**Q1:**

How many times did the bid-ask spread (defined as the ask price less the bid price) widen in this dataset? What proportion of those times did it widen on both sides (bid decreased and ask increased)? What proportion of those times did it widen on one side (bid decreased xor ask increased)? How many times did the bid-ask spread tighten? We define market size as (bidsize + asksize)/2. Report the distribution of market size.

**Response:**

The bid-ask spread widened 233 times in the dataset; only one of them was on both sides, and the rest 232 times was on one side. The bid-ask spread tightened 235 times.

Distribution of market size:

Chart, histogram

Description automatically generated

**Q2:**

Focus on instances in which the mid price changes but the bid-ask spread does not. For an increase (decrease) of the mid price, define the new bid (ask) as the aggressive side, and the other as the defensive side. Report the distribution of the market sizes of the defensive side and aggressive side immediately following a change in mid price. Can you intuitively explain your results?

**Response:**

For instances such that the mid prices changes but the bid-ask spread does not:

As shown in the graphs below, defensive size was larger than the aggressive size. Take an example such that the (bid, ask) was raised from (27.00,27.02) to (27.01,27.03) such that the spread did not change, and the bid size was the aggressive side. Before the change there were already existing orders asked at 27.03 as it was above the best ask, but all the orders bid at 27.01 should be submitted after the change. Thus during a very small time interval the best ask (defensive) size should be larger than the best bid (aggressive) side. This also explains why stock prices are mean-reverting in the very short run.

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

**Q3:**

If, from one update to the next, volume increases from 100 to 150, we know 50 was traded between the two updates. Many trades of varying size at various prices could have happened between the two updates; we only know that the most recent trade happened at the lastprice. Come up with a realistic volume allocation method that allocates the traded volume to price levels. Justify the method and append your allocation to mdLog.csv.

**Response:**

The method I used was evenly split trading volume for every possible price level between the old price and new price. Assume prices levels are only allowed at two decimal places, the formula should be trade\_volume\_at\_each\_price = trades / (abs(last\_price\_old, - last\_price\_new)/0.01 + 1).

I use this method because we have no knowledge about the ask and bid sizes at less priority. For example, if the price moved from 20.00 to 20.05, we know that every ask orders from 20.00 to 20.04 and some ask orders at 20.05 should be consumed, but we only know the ask size at 20.00 prior to the price change. Thus we could only assume there are same amounts of ask orders at {20.00, 20.01, 20.02, 20.03, 20.04, 20.05} and allocate the trading volumes evenly.

I used two ways to keep track of allocated traded volumes:

1. record them in dictionary and collect total traded volumes at each price level over the dataset

Chart, line chart

Description automatically generated

2. Append a new column called ‘trade\_at\_last\_price’ representing the estimation of trade volume at the last price happened between last snapshot to current snapshot.

**Q4:**

We define sizeDelta as the size added or cancelled on a level from one update to the next, net of any traded volume. For example, on update 1, the bidsize was 100. On update 2, the bid price is unchanged and the bidsize is now 70. Suppose, from part 3, that we believe a total of size 20 was traded at the bid.Then we have sizeDeltaAtBid = -10, as in orders totaling size 10 were canceled between update 1 and update 2 on the bid. Note that sizeDelta is only defined for price levels that are unchanged from the previous update. Add to mdLog.csv two additional columns: sizeDeltaAtBid and sizeDeltaAtAsk. Report the distribution of these two variables.

**Response:**

Most of the observations were around zero, but there were also some significant outliers that stretches the x-axis.

Chart

Description automatically generated

Chart

Description automatically generated

**Q5:**

Now we will focus on the aggressive side. Assume the size on the aggressive side we see immediately after a price change is from a single limit order, which we call the top order. We wish to track the performance of top orders. Assume cancellations always happen from the back of the queue and the order may be partially cancelled.

Note that all top orders are inserted at the BBO (Best Bid Offer), as in a buy (sell) top order is always inserted at the best bid (ask). However, not all top orders spend their entire lifespan on the BBO. Report the total number of top orders, the number that do not spend their entire lifespan on the BBO, and the number that do. Out of those that do spend their entire lifespan on the BBO, how many are filled for their original size?

Response:

My program gave me a result of 181 top orders, 57 spend entire lifespan on the BBO, and 30 of those are filled for original size. Notice that as I only have the best bid and ask data, it is very hard to track an order, and I have to make lots of assumptions. I included faily large amount of comments in my code; please check out my comments in codes to have a sense of my thought process!

**Q6:**

We define return as the signed difference between the execution price of an order and the mid price 40 market updates after the time of the final trade. By convention, we normalize these returns by the ticksize, which is 0.01 in this case. Therefore, if a buy order is all traded at 19.20, the bid 40 updates later is 19.20, and the ask 40 updates later is 19.21, then the return is 0.5. Report average returns of top orders that spend their entire lives at the BBO and are completely filled.

**Response:**

The average return is 1.017 under my assumptions and restrictions

**Q8:**

Focusing on top orders that spend their entire lifespan at the BBO and are completely filled, can you write a model for the return of these orders based on the size of such orders and the sizeDelta variables you see on the second update after the order was inserted? Can you improve your model via other predictors with information taken on or before the second update after the order was inserted?

**Response:**

The significance was not quite satisfying. Probably we could include some trends before the top order was inserted, to see if matching or reversing the trends could make the top order more profitable.

Graphical user interface

Description automatically generated with medium confidence

**Q9:**

What are the pros and cons of using the above definition of returns as a metric for order attractiveness? What are some (potentially better) alternatives?

**Response:**

Pros: quite intuitive and understandable

Cons:

1. Such returns may not be attainable as I am not particularly familiar with China stock market short selling regulations. So, if it is a bid order we could definitely use this definition (just sell it 20 updates after acquiring), if it is a sell order we should consider the average cost of acquiring or borrowing. Of course, any transaction costs incurred should be considered as well.

2. Order attractiveness should consider order size. Larger size profits are harder in the capital markets and should be more desirable to us as investors. Probably we could predict PNL (size \* return), or if this makes the Y variable too biased, use adjusted size like sqrt(size).

**Q10:**

Any other interesting observations you want to point out? What assumptions in the problems struck you as unrealistic or overly simplistic?

**Response:**

Observations: top orders seem to make money.

I would say the assumptions to determine top orders are unrealistic but given dataset only with best bid and ask price & sizes that it what it is. If we could have a dataset that is deeper on both ends we could make better assumptions. Also the “volume allocated to this price level” is overly too simple (average of all possible levels); again if we have dataset with deeper bid / ask knowledge like the sizes of top 5 bid prices, we could do much better.