpd=TRUE, cex=0.6, bty='n')

educOfDemocratVoters <- A_explr_Democrats$educ
length_educOfDemocratVoters <- length(educOfDemocratVoters)
paste('Number of Samples of Education Level Of Democratic Voters:', length_edu
cOfDemocratVoters)
#hist(educOfDemocratVoters, xlab="Education Level of Democratic Voters", break
s=20)
#legend("topright", c("1: No High School", "2: HS Graduate", "3: Some Colleg
e", "4: 2-year", "5: 4-year", "6: post-grad"),xpd=TRUE, cex=0.6, bty='n')

educOfRepublicanVoters <- A_explr_Republicans$educ
length_educOfRepublicanVoters <- length(educOfRepublicanVoters)
paste('Number of Samples of Education Level Of Republican Voters:', length_edu
cOfRepublicanVoters)
#hist(educOfRepublicanVoters, xlab="Education Level of Republican Voters", bre
aks=20)
#legend("topright", c("1: No High School", "2: HS Graduate", "3: Some Colleg
e", "4: 2-year", "5: 4-year", "6: post-grad"),xpd=TRUE, cex=0.6, bty='n')
```

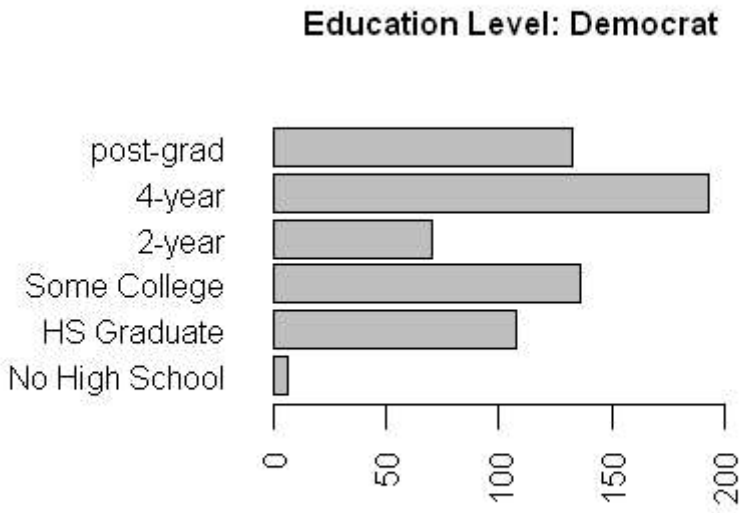
'Number of Samples of Education Level Of All Voters: 1827'

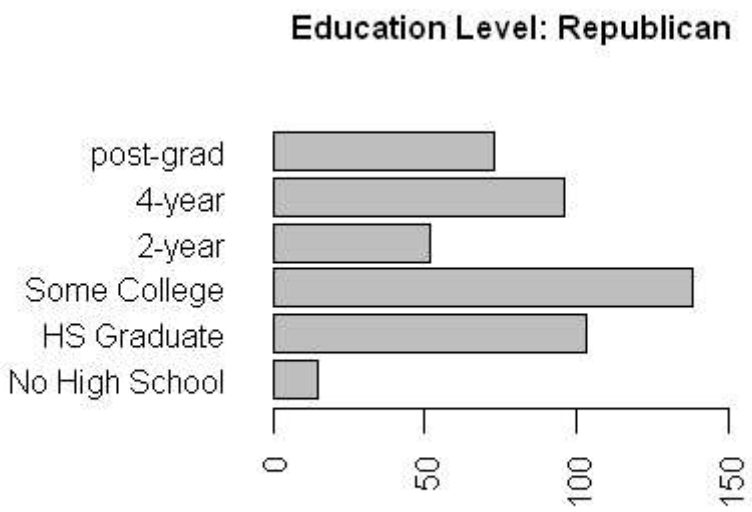
'Number of Samples of Education Level Of Democratic Voters: 648'

'Number of Samples of Education Level Of Republican Voters: 477'

In [26]:
```
# Quick look at Education Level
counts <- table(educOfAllVoters)
par(las=2) # make label text perpendicular to axis
par(mar=c(5,8,4,2)) # increase y-axis margin.
barlabels <- c("No High School", "HS Graduate", "Some College", "2-year", "4-y
ear", "post-grad")
barplot(counts,main ='Education Level: All Voters',horiz=TRUE,names.arg=barlab
els,cex.lab=.8, cex.axis=.8, cex.main=.8, cex.sub=0.8, cex.names=.8, xlim = c(
0, max(counts)+20), xpd = FALSE)

# Quick look at Education Level
counts <- table(educOfDemocratVoters)
par(las=2) # make label text perpendicular to axis
par(mar=c(5,8,4,2)) # increase y-axis margin.
barlabels <- c("No High School", "HS Graduate", "Some College", "2-year", "4-y
ear", "post-grad")
barplot(counts,main ='Education Level: Democrat',horiz=TRUE,names.arg=barlabel
s,cex.lab=.8, cex.axis=.8, cex.main=.8, cex.sub=0.8, cex.names=.8, xlim = c(0,
max(counts)+20), xpd = FALSE)

# Quick look at Education Level
counts <- table(educOfRepublicanVoters)
par(las=2) # make label text perpendicular to axis
par(mar=c(5,8,4,2)) # increase y-axis margin.
barlabels <- c("No High School", "HS Graduate", "Some College", "2-year", "4-y
ear", "post-grad")
barplot(counts,main ='Education Level: Republican',horiz=TRUE,names.arg=barlab
els,cex.lab=.8, cex.axis=.8, cex.main=.8, cex.sub=0.8, cex.names=.8, xlim = c(
0, max(counts)+20), xpd = FALSE)
```


Education Level: All Voters
post-grad
4-year
2-year
Some College
HS Graduate
No High School
0
100
200
300
400


Education Level: Democrat
post-grad
4-year
2-year
Some College
HS Graduate
No High School
0
50
100
150
200



In [27]:
```
medianEducAllVoters <- median(educOfAllVoters)
varEducAllVoters <- var(educOfAllVoters)

paste('median of educOfAllVoters:', medianEducAllVoters)
paste('Variance of educOfAllVoters:', varEducAllVoters)

medianEducDemocratVoters <- median(educOfDemocratVoters)
varEducDemocratVoters <- var(educOfDemocratVoters)

paste('median of educOfDemocratVoters:', medianEducDemocratVoters)
paste('Variance of educOfDemocratVoters:', varEducDemocratVoters)

medianEducRepublicanVoters <- median(educOfRepublicanVoters)
varEducRepublicanVoters <- var(educOfRepublicanVoters)

paste('median of educOfRepublicanVoters:', medianEducRepublicanVoters)
paste('Variance of educOfRepublicanVoters:', varEducRepublicanVoters)
```

'median of educOfAllVoters: 4'

'Variance of educOfAllVoters: 2.23710306219654'

'median of educOfDemocratVoters: 5'

'Variance of educOfDemocratVoters: 2.07818612017479'

'median of educOfRepublicanVoters: 3'

'Variance of educOfRepublicanVoters: 2.15902964959569'

Suitability of the Test:

The research question compares education levels for Democrats/Republicans. Hence, median and distribution of the education level are candidate criteria for comparison. We zeroed in on two test candidates- **the unpaired two-sample t-test** and **Wilcoxon Rank Sum test ('Wilcoxon')**.

We have two independent groups with differing sample sizes. We are interested in determining significance of the median education level difference between Democrat/Republican. Education level is Ordinal. This disallows any t-test, as t-test requires Cardinal data.

The Wilcoxon test compares distribution of Ordinal measurements (e.g., the education level in our case) for two different populations (e.g., Democrat/Republican). The Wilcoxon test aims to detect rank order shifts for the two different populations.

We chose Wilcoxon Rank Sum test because of the Ordinal measurements.
Hypothesis: Null hypothesis for Wilcoxon:

True location = 0

Interpretation for our case: Education level is the same for Democrat/Republican
Alternative hypothesis for Wilcoxon:

True location shift $\neq 0$

Interpretation for our case: Education level is different for Democrat/Republican

Conduct your test. (2 points)

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

In [106]:
```
wilcox.test(educOfDemocratVoters, educOfRepublicanVoters)
```

```
        Wilcoxon rank sum test with continuity correction

data:  educOfDemocratVoters and educOfRepublicanVoters
W = 179980, p-value = 1.339e-06
alternative hypothesis: true location shift is not equal to 0
```

Statistical Significance:

The Wilcoxcon test result shows the p-value of 0.01, less than p-crit of 0.05. Hence, we reject the null hypotheis that education levels are similar for Democrat/Republican. The alternative hypothesis states that the true location shift $\neq$ zero.

Practical Significance:

Since the distributions of the education level are not similar and since the median education level for Democrat (5, 4-year) is greater than the median education level for Republican (3, Some college), we assert Democrats are more educated than Republicans (within population of ANES survey participants).

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.

Conclusion:

According to the null hypotheis for Wilcoxon test, the distributions of the education levels are the same for Republicans/Democrats. Our research question asks if Democrat or Republican are more educated. We used Wilcoxon test to analyze distributions of education levels. The Wilcoxon test yielded the p-value smaller than the significance level corresponding to the 95% Confidence Interval. Hence, we reject the null hypothesis that the distributions of the education levels are the same for Republican/Democrat. Indeed, the median Republican voter education level is smaller than that of the Democrat. Hence, **Democrats are found to be more educated than Republican voters** in the population of ANES survey participants. We cannot extraploate these results to the general U.S. population without leveraging study weights to adjust the survey population to the U.S. population.