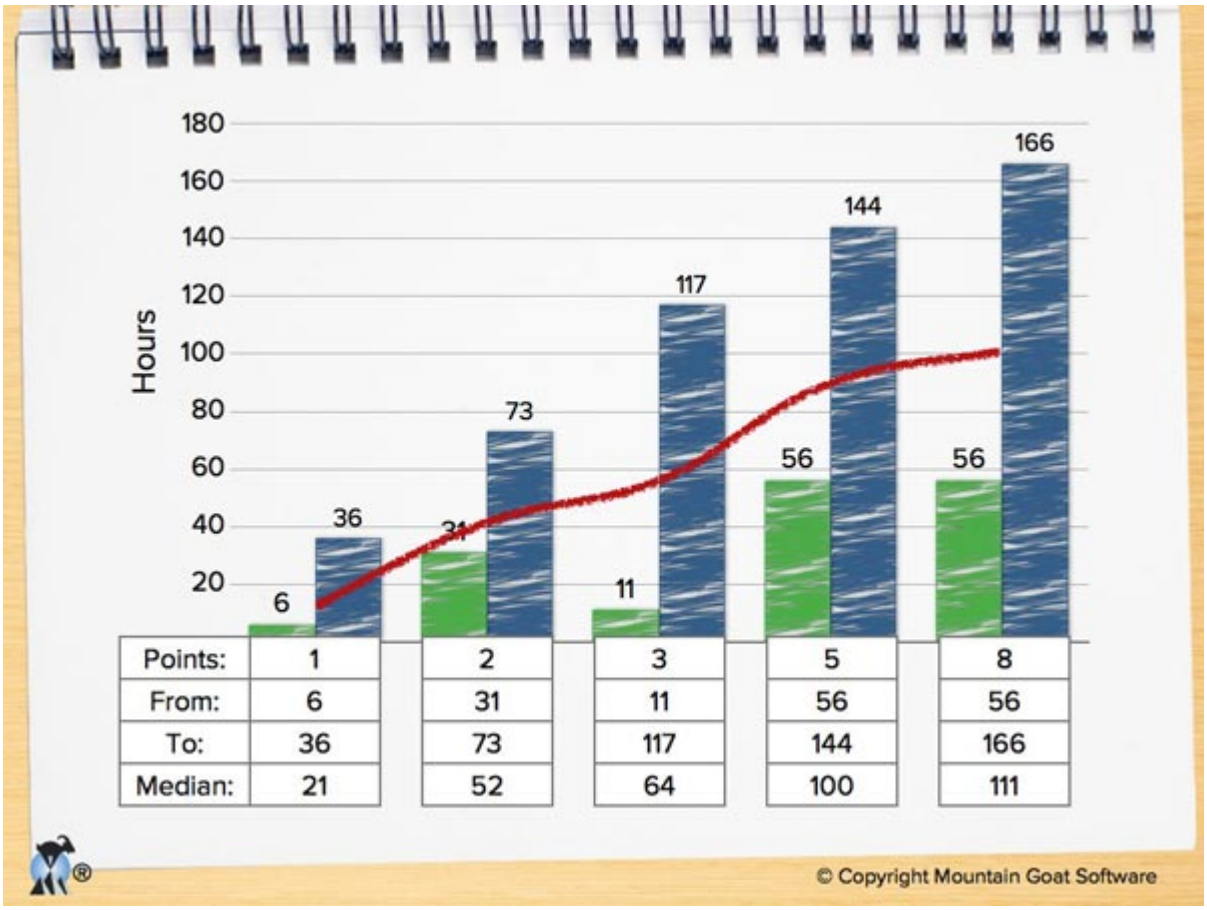
Seeing How Well a Team's Story Points Align from One to Eight

by Mike Cohn • 41 Comments

The topic of how well a team estimates two point stories relative to one point stories (and so on) has come up in a couple of comments and replies on this blog recently, so let's discuss it. Here's a graph showing relevant data from one company:



Each column of data and pair of bars shows data from stories of the size given in the Points row. The first set of data is for one-point stories, the second set of data is for two-point stories and so on. Looking at the one-point data, we see that the median number of hours to complete a one-

point story (at this company) was 21 hours. The shortest (From) took 6 hours and the longest (To) took 36 hours. The shortest and longest are shown in the green and blue bars above the table. The median is graphed as the red line. Let's look at the two-point stories. We see there a median effort of 52 hours and a range from 31 to 73. If we assume the 1-point stories were perfectly estimated, we would expect 2 point stories to have a median of 42 instead of the 52 we see here. Or perhaps the 2-point stories were done perfectly and the 52 is right. In that case, the median for one-point stories should have been 26. Most likely, neither is perfect and the "perfect" estimates are somewhere between.

Either way, the team has drifted a little bit because the median of the 2-pointers does not equal twice the median of the 1-pointers. But let's look at the three-point stories. Three-point stories should presumably be triple the median of the 1-point stories. The median should be 63 and we see that it is 64. Wow. Very close. The five-point stories should presumably have a median of 5x21=105 and they come in at 100. The eight-point stories should presumably have a a median of 8x21=166 but they only come in at 111. But, hold on, we have to make a bit of an adjustment in this thinking for the 5- and 8-point stories. If you use these numbers the way I recommend, you know to think of them as buckets. An 8 is a bucket that holds all stories from six through eight points in size. That is, if you have a story you think is a six, you can't fit it into the five-point bucket so it goes into the eight-point bucket. This means that our five-point stories are really fours and fives. If we have the same number of each, the average size of a story in the five-point bucket is really 4.5 points. Multiplying 4.5 points x 21 hours (length of the 1-point story) give 94.5, off by about 5% from our measured median of 101 hours. Not so bad. Since the 8-point bucket holds stories of size 6, 7, and 8, we can assume the average story given an 8 is really worth 7 points. And 7 points times 21 hours = 147 hours. The 8-point stories should have a median of 147. They don't, they have a median observed value of 111.

So this team has definitely drifted by the time they reach their 8-point stories. (Let me add a sidenote here that rather than simply assuming the 1-point stories were perfect and using their 21 for all this mulitplication, a better approach would be to do linear regression analysis, which can be done with Excel. You can then look at the r-squared value to see how well the values fit. But, I didn't want to go through all that math here and with data like this, we can, I believe, see that things work out pretty well up to 8.) I encourage you to think about collecting data like this at your company. You need to be careful though. Because you'll be collecting actual effort expended on each user story, it's possible that team members feel more than the normal

amount of pressure to finish within any estimates they give. They may then respond by padding their estimates. This defeats the whole purpose.

So, show a graph like the one above to the team and be clear that having this data can help them. For example, the team above could learn that they put 8s on stories that should perhaps have been 5s. (Looking at the data, they could also learn that they estimated correctly but that they really had more sixes in the eight bucket.) Most teams who do this will find that they are good through about 8, as in this example. With some awareness of data like this and some additional practice and calibration, just about any team can get good across a 1-13 range. Beyond that is tough and numbers above 13 (and possibly starting with 13) should be used with caution or only for answering rough, long-term questions like, “Is this project a couple of months or are we talking about a year to do it?”

If you collect data like this for your teams, I would appreciate it if you would share it with me. You can just email it to mike@mountaingoatsoftware.com rather than posting it here. I've been working on some analysis of data like this and the more teams I have, the better.

Posted: September 19, 2011

Tagged: user stories, teams, estimating, story points, metrics