University of Chicago
Booth School of Business
41000: Business Statistics, Autumn 2021: Homework Assignment 2. Due in Week 5

Problem 1: Bayes Gold and Silver Coins

A chest has two drawers. It is known that one drawer has 3 gold coins and no silver coins. The other drawer is known to contain 1 gold coin and 2 silver coins.

You don't know which drawer is which. You randomly select a drawer and without looking inside you pull out a coin. It is gold.

(a) Show that the probability that the remaining two coins in the drawer are gold is 75%.

Solution *Suppose drawer A contains 3G and drawer B contains 1G2S.*

We know the probability $P(G \mid A) = 1$ *and* $P(G \mid B) = 1/3$.

Also, it must be either of two drawers, so P(A) = P(B) = 1/2.

What we are are looking for is $P(A \mid G)$*: the probability that it is drawer A.*

By Bayes' rule,

$$\begin{aligned} P(A \mid G) &= \frac{P(G \mid A) * P(A)}{P(G)} \\ &= \frac{P(G \mid A) * P(A)}{P(G \mid A) * P(A) + P(G \mid B) * P(B)} \text{ (Law of total probability)} \\ &= \frac{1 * \frac{1}{2}}{1 * \frac{1}{2} + \frac{1}{3} * \frac{1}{2}} \\ &= \frac{3}{4} \end{aligned}$$

The probability that it is drawer A is 75%.

Problem 2: The Monty Hall Problem.

This problem is named after the host of the long-running TV show, *Let's Make a Deal*.

A contestant is given a choice of 3 doors. There is a prize (a car, say) behind one of the doors and something worthless behind the other two doors (say two goats).

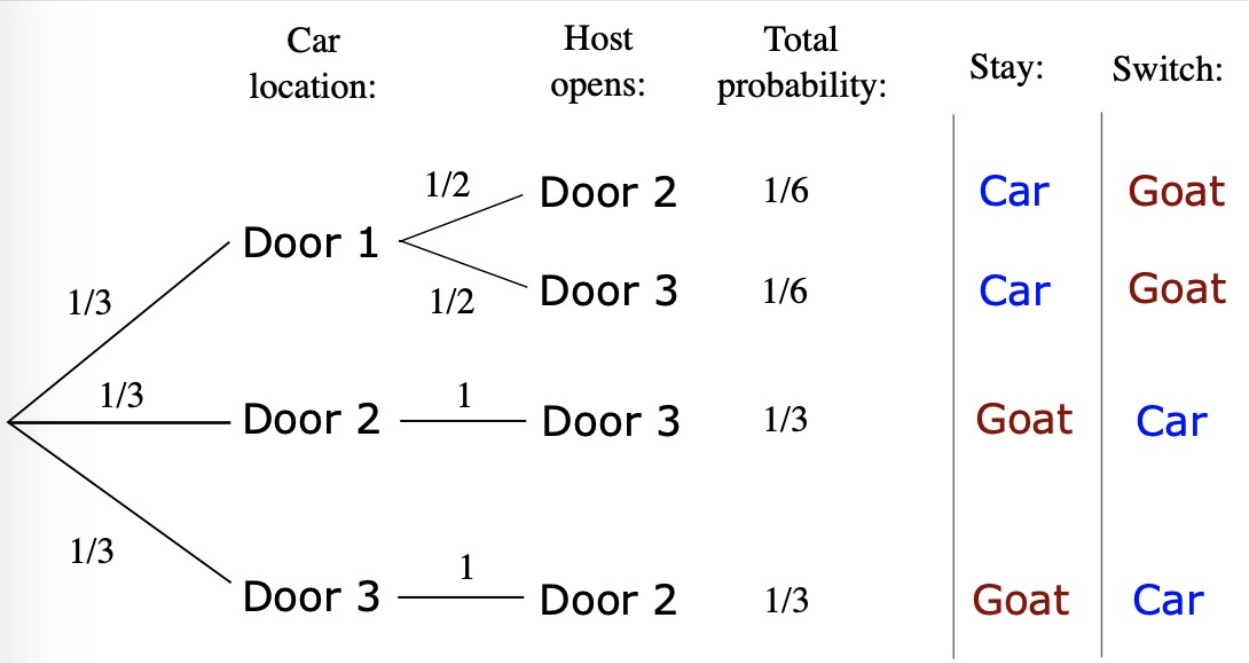
After the contestant chooses a door Monty opens one of the other two doors, revealing a goat.

(a) The contestant has the choice of switching doors. Is it advantageous to switch doors or not?

There is a clear discuss about Monty Hall problem on Wikipedia

Solution

1. *One easy solution by drawing trees*



Because you choose door randomly and the car can be behind each door randomly, it's equivalent to consider the case you ***always*** *select the first door and car is random. If car is behind the first door, Monty will open door 2 or door 3 randomly with equal probability. But if car is not behind the first door, Monty only has one choice. From the tree we see that the probability of winning if you switch is* $\frac{1}{3} + \frac{1}{3} = \frac{2}{3}$.

2. *More complicated stuff. Suppose you plan to switch:*

   You can either pick a winner, a loser, or the other loser when you make your first choice. Each of these options has a probability of $1/3$ *and are marked by a "1" below. In each case, Monty will reveal a loser (X). In the first case, he has a choice, but whichever he reveals, you will switch to the other and lose. But in the other two cases, there is only one loser for him to reveal. He must reveal this one, leaving only the winner. So, if you initially pick a loser, you will win by switching. That is, there is a* $2/3$ *chance of winning if you use the switch strategy.*

	W	L	L
1/3	1	X	
1/3		1	X
1/3		X	1

3. *Not convinced: Try thinking about it this way:*

Imagine the question with 1000 doors, and Monty will reveal 998 wrong doors after you pick one, so you are left with your choice, and one of the remaining 999 doors. Now do you want to stay or switch?

Again, suppose you are going to switch. Define the events

$$W = \textit{the one you switch to is a winner}$$

$$FW = \textit{your first choice is a winner}$$

$$FL = \textit{your first choice is a loser}$$

Since you must pick either a winner or loser with your first choice, and cannot pick both, FW and FL are mutually exclusive and collectively exhaustive. By the rule of total probability:

$$P(W) = P(W \cap FW) + P(W \cap FL) = P(W|FW)P(FW) + P(W|FL)P(FL)$$

$$P(W) = 0 \times \frac{1}{3} + 1 \times \frac{2}{3} = \frac{2}{3}$$

Why is $P(W|FW) = 0$? Because if we choose correctly, and we do switch, we must be on a loser.

Why is $P(W|FL) = 1$? If we first picked a loser, and then switched, we will now have a winner. These both come from above.

There's a longer explanation: Suppose you choose door A and Monty opens B. Consider the events

$$A = \textit{prize is behind A}$$

$$B = \textit{prize is behind B}$$

$$C = \textit{prize is behind C}$$

$$MA = \textit{Monty opens A}$$

$$MB = \textit{Monty opens B}$$

$$MC = \textit{Monty Opens C}$$

Before we choose, each door was equally likely: $P(A) = P(B) = P(C) = 1/3$.

In this case, we know Monty opened B. To decide whether to switch, we want to know if $P(A|MB)$ and $P(C|MB)$ are the same. If they are, then there is no gain in switching:

$$P(A|MB) = \frac{P(A \cap MB)}{P(MB)} = \frac{P(MB|A)P(A)}{P(MB)}$$

We need these components: $P(A) = 1/3, P(MB|A) = 1/2$. This is because you picked A, and Monty can open either B or C. He cannot open A. He can open B or C because the condition in this conditional probability is that the prize is actually behind A.

$$\begin{aligned} P(MB) &= P(MB \cap A) + P(MB \cap B) + P(MB \cap C) \\ &= P(MB|A)P(A) + P(MB|B)P(B) + P(MB|C)P(C) \\ &= \frac{1}{2} \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = \frac{1}{6} + \frac{1}{3} = \frac{1}{2} \end{aligned}$$

I already showed that Pr(MB|A)=1/2. We have that Pr(MB|B)=0 because Monty cannot reveal B if this is where the prize is. If the prize is behind C, he must open B, since you have picked A.

Putting it all together: $P(A|MB) = \frac{P(MB|A)P(A)}{P(MB)} = \frac{\frac{1}{2}\times\frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$

Similarly for C: $P(C|MB) = \frac{P(MB|C)P(C)}{P(MB)} = \frac{1\times\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$

So we are always better off switching!

Problem 3: Descriptive Statistics in R

Download the `superbowl1.txt` and `derby.csv` datasets from the course web-page. The Superbowl contains data data on the `outcome` and the `spread` of all previous Superbowls. The `outcome` is defined as the difference in scores of the favourite minus the underdog. The `spread` is the bookmakers' prediction of the outcome before the game begins.

The Derby dataset consists of all of the results on the `Kentucky Derby` which is run on the first Saturday in May every year at Churchill Downs racetrack.

Install `R` and `RStudio` and answer the following questions:

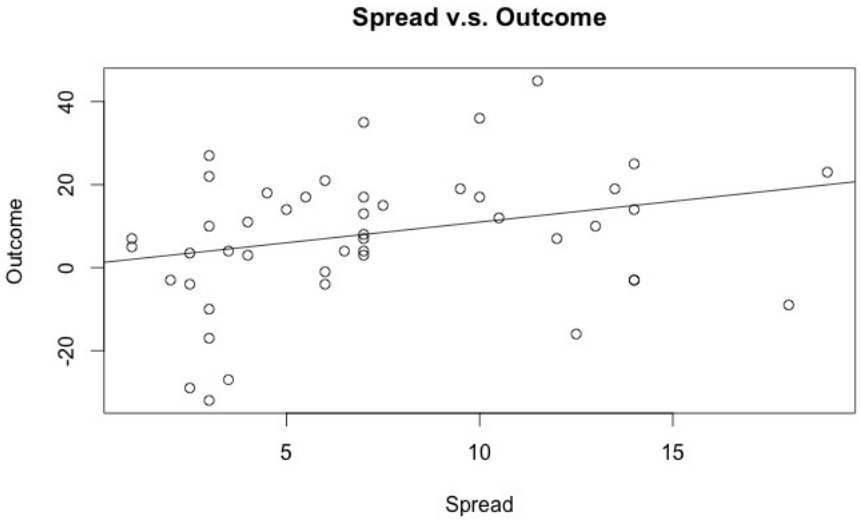
- For the Superbowl data.
  1. Plot the `spread` and `outcome` variables. Calculate means, standard deviations, covariances, correlations and alpha and beta's.
  2. Use a `boxplot` to compare the favourites' score versus the underdog.
  3. Does this data look normally distributed?
- For the Kentucky Derby data.
  (a) Plot a histogram of the winning speeds and times of the horses.

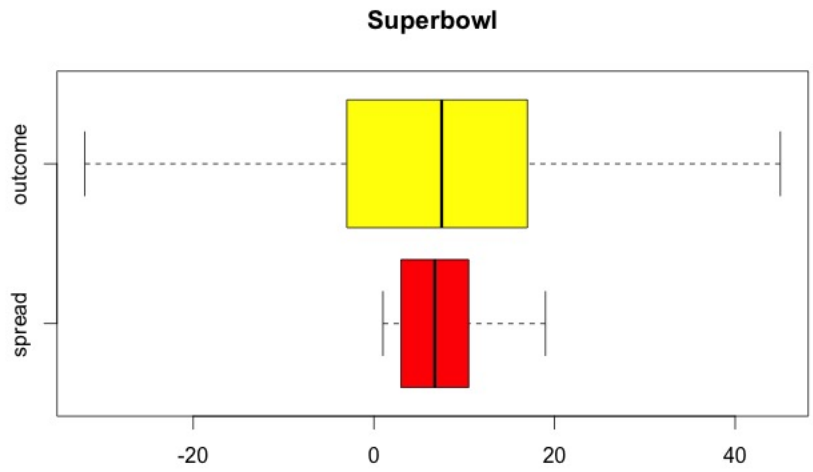
  Why is there a long right-hand tail to the distribution of times?

  (b) Can you identify the outlying horse with the best winning time?

Solution

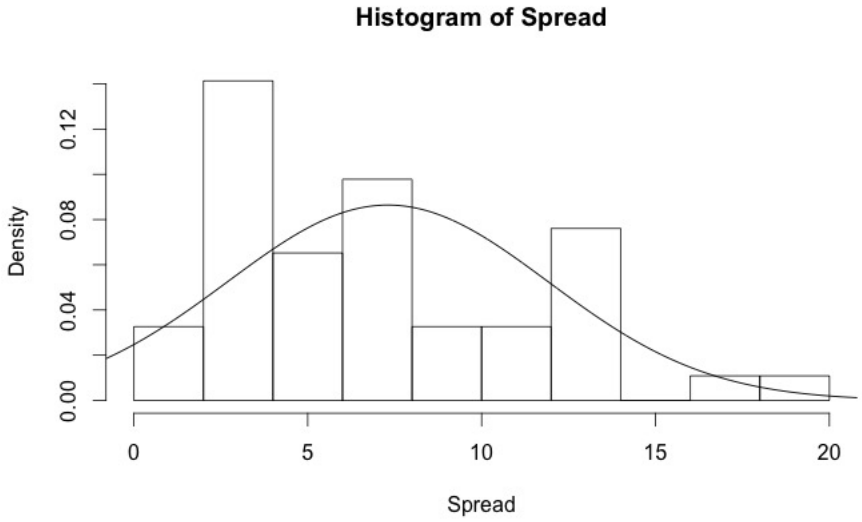
Superbowl






```
> x <- Spread
> y <- Outcome
> mean(x)
[1] 7.304348
> mean(y)
[1] 7.336957
> sd(x)
[1] 4.616971
> sd(y)
[1] 16.14932
> cov(x,y)
[1] 19.82295
> cor(x,y)
[1] 0.2658623
> beta <- cor(x,y)*sd(y)/sd(x)
> alpha <- mean(y)-beta*mean(x)
> beta
[1] 0.9299377
> alpha
[1] 0.5443683
```

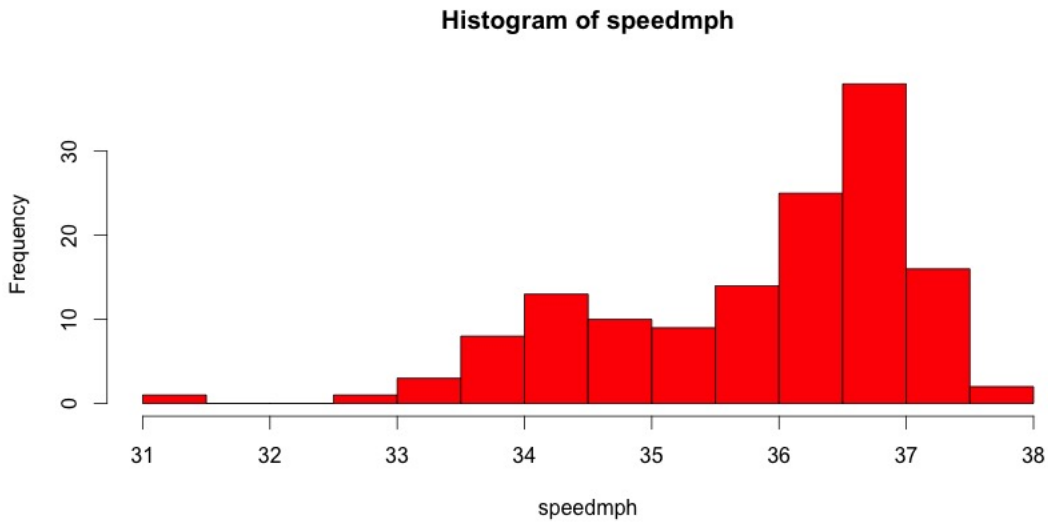
The normal distribution plots are added with sample mean and standard deviation of Outcome and Spread. Compare to the corresponding normal distribution plots, we can conclude that the Outcome is pretty normally distributed but the Spread is not.

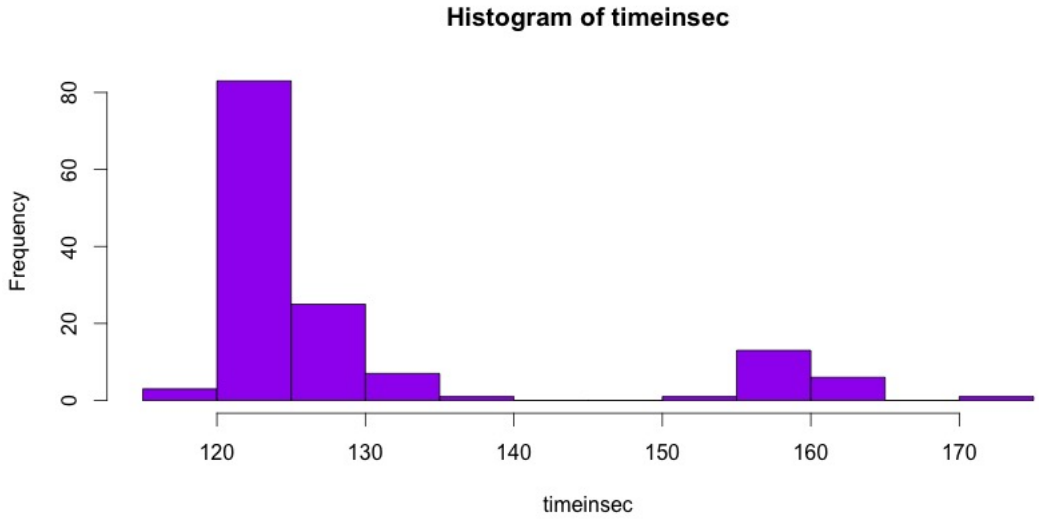


Kentucky Derby:

Here we can find the the outlying horses.

```
> ####################
> # part 2
> ####################
> # to find the left tail observation
> k1 <- which(speedmph == min(speedmph))
> mydata2[k1,]
   year year_num         date  winner mins  secs timeinsec distance speedmph
17 1891       17 May 13, 1891 Kingman    2 52.25    172.25      1.5  31.3498
>
> # to find the best horse
> k2 <- which(speedmph == max(speedmph))
```






```
> mydata2[k2,]
   year year_num     date      winner mins secs timeinsec distance speedmph
99 1973       99 5-May-73 Secretariat    1 59.4     119.4     1.25  37.6884
```

```
####################
## Superbowl
####################

# import superbowl data
mydata1 <- read.table("http://faculty.chicagobooth.edu/nicholas.polson/
teaching/41000/superbowl1.txt",header=T)

# look at data
head(mydata1)
summary(mydata1)

# attach so R recognizes  each variable
attach(mydata1)

######################
# part 1
######################
# plot Spread vs Outcome
plot(Spread,Outcome,main="Spread v.s. Outcome")
# add a 45 degree line to compare
abline(1,1)

# Covariance, Correlation, alpha, beta
x <- Spread
y <- Outcome

mean(x)
mean(y)
sd(x)
sd(y)
cov(x,y)
cor(x,y)
beta <- cor(x,y)*sd(y)/sd(x)
alpha <- mean(y)-beta*mean(x)
beta
alpha

# Regression check
model <- lm(y~x)
coef(model)

####################
# part 2
####################
# Compare boxplot
boxplot(Spread,Outcome,horizontal=T,names=c("spread","outcome"),
        col=c("red","yellow"),main="Superbowl")

######################
# part 3
######################
# see the distribution of outcome and spread through histograms
hist(Outcome,freq=FALSE)
```

```
# add a normal distribution line to compare
lines(seq(-50,50,0.01),dnorm(seq(-50,50,0.01),mean(Outcome),sd(Outcome)))

hist(Spread,freq=FALSE)
lines(seq(-10,30,0.01),dnorm(seq(-10,30,0.01),mean(Spread),sd(Spread)))

######################
## Kentucky Derby
######################

mydata2 <- read.csv("http://faculty.chicagobooth.edu/nicholas.polson/teaching/41000/Kentucky_Derby_2014

# attach the dataset
attach(mydata2)

head(mydata2)
summary(mydata2)

##################
# part 1
##################
# plot a histogram of speedmph
hist(speedmph,col="blue")

# finer bins
hist(speedmph,breaks=10,col="red")
hist(timeinsec,breaks=10,col="purple")

###################
# part 2
###################
# to find the left tail observation
k1 <- which(speedmph == min(speedmph))
mydata2[k1,]

# to find the best horse
k2 <- which(speedmph == max(speedmph))
mydata2[k2,]
```

Problem 4: YahooFinance Berkshire Hathaway

Using the `R` script on the course webpage download daily return data in Warren Buffett's firm Berkshire Hathaway (ticker symbol: `BRK-A`) from 1990 to the present. Using `R` or `RStudio` analyze this data in the following way:

(a) Plot the Historical Price Performance

(b) Calculate the Daily returns. Plot a histogram of the returns. Comment on the distribution that you obtain.

(c) Use the `summary` command to provide statistical data summaries.

Interpret your findings.

Hint: you will find the following `R` commands useful.

```
install.packages("quantmod")
library(quantmod)
```

```
getSymbols("BRK-A", from = "1990-01-01")
```

Solution *Using the daily return data on Berkshire Hathaway's* `BRK-A` *performance we have the following summary statistics for the returns* `ret` *vector.*

```
> summary(ret)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.1209 -0.0058  0.0000  0.0006  0.0063  0.1613
> sd(ret)
[1] 0.0143

> skewness(ret)
[1] 0.868
> kurtosis(ret)
[1] 15
```

On a daily basis, the mean return is 0.06% *or 6 basis points with a standard deviation of* 1.49%.

The returns are slightly positivity skewed with a skewness of 0.868. *However, a zero median implies that returns are 50% positive and 50 % negative.*

Kurtosis greater than 3 means heavy-tail property of the return distribution and there are more extreme price changes compared to normal distribution.

Figure 1 shows the time series plot of stock price and histogram of returns, which is like a random walk. And in the histogram of Figure 2, we can see the heavy tail and positive skewness.

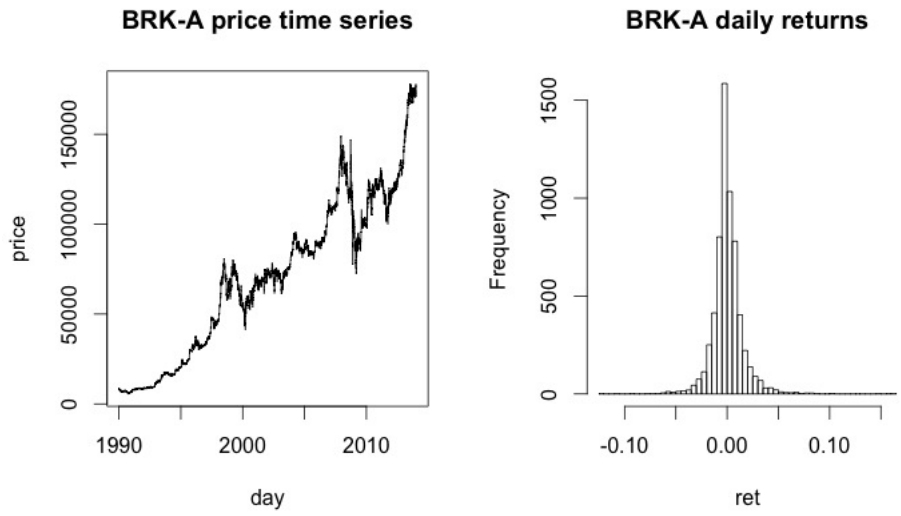


Figure 1: Berkshire Hathaway's Stock Performance

```
# install and import the package fImport
# extract data from yahoo finance
install.packages("quantmod")
install.packages("moments")
library("moments")
library("quantmod")

getSymbols('BRK-A', from = "1990-01-01")
```

```
BRKA = get('BRK-A')
BRKA = BRKA[,4]

# take a look of the data
head(BRKA)

plot(BRKA,type="l",col=20,main="BRKA Share Price",
     ylab="Price",xlab="Time",bty='n')

# calculate the simple return
N <- length(BRKA)
y = as.vector(BRKA)
ret <- rep(NA,N-1) # create a null sequence
for(t in 1:(N-1))
{
  ret[t] <- (y[t+1]-y[t])/y[t]
}

# create summaries of ret for BRK-A
options(digits=3) # control the digit of numbers
summary(ret)
sd(ret)
skewness(ret)
kurtosis(ret)

par(mfrow=c(1,2)) # combine two plots in a 1*2 row

# time series plot of price
# to save the plot,click "Export" ans "save as image"
# create a time series object in R that we can add time domain
y_ts = ts(y, start=c(1990,1,1), end=c(2014,9,30), frequency=252)
ts.plot(y_ts,main="BRK-A price time series",ylab="price",xlab="day")
# histogram of returns
hist(ret,breaks=50,main="BRK-A daily returns")

# plots to show serial correlation in 1st and 2nd moments
# to save the plot,click "Export" ans "save as image"
par(mfrow=c(1,2))
acf(ret,lag.max=10,main="serial correlation in 1st moment")
acf(ret^2,lag.max=10,main="serial correlation in 2nd moment")

dev.off() # stop the above combine function
```

Problem 5: AB Testing

SimCity 5 is one of Electronic Arts (EA's) most popular video games. As EA prepared to release the new version, they released a promotional offer to drive more pre-orders. The offer was displayed on their webpage as a banner across the top of the pre-order page. They decided to test some other options to see what design or layout would drive more revenue.

The control removed the promotional offer from the page altogether. The test lead to some very surprising results. With a sample size of 1000 visitors, of the 500 which got the promotional offer they found 143 people wanted to purchase the games and of the half that got the control they found that 199 wanted to buy the new version of SimCity.

Test at the 1% level whether EA should provide a promotional offer or not

Solution *To see whether the promotion program works, we need to find out whether the proportion of people who want to buy the game is larger. Denote p_1 is the proportion from the experimental group with promotion and p_2 from the control group without promotion. Therefore, the null hypothesis we are testing is*

$$H_0 : p_1 > p_2$$

In the sample, we observe

$$n_1 = n_2 = 500$$

$$\hat{p_1} = 199/500 \text{ and } \hat{p_2} = 143/500$$

To perform the two-sample t-test, we need to calculate the pooled sample standard deviation first.

$$\bar{p} = \hat{p_1} * \frac{n_1}{n_1 + n_2} + \hat{p_2} * \frac{n_2}{n_1 + n_2} = 171/500$$

$$\bar{\sigma} = \sqrt{\bar{p} * (1 - \bar{p}) * (\frac{1}{n_1} + \frac{1}{n_2})} = 0.03$$

Then,

$$\text{t-stat} = \frac{\hat{p_1} - \hat{p_2}}{\bar{\sigma}} = 3.73 > qnorm(0.995) = 2.58$$

Now we can conclude that the promotional program works statistically at the 1% level.

Problem 6: Hypothesis Testing

In a recent article it was claimed that "96% of Americans under the age of 50" spent more than three hours a day on Facebook.

To test this hypothesis, a survey of 418 people under the age of 50 were taken and it was found that 401 used Facebook for more than three hours a day.

Test the hypothesis at the 5% level that the claim of 96% is correct.

Solution *We want to test the proportion of people under the age of 50 on Facebook.*

$$H_0 : p = 96\%$$

The estimated proportion from the survey is

$$\hat{p} = 401/418 = 0.959$$

The standard deviation of the sample proportion is

$$\hat{\sigma} = \sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}} = 0.0096$$

To perform a t-test, the corresponding statistic is

$$t\text{-}stat = \frac{0.959 - 0.96}{0.0096} = -0.1$$

It seems that we cannot reject the null hypothesis at the 5% level.

Problem 7: Paired T-Test

The following table shows the outcome of eight years of a ten year bet that Warren Buffett placed with Protege Partners, a New York hedge fund. Buffett claimed that a simple index fund would beat a portfolio strategy (fund-of-funds) picked by Protege over a ten year time frame. At Buffett's shareholder meeting, he provided an update of the current state of the bet. The bundle of hedge funds picked by Protege had returned 21.9% in the eight years through 2015 and the S&P500 index fund had soared 65.7%.

	SP Index	Hedge Funds
2008	-37.0%	-23.9%
2009	26.6%	15.9%
2010	15.1%	8.5%
2011	2.1%	-1.9%
2012	16.0%	6.5%
2013	32.3%	11.8%
2014	13.6%	5.6%
2015	1.4%	1.7%
cumulative	65.7%	21.9%

(a) Use a paired t-test to assess the statistical significance between the two return strategies

(b) How likely is Buffett to win his bet in two years?

Solution *To conduct a paired t-test, we need define a new variable*

$$D = SP.Index - Hedge.Funds.$$

The null hypothesis we want to test is

$$H_0 : \bar{D} = 0$$

The mean and standard deviation of the difference $\{D_i\}_{i=1}^n$ *are*

$$\bar{D} = 0.0574, \quad s_D = 0.0969$$

```
> SP.Index = c(-37,26.6,15.1,2.1,16,32.3,13.6,1.4)
> Hedge.Funds = c(-23.9,15.9,8.5,-1.9,6.5,11.8,5.6,1.7)
> SP.Index = SP.Index/100
> Hedge.Funds = Hedge.Funds/100
> mean(SP.Index-Hedge.Funds)
[1] 0.057375
> sd(SP.Index-Hedge.Funds)
[1] 0.09687243
```

Therefore the t-score is

$$t = \frac{\bar{D}}{s_D\sqrt{1/n}} = \frac{0.0574}{0.0969 \times \sqrt{1/8}} = 1.6755.$$

And the p-value is 0.094

```
> 2*pnorm(1.6755,lower.tail = F)
[1] 0.09383617
```

The p-value is larger than 0.05. Thus we don't reject the null hypothesis at 5% significance level. In other words, these two return strategies are not significantly different.

If we are to use t-distribution instead of Normal, then corresponding p-value is

$$P(|t_7| > 1.6755) = 0.1378$$

```
> 2*pt(1.6755,df=7,lower.tail = FALSE)
[1] 0.1377437
```

df *is degrees of freedom and we use n - 1 for it.*

In R, you can easily perform t-test using the following command, which gives the same test result as above.

```
> t.test(SP.Index,Hedge.Funds,paired = TRUE)

Paired t-test
data:  SP.Index and Hedge.Funds
t = 1.6752, df = 7, p-value = 0.1378
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.02361238  0.13836238
sample estimates:
mean of the differences
               0.057375
```

For the second question, the exact probability of Buffett winning his bet is calculated by

$$P\left(\frac{(1+S_9)(1+S_{10})(1+65.7\%)}{(1+H_9)(1+H_{10})(1+21.9\%)} > 1\right)$$

where S_9, S_{10}, H_9, H_{10} are returns of two strategies in 9th and 10th year. To derive the distribution of the left hand side expression is difficult. Here we instead give a simulation solution, where the returns are assumed to be i.i.d. Normal variables.

```
> m1 = mean(SP.Index)
> s1 = sd(SP.Index)
> m2 = mean(Hedge.Funds)
> s2 = sd(Hedge.Funds)
> N = 1000
> cum1 = NULL # cumulative return of SP.Index
> cum2 = NULL # cumulative return of Hedge.Funds
> set.seed(410)
> for (i in 1:N)
+ {
+   # simulate returns in 2016 and 2017
+   return1 = c(0.657, rnorm(n=2, mean = m1, sd = s1))
+   # compute the cumulative return
+   cum1 = c(cum1, cumprod(1+return1))
+
+   return2 = c(0.219, rnorm(n=2, mean = m2, sd = s2))
+   cum2 = c(cum2, cumprod(1+return2))
+ }
>
> # compute the prob. of Buffett winning
> mean(cum1>cum2)
[1] 0.9296667
```

The estimated probability is 92.97%, indicating that Buffett is very likely to win his bet. Since the paired t - test suggests no significant difference in average returns, you can also estimate the mean by pooling the data and redo the simulation, which gives a smaller estimate (89.27%).

```
m = mean(c(SP.Index,Hedge.Funds))
m1 = m
```

```
m2 = m
...
> mean(cum1>cum2)
[1] 0.8926667
```