STAT 210 - Project Proposal

Fahad Almatrafi
Yongkang Long
Shucheng Cao
Abdullah Alhareth

October 29, 2020

Applied Statistics and Data Analysis

STAT 210

Professor.Joaquin Ortega Sanchez

Smart Beta?
What is the Return of Investing the newly added stocks in S&P 500?

Introduction.

In stock markets, everyone wants to profit as much as possible. Actually, if one invested in SPY, an exchange-traded fund (also known as ETF) that tracks the S&P 500 index, from the day it was funded and stayed in the market, one will get about 7.6% annualized rate of return (0.2 % per day) (see Figure 1).

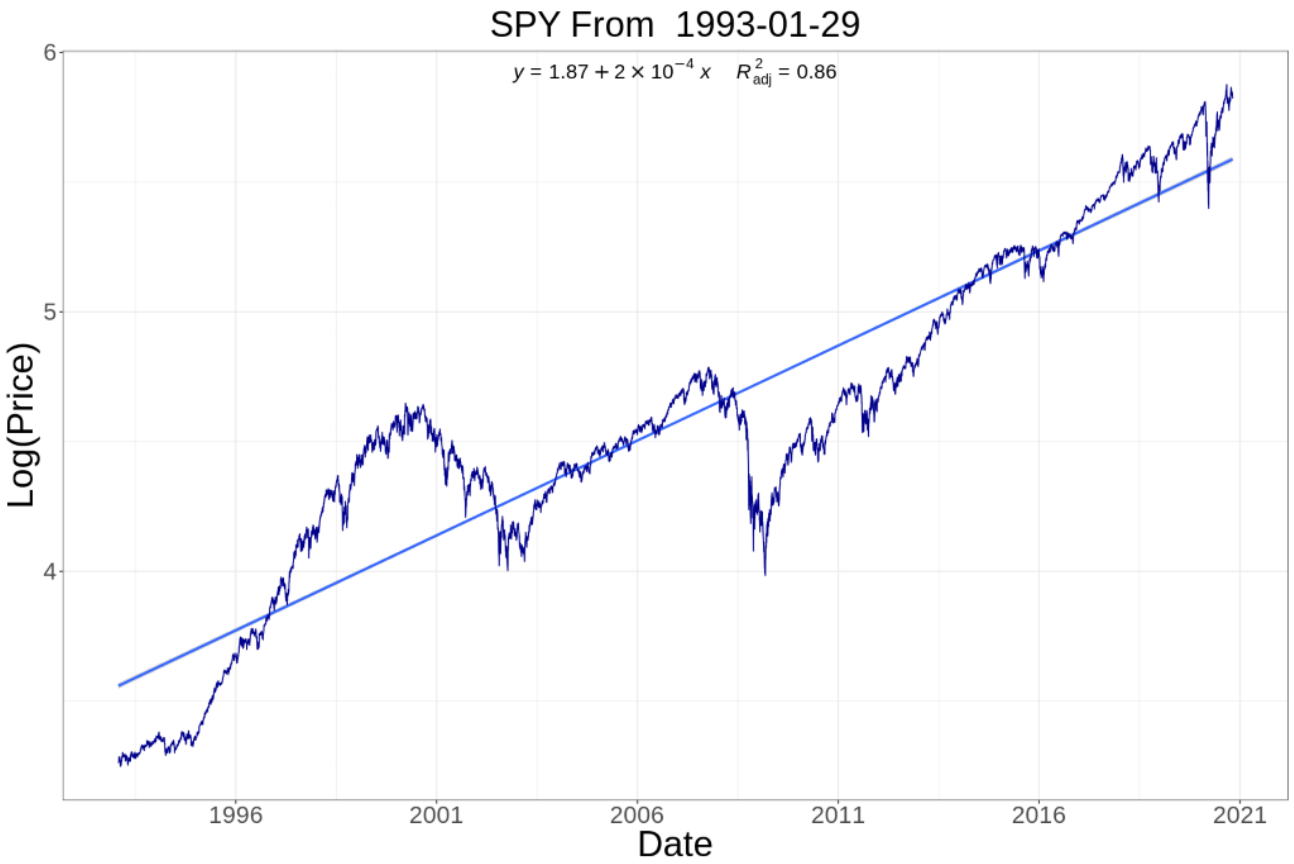

Figure 1..Historical Price of SPY since it was founded

However, some people may not satisfy with the market return; they tried to develop methods to beat the market and earn more money. Unfortunately, historical data showed that most active investment strategies would lose to market in the end. In 2008, Warren Buffet, one of the most successful investors, issued a challenge to the hedge fund industry, saying that no hedge funds can beat an index fund -a fund that invests in all market companies- in the next 10 years. He wins in 2015 because his opponent conceded defeat ahead of the contest's scheduled wrap-up on December 31, 2017. (See figure 2)

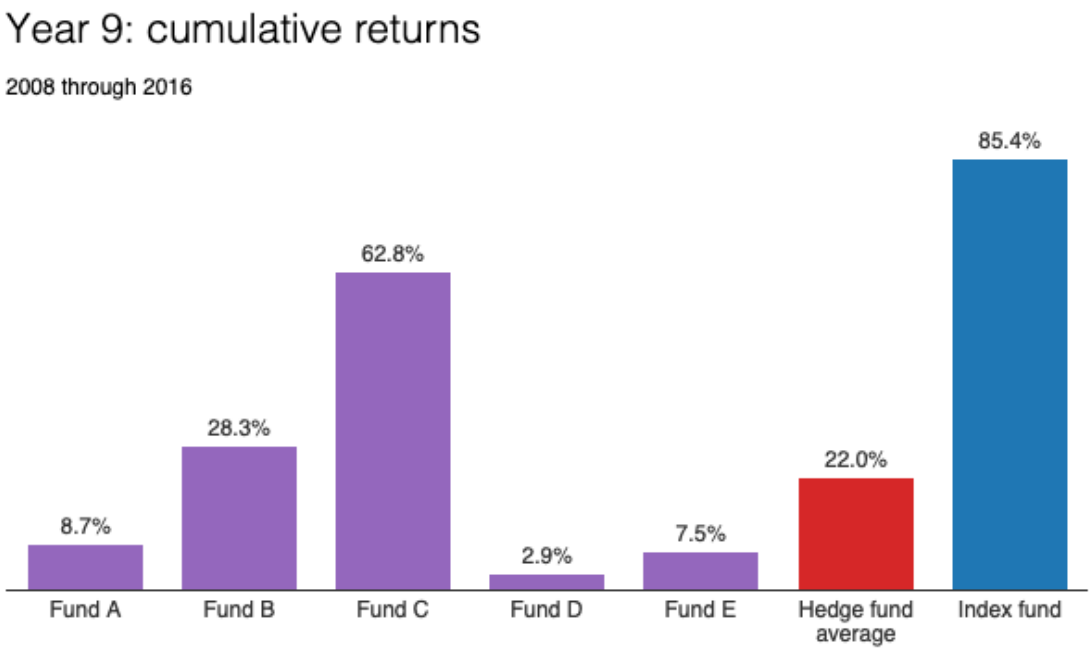

Figure 2.Figure 2.. 9 years cumulative return of 5 hedge funds and Vanguard's S&P 500 Admiral fund Source: BRK 2016 letter.

.

Since most people do not have the ability to know which stocks and when they should be investing and the return of active investment strategies will mostly be lower than the market; therefore, the passive investment strategies are getting more and more popular in recent years. The smart beta strategies are passive intersegment that use alternative index construction rules. For example, use equal weighting instead of market capitalization weighting of the S&P 500 smart beta. Here, we develop a new smart beta strategy that only invests in the newly added stocks in S&P500. We will compare the short-term and long-term returns of our smart beta strategy with SPY. We also measure the volatility to evaluate the risk of our smart beta methods.

Objective.

Our object is to propose a new smart beta strategy that when new stocks are added to S&P 500 companies, we invest in those stokes. We would like to see what is our strategy different from investing in the SPY at the same time. We will use statistical tests to see if the proposed strategies five and ten years return are different from

those of an index fund. We also like to analyze the volatility of the price over time. What's more, we will construct periodic contribution back-test portfolios based on our strategy to analyze and back test portfolio returns, risk characteristics, drawdowns, and rolling returns. Finally, we will get a conclusion that what's the return and risk of our strategy compared to SPY.

Description of the dataset.

In order to compare the newly added companies ROI to the SPY fund, we constructed an R code (see appendix 1) to collect the price data for all the new companies from 1993 until 2020 alongside the SPY prices for the same time period using the R package *quantmod*. Then, the Adjusted price was used for comparison, where we identify the date of joining and calculate a five-year return of investment and a ten-year ROI based on that date and the historical data available. Moreover, another ROI is calculated for the SPY fund for every company based on the joining date of said company. SPY fund historical price data is shown in figure1.

The ROI is calculated according to the formula:

$$ROI = \frac{Profit}{Initial\ cost}$$

And the profit after five years is calculated according to the formula:

$$Profit = Stock\ price\ at\ the\ joining\ day - Stock\ price\ after\ five\ years$$

The ten years profit is calculated in the same way. For a newly added company, the ROI is calculated from the day the company is added, then another complementary ROI is calculated for the SPY fund when investing on the same day.

The dataset constructed using our code is shown in Table1, we used adjusted price for analysis, since the adjusted closing price factors in corporate actions, such as stock splits, dividends, and rights offerings, and therefore it can reflect the truth historical returns.

Table 1 Dataset used in our study

NO	COMPANY	ROI 5	ROI 5,SPY	ROI 10	ROI10,SPY
1	COST	1.365	1.367	2.434	1.608
2	CSCO	11.456	1.803	13.481	1.734
3	KEY	1.698	1.928	2.451	1.932
4	UNM	0.907	1.928	-0.308	1.932
5	MSFT	10.849	2.097	6.952	1.876
…	…	…	…	…	…
…	…	…	…	…	…

The scientific goals, specific hypotheses.

Our goal is to investigate the return on investment (ROI) when investing in SPY funds compared to investing in newly added S&P companies in the short-term (5 years from being added) and the long-term (10 years being added);When we inspect the histograms of the dataset in figure2, we can see that SPY data is more stable then the new companies, when we do a normality test, all groups fail that category. Hence a non-parametric test will be used to compare every two groups.

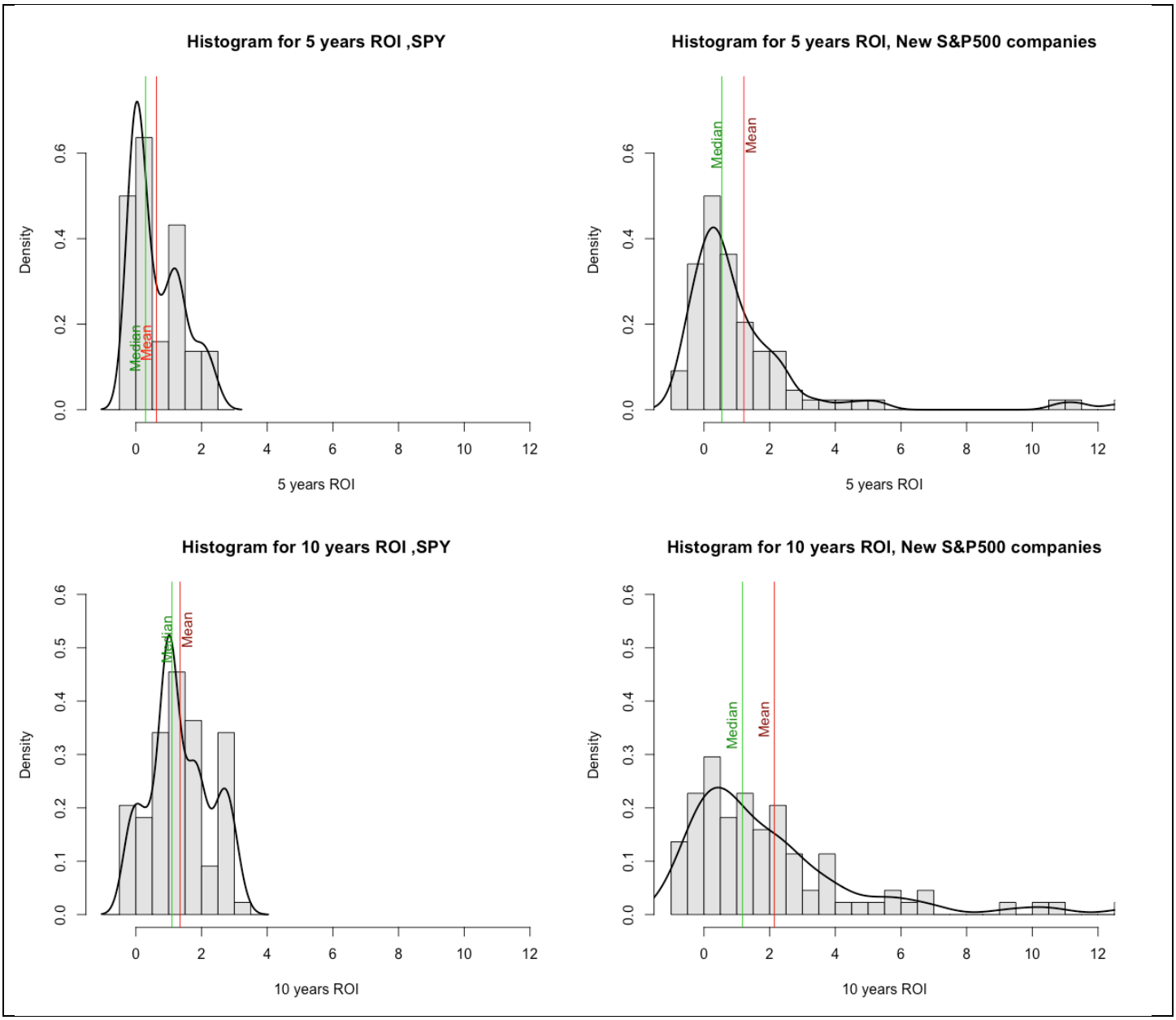

Figure 3. Histograms for five years ROI and ten years ROI. The mean and median are shown in red and green respectively.

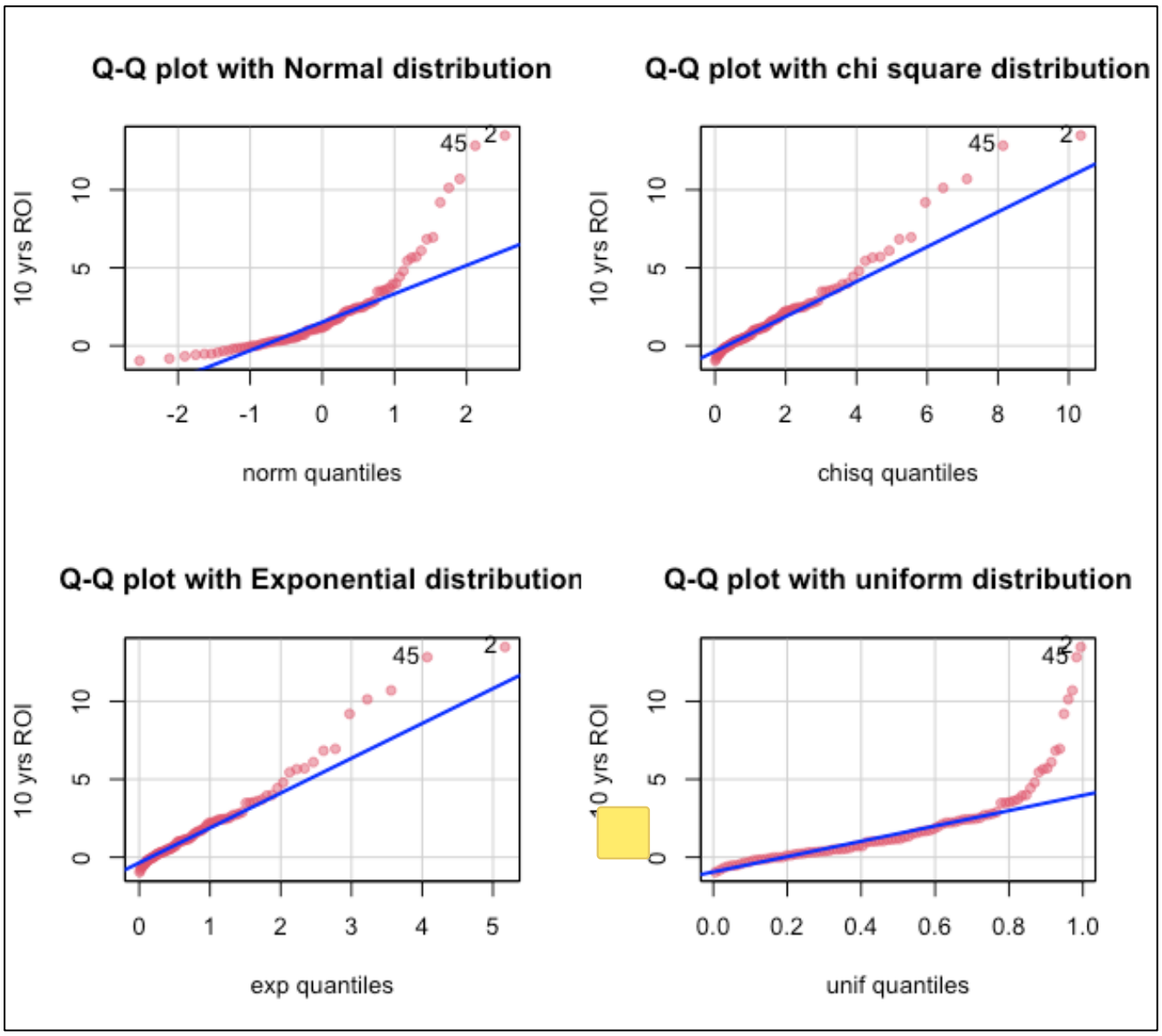

Figure 4. Quantile-Quantile plots for 10 years ROI for the new companies, using different distributions. The uniform distribution is most likely.

Plan for analysis and modeling.

We are planning to continue analyzing the data as shown in the section analysis goal on more data, and applying the tests we had taken in the class, such as the non-parametric tests, and test the homogeneity and ependence of the datasets. Concluding after done many types of tests if the new smart beta strategy can beat the SPY fund or not.

Appendix.

```
library(plyr)
library(quantmod)
library(lubridate)

SP500<-read.csv('sp500_history.txt',header=T)
SP500.added<-SP500[SP500[,'variable']=='added_ticker',] #choose only added tickers after jan 1993
SP500.1993<-SP500.added[SP500.added[,'date']>='1993-02-01',]
SP500.letters<-subset(SP500.1993,select = c(date,value,name))

SPY  = getSymbols("SPY", src = "yahoo", from = "1956-01-01", to = Sys.Date()+1, auto.assign = FALSE)
v <- data.frame("Date" = index( SPY), "Price" =  SPY$SPY.Adjusted)

list.of.companies<-list()
list.of.adjusted<-list()
InvRet<-data.frame(company=NA,ROI5=NA,ROI5_SPY=NA, ROI10=NA,ROI10_SPY=NA)[numeric(0), ]
k=0
for (i in SP500.letters$value) {
  k=k+1
  tryCatch({
    getSymbols(i,src='yahoo',from = "1956-01-01", to = Sys.Date()+1)
    VarName<-paste(i,'1', sep = "")
    yearss<-interval(SP500.letters$date[SP500.letters$value==i], today()) %/% years(1)


    print(i)
    print(yearss[1])
    Dummy<-data.frame(date=index(eval(parse(text = i))), coredata(eval(parse(text = i))))

    FiveYears<-ymd(SP500.letters$date[SP500.letters$value==i]) %m+% years(5)

    FirstPrice<-Dummy[7][Dummy[,'date']==SP500.letters$date[SP500.letters$value==i],]
    FirstPrice.SPY<-v[2][v[,'Date']==SP500.letters$date[SP500.letters$value==i],]
    FivePrice<-Dummy[7][Dummy[,'date']==FiveYears,]
    FivePrice.SPY<-v[2][v[,'Date']==FiveYears,]

    Profit5<-FivePrice-FirstPrice
    Profit5.SPY<-FivePrice.SPY-FirstPrice.SPY
    ROI5<-Profit5/FirstPrice
    ROI5.SPY<-Profit5.SPY/FirstPrice.SPY

    TenYears<-ymd(SP500.letters$date[SP500.letters$value==i]) %m+% years(10)
    TenPrice<-Dummy[7][Dummy[,'date']==TenYears,]
    TenPrice.SPY<-v[2][v[,'Date']==TenYears,]

    Profit10<-TenPrice-FirstPrice
    Profit10.SPY<-TenPrice.SPY-FirstPrice.SPY
    ROI10<-Profit10/FirstPrice
    ROI10.SPY<-Profit10.SPY/FirstPrice.SPY



    de<-data.frame(i,ROI5,ROI5.SPY,ROI10,ROI10.SPY)
    names(de)<-names(InvRet)
    InvRet<-rbind(InvRet,de)

    list.of.companies[[VarName]]<-Dummy[Dummy[,'date']>=SP500.letters$date[SP500.letters$value==i],]
    rm(Dummy,list = i)

  }, error=function(e){}) #tryCatch is to skip error when downloading data
  #list.of.adjusted[[k]]<-(list.of.companies[[k]][c(1,7)])

}
```

```
### PLOTSSS

hist(InvRet$ROI10,freq = F,breaks = seq(-1,ceiling(max(InvRet$ROI10)),0.5),
     main = 'Histogram for 10 years ROI',xlab = '10 years ROI',
     xlim = c(-1,12),ylim = c(0,0.5),col='gray89')
lines(density(InvRet$ROI10),lwd = 2)
abline(v = mean(InvRet$ROI10), col = "red2")
abline(v = median(InvRet$ROI10), col = "green3")
text(mean(InvRet$ROI10),0.4,'Mean',pos=2,srt=90,col='red4')
text(median(InvRet$ROI10),0.4,'Median',pos=2,srt=90,col='green4')

hist(InvRet$ROI10_SPY,freq = F,breaks = seq(-1,ceiling(max(InvRet$ROI10_SPY)),0.5),
     main = 'Histogram for 10 years ROI ,SPY',xlab = '10 years ROI',
     xlim = c(-1,12),ylim = c(0,0.5),col='gray89')
lines(density(InvRet$ROI10_SPY),lwd = 2)
abline(v = mean(InvRet$ROI10_SPY), col = "red2")
abline(v = median(InvRet$ROI10_SPY), col = "green3")
text(mean(InvRet$ROI10_SPY),0.4,'Mean',pos=2,srt=90,col='red4')
text(median(InvRet$ROI10_SPY),0.4,'Median',pos=2,srt=90,col='green4')

qqnorm(InvRet$ROI10, main = 'qq plot for 10 yrs ROI', sub='')
qqline(InvRet$ROI10)


hist(InvRet$ROI5,freq = F,breaks = seq(-1,ceiling(max(InvRet$ROI5)),0.5),
     main = 'Histogram for 5 years Return Of Investment (ROI)',xlab = '5 years ROI',
     xlim = c(-1,12),ylim = c(0,0.75),col='gray89')
lines(density(InvRet$ROI5),lwd = 2)
abline(v = mean(InvRet$ROI5), col = "red2")
abline(v = median(InvRet$ROI5), col = "green3")
text(mean(InvRet$ROI5),0.6,'Mean',pos=4,srt=90,col='red4')
text(median(InvRet$ROI5),0.6,'Median',pos=3,srt=90,col='green4')

hist(InvRet$ROI5_SPY,freq = F,breaks = seq(-1,ceiling(max(InvRet$ROI5_SPY)),0.5),
     main = 'Histogram for 5 years ROI ,SPY',xlab = '10 years ROI',
     xlim = c(-1,12),ylim = c(0,0.75),col='gray89')
lines(density(InvRet$ROI5_SPY),lwd = 2)
abline(v = mean(InvRet$ROI5_SPY), col = "red2")
abline(v = median(InvRet$ROI5_SPY), col = "green3")
text(mean(InvRet$ROI5_SPY),0.2,'Mean',pos=2,srt=90,col='red1')
text(median(InvRet$ROI5_SPY),0.2,'Median',pos=2,srt=90,col='green4')
```