lab2_Chen

Weixuan Chen

2/1/2023

Download the data

```
# Download the data
d2012=read.csv("http://faraway.neu.edu/biostats/lab2_dataset1.csv")
d2008=read.csv("http://faraway.neu.edu/biostats/lab2_dataset2.csv")
d2004=read.csv("http://faraway.neu.edu/biostats/lab2_dataset3.csv")
```

Task 1

1

```
skewness1 <- function (x) {
  x_mean <- mean(x)
  x_sd <- sd(x)
  x_size <- length(x)
  sum <- 0
  for (value in x){
    sum = sum + (value - x_mean)^3
  }
  result <- sum/(x_size*x_sd^3)

  return (result)
}
```

2

2004

```
mean(d2004$population.size)
```

```
## [1] 1017.972
```

```
median(d2004$population.size)
```

```
## [1] 814.5
```

```
skewness1(d2004$population.size)
```

```
## [1] 1.505799
```

2008

```
mean(d2008$population.size)
```

```
## [1] 1038.789
```

```
median(d2008$population.size)
```

```
## [1] 868
```

```
skewness1(d2008$population.size)
```

```
## [1] 1.44054
```

2012

```
mean(d2012$population.size)
```

```
## [1] 993.953
```

```
median(d2012$population.size)
```

```
## [1] 969
```

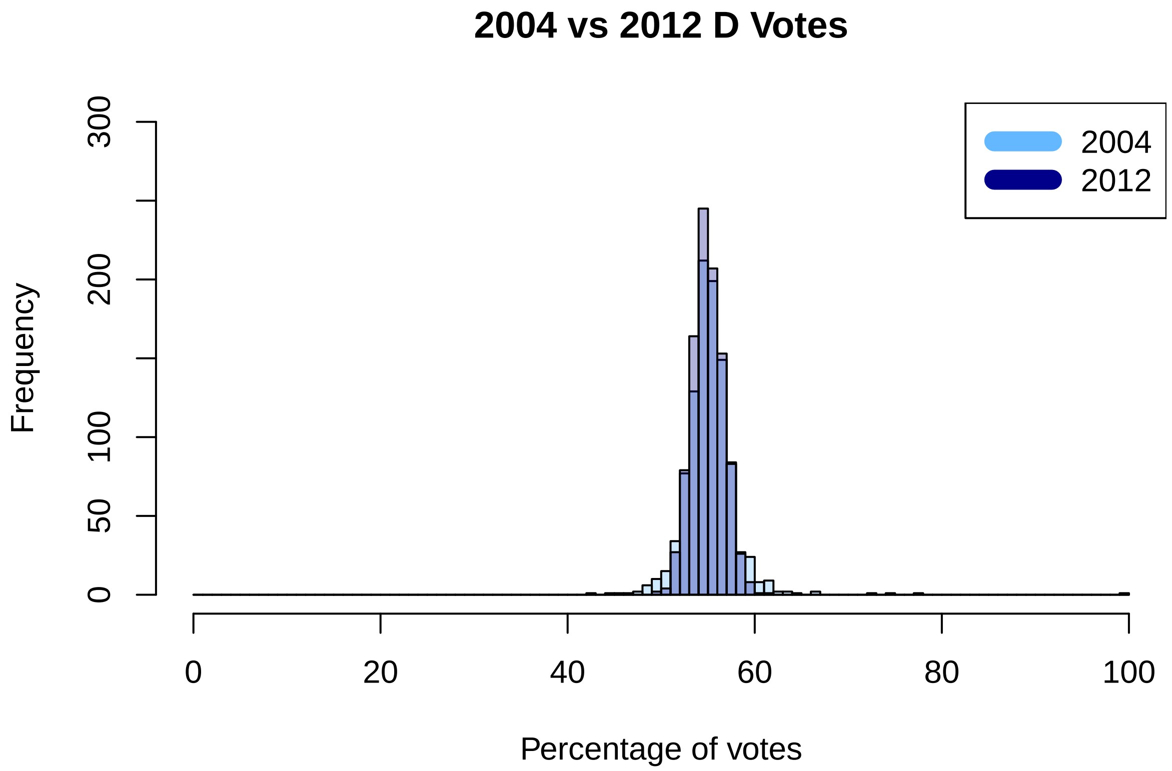
```
skewness1(d2012$population.size)
```

```
## [1] 0.5899025
```

Based on skewness, we can see the metric in 2004 and 2008 are far greater than 0.5, which indicates a strong sknewness of the distribution. This means that most voting precincts are small because most of the population size is smaller than the mean and mean is greater than the median. In 2012, the skewness gets smaller and it is moderate. This means the distribution of the population size of precincts are moderately symmetric and mean and median are close to each other.

3

```
hist(d2004$voted.democrat/d2004$population.size * 100,
     col = adjustcolor("steelblue1", alpha = 0.3), xlim = c(0, 100), ylim = c(0, 300),
     xlab = "Percentage of votes", main = "2004 vs 2012 D Votes", breaks = seq(0, 100, by = 1))
hist(d2012$voted.democrat/d2012$population.size * 100,
     col = adjustcolor("blue4", alpha = 0.3),
     breaks = seq(0, 100, by = 1), add = T)
legend("topright", c("2004", "2012"), col=c("steelblue1", "blue4"), lwd=10)
```



4

The trend of the votes is more centralized towards around 57% percent in 2012 compared to 2004. I don't think there is enough evidence that shows the votes are polarized.

Task 2

1

```
# Use substr to extract first digit
first.digit=substr(as.character(d2012$voted.democrat), start=1, stop=1)
# Convert the first digit to a number
first.digit=as.numeric(first.digit)

digits_table <- table(first.digit)
digits_table
```

```
## first.digit
##   1   2   3   4   5   6   7   8   9
##   4  12 104 257 314 190  92  26   1
```

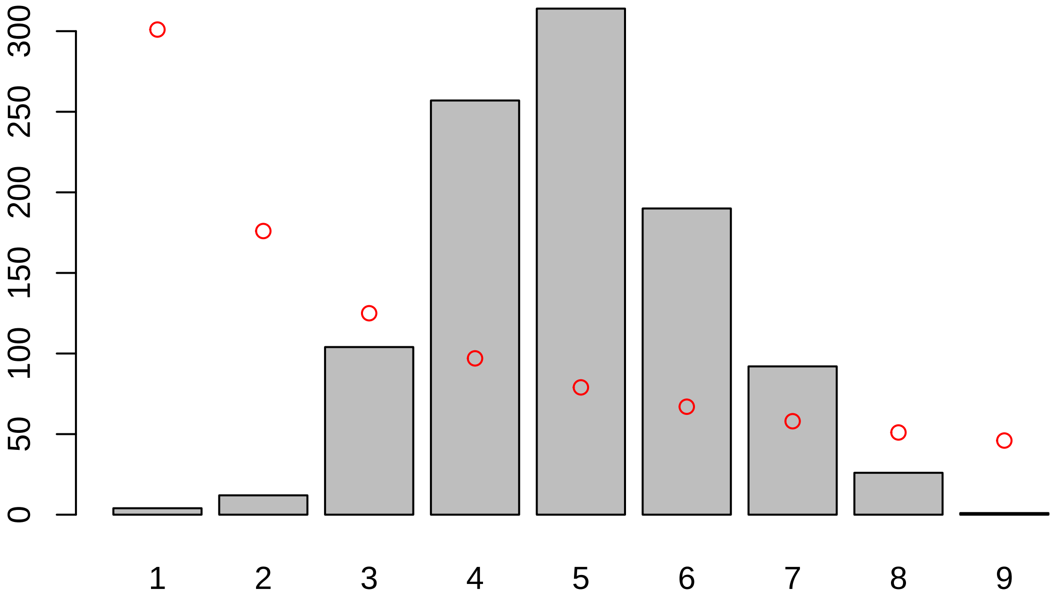
2

```r
# Benford's law
expected=log10(1+1/(1:9))
# Expected count for each digit based on Benford's Law
(expected=round(expected*sum(digits_table)))
```

```
## [1] 301 176 125  97  79  67  58  51  46
```

3

```r
bp=barplot(digits_table, names=1:9)
points(bp, expected, pch=1, col=c("red"))
```



4

Apparently, the first digit from prediction is not consistent with our observed data.

5

H0 = There is no relationship between observed data and expected data H1 = Not H0 (there is a relationship)

```
df_digits <- as.data.frame(digits_table)
df_digits <- cbind(df_digits, expected)
df_digits
```

```
##   first.digit Freq expected
## 1           1    4      301
## 2           2   12      176
## 3           3  104      125
## 4           4  257       97
## 5           5  314       79
## 6           6  190       67
## 7           7   92       58
## 8           8   26       51
## 9           9    1       46
```

```
chisq.test(df_digits$Freq, df_digits$expected)
```

```
## Warning in chisq.test(df_digits$Freq, df_digits$expected): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  df_digits$Freq and df_digits$expected
## X-squared = 72, df = 64, p-value = 0.2303
```

We have p-value > alpha level which we have to not reject the null hypothesis. This means that the two groups of data can be independent.

6

```
expected11=log10(1+1/(1:9))
df_digits <- cbind(df_digits, expected11)
chisq.test(df_digits$Freq, df_digits$expected11)
```

```
## Warning in chisq.test(df_digits$Freq, df_digits$expected11): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  df_digits$Freq and df_digits$expected11
## X-squared = 72, df = 64, p-value = 0.2303
```

We can see the result of p-value is the same as previous one. Since we are using the probability as argument here, we can say the probabilities from the Benford's law is independent of our observed data, which means the observed data does not follow the Benford's law.

Task 3

1

```
# P(Win)
pW=0.5
# P(Favored|Win)
pF.W=0.75
# P(Favored|Loss)
pF.L=0.20
#P(Favored) = P(Favored|Win) * P(Win) + P(Favored|Loss) * P(Loss)
pF = pF.W * pW + pF.L * (1 - pW)
#P(Win|Favoired) =  P(Favored|Win) * P(Win)/P(Favored)
pW.F = pF.W*pW/pF
pW.F
```

```
## [1] 0.7894737
```

I would say if the Democratic candidate is favored(leading in the polls), the chance of wining the election is approximately 79%.

2

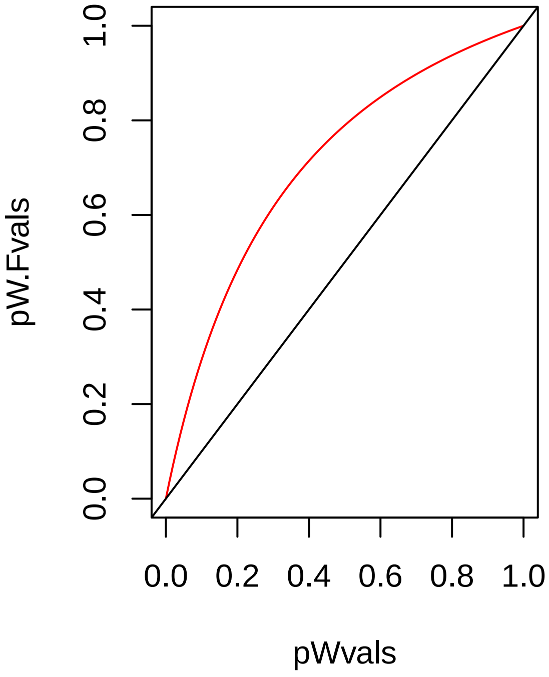
```
# Create a vector of pW values (i.e., P(W))
pWvals=seq(0, 1, length=101)
# Initialize the vector of pW.Fvals (i.e., P(W|F))
pW.Fvals=numeric(101)
pFvals=numeric(101)
for (i in 1:length(pWvals)) {
  # P(Win)
  pW=pWvals[i]
  # P(Favored|Win)
  pF.W=0.75
  # P(Favored|Loss)
  pF.L=0.20
  #P(Favored) = P(Favored|Win) * P(Win) + P(Favored|Loss) * P(Loss)
  pF = pF.W * pW + pF.L * (1 - pW)
  pFvals[i] = pF
  #P(Win|Favoired) =  P(Favored|Win) * P(Win)/P(Favored)
  pW.F = pF.W*pW/pF
  pW.Fvals[i] = pW.F
}

pW.Fvals
```

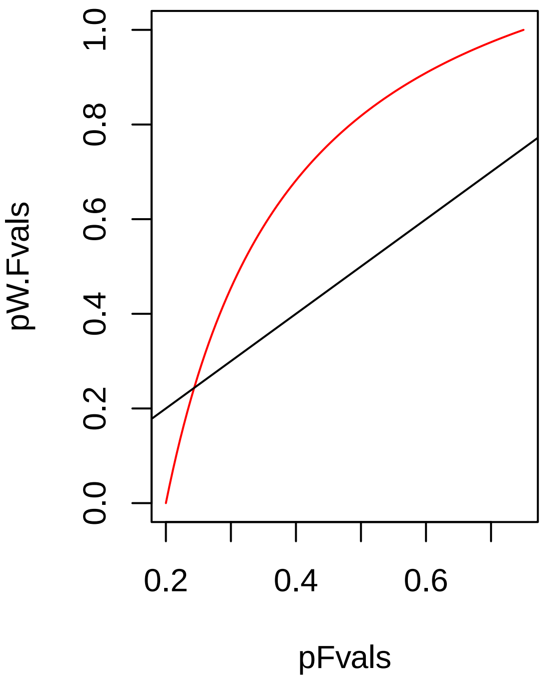
```
##   [1] 0.00000000 0.03649635 0.07109005 0.10392610 0.13513514 0.16483516
##   [7] 0.19313305 0.22012579 0.24590164 0.27054108 0.29411765 0.31669866
##  [13] 0.33834586 0.35911602 0.37906137 0.39823009 0.41666667 0.43441227
##  [19] 0.45150502 0.46798030 0.48387097 0.49920761 0.51401869 0.52833078
## [25] 0.54216867 0.55555556 0.56851312 0.58106169 0.59322034 0.60500695
## [31] 0.61643836 0.62753036 0.63829787 0.64875491 0.65891473 0.66878981
## [37] 0.67839196 0.68773234 0.69682152 0.70566948 0.71428571 0.72267920
## [43] 0.73085847 0.73883162 0.74660633 0.75418994 0.76158940 0.76881134
## [49] 0.77586207 0.78274760 0.78947368 0.79604579 0.80246914 0.80874873
## [55] 0.81488934 0.82089552 0.82677165 0.83252191 0.83815029 0.84366063
## [61] 0.84905660 0.85434174 0.85951941 0.86459286 0.86956522 0.87443946
## [67] 0.87921847 0.88390501 0.88850174 0.89301122 0.89743590 0.90177815
## [73] 0.90604027 0.91022444 0.91433278 0.91836735 0.92233010 0.92622294
## [79] 0.93004769 0.93380615 0.93750000 0.94113091 0.94470046 0.94821021
## [85] 0.95166163 0.95505618 0.95839525 0.96168018 0.96491228 0.96809282
## [91] 0.97122302 0.97430407 0.97733711 0.98032326 0.98326360 0.98615917
## [97] 0.98901099 0.99182004 0.99458728 0.99731363 1.00000000
```

3

```
par(mfrow=c(1,2))
plot(pWvals,pW.Fvals,type='l', col=c("red"))
abline(b=1, a=0)
```



```
par(mfrow=c(1,2))
plot(pFvals,pW.Fvals,type='l', col=c("red"))
abline(b=1, a=0)
```



**I think the problem should be the pW.Fvals vs pFvals, not pWvals. I plotted both graphs but only made comments to pW.Fvals vs pFvals because this shows the relationship while the other does not.

The line we add is y=x, in our case it implies pFvals = pW.Fvals. That's why when the red line is below the black line, P $(W|F) < P(F)$, whereas when the red line is above the black line, $P(W|F) > P(F)$.