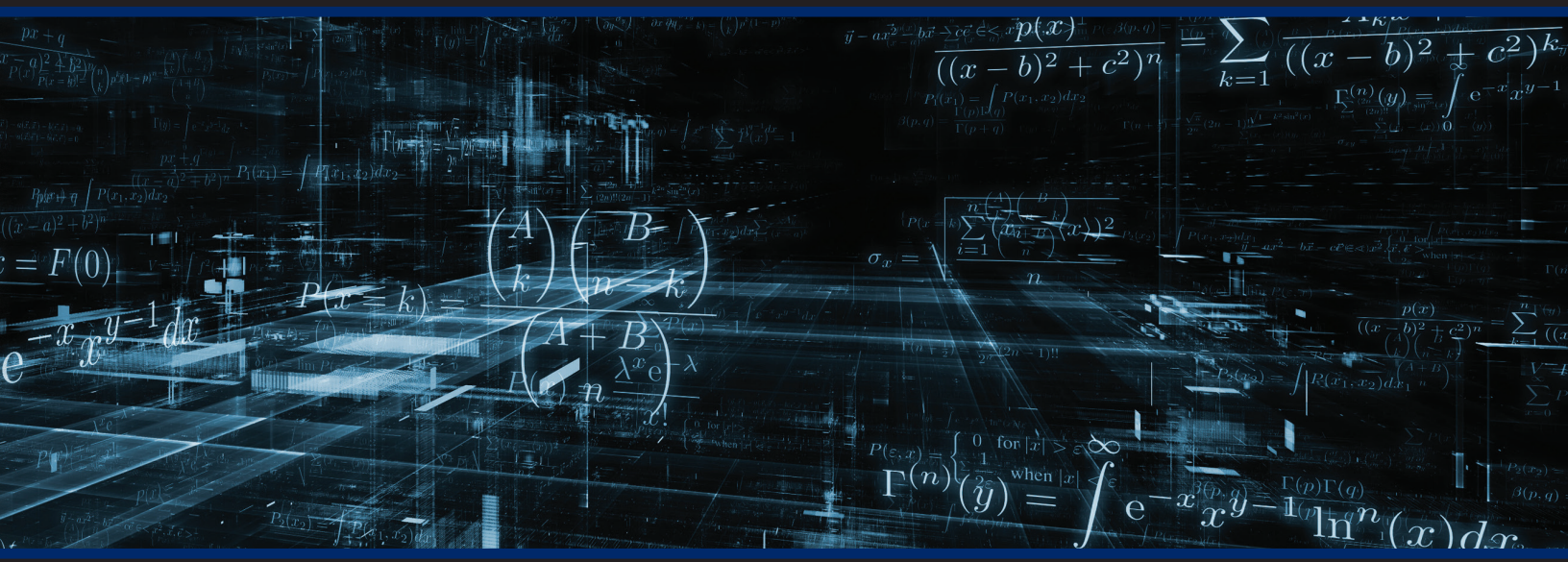
Quantitative Analysis *of* Market Data

a Primer



Adam Grimes

Quantitative Analysis *of* Market Data

a Primer

Adam Grimes

Hunter Hudson Press, New York, New York
MMXV

Published by Hunter Hudson Press, New York, NY.

This book is set in Garamond Premier Pro, using page proportions and layout principles derived from Medieval manuscripts and early book designs, condified in the work of J. A. van de Graaf.

ISBN-13: 978-1511557313

ISBN-10: 1511557311

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Quantitative Analysis *of* Market Data: *a Primer*

If someone separated the art of counting and measuring and weighing from all the other arts, what was left of each of the others would be, so to speak, insignificant.

-Plato

Many years ago I was struggling with trying to adapt to a new market and a new time frame. I had opened a brokerage account with a friend of mine who was a former floor trader on the Chicago Board of Trade (CBOT), and we spent many hours on the phone discussing philosophy, life, and markets. Doug once said to me, "You know what your problem is? The way you're trying to trade ... markets don't move like that." I said yes, I could see how that could be a problem, and then I asked the critical question: "How do markets move?" After a few seconds' reflection, he replied, "I guess I don't know, but not like that"—and that was that. I returned to this question often over the next decade—in many ways it shaped my entire thought process and approach to trading. Everything became part of the quest to answer the all-important question: how do markets really move?

Many old-school traders have a deep-seated distrust of quantitative techniques, probably feeling that numbers and analysis can never replace hard-won intuition. There is certainly some truth to that, but traders can use a rigorous statistical framework to support the growth of real market intuition. Quantitative techniques allow us to peer deeply into the data and to see relationships and factors that

might otherwise escape our notice. These are powerful techniques that can give us profound insight into the inner workings of markets. However, in the wrong hands or with the wrong application, they may do more harm than good. Statistical techniques can give misleading answers, and sometimes can create apparent relationships where none exist. I will try to point out some of the limitations and pitfalls of these tools as we go along, but do not accept anything at face value.

Some Market Math

It may be difficult to get excited about market math, but these are the tools of the trade. If you want to compete in any field, you must have the right skills and the right tools; for traders, these core skills involve a deep understanding of probabilities and market structure. Though some traders do develop this sense through the school of hard knocks, a much more direct path is through quantifying and understanding the relevant probabilities. It is impossible to do this without the right tools. This will not be an encyclopedic presentation of mathematical techniques, nor will we cover every topic in exhaustive detail. This is an introduction; I have attempted a fairly in-depth examination of a few simple techniques that most traders will find to be immediately useful. If you are interested, take this as a starting point and explore further. For some readers, much of what follows may be review, but even these readers may see some of these concepts in a slightly different light.

The Need to Standardize

It's pop-quiz time. Here are two numbers, both of which are changes in the prices of two assets. Which is bigger: 675 or 0.01603? As you probably guessed, it's a trick question, and the best answer is: "What do you mean by bigger?" There are tens of thousands of assets trading around the world, at price levels ranging from fractions of pennies to many millions of dollars, so it is not possible to compare changes across different assets if we consider only the nominal amount of the change. Changes must at least be standardized for price levels; one common solution is to use percentages to adjust for the starting price difference between each asset.

When we do research on price patterns, the first task is usually to convert the raw prices into a return series, which may be a simple percentage return calculated according to this formula:

$$\text{Percent return} = \frac{price_{today}}{price_{yesterday}} - 1$$

In academic research, it is more common to use logarithmic returns, which are also equivalent to

continuously compounded returns.

$$\text{Logarithmic return} = \log\left(\frac{price_{today}}{price_{yesterday}}\right)$$

For our purposes, we can treat the two more or less interchangeably. In most of our work, the returns we deal with are very small; percentage and log returns are very close for small values, but become significantly different as the sizes of the returns increase. For instance, a 1% simple return = 0.995% log return, but a 10% simple return = 9.5% continuously compounded return. Academic work tends to favor log returns because they have some mathematical qualities that make them a bit easier to work with in many contexts, but the most important thing is not to mix percents and log returns.

It is also worth mentioning that percentages are not additive. In other words, a $10 loss followed by a $10 gain is a net zero change, but a 10% loss followed by a 10% gain is a net loss. (However, logarithmic returns *are* additive.)

Standardizing for Volatility

Using percentages is obviously preferable to using raw changes, but even percent returns (from this point forward, simply "returns") may not tell the whole story. Some assets tend to move more, on average, than others. For instance, it is possible to have two currency rates, one of which moves an average of 0.5% a day, while the other moves 2.0% on an average day. Imagine they both move 1.5% in the same trading session. For the first currency, this is a very large move, three times its average daily change. For the second, this is actually a small move, slightly under the average.

It is possible to construct a simple average return measure, and to compare each day's return to that average. (Note that if we do this, we must calculate the average of the *absolute values* of returns so that positive and negative values do not cancel each other out.) This method of average absolute returns is not a commonly used method of measuring volatility because there are better tools, but this is a simple concept that shows the basic ideas behind many of the other measures.

Average True Range

One of the most common ways traders measure volatility is in terms of the average range or Average True Range (ATR) of a trading session, bar, or candle on a chart. Range is a simple calculation: subtract the low of the bar (or candle) from the high to find the total range of prices covered in the trading session. The true range is the same as range if the previous bar's close is within the range of

the current bar. However, suppose there is a gap between the previous close and the high or low of the current bar; if the previous close is higher than the current bar's high or lower than the current bar's low, that gap is added to the range calculation—true range is simply the range plus any gap from the previous close. The logic behind this is that even though the space shows as a gap on a chart, an investor holding over that period would have been exposed to the price change; the market did actually trade through those prices. Either of these values may be averaged to get average range or ATR for the asset. The choice of average length is, to some extent, a personal choice and depends on the goals of the analysis, but for most purposes, values between 20 and 90 are probably most useful.

To standardize for volatility, we could express each day's change as a percentage of the ATR. For instance, if a stock has a 1.0% change and normally trades with an ATR of 2.0%, we can say that the day's change was a 0.5 ATR% move. We could create a series of ATR% measures for each day and average them (more properly, average their absolute values) to create an average ATR% measure. However, there is a logical inconsistency because we are comparing close-to-close changes to a measure based primarily on the range of the session, which is derived from the high and the low. This may or may not be a problem, but it is something that must be considered.

Historical Volatility

Historical volatility (which may also be called either statistical or realized volatility) is a good alternative for most assets, and has the added advantage that it is a measure that may be useful for options traders. Historical volatility (Hvol) is an important calculation. For daily returns:

$$Hvol = \text{StandardDeviation}\left[\ln\left(\frac{p_t}{p_{t-1}}\right)\right] * annualizationfactor,$$

$$annualizationfactor = \sqrt{252}$$

where p = price, t = this time period, t – 1 = previous time period, and the standard deviation is calculated over a specific window of time. Annualization factor is the square root of the ratio of the time period being measured to a year. The equation above is for daily data and there are 252 trading days in a year, so the correct annualization factor is the square root of 252. For weekly and month data, the annualization factors are the square roots of 52 and 12, respectively.

For instance, using a 20-period standard deviation will give a 20-bar Hvol. Conceptually, historical volatility is an annualized one standard deviation move for the asset based on the current volatility. If an asset is trading with a 25% historical volatility, we could expect to see prices within +/–25%

of today's price one year from now, if returns follow a zero-mean normal distribution (which they do not.)

Standard Deviation Spike Tool

The standardized measure I use most commonly in my day-to-day work is measuring each day's return as a standard deviation of the past 20 days' returns. I call this a standard deviation spike (or SigmaSpike™), and use it both visually on charts and in quantitative screens. Here is the four-step procedure for creating this measure:

1. Calculate returns for the price series.
2. Calculate the 20-day standard deviation of the returns.
3. BaseVariation = 20-day standard deviation × Closing price.
4. Spike = (Today's close – Yesterday's close) × Yesterday's BaseVariation.

If you have a background in statistics, you will find that your intuitions about standard deviations do not apply here, because this tools looks at volatility over a short time window; very large standard deviation moves are common. Even large, stable stocks will have a few five or six standard deviation moves by this measure in a single year, which would be essentially impossible if these were true standard deviations (and if returns were normally distributed.) The value is in being able to standardize price changes for easy comparison across different assets.

The way I use this tool is mainly to quantify surprise moves. Anything over about 2.5σ or 3.0σ would stand out visually as a large move on a price chart. After spending many years experimenting with many other volatility-adjusted measures and algorithms, this is the one that I have settled on in my day-to-day work. Note also that implied volatility usually tends to track 20-day historical volatility fairly well in most options markets. Therefore, a move that surprises this indicator will also usually surprise the options market unless there has been a ramp-up of implied volatility before an anticipated event. Options traders may find some utility in this tool as part of their daily analytical process as well.

Some Useful Statistical Measures

The market works in the language of probability and chance; though we deal with single outcomes, they are really meaningful only across large sample sizes. An entire branch of mathematics has been developed to deal with many of these problems, and the language of statistics contains powerful tools to summarize data and to understand hidden relationships. Here are a few that you will find useful.

Probability Distributions

Information on virtually any subject can be collected and quantified in a numerical format, but one of the major challenges is how to present that information in a meaningful and easy-to-comprehend format. There is always an unavoidable trade-off: any summary will lose details that may be important, but it becomes easier to gain a sense of the data set as a whole. The challenge is to strike a balance between preserving an appropriate level of detail while creating that useful summary. Imagine, for instance, collecting the ages of every person in a large city. If we simply printed the list out and loaded it onto a tractor trailer, it would be difficult to say anything much more meaningful than "That's a lot of numbers you have there." It is the job of descriptive statistics to say things about groups of numbers that give us some more insight. To do this successfully, we must organize and strip the data set down to its important elements.

One very useful tool is the histogram chart. To create a histogram, we take the raw data and sort it into categories (bins) that are evenly spaced throughout the data set. The more bins we use, the finer the resolution, but the choice of how many bins to use usually depends on what we are trying to illustrate. Figure 1 shows histograms for the daily percent changes of LULU, a volatile momentum stock, and SPY, the S&P 500 index. As traders, one of the key things we are interested in is the number of events near the edges of the distribution, in the tails, because they represent both exceptional opportunities and risks. The histogram charts show that the distribution for LULU has a much wider spread, with many more days showing large upward and downward price changes than SPY. To a trader, this suggests that LULU might be much more volatile, a much crazier stock to trade.

The Normal Distribution

Most people are familiar, at least in passing, with a special distribution called the normal distribution or, less formally, the bell curve. This distribution is used in many applications for a few good reasons. First, it describes an astonishingly wide range of phenomena in the natural world, from physics to astronomy to sociology. If we collect data on people's heights, speeds of water currents, or income distributions in neighborhoods, there is a very good chance that the data will fit normal distribution well. Second, it has some qualities that make it very easy to use in simulations and models. This is a double-edged sword because it is so easy to use that we are tempted to use it in places where it might not apply so well. Last, it is used in the field of statistics because of something called the central limit theorem, which is slightly outside the scope of this book. (If you're wondering, the central limit theorem says that the means of random samples from populations with any distribution will tend to follow the normal distribution, providing the population has a finite mean and variance. Most of the

assumptions of inferential statistics rest on this concept. If you are interested in digging deeper, see Snedecor and Cochran's *Statistical Methods* [1989].)

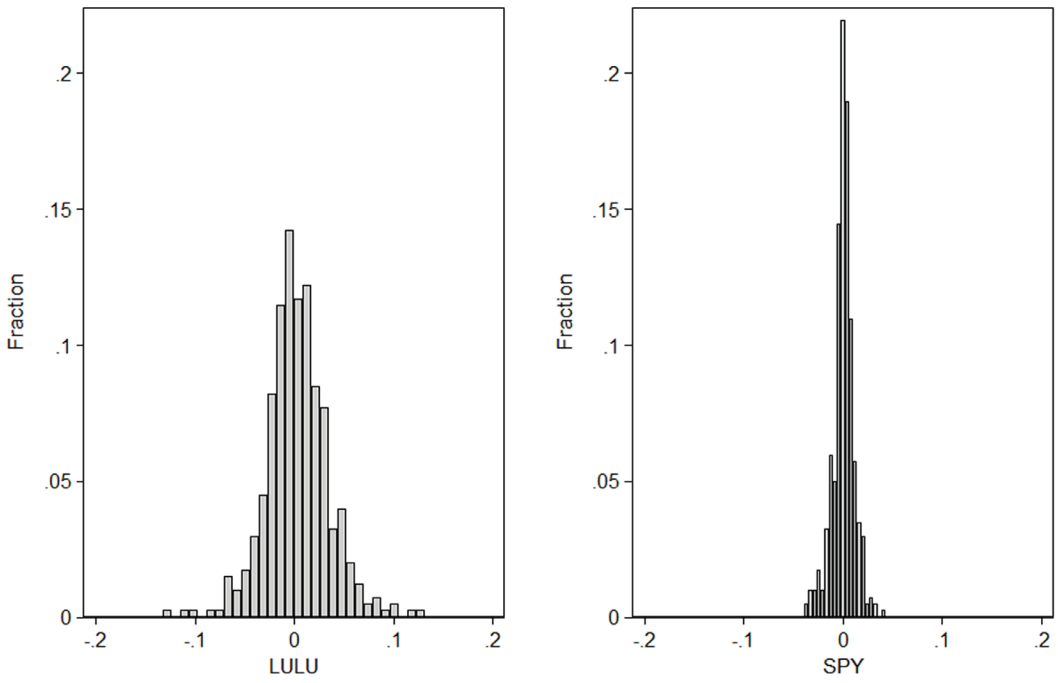


Figure 1 Return distributions for LULU and SPY, 1 June 2009 - 31 December 2010

Figure 2 shows several different normal distribution curves, all with a mean of zero but with different standard deviations. The standard deviation, which we will investigate in more detail shortly, simply describes the spread of the distribution. In the LULU/SPY example, LULU had a much larger standard deviation than SPY, so the histogram was spread further across the x-axis. Two final points on the normal distribution before moving on: Do not focus too much on what the graph of the normal curve looks like, because many other distributions have essentially the same shape; do not automatically assume that anything that looks like a bell curve is normally distributed. Last, and this is very important, normal distributions do an exceptionally poor job of describing financial data. If we were to rely on assumptions of normality in trading and market-related situations, we would dramatically underestimate the risks involved. This, in fact, has been a contributing factor in several recent financial crises over the past two decades, as models and risk management frameworks relying on normal distributions failed.

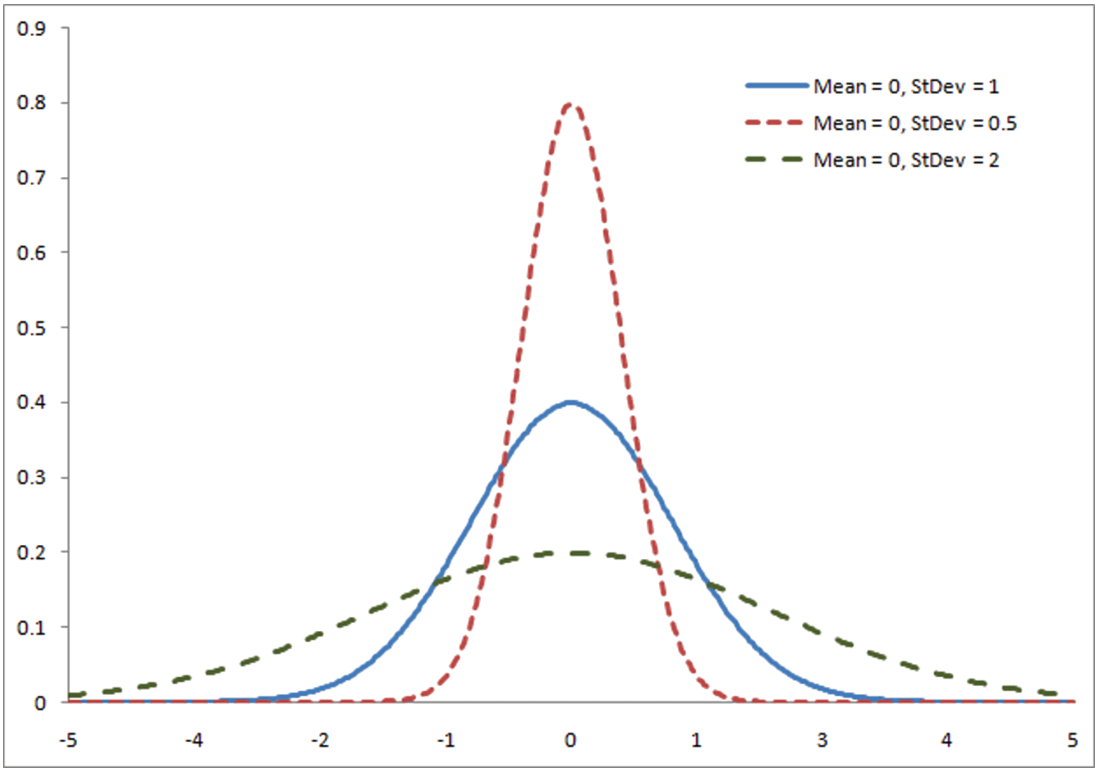


Figure 2 Three Normal Distributions with Different Standard Deviations

Measures of Central Tendency

Looking at a graph can give us some ideas about the characteristics of the data set, but looking at a graph is not actual analysis. A few simple calculations can summarize and describe the data set in useful ways. First, we can look at measures of central tendency, which describe how the data clusters around a specific value or values. Each of the distributions in Figure 2 has identical central tendency; this is why they are all centered vertically around the same point on the graph. One of the most commonly used measures of central tendency is the *mean*, which is simply the sum of all the values divided by the number of values. The mean is sometimes simply called the average, but this is an imprecise term as there are actually several different kinds of averages that are used to summarize the data in different ways.

For instance, imagine that we have five people in a room, ages 18, 22, 25, 33, and 38. The mean age of this group of people is 27.2 years. Ask yourself: How good a representation is this mean?

Does it describe those data well? In this case, it seems to do a pretty good job. There are about as many people on either side of the mean, and the mean is roughly in the middle of the data set. Even though there is no one person who is exactly 27.2 years old, if you have to guess the age of a person in the group, 27.2 years would actually be a pretty good guess. In this example, the mean "works", and describes the data set well.

Another important measure of central tendency is the *median*, which is found by ranking the values from smallest to largest and taking the middle value in the set. If there is an even number of data points, then there is no middle value, and in this case the median is calculated as the mean of the two middle data points. (This is usually not important except in small data sets.) In our example, the median age is 25, which at first glance also seems to explain the data well. If you had to guess the age of any random person in the group, both 25 and 27.2 would be reasonable guesses.

Why do we have two different measures of central tendency? One reason is that they handle outliers, which are extreme values in the tails of distributions, differently. Imagine that a 3,000-year-old mummy is brought into the room. (It sometimes helps to consider absurd situations when trying to build intuition about these things.) If we include the mummy in the group, the mean age jumps to 522.7 years. However, the median (now between 25 and 33) only increases four years to 29. Means tend to be very responsive to large values in the tails, while medians are little affected. The mean is now a poor description of the average age of the people in the room—no one alive is close to 522.7 years old! The mummy is far older and everyone else is far younger, so the mean is now a bad guess for anyone's age.

The median of 29 is more likely to get us closer to someone's age if we are guessing, but, as in all things, there is a trade-off: The median is completely blind to the outlier. If we add a single value and see the mean jump nearly 500 years, we certainly know that a large value has been added to the data set, but we do not get this information from the median. Depending on what you are trying to accomplish, one measure might be better than the other.

If the mummy's age in our example were actually 3,000,000 years, the median age would still be 29 years, and the mean age would now be a little over 500,000 years. The number 500,000 in this case does not really say anything meaningful about the data at all. It is nearly irrelevant to the five original people, and vastly understates the age of the (extraterrestrial?) mummy. In market data, we often deal with data sets that feature a handful of extreme events, so we will often need to look at both median and mean values in our examples. Again, this is not idle theory; these are important concepts with real-world application and meaning.

Measures of Dispersion

Measures of dispersion are used to describe how far the data points are spread from the central values. A commonly used measure is the *standard deviation,* which is calculated by first finding the mean for the data set, and then squaring the differences between each individual point and the mean. Taking the mean of those squared differences gives the *variance* of the set, which is not useful for most market applications because it is in units of price squared. One more operation, taking the square root of the variance, yields the standard deviation. If we were to simply add the differences without squaring them, some would be negative and some positive, which would have a canceling effect. Squaring them makes them all positive and also greatly magnifies the effect of outliers. It is important to understand this because, again, this may or may not be a good thing. Also, remember that the standard deviation depends on the mean in its calculation. If the data set is one for which the mean is not meaningful (or is undefined), then the standard deviation is also not meaningful.

Market data and trading data often have large outliers, so this magnification effect may not be desirable. Another useful measure of dispersion is the interquartile range (IQR), which is calculated by first ranking the data set from largest to smallest, and then identifying the 25th and 75th percentiles. Subtracting the 25th from the 75th (the first and third quartiles) gives the range within which half the values in the data set fall—another name for the IQR is the "middle 50", the range of the middle 50 percent of the values. Where standard deviation is extremely sensitive to outliers, the IQR is almost completely blind to them. Using the two together can give more insight into the characteristics of complex distributions generated from analysis of trading problems. Table 1 compares measures of central tendency and dispersions for the three age-related scenarios. Notice especially how the mean and the standard deviation both react to the addition of the large outliers in examples 2 and 3, while the median and the IQR are virtually unchanged.

Table 1 Comparison of Summary Statistics for the Age Problem

	Example 1	Example 2	Example 3
Person 1	18	18	18
Person 2	22	22	22
Person 3	25	25	25
Person 4	33	33	33
Person 5	38	38	38
The Mummy	Not present	3,000	3,000,000
Mean	27.2	522.7	500,022.7
Median	25.0	29.0	29.0
Standard Deviation	7.3	1,107.9	1,118,023.9
IQR	3.0	6.3	6.3

Inferential Statistics

Though a complete review of inferential statistics is beyond the scope of this brief introduction, it is worthwhile to review some basic concepts and to think about how they apply to market problems. Broadly, inferential statistics is the discipline of drawing conclusions about a population based on a sample from that population, or, more immediately relevant to markets, drawing conclusions about data sets that are subject to some kind of random process. As a simple example, imagine we wanted to know the average weight of all the apples in an orchard. Given enough time, we might well collect every single apple, weigh each of them, and find the average. This approach is impractical in most situations because we cannot, or do not care to, collect every member of the *population* (the statistical term to refer to every member of the set). More commonly, we will draw a small *sample*, calculate statistics for the sample, and make some well-educated guesses about the population based on the qualities of the sample.

This is not as simple as it might seem at first. In the case of the orchard example, what if there were several varieties of trees planted in the orchard that gave different sizes of apples? Where and how we pick up the sample apples will make a big difference, so this must be carefully considered. How many sample apples are needed to get a good idea about the population statistics? What does the distribution of the population look like? What is the average of that population? How sure can we be of our guesses based on our sample? An entire school of statistics has evolved to answer ques-

tions like this, and a solid background in these methods is very helpful for the market researcher.

For instance, assume in the apple problem that the weight of an average apple in the orchard is 4.5 ounces and that the weights of apples in the orchard follow a normal bell curve distribution with a standard deviation of 3 ounces. (Note that this is information that you would not know at the beginning of the experiment; otherwise, there would be no reason to do any measurements at all.) Assume we pick up two apples from the orchard, average their weights, and record the value; then we pick up one more, average the weights of all three, and so on, until we have collected 20 apples. (For the following discussion, the weights really were generated randomly, and we were actually pretty unlucky—the very first apple we picked up was a small one and weighed only 1.07 ounces. Furthermore, the apple discussion is only an illustration. These numbers were actually generated with a random number generator so some negative numbers are included in the sample. Obviously, negative weight apples are not possible in the real world, but were necessary for the latter part of this example using Cauchy distributions. (No example is perfect!)) If we graph the running average of the weights for these first 20 apples, we get a line that looks like Figure 3.

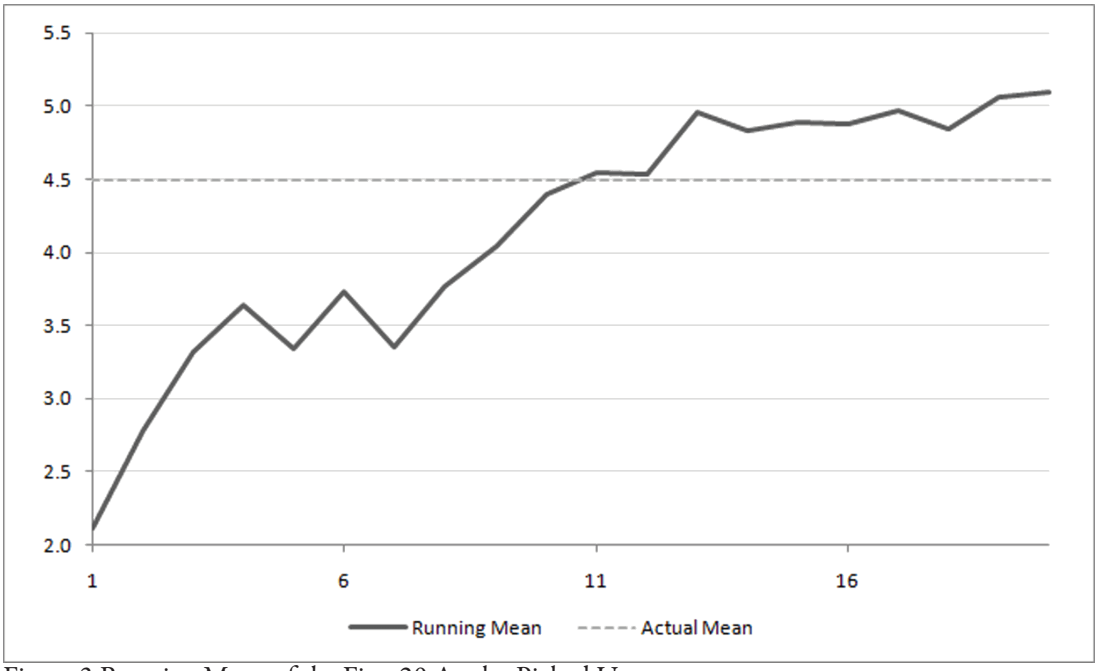


Figure 3 Running Mean of the First 20 Apples Picked Up

What guesses might we make about all the apples in the orchard based on this sample so far? Well, we might reasonably be a little confused because we would expect the line to be settling in on

an average value, but, in this case, it actually seems to be trending higher, not converging. Though we would not know it at the time, this effect is due to the impact of the unlucky first apple; similar issues sometimes arise in real-world applications. Perhaps a sample of 20 apples is not enough to really understand the population; Figure 4 shows what happens if we continue, eventually picking up 100 apples.

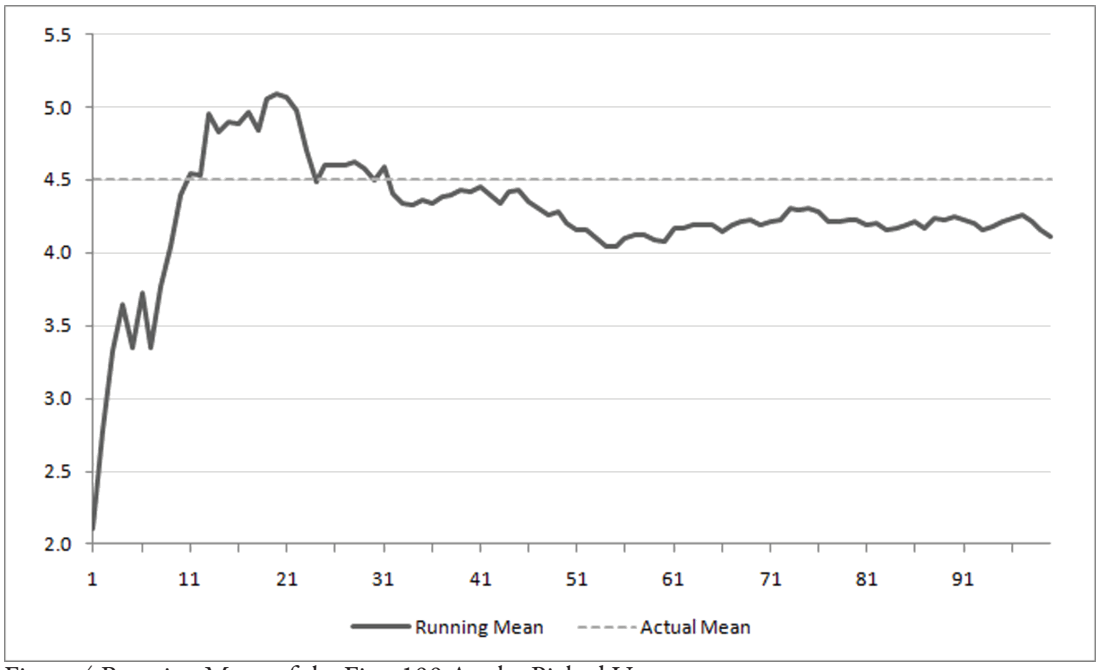


Figure 4 Running Mean of the First 100 Apples Picked Up

At this point, the line seems to be settling in on a value, and maybe we are starting to be a little more confident about the average apple. Based on the graph, we still might guess that the value is actually closer to 4 than to 4.5, but this is just a result of the random sample process—the sample is not yet large enough to assure that the sample mean converges on the actual mean. We have picked up some small apples that have skewed the overall sample, which can also happen in the real world. (Actually, remember that "apples" is only a convenient label and that these values do include some negative numbers, which would not be possible with real apples.) If we pick up many more, the sample average eventually does converge on the population average of 4.5 very closely. Figure 5 shows what the running total might look like after a sample of 2,500 apples.

With apples, the problem seems trivial, but in application to market data there are some thorny issues to consider. One critical question that needs to be considered first is so simple it is often over-

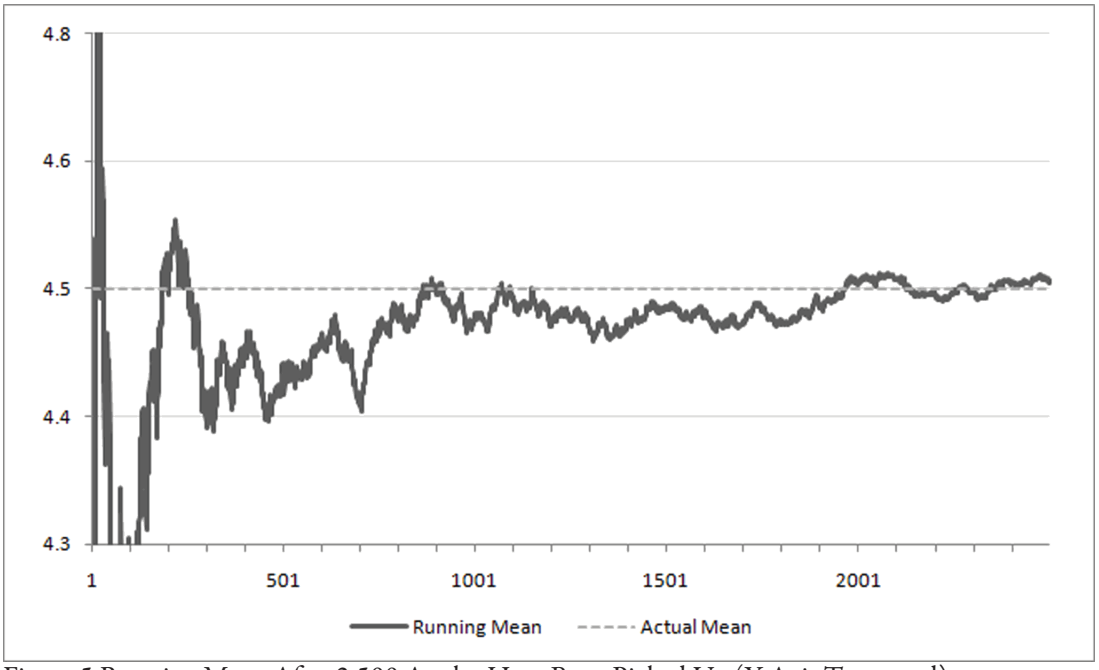


Figure 5 Running Mean After 2,500 Apples Have Been Picked Up (Y-Axis Truncated)

looked: what is the population and what is the sample? When we have a long data history on an asset (consider the Dow Jones Industrial Average, which began its history in 1896, or some commodities for which we have spotty price data going back to the 1400s), we might assume that full history represents the population, but I think this is a mistake. It is probably more correct to assume that the population is the set of all possible returns, both observed and as yet unobserved, for that specific market. The population is everything that has happened, everything that will happen, and also everything that could happen—a daunting concept. All market history—in fact, all market history that will ever be in the future—is only a sample of that much larger population. The question, for risk managers and traders alike, is: what does that unobservable population look like?

In the simple apple problem, we assumed the weights of apples would follow the normal bell curve distribution, but the real world is not always so polite. There are other possible distributions, and some of them contain nasty surprises. For instance, there are families of distributions that have such strange characteristics that the distribution actually has no mean value. Though this might seem counterintuitive and you might ask the question "How can there be no average?" consider the admittedly silly case earlier that included the 3,000,000-year-old mummy. How useful was the mean in describing that data set? Extend that concept to consider what would happen if there were a large

number of ages that could be infinitely large or small in the set? The mean would move constantly in response to these very large and small values, and would be an essentially useless concept.

The Cauchy family of distributions is a set of probability distributions that have such extreme outliers that the mean for the distribution is undefined, and the variance is infinite. If this is the way the financial world works, if these types of distributions really describe the population of all possible price changes, then, as one of my colleagues who is a risk manager so eloquently put it, "we're all screwed in the long run." If the apples were actually Cauchy-distributed (obviously not a possibility in the physical world of apples, but play along for a minute), then the running mean of a sample of 100 apples might look like Figure 6.

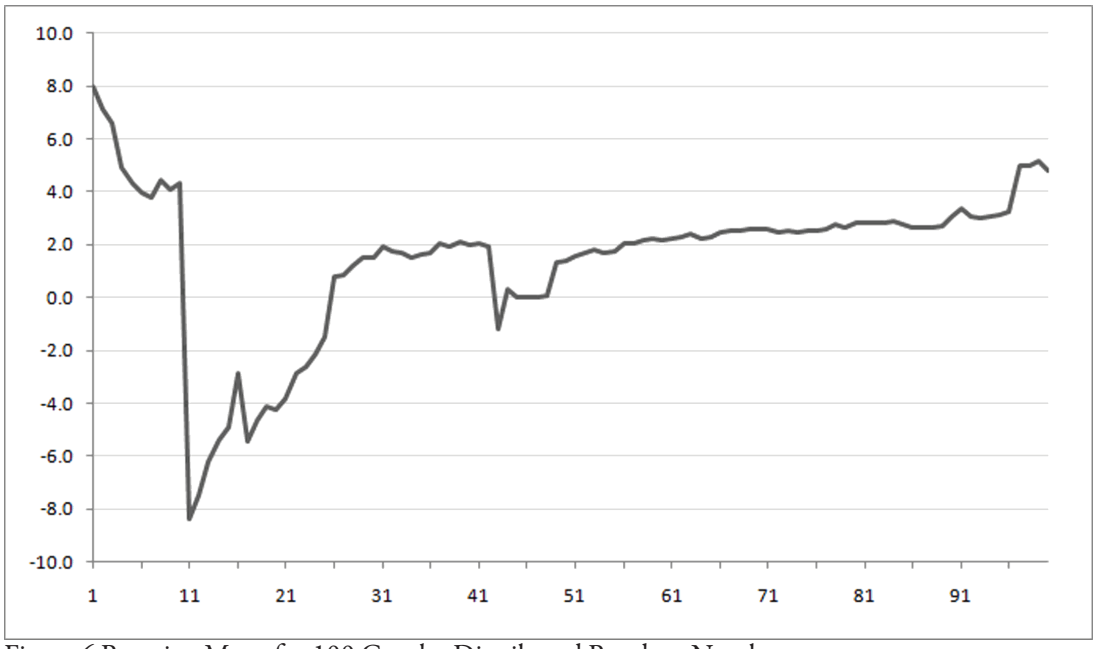


Figure 6 Running Mean for 100 Cauchy-Distributed Random Numbers

It is difficult to make a good guess about the average based on this graph, but more data usually results in a better estimate. Usually, the more data we collect, the more certain we are of the outcome, and the more likely our values will converge on the theoretical targets. Alas, in this case, more data actually adds to the confusion. Based on Figure 6, it would have been reasonable to assume that something strange was going on, but, if we had to guess, somewhere in the middle of graph, maybe around 2.0, might have been a reasonable guess for the mean. Figure 7 shows what might happen if we decide to collect 10,000 Cauchy-distributed numbers in an effort to increase our confidence in the estimate.

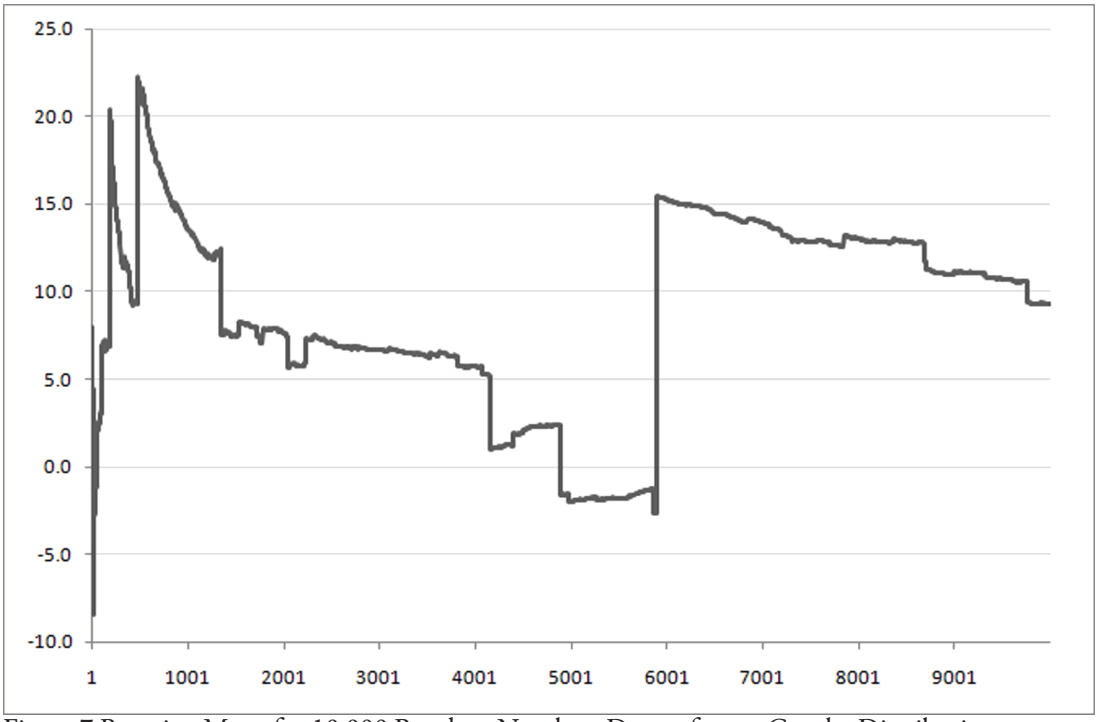


Figure 7 Running Mean for 10,000 Random Numbers Drawn from a Cauchy Distribution

Ouch—maybe we should have stopped at 100. As the sample gets larger, we pick up more very large events from the tails of the distribution, and it starts to become obvious that we have no idea what the actual, underlying average might be. (Remember, there actually *is no mean* for this distribution.) Here, at a sample size of 10,000, it looks like the average will simply never settle down—it is always in danger of being influenced by another very large outlier at any point in the future. As a final word on this subject, Cauchy distributions have undefined means, but the median is defined. In this case, the median of the distribution was 4.5—Figure 8 shows what would have happened had we tried to find the median instead of the mean. Now maybe the reason we look at both means and medians in market data is a little clearer.

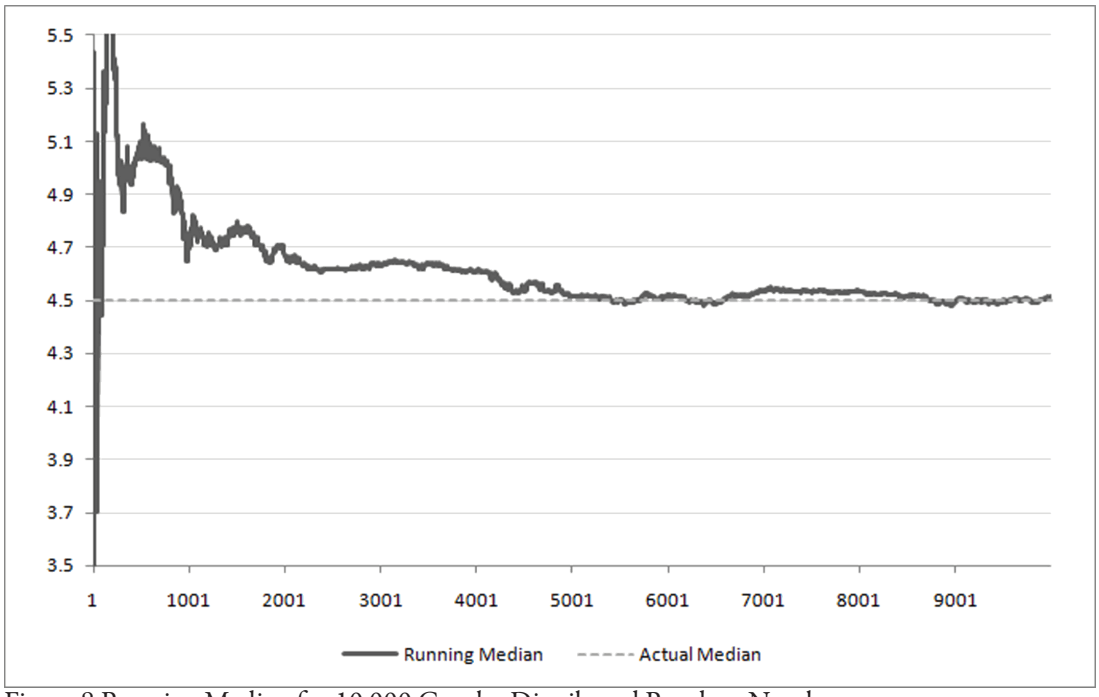


Figure 8 Running Median for 10,000 Cauchy-Distributed Random Numbers

Three Statistical Tools

The previous discussion may have been a bit abstract, but it is important to think deeply about fundamental concepts. Here is some good news: some of the best tools for market analysis are simple. It is tempting to use complex techniques, but it is easy to get lost in statistical intricacies and to lose sight of what is really important. In addition, many complex statistical tools bring their own set of potential complications and issues to the table. A good example is data mining, which is perfectly capable of finding *nonexistent* patterns—even if there are no patterns in the data, data mining techniques can still find them! It is fair to say that nearly all of your market analysis can probably be done with basic arithmetic and with concepts no more complicated than measures of central tendency and dispersion. This section introduces three simple tools that should be part of every trader's tool kit: bin analysis, linear regression, and Monte Carlo modeling.

Statistical Bin Analysis

Statistical bin analysis is by far the simplest and most useful of the three techniques in this section. If we can clearly define a condition we are interested in testing, we can divide the data set into

groups that contain the condition, called the signal group, and those that do not, called the control group. We can then calculate simple statistics for these groups and compare them. One very simple test would be to see if the signal group has a higher mean and median return compared to the control group, but we could also consider other attributes of the two groups. For instance, some traders believe that there are reliable tendencies for some days of the week to be stronger or weaker than others in the stock market. Table 2 shows one way to investigate that idea by separating the returns for the Dow Jones Industrial Average in 2010 by weekdays. The table shows both the mean return for each weekday as well as the percentage of days that close higher than the previous day.

Table 2 Day-of-Week Stats for the DJIA, 2010

Weekday	% Close Up	Mean Return
Monday	59.57%	0.22%
Tuesday	53.85%	(0.05%)
Wednesday	61.54%	0.18%
Thursday	54.90%	(0.00%)
Friday	54.00%	(0.14%)
All	56.75%	0.04%

Each weekday's statistics are significant only in comparison to the larger group. The mean daily return was very small, but 56.75% of all days closed higher than the previous day. Practically, if we were to randomly sample a group of days from this year, we would probably find that about 56.75% of them closed higher than the previous day, and the mean return for the days in our sample would be very close to zero. Of course, we might get unlucky in the sample and find that it has very large or small values, but, if we took a large enough sample or many small samples, the values would probably (very probably) converge on these numbers. The table shows that Wednesday and Monday both have a higher percentage of days that closed up than did the baseline. In addition, the mean return for both of these days is quite a bit higher than the baseline 0.04%. Based on this sample, it appears that Monday and Wednesday are strong days for the stock market, and Friday appears to be prone to sell-offs.

Are we done? Can the book just end here with the message that you should buy on Friday, sell on Monday, and then buy on Tuesday's close and sell on Wednesday's close? Well, one thing to remember is that, no matter how accurate our math is, it describes only the data set we examine. In this case, we are looking at only one year of market data, which might not be enough; perhaps some things change year to year and we need to look at more data. Before executing any idea in the market, it is important to see if it has been stable over time and to think about whether there is a reason it should

persist in the future. Table 3 examines the same day-of-week tendencies for the year 2009.

Table 3 Day-of-Week Stats for the DJIA, 2009

Weekday	% Close Up	Mean Return
Monday	56.25%	0.02%
Tuesday	48.08%	(0.03%)
Wednesday	53.85%	0.16%
Thursday	58.82%	0.19%
Friday	53.06%	(0.00%)
All	53.97%	0.07%

Well, if we expected to find a simple trading system, it looks like we may be disappointed. In the 2009 data set, Mondays still seem to have a higher probability of closing up, but Mondays actually have a *lower* than average return. Wednesdays still have a mean return more than twice the average for all days, but, in this sample, Wednesdays are more likely to close down than the average day is. Based on 2010, Friday was identified as a potentially soft day for the market, but in the 2009 data, it is absolutely in line with the average for all days. We could continue the analysis by considering other measures and using more advanced statistical tests, but this example suffices to illustrate the concept. (In practice, a year is probably not enough data to analyze for a tendency like this.) Sometimes just dividing the data and comparing summary statistics are enough to answer many questions, or at least to flag ideas that are worthy of further analysis.

Significance Testing

This discussion is not complete without some brief consideration of significance testing, but this is a complex topic that even creates dissension among many mathematicians and statisticians. The basic concept in significance testing is that the random variation in data sets must be considered when the results of any test are evaluated, because interesting results sometimes appear by random chance. As a simplified example, imagine that we have large bags of numbers, and we are only allowed to pull numbers out one at a time to examine them. We cannot simply open the bags and look inside; we have to pull samples of numbers out of the bags, examine the samples, and make educated guesses about what the populations inside the bag might look like. Now imagine that you have two samples of numbers and the question you are interested in is: "Did these samples come from the same bag (population)?" If you examine one sample and find that it consists of all 2s, 3s, and 4s, and you compare that with another sample that includes numbers from 20 to 40, it is probably reasonable to conclude that they came from different bags.

This is a pretty clear-cut example, but it is not always this simple. What if you have a sample that has the same number of 2s, 3s, and 4s, so that the average of this group is 3, and you are comparing it to another group that has a few more 4s, so the average of that group is 3.2? How likely is it that you simply got unlucky and pulled some extra 4s out of the same bag as the first sample, or is it more likely that the group with extra 4s actually did come from a separate bag? The answer depends on many factors, but one of the most important in this case would be the sample size, or how many numbers we drew. The larger the sample, the more certain we can be that small variations like this probably represent real differences in the population.

Significance testing provides a formalized way to ask and answer questions like this. Most statistical tests approach questions in a specific, scientific way that can be a little cumbersome if you haven't seen it before. In nearly all significance testing, the initial assumption is that the two samples being compared did in fact come from the same population, that there is no real difference between the two groups. This assumption is called the *null hypothesis*. We assume that the two groups are the same, and then look for evidence that contradicts that assumption. Most significance tests consider measures of central tendency, dispersion, and the sample sizes for both groups, but the key is that the burden of proof lies in the data that would contradict the assumption. If we are not able to find sufficient evidence to contradict the initial assumption, the null hypothesis is accepted as true, and we assume that there is no difference between the two groups. Of course, it is possible that they actually are different and our experiment simply failed to find sufficient evidence. For this reason, we are careful to say something like "We were unable to find sufficient evidence to contradict the null hypothesis," rather than "We have proven that there is no difference between the two groups." This is a subtle but extremely important distinction.

Most significance tests report a p-value, which is the probability that a result at least as extreme as the observed result would have occurred if the null hypothesis were true. This might be slightly confusing, but it is done this way for a reason; it is very important to think about the test precisely. A low p-value might say that, for instance, there would be less than a 0.1% chance of seeing a result at least this extreme if the samples came from the same population, if the null hypothesis were true. It could happen, but it is unlikely, so, in this case, we say we reject the null hypothesis, and assume that the samples came from different populations. On the other hand, if the p-value says there is a 50% chance that the observed results could have appeared if the null hypothesis were true, that is not very convincing evidence against it. In this case, we would say that we have not found significant evidence to reject the null hypothesis, so we cannot say with any confidence that the samples came from different populations. In actual practice, the researcher has to pick a cutoff level for the p-value

(.05 and .01 are common thresholds, but this is only due to convention.) There are trade-offs to both high and low p-values, so the chosen significance level should be carefully considered in the context of the data and the experiment design.

These examples have focused on the case of determining whether two samples came from different populations, but it is also possible to examine other hypotheses. For instance, we could use a significance test to determine whether the mean of a group is higher than zero. Consider one sample that has a mean return of 2% and a standard deviation of 0.5%, compared to another that has a mean return of 2% and a standard deviation of 5%. The first sample has a mean that is 4 standard deviations above zero, which is very likely to be significant, while the second has so much more variation that the mean is well within one standard deviation of zero. This, and other questions, can be formally examined through significance tests. Return to the case of Mondays in 2010 (Table 15.2), which have a mean return of 0.22% versus a mean of 0.04% for all days. This might appear to be a large difference, but the standard deviation of returns for all days is larger than 1.0%. With this information, Monday's outperformance is seen to be less than one-fifth of a standard deviation—well within the noise level and not statistically significant.

Significance testing is not a substitute for common sense and can be misleading if the experiment design is not carefully considered. (Technical note: The t-test is a commonly used significance test, but be aware that market data usually violates the t-test's assumptions of normality and independence. Nonparametric alternatives may provide more reliable results.) One other issue to consider is that we may find edges that are statistically significant (i.e., they pass significance tests), but they may not be economically significant because the effects are too small to capture reliably. In the case of Mondays in 2010, the outperformance was only 0.18%, which is equal to $0.18 on a $100 stock. Is this a large enough edge to exploit? The answer will depend on the individual trader's execution ability and cost structure, but this is a question that must be considered.

Linear Regression

Linear regression is a tool that can help us understand the magnitude, direction, and strength of relationships between markets. This is not a statistics textbook; many of the details and subtleties of linear regression are outside the scope of this book. Any mathematical tool has limitations, potential pitfalls, and blind spots, and most make assumptions about the data being used as inputs. If these assumptions are violated, the procedure can give misleading or false results, or, in some cases, some assumptions may be violated with impunity and the results may not be much affected. If you are interested in doing analytical work to augment your own trading, then it is probably worthwhile to

spend some time educating yourself on the finer points of using this tool. Miles and Shevlin (2000) provide a good introduction that is both accessible and thorough—a rare combination in the literature.

Linear Equations and Error Factors

Before we can dig into the technique of using linear equations and error factors, we need to review some math. You may remember that the equation for a straight line on a graph is:

$y = mx + b$

This equation gives a value for y, to be plotted on the vertical axis, as a function of x, the number on the horizontal axis; x is multiplied by m, which is the slope of the line; higher values of m produce a more steeply sloping line. If m = 0, the line will be flat on the x-axis, because every value of x multiplied by 0 is 0. If m is negative, the line will slope downward. Figure 9 shows three different lines with different slopes. The variable b in the equation moves the whole line up and down on the y-axis. Formally, it defines the point at which the line will intersect the y-axis where x = 0, because the value of the line at that point will be only b. (Any number multiplied by x when x = 0 is 0.) We can safely ignore b for this discussion, and Figure 9 sets b = 0 for each of the lines. Your intuition needs to be clear on one point: the slope of the line is steeper with higher values for m; it slopes upward for positive values and downward for negative values.

One more piece of information can make this simple, idealized equation much more useful. In the real world, relationships do not fall along perfect, simple lines. The real world is messy—noise and measurement error obfuscate the real underlying relationships, sometimes even hiding them completely. Look at the equation for a line one more time, but with one added variable:

$$y = mx + b + \varepsilon$$
$$\varepsilon \sim i.i.d.\ N(0, \sigma)$$

The new variable is the Greek letter epsilon, which is commonly used to describe error measurements in time series and other processes. The second line of the equation (which can be ignored if the notation is unfamiliar) says that ε is a random variable whose values are drawn from (formally, are "independent and identically distributed [i.i.d.] according to") the normal distribution with a mean of zero and a standard deviation of sigma. If we graph this, it will produce a line with jitter, as the points will be randomly distributed above and below the actual line because a different, random ε is added to each data point. The magnitude of the jitter, or how far above and below the line the points are scattered, will be determined by the value of the standard deviation chosen for σ. Bigger values will result in more spread, as the distribution for the error component has more extreme values (see

Figure 2 for a reminder.) Every time we draw this line it will be different, because ε is a random variable that takes on different values; this is a big step if you are used to thinking of equations only in a deterministic way. With the addition of this one term, the simple equation for a line now becomes a random process; this one step is actually a big leap forward because we are now dealing with uncertainty and stochastic (random) processes.

Figure 10 shows two sets of points calculated from this equation. The slope for both sets is the same, m = 1, but the standard deviation of the error term is different. The solid dots were plotted

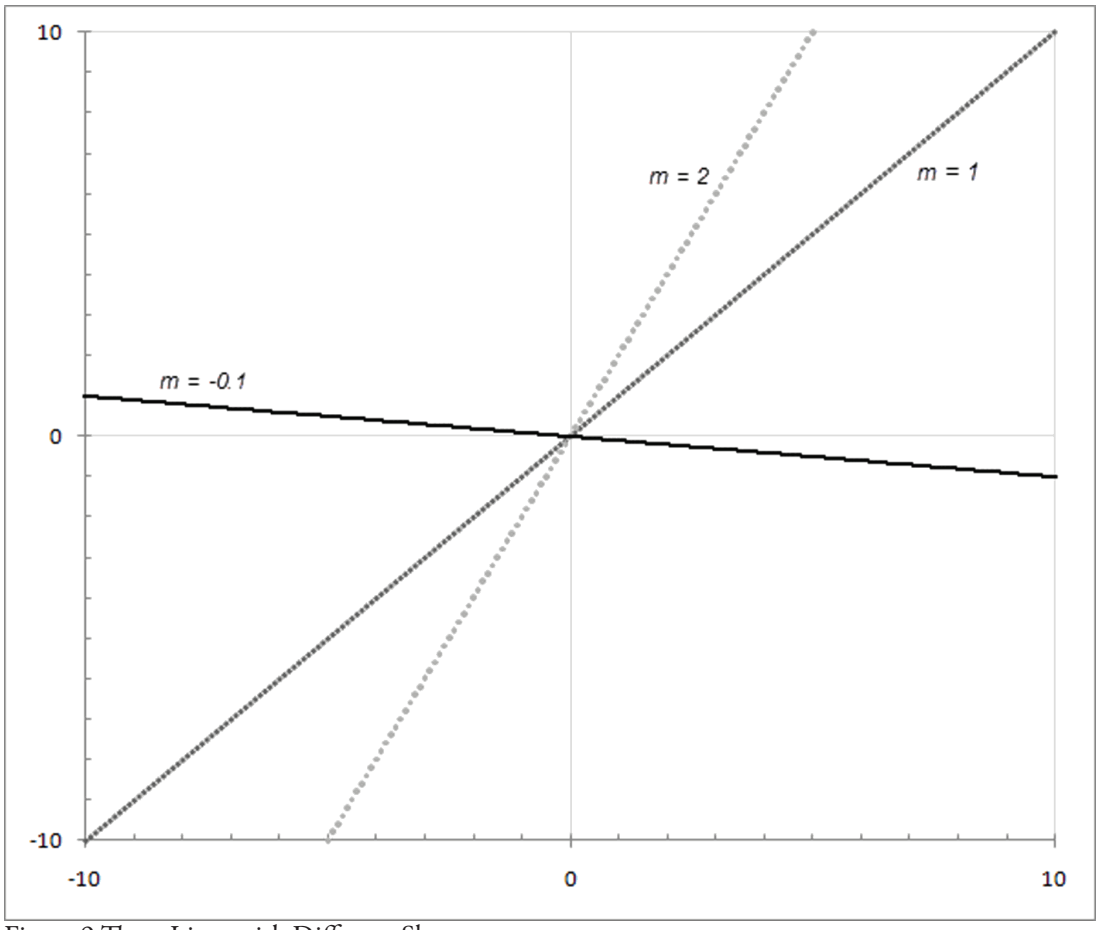


Figure 9 Three Lines with Different Slopes

with a standard deviation of 0.5, and they all lie very close to the true, underlying line; the empty circles were plotted with a standard deviation of 4.0, and they scatter much farther from the line. More variability hides the underlying line, which is the same in both cases—this type of variability is

common in real market data, and can complicate any analysis.

Regression

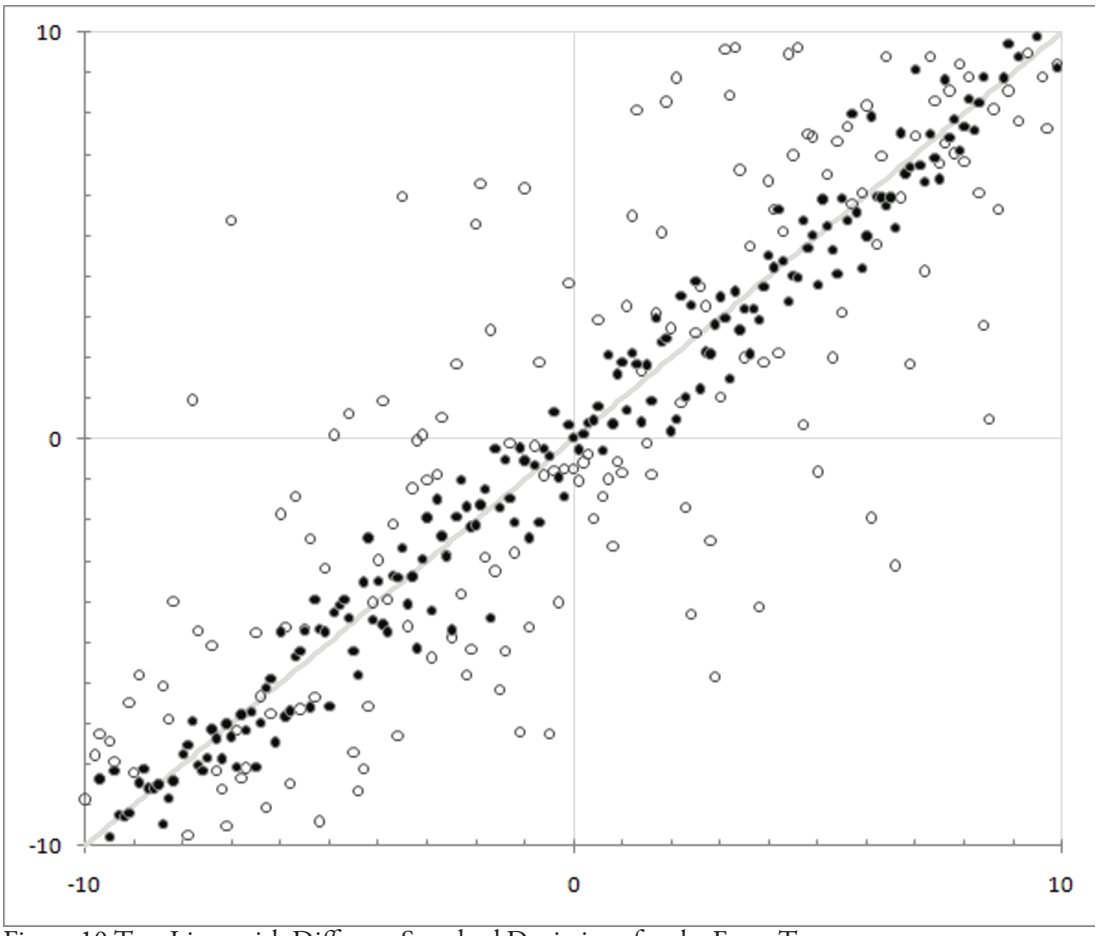


Figure 10 Two Lines with Different Standard Deviations for the Error Terms

With this background, we now have the knowledge needed to understand regression. Here is an example of a question that might be explored through regression: Barrick Gold Corporation (NYSE: ABX) is a company that explores for, mines, produces, and sells gold. A trader might be interested in knowing if, and how much, the price of physical gold influences the price of this stock. Upon further reflection, the trader might also be interested in knowing what, if any, influence the U.S. Dollar Index and the overall stock market (using the S&P 500 index again as a proxy for the entire market) have

on ABX. We collect weekly prices from January 2, 2009, to December 31, 2010, and, just like in the earlier example, create a return series for each asset. It is always a good idea to start any analysis by examining summary statistics for each series. (See Table 4.)

Table 4 Summary Statistics for Weekly Returns, 2009–2010

Ticker	N=	Mean	StDev	Min	Max
ABX	104	0.4%	5.8%	(20.0%)	16.0%
SPX	104	0.3%	3.0%	(7.3%)	10.2%
Gold	104	0.4%	2.3%	(5.4%)	6.2%
USD	104	0.0%	1.3%	(4.2%)	3.1%

At a glance, we can see that ABX, the S&P 500 (SPX), and Gold all have nearly the same mean return. ABX is considerably more volatile, having at least one instance where it lost 20 percent of its value in a single week. Any data series with this much variation, measured by a comparison of the standard deviation to the mean return, has a lot of noise. It is important to notice this, because this noise may hinder the usefulness of any analysis.

A good next step would be to create scatterplots of each of these inputs against ABX, or perhaps a matrix of all possible scatterplots as in Figure 11. The question to ask is which, if any, of these relationships looks like it might be hiding a straight line inside it; which lines suggest a linear relationship? There are several potentially interesting relationships in this table: the Gold/ABX box actually appears to be a very clean fit to a straight line, but the ABX/SPX box also suggests some slight hint of an upward-sloping line. Though it is difficult to say with certainty, the USD boxes seem to suggest slightly downward-sloping lines, while the SPX/Gold box appears to be a cloud with no clear relationship. Based on this initial analysis, it seems likely that we will find the strongest relationships between ABX and Gold and between ABX and the S&P. We also should check the ABX and U.S. Dollar relationship, though there does not seem to be as clear an influence there.

Regression basically works by taking a scatterplot and drawing a best-fit line through it. You do not need to worry about the details of the mathematical process; no one does this by hand, because it could take weeks to months to do a single large regression that a computer could do in a fraction of a second. Conceptually, think of it like this: a line is drawn on the graph through the middle of the cloud of points, and then the distance from each point to the line is measured. (Remember the ε's that we generated in Figure 11? This is the reverse of that process: we draw a line through preexisting points and then measure the ε's (often called the errors).) These measurements are squared, by the same logic that leads us to square the errors in the standard deviation formula, and then the sum

of all the squared errors is calculated. Another line is drawn on the chart, and the measuring and squaring processes are repeated. (This is not precisely correct. Some types of regression are done by a trial-and-error process, but the particular type described here has a closed-form solution that does not require an iterative process.) The line that minimizes the sum of the squared errors is kept as the best fit, which is why this method is also called a least-squares model. Figure 12 shows this best-fit line on a scatterplot of ABX versus Gold.

Let's look at a simplified regression output, focusing on the three most important elements. The

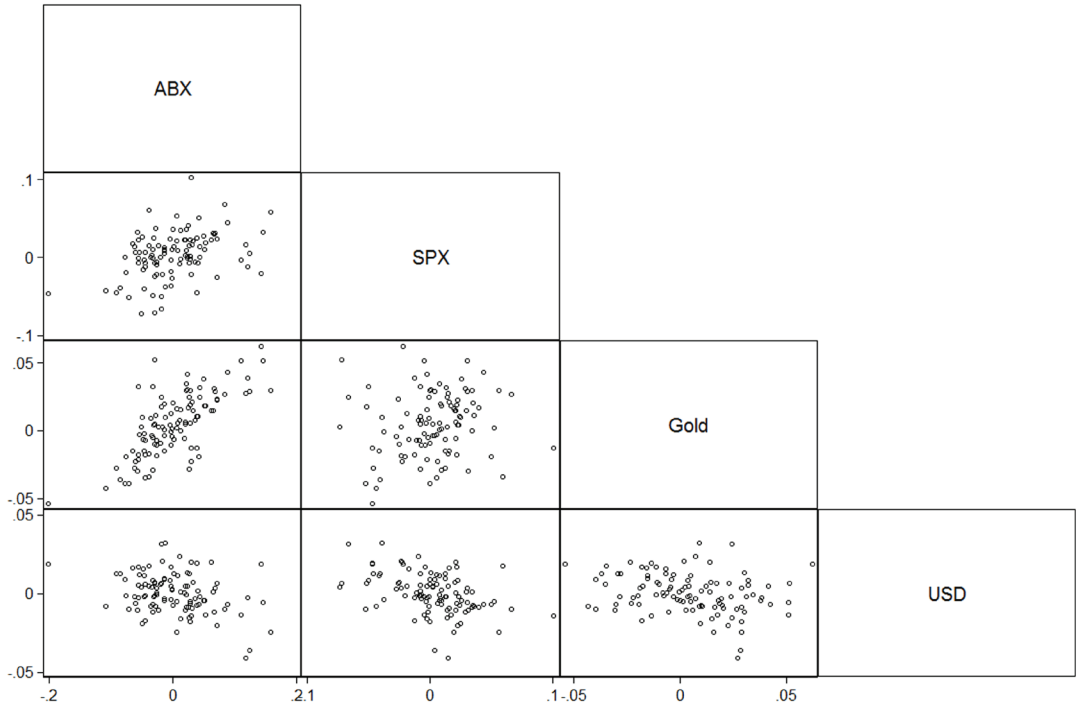


Figure 11 Scatterplot Matrix (Weekly Returns, 2009–2010)

first is the slope of the regression line (m), which explains the strength and direction of the influence. If this number is greater than zero, then the dependent variable increases with increasing values of the independent variable. If it is negative, the reverse is true. Second, the regression also reports a p-value for this slope, which is important. We should approach any analysis expecting to find no relationship; in the case of a best-fit line, a line that shows no relationship would be flat because the dependent variable on the y-axis would neither increase nor decrease as we move along the values of the independent variable on the x-axis. Though we have a slope for the regression line, there is usually also a

lot of random variation around it, and the apparent slope could simply be due to random chance. The p-value quantifies that chance, essentially saying what the probability of seeing this slope would be if there were actually no relationship between the two variables.

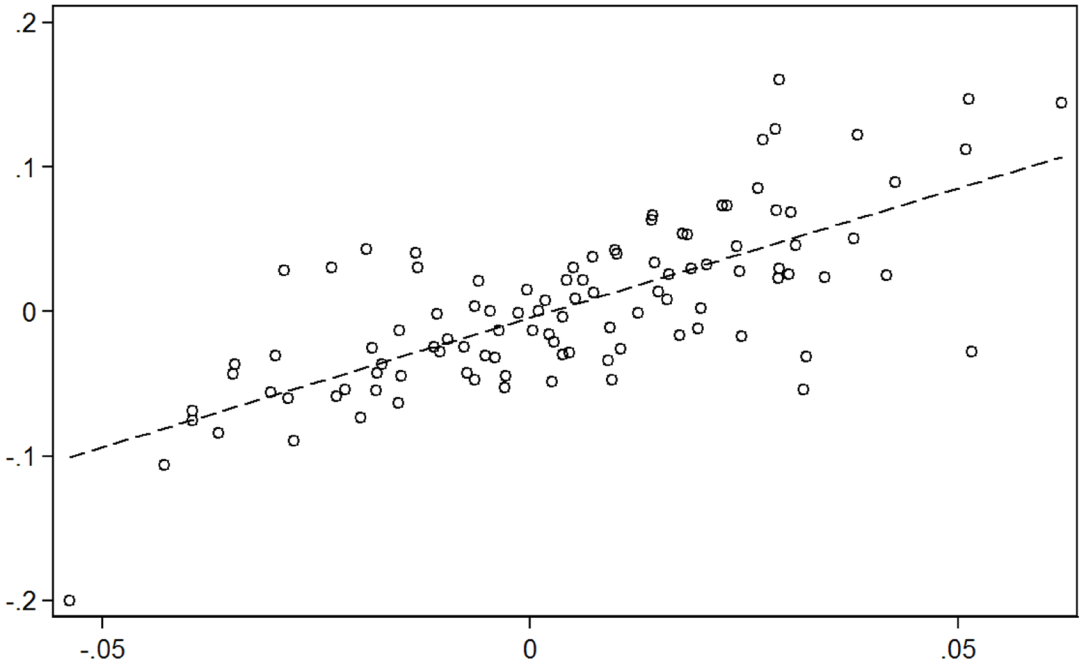


Figure 12 Best-Fit Line on Scatterplot of ABX (Y-Axis) and Gold

The third important measure is R^2 (or R-squared), which is a measure of how much of the variation in the dependent variable is explained by the independent variable. Another way to think about R^2 is that it measures how well the line fits the scatterplot of points, or how well the regression analysis fits the data. In financial data, it is common to see R^2 values well below 0.20 (20%), but even a model that explains only a small part of the variation could elucidate an important relationship. A simple linear regression assumes that the independent variable is the only factor, other than random noise, that influences the values of the independent variable—this is an unrealistic assumption. Financial markets vary in response to a multitude of influences, some of which we can never quantify or even understand. R^2 gives us a good idea of how much of the change we have been able to explain with our regression model. Table 5 shows the regression output for regressing the S&P 500, Gold, and the U.S. Dollar on the returns for ABX.

Table 5 Regression Results for ABX (Weekly Returns, 2009–2010)

	Slope	R^2	p-Value
SPX	0.75	0.15	0.00
Gold	1.79	0.53	0.00
USD	(1.53)	0.11	0.00

In this case, our intuition based on the graphs was correct. The regression model for Gold shows an R^2 value of 0.53, meaning the 53% of ABX's weekly variation can be explained by changes in the price of gold. This is an exceptionally high value for a simple financial model like this, and suggests a very strong relationship. The slope of this line is positive, meaning that the line slopes upward to the right—higher gold prices should accompany higher prices for ABX stock, which is what we would expect intuitively. We see a similar positive relationship to the S&P 500, though it has a much weaker influence on the stock price, as evidenced by the significantly lower R^2 and slope. The U.S. dollar has a weaker influence still (R^2 of 0.11), but it is also important to note that the slope is negative—higher values for the U.S. Dollar Index lead to lower ABX prices. Note that p-values for all of these slopes are less than zero (they are not actually 0.00 as in the table; the values are just truncated to 0.00 in this output), meaning that these slopes are statistically significant.

One last thing to keep in mind is that this tool shows relationships between data series; it will tell you when, if, and how two markets move together. It cannot explain or quantify the causative link, if there is one at all—this is a much more complicated question. If you know how to use linear regression, you will never have to trust anyone's opinion or ideas about market relationships. You can go directly to the data, and ask it yourself.

Monte Carlo Simulation

So far, we have looked at a handful of fairly simple examples and a few that are more complex. In the real world, we encounter many problems in finance and trading that are surprisingly complex. If there are many possible scenarios, and multiple choices can be made at many steps, it is very easy to end up with a situation that has tens of thousands of possible branches. In addition, seemingly small decisions at one step can have very large consequences later on; sometimes effects seem to be out of proportion to the causes. Last, the market is extremely random and mostly unpredictable, even in the best circumstances, so it very difficult to create deterministic equations that capture every possibility. Fortunately, the rapid rise in computing power has given us a good alternative—when faced with an exceptionally complex problem, sometimes the simplest solution is to build a simulation that tries to capture most of the important factors in the real world, run it, and just see what happens.

One of the most commonly used simulation techniques is Monte Carlo modeling or simulation, a term coined by the physicists at Los Alamos in the 1940s in honor of the famous casino. Though there are many variations of Monte Carlo techniques, a basic pattern is:

- Define a number of choices or inputs.
- Define a probability distribution for each input.
- Run many trials with different sets of random numbers for the inputs.
- Collect and analyze the results.

Monte Carlo techniques offer a good alternative to deterministic techniques, and may be especially attractive when problems have many moving pieces or are very path-dependent. Consider this an advanced tool to be used when you need it, but a good understanding of Monte Carlo methods is very useful for traders working in all markets and all time frames.

Be Careful of Sloppy Statistics

Applying quantitative tools to market data is not simple. There are many ways to go wrong; some are obvious, but some are not obvious at all. The first point is so simple it is often overlooked—think carefully before doing anything. Define the problem, and be precise in the questions you are asking. Many mistakes are made because people just launch into number crunching without really thinking through the process, and sometimes having more advanced tools at your disposal may make you more vulnerable to this error. There is no substitute for thinking carefully.

The second thing is to make sure you understand the tools you are using. Modern statistical software packages offer a veritable smörgåsbord of statistical techniques, but avoid the temptation to try six new techniques that you don't really understand. This is asking for trouble. Here are some other mistakes that are commonly made in market analysis. Guard against them in your own work, and be on the lookout for them in the work of others.

Not Considering Limitations of the Tools

Whatever tools or analytical methodology we use, they have one thing in common: none of them are perfect—they all have limitations. Most tools will do the jobs they are designed for very well most of the time, but can give biased and misleading results in other situations. Market data tend to have many extreme values and a high degree of randomness, so these special situations actually are not that rare.

Most statistical tools begin with a set of assumptions about the data you will be feeding them;

the results from those tools are considerably less reliable if these assumptions are violated. For instance, linear regression assumes that there is a relationship between the variables, that this relationship can be described by a straight line (is linear), and that the error terms are distributed i.i.d. N(0, σ). It is certainly possible that the relationship between two variables might be better explained by a curve than a straight line, or that a single extreme value (outlier) could have a large effect on the slope of the line. If any of these are true, the results of the regression will be biased at best, and seriously misleading at worst.

It is also important to realize that any result truly applies to only the specific sample examined. For instance, if we find a pattern that holds in 50 stocks over a five-year period, we might assume that it will also work in other stocks and, hopefully, outside of the five-year period. In the interest of being precise, rather than saying, "This experiment proves ... ," better language would be something like "We find, in the sample examined, evidence that ..."

Some of the questions we deal with as traders are epistemological. Epistemology is the branch of philosophy that deals with questions surrounding knowledge about knowledge: What do we really know? How do we really know that we know it? How do we acquire new knowledge? These are profound questions that might not seem important to traders who are simply looking for patterns to trade. Spend some time thinking about questions like this, and meditating on the limitations of our market knowledge. We never know as much as we think we do, and we are never as good as we think we are. When we forget that, the market will remind us. Approach this work with a sense of humility, realizing that however much we learn and however much we know, much more remains undiscovered.

Not Considering Actual Sample Size

In general, most statistical tools give us better answers with larger sample sizes. If we define a set of conditions that are extremely specific, we might find only one or two events that satisfy all of those conditions. (This, for instance, is the problem with studies that try to relate current market conditions to specific years in the past: sample size = 1.) Another thing to consider is that many markets are tightly correlated, and this can dramatically reduce the effective sample size. If the stock market is up, most stocks will tend to move up on that day as well. If the U.S. dollar is strong, then most major currencies trading against the dollar will probably be weak. Most statistical tools assume that events are independent of each other, and, for instance, if we're examining opening gaps in stocks and find that 60 percent of the stocks we are looking at gapped down on the same day, how independent are those events? (Answer: not independent.) When we examine patterns in hundreds of stocks, we

should expect many of the events to be highly correlated, so our sample sizes could be hundreds of times smaller than expected. These are important issues to consider.

Not Accounting for Variability

Though this has been said several times, it is important enough that it bears repeating: It is never enough to notice that two things are different; it is also important to notice how much variation each set contains. A difference of 2% between two means might be significant if the standard deviation of each were half a percent, but is probably completely meaningless if the standard deviation of each is 5%. If two sets of data appear to be different but they both have a very large random component, there is a good chance that what we see is simply a result of that random chance. Always include some consideration of the variability in your analyses, perhaps formalized into a significance test.

Assuming That Correlation Equals Causation

Students of statistics are familiar with the story of the Dutch town where early statisticians observed a strong correlation between storks nesting on the roofs of houses and the presence of newborn babies in those houses. It should be obvious (to anyone who does not believe that babies come from storks) that the birds did not cause the babies, but this kind of flimsy logic creeps into much of our thinking about markets. Mathematical tools and data analysis are no substitute for common sense and deep thought. Do not assume that just because two things seem to be related or seem to move together that one actually causes the other. There may very well be a real relationship, or there may not be. The relationship could be complex and two-way, or there may be an unaccounted-for and unseen third variable. In the case of the storks, heat was the missing link—homes that had newborns were much more likely to have fires in their hearths, and the birds were drawn to the warmth in the bitter cold of winter.

Especially in financial markets, do not assume that, because two things seem to happen together, they are related in some simple way. These relationships are complex, causation usually flows both ways, and we can rarely account for all the possible influences. Unaccounted-for third variables are the norm, not the exception, in market analysis. Always think deeply about the links that could exist between markets, and do not take any analysis at face value.

Too Many Cuts Through the Same Data

Most people who do any backtesting or system development are familiar with the dangers of overoptimization. Imagine you came up with an idea for a trading system that said you would buy when a set of conditions is fulfilled and sell based on another set of criteria. You test that system on

historical data and find that it doesn't really make money. Next, you try different values for the buy and sell conditions, experimenting until some set produces good results. If you try enough combinations, you are almost certain to find some that work well, but these results are completely useless going forward because they are only the result of random chance.

Overoptimization is the bane of the system developer's existence, but discretionary traders and market analysts can fall into the same trap. The more times the same set of data is evaluated with more qualifying conditions, the more likely it is that any results may be influenced by this subtle overoptimization. For instance, imagine we are interested in knowing what happens when the market closes down four days in a row. We collect the data, do an analysis, and get an answer. Then we continue to think, and ask if it matters which side of a 50-day moving average it is on. We get an answer, but then think to also check 100- and 200-day moving averages as well. Then we wonder if it makes any difference if the fourth day is in the second half of the week, and so on. With each cut, we are removing observations, reducing sample size, and basically selecting the ones that fit our theory best. Maybe we started with 4,000 events, then narrowed it down to 2,500, then 800, and ended with 200 that really support our point. These evaluations are made with the best possible intentions of clarifying the observed relationship, but the end result is that we may have fit the question to the data very well and the answer is much less powerful than it seems.

How do we defend against this? Well, first of all, every analysis should start with careful thought about what might be happening and what influences we might reasonably expect to see. Do not just test a thousand patterns in the hope of finding something, and then do more tests on the promising patterns. It is far better to start with an idea or a theory, think it through, structure a question and a test, and then test it. It is reasonable to add maybe one or two qualifying conditions, but stop there.

Out-of-sample testing

In addition, holding some of the data set for out-of-sample testing is a powerful tool. For instance, if you have five years of data, do the analysis on four of them and hold a year back. Keep the out-of-sample set absolutely pristine. Do not touch it, look at it, or otherwise consider it in any of your analysis—for all practical purposes, it doesn't even exist.

Once you are comfortable with your results, run the same analysis on the out-of-sample set. If the results are similar to what you observed in the sample, you may have an idea that is robust and could hold up going forward. If not, either the observed quality was not stable across time (in which case it would not have been profitable to trade on it) or you overoptimized the question. Either way, it is cheaper to find problems with the out-of-sample test than by actually losing money in the mar-

ket. Remember, the out-of-sample set is good for one shot only—once it is touched or examined in any way, it is no longer truly out-of-sample and should now be considered part of the test set for any future runs. Be very confident in your results before you go to the out-of-sample set, because you get only one chance with it.

Multiple Markets on One Chart

Some traders love to plot multiple markets on the same charts, looking at turning points in one to confirm or explain the other. Perhaps a stock is graphed against an index, interest rates against commodities, or any other combination imaginable. It is easy to find examples of this practice in the major media, on blogs, and even in professionally published research in an effort to add an appearance of quantitative support to a theory.

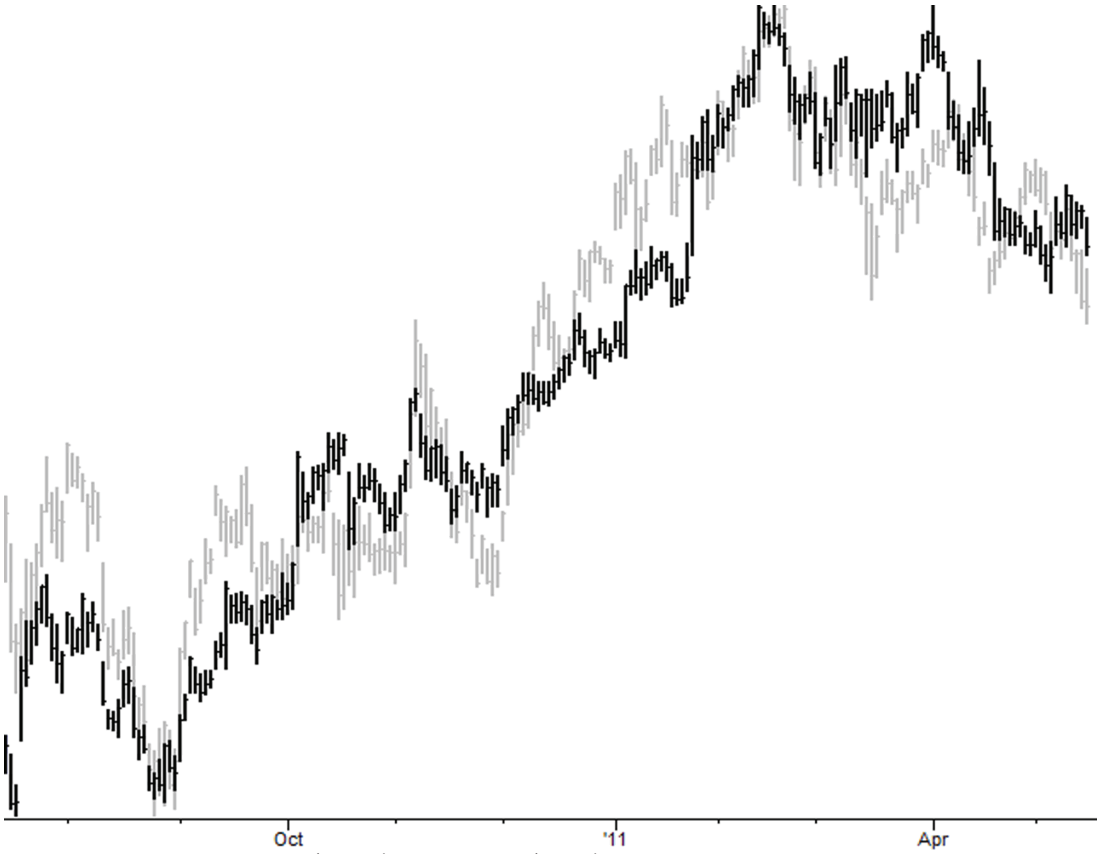


Figure 13 Harley-Davidson (Black) versus XLF (Gray)

This type of analysis nearly always looks convincing, and is also usually worthless. Any two financial markets put on the same graph are likely to have two or three turning points on the graph, and it is almost always possible to find convincing examples where one seems to lead the other. Some traders will do this with the justification that it is "just to get an idea," but this is sloppy thinking—what if it gives you the wrong idea? We have far better techniques for understanding the relationships between markets, so there is no need to resort to techniques like this. Consider Figure 13, which shows the stock of Harley-Davidson, Inc. (NYSE: HOG) plotted against the Financial Sector Index (NYSE: XLF). The two seem to track each other nearly perfectly, with only minor deviations that are quickly corrected. Furthermore, there are a few very visible turning points on the chart, and both stocks seem to put in tops and bottoms simultaneously, seemingly confirming the tightness of the relationship.

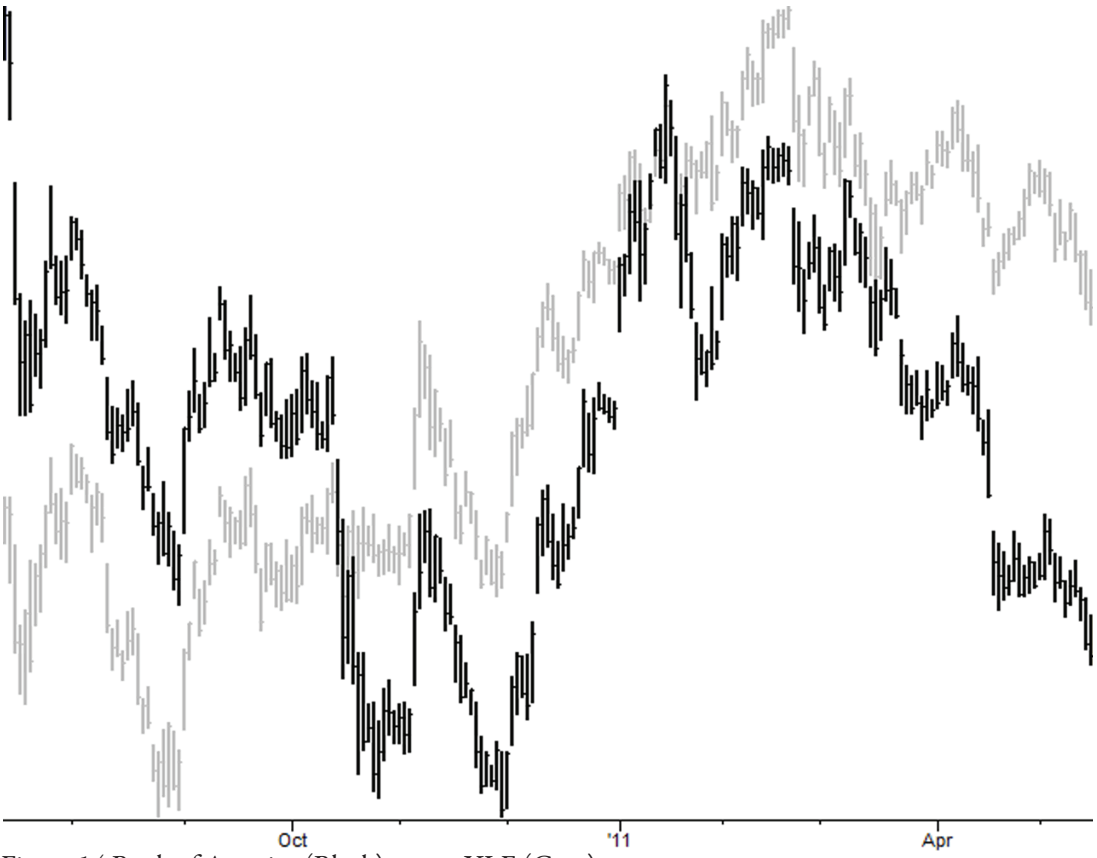


Figure 14 Bank of America (Black) versus XLF (Gray)

For comparison, now look at Figure 14, which shows Bank of America (NYSE: BAC) again

plotted against the XLF over the same time period. A trader who was casually inspecting charts to try to understand correlations would be forgiven for assuming that HOG is more correlated to the XLF than is BAC, but the trader would be completely mistaken. In reality, the BAC/XLF correlation is 0.88 over the period of the chart whereas HOG/XLF is only 0.67. There is a logical link that explains why BAC should be more tightly correlated—the stock price of BAC is a major component of the XLF's calculation, so the two are strongly linked. The visual relationship is completely misleading in this case.

Percentages for Small Sample Sizes

Last, be suspicious of any analysis that uses percentages for a very small sample size. For instance, in a sample size of three, it would be silly to say that "67% show positive signs," but the same principle applies to slightly larger samples as well. It is difficult to set an exact break point, but, in general, results from sample sizes smaller than 20 should probably be presented in "X out of Y" format rather than as a percentage. In addition, it is a good practice to look at small sample sizes in a case study format. With a small number of results to examine, we usually have the luxury of digging a little deeper, considering other influences and the context of each example, and, in general, trying to understand the story behind the data a little better.

It is also probably obvious that we must be careful about conclusions drawn from such very small samples. At the very least, they may have been extremely unusual situations, and it may not be possible to extrapolate any useful information for the future from them. In the worst case, it is possible that the small sample size is the result of a selection process involving too many cuts through data, leaving only the examples that fit the desired results. Drawing conclusions from tests like this can be dangerous.

Summary

Quantitative analysis of market data is a rigorous discipline with one goal in mind—to understand the market better. Though it may be possible to make money in the market, at least at times, without understanding the math and probabilities behind the market's movements, a deeper understanding will lead to enduring success. It is my hope that this little book may challenge you to see the market differently, and may point you in some exciting new directions for discovery.

About the Author

Adam Grimes is a trader, author, and blogger with experience trading virtually every asset class in a multitude of styles and approaches, from scalping to long-term investing. He currently is the CIO of Waverly Advisors (http://www.waverlyadvisors.com), a New York-based research and asset management firm, where he oversees the firm's investment work and writes daily research and market notes that help the firm's clients manage risk and find opportunities.

Adam also blogs and podcasts actively at http://www.adamhgrimes.com, and is the author of *The Art and Science of Technical Analysis: Market Structure, Price Action, and Trading Strategies* (Wiley, 2012). His work focuses on the intersection of human experience and intuition with the statistical realities of the marketplace—seeking a blend of quantitative and discretionary approaches to trading and investing.

Adam is also an accomplished musician, having worked as a professional composer, and classical keyboard artist specializing in historically-informed performance practices. He is also a classically-trained French chef, and served a formal apprenticeship with chef Richard Blondin, a disciple of Paul Bocuse.

Suggested Reading

Campbell, John Y., Andrew W. Lo, and A. Craig MacKinlay. The Econometrics of Financial Markets. Princeton, NJ: Princeton University Press, 1996.

Conover, W. J. Practical Nonparametric Statistics. New York: John Wiley & Sons, 1998.

Feller, William. An Introduction to Probability Theory and Its Applications. New York: John Wiley & Sons, 1951.

Lo, Andrew. "Reconciling Efficient Markets with Behavioral Finance: The Adaptive Markets Hypothesis." Journal of Investment Consulting, forthcoming.

Lo, Andrew W., and A. Craig MacKinlay. A Non-Random Walk Down Wall Street. Princeton, NJ: Princeton University Press, 1999.

Malkiel, Burton G. "The Efficient Market Hypothesis and Its Critics." Journal of Economic Perspectives 17, no. 1 (2003): 59–82.

Mandelbrot, Benoît, and Richard L. Hudson. The Misbehavior of Markets: A Fractal View of Financial Turbulence. New York: Basic Books, 2006.

Miles, Jeremy, and Mark Shevlin. Applying Regression and Correlation: A Guide for Students and Researchers. Thousand Oaks, CA: Sage Publications, 2000.

Snedecor, George W., and William G. Cochran. Statistical Methods, 8th ed. Ames: Iowa State University Press, 1989.

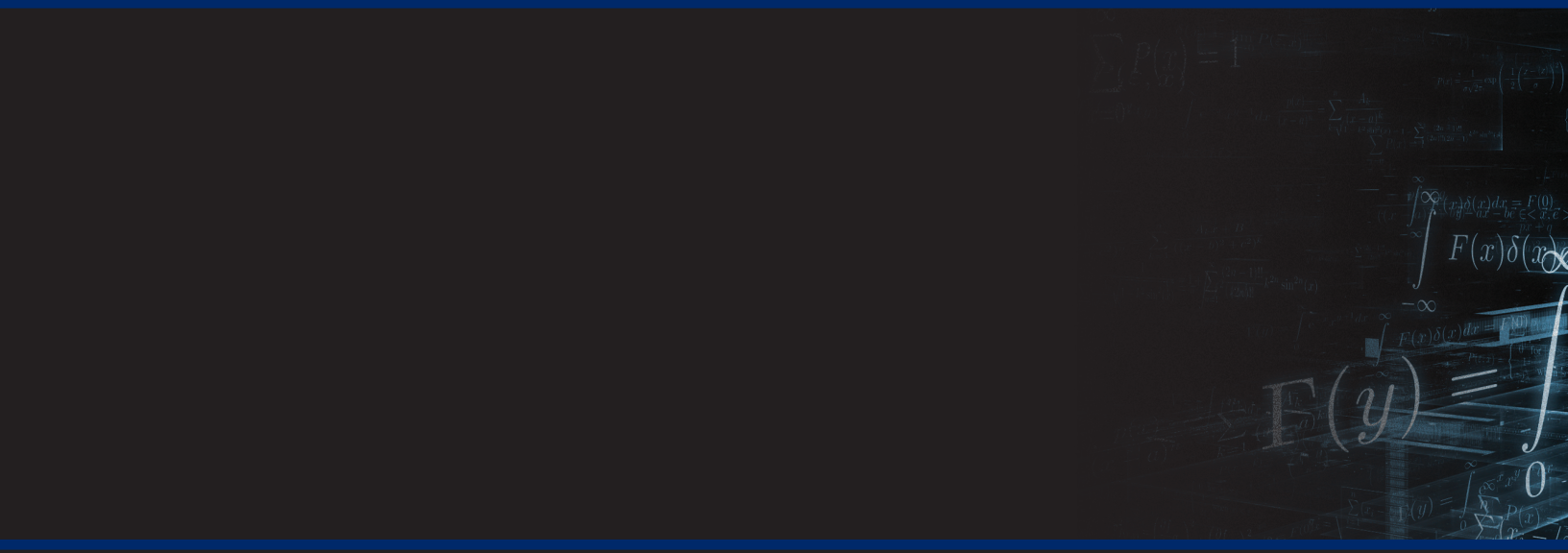
Taleb, Nassim. Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets. New York: Random House, 2008.

Tsay, Ruey S. Analysis of Financial Time Series. Hoboken, NJ: John Wiley & Sons, 2005.

Wasserman, Larry. All of Nonparametric Statistics. New York: Springer, 2010.

Financial Investments • Mathematics

$14.99



Hunter Hundson Press