Learning Project Completion Probabilities

The Problem

In a project file, for a given task, the assigned team or an estimator provides a best-case, expected, and worst-case three-point estimate for the time or cost to complete. While the task is in progress, the team provides status updates with completion percentages at possibly irregular intervals. The challenge is to use this incoming data to determine the true probability of completing the task on time and assess the likelihood of meeting the original expected date. This process μst then be scaled across all tasks to forecast a completion probability for the entire project.

The Main Idea

Any process that unfolds over time can be described in terms of a simple relationship: progress = rate × time. Whether measuring physical distance traveled, percentage of work completed, money spent, or units produced, the underlying idea is the same—an estimate of the total effort implies an expected rate at which progress occurs for the time of costs consumed (Equation 1).

$$Completion\ Velocity = \frac{Progress}{Duration}$$

$$Spend\ Velocity = \frac{Progress}{Actual\ Spend}$$

Equation 1

A progress update that combines actual spending or elapsed time with the percent complete provides insight into the actual rate of progress. For example, if a task is 40% complete after 3 weeks, the implied rate is roughly 13.3% per week. Meanwhile, spending $200,000 to reach 25% completion on a $1,000,000 task indicates a spending rate of $8,000 per percentage point of progress. These updates clearly show how quickly work is advancing or resources are being used, in comparison to the rates estimated initially.

Once the velocity has been adjusted based on actual observations, it can be combined with the remaining effort to directly estimate the final time and cost to complete. For example, if 40% of a task is finished and the updated velocity shows 12% progress per week, the remaining 60% suggests about five more weeks of work. The same logic applies to costs: if the observed spending rate is $90,000 per month and $450,000 worth of work remains, the projected total cost at completion is roughly $900,000. In this way, updated velocity offers a continually improving basis for predicting both schedule and budget outcomes, turning raw progress data into actionable forecasts.

The initial estimates and updates are exactly the kind of data captured and tracked in project management and Earned Value Management (EVM) tools, where planned versus actual rates of progress and spending are monitored to assess performance, forecast outcomes, and update the likelihood of meeting schedule and cost targets.

A key principle underlying this paper is that the initial prediction of an uncertain variable and observations of it are exactly the setting for applying Bayes' theorem: The initial estimate is the prior, and the ratio of percent completed to actual costs or spends are the observations. By treating these updates as evidence, we can apply Bayesian refinement to continually update the probability distributions of the velocities. Dividing this updated velocity into the remaining effort yields a remaining cost or duration.

- Initial estimates of time or cost assume a certain velocity (rate of progress or spend).
- Progress updates with actual duration, spend, and percent complete are observations of that velocity.
- These observations allow Bayesian refinement, updating the prior (initial estimate) into an improved probability distribution of velocity.
- With the updated velocity, the remaining effort can be divided by that rate to forecast the eventual time and cost to complete.

Further, applying the ideas from Appendix 1,

- The initial estimate is an uncertain variable and so is suitable as a Bayesian prior.
- All the arithmetic calculations described above are carried out using a Monte Carlo siµlation, as described in Appendix 1.

The Methodology

To explain how the parameter learning algorithm works, we will use a time-to-complete example. Suppose that in a project file, for a given task, at the onset, the assigned team or an estimator provides a best-case, expected (aka deterministic) case, and a worst-case 3-point estimate for time to complete. While the task is in progress, the team provides the status date and completion percentage at (possibly) irregular intervals.

We need to determine how to use the initial estimate and actuals to find the probability of completing the task within a specific time frame and the likelihood of meeting the planned completion date. For example, suppose the team estimates they will probably finish a task in 12 weeks, with a best-case estimate of 10 weeks and a worst-case of 15 weeks. A linear completion velocity would be about 8.3% per period.

They report percent complete at possibly irregular intervals (Table 1).

Period	Percent Complete
2	15
3	40
5	40
8	70
10	90

Table 1

Examining Table 1, we observe that it is possible that there is no progress between reporting periods.

We will use the initial estimates and the observed progress to learn the team's velocity and time to complete the task. This process is described in detail below.

The team's velocity is uncertain, so we need to find its PDF. We will learn by applying the Bayesian parameter learning (BPL) method introduced in Appendix 1. To do this, we will use a technique found in (Fenton and Neil 2019). The main idea is to choose a model-parametrized PDF and treat its parameters as random variables and then derive the resulting PDF. In this, we use a unimodal model whose parameters are its *location* and *dispersion,* as described in Appendix 1.

Here are the steps of the velocity learning algorithm:

1. Pick a model parameterized PDF whose parameters are to be learned
2. Use the original estimates to build the prior location and dispersion PDFs
3. For each observed data point:
    3.1. Stage the actuals to be suitable as observations for applying a 2-D version of Bayes theorem.
    3.2. Build a 2-D likelihood grid by using the numerator of the Bayes' theorem ratio
    3.3. Compute the normalized marginals of the grid to get PMFs of the location and dispersion parameters.
    3.4. Interpolate the PMFs to get the PDFs of the location and dispersion PDFs
    3.5. Save these to be priors for the next observation
4. Use a Monte Carlo siµlation with the PDFs and the model to construct the velocity PDF.

5. Use the velocity PDF and Monte Carlo simulation to divide the remaining percentage to be completed by the velocity to calculate the PDF of the current duration or expenditure needed to complete the effort.

Let's go through each of these steps:

Step 1: Pick the model PDF

We will choose between two PDF models with location and dispersion parameters: the normal and the logistic. In both cases, the location is the mean, μ, and the dispersion is the standard deviation, σ. Note that even though these are symmetric PDFs, the eventual velocity PDF and the predictions of cost and duration will likely be skewed.

Step 2: Build the parameter priors

The method requires the original estimates to be uncertain variables. This can be determined in a couple of ways:

1. Triangular Distributions.

   Project management tools often gather 3-point estimates for task durations or costs (best case, planned, worst case). The assumption is that the team believes the actual time or cost is unlikely to fall below the best-case scenario or exceed the worst-case scenario, with the planned case being the most probable. It is also common in risk analysis to represent this input as triangular distributions (Mun 2003). See Figure 1 for an example. These are explained more thoroughly in Appendix 1. For instance, if the expected (also known as deterministic) effort is 12 team periods, the low is 10, and the high is 15, then the PDF is shown in Figure 1.

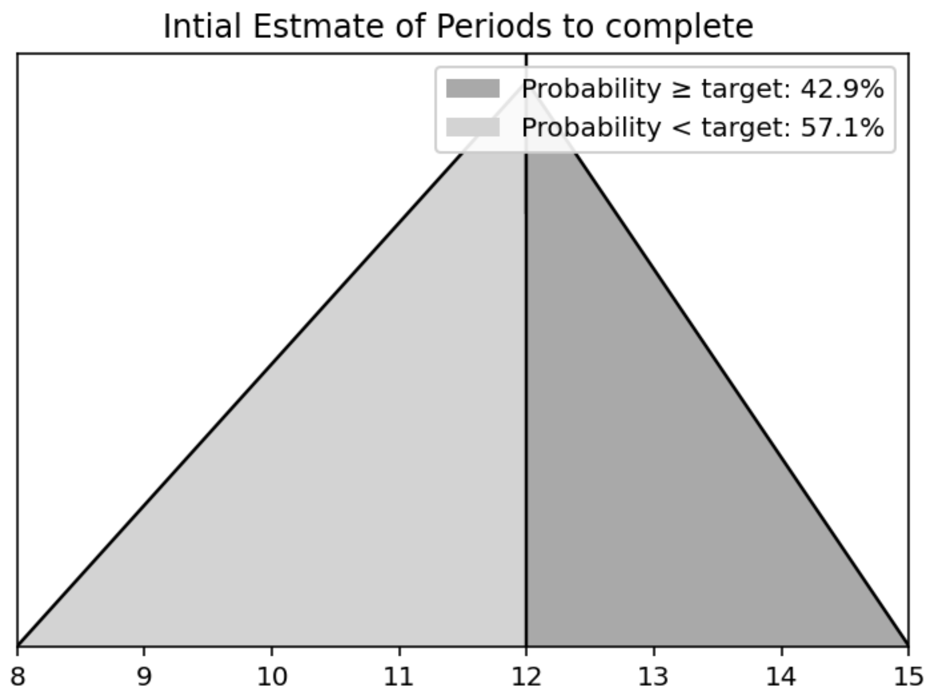

Figure 1

2. Empirical PDFs

   These can be based on team history or may be supplied by using the estimation techniques in Chapter 5.

A key insight is that these time or cost to complete estimates imply a corresponding velocity. They imply that the task will be 100% completed within the estimated cost and time. So the implied velocity is $\frac{100\%}{12\ periods} \cong \frac{8.33\%}{period}$. So, in general, the initial velocity is given by Equation 2.

$$Prior\ Velocity = \frac{1}{Initial\ Completion\ Estimation}$$

Equation 2

Since the initial estimate involves an uncertain variable, to determine the prior velocity, one needs to use Monte Carlo simulation. See Figure 2. Note that the expected effort level and the deterministic plan's progress per period is 1/12 = 0.08333… So, as expected, peaks around 0.083.

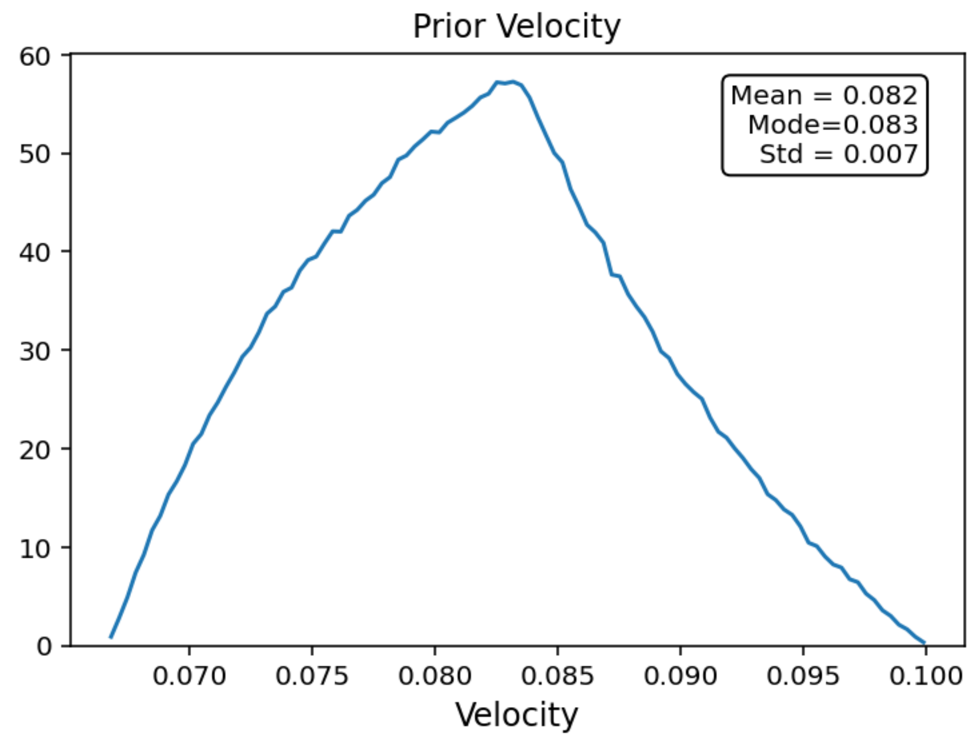

Figure 2

Another key insight is that Figure 2 gives the prior location, μ, for the velocity.

To set the prior for σ, we use a uniform distribution, also known as an uninformative prior, with a low value of 0.01 and a high value of 10 times the standard deviation of μ, as shown in Figure 3.

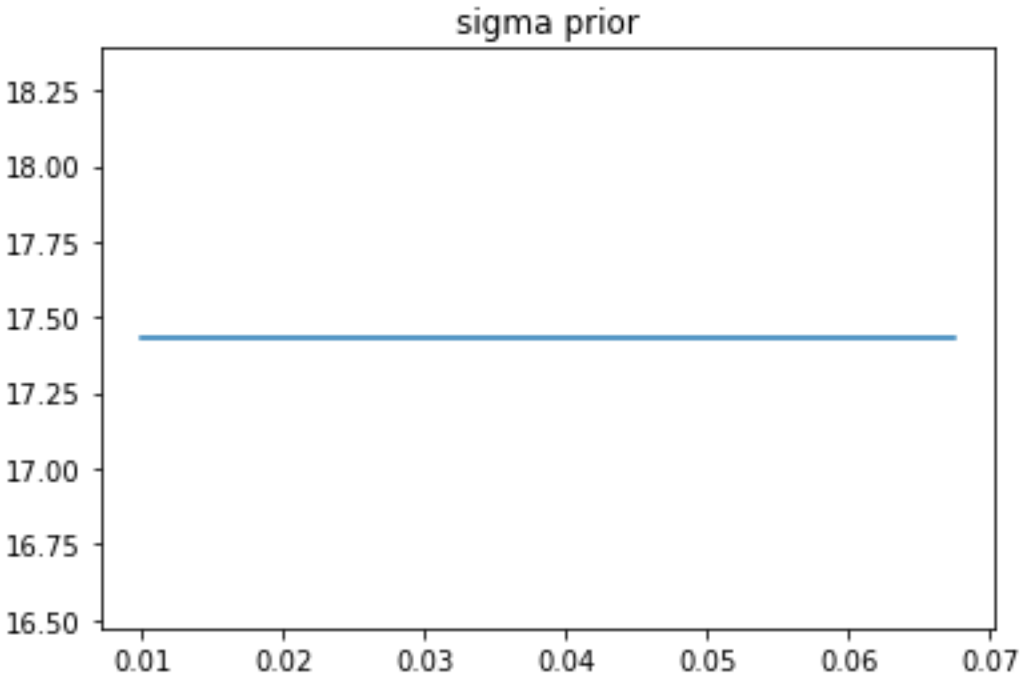

Figure 3

Step 3.1:

The project data comes in the form of (actuals, percent completes). We need to convert these into velocities. To do this, we use Equation 3 where the actual is the elapsed duration or money spent.

$$Velocity\ observation = \frac{\Delta(percent\ complete)}{\Delta(actual)}$$

Equation 3

Step 3.2

Since the and are independent, we can apply the version of Bayes' theorem in Equation 4.

$$P(\mu, \sigma | oberserved\ velocity) = \frac{P(obersved\ velocity|\mu, \sigma)P(\mu)\mathrm{P}(\sigma)}{P(Observed\ Velocity)}$$

Equation 4

Step 3.2 Build a 2-D likelihood grid by using the numerator of the Bayes' theorem ratio

We start this calculation once we have the first observation from the staged data. In our case, *P(μ)* is its prior PDF, and *P(σ)* is its prior, as shown in Figure 2 and Figure 3. To apply , to learn the updated *P(μ)* and P(σ), we use the discrete grid method (Kruschke 2014) to Equation 4 to compute the updated PDFs of μ and σ.

Discretize the support of the σ and μ priors into *equal* intervals, ($\mu_1,\ldots,\mu_n$) and (σ_1, ...), and build a 2D array $L_{i,j}$ by inserting the pairs ($\mu_{,i}$, $\sigma_{,j}$) into the numerator of Equation 4. We can ignore the denominator because that term is accounted for when we normalize the margin in step 3.3. For example, if we are using a normal likelihood model, we use Equation 5 to build the likelihood grid.

$$L_{i,j} = \frac{1}{\sqrt{2\pi{\sigma_j}^2}} e^{-\frac{1}{2}\left(\frac{obs-\mu_i}{\sigma_j}\right)^2} P(\mu_i)P(\sigma_j)$$

Equation 5

Step 3.3 Compute the normalized marginals of the grid to get PMFs of the location and dispersion parameters.

Sum the rows and columns of the $L_{i,j}$ grid to get a discretized version of the likelihoods of the parameters:

$$l(\mu_i) = \sum_j L_{i,j}\,, \qquad l(\sigma_j) = \sum_i L_{i,j}$$

Step 3.4 Interpolate the PMFs to get the PDFs of the location and dispersion PDFs

To get the posterior PDFs of the parameters, interpolate the pairs (μ_i, $l(\mu_i)$) and (σ_j, $l(\sigma_i)$) into functions and normalize each of the functions so that their integrals are equal to 1. These normalized functions are the updated PDFs of the parameters.

Step 3. 5 Save these to be priors for the next observation

The PDFs found in Step 3.4 serve as the new priors for updating them with the next observation. This iterative method is called Bayesian refinement.

Step 4: Use a Monte Carlo simulation with the PDFs and the model to construct the velocity PDF.

Once we have the PDFs of the parameters, we can find the velocity at each iteration by using the parameters in the PDF of the velocity. This is done using Monte Carlo simulation:

- Make arrays of say 10,000 random samples of the μ PDF and the σ PDF
- For each pair (μ_i, σ_i) of the μ and σ samples, substitute them into Equation 5, and take a few random samples (r_i,) of that PDF.
- Build a histogram of the (r_n) array, normalize it, and interpolate it to get the new velocity PDF.

Step 5: Use the velocity PDF and Monte Carlo simulation to divide the remaining percentage to be completed by the velocity to calculate the PDF of the current duration or expenditure needed to complete the effort.

With the PDF of the velocity to find the time to complete, we simply need to divide the remaining percentage of work by the velocity.

$$time\ to\ complete = \frac{Remaining\ Effort}{Velocity}$$

Equation 6

Since the velocity is described by a PDF, we again use a Monte Carlo simulation to generate the PDF of the time to complete. This is performed at each refinement for reporting purposes.

Project-Wide Certainty: Rolling Up Task Estimations

The true value for a project manager lies in extending task duration and cost PDFs to the entire project. This involves "rolling up" the probabilistic durations of all individual subtasks to summary tasks and the entire project, while respecting their dependencies, to generate a single, comprehensive PDF for the overall project completion time, as well as those at the summary task level.

Cost Estimations.

In many, if not most, cases, the costs of the tasks are independent. In that case, one can roll up the task costs by simply adding the uncertain variable using a Monte Carlo simulation. If not, one would need to build a dependency model, such as a Bayesian network.

Durations Estimations

A project is more than the sum of its parts. It is a complex network of interconnected tasks. Simply adding the expected completion times of all subtasks would be a critical error, as it ignores the fundamental role of task dependencies. A project file, feature, and epic plan define the logical sequence of work. One task must finish before another can begin (a

Finish-to-Start dependency), or two tasks might need to finish simultaneously (a Finish-to-Finish dependency). This web of dependencies creates multiple paths through the project, with the longest path—known as the critical path—determining the project's minimum duration. Because the duration of each task is not a fixed number but a probability distribution, the critical path itself is not constant; it can vary from one scenario to another. The challenge, therefore, is to combine these individual task PDFs in a way that respects the project's network logic to generate a meaningful project-level forecast.

In this case, to account for these dependencies, the project management tool would need to include a Monte Carlo simulator that models running the entire project, sampling the times to complete at the task level to get a sample of project completion times, and then using them to create an empirical distribution of the project completion. At this writing the Intaver Institute's Risky Project includes both the velocity learning and schedule rollup features.

Reporting

There are two fundamental questions:

1. "How likely is it that I will deliver on time and/or on budget?"
2. "How late and/or over budget am I likely to be?"

To demonstrate visualizations that can be used to answer these questions, we consider the visualizations shown when using the prior velocity from Figure 1 and the observations in Table 2.

Period	% Complete
1.00	7.5%
2.00	15.0%
3.00	20.0%
4.00	25.0%
5.00	40.0%
6.00	60.0%
7.00	70.0%
8.00	80.0%
9.00	90.0%
10.00	95.0%
11.00	99.0%

Table 2

Recall from Figure 2 that to stay on track, the project should be about 8.2% complete per period. Note that the project is falling behind through period 4, but is catching up in period

5 and is running slightly ahead for the remaining periods. Visualizing the impact of this history on the certainty of meeting the plan is shown in the status charts shown below.

Status Charts

Consider Figure 4, which displays the answers to the two questions. Each shows the PDF of the periods-to-complete PDF after three observations. On the left is the risk of missing the target, in this case, about 87%. On the right is shown the same PDF with vertical lines at various percentile points. There is a 20% chance that it will arrive no later than 13 periods, just missing the target, and there is an 80% chance that it will be no later than about 17 periods.

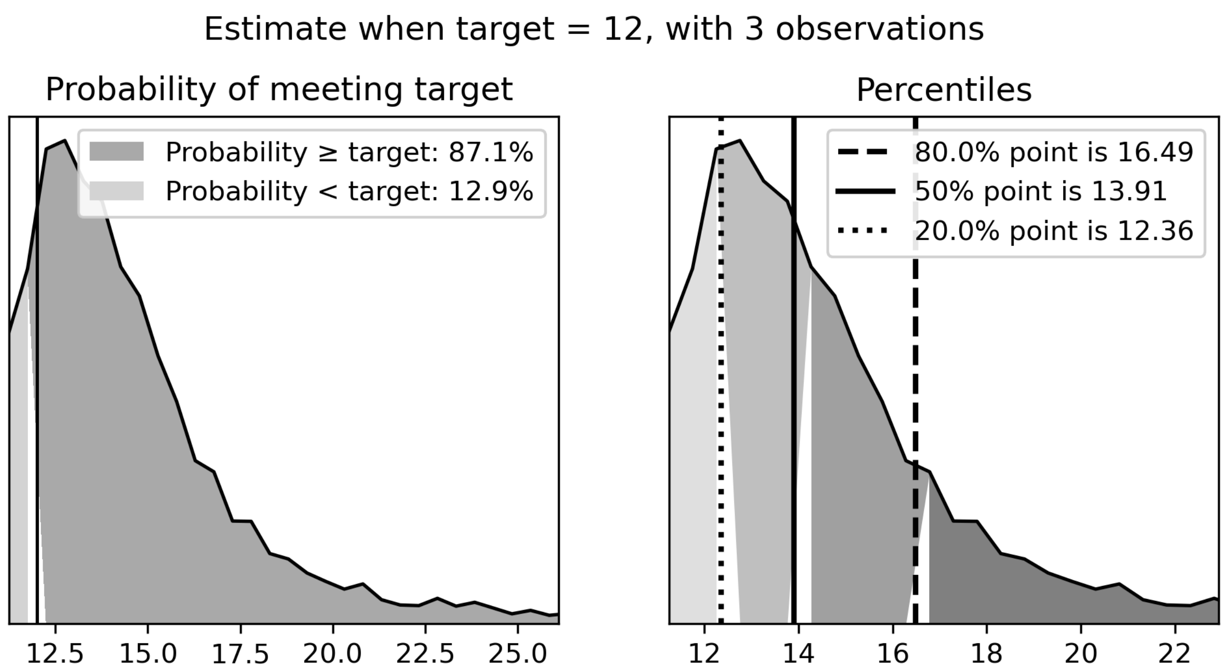

Figure 4

Figure 5 shows how the situation has improved after 6 observations: At that point it is 63% likely the project will meet its target. Much of the uncertainty has been worked off.

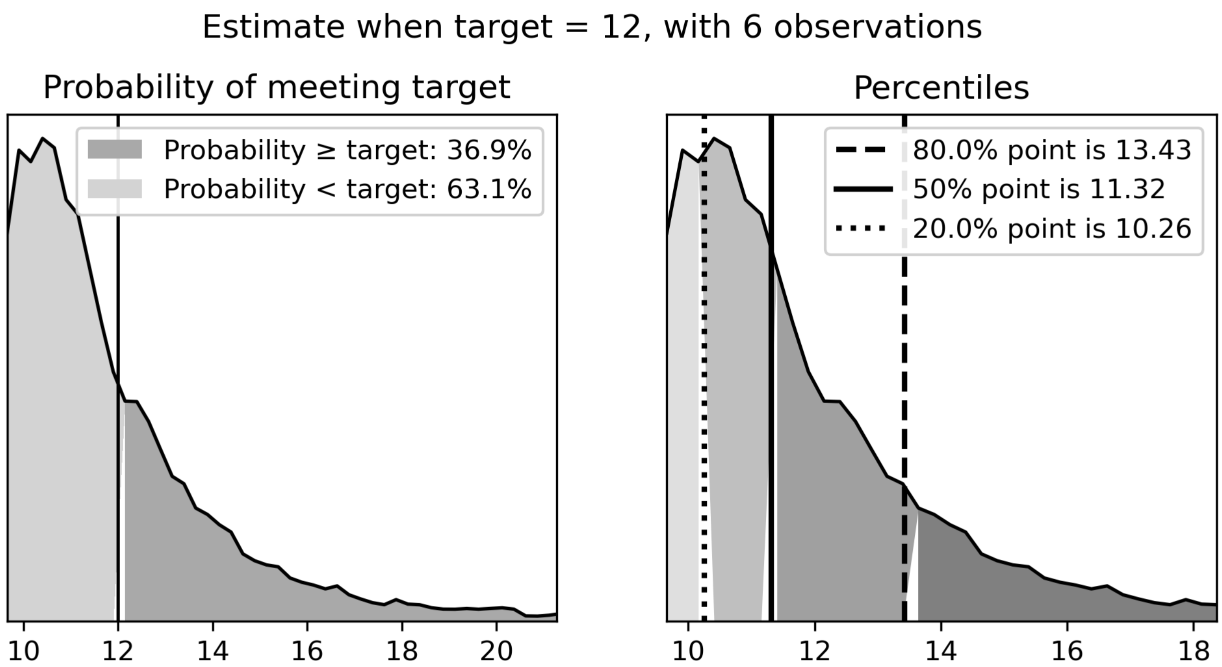

Figure 5

Figure 6 shows the two views of the time-to-complete PDF at 9 observations. At this point, there is scant probability of missing the target. In fact, it is over 80% likely that it will be completed one period early.

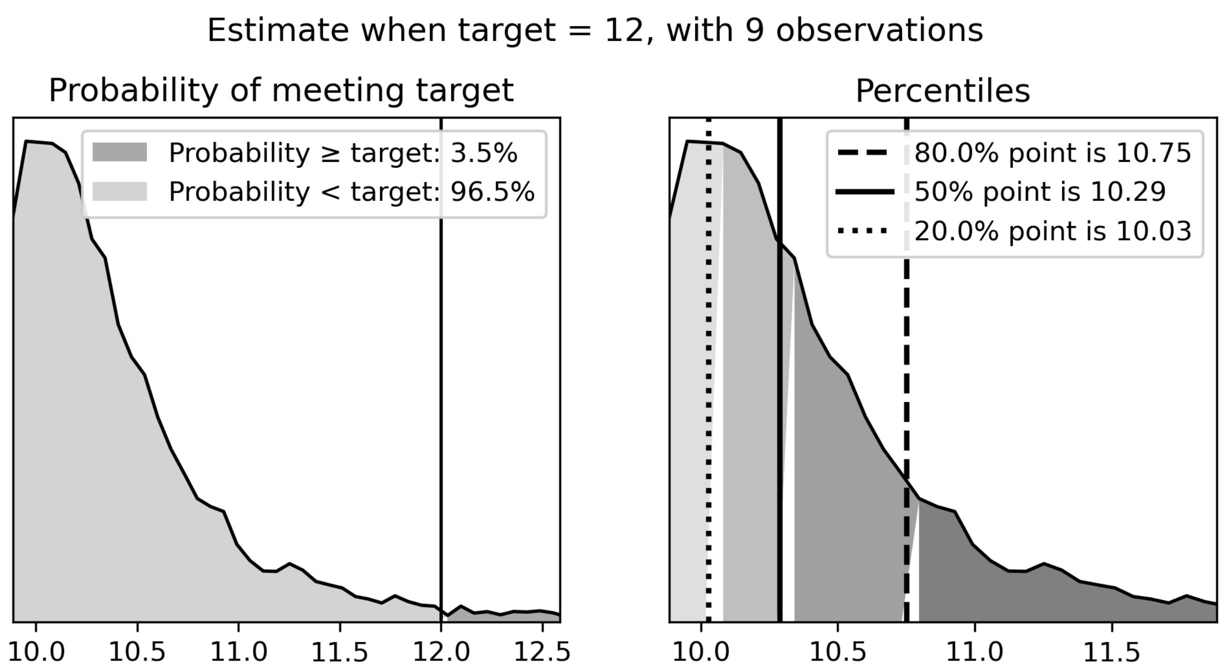

Figure 6

Trend Charts

The status charts may be too technical for executive communications. A clearer visualization can be found in Figure 7 and Figure 8. The first summarizes the results of all 12 of the left status charts. The second summarizes the right-hand charts.

The dashed work-off plan reflects that the initial uncertainty of completing on time was 40%, as shown in Figure 1. The line indicates the reduction of uncertainty if it follows a linear path. The filled-in graph shows the uncertainty of meeting the target at each reporting period. Above the line indicates too much residual risk and suggests the team is not on track to meet the target. The area below the line shows the team is on track to be on time. In this example, the team was at risk at period 4 but took action to recover at that time, with a steep downward slope to get fully on track by period 8.

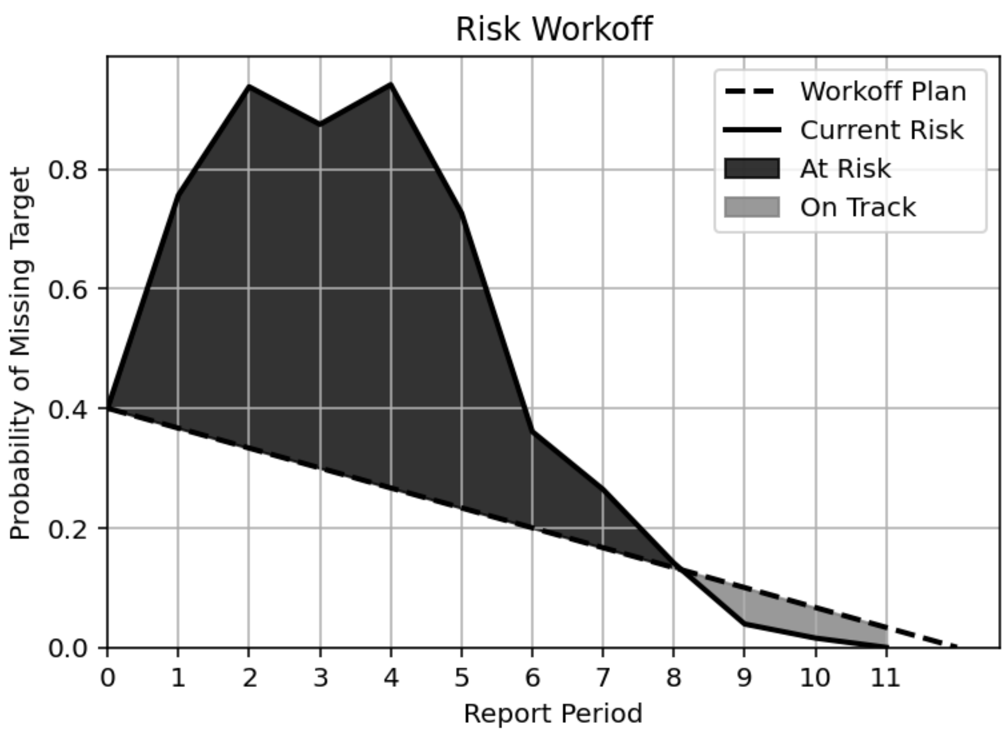

Figure 7

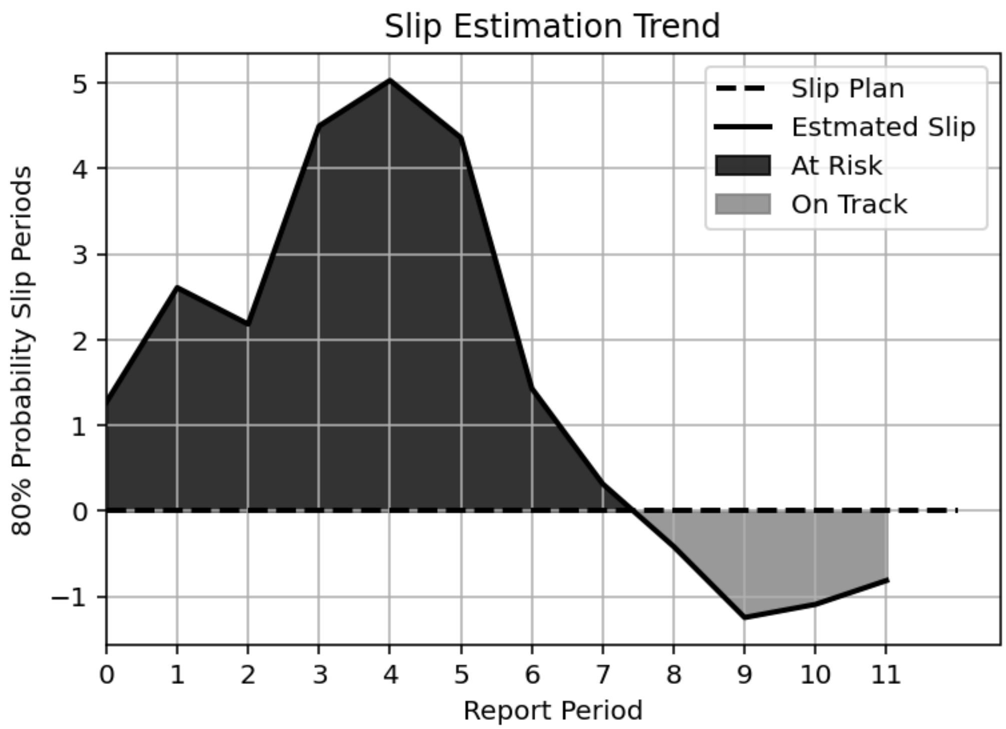

Figure 8

In Conclusion

The task, summaries, and project completions are invaluable tools for project managers and stakeholders. It moves the conversation away from a single, misleadingly precise end date and towards a more realistic discussion about risk and probability. With these distributions, a project manager can confidently answer such critical questions as:

- What is the probability of completing by our contractual deadline?"
- What is the probability of meeting our project Budget?"
- "What is the date we are 90% certain we will be finished by?"
- “Which activities are contributing the most to the overall schedule, and so need management attention? This is found by viewing the PDFs of the summary tasks

References

Fenton, N., and M. Neil. 2019. *Risk Assessment and Decision Analysis with Bayesian Networks*. Second. CRC Press.

Kruschke, J. 2014. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS*. And Stan 2nd. Academic Press.

Mun, J. 2003. *Applied Risk Analysis*. John Wiley & Sons.

Appendix 3 Learning Completion Probabilities