Hypothesis Testing of Drink Preference

Reading Survey Data and libraries

```
data <- read.csv('coffeetea.csv')
library(dplyr)
```

Preparing Raw and Unstructered Data to Useful data

- The First five email address are not present (We added the field later). Assigning random email addresses to the first 5 rows

```
finaldata = data[!duplicated(data$Email.Address), ]
```

- We performed the 1st exercise to make the email address column not null. Now, we remove the duplicates with respect to the email address from the 2nd exercise above

Exploratory Data Analysis

```
summary(finaldata[,2:7])
```

```
##  Do.you.prefer.tea.or.coffee. How.many.times.do.you.drink.it.on.daily.basis.
##  Coffee:31                    1        :31
##  None  : 3                    2        :20
##  Tea   :32                    3 or more: 5
##                               Never    :10
##
##
##
##     Gender     Nationality    Age                      Email.Address
##  Female:32   American: 2   >30  : 1   301194sach@gmail.com    : 1
##  Male  :34   Chinese :17   18-20: 3   8888@husky.nau.edu      : 1
##              Indian  :42   21-30:62   a@b.com                 : 1
##              Taiwan  : 4              abc70177@hotmail.com    : 1
##              Thai    : 1              amalsharma24@gmail.com  : 1
##                                       angikasingh54@gmail.com: 1
##                                       (Other)                 :60
```

- We have collected 6 attributes of data namely

- 1. Do you prefer Coffee?

- 2. Number of drinks/day

- 3. Gender

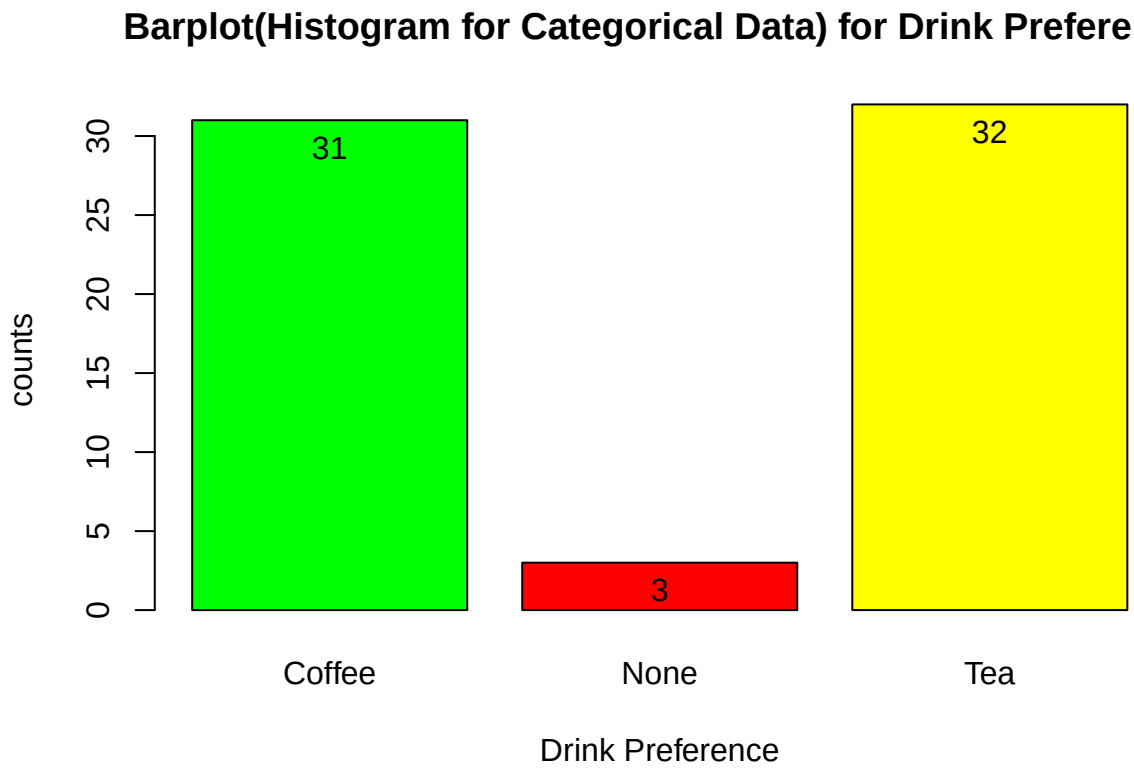
- 4. Nationality
- 5. Age
- 6. Email Address

Our main attribute (on which we are going to perform our analysis) is Drink Preference (1st attribute)

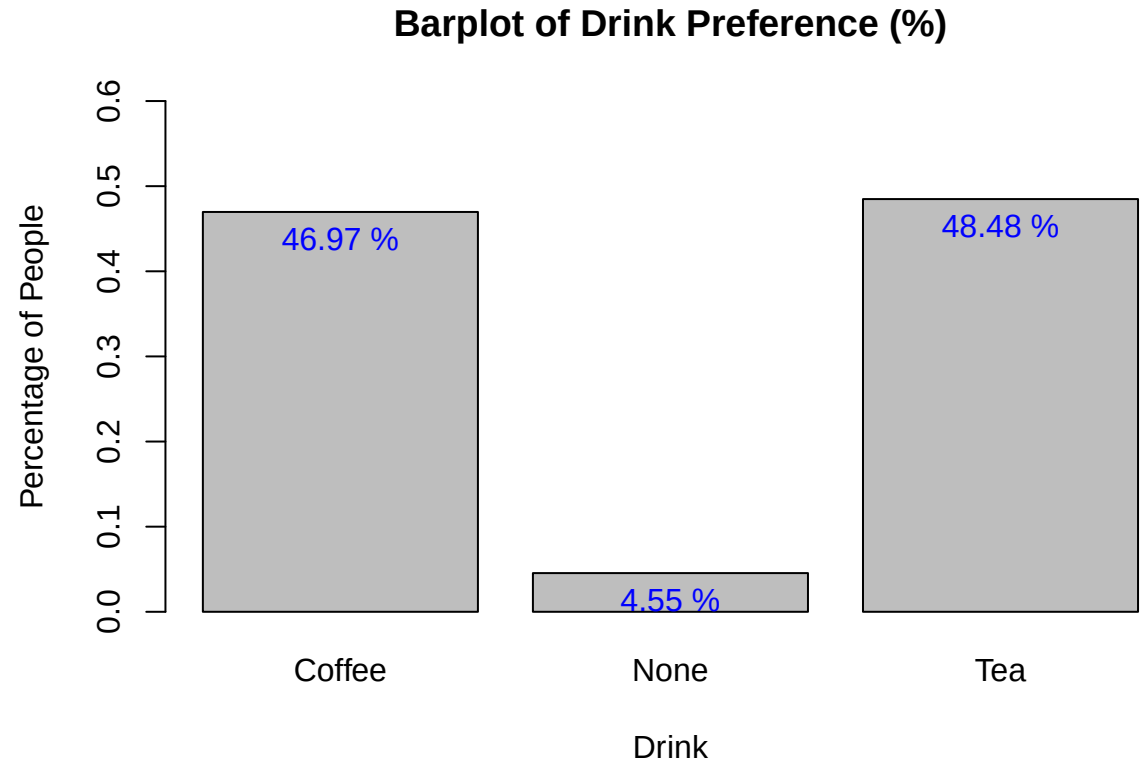
- We can see a high level summary below for all the attributes.
- We shall describe the main attribute's description in the next block
- We got a good Male:Female ratio which is 34:32
- We also got the survey response from diverse Nationalities like American, India, Chinese, Thailand, and Taiwan
- More than 95% of our age-group is 21-30 (which is a slight limitation to our survey)

Drink Preference Analysis

```
counts = table(finaldata$Do.you.prefer.tea.or.coffee.)

x = barplot(counts, main="Barplot(Histogram for Categorical Data) for Drink Preference",
  xlab="Drink Preference", ylab='counts', col = c("green","red", "yellow"), beside = TRUE)

text(x, counts, labels = counts ,cex=1, pos = 1)
```



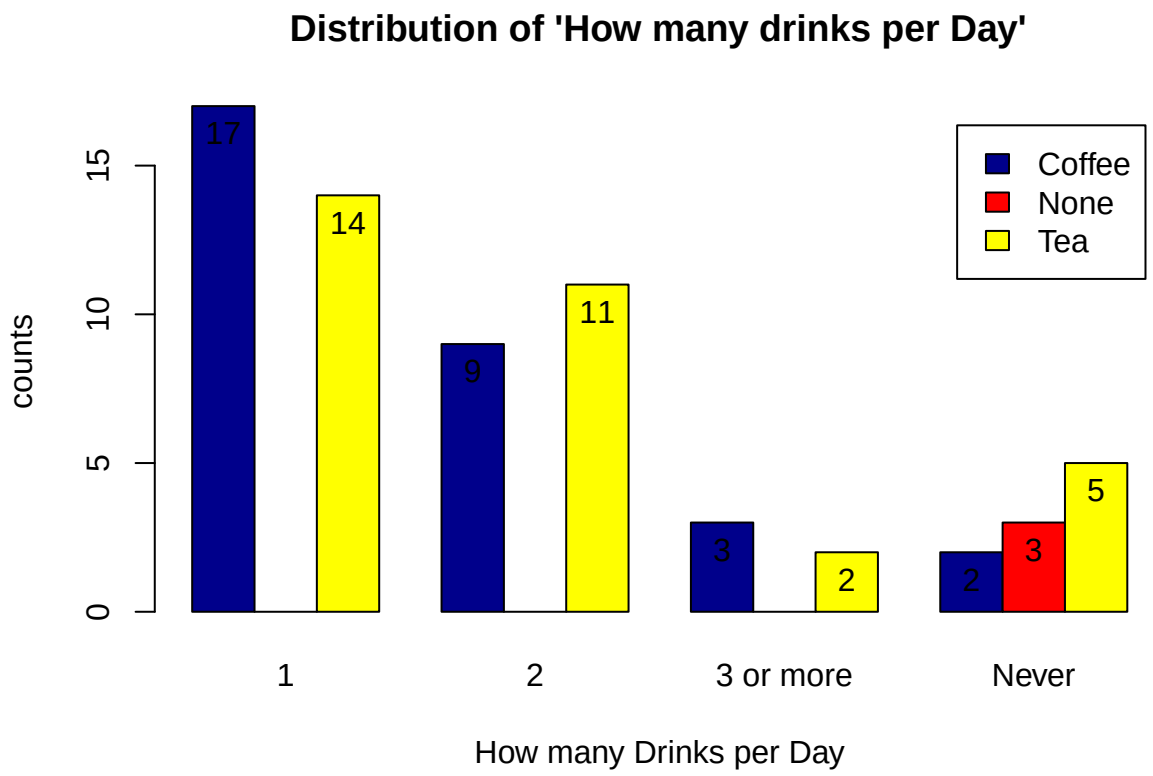

```
x = barplot(prop.table(table(finaldata$Do.you.prefer.tea.or.coffee.)), main = 'Barplot of Drink Prefere

text(x, prop.table(table(finaldata$Do.you.prefer.tea.or.coffee.)), labels = paste(round(prop.table(tabl
```



Barplots are the histograms of Categorical Variables

We see that 46.97% of of our sample prefers Coffee, while 48.48% perfer Tea and 4.55% doesn't prefer any of the drink

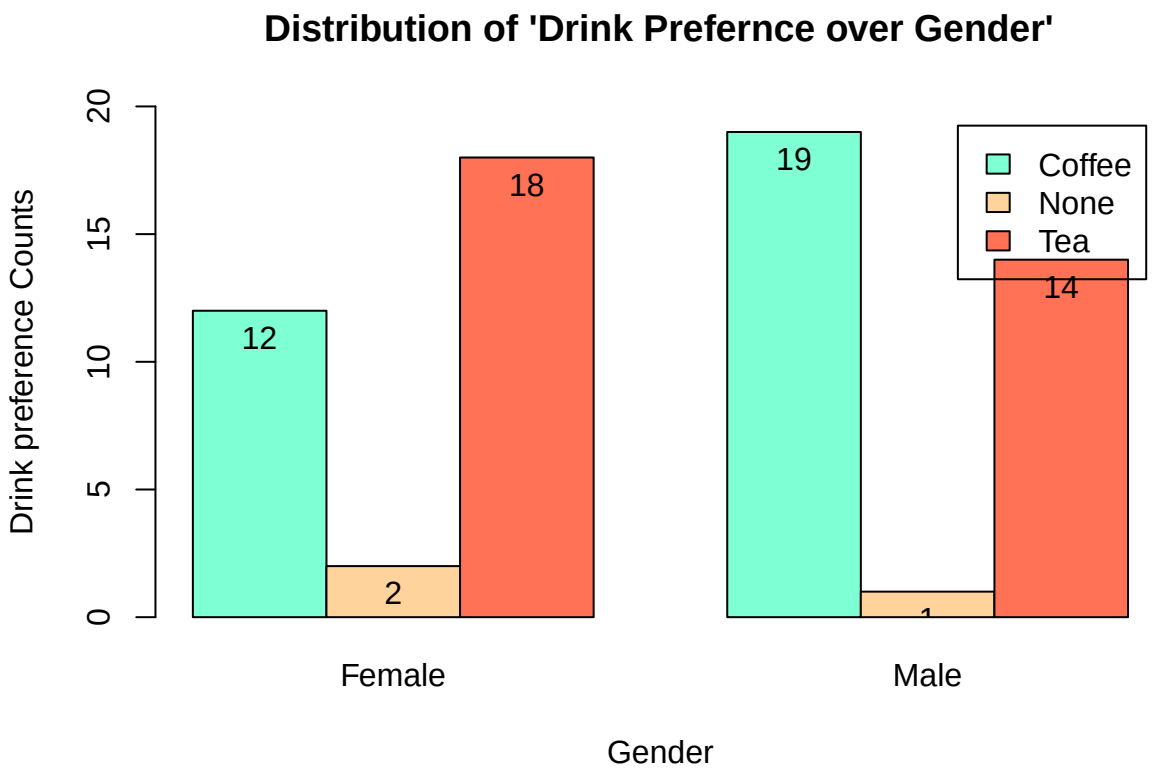
```
counts = table(finaldata$Do.you.prefer.tea.or.coffee., finaldata$How.many.times.do.you.drink.it.on.daily

x = barplot(counts, main="Distribution of 'How many drinks per Day'",
  xlab="How many Drinks per Day", ylab='counts', col = c("darkblue","red","yellow"),legend = rownames(c

text(x, counts, labels = counts ,cex=1, pos = 1)
```



The distribution of the drinks per day with respect to Drink Preference is shown in the graph.

- More Coffee drinkers have 1 drink per day than tea drinkers
- While more Tea drinkers have 2 drinks per day than coffee drinkers
- 7 people have a preference over Coffee or Tea but don't drink it on a daily basis
- 3 people have no preference of tea or coffe and doesn't drink them on a daily basis

```
counts = table(finaldata$Do.you.prefer.tea.or.coffee., finaldata$Gender)

x = barplot(counts, main="Distribution of 'Drink Prefernce over Gender'",
  xlab="Gender", ylab='Drink preference Counts', col = c('aquamarine', 'burlywood1', 'coral1'), legend =

text(x, counts, labels = counts ,cex=1, pos = 1)
```



The distribution drink preference with respect to the gender is shown in the above graph.

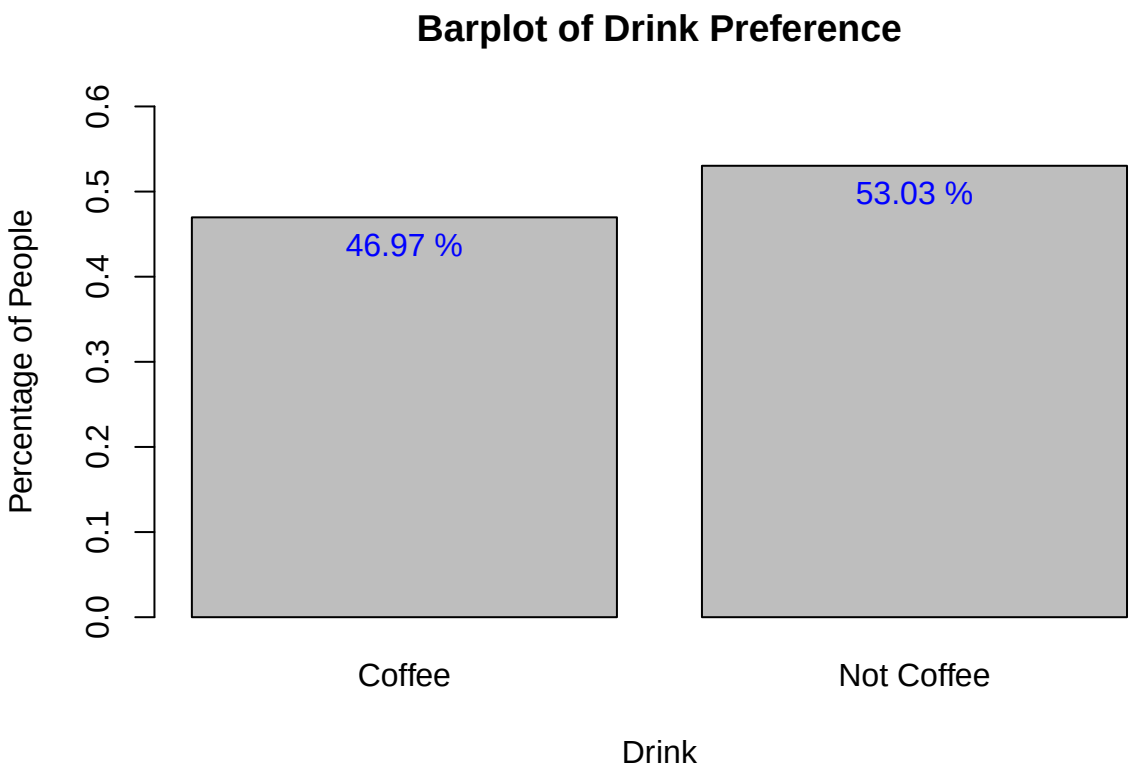
- It's an interesting fact that more Males prefer Coffee than Females.
- While more Females prefer Tea than Males

Altering the Drink Preference column to a bi-valued column. - According to our statistical analysis, the column should have 2 values - 'Coffee', 'Not Coffee'

```
finaldata$Do.you.prefer.tea.or.coffee. = as.character(finaldata$Do.you.prefer.tea.or.coffee.)

finaldata$Do.you.prefer.tea.or.coffee.[finaldata$Do.you.prefer.tea.or.coffee. != 'Coffee'] = 'Not Coffe
```

Now let's look at the distribution of how the drink preference looks like

```
x = barplot(prop.table(table(finaldata$Do.you.prefer.tea.or.coffee.)), main = 'Barplot of Drink Prefere

text(x, prop.table(table(finaldata$Do.you.prefer.tea.or.coffee.)), labels = paste(round(prop.table(tabl
```



We can simply say that 46.97% prefer Coffee while 53.03% doesn't prefer Coffee

Statistical Analysis

Null Hypothesis: The proportion of people who prefer coffee will be equal to the proportion of people who don't prefer Coffee

Alternate Hypothesis: There proportion of people who prefer cofffe will not be equal to the proportion of people who don't prefer coffee

- Sample Size n

```
n = length(finaldata$Do.you.prefer.tea.or.coffee.)
```

- Sample size n = 66

Here Null Hypothesis says that P = P0. So we have P = P0 = 0.5

Calculating P_hat

```
p0 = 0.5
p_hat = length(subset(finaldata$Do.you.prefer.tea.or.coffee., finaldata$Do.you.prefer.tea.or.coffee. ==

print(p_hat)
```

```
## [1] 0.469697
```

Confidence Intervals

- The Confidence Intervals for the above are p_hat - E and p_hat + E. Where E is the Error.

CI: (p_hat-E, p_hat+E)

where E = z * sqrt(p_hat*q_hat / n)

and We have q_hat = 1 - p_hat

Our Confidence level is 95%

Hence alpha = 0.05 and alpha/2 = 0.025 Hence we have to calculate z-value for 1-0.025 = 0.975

```
E = qnorm(0.975)*sqrt(p_hat*(1-p_hat)/n)
print(E)
```

```
## [1] 0.1204057
```

Hence the Error for the confodence interval is 0.12

```
Lower_CI = p_hat - E
Upper_CI = p_hat + E

print(Lower_CI)
```

```
## [1] 0.3492913
```

```
print(Upper_CI)
```

```
## [1] 0.5901027
```

Lower Confidence Interval = 0.34929

Upper Confidence Interval = 0.59010

We can say that, with 95% confidence, the population proportion of people who prefer coffee will definetly be in the range (0.349, 0.590)

Calculating Test Statistic

```
z = (p_hat - p0)/sqrt(p0*(1-p0)/n)
print(z)
```

```
## [1] -0.492366
```

Calculating p-value

```
p_value = 2*pnorm(z)
print(p_value)
```

```
## [1] 0.6224607
```

p-value is 0.62 which greater than our confidence level 0.05

We fail to reject the Null Hypothesis

Statistical Analysis using Traditional Method

*The Yates continuity correction is disabled for pedagogical reasons.

```
finaltest = prop.test(length(subset(finaldata$Do.you.prefer.tea.or.coffee., finaldata$Do.you.prefer.tea

finaltest
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  length(subset(finaldata$Do.you.prefer.tea.or.coffee., finaldata$Do.you.prefer.tea.or.coffee. =
## X-squared = 0.24242, df = 1, p-value = 0.6225
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3543043 0.5884231
## sample estimates:
##        p
## 0.469697
```

The p_value obtained in the traditional test is also 0.6225. While it is greater than our confidence level 0.05, We interpret the following

There is a much higher probability (0.6225) that the null hypothesis (the proportion of people who prefer coffee is equal to the proportion of people who prefer Tea) is TRUE. when compared to our probability (0.05)

Hence we fail to reject the Null Hypothesis.

```
print(paste("Test Statistic: ",finaltest$statistic))
```

```
## [1] "Test Statistic:  0.242424242424242"
```

```
print(paste("Parameter: ",finaltest$parameter))
```

```
## [1] "Parameter:  1"
```

```
print(paste("P_Value: ",finaltest$p.value))
```

```
## [1] "P_Value:  0.622460655893454"
```

```
print(paste("Null Value: ",finaltest$null.value))
```

```
## [1] "Null Value:  0.5"
```

```
print(paste("Confidence Intervals : ",finaltest$conf.int))
```

```
## [1] "Confidence Intervals :  0.354304284963777"
## [2] "Confidence Intervals :  0.588423142734014"
```

```
print(paste("Alternative: ", finaltest$alternative))
```

```
## [1] "Alternative:  two.sided"
```

```
print(paste("Method: ", finaltest$method))
```

```
## [1] "Method:  1-sample proportions test without continuity correction"
```

```
knitr::opts_chunk$set(echo = TRUE)
data <- read.csv('coffeetea.csv')
library(dplyr)

finaldata = data[!duplicated(data$Email.Address), ]
summary(finaldata[,2:7])

counts = table(finaldata$Do.you.prefer.tea.or.coffee.)

x = barplot(counts, main="Barplot(Histogram for Categorical Data) for Drink Preference",
  xlab="Drink Preference", ylab='counts', col = c("green","red", "yellow"), beside = TRUE)

text(x, counts, labels = counts ,cex=1, pos = 1)
x = barplot(prop.table(table(finaldata$Do.you.prefer.tea.or.coffee.)), main = 'Barplot of Drink Preferer

text(x, prop.table(table(finaldata$Do.you.prefer.tea.or.coffee.)), labels = paste(round(prop.table(table

counts = table(finaldata$Do.you.prefer.tea.or.coffee., finaldata$How.many.times.do.you.drink.it.on.daily

x = barplot(counts, main="Distribution of 'How many drinks per Day'",
  xlab="How many Drinks per Day", ylab='counts', col = c("darkblue","red","yellow"),legend = rownames(co

text(x, counts, labels = counts ,cex=1, pos = 1)
counts = table(finaldata$Do.you.prefer.tea.or.coffee., finaldata$Gender)

x = barplot(counts, main="Distribution of 'Drink Prefernce over Gender'",
  xlab="Gender", ylab='Drink preference Counts', col = c('aquamarine', 'burlywood1', 'coral1'), legend =

text(x, counts, labels = counts ,cex=1, pos = 1)

finaldata$Do.you.prefer.tea.or.coffee. = as.character(finaldata$Do.you.prefer.tea.or.coffee.)

finaldata$Do.you.prefer.tea.or.coffee.[finaldata$Do.you.prefer.tea.or.coffee. != 'Coffee'] = 'Not Coffe
x = barplot(prop.table(table(finaldata$Do.you.prefer.tea.or.coffee.)), main = 'Barplot of Drink Preferer

text(x, prop.table(table(finaldata$Do.you.prefer.tea.or.coffee.)), labels = paste(round(prop.table(table
n = length(finaldata$Do.you.prefer.tea.or.coffee.)

p0 = 0.5
p_hat = length(subset(finaldata$Do.you.prefer.tea.or.coffee., finaldata$Do.you.prefer.tea.or.coffee. ==

print(p_hat)

E = qnorm(0.975)*sqrt(p_hat*(1-p_hat)/n)
print(E)

Lower_CI = p_hat - E
Upper_CI = p_hat + E
```

```r
print(Lower_CI)
print(Upper_CI)
z = (p_hat - p0)/sqrt(p0*(1-p0)/n)
print(z)

p_value = 2*pnorm(z)
print(p_value)

finaltest = prop.test(length(subset(finaldata$Do.you.prefer.tea.or.coffee., finaldata$Do.you.prefer.tea

finaltest

print(paste("Test Statistic: ",finaltest$statistic))

print(paste("Parameter: ",finaltest$parameter))

print(paste("P_Value: ",finaltest$p.value))

print(paste("Null Value: ",finaltest$null.value))

print(paste("Confidence Intervals : ",finaltest$conf.int))

print(paste("Alternative: ", finaltest$alternative))

print(paste("Method: ", finaltest$method))
```