Analysis of TAQ trade data

Dayu Yang(independent work)

April 2021

Abstract

Following AG paper, I reproduce the analysis process in this paper using data of the TAQ trading data of Intel in Jan 1997. Later, I made a comparison across both the time and industry. Result shows the stock of Intel indicates the company is becoming more mature, predictable and stable, Also, results show that under the wide use of quantitative trading technology, people may split one large bids into small bids to achieve their desired positions. Finally, the market performance of TSMC, a major competitor of Intel can hardly compete with INTEL in Jan 2013.

1 INTC in Jan 1997

1.1 Intraday patterns

Here is how I compute (log)return:

$$R_i = ln(\frac{price}{price.shift(1)}) = ln(price_{last}) - ln(price_{first})$$

under such return computation, for each 15 interval:

$$R_{15min} = \sum_i ln(R_i)$$

further, if we want to obtain return from log return, we could use:

$$\sum_t log(\frac{P_t}{P_{t-1}}) = log(\prod_t \frac{P_t}{P_{t-1}}) = log(\frac{P_{15}}{P_1})$$

The mean share volumn could be obtained by

$$\frac{sum(SIZE)}{num_{trade}} = mean(SIZE)$$

Similarly, we could get mean squared return by using:

$$\frac{\sum_n (return - mean(return))^2}{n}$$

There are two types of variance estimation, unbiased or biased version. I use biased one which is:

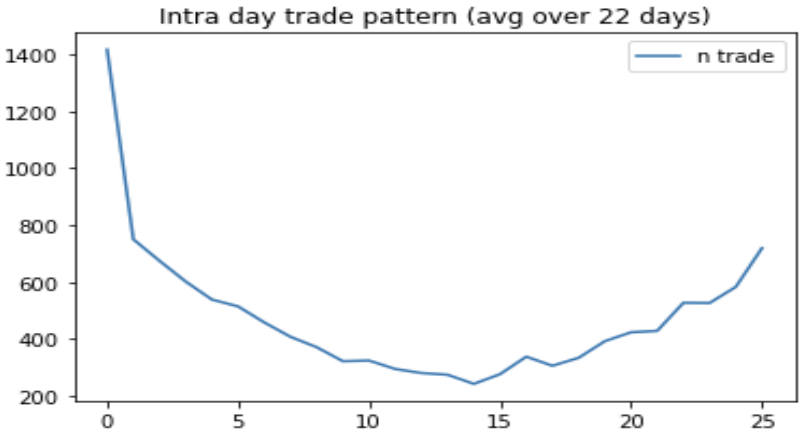
$$var\ return = var([log(\frac{P_2}{P_1}), log(\frac{P_3}{P_2}), log(\frac{P_4}{P_3})...log(\frac{P_t}{P_{t-1}})])$$

$$t \in one\ 15min\ interval$$

1.2 trade

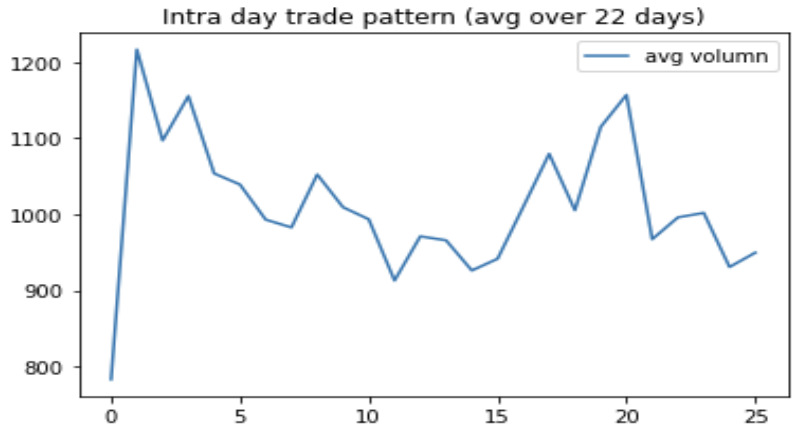
First interesting fact we could clearly notice is that the highest number of trades is at the beginning of a trading day. Another high peak is at the end of a trading day. That's is, the graph of the trade number against time in a day forums a U-shaped pattern, with occasional short-term volume spikes near noon(x-axis = 16). This pattern may be explained by several reasons, but almost all reasons are associated with "information accumulation" overnight. First, at the beginning of the day, imbalances from overnight trading in foreign markets between the price overseas and in the US may provide a brief opportunity for spatial arbitrage. Second, public companies and the government maintain operation after the close of the stock market. They may settle important decisions during that period and significantly affect the analysts' expectation of the company. For example, UP Fintech Holding(TIGR) is an leading online brokerage firm focusing on global Chinese investors, after Biden administration announced they will maintain part of the trading restriction with China at night, which bring no relief to tensions between China and the US. The stock price decreased 10% right after the market is open the next day. Another major cause of the U-shape curve is that day traders usually prefer to have no overnight exposure, which follows one of the fundamentals of technical analysis suggestions: holding a security overnight represents a huge commitment. Thus they open positions after the open and close positions after open.

Figure 1: intraday pattern of the number of trades



Source: WRDS TAQ dataset

Figure 2: intraday pattern of volumn



Source: WRDS TAQ dataset

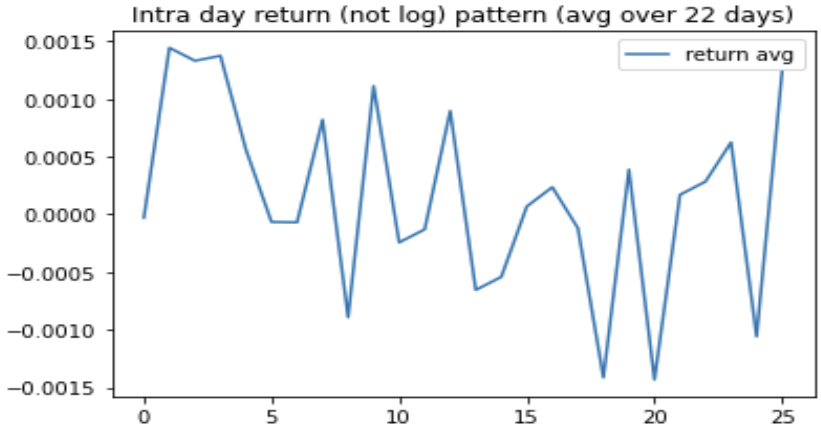
1.2.1 volume

For trading volume, it shows like a M-curve instead of U.(M could be roughly described as a "narrow" U). One of the reasons is that major players on the market are more cautious than retail investors. It is no doubt that major institutions also encountered "information accumulation", but they have to make sure the bids could be successfully executed.(Since their trading volumes are huge, their bids may be partially execute for a relatively long time). That's why they tend to leave some time as a cushion.

1.2.2 return

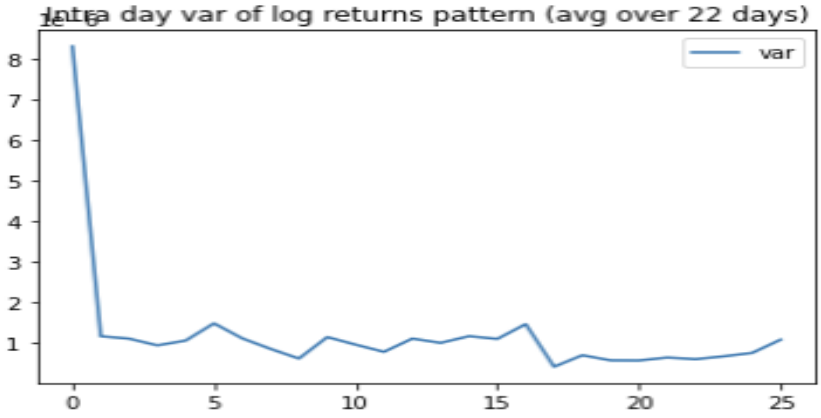
First pattern we observe from returns is actually "no pattern", which means they are very random and continuously fluctuate around 0 during all 26 15min-interval for a day. First, this has been well-illustrated by a lot of financial theories. Many economists describe the return of the stock market as a brownian motion in a short period of time, since there are a tremendous amount of bidders in the market. In other words, there always are a huge number of long traders and short traders. Long traders buy stock and push the price goes higher(return becomes positive) and higher. Once the price reaches the threshold of the short traders, they will become more dominant in the market and push the price to be lower(and the return becomes negative). Second interesting pattern is that the returns at the start and the end of the day are averagely positive. This gives us more information about trading columns – long trades are more than short trades at the begin and end of the day.

Figure 3: intraday pattern of return



Source: WRDS TAQ dataset

Figure 4: intraday pattern of variance of returns



Source: WRDS TAQ dataset

1.2.3 variance

Figure 4 shows the variance of return, as expected, the variance is at its peak right after the market is open, since there are far more bids and trade volumes at that time. However, more interestingly, we should keep in mind that INTEL did a certain thing overnight, which means we could assume that the orthant of the influence triggered by this thing is settled (negative or positive) but unknown, like a parameter of a statistical distribution. And somehow, everyone received different information from their information sources. For example, some media may describe this as an event with positive impact others may be the opponents. If we assume every trader grasp the same piece of information, for example, intel announced their quarterly financial report yesterday evening. That leads a conclusion that: traders have a huge divergence of opinions about it, and further causing a huge variance of return at the beginning of a trade day. Informally, we may describe this specific period as a "information correction" period. That is, everyone holds a different view at the very beginning, but after observing the tendency of the stock price and the emerging news published in the morning, they are converging to an "agreement" about what INTEL did last night, and this "mutual agreement" cause the peaceful variance of return curve after 10 am.

1.3 Overnight return

Overnight returns are defined as:

Table 1: **Overnight return pattern of INTC in Jan 1997**

	mean	var
C-O return	0.005693	0.000043
C-C return	0.010700	0.000373
O-C return	0.004229	0.000310

Figure 5: Descriptive Statistics of INTC 1997

sum_log_return stat:

	mean	var	skew	kurtosis	m3	m4	m5	m6
0	0.000144	0.000012	0.223593	2.483746	8.823317e-09	7.367307e-10	1.640987e-12	9.823959e-14

avg_volumn stat:

	mean	var	skew	kurtosis	m3	m4	m5	m6
0	1011.602854	73140.178195	2.022717	8.690687	4.001005e+07	6.253916e+10	8.822603e+13	1.425413e+17

trade_count stat:

	mean	var	skew	kurtosis	m3	m4	m5	m6
0	473.687063	118805.396826	5.90515	71.025027	2.418161e+08	1.044843e+12	4.927034e+15	2.387479e+19

Source: WRDS TAQ dataset

$$COreturn = \frac{Open_t}{Close_{t-1}}$$

$$CCreturn = \frac{Close_t}{Close_{t-1}}$$

$$OCreturn = \frac{Close_t}{Open_t} \ \ (sameday)$$

First, the product of 1+ Close-to-Open return and 1+Open-to-Close return is roughly equal to 1+ Close-to-Close return, which endorsed that our estimates are correct. Secondly, we could observe that the mean Close-to-Open return is very similar to Open-to-Close return, and the Close-to-Close return is almost twice as much as any of them. That may indicate that return is positively and proportionally related to the length of time. More interestingly, it is clock time instead of trading time. That could lead many interesting conclusions under different assumptions. For example, if we assume the stock is the comprehensive representation of every information traders could grasp from the market. Is the generation of information dominated by the operation of the company itself or the financial market? From our observation, the conclusion is "the company itself" since the return is proportional to the time the company operates instead of the stock market operates.

In contrast, the variance is disproportionate to the clock time. We could found the variance of Open-to-Close return is much larger than Close-to-Open return. Integrating the conclusion we established from mean, that means the uncertainty of return is mostly generated from trading, or from the stock market.

In conclusion, different moments of a distribution tell us different levels of information about the random variable, return, in this case. That's the reason we will discuss a quantitative comparison of all kinds of statistics in the next section.

1.4 Statistics

1.4.1 descriptive statistics

First, we could notice Figure 5 that the magnitude of all statistics among return, volume and number of trades are significantly different from each other. That is because we didn't standardize those numbers before the computation, which means any comparison across those three variables will not make much sense.

For Return, we could observe the mean and variance is quite small since the time interval is only 15 minutes long. Obviously, we won't expect any huge return rate for such a mature company, INTEL, in

Table 2: **regression results of INTC in Jan 1997**

	betas	gammas	R2
reg5a	-1.5566	-	0.01236
reg5b	-	5.0251e-09	0.05810
reg5c	6.81e-10	5.1745e-09	0.05919

such a short period of time. However, the mean of returns is positive, which indicates that most traders hold a positive attitude towards the future of INTEL. Moreover, skewness of returns are positive from my computation, which means in the large portion of trading time, the INTEL stock will only yield a relatively small return. This may not be a good news for some ambitious investors. Then, the kurtosis is both positive from my computation as well as the author's result. That is actually a good news for investors since a positive kurtosis indicates a flat tail of the return of INTEL stock. In other words, the return rate of INTC is quite stable. Even more stable than a normal distribution.

For Volume, and trade, both of them have positive skewness and kurtosis, which means a large number of trades came from a small period of time. That could be either the beginning of a day or the end. On the other hand, the tiny portion of traders contributes a large portion of trading volume. That implies the institutions and retail investors may play different roles in the stock market.

1.4.2 regression

First, we should deploy a pre-regression for obtaining the randomness of return using:

$$Y_t = \sum_{j=1}^{12} \delta_j Y_{t-j} + \varepsilon_t$$

And then, compute some variables we need for three formal regressions later:

$$|\hat{\sigma}_t|$$

= residual of 1st reg

$$\Delta V_t = volumn_t - volumn_{t-1}$$

$$\Delta T_t = Ntrade_t - Ntrade_{t-1}$$

We should notice that there are totally 12 auto-regression term for each regression:

$$\phi_j |\hat{\sigma}_{t-j}|$$

For formal regression, there are three types of them:

$$Volumn + auto$$

$$N_trade + auto$$

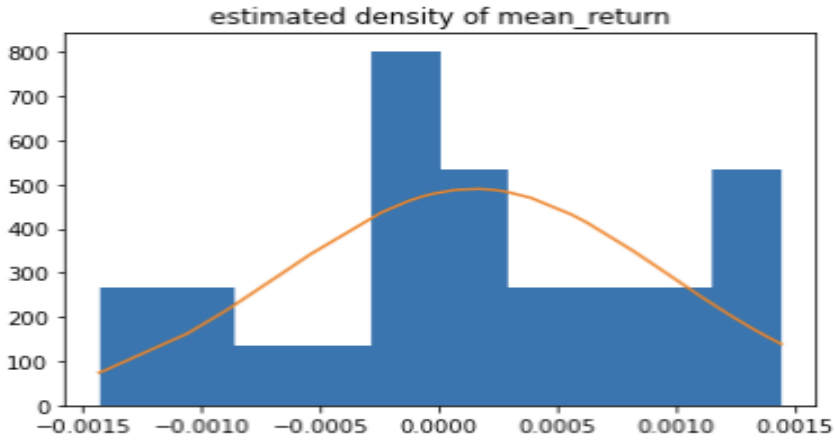
$$Volumn + N_trade + auto$$

Formally, their equations are:
$|\hat{\sigma}_t| = \alpha + \beta \Delta V_t + \sum_{j=1}^{12} \rho_j |\hat{\sigma}_{t-j}| + \eta_t^1$
$|\hat{\sigma}_t| = \alpha + \gamma \Delta T_t + \sum_{j=1}^{12} \rho_j |\hat{\sigma}_{t-j}| + \eta_t^2$
$|\hat{\sigma}_t| = \alpha + \beta \Delta V_t + \gamma \Delta T_t + \sum_{j=1}^{12} \rho_j |\hat{\sigma}_{t-j}| + \eta_t^3$

From Table 2, the first question could be possibly answered by table XX is that Volume and number of Trades, which one has more explanatory power over the randomness of return? We could clearly see the same pattern of R-square from the author's result and mine, which is that regression5b, which includes ΔT has significantly larger R-square than regression5a, which includes ΔV. Besides, For the purpose of further uncovering if the interaction term between trades and Volumes could provides some explanation of the randomness of return, we used the regression5c to fit the data. Result shows Adjusted R-square did not increase.

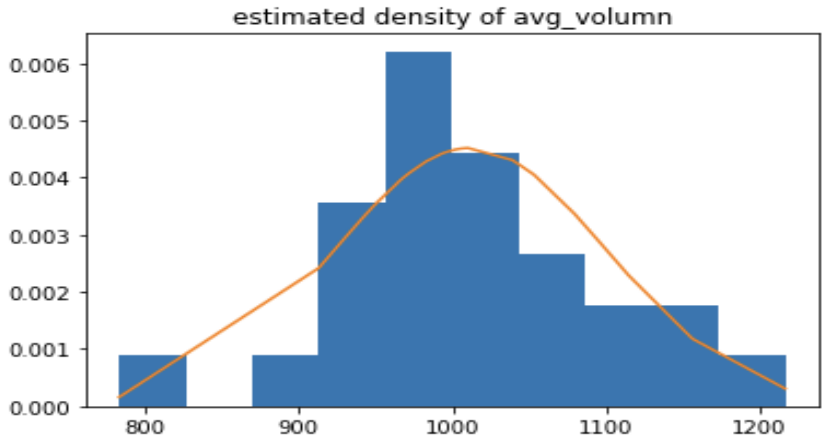
Following the conclusion, we already know that ΔT has certain explanatory power over the randomness of return, and volume does not. The next question is that will the increase of ΔT leads to the increase of $|\sigma|$? By looking at the regression coefficient, γ, we could safely conclude the increase of gap between the number of trades of 2 close trades will increase the randomness of return. This conclusion follows the

Figure 6: Estimated density of return of INTC 1997



Source: WRDS TAQ dataset

Figure 7: Estimated density of volumn of INTC 1997



Source: WRDS TAQ dataset

logic that the change of number of trades means new information(either bad or good) was obtained by investors and let them either buy or sell more stock than usual. On the other hand, "new information" means change of the current operation status of the company. For example, If INTEL suddenly receives a large deal from an OEM(original equipment manufacturer) in PC(personal computer), the investor will certainly increase the expectation of the value of INTEL's stock. Long traders will start to buy more INTEL stock, which results in the positive unpredicted move of its return.

By the way, the coefficient of ΔV is negative which is different from the author's result, but I think that also makes sense. First, we have know that ΔV has no explanatory power over $|\sigma|$, which indicates that the coefficient of it, β, should be not significant which exactly matches my regression result. Further, if an independent variable is not statistically significant in a regression, we could not tell anything from its coefficient.

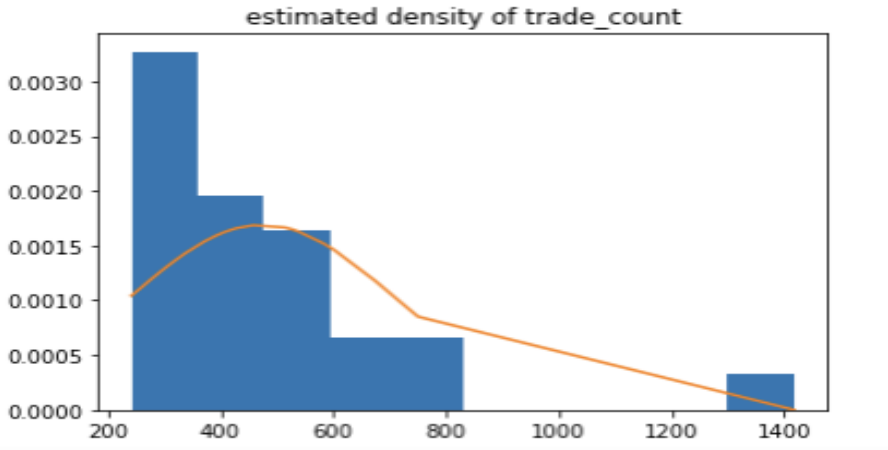
1.5 estimated density

Using kernel function

$$f(x) = \frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}}$$

, for return, my result is very similar to the result in the AG paper. That is, the estimated density of return is almost symmetric and more peaked than the Gaussian and exhibit leptokurtosicity. For the estimated density of trading volume, my result also share the same pattern with the paper. There is a clearly "second

Figure 8: Estimated density of the number of trades of INTC 1997



Source: WRDS TAQ dataset

Table 3: **P values of KS-stat of density estimates**

	P value
return	1.6e-06
volume	0.0
number of trades	0.0

high peak" at the right of the highest peak, which means there is more weight of density on the right part of the estimated distribution compared to the left. For the estimated density of the number of trades, my result also shares exactly the same pattern with the author – a distribution with large positive skewness. Following the K-S test, the hypothesis 0 of all three variables are rejected, which means none of the three variables are normally distributed. However, the return has the largest P value compared to the rest, which indicates that the density of return is more like a normal distribution compared to volume and the number of trades. A little bit more thought about this result is that: this may relate to the famous central limit theorem. Without formal proof, intuitively, return of a stock is the "ultimate result" of all kinds of information and trading on the market. That is saying, mean return is a function of all other random variables, such as information, volume, number of trades, number of bids, number of investors... Finally, we took the mean of all of the ultimate random variable, returned. It's plausible that this ultimate random variable will very likely follow the normal distribution.

1.6 trade-time interval

Figure 9 is the density plot of return under trading-time intervals. Compared with returns under clock-time interval, we could clearly observe that returns are more centralized, or, more leptokurtic. In other words, clock-time return shows fat tails in density estimate. One fact for clock-time returns is that the number of trades are disproportionate to the clock time.(See figure 1 in section 1) That is saying, at the end and beginning of a trade day, there are huge amounts of trades in the 15-min interval. Another phenomenon we captured from figure XX is that return may be positively related to the number of trades. Those two conclusions from section 1 help us explain why there are some "outliers" that exist on the tail of the density estimation under clock-time, and the fat tails disappear when the intervals are separated under equal-number-of-trade basis.

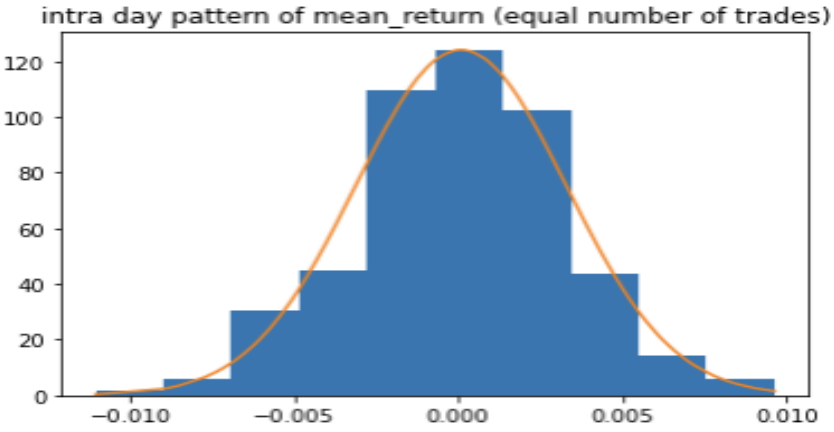
2 comparison

2.1 cross time comparison

2.1.1 intraday pattern comparison

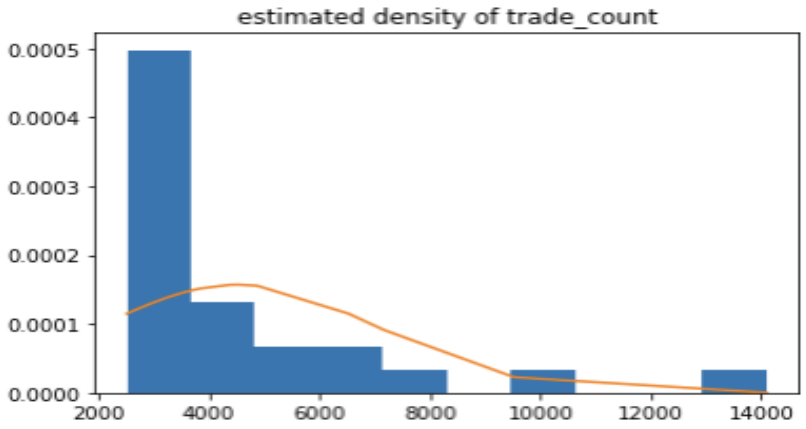
Most intraday patterns still exist except for a number of trades, which looks like a mirror symmetry where we observed in Jan 1997. Traditional theory told us, in the last minutes of trading, the increase of trading can often be attributed to certain kinds of traders closing out their position before the end of the day. For

Figure 9: trade-time return estimated density of INTC Jan 1997



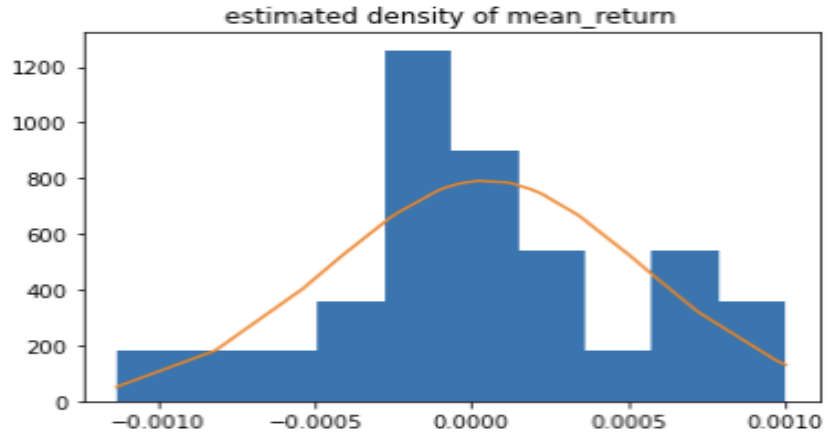
Source: WRDS TAQ dataset

Figure 10: intraday pattern of the number of trades of INTC Jan 2013



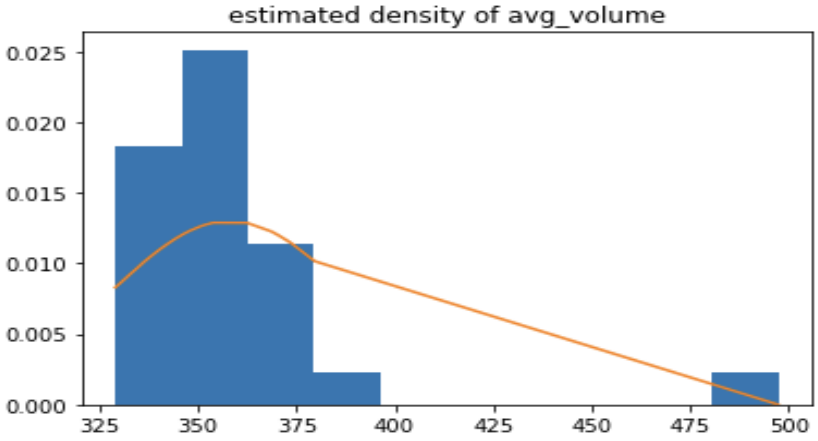
Source: WRDS TAQ dataset

Figure 11: intraday pattern of return of INTC Jan 2013



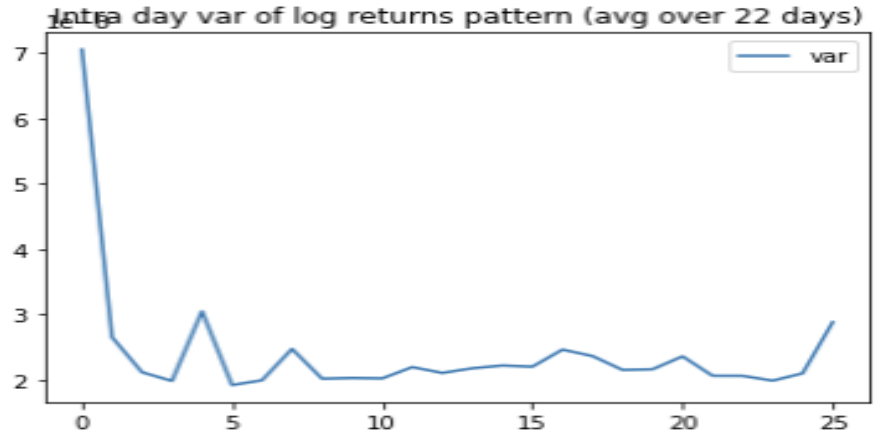
Source: WRDS TAQ dataset

Figure 12: intraday pattern of volume of INTC Jan 2013



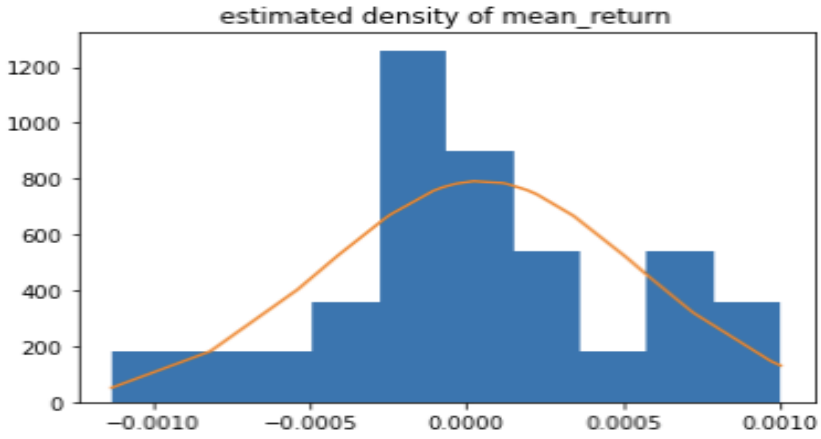
Source: WRDS TAQ dataset

Figure 13: intraday pattern of variance of INTC Jan 2013



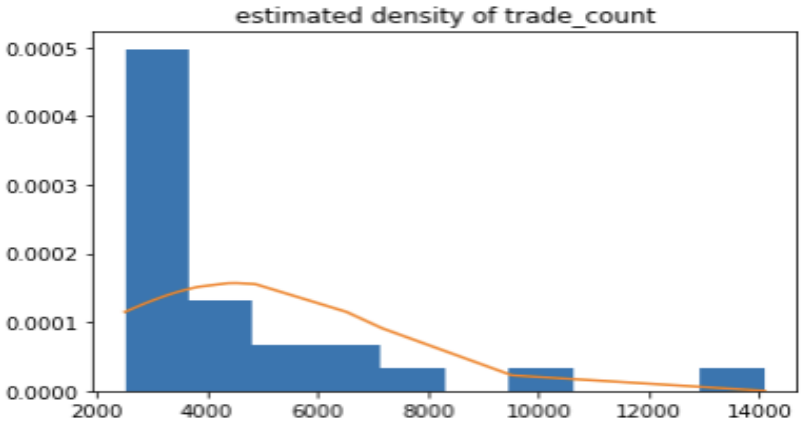
Source: WRDS TAQ dataset

Figure 14: estimated density of return of INTC Jan 2013



Source: WRDS TAQ dataset

Figure 15: estimated density of the number of trades of INTC Jan 2013



Source: WRDS TAQ dataset

example, if I don't want to take the risk of a large price movement at the start of the next day affecting me, I would need to completely close my position. However, why is the number of trades higher than the beginning of a day? During the 13 years, the number of retail investors kept surging. Since the number of trades does not consider the size of each trade, I guess retail investors tend to trade after their work. Intuitively speaking, I, as a retail investor, rarely trade in the morning since I have something else to work with, but only trade after work.

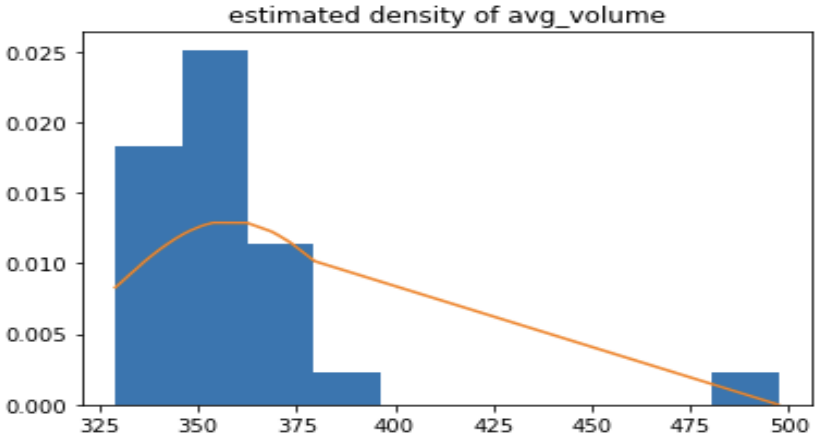
2.1.2 estimated density comparison

Another pattern change significantly over the 13 years is how an investor achieves a certain position. From figure 2 we could say that the magnitude of average volumes is much less than what we observe in Jan 1997. For example, most volumes are between the interval of 325 and 400 in Jan 2013. However, they are most between 800 and 1200. In the meanwhile, the magnitude of the number of trades increased a lot, from roughly between 200 and 800 in Jan 1997 to between 2000 to 8000 in Jan 2013. Those two changes shows investors tend to split their single large bids into several smaller bids to achieve a more accurately controlled portfolio. The rising quantitative trading technology may help explain this tendency.

2.1.3 statistics comparison

First, the mean return of INTC decreased from 0.00144 in Jan 1997 to 0.00044 in Jan 2013, which means investing in INTC stock may not be a good choice for those ambitious investors. However, the variance

Figure 16: estimated density of volume of INTC Jan 2013



Source: WRDS TAQ dataset

Figure 17: descriptive statistics of INTC Jan 2013

	mean	var	skew	kurtosis	m3	m4	m5	m6
0	0.000044	0.000004	0.965866	9.55291	7.973184e-09	2.094262e-10	1.999993e-12	3.339485e-14

avg_volume stat:

	mean	var	skew	kurtosis	m3	m4	m5	m6
0	358.042044	6120.30097	2.51152	14.065846	1.202530e+06	6.392539e+08	3.713002e+11	2.419674e+14

trade_count stat:

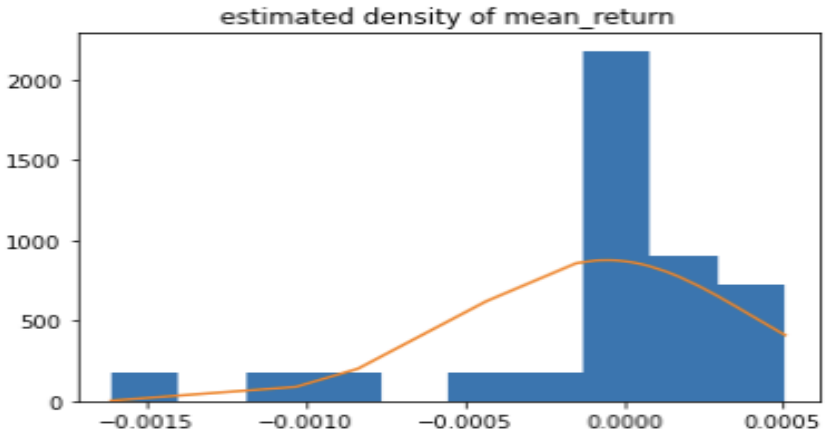
	mean	var	skew	kurtosis	m3	m4	m5	m6
0	4515.214286	1.672114e+07	6.952597	77.171468	4.753853e+11	2.241566e+16	1.165045e+21	6.325249e+25

Source: WRDS TAQ dataset

Table 4: **regression results of INTC in Jan 2013**

	betas	gammas	R2
reg5a	0.881080	-	0.163883
reg5b	-	0.876652	0.752578
reg5c	-1.044326	0.878782	0.753353

Figure 18: estimated density of TSM Jan 2013



Source: WRDS TAQ dataset

of return also decreased from 0.000012 to 0.000004. The decrease of both mean and variance of return indicates that INTEL is becoming a more stable company and tech giant. People no longer have high expectations for the huge increase of its performance. Skewness also indicates Intel is becoming more and more mature and stable, which means skewness is increasing and a high return rate is much rare compared to 1997.

About the explanatory variable of randomness of return, from Table 4, I found the conclusion we found in section 1.4.1 also still holds 16 years later.

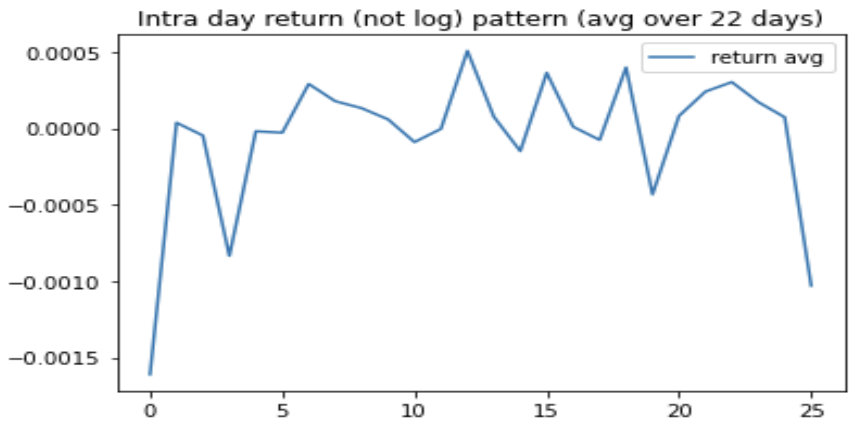
2.2 cross industry comparison

I am always excited to compare the business of TSM and INTC. They are the only remaining two of three major plays in the semiconductor manufacturing industry(the third is samsung). Without TSMC(Taiwan Semiconductor Manufacturing Company), I think the current technology will fall back at least 3 years. TSMC, a Taiwanese chipmaker, runs a totally different business mode compared with INTEL. They do not design chips, but only be responsible for manufacturing, packaging and testing. Although INTEL designs chips as AMD, it also owns its own semiconductor manufacturing plants.

From Figure 18, first thing I noticed is that the estimated density of return of TSM is significantly different with INTEL: the magnitude of average return is much lower than INTEL in 2013. At 2013, intel announce its 3rd generation intel core processor, which use 22nm manufacturing technology. However, TSM is still struggling in 28nm. We have to know that intel is more targeting PC and industrial high performance computation clusters, and TSMC mainly produce chips for mobile devices including chips for Qualcomm Inc and Taiwan's MediaTek Inc. Chips for mobile devices generally have more strict energy upper bound due to its limited battery size, but TSMC, as a main supplier for chips of mobility devices, has no ability to produce chips made by smaller manufacturing technology. We could guess that they didn't sell well compared with Intel in 2013. In the meanwhile, Investors generally did not have a high expectation for their stock price.

We could found investors' pessimistic attitude toward TSM in the intraday pattern of return as well. It shapes like a M instead of W where we obtain from INTEL at the same time.

Figure 19: intraday pattern of return of TSM Jan 2013



Source: WRDS TAQ dataset