

Table of Contents

How Can We Forecast When We Do Not Have a Lot of Data?	3
---	----------

How Can We Forecast When We Do Not Have a Lot of Data?

“How much data do we need to generate a useful metric?”

Many people think that you need a lot of data in order to get statistically useful data. For example, we hear a lot about “Big Data” and how we can mine huge volumes of data to produce interesting information. Many of us assume that in order to do something meaningful, and with any accuracy at all, we also will need a lot of data. It turns out that in many cases you really do not need much data to produce useful and interesting results.

For example, let's say we just started working in an agile fashion. We have 6 sprints under our belt and someone is trying to figure out what might happen in the future. Do we have sufficient data to be able to do anything useful?

Let's try working through an example. Say the velocity history of a team over the past 6 sprints is 20, 25, 23, 15, 30, and 27. We can say, based on this data that the worst this team could perform for the next 4 sprints is 4 times lowest number (15) for a total 60 points delivered. Again, we don't think this is likely, but it is possible. We can say the best this team could perform for the next 4 sprints is 4 times the highest number (30) for a total of 120 points delivered. Again not likely, but possible. The average of this sequence is 23.3 points which after 4 sprints would give a total of ~93 points delivered.

But we actually have more information than this. When we have completed one Sprint, we have a single known velocity, in this case 20. What this means is that in the next Sprint the velocity number we will have will either be higher than that (20) number, or lower, with about a 50% chance of either event happening and, ignoring chance that number is exactly the same, 100% it will be either bigger or smaller than the number we have seen so far.

For the second sprint, we will have two velocities, the 20 we just talked about, and a 25. If we are trying to understand what velocity we will have in the future based on these two numbers, we have about a third (33%) chance that future numbers are below the 20, another third that they are between 20 and 25, and another third that it will be bigger than 25. In other words, there is a 33% chance that future numbers will be in the range of 20 to 25, and a 66% (2 x 33%) chance it will be outside that range.

Here's what is interesting. By having two data points instead of one we've gone from 100% uncertainty to the understanding that we know that there is a 66% uncertainty that the next number will not be in the 20 to 25 range. Now I realize this is not useful yet, but watch what happens as complete Sprints and have more velocity numbers:

Sprint	Velocity	Min - Max Range So Far	Chance Next Sprint's Velocity is Outside This Range
1	20	20 - 20 (1 data point)	100%
2	25	20 - 25 (2 data points so set min and max)	66%
3	23	20 - 25 (no change)	50%
4	15	15 - 25 (new minimum)	33%
5	30	15 - 30 (new maximum)	25%

Sprint	Velocity	Min - Max Range So Far	Chance Next Sprint's Velocity is Outside This Range
6	27	15 - 30 (no change)	20%

In other words, by the sixth Sprint, we only have a 20% concern that the next Sprint's velocity is going to be outside the 15 - 30 point range and we are 80% sure it will be in that range.

That's a pretty small amount of data to generate a lot of understanding of the data we have. And the thinking approach can be applied to all kinds of metrics and at all levels of the planning process - portfolio level epics, program level features etc.

[Estimates](#), [Forecast](#), [Data](#), [Metrics](#), [Probability](#), [FAQ](#), [Points](#)

From:

<https://www.hanssamios.com/dokuwiki/> - Hans Samios' Personal Lean-Agile Knowledge Base

Permanent link:

https://www.hanssamios.com/dokuwiki/how_can_we_forecast_when_we_do_not_have_a_lot_of_data

Last update: 2020/06/04 09:15

