R HW6

Due: 2020/1/1 23:59

Notice

This assignment requires more descriptive answers rather than codes. You still have to run some codes and hand in a pdf file, but what you need to answer are more related to statistical or programming concepts.

The purpose of the 6^{th} homework for R is to consolidate what we've learned during the 5 lessons. If you have achieved full grades in the previous 5 homeworks, then you can skip this homework. If you didn't, this homework is relatively easy, then it may be a good chance for you to do it.

A. Basic Data Types in R

I throw a 6 sided fair dice for n times and record the sum of numbers from the n outcomes I get, denote as $n\bar{X_n}$ or $\sum_{i=1}^{n} X_i$. I want to repeat this process for $t = 10000$ times and see how $\sum_{i=1}^{n} X_i$ behaves.

Let $n = 10, 100, 1000$ respectively. Run the following codes and answer the questions.

```
dice = function(n){
  X = sample(1:6, size = n, replace = T)
  return(mean(X))
}

Xbar10 = replicate(10000, dice(10))
Xbar100 = replicate(10000, dice(100))
Xbar1000 = replicate(10000, dice(1000))
```

1. Are `Xbar10`, `Xbar100`, `Xbar1000` scalars? Or are they vectors? What are the length of them? (You can use `length()` to find out.)

2. What are the means and standard deviations of `Xbar10`, `Xbar100`, `Xbar1000`? Does the standard deviations get samller when n gets larger?

3. Does the distribution for $\sum_{i=1}^{n} X_i$ look like normal as n gets larger? Why? (i.e. What theorem supports this result?)

4. If we see `Xbar10`, `Xbar100`, `Xbar1000` as random variables, denote as $\bar{X}_{10}$, $\bar{X}_{100}$, $\bar{X}_{1000}$, what are the theoretical means and variance of these random variables? Are they consistent with the result in (2.)?

B. Graph

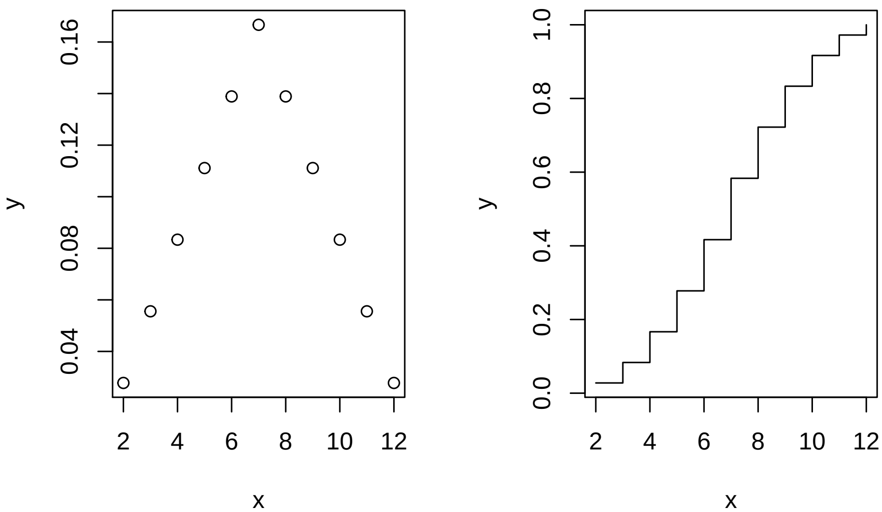
Let's continue the setup in **(A.)** Let X_1 be a r.v. denoting the outcome number from the first throw. Also, let X_2 be a r.v. denoting the outcome number from the second throw. There are 2 throws only, i.e. what we are interested in is the distribution of $X_1 + X_2$

Note that the possible realizations for r.v. $X_1 + X_2$ would be : $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

The graph of p.m.f. & c.d.f. for $X_1 + X_2$ are:

```
par(mfrow = c(1,2))
x = 2:12
y = c(1/36, 2/36, 3/36, 4/36, 5/36, 6/36, 5/36, 4/36, 3/36, 2/36, 1/36)
plot(x, y, type = 'p')
```

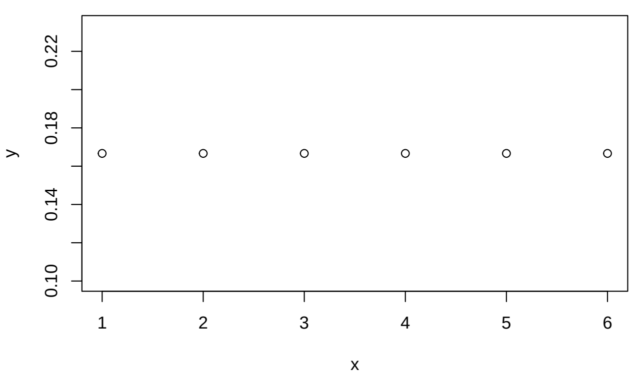
```r
x = 2:12
y = c(1/36, 3/36, 6/36, 10/36, 15/36, 21/36, 26/36, 30/36, 33/36, 35/36, 1)
plot(x, y, type = 's')
```




```r
graphics.off()
```

Note that the graph of p.m.f. for $X_i, i = 1, 2$ is:

```r
x = 1:6
y = rep(1/6, 6)
plot(x, y, type = 'p')
```



5. Try to explain why the p.m.f. of $X_1 + X_2$ looks like a triangle. What do you find from the p.m.f. of X_i to the p.m.f. of $\sum_{i=1}^{2} X_i$? What would you expect when $i = 1, 2, \ldots, n$ and n is large? Is this result related to CLT?

6. We graph the p.m.f. with `type='p'`. Briefly explain why.

7. We graph the c.d.f. with `type='s'` instead of `type='l'`. Give a brief explaination.

C. Bootstrap & Permutation(Hypothesis) Test

Recall the example in HW5:

```r
Verizon = read.csv("http://sites.google.com/site/chiharahesterberg/data2/Verizon.csv")
Time.ILEC = subset(Verizon, select = Time, Group == "ILEC", drop = T)
Time.CLEC = subset(Verizon, select = Time, Group == "CLEC", drop = T)
```

```r
B = 10^4
time.ratio.mean = numeric(B)
for(i in 1:B){
  ILEC.sample = sample(Time.ILEC, 1664, replace = TRUE)
  CLEC.sample = sample(Time.CLEC, 23, replace = TRUE)
  time.ratio.mean[i] = mean(ILEC.sample)/mean(CLEC.sample)
}
```

We can get the bootstrap standard error easily from simulation.

```r
sd(time.ratio.mean)
```

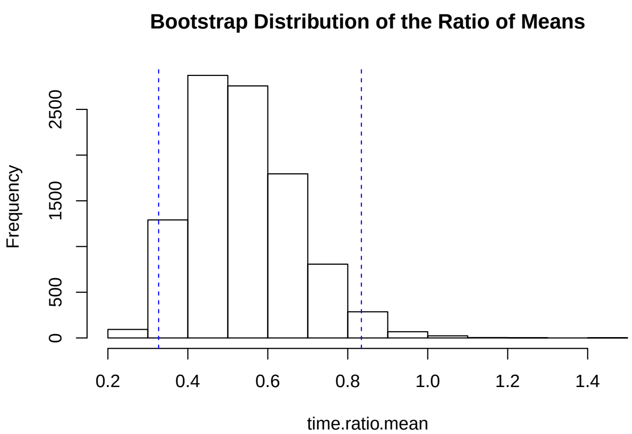
```
## [1] 0.1323682
```

Given $\alpha = 0.05$, we can construct a bootstrap interval estimate.

```r
#The Interval Estimate with 95% Confidence
quantile(time.ratio.mean, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.3269473 0.8340970
```

```r
L = quantile(time.ratio.mean, 0.025)
U = quantile(time.ratio.mean, 0.975)
hist(time.ratio.mean, main="Bootstrap Distribution of the Ratio of Means")
abline(v=L, col = "blue", lty = 2)
abline(v=U, col = "blue", lty = 2)
```



8. Why we construct the bootstrap interval by using `quantile()` instead of writing $\bar{X}_{ILEC}/\bar{X}_{CLEC} \pm 1.96 \times se$ where $se =$ `sd(time.ratio.mean)`? Think about what we know from out statistic of interest and where do 1.96 come from.

9. What does this bootstrap interval estimate tell us? Give at least one statistical insight.

In permutation test, we're actually doing the hypothesis test where:

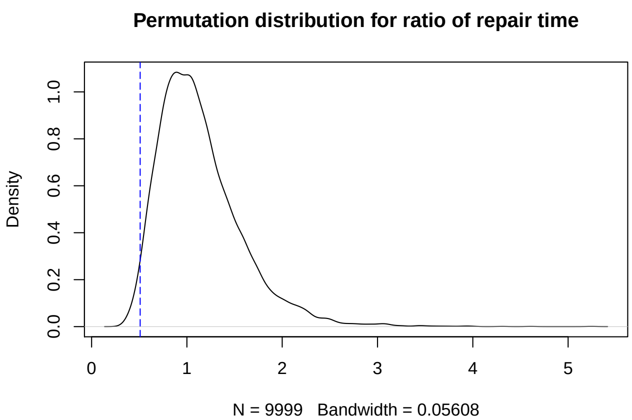
$$H_0 : \text{The repair time for ILEC is equal to the repair time for CLEC}$$

$$H_a : \text{The repair time for ILEC is less than the repair time for CLEC}$$

```r
repairTime = Verizon$Time
observed = mean(Time.ILEC)/mean(Time.CLEC)
```

```r
N = 10000-1  #set number of times to repeat this process
result = numeric(N) # space to save the random differences
for(i in 1:N){
  index = sample(1687, size=1664, replace = FALSE) # sample of numbers from 1:1687
  result[i] = mean(repairTime[index])/mean(repairTime[-index])
}

plot(density(result), main = "Permutation distribution for ratio of repair time")
abline(v = observed, col = "blue", lty=5)
```



10. Is the distribution we graph under H_0 or H_a? How to interpret the area which is at the left hand side of the blue dash line under the curve?