Using R for statistical analysis

Code

Importing data and libraries

Hide

```
library(readr)
poll84 <- read_csv("E:/R/poll84.csv")
library(tidyverse)
library(flextable)
library(janitor)
library(car)
```

Part a

Hide

```
colnames(poll84)[which(names(poll84) == "female")] <- "Sex"
poll84$Sex <- factor(poll84$Sex,
                     levels=c(0,1),
                     labels=c("Male","Female"))
myTab <- poll84 %>% group_by(Sex) %>% summarise(N=n(),Min=min(age),Max=max(age),Mean=mean(age),S
D=sd(age),
                     Q1=quantile(age,.25),
                     Median=median(age),
                     Q3=quantile(age,.75), IQR=Q3-Q1)
flexv1 <- mytab1 %>% flextable
flexv1 <- set_caption(flexv1, caption = "Summary Statistics of the AGE for each SEX")
print(flexv1)
```

Summary Statistics of the AGE for each SEX

Sex	n	Mean	sd	Q1	Median	Q3
Male	541	44.43253	15.86522	32	41	57
Female	668	44.88323	17.07576	31	41	59

Hide

```
NA
```

From the results above we can conclude that the number of female voters are more than that of male voters.The average age of both male and female voters is the same. The standard deviation of females is greater in value than that of males indicating that there is a larger dispersion of data close to the mean value as compared to males.This can be further supported by the Interquartile Range(IQR), as the male voter’s IQR value is lower, this shows that the middle values are clustered more tightly.

Part b

sub-part 1

Hide

```
poll84$vote <- factor(poll84$vote,
                      levels=c(0,1),
                      labels=c("Democrats","Republicans"))
mytab2 <- tabyl(poll84, income, vote)
flexv2 <- mytab2 %>% flextable
flexv2 <- set_caption(flexv2, caption = "Voting Intention per Income Level")
print(flexv2)
```

Voting Intention per Income Level

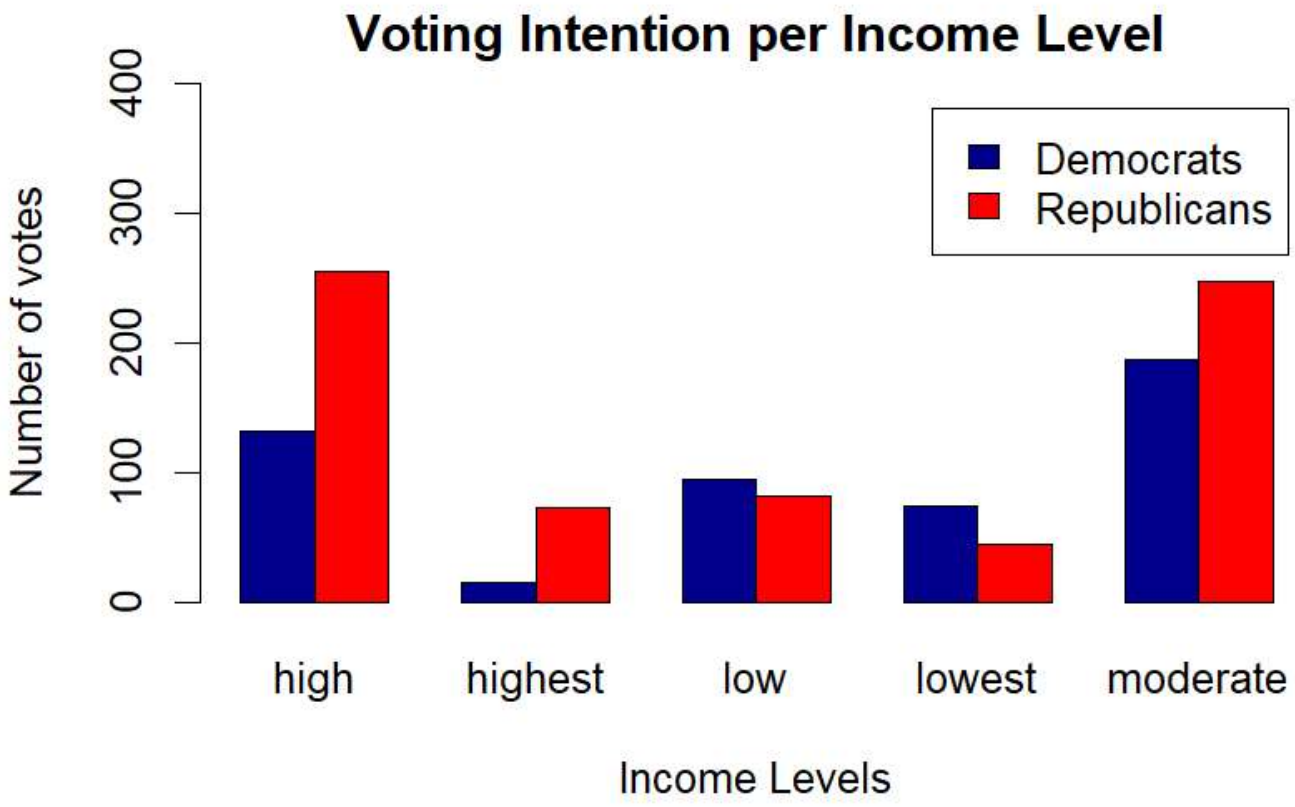
income	Democrats	Republicans
high	132	255
highest	16	74
low	95	82
lowest	75	46
moderate	187	247

From the above table it is clear that the Republicans can gain more number of votes as compared to the Democrats as per the poll. People belonging to high and highest income groups have larger number of voting intentions inclined towards the Republicans. On the other hand, people from the low and lowest income groups seem to be more in favor of the Democrats. Although of all the income groups, maximum voting intentions for the Democrats come from the moderate income group, it is still less than that for the Republicans.

sub-part 2

Hide

```
barplt <- table(poll84$vote, poll84$income)
barplot(barplt, main="Voting Intention per Income Level",
        xlab="Income Levels",ylim = c(0,400),ylab = "Number of votes",col=c("darkblue","red"),le
gend = rownames(barplt), beside=TRUE)
```



The above graph illustrates voting intentions(X-axis) for each income level(Y-axis) for the Democrats and the Republicans.As we go from the lowest to the high income group we can see a significant rise in voting intentions but there is also a sharp drop from high to higher income class in the favor of the Republicans. However, in case of the democrats, there is a fluctuation while tracing the income levels from lowest to highest. From this graph, it is evident that the high and highest income groups have voting intentions for the Republicans with a clear majority. On the flip side, voting intentions for the low and lowest income groups tend to be in the favor of the Democrats with a slight majority. In case of moderate income group, the number of votes that the Republicans can gain is almost similar as that of the high income group.Democrats can receive maximum number of votes from the moderate income group.

Part c

Hide

```
#case1:For Republicans
prop.test(x=704, n=1209, correct = F, p=0.58)
```

```
    1-sample proportions test without continuity
    correction

data:  704 out of 1209, null probability 0.58
X-squared = 0.026241, df = 1, p-value = 0.8713
alternative hypothesis: true p is not equal to 0.58
95 percent confidence interval:
 0.5542819 0.6097957
sample estimates:
        p
0.5822994
```

Hide

```
binom.test(x=704, n=1209, p=0.58)
```

```
    Exact binomial test

data:  704 and 1209
number of successes = 704, number of trials = 1209,
p-value = 0.8842
alternative hypothesis: true probability of success is not equal to 0.58
95 percent confidence interval:
 0.5539139 0.6102837
sample estimates:
probability of success
             0.5822994
```

Hide

```
#case2:For Democrats
prop.test(x=505, n=1209, correct = F, p=0.4177)
```

```
    1-sample proportions test without continuity
    correction

data:  505 out of 1209, null probability 0.4177
X-squared = 1.6663e-09, df = 1, p-value = 1
alternative hypothesis: true p is not equal to 0.4177
95 percent confidence interval:
 0.3902043 0.4457181
sample estimates:
        p
0.4177006
```

Hide

```
binom.test(x=505, n=1209, p=0.4177)
```

```
    Exact binomial test

data:  505 and 1209
number of successes = 505, number of trials = 1209,
p-value = 1
alternative hypothesis: true probability of success is not equal to 0.4177
95 percent confidence interval:
 0.3897163 0.4460861
sample estimates:
probability of success
             0.4177006
```

Confidence Intervals

In case 1, as the confidence intervals include 58%, we can conclude that the voting intention of the poll is compatible with the election result. Similarly in case 2, as the confidence intervals include 41.77% we can say that this further justifies the compatibility of the poll with the election result.

Hypothesis testing for proportion

Lets assume that we test the above at 5% significance level: **H_0**: voting intention of the poll is compatible with the election result
H_1: voting intention of the poll is not compatible with the election result

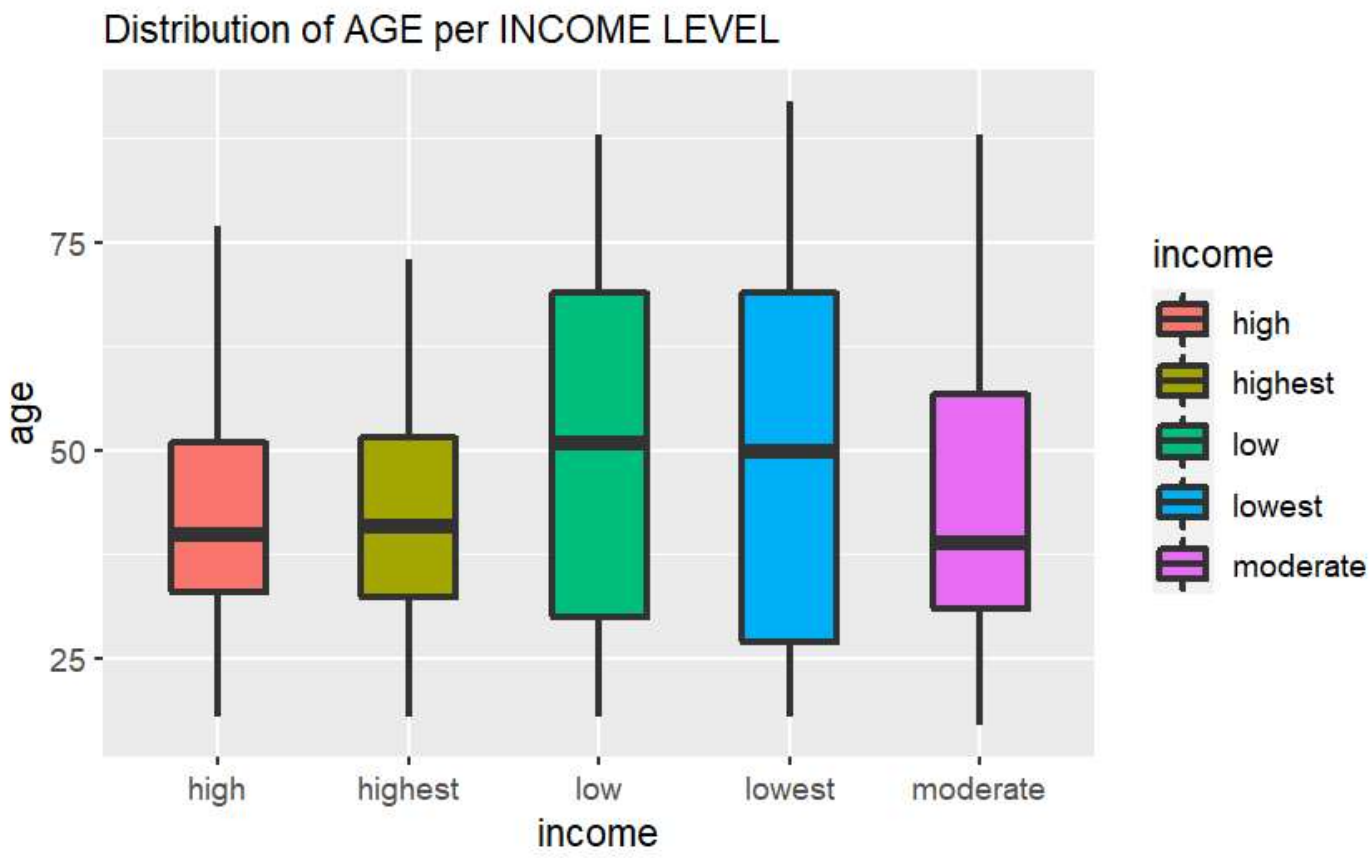
$$H_0 : \pi = 0.58 \text{ vs } H_1 : \pi \neq 0.58$$

In case 1, as the p-value = 0.8842 is > 0.05 and in case 2, as the p-value = 1 is > 0.05 there is no evidence to reject the null hypothesis at 5% significance level. Hence we conclude that voting intention of the poll is compatible with the election result.

Part d

Hide

```
poll84 %>% ggplot(aes(x=income, y=age, fill=income)) + geom_boxplot(width=0.5,lwd=1)+ labs(subti
tle="Distribution of AGE per INCOME LEVEL")
```



From the above graph it is clear that the minimum age limit is same for all the income groups. This can be because of the general minimum age criteria for voting. Whereas the maximum age varies in all income groups. The mean age for high, highest and moderate income levels is almost similar and so is the case for the low and lowest income levels with low-lowest income groups having greater mean value for age than that of the high-highest income groups. The moderate income group seems to have the lowest mean value of age. There is a significant difference between the Interquartile Range(IQR) in all the income groups as low-lowest income groups have larger IQR as compared to high-highest income groups.

Part e

In order to do the ANOVA test, first we test the equality of the variances

Hypothesis

Lets assume that we test the below at 5% significance level: $\mathbf{H_0}$: Variance is equal in all income levels $\mathbf{H_1}$: Variance is not equal in all income levels

Hide

```
leveneTest(age~income, data=poll84)
```

```
Warning in leveneTest.default(y = y, group = group, ...) :
  group coerced to factor.
```

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    4  44.249 < 2.2e-16 ***
      1204
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since the p-value of the Levene's test is 2.2e-16 < 0.05 we reject the null hypothesis and conclude that the assumption that the variance is equal in all income groups cannot be considered.

Hypothesis

Lets assume that we test the below at 5% significance level:

$$H_0 : \sigma_l^2 = \sigma_L^2 = \sigma_m^2 = \sigma_h^2 = \sigma_H^2.$$

H_1: At least one is different

Hide

```
fit<-aov(age ~ income,data=poll84)
summary(fit)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
income         4  13545    3386   12.86 2.94e-10 ***
Residuals   1204 316922     263
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since the P-value is 2.94e-10 < 0.05, this is an evidence to reject the null hypothesis. Hence we conclude that there is at least one income group with different mean(age) and the average age is not equal in all income groups.