The Need for Speed

The Importance of Low Latency in a High Frequency Trading Environment

Silvia Tas[1] and Sofia Ternby[2]
Bachelor's Thesis in Finance
Stockholm School of Economics
Spring 2012

Abstract

With the emergence of algorithmic trading, the importance of low latency has suddenly become a key success factor in capital markets. Many have realized having a "slow" trading system poses a new kind of risk and potential loss. We will in this thesis give an example where latency becomes a financial risk factor, potentially affecting the profitability of a bank / liquidity provider by creating a model and studying the Foreign Exchange market setting. The results show that cost increases with latency, more specifically that there is an increasing marginal cost of latency. The results also stress the importance of adapting to the new conditions in the market where high frequency trading is present.

Keywords: Latency, Speed Arbitrage, Foreign Exchange Market

Tutor: Paolo Sodini

Presentation/Submitted: 2012-05-25

[1] 21874@student.hhs.se
[2] 21843@student.hhs.se

Acknowledgements

We would like to thank our tutor Associate Professor Paolo Sodini for his guidance and advice in the process of writing this thesis. Many thanks must also be given to *Skandinaviska Enskilda Banken* AB, SEB for giving us access to valuable market data and analytical tools. To Pablo Landherr at SEB for being our main tutor at the bank and for sharing great insights. And lastly to Dr. Pär Hellström at SEB for acting as a sounding board and for pointing us in the right direction when inspiration sometimes failed to appear.

Table of Contents

1. Introduction

1.1 Background

After releasing the Bank of International Settlement (BIS) report on the FX[3] market in 2010 the BIS organization asked the 4 trillion dollar question referring to the fact that by 2010 the average daily turnover on the global Foreign Exchange market reached $4.0 trillion. This was an increase of 20 percent compared to 2007[4]. One contributing factor was the increased trading activity by "other financial institutes". The growth in this category included the increasing occurrence of high frequency traders and electronic trading, particularly algorithmic trading[5]. This thesis aims to explore how the increase in electronic trading and more specifically how algorithmic trading has affected the market microstructure and market participants in the FX market.

1.1.1 Latency and the Need for Speed

With the emergence of algorithmic trading, the importance of low latency has suddenly become a key success factor in capital markets. Latency can be defined in several ways and the definition depends on the context. The question of latency is often applied on a millisecond or microsecond scale, an environment where no human can compete and algorithms rule the world. Synonymous with delay, latency means a time delay in a system. Latency can be divided into two groups; *spatial latency* and *internal latency*.

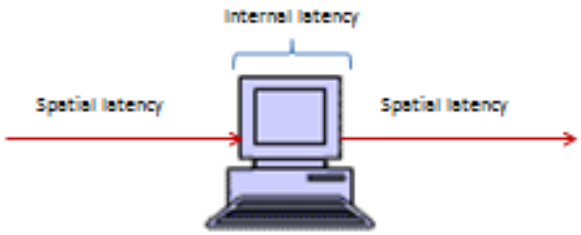


Figure 1- Illustrating the difference between spatial and internal latency. Internal latency occurs inside a system (computer) and spatial latency is a function of distance.

An example of *spatial latency* is the time it takes for a market quote, sent from New York, to reach a counterpart in London. Components affecting the spatial latency are the distance between the

[3] A shortening for the Foreign Exchange market is FX market and will be used throughout the thesis
[4] See BIS-report; *Global Foreign Exchange market activity in 2010*
[5] See BIS-report; *The $4 trillion question: what explains FX growth since the 2007 survey?*

physical locations as well as the characteristics of the media used to transfer the information. Different media such as electrical signals, light (optical) signals, as well as radio signals are used to transfer information, where each media type has its own advantages and disadvantages. Co-location has become a common way of reducing the spatial latency, where trading entities place algorithmic engines in close proximity to the market place. An advantage of co-location is getting important information first and using this to your advantage. Many IT service companies offer proximity services to reduce the physical distance and thus the spatial latency. However, when it comes to the Foreign Exchange market, it is difficult to reduce all spatial latency since the FX market is characterized by geographic dispersion with major trading hubs all over the world. Today the roundtrip time between London and New York ranges from 65[6] milliseconds and up. However different companies compete to be first with sub-60 milliseconds cables over the Atlantic, saving a few precious milliseconds in a highly competitive world[7]. Even though the fight for reducing the latency of the cross-Atlantic cable is fierce, the speed of light will always be the limit when it comes to spatial latency[8].

Internal latency is the time it takes for a system to process received information and act accordingly, i.e. delay due to data processing. A "normal latency level" is difficult, or maybe impossible to define, because it depends on the system, its purpose etc. The fastest systems in the world right now is said to be ultra-low latency defined as below 10 microseconds. However, a more common latency level is probably still in the millisecond range. In this thesis, the focus is primarily the internal latency of FX trading systems.

1.1.2 Foreign Exchange Market Structure and Characteristics

To fully understand the model presented here in, it is important to define the institutional setting of t*he Foreign Exchange market*. The FX market is divided into two participating groups; the market makers or interbank market and the retail market or customers (Frankel 1996). The interbank market consists of the largest commercial banks and securities dealers and is normally characterized by very tight spreads. The spread is defined as the difference between the bid and ask price. The major banks typically trade both in the interbank market and with its retail clients. Retail clients are quoted a price, which consists of the interbank quoted price plus an additional mark-up. This markup results in a slightly wider retail spread. The banks act as both a market maker and taker in the interbank market.

[6] http://www.telegraph.co.uk/technology/news/8753784/The-300m-cable-that-will-save-traders-milliseconds.html#disqus_thread
[7] http://low-latency.com/
[8] The geographic distance between London and New York is 5576 kilometers, corresponding to a theoretical minimum latency of 18.6 millisecond (using the speed of light)

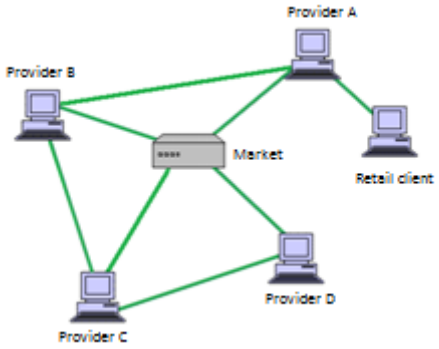


Figure 2- Illustrates how the Foreign Exchange market participants are connected. The providers (representing the interbank participants) can be connected both bilaterally (direct link) and through market-like trading venues. The retail clients are connected to one or more providers depending on the size of the retail client.

Currencies are traded in pairs, each currency with its own international three-letter code. For example trading Euro against the US Dollar has a code of EUR/USD. Characterized by a geographic dispersion, the Foreign Exchange market is open 24/7[9] with the major trading hubs located in London, New York and Singapore.

There are three characteristics particularly distinguishing the FX Market; enormous trading volume, low trade transparency and most of the traded volume is done between dealers (Lyons 2001). The FX market is also characterized by an over-the-counter, OTC, trading structure (Lyons 2001). Market makers are connected through different electronic trading platforms instead of a single exchange. Trading platforms link buyers to sellers for different currencies. Banks and other FX providers can use a number of venues as contact points with the market including among others EBS, Reuters and Currenex. All the mentioned platforms stream out current market prices to the connected market makers, showing at what price the market is willing to sell and/or buy a specific instrument at that point in time.

The Bank for International Settlements (the BIS organization) is a good source of information regarding the FX market. Every third year they publish a report about the state of the FX market including turnover categorized by execution method, product, participants, and currency pair. The report, which compiles information from over 53 central banks[10] all over the world, is a good indicator of the current market situation and where it is heading. The report published in 2010 shows, among other things, a major increase in overall turnover and indicated the execution method called "electronic methods" being one of the major reasons. Electronic execution

[9]The market is only closed from Friday 5.00 PM New York time to early Sunday morning Singapore time.
[10] See BIS-report; *The $4 trillion question: what explains FX growth since the 2007 survey?*

methods include single and multi-bank trading systems and the use of algorithmic trading and represented 41.3 percent of the executions by 2010[11].

Algorithmic trading, a growing trading strategy, is implemented using a system built on mathematical models that execute orders in the market without human intervention. The algorithms have strict rules for making the execution decision and strive to optimize timing, thus the price of the order. The algo's[12] ability to make trading decisions in just a few milliseconds changes the rules in trading compared to 10 years ago when most trading were done by phone via brokers in the interbank market. Through the growth of electronic trading it is hard to tell whether or not the counterparty is a human or a computer. The change towards electronic execution has made the importance of low latency a growing issue for market participants and a fast system is a vital part of operations.

We will in this thesis give an example where latency becomes a financial risk factor, potentially affecting the profitability of a provider[13]. The focus will be on the interbank market that has the highest trading volumes and the tightest spreads, implying a high liquidity and market activity. The frequent price updates, sometimes occurring only milliseconds apart, contribute to the high liquidity and market activity.

1.1.3 Latency arbitrage

Low latency and co-location has become the latest competitive strategy in the financial markets. Interbank dealers, brokerage firms and hedge funds compete to be as close to the market as possible and to introduce ever-faster algorithmic engines. Low latency enables these parties to gain a profit by capturing arbitrage opportunities that occur when the price of an instrument from one provider crosses that of another at a certain point in time[14]. In such a case the trading firms can explore such arbitrage opportunity by buying a currency at a low price and selling it at a higher price. It is argued in this thesis that, a contributing factor to the occurrence of such arbitrage opportunities is internal latency. Furthermore it is argued that latency exposes an interbank dealer to potential loss as counterparties might exploit arbitrage opportunities.

If latency could potentially create a loss through "giving" away arbitrage opportunities, then trading entities/providers that want to minimize losses need to adapt to algorithmic trading and apply low latency trading strategies (Riordan 2010).

[11] See BIS-report; *Global Foreign Exchange market activity in 2010*

[12] Algo = Algorithmic system/engine

[13] Throughout this thesis a provider will be used to represent trading entities, such as banks, that act as market makers, liquidity provider as well as potentially having retail clients.

[14] A client can explore an arbitrage opportunity when there is a gap between the provider's offer(bid) and the market bid (offer)

1.2 Aim of thesis

This study investigates how the level of latency affects the profitability of a provider and how it can result in arbitrage opportunities for counterparties. More specifically we define our questions as follows:

How is the profitability of a provider affected by latency? Does potential loss increase with the level of latency?

To answer the research question a simple model has been created, that has been simulated using market data from the Foreign Exchange market. The result from the model quantifies a measure for cost of latency, which is the number of occurrences of arbitrage opportunities combined with the size of the arbitrage opportunities. Simulation results are further analyzed by regressions and hypothesis testing.

2. Previous Literature/ Related Work

So far cost of latency is a relatively unexplored subject in academia even though it is a heavily discussed subject in business. The research done on the cost of latency has been limited to the equity market, which is true for much of the market microstructure research. This might be due to the fact that equities have been predominantly exchange traded and has resulted in a more transparent and open market compared to the fairly closed interbank market in FX. [15]

Moallemi and Saglam (2010) quantify the cost of latency in trade execution by using a theoretical model. Their study is made using data regarding NYSE common stocks from 1995 and 2005. Since like most papers on the subject has been done on the equity market their results become relatively irrelevant for this thesis. However, Moallemi et Al suggest that the importance of latency increases as more volatile or liquid (i.e. tight spread) assets are being traded, which is similar to the findings presented in this thesis. This could imply that the relationship between latency and volatility and/or spread is true for all assets (or at least for Equities and Foreign Exchange).

Moallemi and Saglam (2010) seem to be one of the few persons quantifying cost of latency. Other areas of the effect of latency have been explored more extensively. There are some papers studying latency impact on liquidity, market quality and price discovery in the equity market. However, no conclusion can be made on whether or not these results can be applied to the Foreign Exchange market, thus making it difficult to apply their findings here in (which aim is to explore cost and not market quality associated with latency).

[15] In the Equity market a private person can trade on the exchange. On the interbank market however, the FX market is not available for private persons.

Moreover, as all of the papers associated with latency are on a graduate level, i.e. not published in any major economic journals[16], there are weaknesses related to using them as sources in this thesis.

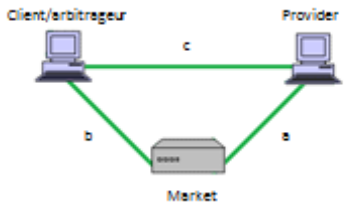
Rather than working with a solid foundation of related work and previous literature, this thesis is exploring relatively unknown territory in academia. The lack of work on the subject does not mean that the subject of latency is irrelevant. It is due to the fact that it is a relatively new phenomenon as a result of technological advances in recent years. Furthermore, as experienced in this study, quantifying cost of latency can be tricky. Creating a model explaining the total cost of latency will need some consideration (which is outside of the scope of this thesis).

3. Methodology

3.1 Model

To explore cost of latency a simple model of reality has been created using the intuition about how arbitrage opportunities occur. The model is a simulation done using real time market data and is explained below.

It is intuitive to think that if there are plenty of arbitrage opportunities in the market, a trader could earn a fortune by buying low and selling high. In the Foreign Exchange market, there are three parties of interest; the client, the provider (bank) and the market. The provider streams out a price to its client for a certain volume of a currency pair. The client can either accept the quoted price and thus make a trade with the provider, or turn the price down in the hope of getting a better price/deal elsewhere, for instance in the market. The relationship between the three parties is shown in the picture below.



a = distance (in time) from market to the provider
b = distance (in time) from market to the client
c = distance (in time) from the client to the provider

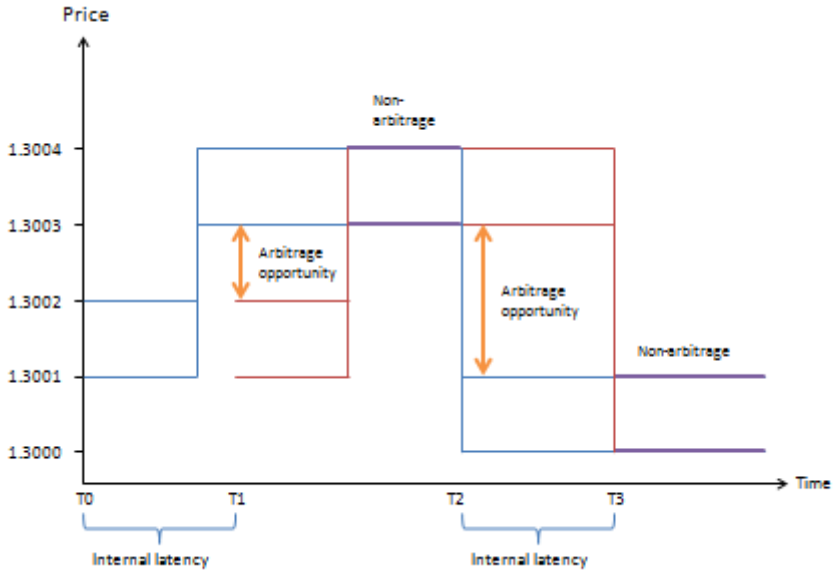
Figure 3 – Illustration of the relationship between the market participants in the model. It is assumed that the client/arbitrageur is connected to both the "market" / the platforms as well as to the provider.

If all three parties have the same quality of wire, that is the same circumstances to obtain the market information at the same time, then spatial latency should not matter, meaning the sum of

[16] where they have been subject of peer to peer review before publishing their findings

b and c equals a (a≈b+c), assuming that distance c is large. Or rather if the internal latency of the provider is larger than the internal latency of the client ($l_p \geq l_c$) then it would follow that $a + l_p \geq b + c + l_c$. Where l_p is the internal latency of the provider and l_c is the internal latency of the client. This implies that spatial latency does not matter and thus only the internal latency will be analyzed in this thesis.

The most basic assumption in the described model is the assumption that in a simplified world, the price that the provider streams out to its clients[17] is exactly the same as the market price, meaning no additional spread is added. In reality however, the price the provider offers to its clients may differ from the market quote reflecting the fact that the provider wants to maximize profit or maximize flows by quoting inside the market. However to simplify the model it is assumed that there is no incentive for the provider to do this. The only difference between the streamed out price and the market price is the time at which the provider streams out his price (see Figure 4)



Blue line= market bid and offer

Red lines = provider bid and offer

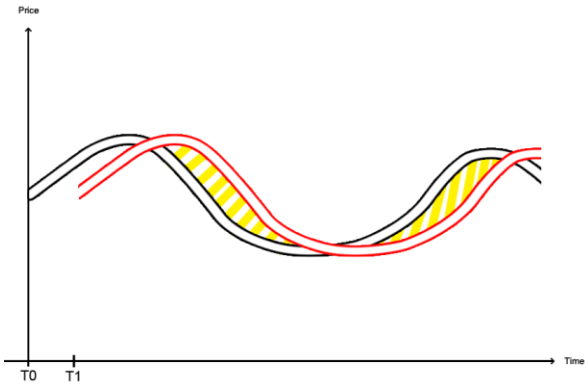
Purple line = when market and provider bid and offer are the same

Figure 4 -Illustrates the relationship between a provider's price and the market price, when latency exists. When market bid is larger than provider's offer (like in t_1) or when the market offer is smaller than provider's bid (like in t_2) an arbitrage opportunity exists. For a numerical example see explanation of Figure 1 in the Appendix.

Looking at the model above, at t_0 the provider receives the price from the market. At t_1, the provider streams out the exact same price as the market had in t_0. The difference between t_0 and t_1

[17] The clients can be both retail and market makers. Remember a market maker can both trade on the interbank market and with retail client. Retail clients using algorithmic trading have been less unlikely but there is a trend towards a wider use of algorithmic trading among all kinds of trading entities.

is due to the internal latency of the provider. If the provider has a low latency, meaning the time for the price to go through the pricing system of the provider and back to reach its receiver is short, then the difference between the market price in t_0 and t_1 will be small. This is due to the fact that it is less likely that the price has moved substantially between t_0 and t_1. It then follows that if the latency for the provider is higher than the latency for the client, that is $a + l_p > b + c + l_c$ and the market quote has moved away from the original spread, it can be argued that the provider is "giving away" a speed arbitrage opportunity. If the client has a lower latency than the provider, the client will have knowledge of the fact that the price has moved in a favorable way and will be able to send an order to the provider that will be accepted by the provider before the new market quote has reached the provider. In this situation the provider has accepted an order even though the market price has moved in an unfavorable way. The arbitrageur[18] is exploring a speed arbitrage opportunity, assuming the market quote has increased, by buying low from the provider and selling high in the market. The same goes if the market price has decreased in t_1. Then the arbitrageur can buy at a low price in the market and sell at a high price to the provider[19]. Figure 5 also illustrates when arbitrage opportunities occur. The difference between the two figures is the fact that the price moves gradually in Figure 5, which is visually easier to comprehend. However, Figure 4 is more accurate in price movements (discrete jumps in price, i.e. not continuous).



Red curve: The provider's bid-offer

Black curve: Market price feed bid-offer

t_1-t_0: internal latency of the provider

Figure 5- The yellow area represents when an arbitrage opportunity occurs, i.e. when the market bid (lower black line) is above the providers offer (upper red line) or vice versa. Note that is a snapshot of a half a couple of hundred milliseconds. The time from t_0-t_1 is between 50-200 milliseconds depending on what latency level you are investigating. Due to the latency of the provider, it is not until t_1 that the provider streams out the market price from t_0.

[18] The Arbitrageur is in fact the client that is trading with the provider. The Arbitrageur is exploiting the fact that the provider misprices due to latency

[19] Here it is assumed that the trader can observe the new market price at the same time as the streamed out price of the bank.

The simulation is merely a time shift applied to the market price. Each time a new market quote arrives it is compared to what the price is after x milliseconds latency (where x is set to 100, 250 and 500 milliseconds). It can be noted that this model is simplified. The length of time that arbitrage opportunities exist has not been taken into consideration. It is assumed that each time there exists an arbitrage opportunity, after time shifting the market quote, the arbitrageur manages to act once (i.e. trade once).

3.2 Data set

3.2.1 Market data

To test the validity of the model, Reuters, Currenex and EBS Live have been considered as data sources. Even though access to all the major venues price feeds been available, the presented work will only treat data from Currenex. This restriction was made because EUR/USD, which is the chosen currency in this study, is not very actively traded on Reuters. EBS Live was excluded because price updates are throttled to one per 100 milliseconds. These two factors, lack of timely price updates and periodic instead of real-time price updates, have an adverse effect on the results. Therefore the obvious choice was the Currenex price feed.

EUR/USD is the most traded currency pair in the world (with a daily turnover of 1,101 billion dollars by April 2010 representing 28 % of the world trade of currencies)[20]. Since it is the most traded currency pair in the world, it is very liquid and the price is updated more frequently than less liquid currency pairs. As implied by the model, the more frequently the currency pair price is updated, the more speed arbitrage opportunities should occur. However, since this thesis only explores one currency pair, EUR/USD, it will of course not explore the difference in the occurrence of arbitrage opportunities between currency pairs. Still, testing the model on the most active currency pair seems reasonable.

Some of the platforms stream out a multilevel order book with different price levels at different quantity levels. The dataset has been limited and only the top of book prices have been used[21]. The reason for using top of book is simple; it enables further simplifying of the model while at the same time still reflecting reality.

The selected time window is relatively small; one month's worth of data has been used. The month selected was March 2012, which consisted of 22 trading days. The time window per day was also narrowed down to represent a trading day (in Sweden between 08.00-17.00). The simulation of the model (explained in Section 3.1) was done on three different latency levels; 100,

[20] See BIS-report; *Global foreign exchange market activity in 2010*

[21] Top of book = the best price level with the tightest spread between buy and sell side

250 and 500 milliseconds. The selected latency levels do not necessarily reflect common latency levels, since it for apparent reasons is undisclosed information by market participants[22]. Instead, they are randomly chosen to visualize that the potential loss (cost) increases with latency. Even though there are probably not many still left with a latency level of 500 milliseconds, it is good to be able to see how much latency dispersion[23] can cost.

3.2.2 Explanatory variables

To be able to answer the research question about how latency affects profit, three main variables representing cost have been identified.

1. The average size of arbitrage opportunities
2. The number of occurred arbitrage opportunities
3. The ratio between number of arbitrage opportunities and the total number of observations for the same time period. The ratio is calculated as

$$\text{Ratio} = \frac{\text{Number of arbitrage opportunities}}{\text{Total number of observations}} \qquad (1)$$

These variables are quantified by simulation of our model explained and their properties are displayed in Table 1-2 in Appendix. The *Average Size* variable represents the average size of the arbitrage opportunities given such opportunity occurred. (Excluded from the data set in this case are those observations that did not result in an arbitrage opportunity).

3.3 Hypothesis testing

Firstly, hypothesis testing is done to see that arbitrage opportunities (the measure of potential loss) exist in the model for the different latency levels. Secondly, the relationship between the three cost variables and latency is investigated. The hypothesis is that higher latency increases the risk of giving away arbitrage opportunities and thus losing money.

[22] The information being undisclosed is due to the fact that provider$_1$ do not want provider$_2$ to know the latency level that they have in respectively systems. I.e. it is something you do not want your competition to know.

[23] Even systems that have been tuned to provide very low latency exhibit some latency variation or dispersion. Many systems show increased latency under high load and some systems have occasional latency outliers or spikes due to the housekeeping activities in the system that interferes with the processing of transactions. It is therefore important to control the latency variation in addition to the average latency. A latency spike may occur at a time when the best possible latency is required to protect against latency arbitrage attempts from clients and competitors.

To test the model and hypothesis, t-statistics shows whether or not the average sizes of arbitrage opportunities are significantly different from zero. The null hypothesis thus become:

$$H_0: \mu_{number,} \mu_{ratio,} \mu_{avg\_size} > 0 \qquad (2)$$

Thereafter, the relationship between the cost variables and different latency levels is investigated by setting up the following null hypothesis:

$$H_0: \text{ Average Size}_{100} > \text{Average\_Size}_{250} > \text{Average\_Size}_{500} \qquad (3)$$

$$H_0: \text{Number}_{100} < \text{Number}_{250} < \text{Number}_{500} \qquad (4)$$

$$H_0: \text{ Ratio}_{100} < \text{Ratio}_{250} < \text{Ratio}_{500} \qquad (5)$$

Hypothesis regarding the relationship of the average size of arbitrage opportunities for different latency levels (Equation 3), tests if the average arbitrage size of arbitrage opportunities decrease as latency increases. Even though it might look intuitively wrong, the results will further explain the implications of this null hypothesis. The last two null hypotheses test if the number of arbitrage opportunities and the ratio of arbitrage opportunities to total number of observations increase as latency increases.

Since the p-value indicates the lowest significance level at which the null hypothesis would be rejected, this value will be looked at primarily.

As shown in Empirical Results, the null hypothesis tested resulted in statistical proven existence of arbitrage opportunities. Given that the hypothesis failed to be rejected further studies was made to investigate which factors potentially affect the occurrence of arbitrage opportunities. This is done using regression analyzes.

3.4 Regression model

OLS regression lines are used in the study of the relationship between existing arbitrage opportunities and the market volatility and its relation to the bid-ask spread. Volatility and bid-ask spreads are two common structure components that help explain the market in a decent way. By using OLS regression, unbiasedness and consistency can be derived and thus more precise coefficients can be identified.

In order to study the maximum correlation that volatility and bid-ask spreads have on arbitrage opportunities, four regressions were done with different dependent variables.

$$Average\ Size = \beta_0 + \beta_1 Volatility + \beta_2 AvgSpread + \varepsilon_i \qquad (6)$$

$$Exist = \beta_0 + \beta_1 Volatility + \beta_2 AvgSpread + \varepsilon_i \qquad (7)$$

$$\text{Ratio} = \beta_0 + \beta_1 Volatility + \beta_2 AvgSpread + \varepsilon_i \quad (8)$$

$$\text{Number} = \beta_0 + \beta_1 Volatility + \beta_2 AvgSpread + \varepsilon_i \quad (9)$$

In the first regression, the average size of the arbitrage opportunities is the dependent variable (see Table 5). The dependent variable for this regression is named *Average Size*. In the second regression the dependent variable acts as a binary variable, that is it takes on the value 1 if there exist an arbitrage opportunity and 0 otherwise. This variable is called *Exist*. In the third regression, *Ratio* is used as dependent variable (as defined in Equation 1) and the fourth and final regression has only the number of arbitrage opportunities as dependent variable, called *Number*. Constant for all four regressions are the independent variables, volatility and spread, described below. Results of these four regressions can be found in Table 5 (Panel A-C).

3.4.1 Volatility

Volatility is one of the most common measure for financial risk and an important tool when evaluating portfolio selection and asset prices. Since prices on the market changes frequently, it is of interest to look closely at the volatility of the price changes. To be able to do this, the realized volatility approach will be used where the calculation is based on the volatility of historical observations. A traditional way to estimate volatility is to calculate the standard deviation of historical observations. Other methods such as ARCH and GARCH models that estimate a fictive variable of volatility as a starting point for estimating the volatility can also be used (Andersen et Al 2000). However since we have high frequency data we will not use the mentioned methods, instead the following 2-step formula that Andersen et Al call the realized volatility has been used:

1. A return is calculated

$$R_t = \frac{P_t - P_{t-1}}{P_t} = = \frac{P_t}{P_{t-1}} - 1 \quad (10)$$

where R_t is the return for the period t, in this case t represent minute intervals, , P_t is the price at *t* and P_{t-1} is the price at *t-1*.

2. The volatility can be calculated using

$$Vol = \sum_{t=1}^{T} |r_t * r_{t+1}| \quad (11)$$

According to Andersen et Al, this method gives a better estimation of volatility since it is based on assumption about the volatility of the market price. The traditional models rely on the assumption that the returns are statistically distributed and thus, the validity of volatility becomes less precise using the latter. Andersen et Al points out some benefit of using their approach. They claim that it is hard to get rid of the measurement error in the calculations using the traditional method; hence the results may not always be as reliable as one think. Therefore it is better, when evaluating high frequency data to use the realized volatility. The discussed issues do not occur then since the frequency of the dataset is so high. Thus the measurement error approaches zero and the volatility measure becomes more reliable and correct.

3.4.2 Average Spread

The FX market is a very liquid market, which is reflected in the tight bid-offer spread, hence the bid-offer spread acts as a measure of liquidity. Liquidity is an important determinant to market behavior and there are different models for measuring liquidity, one being the bid-offer spread (O'Hara 1995). A tighter spread indicates a more liquid market.

The FX market order book consists of different liquidity levels with different spreads. Top of book, the main focus in this thesis, has the tightest spread (thus representing the "best bid-offer") and is also characterized by a lower supply of volume.

Implied by simple logic, results should show that as the spread becomes wider the number of arbitrage opportunities decreases. To explore this, the spread's correlation to arbitrage opportunity size and frequency is tested in OLS regressions. The average spread variable in this thesis is an average of the difference between offer and bid during each hour, see below.

$$\text{Spread} = (\text{Offer} - \text{Bid}) \qquad (12)$$

4. Empirical results

With the main objective to explore whether or not latency has an effect on profitability, the main results become the simulations and the hypothesis testing done on the results. The cross-sectional result has an explanatory function, describing all variables and also our data set. The regressions and its results have a complementary function. It aims to investigate which factors affect the main result.

4.1 Model simulation results

Table 1 and Table 2 show cross-sectional descriptive for the main results from the model simulations for the three different latency levels. The results are divided by weekday as well as

statistics for the week as a whole. Descriptive statistics for the latter are captured by the variable *Total* in the table. What can be observed, referring to Table 2 Panel B is that the number of arbitrage opportunities occurrences increases with latency. This is indicated by the increasing mean and maximum amount of arbitrage opportunities presented in the table. However the average size of the arbitrage opportunities does not seem to increase with the size of the latency, as one would intuitive think. It rather seems to be the opposite trend since the mean of the total week decreases as latency increases (see Table 1 Panel A-C). It is important to note that the results for the average size of the arbitrage opportunities are expressed in per million traded. The assumption is that every time an arbitrage opportunity occurs, the arbitrageur manages to trade 1 million dollar on the market and with the provider simultaneously[24].

Hypothesis tests have been performed for the three variables representing cost of latency[25], to discern if arbitrage opportunities exist and if the average size of the arbitrage opportunities is greater than zero[26]. The results are shown in Table 3 Panel A-C and support the hypothesis that the cost variables are greater than zero since the p-value is .0000. This is true for all the latency levels tested.

To support the hypothesis that the average size of arbitrage opportunities do not increase with latency, hypothesis testing was performed as shown in Table 4 Panel A[27]. The results show that we fail to reject the null hypothesis with a p-value of .0000, meaning the average size of arbitrage opportunities when having a latency of 100 milliseconds is higher than the average size of arbitrage opportunities when having a latency of 250, and thus 500 milliseconds.

Still, as can be seen in Figure 2, potential loss[28] is implied to increase with latency. As it is graphically illustrated in Figure 2, there is a convexity in the relationship between latency and potential loss with each dot in the diagram representing one trading day worth of accumulated potential loss. This implies an increasing marginal cost of latency. The potential loss due to arbitrage, during a trading day, with a latency of 50 milliseconds, is implied by Figure 2 to be between 5 000 and 28 000 USD. With a latency of 1 second, the indicated potential loss has a larger spread ranging from 25 000 to 105 000 USD.

[24] This means that with a arbitrage opportunity size of 1 basis point (.0001) results in a the per 1 million dollar profit of 100 USD for the arbitrageur.
[25] Assuming that our variables representing cost of latency is a good measure of cost of latency.
[26] The test performed is H_0: $\mu_{number}, \mu_{ratio}, \mu_{avg\_size} > 0$
[27] The performed test is: H_0: avg_size100ms > avg_size250ms > avg_size500ms
[28] Potential loss is based on the above assumption about 1 million traded each time an arbitrage opportunity occurs. Realistically top of book supplied volume may be as low as 100.000 USD and can also be larger than 1 million. Thus it is only an assumption that the arbitrageur will be able to trade 1 million, having in mind that sometimes it may be able to trade less and sometimes more.

In Table 2 Panel A-C, the cross-sectional descriptive for the second and third measure of cost of latency; *Number* and *Ratio* is presented. The table also includes the descriptive for total number of observations per hour, that is number of price updates per hour. The variable is called *Total* in the table. The average number of price updates per hour, during the observed time period, was around 2100[29] (Table 2 Panel C) with a low of 926 updates per hour, up to 4952 representing a very active hour. The results show that there is quite a large dispersion in the number of price updates per hour.

In Table 2 Panel A, the variable *Ratio*, represents the number of price updates resulting in arbitrage opportunities compared to the total number of price updates per hour. It is shown in the table that on average only .40 percent of the observations results in arbitrage opportunities (for a latency level of 100 ms). This may seem low but it still results in costs that could be significant. Furthermore, the ratio part of the cross-sectional descriptive implies that this ratio increases with latency.

Table 4, Panel A-C, shows the results from further investigating if the cost increases with latency by doing more hypothesis testing[30]. With low significance levels (p-value of .000) we fail to reject the null hypothesis that both the number of arbitrage opportunities and the ratio variable increases with latency. Once again implying that cost increases with latency.

Concluding, the results show that arbitrage opportunities would exist under the assumptions made in the model, resulting in a potential loss for the provider in question. The results also show that cost is an increasing function of latency.

4.2 Regression Results

The regression results are presented in Table 5. In the first regression where *Average Size* acts as a dependent variable, the regression result points out a predicted average size of arbitrage opportunities of 222 USD if assuming the volatility on the market and the bid-ask spread equals zero, i.e. the interpretation of the constant[31]. However, the interpretation of only the constant is not meaningful since in reality it is normally not the case that market volatility and bid-ask spread equals zero. Looking at the slope coefficients are more interesting; observable is that there is a positive relationship between the average size of the arbitrage opportunities and the volatility on the market and a negative relationship between the average size of the arbitrage opportunities and

[29] Based on the same data set, the number of total observations per hour should not differ, however the simulations creates a small difference. This does not affect the results.

[30] H_0: number100ms < number250ms < number500ms
H_0: ratio100ms < ratio250ms < ratio500ms

[31] The estimated arbitrage profit of 222 USD assumes that the arbitrageur trades a volume of 1 million in the market simultaneously as trading with the provider in the same volume.

the bid-ask spread. However the latter is statistically insignificant meaning we cannot for sure say that there is a negative relationship between the average size of the arbitrage opportunities and the bid-ask market spread. Due to the fact that the regression, with average size as the dependent variable, has a very low R^2 one can conclude that there are other factors, than volatility and bid-ask spread that affects the average size of arbitrage opportunities. According to the regression results, the two independent variables only explain 4.5 percent of the total variation in the average size of arbitrage opportunities. Generally, a low value of R^2, indicate that it is hard to predict the individual effect of the independent variables on the dependent one. Therefore it is hard to conclude for sure that the volatility is a variable that has an impact on the arbitrage size although the regression results shows a strong positive relationship that also is statistically significant. It is also important to have in mind that R^2 has some drawbacks; it increases when more independent variables are added to the regression. Thus, in this particular regression with only two independent variables, it is not surprising that the R^2 is low and since it never decreases when more variables are added it makes it a poor tool when deciding how many variables should be added to the model.

Worth noticing is that the average size of arbitrage opportunities decreases when latency increases, when looking at the constants (β_0). With a latency of 100 milliseconds the average size of arbitrage opportunities is 222 USD compared to only 151 USD when the latency is 250 milliseconds. This conclusion is not intuitive; the logical intuition is that with a higher latency level, the average size of the arbitrage profit increases. However it is important to distinguish between the size of the arbitrage opportunities and the number of arbitrage opportunities. As the results show, the number of arbitrage opportunities increases with higher latency level, but this does not necessarily means that the size of the arbitrage opportunities increases as well.

In the second regression where the dependent variable is a binary variable taking on the value 1 if there exist arbitrage opportunities and 0 otherwise, the interpretation becomes different from regression 1. Still assuming the assumption made in the first regression holds, this regression with a binary dependent variable is called the linear probability model (LPM) meaning the slope coefficients measures the predicted change in the binary variable when the independent variables increase by one unit. As can be seen from the regression results, if volatility marginally changes with one unit[32] the probability of an arbitrage opportunity existing increases with .103 with a latency level of 100 milliseconds. This is not surprising since when the volatility increases, there is more activity on the market and thus the probability of arbitrage opportunities existing increases.

[32] In the case of volatility for EUR/USD a change in one unit translates to change in the seventh decimal, that is .000001

The same reasoning goes for the spread; a change in spread of one basis point[33], that is .0001, results in a decrease of the probability of an existing arbitrage opportunity of 1.225. Intuitively, it makes sense that if the spread becomes wider, the probability of arbitrage opportunities existing decreases. However, the size of the coefficient for spread (-1.225) does not make sense. This is due to the fact that the model used, the LPM do not limit the probability to be between 0 and 1, it can be less than 0 or greater than 1. If only looking at the constant, assuming both volatility and spread equals zero, the probability of an arbitrage opportunity occurring equals more than 200 % when having a latency level of 100 milliseconds. This conclusion is statistically significant and means that when having a low latency the probability of exploring an arbitrage opportunity increases[34]. Again, this result does not make intuitively sense, however even though probability of arbitrage opportunities seem to decreases as latency increases cost has still been proven to increase with latency.

Moving on to the third regression with *Ratio* as dependent variable the only variable worth commenting on, that is statistically significant is the volatility. There exists a clear positive relationship between the variable *Ratio*[35] and the market volatility. If the volatility of the market increases by .1 (volatility mean is 1.2043 and thus it is reasonable to think that the volatility increases by .01 and not 1) the ratio would increase by .0003[36] for a latency level of 100 milliseconds. With a mean for ratio of .004, a change of .0003 is relatively large. The variable spread is only significant when having a latency level of 500 and 100 milliseconds but unfortunately not for a sufficiently high significant level and since the change is also very small it is not of interest to interpret this specific result further.

In the last regression with *Number* as dependent variable the same conclusions as above considering the relationship between the volatility and latency respectively the spread and latency can be drawn. That there is positive relationship to the volatility and a negative relationship to the spread for all investigated latency levels. The change in number of arbitrage opportunities when volatility increases with .1 is in the fourth regression implied to be 1.276[37] for 100 milliseconds (corresponding number for 500 milliseconds is 2.601). It is also shown that volatility's effect on number of arbitrage opportunities occurring increases with latency. In comparison to the other three regressions, this regression is the only one that has a R^2 higher than 50 percent (looking at latency levels of 250 and 500 milliseconds an even higher R^2 is obtained) meaning in this context

[33] In EUR/USD this means that 1 basis point is a 100th of a cent

[35] Which is, in this regression, the measure of potential loss due to latency

[36] .1*.003=.1* $\beta_{vol.}$ With a mean for volatility of around 1.0 a change of 0.1 is feasible.

[37] .1*12.759

that the chosen explanatory variables are suitable measures for explaining the number of arbitrage occurrences. Implied by the regression result for regression four is that when volatility and spread are zero, then the average number of arbitrage opportunities during an hour equal to 37.8 (for a latency level of 100 milliseconds). However, the constants are not statistically significant, thus any real conclusions cannot be drawn.

Since our sample is very small, the assumption that there is no perfect collinearity among the independent variables may not hold. Also since the volatility can be expressed as an exact linear function of bid-ask spread, the assumption about perfect collinearity will be violated and thus the assumption of zero conditional mean, that is the error term has an expected value of zero given any values of the independent variables is also violated. This is probably the reason to why the bid-ask spread is statistically insignificant.

4.3 Cross-sectional descriptive of the independent variables

Looking at the cross sectional descriptive for the volatility, generally there is no observable difference between the weekdays. The mean for an average weekday of each hour, referring to Table 6 is approximately the same and around 1.0[38]. On Friday and Monday on the other hand the volatility is more dispersed, as the value of the kurtosis is high for those days. A high value of kurtosis implies the tails in the distribution are fat resulting in more dispersion. A possible interpretation of the high kurtosis value for Friday, i.e. an explanation to why the volatility is more spread on Friday can be due to the Non-Farm payrolls[39] statistics being released in the US first Friday of the month, resulting in more activity on the market and thus increasing market volatility. Regarding Monday, it is hard to conclude anything but it could potentially be due to events that may have happened during the weekend, resulting in change expectations of the future thus resulting in active trading.

Going on analyzing the spread which results can be found in Table 6 the estimated cross sectional descriptive are roughly the same for all weekdays. The spread seems to not have an impact on what weekday there is when looking at the means and variance. The only parameter that has a little difference in weekdays is skewness. [40]An average Thursday in March has a positive skewness of .3094 compared to the negative ones in the other weekdays. The reason to why this is the case is very hard to say because there is no observable trend that are shown in the results.

[38] Note that the volatility has been multiplied by 1 000 000 to be easier to observe and analyze, meaning the volatility is as small as .0000001.
[39] The nonfarm payrolls is seen as a good indicator for the US economy, because the number presented tells us how many jobs have been created or lost in the economy for the last month (excluding jobs from the farming industry)
[40] Skewness is a measure of the symmetry of a distribution. When the skewness is zero, the distribution is followed by a standard normal distribution.

Although the other weekdays have a negative skewness, the difference of the skewness is quite large when comparing the different weekdays. Monday for example has a negative skewness of -.8816 while Wednesday has a negative skewness of -.0542. The spread of the skewness is quite large within weekdays meaning the dispersion to the mean is relatively day specific. The value of kurtosis is roughly the same for all weekdays. [41]

As can be seen from the graph of the distribution of the spread in Figure 5, the distribution of the spread in basis points has the same shape as a normalized curve. The mean is roughly 1.25 basis points meaning the spread is on average 1.25. As the kurtosis for a standard normal distribution is three, one can conclude that the spread is near a normally distribution, which is also supported by the histogram in the Appendix (Figure 5).

The cross-sectional descriptive of spread for EUR/USD for march 2012 in Table 6 supports the notion that the FX market is characterized by a tight spread (i.e. high liquidity).

5. Implications and conclusions

The results show that the model did simulate the occurrence of arbitrage opportunities and that the variables representing cost increases with latency. It is argued in this thesis that the model applied could reflect real potential situations, meaning also the results can reflect real state of affairs. The fact that the cost variables are statistically significantly larger than zero supports the notion that arbitrage opportunities exist in our model world.

The regression results primarily showed that there is a positive relationship between volatility and the occurrence of arbitrage opportunities, implying that the level of volatility in the market matters. Not only was there a positive relationship between volatility and arbitrage opportunities, this relationship increases with latency. Due to statistically insignificant results for the bid-ask spread coefficients in the regressions, no conclusions could be drawn regarding the relationship between the occurrence of arbitrage opportunities and the size of the bid-ask spread. Furthermore, the low R^2-value of the regressions means that factors other than volatility and spread affect the occurrence of arbitrage opportunities.

As shown, there is a convexity in the relationship between potential loss per day and latency, which implies an increasing marginal cost of latency. This would mean that the benefits of reducing latency have a decreasing marginal return. Moving from 500 milliseconds to 250

[41] Kurtosis is a measure of peakness. It shows the heaviness of the tails relative to a normal distribution. A high kurtosis indicates that the distribution is peaked near the mean resulting in heavy tails and vice versa. For a standard normal distribution, the kurtosis is three.

milliseconds saves more than moving from 250 milliseconds to only microseconds. It can always be questioned if the small sample in this thesis reflects the population. However the small sample functions as an indicator that further research should be done to verify this. If the increasing marginal cost of latency (or decreasing marginal return of reducing latency) can be proven, this will have implications on how trading entities should make IT investments decisions. This could also serve as a reason for choosing an algorithmic trading strategy.

In conclusion, this study has found that latency is a financial risk factor, as it affects the occurrence of arbitrage opportunities. This conclusion coincides with the fact that some trading entities spend substantial resources to get a few tenths of a millisecond lower latency.

6. Further research

It is apparent from the results of this thesis that latency can affect the profitability of a bank/provider and that there is much reason to future investigate this, fairly new, risk factor in modern trading. This thesis was limited to testing a simple model to see if latency can be a financial risk factor in trading. Latency is thus far a relatively unexplored subject in academia and there is still much research to be done. We have demonstrated that speed arbitrage opportunities can occur when banks/providers have a certain latency level. However, the sample was relatively small and even though it gave statistically significant result; there are benefits of further expanding the time window.

The study shows that latency can be seen as a financial risk factor, adversely affecting the profitability. It is important to note that this study was made on only one currency pair and even though it is the most traded currency pair in the world, adding more currency pairs in the model would probably increase the potential loss due to arbitrage. An increase in the potential loss, due to taking more currency pairs into consideration, would therefore also increase the severity of latency as a risk factor. Thus, expanding the model in this thesis would increase latency impact on profitability.

Furthermore, it is of interest to investigate the occurrence of triangle arbitrage[42], where mispricing is taken advantage of by trading three currencies simultaneously. The potential loss would possibly increase when taking this type of speed arbitrage into account.

It would also be valuable to explore additional latency levels as well as exploring the difference between different currency pairs. The low R^2 also implies that other factors than spread and

[42] An example of triangle arbitrage: Suppose you trade EUR/CHF and then trade EUR/USD and USD/CHF. Even though you are trading between the same currency pair, EUR/CHF, there can be mispricing involving USD that a speed arbitrageur can take advantage of. See Appendix, Figure 4 for illustration.

volatility affect the occurrence of speed arbitrage opportunities, which can be added to potential sources for further research.

7. References

O'Hara, M. (1995). *Market Microstructure Theory.* Blackwell Publishers: Cambridge, Massachusetts.

Sarno Lucio, Taylor Mark P. (2001) *The Microstructure of the Foreign-Exchange Market.* Publishers: Princeton University New Jersey

Lyons, Richard K. (2001) *The Microstructure Approach to Exchange Rates.* Publishers: The MIT Press, Cambridge, Massachusetts.

Frankel, Jeffrey A. Galli Giampaolo, Gioannini Alberto. (1996) *The Microstructure of Foreign Exchange Markets*. Publisher: University of Chicago Press

Hasbrouck Joel, Saar Gideon, *Low Latency Trading*. (2010). Publisher: Stern School of Business and Cornell University

Ciamac C. Moallemi, Mehmet Sa˘glam, *The Cost of Latency*. (2009), Colombia University

Riordan Joseph Ryan. *The Economics of Algorithmic Trading* (2009), Karlsruhe Institute of Technology.

Aldridge, I.,(2008) *High Frequency Trading–A Practical Guide to Algorithmic Strategies and Trading System.* Publisher: John Wiley & Sons, Inc., New Jersey

Barndorff-Nielsen, Ole E., Shephard Neil. *Power and Bipower Variation with Stochastic Volatility and Jumps* (2003)

Arnuk Sal, Saluzzi Joseph. *Latency Arbitrage: The Real Power Behind Predatory High Frequency Trading* (2009)

Andersen, T, Bollerslev, T, Diebold, F, Ebens, H. *The Distribution of Stock Return Volatility* (2000)

Bank for International Settlements. *High-frequency trading in the foreign exchange market*. (September 2011)

Bank for International Settlements, Triennial Central Bank Survey, *Report on the global foreign exchange market activity in 2010*

Bank of International Settlements, BIS Quarterly Review December 2010, *The $4 trillion question: what explains FX growth since the 2007 survey?*

Wooldrige, J.M., (2009). *Introductory Econometrics: A Modern Approach (4th edition).* Publisher: South Western

Newbold, P., Carlson L.W., & Thorne B (2010). *Statistics for Business and Economic (7th edition).* Publisher: Pearson Education

8. Appendix

Table 1. Cross-Sectional Descriptive for the Average Size of Arbitrage Opportunities

The variable *Average Size* measures the average size of the arbitrage opportunities explored and is expressed in per million traded. The variable properties are divided into week days in order to see potential differences between different days of the week. The descriptive also shows the week as a whole. The average size of arbitrage opportunities and its properties differ between the latency levels and are therefore divided into 3 panels. The natural logarithm of the dependent variables was not used although the dependent variables reflect prices (or more correct profits). This is due to the fact that the variables seem to already look like a standard normalized curve, without having taken the logarithm of the variables (see Figure 3).

Panel A: Latency of 100 ms

Weekday	N	Min	Max	Mean	Variance	Std. Dev.	Skewness	Kurtosis
Monday	596	10	400	148.305	4 308	65.639	1.358	5.712
Tuesday	347	10	400	116.397	2 399	48.984	2.756	13.500
Wednesday	254	10	900	115.748	6 815	82.553	7.149	65.750
Thursday	538	10	900	122.955	9 961	99.808	6.068	43.810
Friday	621	30	800	115.024	3 426	58.540	7.015	69.786
Total	1788	10	900	120.727	6 426	80.167	6.018	50.323

Panel B: Latency of 250 ms

Weekday	N	Min	Max	Mean	Variance	Std. Dev.	Skewness	Kurtosis
Monday	482	10	500	125.021	4 869	69.785	3.0424	12.651
Tuesday	423	10	400	115.887	2 667	51.647	3.144	15.622
Wednesday	382	10	900	109.215	3 179	56.391	8.081	104.833
Thursday	774	10	900	116.111	4 718	68.691	7.083	71.895
Friday	895	10	900	117.575	3 455	58.779	5.980	62.744
Total	2965	10	900	117.032	3 871	62.218	5.756	56.151

Panel C: Latency of 500 ms

Weekday	N	Min	Max	Mean	Variance	Std. Dev.	Skewness	Kurtosis
Monday	699	10	600	121.059	4 689	68.479	3.311	15.506
Tuesday	632	10	500	120.221	3 947	62.830	2.806	12.079
Wednesday	560	10	900	108.054	2 399	48.985	8.585	126.465
Thursday	947	10	400	115.079	2 468	49.682	2.888	13.822
Friday	1148	10	900	117.491	3 653	60.445	5.148	48.383
Total	4196	10	900	116.513	3 539	59.491	4.594	38.475

Table 2. Cross Sectional Descriptive for Variables Ratio, Number and Total

The variable *Ratio* is calculated by dividing the number of observations that resulted in arbitrage opportunities with the total number of observations including those observations that did not result in an arbitrage profit. *Number* is measuring the number of observations resulting in an arbitrage profit (the numerator in the Ratio variable). *Total* is a variable measuring the number of price updates per hour (the total number of observations in an hour). All measures are on per hour basis. Note that the total number of observations in each table does not vary between latency levels, because the total number of observations reflects the number of price updates per hour, independent of latency. The only difference observable for the variable *Total* is a small difference in the mean comparing different latency levels. This is due to simulations in the model. However this does not affect any conclusions. Remember that the variable *Ratio* is a ratio measured in decimal form and *Number* counts the number of arbitrage opportunities existing.

Panel A: Descriptive for the variable *Ratio* for 100, 250 and 500 milliseconds

	100 ms	250 ms	500 ms
N	198	198	198
Min	0	0	0
Mean	0.004	0.006	0.009
Max	0.022	0.026	0.038
Variance[43]	0.0000	0.0000	0.0000
Std. Dev.	0.0041	0.0051	0.006
Skewness	1.327	1.280	1.328
Kurtosis	4.937	4.937	5.705

[43] Note that the variance is very small, it may be a variance on the 7th decimal but since it is very small it approximates to zero

Panel B: Descriptive for the variable *Number* for 100, 250 and 500 milliseconds

	100 ms	250 ms	500 ms
N	198	198	198
Min	0	0	0
Mean	10.924	14.661	22.252
Max	106	128	189
Variance	188.2	319.7	575.0
Std. Dev.	13.718	17.881	23.979
Skewness	3.157	3.222	3.422
Kurtosis	18.228	18.223	20.676

Panel C: Descriptive for the variable *Total* for 100, 250 and 500 milliseconds

	100 ms	250 ms	500 ms
N	198	198	198
Min	926	926	926
Mean	2164.7	2118.1	2135.7
Max	4952	4952	4952
Variance	798550	508018	497395
Std. Dev.	893.6	712.8	705.3
Skewness	3.839	1.128	1.064
Kurtosis	31.362	4.629	4.579

Table 3. Hypothesis Testing That There Exist Arbitrage Opportunities

The t-tests are testing whether or not there exist arbitrage opportunities for three variables. *Average Size* which measures the average size of the arbitrage profit if the arbitrageur is able to explore all the available arbitrage opportunities. It is expressed in per million traded. *Number* is measuring the number of observations resulting in an arbitrage profit (for the arbitrageur, loss for the provider) and *Ratio* is calculated by dividing the number of observations that resulted in arbitrage opportunities with the total number of observations including those observations that did not result in an arbitrage profit. Remember that the variable *Average Size* is measured in USD, while the *Ratio* is a ratio measured in decimal form and *Number* counts the number of arbitrage opportunities existing.

Panel A: H_0: $\mu_{\text{Average Size}} > 0$

	100 ms	250 ms	500 ms
N	198	198	198
Degrees of Freedom	197	197	197
Mean	97.1974	99.9611	106.3463
Std. Dev.	0.0178	0.0133	0.0082
t-stat	24.1978	33.3076	57.3663
p-value	0.0000	0.0000	0.0000

Panel B: H_0: $\mu_{\text{Number}} > 0$

	100 ms	250 ms	500 ms
N	198	198	198
Degrees of Freedom	197	197	197
Mean	10.9242	14.6616	22.2525
Std. Dev.	13.7176	0.0056	0.0076
t-stat	11.2059	11.5375	13.0578
p-value	0.0000	0.0000	0.0000

Panel C: H_0: $\mu_{Ratio} > 0$

	100 ms	250 ms	500 ms
N	198	198	198
Degrees of Freedom	197	197	197
Mean	0.0043	0.0058	0.0091
Std. Dev.	0.0041	0.0051	0.0065
t-stat	15.0260	16.3081	19.7776
p-value	0.0000	0.0000	0.0000

Table 4. Hypothesis Testing Showing That Results from Different Latency Levels For Variables Are Different From One Another

The t-tests are testing whether or not the variables *Average Size, Number and Ratio* for a specific latency level is greater than one another. In Panel A, the null hypothesis suggests that the average size of arbitrage opportunities decreases as latency increases. However, in panel B and C the null hypothesis is that the number of arbitrage opportunities and the ratio of arbitrage opportunities compared to total number of observations increases as latency increases. Remember that the variable *Average Size* is measured in USD, while the *Ratio* is a ratio measured in decimal form and *Number* counts the number of arbitrage opportunities existing

Panel A: H_0: Average Size$_{100}$ > Average_Size$_{250}$ > Average_Size$_{500}$

	Average Size$_{100}$< Average Size$_{250}$	Average Size$_{100}$< Average Size$_{500}$	Average Size$_{250}$< Average Size$_{500}$
N	1788	1788	1788
Degrees of Freedom	1787	1787	1787
t-stat	12.408	13.498	17.708
p-value	0.000	0.000	0.000

Panel B: Ho: Number$_{100}$ < Number$_{250}$ < Number$_{500}$

	Number$_{100}$<Number$_{250}$	Number$_{100}$<Number$_{500}$	Number$_{250}$<Number$_{500}$
N	198	198	198
Degrees of Freedom	197	197	197
t-stat	-7.810	-12.681	-13.284
p-value	0.000	0.000	0.000

Panel C: H_0: Ratio$_{100}$ < Ratio$_{250}$ < Ratio$_{500}$

	Ratio$_{100}$<Ratio$_{250}$	Ratio$_{100}$<Ratio$_{500}$	Ratio$_{250}$<Ratio$_{500}$
N	198	198	198
Degrees of Freedom	197	197	197
t-stat	-9.187	-17.378	-16.066
p-value	0.000	0.000	0.000

Table 5. Main regressions

The dependent variable (a) is a measure of the average size of the arbitrage opportunities an arbitrageur can obtain if he is able to explore all the available arbitrage opportunities. This variable is given in absolute numbers. The variable (b) is a binary variable taking on the value 1 if there exists an arbitrage opportunity and 0 otherwise. The variable (c) is a calculated ratio where the numerator is the number of observation resulted in an arbitrage opportunity and the denominator is the total number of observations. The variable (d) represents the number of arbitrage opportunities occurring (the numerator in the (c) variable). Volatility and spread are the independent variables that aim to explain the dependent variable in each regression. Note that the volatility has been multiplied by 1 000 000 to easier observe the dispersion or the spread that the volatility measures, (the volatility is so small that it is often only one the 7th decimal .0000001) and the spread by 10 000 to get it in basis points. Everything is on a per hour basis.

Panel A. Latency level of 100 ms.

Coefficients	Average Size (a)	Exist (b)	Ratio (c)	Number (d)
Volatility	12.669***	0.103**	0.003***	12.759***
Average Spread	-111.564	-1.225*	-0.011*	-33.518**
Constant	222.236**	2.284***	0.015**	37.808*
R2	0.045	0.106	0.313	0.503
N	198	197	198	198

**p<0.05, **p<0.01, ***p<0.001*

Panel B. Latency level of 250 ms.

Coefficients	Avgerage Size (a)	Exist (b)	Ratio (c)	Number (d)
Volatility	14.047***	0.066*	0.004***	19.292***
Average Spread	-54.211	-0.478	-0.007	-16.439
Constant	151.315*	1.445*	0.009	12.311
R2	0.070	0.051	0.393	0.616
N	198	198	198	198

**p<0.05, **p<0.01, ***p<0.001*

Panel C. Latency level of 500 ms.

Coefficients	Average Size (a)	Exist (b)	Ratio (c)	Number (d)
Volatility	8.805***	0.010	0.005***	26.014***
Average Spread	-19.766	-0.099	-0.019**	-46.075*
Constant	120.680***	1.103***	0.026**	49.117
R2	0.065	0.010	0.434	0.654
N	198	198	198	198

**p<0.05, **p<0.01, ***p<0.001*

Table 6. Cross-Sectional Descriptive over Volatility for March 2012

The volatility is expressed in average volatility per hour. The descriptive is divided into which day of the week as well as the week as a whole. By separating the days, differences in volatility depending on the day are illustrated. Each day in the data set consists of 9 trading hours. Note that for March there were five Thursdays and Fridays and only four Mondays, Tuesdays and Wednesdays. Also note that the volatility has been multiplied by 1 000 000 to easier observe the dispersion or the spread that the volatility measures, (the volatility is so small that it is often only one the 7th decimal .0000001).

Volatility	N	Min	Mean	Max	Variance	Std. Dev.	Skewness	Kurtosis
Monday	36	0.3055	1.1867	6.2010	1.0356	1.0176	3.5782	17.6716
Tuesday	36	0.3649	0.9976	2.2403	0.1920	0.4382	0.7640	3.1346
Wednesday	36	0.3665	1.1155	4.2764	0.4283	0.6545	3.1704	16.206
Thursday	45	0.3822	1.3318	2.9732	0.2931	0.5414	1.1340	4.8261
Friday	45	0.4324	1.2746	5.0769	0.6064	0.77872	2.7903	13.8661
Total	189	0.3055	1.2043	6.2010	0.5182	0.7199	3.2659	19.6724

Table 7. Cross-Sectional Descriptive over Spread in Basis Points per hour for March 2012

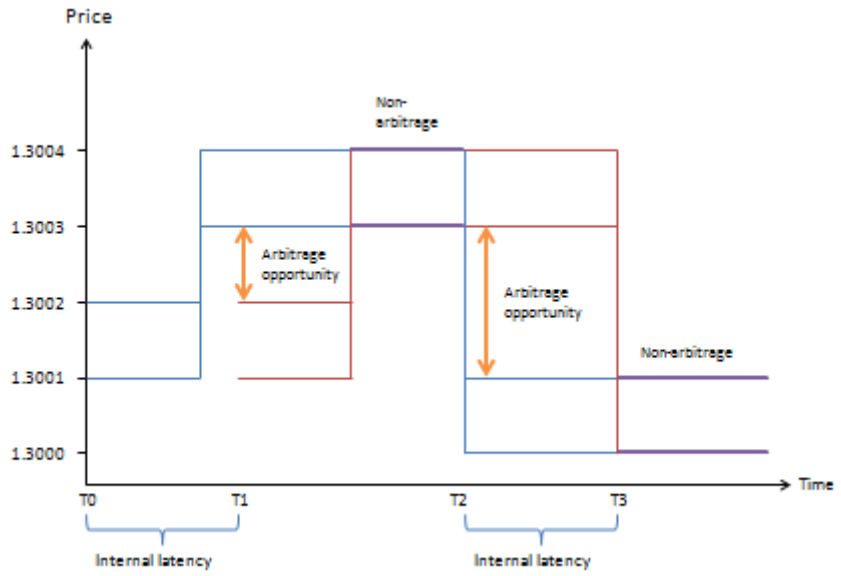
The cross-sectional descriptive below shows the properties of average spread per hour expressed in basis points[44]. Exactly like the cross-sectional descriptive on volatility the table consists of each day separate from each other as well as the week as a whole. Each day in the data set consists of 9 trading hours. Note that for March there were five Thursdays and Fridays and only four Mondays, Tuesdays and Wednesdays. Note the spread has been multiplied by 10 000 to get it in basis points.

Spread	N	Min	Mean	Max	Variance	Std. Dev.	Skewness	Kurtosis
Monday	36	1.1443	1.2549	1.3102	0.0018	0.0430	-0.8816	3.0096
Tuesday	36	1.1207	1.2528	1.3886	0.0046	0.0683	-0.1288	2.2380
Wednesday	36	1.1360	1.2448	1.3567	0.0026	0.0512	-0.0542	2.8392
Thursday	45	1.1712	1.2689	1.3806	0.0024	0.0490	0.3094	2.5665
Friday	45	1.1155	1.2565	1.3589	0.0025	0.0506	-0.5079	3.1912
Total	189	1.1155	1.2563	1.3886	0.0028	0.0533	-0.2234	2.9489

[44] For EUR/USD a basis point is a 100th of a cent.

Graphs and illustrations

Figure 1. Numerical example of Figure 4 in Section 3.1



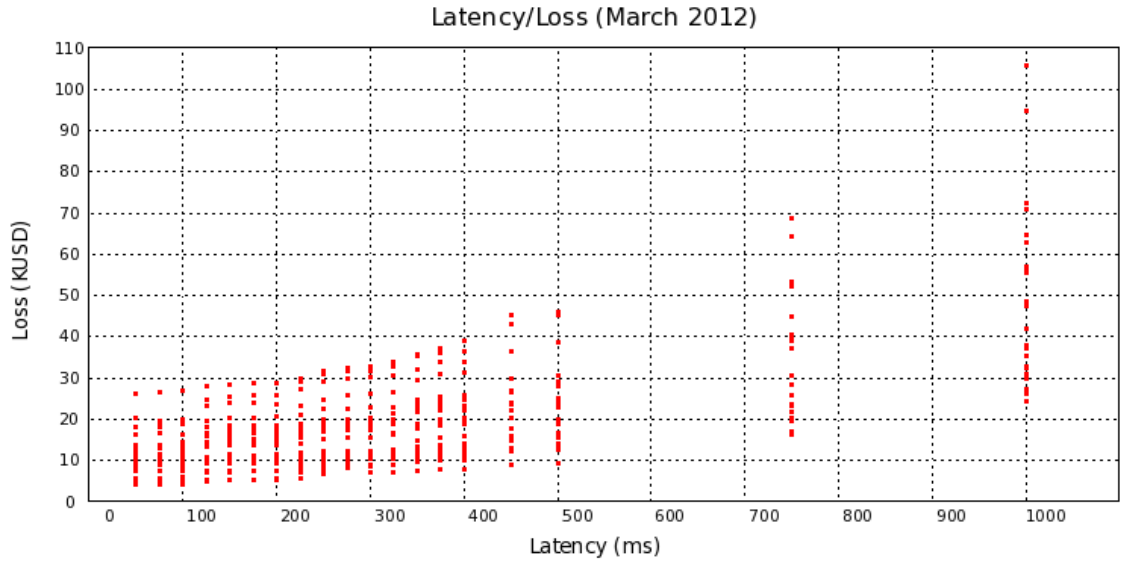
Currency prices move in discrete steps. As a result, a snapshot of a few hundred milliseconds would like the figure above. At t_0 the market streams an offer for the client to buy at 1.3002 and a bid to sell at 1.3001. This price is streamed out by the provider in t_1. However by that time the market quote has moved to an offer of 1.3004 and a bid of 1.3003. Thus an arbitrage opportunity of 0.0001 (1.3003-1.3002) occurs due to the fact that the arbitrageur can buy at 1.3002 from the provider and sell at the higher price of 1.3003 in the market.

From the millisecond when the market moves to 1.3003/1.3004 it takes the provider the internal latency of the provider to start streaming out this price, during this time they expose themselves to arbitrageurs. When market and provider stream out the same price no arbitrage opportunities exist, illustrated by the purple line in the figure.

The second arbitrage opportunity in the figure illustrates when the market depreciates. In t_2 the market offer is lower than the providers bid. Thus the arbitrageur can buy low from the market and sell high to the provider. In t_3 the provider has adjusted its price to reflect the price from t_2. Thus eliminating any arbitrage opportunity.

Figure 2. Graph over the relationship between latency level and cost of latency

Plotted in the graph below where potential loss is in thousands of dollars on the y-axis and different latency levels on the x-axis, one can notice that the potential loss is implied to increase with latency. Potential loss is based on the above assumption about 1 million traded each time an arbitrage opportunity occurs. The figure shows a convexity in relationship between latency and potential loss with each dot in the diagram representing one trading day worth of accumulated potential loss. The potential loss due to arbitrage during a trading day, with a latency of 50 milliseconds, is implied by the figure to be between 5 000 and 28 000 USD. With a latency of 1 second, the potential loss seem to have much larger spread ranging from 25 000 to 105 000 USD[45].

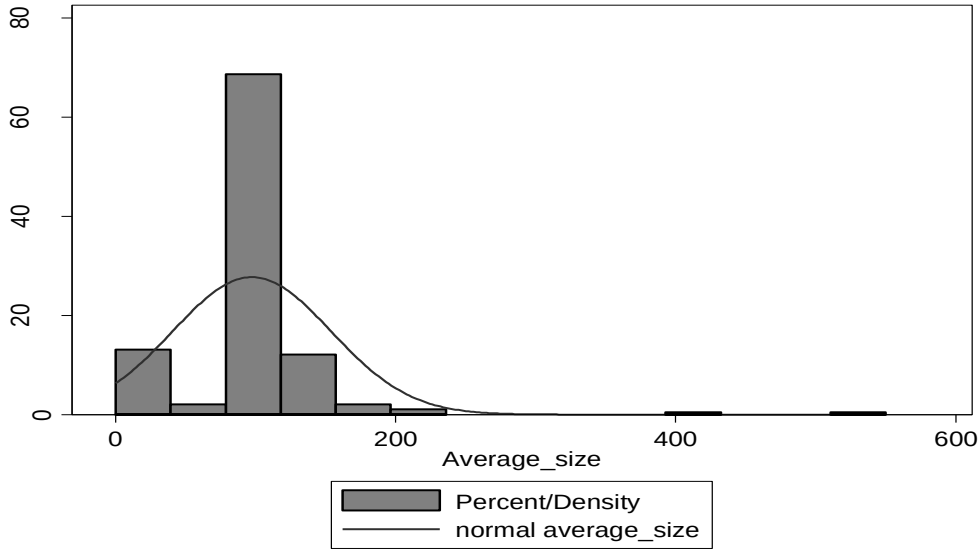


[45] Again, based on the assumption that the arbitrageur manages to trade 1 million with the provider and on the market simultaneously every time an arbitrage opportunity occurs. However, it should be noted that the supplied volume in top of book for Currenex can be as low as 100.000 and can be higher than 1 million. Therefore it is an assumption of a trading volume of 1 million, having in mind that the traded volume may sometime be smaller and sometimes higher.

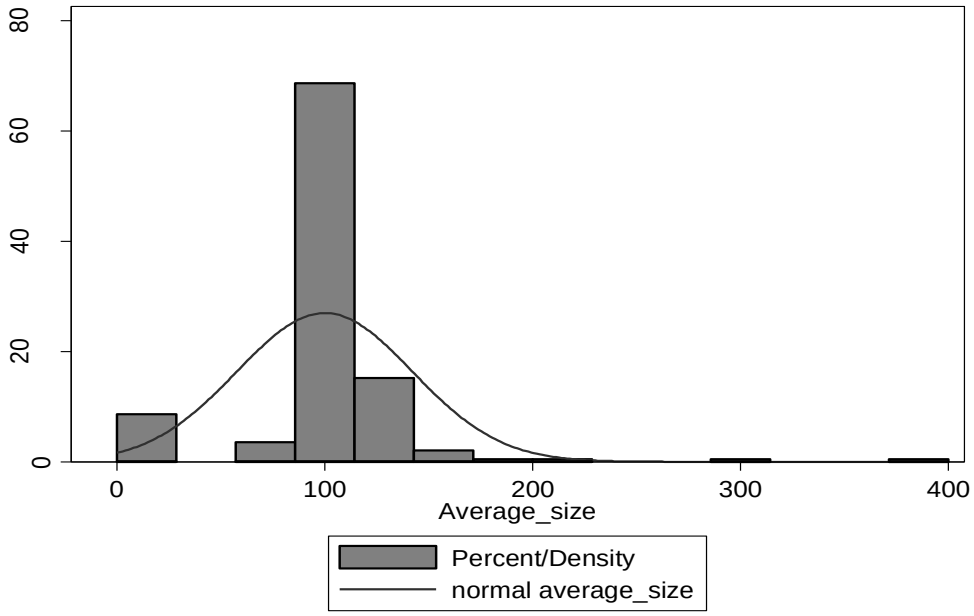
Figure 3. Histogram of the Average Size of Arbitrage Opportunities per Average Hour in March

The graphs show the distribution per hour of the average size of the arbitrage opportunities given by the variable *Average Size*. The variable is measured in absolute terms, that is in USD. As can be seen from the three following graph, latency level do not have a great impact on the distribution. In near 70 percent of the cases, the potential arbitrage profit (loss for the provider) lays around 100 USD.

Panel A: Latency of 100 ms



Panel B: Latency of 250 ms



Panel C: Latency of 500 ms

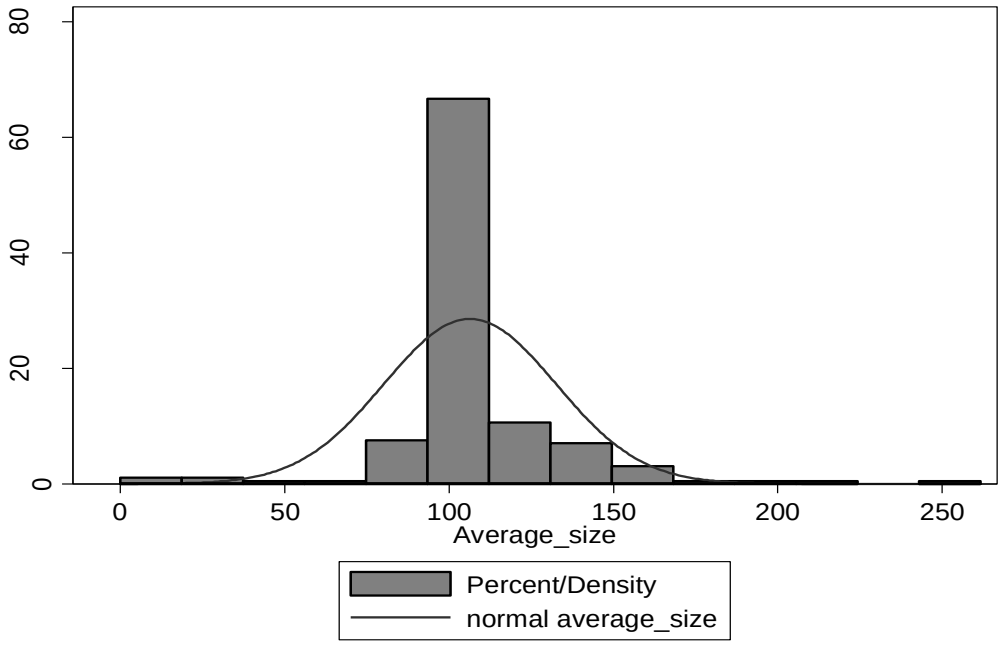


Figure 4. Distribution of Spread Given in Basis Points

The graph shows the distribution of the spread in basis points. For EUR/USD a basis point is a 100th of a cent.

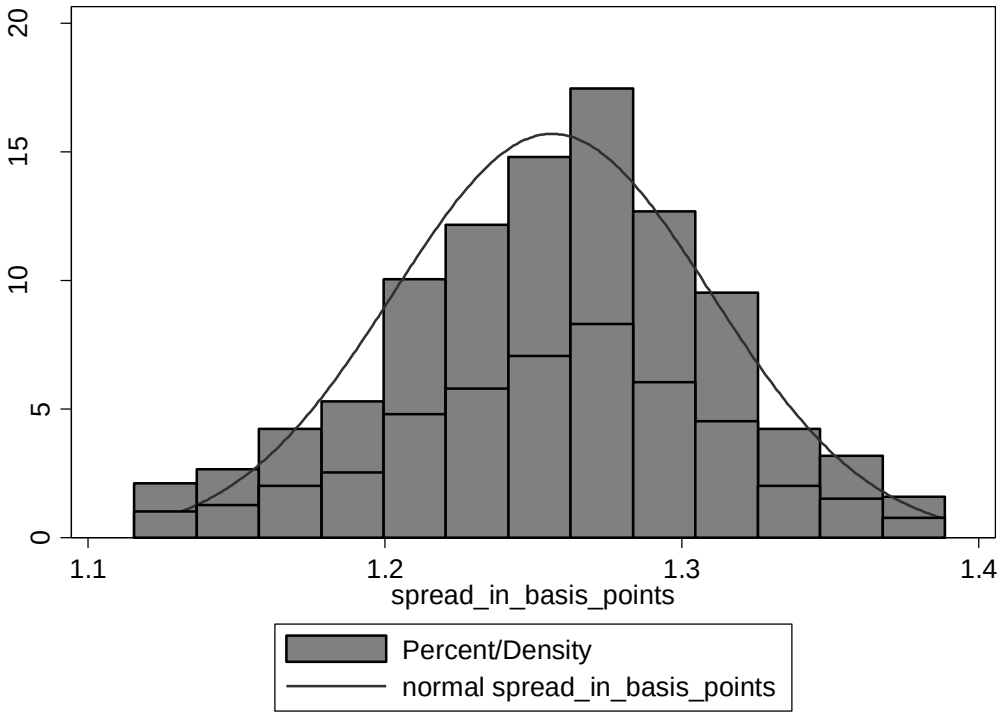


Figure 5.

Illustration of triangle arbitrage

In the example below the triangle arbitrage consists of three currencies; CHF, EUR and USD. By trading all three currency pairs simultaneously mispricing across the three currencies are taken advantage of.