Project #1

Winter Nguyen - htn937

2025-09-22

```r
library(nimble)
```

Problem #1 (55 points)

Download the historical stock prices of *Walmart* for the last 252 (or so) trading days, i.e., for a one-year period from our course website. Do the same for *IBM*.

```r
#stockW is the Walmart stock data,
#StockI is the IBM stock data
stockW=read.csv("/Users/minhtuyennguyen/R/WMT.csv")
stockI=read.csv("/Users/minhtuyennguyen/R/IBM.csv")

#closeW is the Walmart close price, CloseI is IBM close price
closeW=as.numeric(stockW$Close)
closeI=as.numeric(stockI$Close)
```

(5 points) Draw the time-plot of the evolution of the closing stock prices (not the adjusted) for both of the stocks on the **same coordinate system**. You do **not** need to put the calendar days on the horizontal axis, but you **do** need to label your axes and give your time-plot a title indicating the dates. Make sure that you plot the two trajectories in **different** colors indicating in the text which color corresponds to which company. Include the legend in your plot.

```r
#I create a scatter plot with walmart, IBM close prices be the y-axis
#time (days) is the x-axis
plot(closeW,type="l", lwd=2, col="blue",
     xlab="Time (days)",
     ylab= "Closing price ($)",
     main="Comparision Walmart and IBM stock closing prices")

lines(closeI, col="red", lwd=2, type="l", )

legend("topleft", legend=c("Walmart", "IBM"),
       col=c("blue", "red"), lty=1, lwd=2)
```

Comparision Walmart and IBM stock closing prices
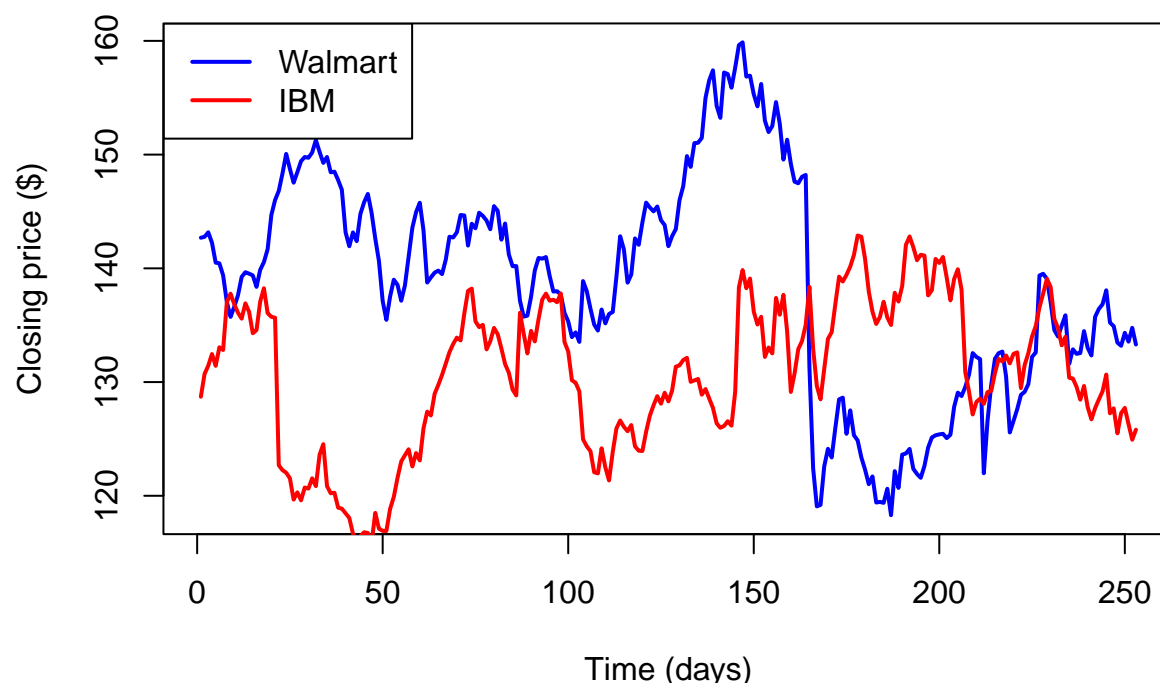
The plot shows that Walmart and IBM stock prices fluctuate over the one-year period. Comparing them visually helps to understand trends and relative volatility. There are periods where they have the same patterns but either increase or decrease in different level. There is probably an event happened to Walmart that decreased its stock market tremendously. However, we don't know which date it is because the limitation of time on the x-axis is represented as index days, not calendar dates, so seasonality or specific events aren't visible. Besides, using raw prices instead of adjusted prices ignores dividends and splits, which can distort comparability.

The **simple daily return** of the stock over a day indexed by t is defined as

$$\frac{\text{price at end of day } t - \text{price at end of day } (t-1)}{\text{price at end of day } (t-1)}$$

(5 points) Construct the vector of simple daily returns over the last year for both stocks. With R denoting the daily simple return, the *daily volatility* of the stock is defined as the standard deviation of R. **Assuming independent, identically distributed** daily returns, what is the **daily volatility** point estimate for the daily volatility of both stocks?

```
#RW is Walmart Return, RI is IBM Return
RW=diff(closeW) / closeW[-length(closeW)]

RI=diff(closeI) / closeI[-length(closeI)]

#volW is Walmart voltality, volI is IBM voltality
volW=sd(RW,na.rm=TRUE)
volI=sd(RI,na.rm=TRUE)
```

The volatility of Walmart is 1.6027952%, and the volatility of IBM is 1.5172089%. Both stocks have similar daily volatility, meaning their daily risk levels are close. Walmart is slightly more volatile. One limitation to point out is the assumption that returns are independent and identically distributed is unrealistic.

(10 points) If you wanted to study the *relationship* between the returns of the two stocks, which plot would you create? What (if anything) can you say after looking at the plot?

```
# I create a scatter plot to show relationship between Walmart returns and IBM returns
plot(RW,RI,
     xlab="Walmart Returns",
     ylab="IBM Returns",
     main="Walmart vs IBM stock returns",
     col="brown", cex = 0.5, pch=19)
abline(lm(RW~RI),lwd=4)
```
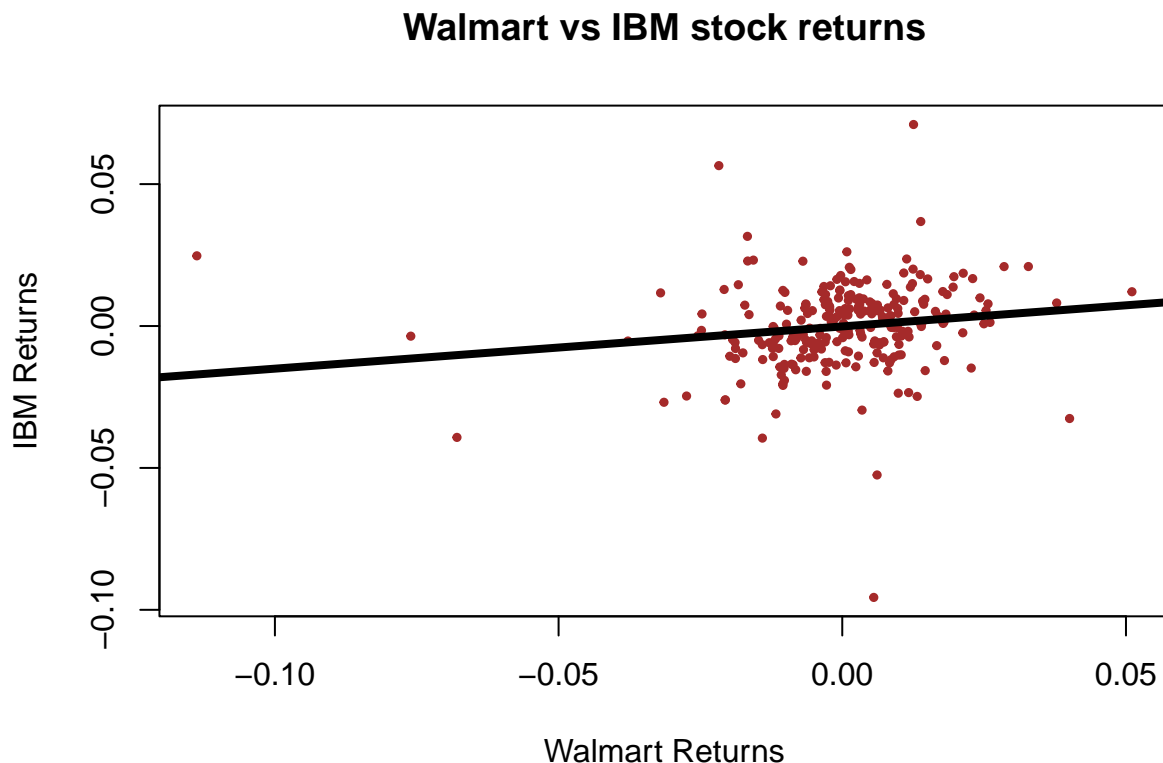

I would use the scatterplot to study the relationship between the returns of the two stock. The plot has a cluster with an upward line indicates a weak positive relationship: when Walmart has positive returns, IBM tends to move in the same direction, but not strongly. The cluster indicates they have similar returns.

(5 points) If you wanted to provide one single, unitless quantity which provided you a measure of association between the returns of *Walmart* and *IBM*, which value would you report?

```
#I calculate the correlation between Walmart returns and IBM returns
cor_WI=cor(RW,RI)
```

I would choose the correlation. The correlation between Walmart and IBM returns is 0.140869, which is a weak positive relationship. The limitation is correlation can vary over time, there is some period, for

example, 100 days to 150 days, I believed they have a stronger correlation. One value of correlation does not tell us their permanent relationship.

(5 points) You create a portfolio in which half your wealth is maintained in the *Walmart* stock and the remaining half of your wealth is kept in the *IBM* stock. What is your estimate of the daily volatility of this portfolio?

```
#I define w as weight of wealth is maintained in the Walmart stock
#(1-w) would be weight in IBM stock, since it is half and half, I set w=0.5
w = 0.5
varI=volI^2
varW=volW^2
cov_WI=cov(RW,RI)

port_vol = sqrt((w^2)*varW +
        ((1-w)^2)*varI +
        2*w*(1-w)*cov_WI)
```

The point estimate of the daily volatility of this portfolio is 0.0117856. The portfolio is less volatile than either stock individually. This means when we invest 50% in Walmart and 50% in IBM, the risk would be lower than when we go all in in either stock individually. However, lower risk doesn't mean we would have highest profit.

(25 points) You create a portfolio in which the weight w of your wealth is maintained in the *Walmart* stock and the remainder of your wealth is kept in the *IBM* stock. Create a function in R which calculates the daily volatility of this portfolio as it depends on w? Plot the portfolio's volatility as a function of w for w ranging from -2 to 2. **Yes, it's perfectly acceptable to have negative weights; if you want to know why, come ask in office hours.** In the same plot, add two horizontal lines at the values of the daily volatilities of *Walmart* and *IBM*. Now, review the optimal two-stock portfolio work we did in the first week of classes. Add the horizontal line corresponding to the optimal portfolio whose *Walmart* weight is $\hat{w} = \hat{\alpha}$ given by Hastie and Tibshirani. Add the vertical line corresponding to the optimal $\hat{w}$.

```
#I define the function that calculates the portolio voltality using the input w
varI=volI^2
varW=volW^2
cov_WI=cov(RW,RI)

port_vol_func = function(w) {
  sqrt((w^2)*varW +
        ((1-w)^2)*varI +
        2*w*(1-w)*cov_WI)
}
#I define "alpha" as the Walmart weight given by Hastie and Tibshirani
alpha = (varI - cov_WI) / (varW + varI - 2*cov_WI)
port_vol_alpha=port_vol_func(alpha)

#I plot the scatter plot with response is portfolio volatility
#predictor is Walmart weight ranging from -2 to 2
#Note: negative weights represent short-selling Walmart (borrowing shares to sell),
#       while weights > 1 mean hold long positions in Walmart
#difference between point is 0.1 to ensure the efficiency and effectiveness
w_seq = seq(-2, 2, 0.1)
vol_seq = sapply(w_seq, port_vol_func)
plot(w_seq,vol_seq, col="purple",
     xlab = "Weight in Walmart (w)",
```

```
        ylab = "Portfolio Volatility",
        main = "Portfolio Volatility as Function of w", lwd=2, type="l")
abline(h=port_vol_alpha, col="red", lwd=2, lty=2)
abline(v=alpha, col="blue", lwd=2, lty=2)
abline(h=volW, col="brown", lwd = 2, lty=2)
abline(h=volI, col="darkgreen", lwd = 2, lty=2)
legend("topleft", legend=c("Optimal w", "Optimal portfolio","IBM volatility", "Walmart volatility","Por
        col=c("blue", "red","darkgreen","brown","purple"), lty=c(2,2,2,2,1), lwd=2)
points(alpha,port_vol_alpha,pch=19 )
text(alpha, port_vol_alpha,
        labels = paste0("(", round(alpha, 3), ", ", round(port_vol_alpha, 4), ")"),
        pos = 4, col = "darkgreen")
```

Portfolio Volatility as Function of w
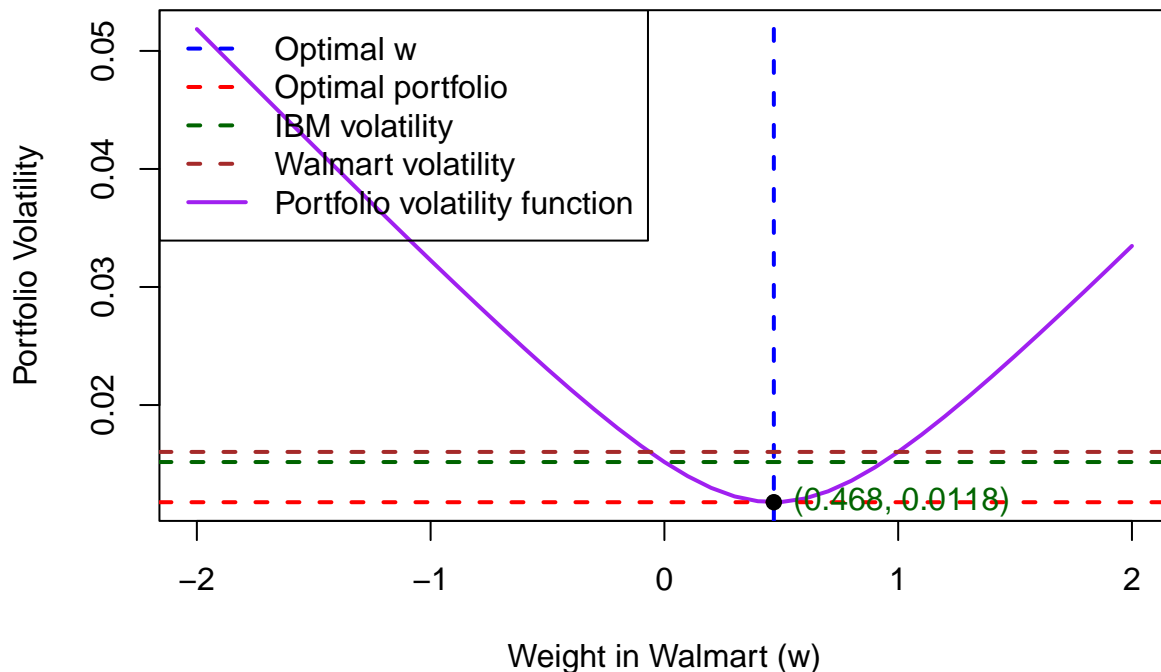
The U-shaped curve shows how portfolio risk varies with Walmart's weight. The optimal portfolio weight minimizes volatility at 0.4681026, producing volatility ~ 0.0117675. Notice that, 0.0117675 is lowest portfolio volatility, lower than either stock individually, and the 50/50 portfolio. With this weight value, we have the lowest risk. By holding ~46.8102555% in Walmart and 53.1897445% IBM, you reduce daily risk to 1.1767502%. Minimizing volatility ignores expected returns. The portfolio is "safest," but not necessarily the most profitable.

Problem #2 (45 points)

Download the historical prices for the *NASDAQ Composite* index for the **maximum** time period and at the **monthly** frequency available on our course website.

```
stockndq=read.csv("/Users/minhtuyennguyen/R/ndq.csv")
```

(5 points) Use the above definitions of returns, changing the period length to a month (rather than a day). Assuming, as usual **independent, identically distributed** monthly returns, what is your *point estimate* for the mean monthly return of *NASDAQ*?

```
#calculate mean monthly return
close_month=stockndq$Close
return_month=diff(close_month)/close_month[-length(close_month)]
mean_month= mean(return_month)
```

The mean monthly return of NASDAQ is 1.0147329%. On average, the NASDAQ index grew by about 1.0147329% per month over the sampled period, showing strong long-term upward momentum.

(5 points) Assuming, as usual **independent, identically distributed** monthly returns, what is your *point estimate* for the monthly volatility of *NASDAQ*?

```
#calculate monthly volatility
vol_month=sd(return_month)
```

The monthly volatility is 6.1655216%. However, this is over a long periods of time in decades, averages smooth out volatility; individual months may differ drastically.

(20 points) You know that there is uncertainty in point estimation due to sampling variability. So, you want to provide a **confidence interval**.

For more about confidence intervals in general, please watch:

Confidence intervals

A follow up video for the mean parameter is

Inference for the mean

Let n stand for the number of observations. Let $\hat{\mu}$ denote our point estimate for the mean parameter. Let s denote the sample standard deviation. At a confidence level C, the structure of a **confidence interval** is

$$(\hat{\mu} - \frac{s}{\sqrt{n}}z^*, \hat{\mu} + \frac{s}{\sqrt{n}}z^*)$$

where z^* is a critical value of the standard normal distribution such that $\mathbb{P}[-z^* < Z < z^*] = C$ with $Z \sim N(0,1)$.

You should input `?qnorm` into the console in `R` to learn more about the different functions which have to do with the normal distribution in `R`.

Provide confidence intervals for the monthly mean return at the confidence level $C = 0.95$.

```
#calculate the 95% confidence interval
CI=c(mean_month - qnorm(0.975)*vol_month/sqrt(length(return_month)),
     mean_month + qnorm(0.975)*vol_month/sqrt(length(return_month)))
```

The 95% confidence interval is (0.4463399% , 1.5831258%). We are 95% confident the true average monthly return lies between 0.4463399% and 1.5831258%. This supports the conclusion that NASDAQ typically delivers positive monthly growth. This is based on the assumption that the return has a normal distribution which might be not case in financial returns.

(5 points) There is a "shortcut" for constructing confidence intervals in this case which uses the `t.test` function. You should input `?t.test` into the console in `R` to learn more about it. Create the 95%−confidence interval using this command and compare to your result from the "pedestrian" implementation above.

```
# comparing t.test 95% confidence interval and computed 95% confidence interval
t.test(return_month, conf.level = 0.95)
##
##  One Sample t-test
##
## data:  return_month
## t = 3.4991, df = 451, p-value = 0.0005133
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.004448105 0.015846553
## sample estimates:
##  mean of x
## 0.01014733
```

My 95% confidence interval (0.0044634 , 0.0158313) is approximately equal to the 95% confidence interval from the t.test, which indicates that my confidence interval is correct. I used "qnorm" instead of "qt" because we have a large sample size that is greater than 30.

(10 points) Finally, create a **bootstrap** $2SE$ confidence interval and a 95% percentile interval for the monthly mean return. Please, set your seed to 1 for comparison. Plot the histogram of the bootstrap point estimates for $N = 10^5$ draws. Indicate your confidence intervals on the histogram. Compare your confidence intervals to what you obtained above without resampling.

```
# create bootstrap, resample 10^5 times
# Define the bootstrap statistic function
# data     = the vector of observed returns
# indices = a resampled set of positions (with replacement) from data
# The function computes the mean return of the resampled data
boot_mean_fn <- function(data, indices) {
  mean(data[indices])
}
library(boot)
##
## Attaching package: 'boot'
## The following object is masked from 'package:nimble':
##
##     logit
set.seed(1)
boot_out=boot(return_month, statistic= boot_mean_fn, R=10^5)

# calculate bootstrap 2SE confidence interval
ci_2se=c(boot_out$t0 - 2*sd(boot_out$t), boot_out$t0 +2*sd(boot_out$t))

# create the histogram of bootstrap means
hist(boot_out$t, breaks = 75, col = "lightblue",
    main = "Bootstrap Distribution of Mean Monthly Return",
    xlab = "Bootstrap Means")

#calculate bootstrap 95% confidence interval
ci_95=quantile(boot_out$t, c(0.025,0.975))

#add line corresponding to each confidence interval
abline(v=ci_2se, col="red", lwd =0.5)
abline(v=CI, col="blue", lwd=0.5)
```
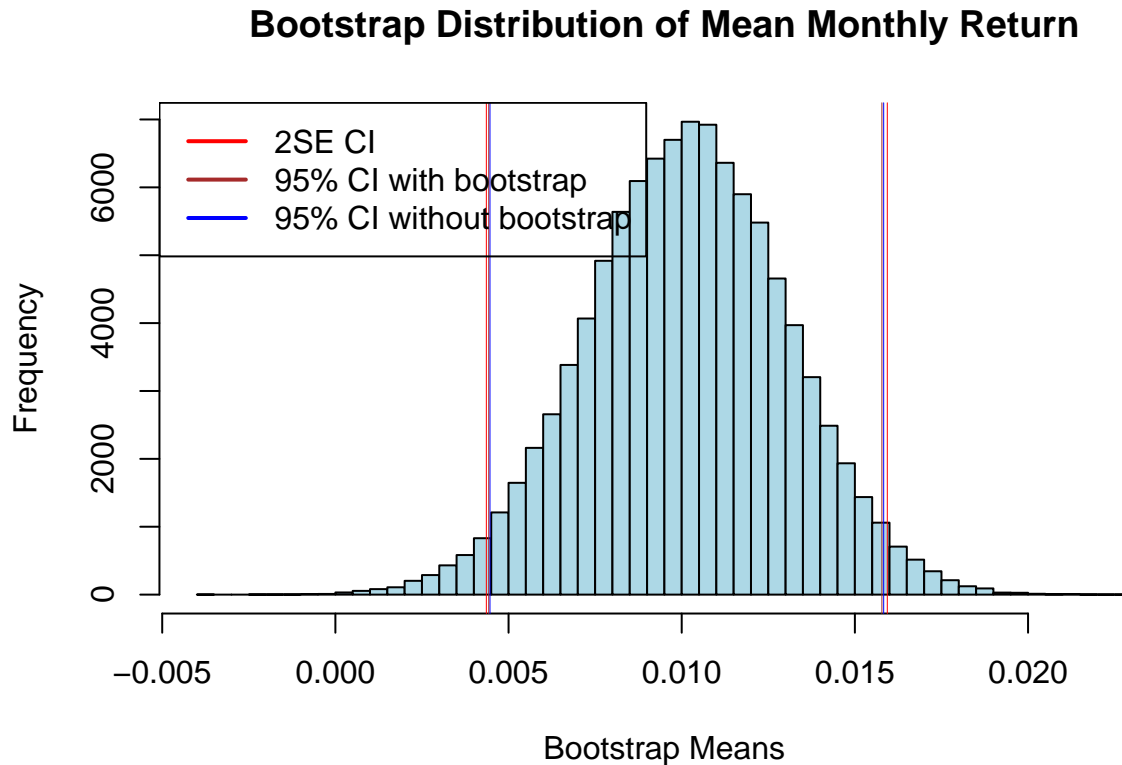
```
abline(v=ci_95, col="brown", lwd=0.5)

legend("topleft",
       legend = c("2SE CI", "95% CI with bootstrap", "95% CI without bootstrap"),
       col    = c("red", "brown", "blue"),
       lwd    = 2,
       lty    = 1)
```

Bootstrap Distribution of Mean Monthly Return

The 95% confidence interval without resampling is $(0.4463399\% , 1.5831258\%)$. The 95% confidence interval with resampling is $(0.4427623\% , 1.5779489\%)$. The bootstrap 2SE confidence interval is $(0.435888\% , 1.5935777\%)$. They are approximately the same with the maximum difference is 0.0104519% for the lower bound and 0.0156288% for the upper bound, meaning that the estimated mean is robust.

The computed 95% confidence interval is a parametric approach with assumption that the samples have normal distribution by the Central Limit Theorem (the sample size is large n = 452). While the bootstrap confidence interval is non-parametric approach without this assumption. The similarity among these intervals indicates that the sampling distribution of the monthly mean returns is approximately normal (as expected by the CLT), not that the returns themselves are normal. The bootstrap 2SE interval is also similar, which is consistent with an approximately symmetric sampling distribution of the mean, where the 2 standard errors from the mean roughly correspond to a 95% interval of a normally distributed sample. In short, the agreement among the confidence intervals is explained by the Central Limit Theorem.

However, resampling ignores structural breaks or trends in financial markets.