THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

SCHOOL OF SCIENCE

Using Probability Theory and the Discrete Fourier Transform to Develop Equity Trading Strategies

Noah Chough
Spring 2023

A thesis
submitted in partial fulfillment
of the requirements
for baccalaureate degrees
in Mathematics
with honors in Mathematics

Reviewed and approved* by the following:

Joseph Previte, Ph.D.
Associate Professor, Mathematics
Thesis Supervisor

Daniel Galiffa, Ph.D.
Associate Professor, Mathematics
Honors Adviser

*Signatures are on file in the Schreyer Honors College.

Abstract

Although stock prices depend on a wide array of economic factors, the purpose of this research is to analyze them from a mathematical standpoint. This analysis consists of two approaches: one of probability theory, and one of Fourier analysis.

First, probability theory is used to calculate the theoretical probabilities of short-term trends in random data. Then, these are compared to the probabilities of the same short-term trends in stock price data. This comparison is accomplished via the Kolmogorov-Smirnov test, which reveals that the probabilities differ from their theoretical values in random data.

Second, Fourier analysis, accomplished via the Discrete Fourier Transform, is used to detect periodicity in data. If there is periodicity in stock price data, this could hint at where optimal entry and exit points may lie. However, the data analyzed here does not exhibit any notable periodicities.

Hypothetically, a combination of the above strategies - the probability of short-term trends, and periodic behavior - could be used to construct a profitable trading strategy. Such a strategy could be applied in the short term on a daily level, or over a longer time period, depending on the frequency of observations.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Rationale

The purpose of this research is to analyze stock price data from two mathematical perspectives: one of probability theory to determine the probability of short-term trends, and one of Fourier analysis to look for periodicity.

Consider that short-term trends have a certain probability of occurring in sequences of random numbers. By comparing the probability of these trends in sequences of random numbers to the probability of the same trends in stock data, one could analyze the difference to determine the odds of another decrease or increase in the data. In order to compare the incidence of these probabilities, a statistical test known as the Kolmogorov-Smirnov test can be applied.

The Fourier transform can be applied to break down a function into its frequency components. Here, the discrete version of the Fourier transform is utilized to break down a sequence of a stock's prices into its frequency components. If a sequence has certain frequencies inherent in it, these correspond to periodicities; that is, intervals over which the data is periodic/has the same pattern. By looking for these frequencies and their corresponding periods, one may be able to determine how long certain trends in the data are.

By combining these two approaches, one may be able to create a profitable trading strategy that is solely based on mathematical trends in the data.

From an economic perspective, it may seem more logical to look at the causal factors that influence stock prices and use those to predict the behavior of stock prices. However, that is not the purpose of this project. Here, the interest lies in looking for potential mathematical trends that may exist in stock data.

Lastly, it is worth mentioning that the topics discussed in this paper assume some familiarity with basic calculus and statistics. In particular, a basic understanding of probability and some familiarity with integral calculus are necessary to understand the approaches presented here.

1.2 Data

The stock data studied in this analysis were the daily closing prices of the 30 companies that make up the Dow Jones Industrial Index. These particular stocks were chosen because their average is typically used to describe the behavior of the stock market as a whole. Furthermore, since many of the analyses in this research were done by hand and not automated, 30 was a manageable number to work with.

The raw stock price data was obtained from yahoo! finance in the form of .csv files that are readily available for download.

Chapter 2

Probability of Short-Term Trends

2.1 Probability of Short-Term Trends in Sequences of Random Numbers

Consider a sequence of random numbers, where the probability that any two terms are equal is essentially zero. Given that there was a decrease from the previous term to the current term, what is the probability that there will be another decrease?

Without loss of generality, call the "current" term x_n, and think of a decrease from the previous term x_{n-1} to x_n as the expression $x_{n-1} > x_n$. Furthermore, think of another decrease to the next term as the expression $x_{n-1} > x_n > x_{n+1}$. Now, the probability we are interested in in can be expressed as the following conditional probability:

$$P(x_{n-1} > x_n > x_{n+1} \mid x_{n-1} > x_n) \tag{2.1}$$

that is, the probability that $x_{n-1} > x_n > x_{n+1}$ given that $x_{n-1} > x_n$.

2.1.1 Evaluating the Conditional Probability

Conditional probability, the probability of an event A occurring given that an event B has occurred, is defined as the following:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.2}$$

With this in mind, the conditional probability from above can be expressed as follows:

$$\begin{aligned} P(x_{n-1} > x_n > x_{n+1} \mid x_{n-1} > x_n) &= \frac{P(x_{n-1} > x_n > x_{n+1} \ \cap \ x_{n-1} > x_n)}{P(x_{n-1} > x_n)} \\ &= \frac{P(x_{n-1} > x_n > x_{n+1})}{P(x_{n-1} > x_n)} \end{aligned} \tag{2.3}$$

In order to calculate the value of these probabilities in a sequence of random numbers, consider the following combinatorial perspective.

Suppose that each of the terms represents the height of a block. Because the probability that any two terms are equal is essentially zero, it can be assumed that the blocks are of varying, unequal heights. Now, the probabilities from above can be rephrased as the following question: when arranging the blocks, what is the probability that they are in decreasing order?

First, we will examine the case involving two terms, x_{n-1} and x_n. Treating these as the heights of two blocks, they can be arranged in $2! = 2$ ways. However, the taller one comes first in only 1 of those arrangements. Thus, the probability that $x_{n-1} > x_n$, that is, $P(x_{n-1} > x_n)$, is $\frac{1}{2}$.

Next, we will construct a similar argument to examine the case involving three terms, x_{n-1}, x_n, and x_{n+1}. Treating these as the heights of three blocks, they can be arranged in $3! = 6$ ways. However, they are arranged in decreasing order in only 1 of those arrangements. Thus, the probability that $x_{n-1} > x_n > x_{n+1}$, that is, $P(x_{n-1} > x_n > x_{n+1})$, is $\frac{1}{6}$.

With these results in mind, the conditional probability from (2.3) can now be evaluated.

$$P(x_{n-1} > x_n > x_{n+1} \mid x_{n-1} > x_n) = \frac{P(x_{n-1} > x_n > x_{n+1})}{P(x_{n-1} > x_n)} = \frac{\left(\frac{1}{6}\right)}{\left(\frac{1}{2}\right)} = \frac{1}{3} \tag{2.4}$$

Therefore, the probability that there will be another decrease given that there was a decrease from the previous term to the current term is $\frac{1}{3}$.

2.1.2 Expanding the Conditional Probability

The same approach can be expanded upon to determine the probability of another decrease occurring after more than one previous decrease(s). Consider a similar conditional probability as the one from before, but now spanning back k previous decreases

$$P(x_{n-k} > ... > x_n > x_{n+1} \mid x_{n-k} > ... > x_n) \tag{2.5}$$

where $k \geq 1$.

Evaluating this via the same combinatorial perspective as before yields the following result:

$$\begin{aligned} P(x_{n-k} > ...x_n > x_{n+1} \mid x_{n-k} > ... > x_n) &= \frac{P(x_{n-k} > ...x_n > x_{n+1} \cap x_{n-k} > ... > x_n)}{P(x_{n-k} > ... > x_n)} \\ &= \frac{P(x_{n-k} > ... > x_n > x_{n+1})}{P(x_{n-k} > ... > x_n)} \\ &= \frac{\left(\frac{1}{(k+2)!}\right)}{\left(\frac{1}{(k+1)!}\right)} \\ &= \frac{(k+1)!}{(k+2)!} \\ &= \frac{1}{k+2} \end{aligned} \tag{2.6}$$

Thus, given k previous decreases, the probability of another decrease is equal to $\frac{1}{k+2}$. This result is evident in the previous section, where, given 1 previous decrease, the probability of another decrease was found to equal $\frac{1}{3}$.

2.1.3 Reverse Direction

It is worth noting that the probability of another increase given $k \geq 1$ previous increases can be evaluated in the same manner. The only difference is that, instead of counting the number of arrangements in which the blocks are decreasing, count the number of arrangements where the blocks are increasing. Similar to the previous result, though, the blocks are increasing in only 1 arrangement, and the result for increases is the same as the result for decreases.

2.1.4 Verifying the Probability with Random Numbers

In order to verify the theoretical value of this probability, a sequence of random numbers can be generated, and the incidence of both conditions can be checked and counted to experimentally determine the value of the conditional probability.

Below is the pseudocode that was used to calculate the experimental value of the conditional probability. Although this was accomplished in C++, the same concept can be implemented in other programming languages and in Microsoft Excel.

Note that this corresponds to the (conditional) probability of another decrease given a decrease from the previous term to the current term; that is, the $k = 1$ case of equation (2.6).

```
generate N random numbers

create an array of length 3

fill the array with the first 3 random numbers

if there is a decrease between the first two values:
  add 1 to the count of event B

  if there is a decrease between all three values:
    add 1 to the count of event A

else
  shift the array by 1 index

the experimental value of the conditional probability is equal to
the count of event A divided by the count of event B
```

Here are a few results that were generated for $N = 1000$:

```
P(A|B) = 152/493 = 0.308316
P(A|B) = 176/505 = 0.348515
P(A|B) = 173/493 = 0.350913
```

...and a few more that were generated for $N = 10000$:

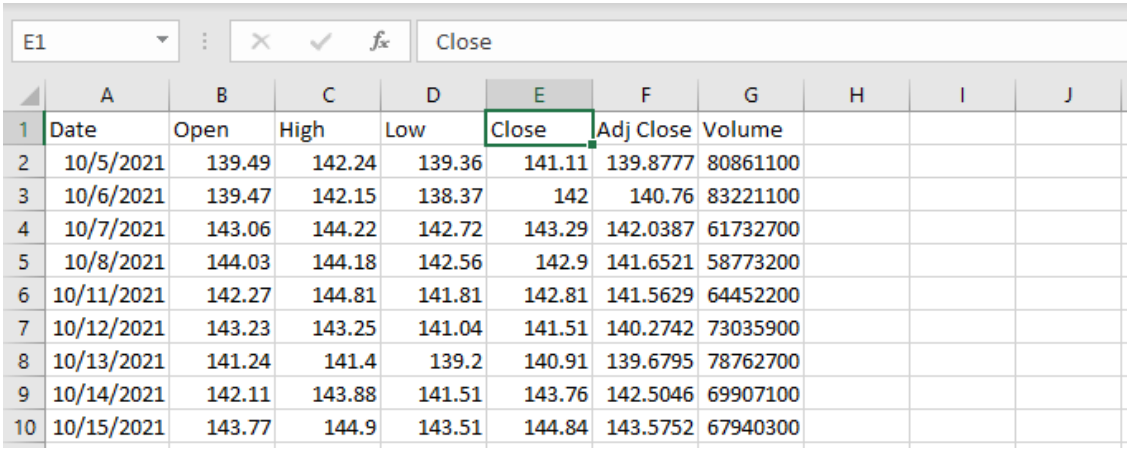
```
P(A|B) = 1635/5001 = 0.326935
P(A|B) = 1674/4996 = 0.335068
P(A|B) = 1692/5012 = 0.33759
```

As N increases, the experimental value of the conditional probability approaches $\frac{1}{3}$, which appears to confirm the theoretical value calculated in the previous section.

2.2 Evaluating the Probability in Stock Data

In the previous sections, the value of the conditional probability in sequences of random numbers was examined. Now, we will examine the value of the conditional probability for stock data.

As mentioned in the introduction, the stocks examined in this analysis were the 30 stocks that make up the Dow Jones Industrial Index, and the past 1 year of their daily closing prices - approximately 252 observations - were analyzed here. This data was obtained in the form of .csv files available for download from yahoo! finance. Here is one such .csv file for AAPL:

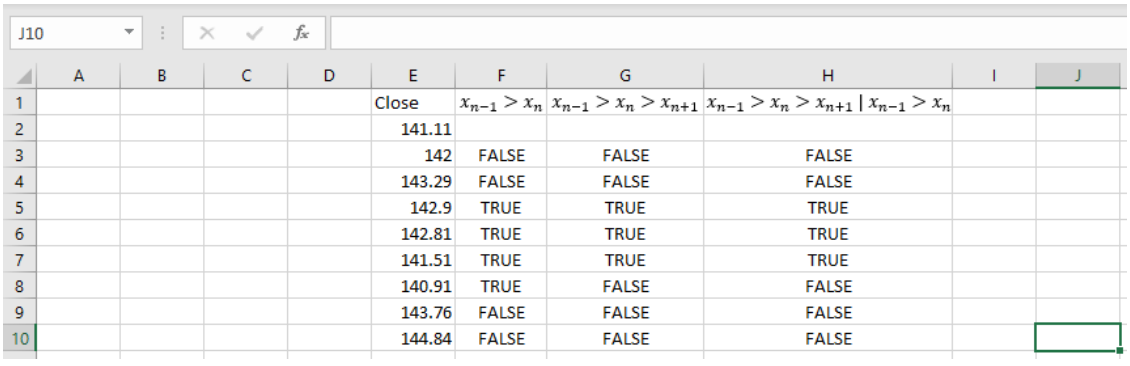


E1 | Close

	A	B	C	D	E	F	G	H	I	J
1	Date	Open	High	Low	Close	Adj Close	Volume			
2	10/5/2021	139.49	142.24	139.36	141.11	139.8777	80861100			
3	10/6/2021	139.47	142.15	138.37	142	140.76	83221100			
4	10/7/2021	143.06	144.22	142.72	143.29	142.0387	61732700			
5	10/8/2021	144.03	144.18	142.56	142.9	141.6521	58773200			
6	10/11/2021	142.27	144.81	141.81	142.81	141.5629	64452200			
7	10/12/2021	143.23	143.25	141.04	141.51	140.2742	73035900			
8	10/13/2021	141.24	141.4	139.2	140.91	139.6795	78762700			
9	10/14/2021	142.11	143.88	141.51	143.76	142.5046	69907100			
10	10/15/2021	143.77	144.9	143.51	144.84	143.5752	67940300			

Figure 2.1: Stock data .csv file from yahoo! finance

For simplicity's sake, this data was analyzed directly in Microsoft Excel. In the image below, the condition $x_{n-1} > x_n$ occurred was checked in column `F` using the command `=E2>E3`; the condition $x_{n-1} > x_n > x_{n+1}$ was checked in column `G` using the command `=AND((E2>E3),(E3>E4))`; the statement $x_{n-1} > x_n > x_{n+1} \mid x_{n-1} > x_n$ was checked in column `H` using the command `AND(F3,G3)`.



J10

	A	B	C	D	E	F	G	H	I	J
1					Close	$x_{n-1} > x_n$	$x_{n-1} > x_n > x_{n+1}$	$x_{n-1} > x_n > x_{n+1} \mid x_{n-1} > x_n$		
2					141.11					
3					142	FALSE	FALSE	FALSE		
4					143.29	FALSE	FALSE	FALSE		
5					142.9	TRUE	TRUE	TRUE		
6					142.81	TRUE	TRUE	TRUE		
7					141.51	TRUE	TRUE	TRUE		
8					140.91	TRUE	FALSE	FALSE		
9					143.76	FALSE	FALSE	FALSE		
10					144.84	FALSE	FALSE	FALSE		

Figure 2.2: Excel commands used to calculate the conditional probability

After copying those commands to span the entire length of the data, the count of each one was tallied up and used to calculate the value of the conditional probability. This was accomplished by using the command `=COUNTIF(E3:E252, "TRUE")`, which counted the number of times each condition was true.

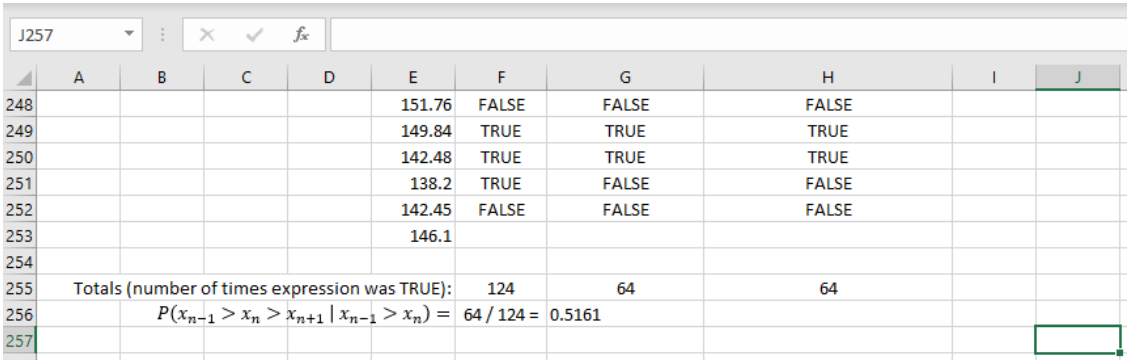


Figure 2.3: Excel output of the conditional probability result

Notice that, in this case, the value of the conditional probability was 0.5161 - slightly greater than $\frac{1}{2}$, and undeniably different from the value of $\frac{1}{3}$ that was observed in random data. This result was similar to the results observed for the rest of the 30 stocks. All of them had a probability greater than $\frac{1}{3}$, with most having a probability of slightly below $\frac{1}{2}$.

This motivates the following question. though: how much do these probabilities differ from 1/3? Is there some statistical method that can be applied to conclude whether they differ from 1/3 by a statistically significant amount?

2.3 Comparing the Probabilities with the Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test is used to compare a sample cumulative distribution function (CDF), $F_n(x)$, to an existing CDF, $F(x)$, by measuring the distance at the value of x where the sample differs from the existing by the greatest magnitude. This distance:

$$D_n = \sup |F_n(x) - F(x)| \tag{2.7}$$

is known as the Kolmogorov-Smirnov test statistic, and can be tested at various significance levels to determine how close the sample CDF is to the existing CDF.

The null and alternative hypotheses of the Kolmogorov-Smirnov test are as follows:

$$\begin{aligned} H_0: &\quad \text{the sample CDF could have come from the existing CDF} \\ H_a: &\quad \text{the sample CDF could not have come from the existing CDF} \end{aligned} \tag{2.8}$$

and can be tested at various levels of significance. [1]

The null hypothesis of the Kolmogorov-Smirnov test, that the sample probability distribution comes from the existing probability distribution, is rejected at level α if

$$\begin{aligned} \sqrt{n} D_n &> K_\alpha \\ D_n &> \frac{K_\alpha}{\sqrt{n}} \end{aligned} \tag{2.9}$$

The values of $\frac{K_\alpha}{\sqrt{n}}$ can be found in tables, so the test statistic - that is, the difference D_n - can be directly compared with them.

In order to apply the Kolmogorov-Smirnov test to the probability of interest here, though, we need to generate a CDF and determine what existing CDF to compare it to.

2.3.1 Existing CDF

The conditional probability in question, $P(x_{n-1} > x_n > x_{n+1} \mid x_{n-1} > x_n)$, has a probability of $\frac{1}{3}$ for random x_i. Each check for the expression $x_{n-1} > x_n > x_{n+1}$ being true given that $x_{n-1} > x_n$ is true can be thought of as a Bernoulli trial, because it will either be true or false. Therefore, the distribution of the conditional probability in sequences of random numbers can be thought of as a Binomial distribution with $p = \frac{1}{3}$.

The CDF of a binomial distribution is the following:

$$P(X \leq k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i} \tag{2.10}$$

where $P(X \leq k)$ represents the probability of random variable X being less than k after n trials. As mentioned above, $p = \frac{1}{3}$, and n was chosen to be 100.

Recall that each stock data file contained 252 observations. Additionally, note that in order to count the incidence of the expression $x_{n-1} > x_n > x_{n+1}$ being true, the expression $x_{n-1} > x_n$ must be true first. As demonstrated in Section 2.1.1, the expression $x_{n-1} > x_n$ has a probability of $\frac{1}{2}$ of occurring in a sequence of random numbers. So, given 252 trials, one would theoretically expect this expression to be true 126 times. In practice, when looking at the stock data, it was found that this condition was always true at least 100 times. So, with this in mind (and also because 100 is a convenient number to work with) it was chosen for n.

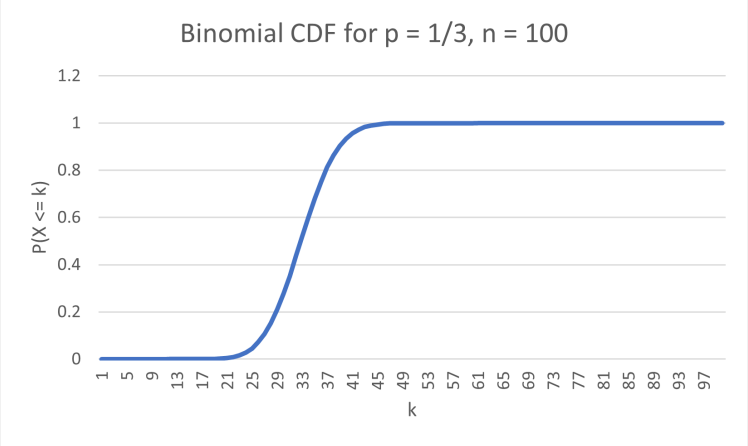


Figure 2.4: CDF of a Binomial distribution with $p = \frac{1}{3}$ and $n = 100$

An explanation of the above CDF can be thought of as follows: This represents the cumulative probability of the expression $x_{n-1} > x_n > x_{n+1}$ being true X times out of k trials. By comparing this to a CDF generated from the value of the conditional probability in stock data, we can see if the cumulative probabilities are similar via the Kolmogorov-Smirnov test.

2.3.2 Experimental CDF

Now that the existing CDF is known, all that remains is to generate the sample CDF of the probability in stock data.

The sample CDF of the conditional probability in stock data was generated in the following manner. First, each of the 30 stocks was analyzed, and both conditions were counted until the condition $x_{n-1} > x_n$ was true 100 times. Then, the corresponding number of times that $x_{n-1} > x_n > x_{n+1}$ was true was recorded. This was done because, in order to compare this CDF with the one from the previous section, n must be the same, and $x_{n-1} > x_n > x_{n+1}$ being true given that $x_{n-1} > x_n$ was the Bernoulli trial that the binomial CDF in the previous section was based upon.

Bear in mind, though, that this process only generates one CDF. In order to account for this, larger data sets containing the past 5 years' worth of stock data were downloaded, and various intervals of sampling - every day, every other day, and every third day - were applied to generate multiple CDFs through which the conditional probability could be examined. Furthermore, to account for a difference over time, the most recent results were sampled ("First 100"), along with the most historical results ("Last 100"). Here is a table containing the results:

9		Every day		Every other day		(shifted 1 day)		Every third day		(shifted 1 day)		(shifted 1 more day)	
10	Symbol	First 100	Last 100	First 100	Last 100	First 100	Last 100	First 100	Last 100	First 100	Last 100	First 100	Last 100
11	AXP	47	56	45	52	44	51	47	51	39	55	39	44
12	AMGN	48	51	46	46	44	50	46	47	46	55	40	52
13	AAPL	49	53	45	50	41	46	38	42	43	46	37	41
14	BA	39	54	43	54	48	57	43	51	49	57	47	49
15	CAT	47	57	49	46	44	47	55	43	49	41	50	52
16	CSCO	37	56	43	51	48	48	50	40	49	44	48	47
17	CVX	40	45	47	49	46	43	39	44	50	41	50	38
18	GS	46	56	57	52	50	54	59	53	52	49	44	54
19	HD	50	56	47	50	52	51	48	45	43	48	46	43
20	HON	41	54	48	51	36	46	46	55	46	48	42	46

Figure 2.5: Tallies of the condition given 100 triggers

Keep in mind that each column corresponds to the generation of 1 CDF.

From here, the CDFs were generated in the following manner:

	37	0	0	0
	38	0	0	0
	39	1	0.033333	0.033333
	40	3	0.1	0.133333
	41	0	0	0.133333
	42	3	0.1	0.233333
	43	2	0.066667	0.3
	44	2	0.066667	0.366667
	45	3	0.1	0.466667
	46	2	0.066667	0.533333
	47	2	0.066667	0.6
	48	2	0.066667	0.666667
	49	2	0.066667	0.733333
	50	1	0.033333	0.766667
	51	2	0.066667	0.833333
	52	1	0.033333	0.866667
	53	1	0.033333	0.9
	54	0	0	0.9
	55	3	0.1	1
	56	0	0	1

Figure 2.6: Columns used to generate the sample CDF

The screenshot above depicts the structure used to generate the sample CDF.

The first column represents the values of k, where k represents the number of times that the main condition, $x_{n-1} > x_n > x_{n+1}$, was true out of 100 trials where $x_{n-1} > x_n$ was true. The second column tallies up the measurements observed in the stock data. The third column divides the tallies by 30, turning them into probabilities. This column could be thought of as the probability distribution function (PDF) of the conditional probability for the given sampling method. The fourth and final column kept a running total of the probability to ultimately generate a CDF.

Visually, a CDF generated via this process looks like the following:

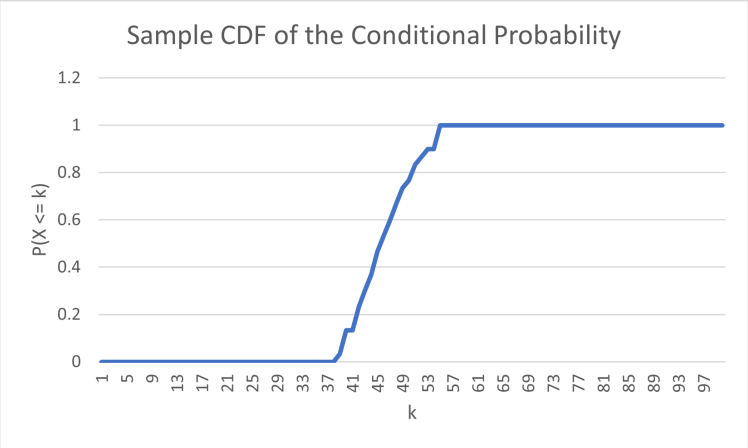


Figure 2.7: Sample CDF generated via this process

Now, we have a sample CDF that can be compared with the existing binomial CDF.

2.3.3 Running the Kolmogorov-Smirnov Test

In order to run the Kolmogorov-Smirnov test, the difference between each term of the sample CDF and the existing binomial CDF was analyzed. These differences were stored in a column, and the maximum difference in that column was the Kolmogorov-Smirnov test statistic.

Visually, this looks like the following:

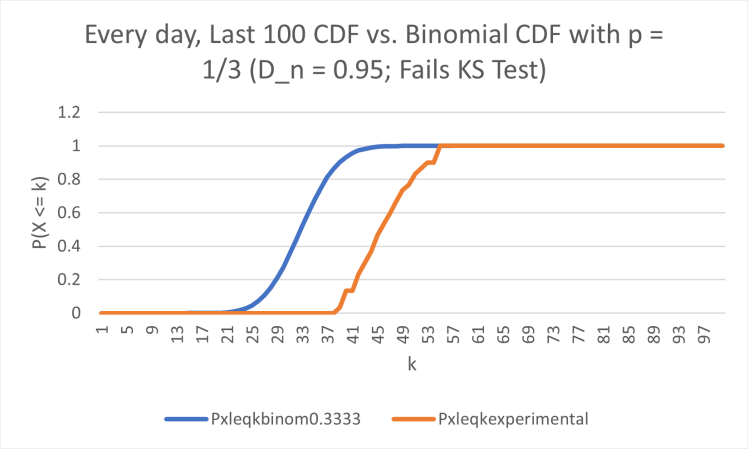


Figure 2.8: Visual representation of the Kolmogorov-Smirnov test

In the graph above, the Kolmogorov-Smirnov test indicated to reject the null hypothesis that the sample distribution of the conditional probability could have come from a binomial distribution with p = 1/3. Here is another graph visually depicting the Kolmogorov-Smirnov test (note that this comparison also fails the test):

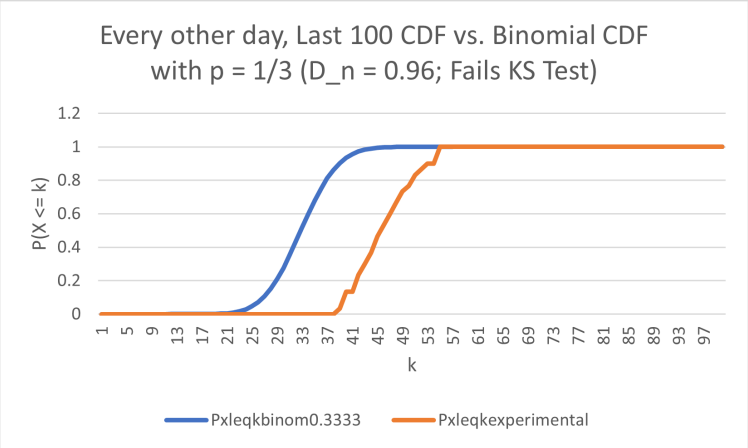


Figure 2.9: Another visual representation of the Kolmogorov-Smirnov test

After observing these results, it was clear that the experimental CDF was not close to the binomial CDF with $p = \frac{1}{3}$. In each sampling method, the sample CDF never passed the Kolmogorov-Smirnov test when compared to the binomial CDF with $p = \frac{1}{3}$. However, it was noticed that sometimes, the experimental CDF did look close to a binomial CDF with $p = \frac{1}{2}$. Out of curiosity, it was compared to a binomial CDF with $p = \frac{1}{2}$, and it "passed" the Kolmogorov-Smirnov test on numerous occasions:

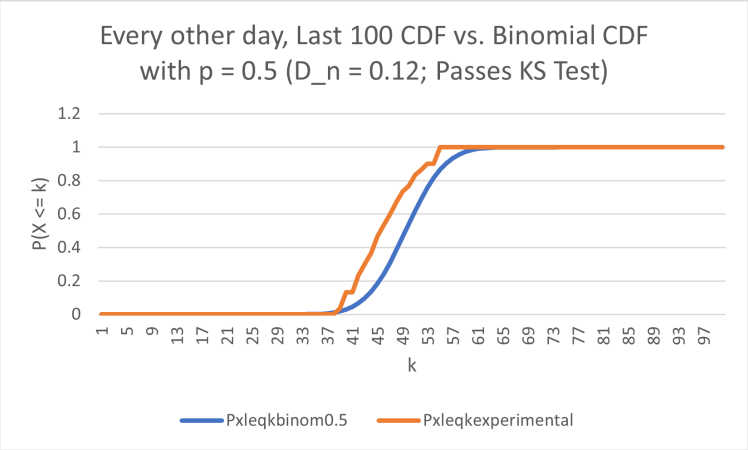


Figure 2.10: A "pass" of the Kolmogorov-Smirnov test

However, the main takeaway is that the sample CDF never "passed" when compared to the binomial distribution with $p = \frac{1}{3}$; the decision at $\alpha = 0.05$ was always to reject the null hypothesis that the sample CDF could have come from the existing binomial CDF. All in all, there was no trial in which the experimental CDF could have come from a binomial CDF with $p = \frac{1}{3}$.

Chapter 3

Fourier Analysis

3.1 Fourier Transform

Fourier analysis, the practice of breaking down a function into its frequency components, is accomplished via the Fourier Transform. The Fourier Transform of a function $f(t)$ is defined as follows:

$$f(v) = \int_{-\infty}^{\infty} f(t) e^{-2\pi i v t} dt \tag{3.1}$$

In practice, $f(t)$ is not usually known; more often than not, a sequence of data points is the object of interest.

This motivates the Discrete Fourier Transform, which, instead of decomposing a function into its frequency components, decomposes a sequence of numbers into its frequency components.

3.1.1 Discrete Fourier Transform

The discrete Fourier transform of a sequence $\{f_k\}_{k=0}^{N-1}$ is defined as follows:

$$F_n = \sum_{k=0}^{N-1} f_k \, e^{-2\pi i n k/N} \tag{3.2}$$

where F_n represents the frequency component corresponding to that value of n. [2]

In order to understand the mechanism by which the discrete Fourier transform reveals periodicity, recall Euler's identity, $e^{i\theta} = \cos(\theta) + i \sin(\theta)$, and notice that the quantity $f_k \, e^{-2\pi i n k/N}$ can be rewritten in the following manner:

$$\begin{aligned} f_k \, e^{-2\pi i n k/N} &= f_k \left(\cos(-2\pi n k/N) + i \sin(-2\pi n k/N)\right) \\ &= f_k \cos(-2\pi \frac{n}{N} k) + i f_k \sin(-2\pi \frac{n}{N} k) \end{aligned} \tag{3.3}$$

Notice that the angular frequency in the $\cos$ and $\sin$ functions is $\omega = 2\pi \frac{n}{N}$, which corresponds to an ordinary frequency of $f = \frac{n}{N}$. If $\frac{n}{N}$ corresponds to a frequency that is present in the data, then the sum corresponding to that value of n will be large. This is best observed in a plot of the complex modulus of F_n, where large values - peaks in the graph - indicate frequencies that are inherent in the sequence $\{f_k\}$.

3.1.2 Examples with Manufactured Periodic Data

To demonstrate the discrete Fourier transform's ability to reveal periodicity in data, it will be applied to sequences of known periodicity, and its output will be analyzed.

An Example with One Periodicity

Consider the function $f(x) = \sin\left(\frac{2\pi}{4}x\right)$ evaluated at the integers from 1 to 100. The plot of $f(x)$, with the outputs of the function at the integers from 1 to 100 denoted by dots, is as follows:

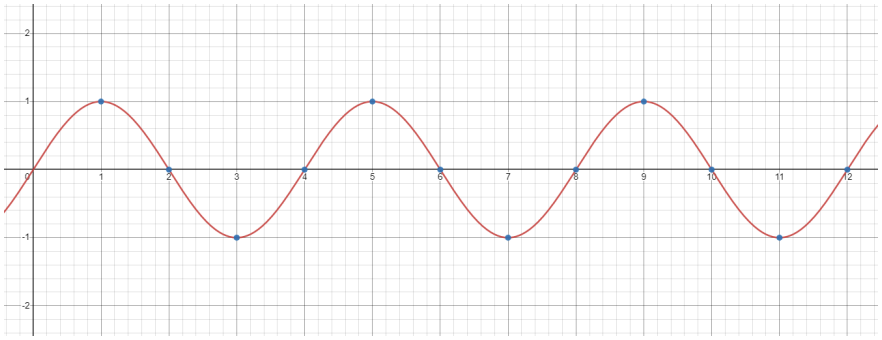


Figure 3.1: Plot of $f(x) = \sin\left(\frac{2\pi}{4}x\right)$

By definition, $f(x)$ is periodic every 4 terms. Now, observe the discrete Fourier transform of the sequence $\{f_k\}$, where f_k is the set of the outputs of $f(x)$ at the integers from 1 to 100:

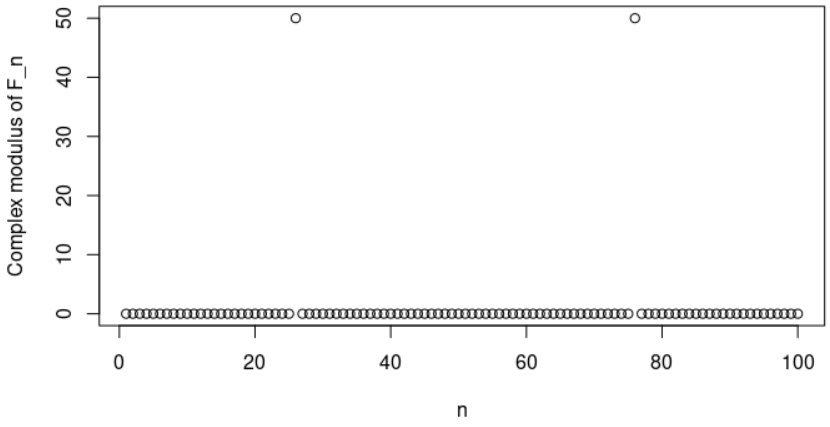


Figure 3.2: Plot of the discrete Fourier transform of $\{f_k\}$

Notice the peaks at $n = 25$ and $n = 75$. Because the output of the discrete Fourier transform is symmetric, these correspond to the same frequency of $\frac{n}{N} = \frac{25}{100} = 0.25$ (that is, a period of 4).

An Example with Multiple Periodicities

Consider the function $g(x) = 5\sin\left(\frac{2\pi}{50}x\right) + \sin\left(\frac{2\pi}{5}x\right)$ evaluated at the integers from 1 to 100. The plot of $g(x)$, with the outputs of the function at the integers from 1 to 100 denoted by dots, is as follows:

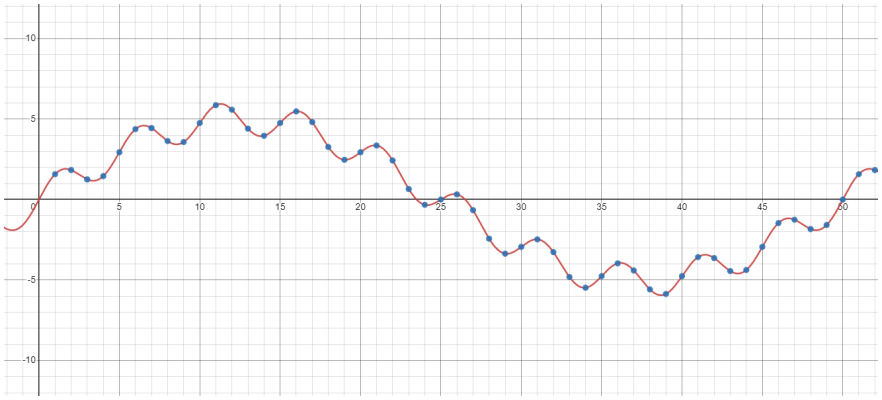


Figure 3.3: Plot of $g(x) = 5\sin\left(\frac{2\pi}{50}x\right) + \sin\left(\frac{2\pi}{5}x\right)$

By definition, $g(x)$ is periodic every 50 terms, and is also periodic every 5 terms. Now, observe the discrete Fourier transform of the sequence $\{g_k\}$, where g_k is the set of the outputs of $g(x)$ at the integers from 1 to 100:

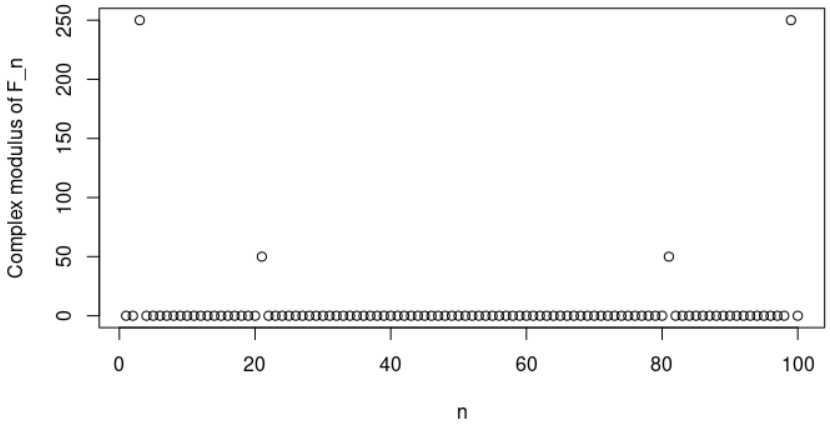


Figure 3.4: Plot of the discrete Fourier transform of $\{g_k\}$

Notice the peaks at $n = 2$ and $n = 20$ (symmetric peaks at $n = 98$ and $n = 80$, respectively). The larger peak at $n = 2$ indicates that a strong frequency present in the data is $\frac{n}{N} = \frac{2}{100} = 0.02$ (that is, a period of 50), and the smaller peak at $n = 20$ indicates that another frequency present in the data is $\frac{n}{N} = \frac{20}{100} = 0.2$ (that is, a period of 5).

3.2 Application to Stock Data

Now that the discrete Fourier transform's ability to detect periodicity in a sequence has been recognized, it will be applied to stock data. Similar to the analysis in Chapter 2, the data analyzed here was 1 year of daily closing prices for each of the 30 stocks that make up the Dow Jones Industrial Average.

3.2.1 Demonstration of the Process for One Stock

Here is a plot of 1 year's worth of daily closing prices for AAPL:

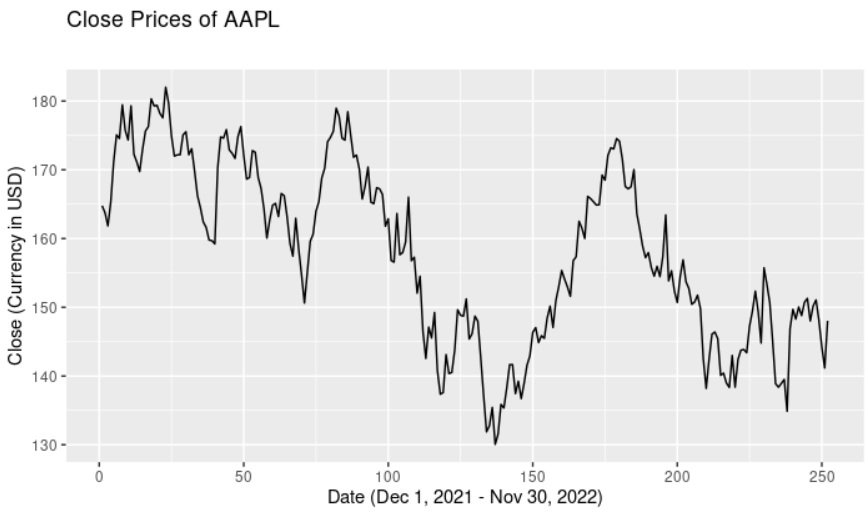


Figure 3.5: Plot of 1 year of daily closing prices for AAPL

...and here is the plot of the discrete Fourier transform of this sequence of closing prices:

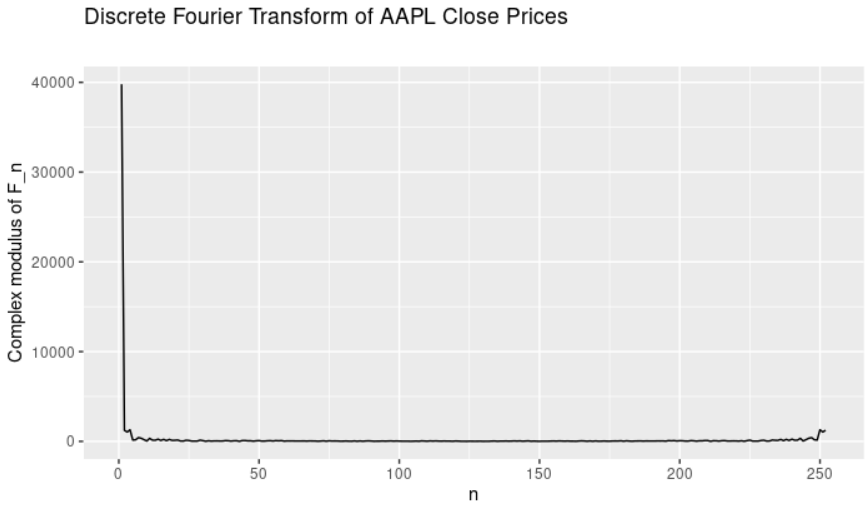


Figure 3.6: Plot of the discrete Fourier transform of the AAPL closing prices

Unlike the examples that were manufactured to be periodic, the discrete Fourier transform of this sequence does not exhibit any peaks. Although there is a peak at $n = 0$, it is not significant for the following reasons:

By definition, the output of the discrete Fourier transform at $n = 0$ is just the sum of all data points in the sequence $\{f_k\}$:

$$F_{(0)} = \sum_{k=0}^{N-1} f_k \, e^{-2\pi i(0)k/N} = \sum_{k=0}^{N-1} f_k \, e^{(0)} = \sum_{k=0}^{N-1} f_k \tag{3.4}$$

Furthermore - from a more intuitive point of view - a frequency of $\frac{n}{N} = \frac{(0)}{N} = 0$ does not have any sort of useful interpretation. Thus, the peak at $n = 0$ can be disregarded.

With this in mind, consider the previous plot, but now without the value at $n = 0$:

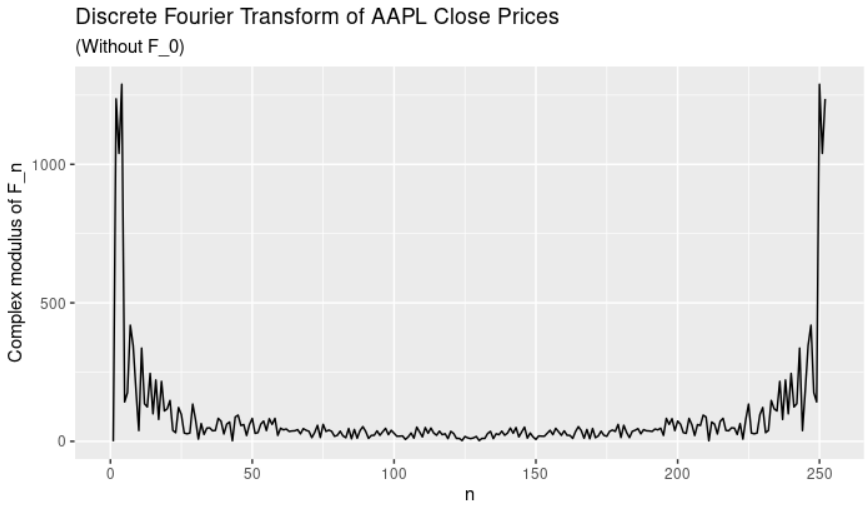


Figure 3.7: Same plot as before, but now with F_0 omitted

Although there are some peaks visible, they are small in comparison to the peak at $n = 0$, and their interpretation as frequency components would not translate well to a strategy involving short-term trends, as these hypothetical frequencies would correspond to longer-term periods.

In order to combine this approach with the probability of short-term trends, one would be interested in looking for shorter-term periodicities, such as those corresponding to periods of 3, 4, 5, or 6 days. These periodicities would correspond to peaks at $n = 84$, $n = 63$, $n \approx 50$, and $n = 42$, respectively, in these 1 year stock data sets. However, when the plot of the discrete Fourier transform is zoomed in to this interval:

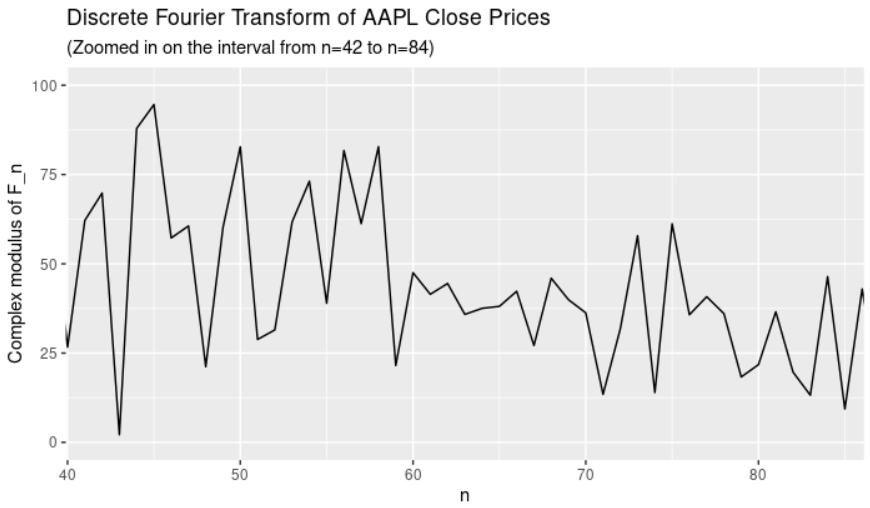


Figure 3.8: Same plot as before, but now zoomed in to where short-term periodicities would be

...no peaks are apparent. Thus, we observe that AAPL does not exhibit any useful periodicities over this year of closing prices.

3.2.2 Overall Results

The results presented for AAPL are representative of the results for the rest of the stocks examined in this analysis. Each stock was examined in the same manner as demonstrated with AAPL, but none exhibited any periodicity. All results were similar in nature to the ones depicted here; none of significance were found.

Chapter 4

Conclusions

4.1 Conclusions from Chapter 2

In Chapter 2, the probability of two consecutive decreases given one decrease occurring in a sequence of random numbers was calculated, and this was then compared to the probability of the same short-term trend occurring in stock data. After noticing that the probability in stock data differed from the probability in a sequence of random numbers, the Kolmogorov-Smirnov test was applied to test whether the probabilities differed by a statistically significant amount. At the $\alpha = 0.05$ significance level, it was found that the value of the probability in all 30 stocks examined differed from the theoretical value of $\frac{1}{3}$ calculated and observed for sequences of random numbers.

The main conclusion to be drawn from this section is that stock price data is not drawn from a random distribution. Although this may seem obvious - stock data is undeniably influenced by many economic factors - the merit of this section lies in proving this from a mathematical standpoint. Additionally, the fact that the value of the probability of two consecutive decreases given one decrease in stock data was close to $\frac{1}{2}$ may indicate that the chance of a stock's price decreasing given a previous decrease is essentially a coin toss; that is, a previous decrease does not better nor worsen the odds of any particular outcome.

4.2 Conclusions from Chapter 3

In Chapter 3, the utility of the Discrete Fourier Transform at detecting periodicity in data was examined, and then it was applied to stock price data. However, no periodicities were observed in any of the 30 stocks examined. Even when the plots of the discrete Fourier transforms were zoomed in on where short-term periodicities would lie, no peaks were found that would indicate any sort of notable periodicity.

The main conclusion to be drawn from this chapter is that stock data does not exhibit periodicity. Again, this may seem like an obvious conclusion to make; if stock data contained a notable periodicity, this would likely be capitalized upon by the market. Realistically speaking, though: as the economy evolves, it is highly unlikely that a stock's price would exhibit periodicity.

4.3 Combining Both Results to Create a Trading Strategy

Theoretically speaking: if short-term trends have a certain probability of occurring in stock data, and if stock data is periodic over some interval, then these traits could be capitalized on in order to create a profitable trading strategy.

In practice, however, the probability of the short-term trends examined in this analysis was close to $\frac{1}{2}$; that is, essentially a coin toss, and none of the stocks examined exhibited any periodicity. Hypothetically, a profitable strategy over a longer term may be possible using these concepts (with great emphasis on net profit and long term); realistically speaking, though, a strategy involving these attributes would not outperform benchmark indices such as the S&P 500.

4.4 Suggestions for Future Study

The approaches introduced in this research could be expanded upon in the following manners.

First of all, the same concepts presented here could be applied over shorter timeframes (e.g. daily, weekly) with more frequent observations. Furthermore, a larger and more diverse array of stocks than just the 30 stocks in the Dow Jones Industrial Average could be analyzed.

Next, the probability of short-term trends could be examined further. Although it is not statistically close to the probability of $\frac{1}{3}$ found in random numbers, it was certainly close to $\frac{1}{2}$. The probability could be determined for more stocks, and this could be analyzed to see how the probability is distributed. From there, it could be analyzed if its fall within a certain bound, and how far out of that bound it has to fall in order to be profitable.

Lastly, the output of the discrete Fourier transform could be examined in more detail. Although no notable peaks were noticed, it remains to be seen if smaller peaks/behaviors can be used to generate profit over a longer term.

Bibliography

[1] Eric W. Weisstein. Kolmogorov-smirnov test. https://mathworld.wolfram.com/Kolmogorov-SmirnovTest.html. Accessed: 2023-04-03.

[2] Eric W. Weisstein. Discrete fourier transform. https://mathworld.wolfram.com/DiscreteFourierTransform.htm Accessed: 2023-04-03.

Noah S. Chough

nsc5179@psu.edu

EDUCATION

Penn State Erie, The Behrend College
Bachelor of Science in Applied Mathematics — *Graduation: Spring 2023*
Certificate in Actuarial Mathematics and Statistics
Schreyer Honors College

RELEVANT EXPERIENCE

Peer Tutor, Penn State Erie — *Spring 2020*

- Helped clarify difficult subjects and communicated effective studying methods to students

Math Grader, Penn State Erie — *Fall 2019*

- Graded students' Calculus II homework assignments and noted areas for improvement

RESEARCH EXPERIENCE

Undergraduate Honors Thesis, Schreyer Honors College — *Fall 2022 - Spring 2023*

- Analyzed stock price data using probability theory and the Discrete Fourier Transform
- Presenting research at Penn State Erie's Sigma Xi conference in Spring 2023

HONORS/AWARDS

Member, Pi Mu Epsilon — *Fall 2021 - Present*
Member, Lambda Sigma Honor Society — *Fall 2019 - Spring 2020*