

./ Possibility and Probability

A Python programmer with a personality thinking about space exploration

15 March 2010

The iPad and the German Tank Problem

by [nickadmin](#)

Based on order numbers the day the iPad was made available for pre-order, several sites were speculating that Apple was selling about 50,000 iPads per hour. This reminded me of something I heard about back when the iPhone was just getting started: The [German Tank Problem](#). Its basically a way of statistically estimating the maximum number of "items" that have been produced based on the serial numbers of the item. This method works when the serial numbers are in sequential order, and basically allows you to produce a somewhat realistic estimation (as opposed to a wild-ass guess). Estimation is one of those skills that most people (myself included) could stand to improve a little bit. In the wikipedia link above, this is shown by the intelligence estimates the Allies had for the number of tanks the Germans were producing in the middle of WWII. The initial estimates were really high (1,000 per month), but using statistics based on the serial numbers of crankcases from captured or destroyed German tanks showed that the number might be lower (around 200 per month, or 1/5th of the original. After the war when the factory records were looked at, the true number was a lot closer to 200 than 1,000.

Math FTW

The formula is pretty easy:

$$N=m + (m/k) -1$$

Where m is the largest serial number observed, and k is the number of serial numbers seen. The [variance](#) is roughly equal to:

$$(N^2)/(k^2)$$

Which means that the [standard deviation](#) is roughly:

$$N/k$$

So, how does this apply to the iPad's initial orders? Based on the data points of 2 known orders spaced 50,000 numbers apart (keeping in mind these order numbers also probably

included orders for items other than iPads). So plugging those two numbers into the equation we get $N =$

$$m=50000 \gg k=2 \gg N = m + (m/k)-1 \gg N \text{ 74999} \gg$$

Wow, almost 75,000... That seems like a really big number. The question we should then ask ourselves is "How realistic is this number?" Using the standard deviation and variance we could find out how spread out our numbers are:

$$(N^2)/(k^2) \text{ 1406212500L} \gg N/k \text{ 37499} \gg$$

Wow. Those numbers huge. And that is a very bad thing. The bigger the standard deviation and variance, the more less accurate the estimation. Another way to approach this analysis is to look at the confidence interval and see how big it is. The wikipedia article has a handy formula for finding the [confidence interval](#) which leads us to the estimation that there are (based on $k=2$ and $m=50,000$) between 50,000 and 225,000 iPads ordered!

Those numbers are lying

Why? There's two reasons: The main one is that our sample size (of 2 orders) is waaaaaaaay too small. Its like trying to guess how big your grocery bill is by averaging the price of two items, and then multiplying it times the number of things you bought. That estimation will be way off. As an example, if there were 20 orders to work with ($k=20$), the high end of the confidence interval would shrink to 58,000. But that leads to the second reason why we can't trust these numbers: We don't know the lower bound. In other words, yes, there could have been 50,000 orders between the first and last data points. But what if only half of them were for iPads? That would mean that m is actually 25,000 which would drastically skew the numbers down. Remember that N that was almost 75,000? With m at 25,000 (and keeping $k=2$) N drops to 37,499 which is half the original estimate!

So how many were sold?

That's a really good question. Knowing Apple and how people love their products, I bet they sold a TON of iPads. But based off these rough numbers we see in the news, we can't really draw a good conclusion. We can get a couple of estimates which are better than nothing, but they are so numerically shaky (huge standard deviation and enormously questionable confidence interval) that they strain believability. More data points will help establish an upper bound (i.e. the estimated maximum number sold), but without a lower bound to keep us grounded the numbers will still look really huge.

tags: