Lab5_comp_stat

Gowtham(gowku593) & Biswas (bisku859)

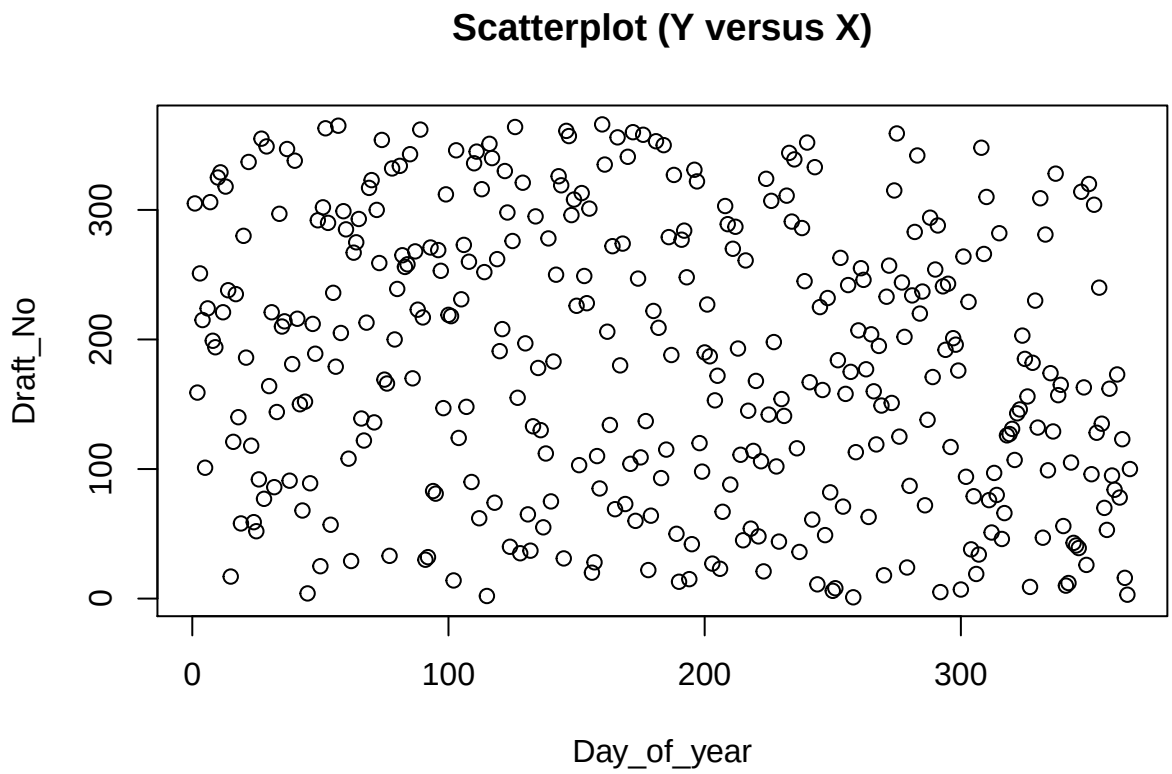
3/29/2020

Question 1: Hypothesis testing

In 1970, the US Congress instituted a random selection process for the military draft. All 366 possible birth dates were placed in plastic capsules in a rotating drum and were selected one by one. The frst date drawn from the drum received draft number one, the second date drawn received draft number two, etc. Then, eligible men were drafted in the order given by the draft number of their birth date. In a truly random lottery there should be no relationship between the date and the draft number. Your task is to investigate whether or not the draft numbers were randomly selected. The draft numbers (Y=Draft_No) sorted by day of year (X=Day_of_year) are given in the file lottery.xls.

1.

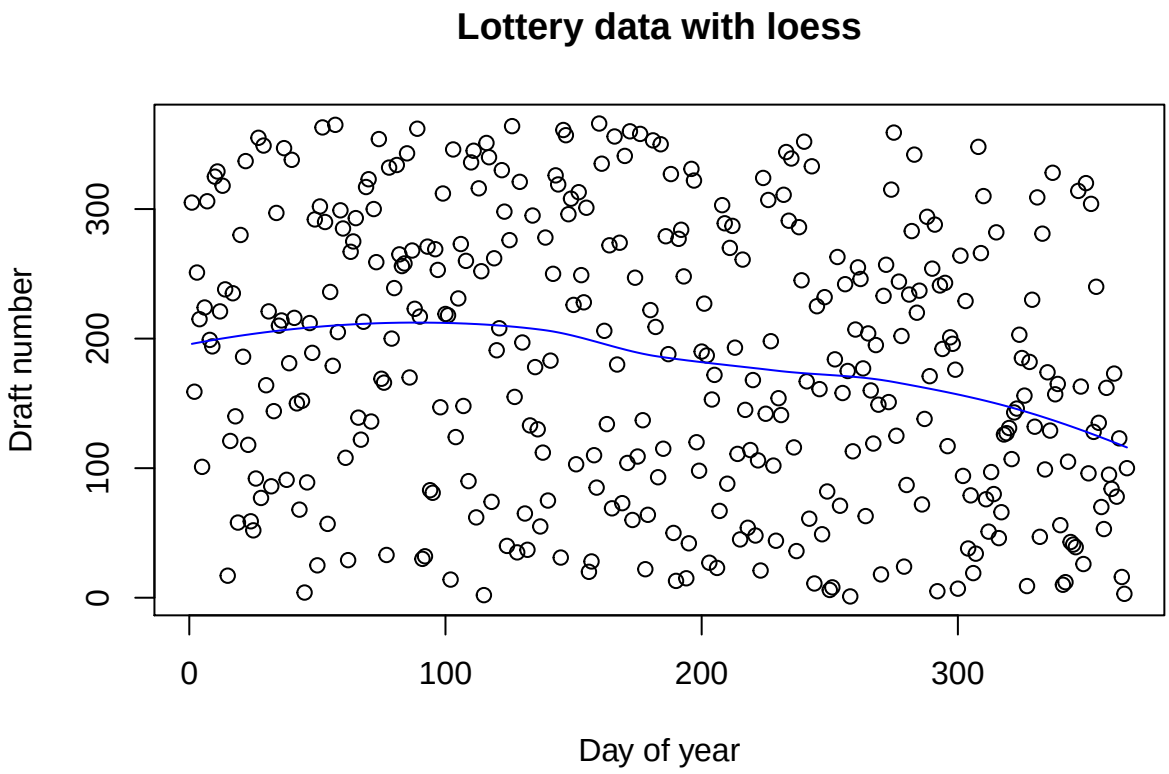
Make a scatterplot of Y versus X and conclude whether the lottery looks random



Looking at the scatterplot, we see that the data is distributed all over the plot . Its difficult to find any trend by visual inspection . Therefore, It is easier to can conclude that the probability of random selection for the military draft is very high.

2.

Compute an estimate $\hat{Y}$ of the expected response as a function of X by using a loess smoother (use loess()), put the curve $\hat{Y}$ versus X in the previous graph and state again whether the lottery looks random.

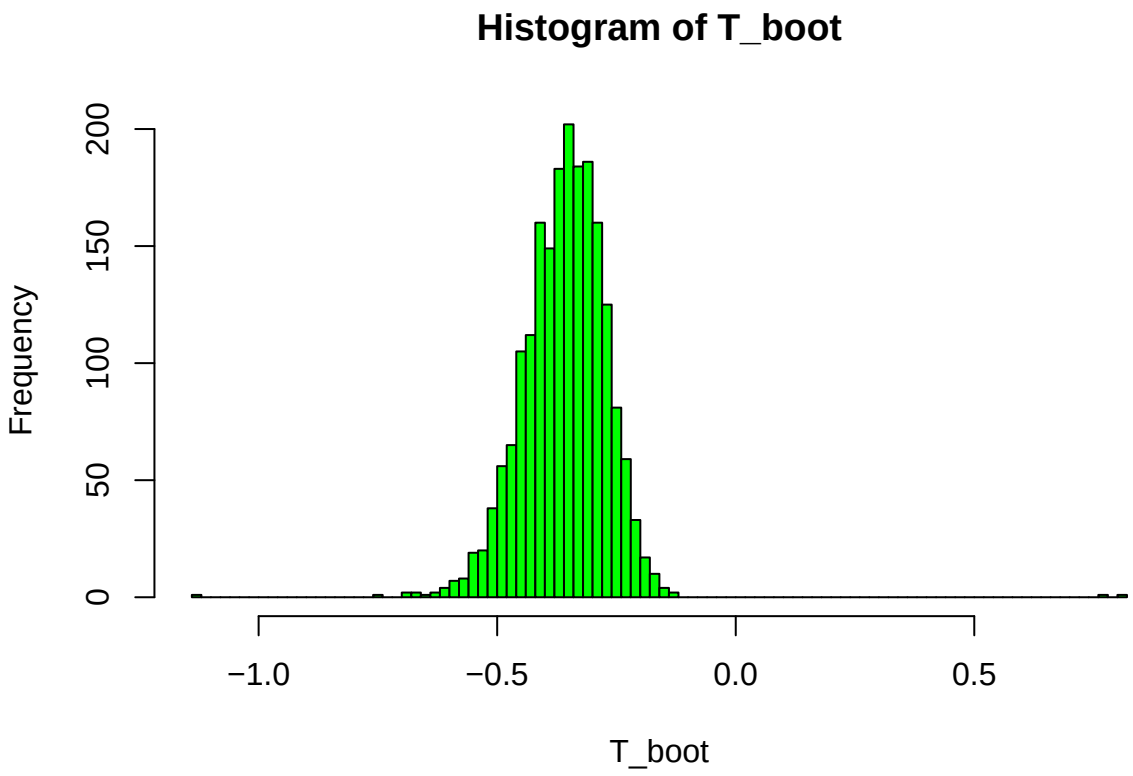


Looking at the graph, we see a declining trend alongwith blue line (computed model) especially for the second half of the year, which establishes a relationship between the date and the draft number. This indicate that there could be some process uncertainty and therefore the lottery draw was not truely random.

3.

To check whether the lottery is random, it is reasonable to use test statistics $T = \frac{\hat{Y}(X_b)-\hat{Y}(X_a)}{X_b-X_a}$, where $X_b = argmax_x Y(X), X_a = argmin_x Y(X)$

If this value is significantly greater than zero, then there should be a trend in the data and the lottery is not random. Estimate the distribution of T by using a non-parametric bootstrap with B = 2000 and comment whether the lottery is random or not. What is the p-value of the test?




```
## The pvalue : 0.001
```

Let us assume that the Lottery is random given by null hypothesis.Hence, our hypothesis will be defined as under : Null hypothesis, $H_0 : t \leq 0$ Lottery is random alternative , $H_a : t = 0$ Lottery is not random

Let the significance level (widely considered unless specified otherwise) is as below: significance level = 5 % (or 0.05)

The calcultated p_value = 0.001.

Since, we got very low P value which is much less than the significance level, therefore we shall reject the hypothesis and will conclude that the lottery was not random, which is opposite to our initial assumption.

Also, looking at the distribution of T-values on histogram, it suggests that most of the T-values are negative or less than zero (only few are at extreme right of plot).It further give us the impression that the probability of T=0 is very less. However, upon further computing the P values (from few positive T values), we get very less value (reject our null hypothesis) which finally confirms that there was no evidence that the lottery was random.

1.4 Implement a function depending on data and B that tests the hypothesis

H_0: Lottery is random versus H_1: Lottery is non-random by using a permutation test with statistics T. The function is to return the p-value of this test. Test this function on our data with B = 2000.

```
## P value : 0.154
```

Upon implementing the function and testing it with B=2000, we have got the p-value of 0.154. This is much higher than the limit of 0.05 (or 5% significance level). It is therefore, we can not reject the null hypothesis here. It does not dismiss the assumption that the lottery was random.

5.

Make a crude estimate of the power of the test constructed in Step 4: (a) Generate an obviously non(random) dataset with n = 366 observations by using same X as in the original data set and $Y(x) = max(0, min(\alpha x + \beta, 366))$, where $\alpha = 0.1$ and $\beta \sim \mathcal{N}(183, sd = 10)$. (b) Plug these data into the permutation test with B = 200 and note whether it was rejected.

```
## The p value is  0.005
```

With B=200 and plugging the data in permutatioin test, We got p value as 0.005 which is very near to zero(0).

(c) Repeat Steps 5a-5b for $\alpha = 0.2, 0.3, \cdots, 10$. What can you say about the quality of your test statistics considering the value of the power?

```
## P value is  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

We got zero (0) p-vale for all 100 α values i.e sequence(0.1 to 10) differ by a constant interval of 0.1. As the p-value for all alpha is below 0.05 (and moreover 0), we shall reject the null hypothesis for each one.

From this we can not reject our initial assuption that lottery was random.

Since the power is the probability that it rejects the null hypothesis and we are getting 100% correct rejection here i.e probability =1, for each alpha case.The p-value is also zero (well within 0.05 band), we can therefore conclude that the quality of our test statistics are very good.

Question 2: Bootstrap, jackknife and confidence intervals

The data you are going to continue analyzing is the database of home prices in Albuquerque,1993. The variables present are Price;Sqft: the area of a house; FEATS: number of features such as dishwasher, refrigerator and so on; Taxes: annual taxes paid for the house. Explore the file prices1.xls.

1.

Plot the histogram of Price. Does it remind any conventional distribution? Compute the mean price.

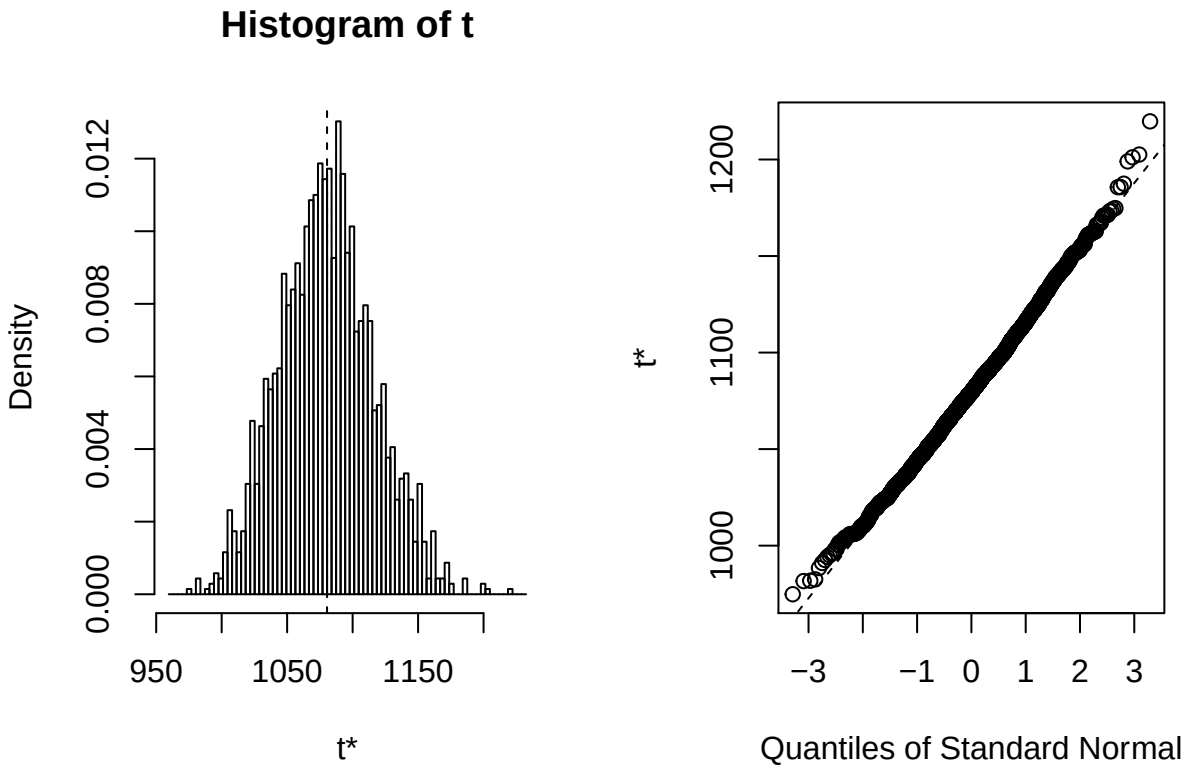

```
## The mean price is  1080.473
```

It histogram shows a decreasing treand towards right , with a steep climb on left. So, the graph is bit like a chi-square distribution (special case of gamama distribution) and follows for almost 75% of trend because the tail on right again picks up unlike the normal chi-square.

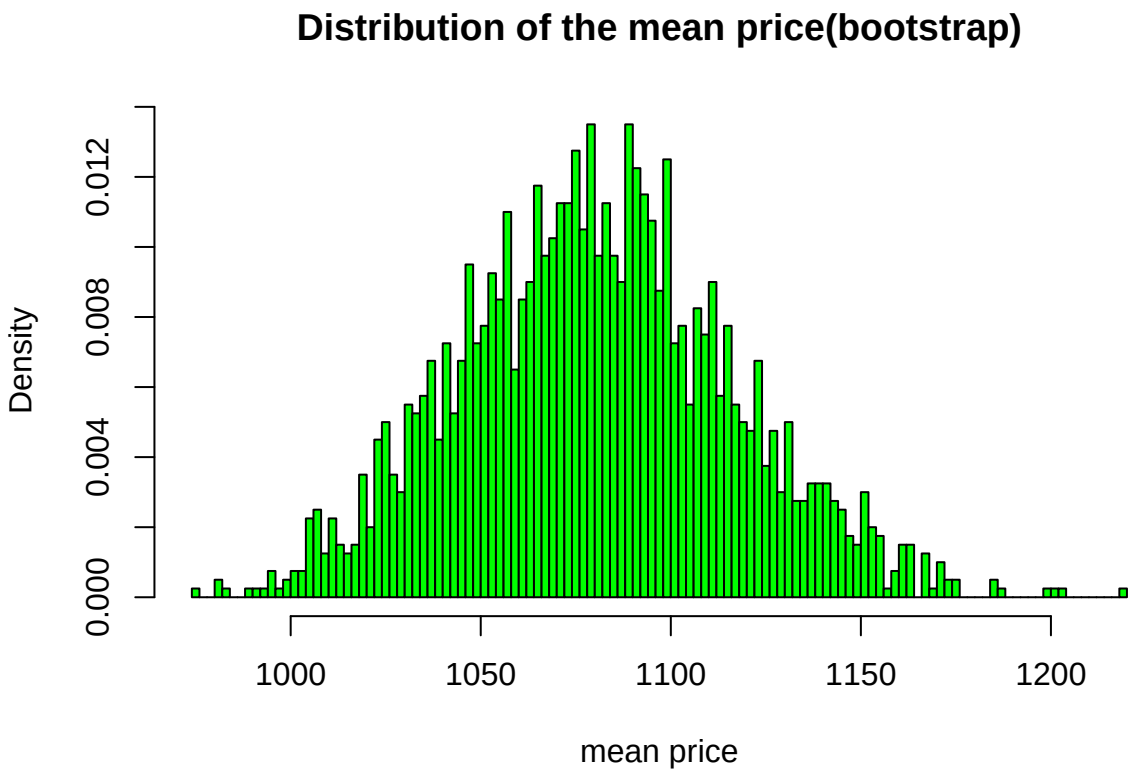
In short, it has 75% resemblance to chi-square. We checked many but are not sure if it matches any standard graph or distribution type.

2.

Estimate the distribution of the mean price of the house using bootstrap. Determine the bootstrap bias correction and the variance of the mean price. Compute a 95% confidence interval for the mean price using bootstrap percentile, bootstrap BCa, and first order normal approximation.(Hint: use boot(), boot.ci(), plot.boot(), print.bootci())




```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = data, statistic = f_boot, R = 2000)
##
##
## Bootstrap Statistics :
##     original     bias    std. error
## t1* 1080.473 -0.2072727    35.83778
```




```
## Bias-correction is given by  1080.68

## The variance of mean price: 1284.346

## [1] "The bootstrap percentile, BCa and first order normal approximation is given by  :"

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_res, type = c("perc", "bca", "norm"))
##
## Intervals :
## Level      Normal             Percentile            BCa
## 95%   (1010, 1151 )   (1011, 1153 )   (1017, 1157 )
## Calculations and Intervals on Original Scale
```

3.

Estimate the variance of the mean price using the jackknife and compare it with the bootstrap estimate

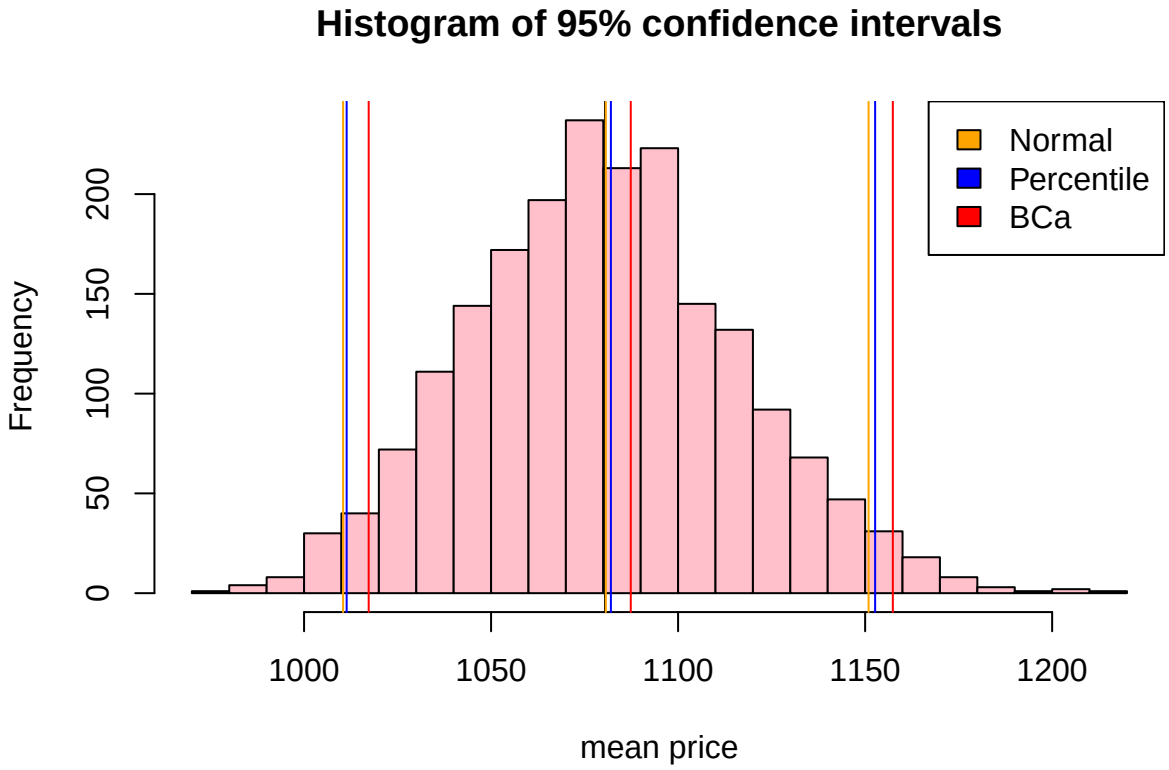
```
## The mean price variance by Jackknife is : 1320.911

## The difference in variance between Jackknife and bootstrap is 36.56461
```

Using jackknife we got variance = 1320.911 as against 1284.346 reported by bootstrap. The difference between both variance is approx 36. This suggests that variance of jackknief is approx. 2.8% higher than bootstrap.

4.

Compare the confidence intervals obtained with respect to their length and the location ofthe estimated mean in these intervals.

```
## Warning in boot.ci(boot.out = boot_res): bootstrap variances needed for
## studentized intervals
```



Herein we have plotted all three mean price related to bootstrap percentile, bootstrap BCa, and first order normal approximation alongwith the confidence band.

Looking at the histogram and the confidence interval, we can say that Normal and percentile are narrowly following each other. The BCa value are far apart. The length of all intervals are almost the same.

The general trend is that the normal interval is lowest which is followed by percentile interval and BCa is at highest level.

We have further marked the oveall mean of price in black.The most nearest to the same is Normal mean marked at almost middle of histogram in discussion.

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE)
library(readxl)
library(boot)
library(bootstrap)

data=read_excel("lottery.xls")
plot(x=data$Day_of_year,y=data$Draft_No ,
     main="Scatterplot (Y versus X)",xlab = "Day_of_year",ylab="Draft_No")

X<-data$Day_of_year
Y<-data$Draft_No
dataframe<-data.frame(X=data$Day_of_year,Y= data$Draft_No)
# using loess() as per instruction for model and plotting
loess_res<-loess(Y~X)
plot(X,Y,main="Lottery data with loess",xlab="Day of year", ylab="Draft number" )
points(X,loess_res$fitted,col="blue",type="l")


RNGversion("3.6.2")
set.seed(12345, kind = "Mersenne-Twister", normal.kind = "Inversion")

# As stated for Xa and Xb for maximum and minimum (out of 366 outcomes)respectively
# we choose accordingly
t_stat<-function(data_new){
  # X and Y dependent on data_new we provide for generic function
X<-data_new$Day_of_year
Y<-data_new$Draft_No
# calculating loess model inside the function
loess_res<-loess(Y~X, data=data_new)
#index of minimum
Xa <- which.min(loess_res$fitted)
Xa_min_day<-X[Xa]
#index of maximum
Xb <- which.max(loess_res$fitted)
Xb_max_day<-X[Xb]
Yhat_Xa <- loess_res$fitted[Xa]
Yhat_Xb <- loess_res$fitted[Xb]
# As per given equation, we calculate T
T_value<- (Yhat_Xb - Yhat_Xa)/(Xb_max_day -Xa_min_day)
return(T_value)
}


# Estimation of  the distribution of T by using a non-parametric bootstrap
B<- 2000 # given
T_boot<-c()
n_row<-nrow(data)
i=1
while(i<=B){
  #sampling of data
  row_sample<-sample(n_row,n_row,replace=TRUE)
```

```
  data_new<-data[row_sample,]
  T_boot[i]<-t_stat(data=data_new)
  i=i+1
}
hist(T_boot,100,col="green")


# for calculating  p value from distribution

count<-0
for (i in 1:length(T_boot)){
  if(T_boot[i]>0){
    count=count+1
  }
}
p_value <-(count/B)
cat("The pvalue :", p_value)

# Building up a permutation function depending on data and B .
RNGversion("3.6.2")
set.seed(12345, kind = "Mersenne-Twister", normal.kind = "Inversion")
permutation_test<-function(data,B){
  loess_res<-loess(Y~X, data = data )
  Y_fit<-loess_res$fitted
  Y_max<-match(max(Y_fit),Y_fit)
  Y_min<-match(min(Y_fit), Y_fit)
  Xb<-data$X[Y_max]
  Xa<-data$X[Y_min]
  T_original<-(max(Y_fit)-min(Y_fit))/(Xb -Xa)
  T_permutation<-c(length=B)
  n<-nrow(data)
  for(b in 1:B){
      data$Y<-sample(data$Y,n)
      loess_res<-loess(Y~X,data)
      Y_fit<- fitted(loess_res )
      Y_max<-match(max(Y_fit),Y_fit)
      Y_min<-match(min(Y_fit),Y_fit)
      Xb<-data$X[Y_max]
      Xa<-data$X[Y_min]
      T_permutation[b]<-(max(Y_fit)-min(Y_fit))/(Xb -Xa)
      }
  permutation_list<-list(t_statistics=T_permutation,
                         p_value=mean(abs(T_permutation )>abs(T_original)))
  return(permutation_list)
}
final_res<-permutation_test(data=dataframe,B=2000)
t_stat<-final_res$t_statistics
cat("P value :",final_res$p_value)

# a. Generate an obviously non(random) dataset with given conditions
RNGversion("3.6.2")
set.seed(12345, kind = "Mersenne-Twister", normal.kind = "Inversion")
n<-366  #given
```

```r
alpha<- 0.1 #given
Y<-c() # to be calculated
i=1
while(i<=length(X)){
beta<- rnorm(1,mean=183,sd=10)  # given Beta is a normal distribution
equation<- min(alpha*X[i]+beta,366)
Y[i]<-max(0,equation)  # the given maximum condition
i=i+1
}
data<-data.frame(X,Y)  # dataframe of X and calculated Y
#plot(X,Y) # to check on the graph

#b. Plug data in permutation_test, B=200
RNGversion("3.6.2")
set.seed(12345, kind = "Mersenne-Twister", normal.kind = "Inversion")
result<-permutation_test(data,200)
cat("The p value is ",result$p_value)

# for sequence of alpha , the step 5a and 5b need to be repeated
RNGversion("3.6.2")
set.seed(12345, kind = "Mersenne-Twister", normal.kind = "Inversion")
alpha_sequence<-seq(0.1,10,by=0.1)
len_seq<-1:length(alpha_sequence)
p_value<-c()
for(seq in alpha_sequence){
   n<-366  #given
   alpha<- seq #given
   Y<-c() # to be calculated
   i=1
   while(i <=length(X)){
   beta<- rnorm(1,mean=183,sd=10)  # given Beta is a normal distribution
   equation<- min(alpha*X[i]+beta,366)
   Y[i]<-max(0,equation)  # the given maximum condition
   i=i+1
   }
   # the new data frame with new value of Y from loop above
   data<-data.frame(X,Y)
   # checking p value inside ongoing each alpha and saving out of loop for respective p value
   result<-permutation_test(data,200)
   p_val<-result$p_value
   p_value[len_seq]<-p_val
}
cat("P value is ",p_value)

data=read_excel("prices1.xls")
hist(data$Price,xlab = "price",main="histogram of price",col = "green")
cat("The mean price is ",mean(data$Price))
mean_X<-mean(data$Price)

RNGversion("3.6.2")
set.seed(12345, kind = "Mersenne-Twister", normal.kind = "Inversion")
f_boot<-function(data,indices){
data<-data[indices,]
```

```r
f_boot<-mean(data$Price)
return(f_boot)
}

#using boot
boot_res<- boot(data=data,statistic=f_boot,R=2000)
plot(boot_res)
print(boot_res)
# Distribution of the mean price(bootstrap)
hist(boot_res$t,100,main="Distribution of the mean price(bootstrap)",
     col="green",probability = TRUE,xlab="mean price")

# Bias-correction is given by :
bias_correction <- 2 *(mean(data$Price)) - mean(boot_res$t)
cat("Bias-correction is given by ",bias_correction,"\n")

# variance of the mean price
boot_variance<-var(boot_res$t)
cat("The variance of mean price:",boot_variance,"\n")

#
boot_table <- boot.ci(boot_res, type = c("perc","bca","norm"))
print("The bootstrap percentile, BCa and first order normal approximation is given by  :")
print(boot_table)



jack_res<-jackknife(x=1:nrow(data),theta=f_boot,data=data)
std_error<-jack_res$jack.se
jack_variance<-std_error^2  # variance is square of standard error
cat("The mean price variance by Jackknife is :",jack_variance,"\n")

difference_variance <-jack_variance-boot_variance
cat("The difference in variance between Jackknife and bootstrap is",difference_variance,"\n")

# Representation of all components in histogram that we have calculated for boot_table
hist(boot_res$t, main="Histogram of 95% confidence intervals", xlab="mean price",col="pink",30)
Confidence_intervals<- boot.ci(boot.out = boot_res)
# using abline functions to add straight line on histogram and marking all mean for all three
# v parameter of abline adds x-value(s) for vertical line(s)
abline(v = mean(data$Price), col = "black",pch=5) # the overall mean
abline(v = Confidence_intervals$normal[2:3], col = "orange")
abline(v = mean(Confidence_intervals$normal[2:3]), col = "orange")
abline(v = Confidence_intervals$percent[4:5], col = "blue")
abline(v = mean(Confidence_intervals$percent[4:5]), col = "blue")
abline(v = Confidence_intervals$bca[4:5], col = "red")
abline(v = mean(Confidence_intervals$bca[4:5]), col = "red")
legend(x = "topright", legend = c("Normal", "Percentile", "BCa"), fill = c("orange", "blue", "red"))
```