Chapter 6
Uncertainty of Time Predictions

There is a degree of uncertainty in all time predictions. We would not even use the term *predict* if there were no uncertainty in our statements about the usage of time for future tasks. A realistic view of this uncertainty is essential, as illustrated in the following real-life case.

A client was hiring a consultancy company to develop a large and costly IT system. The client suggested a risk-sharing contractual model designed so that the cost of work beyond the predicted time usage (used to calculate the target price) would be split 50–50 between the consulting company and the client. The consulting company, however, was so confident about the accuracy of its time prediction that it suggested an alternative risk sharing model. In its model, which eventually was chosen for the project, the client covered more of the cost if there was a time overrun of less than 30%. The consultancy company would, on the other hand, not get paid at all for work that exceeded the 30% time overrun. Apparently, the consulting company believed that the actual time usage exceeding the predicted time usage by more than 30% was extremely unlikely. What happened? The exact figure was not made public, but the consultancy company suffered a large financial loss that could have been avoided if it had accepted the initial suggestion to split all additional costs 50–50. One might think that this was mainly a negative outcome for the consultancy company. After all, the client did not have to pay anything for the additional work. However, it is very difficult to collaborate with contractors who do not get paid for work done. The consequence was a decrease in quality and endless discussions about whether a feature of the software was specified in the contract or should be considered a change order, leading to extra payment. In such cases, the client often ends up paying much of the overrun, receives a product of low quality and low usefulness, and spends costly time in discussions and perhaps even in court to settle disagreements related to payments—all of this due to underestimating the uncertainty of time predictions.

It will probably not come as a great surprise that people generally trust the accuracy of their time predictions to a greater extent than they should; that is, people tend to be overconfident more often than underconfident regarding the uncertainty of their time predictions [1]. This tendency was illustrated in a study in which students were given a software programming assignment [2]. Before completing the assignment,

T. Halkjelsvik and M. Jørgensen, *Time Predictions*, Simula SpringerBriefs on Computing 5, https://doi.org/10.1007/978-3-319-74953-2_6

they were instructed to predict the upper limit (maximum) of time usage, a value they were instructed to be 99% sure not to exceed, and the lower limit (minimum), a value corresponding to the amount of time they were 99% sure of surpassing. This minimum–maximum interval is a 98% prediction interval, that is, an interval that should be 98% likely to include the actual time usage. If the students' confidence levels were realistic, one should expect the actual time usage to be outside the intervals for only 2% of the students. After completing the programming task, however, as many as 43% of the students failed to include the actual time usage in their prediction intervals.

Predicting in groups may reduce but does not remove overconfidence. We once asked software professionals, first individually and then in teams, to predict the time usage and provide a 90% prediction interval for a software development project [3]. The prediction intervals given individually were the most overconfident, with only 10% of the actual time usages inside the interval (the intervals should have covered as much as 90%). The team discussion-based prediction intervals were better, including 40% of the actual time usage but still indicating overconfidence. It appeared as team members with conservative time predictions and wider individual prediction intervals had a greater influence on the teams' predictions than the most optimistic and overconfident members did.[1] There is, of course, no guarantee that this will happen in other contexts. Group work may sometimes increase willingness to take risks, as well as probably the overconfidence in time predictions.

Take home message: Expect people, even when working in groups, to be overconfident in the accuracy of their time predictions (i.e., they give too narrow time prediction intervals).

6.1 Why Are We Overconfident?

Why are people overconfident in the accuracy of their time predictions? One explanation is that most people have no way of knowing whether they are 50%, 90%, or 99% confident. We are simply not equipped with good skills, or intuition, in understanding and assessing confidence levels expressed as probabilities.[2]

A demonstration of our limited ability in assessing confidence levels as probabilities is reported in a study where one group of participants was instructed to give

[1] Note that whether the actual time usage of a single project is inside or outside a prediction interval is not necessarily a good indicator of an individual's ability to produce realistic intervals. A very high actual project time usage far outside the boundaries of the prediction intervals may, for example, be a consequence of extremely bad luck with the project, that is, a rare situation not meant to be included by the prediction interval.

[2] It is astonishing that so many prediction models, time prediction practices, and project management textbooks are based on the assumption that people are equipped with good judgment on, for example, the minimum and maximum time usages connected with 98 or 90% confidence. This is an unfortunate example of the lack of transfer of research results from the domain of psychology to the domain of engineering and project management.

99% prediction intervals, a second group 90% prediction intervals, a third group 75% prediction intervals, and a fourth group 50% prediction intervals. We would expect that the groups with higher confidence levels also gave wider intervals. This was not the case. All groups gave, on average, about the same time prediction intervals. Other studies have shown similar results. People seem to provide what they consider to be a reasonable low time prediction (the minimum value) and a reasonable high time prediction (the maximum value) more or less irrespective of the confidence level requested [4].

Given the above results, in many contexts, it may be more correct to talk about the *ignorance* of confidence levels in time predictions rather than *overconfidence* [5]. If ignorance of confidence levels is common, we will find that the higher the confidence level you ask for, the more overconfident people will appear to be. For typical software development work, for example, a possible rule of thumb based on various published and unpublished studies is that people's perceptions of minimum and maximum time usage frequently correspond to a hit rate (proportion of actual values included in the predicted minimum–maximum interval) of 50–70% [6]. This means that most people will be overconfident when asked to give a 90% confidence interval, but not when asked for a 50% confidence interval.[3]

Another factor contributing to too narrow, apparently overconfident time prediction intervals is the wish to provide useful information and to be seen as competent (see also our discussion on this topic in Sect. 3.1). A software professional we once interviewed reported that '[wide prediction intervals] will be interpreted as a total lack of competence and has no informative value to the project manager. I'd rather have fewer actual values inside the minimum–maximum interval than providing meaningless, wide effort prediction intervals'. This desire to provide informative intervals seems to be reinforced by managers. In one study, software managers received information about the prediction intervals of two software developers. One of the developers was described as giving wide 90% confidence time prediction intervals and attaining a hit rate of 80%. The other developer gave more narrow 90% prediction intervals of time usage and attained a hit rate of only 60%. Despite the higher degree of realism of the wide interval, most managers preferred the narrow interval and believed that the developer who provided this was more skilled and had greater knowledge of the task. Most managers even believed that the developer with the narrower interval had more knowledge about the task's uncertainty [3]. Furthermore, a study on environmental change predictions suggested that people can perceive narrow confidence interval as signals of *high confidence* (e.g. 90% sure), whereas, in reality, narrow intervals are less likely to include the actual outcome than wide intervals are and should be associated with lower confidence [7].

Take home message 1: Overconfidence in the accuracy of time predictions is often more correctly described as ignorance of confidence levels expressed as probabilities. When unaided by historical data, people tend to give the same minimum and

[3]Be aware that the given rule of thumb of a 50–70% hit rate does not apply to all sorts of work. Different contexts will have different hit rates.

maximum time usage values for widely different confidence levels, for example, little difference between 90 and 50% prediction intervals.

Take home message 2: Overconfidence may also be caused by a desire to give informative time prediction intervals and to appear competent.

6.2 What Can We Do to Avoid Overconfidence?

Ignorance of confidence levels, the wish to be informative, and the desire to be perceived as competent may, as claimed in the previous section, explain to some extent the typical overconfidence in time prediction accuracy in high-uncertainty situations. The more important question is whether we can do something about it. While it may be hard to avoid overconfidence altogether, there are methods that seem to improve the realism of time prediction intervals.

6.2.1 The Use of Alternative Interval Prediction Formats

Instead of asking for a minimum–maximum time interval corresponding to 90% confidence, we could turn the question around. We can ask for the level of confidence for a given minimum–maximum time usage interval. If you think 100 work hours is a reasonable time prediction for a certain job, then you could use the time usage corresponding to, say, 50–200% of this prediction (in this case the interval between 50 and 200 work hours) and ask for the probability of the actual time usage falling between 50 and 200% of the predicted time. Studies report a remarkable reduction in overconfidence when using this alternative request format for *wide* time usage intervals [6]. When, on the other hand, the time prediction intervals were *narrow*, this request format did not increase realism [8]. For example, software developers believed that it was at least 60% probable that the actual use of time would be within ±10% of the predicted use of time. In reality, only 15%- and not 60%—of the projects fell within the ±10% prediction interval.

A potential advantage of the alternative format is that it eases the use of historical data. In situations with tasks of varied size and complexity, it is difficult to use historical data on time usage to calculate a 90% prediction interval for a new task. It is easier to use the historical data from tasks of various sizes and degrees of complexity to develop a distribution for the time prediction error. The latter can, together with the alternative interval prediction method, be used to find prediction intervals and pX predictions. This method is illustrated below.

Think about one type of work you often do. How often have you ended up spending more than twice (200%) the predicted time on this type of work? Say that this happens about 5% of the time. This means that it does *not* happen about 95% of the time. In other words, given that history is a good predictor of future prediction errors, you can

be 95% confident that you will not exceed twice the work effort of your predicted time usage. This value is then your p95 prediction value. Use the same reasoning to add a lower bound. Look back on previous tasks and assess how often you spent less than 80% of the predicted time (20% underrun) on the task. If you spent less than 80% of your predicted time usage about 5% of the time, your p5 prediction corresponds to 80% of your prediction. By using these two values, the p5 and p95 predictions, as the interval limits, we obtain a (95–5)% = 90% time prediction interval. Given relevant historical data, the actual time usage will be 90% likely to be between 80% (p5) and 200% (p95) of the predicted time. If the predicted time is 100 work hours, the 90% prediction interval is 80–200 work hours.

If you are a manager and ask your employees to use the technique above when predicting work hours, you might end up with a p25 prediction here and a p90 prediction there. If what you really want is, for example, the p50 prediction (which you want to use as your planned time usage) and the p85 prediction (which you want to use as your budgeted time usage), what should you do?

The distribution of potential time usages can be derived from any two pX predictions, given a few assumptions about the underlying distribution of outcomes [9]. As discussed in Sect. 3.4, the distribution of time usage is typically right-skewed with a long tail. We find that a lognormal distribution fits this characteristic well.[4] Based on the assumption of lognormal time usage distributions we have developed a simple tool that helps you derive prediction intervals and any pX prediction based on a time prediction (your best guess of a task's time usage) and two pX values based on past prediction accuracy (i.e. the amounts of overrun and underrun you typically experience with similar tasks).[5]

Example[6]: Assume that you receive a new task and predict it will take about 10 hours. You think back on similar tasks and recognize that, in two out of 10 similar cases, you spent 90% or less of the predicted time (an underrun of 10% or more) and, in eight out of the 10 cases, you spent 150% or less of the predicted time (an overrun of 50% or less). Two pX values associated with your time prediction are easily calculated from this information (see Table 6.1).

The above two pX values can be used to calculate the probability distribution of time usage. This distribution can be displayed as a *density* distribution, as in Fig. 6.1, or as a *cumulative* probability distribution, as in Fig. 6.2. In Fig. 6.1, we see that the most likely time usage is 10.6 hours, that is, a little bit more than the predicted time usage. The model does not care what you think your initial point estimate represents, for instance, whether it is meant to be the most likely, mean, or median time usage or something else. The only thing that matters is that you are reasonably consistent in what you mean with a time prediction and how you measure your prediction error. This means that, even if you think you provided a prediction of the most likely time,

[4] There may, however, be a few cases in which we need distributions that can accommodate thicker tails (more 'black swans') than can be represented with a lognormal distribution. See [10].

[5] See www.simula.no/~magnej/Time_predictions_book.

[6] This example is included in the aforementioned tool (the Excel spreadsheet).

Table 6.1 Find two pX predictions based on past time prediction accuracy

Observation	Corresponding pX	pX value when the prediction is 10 hours
In 2 out of 10 times, we spent 90% or less of the predicted time usage (10% or more underrun)	p20	90% of the predicted time usage (0.9×10 hours = 9 hours)
In 8 out 10 times, we spent 150% or less of the predicted time usage (50% or less overrun)	p80	150% of the predicted time usage (1.5×10 hours = 15 hours)

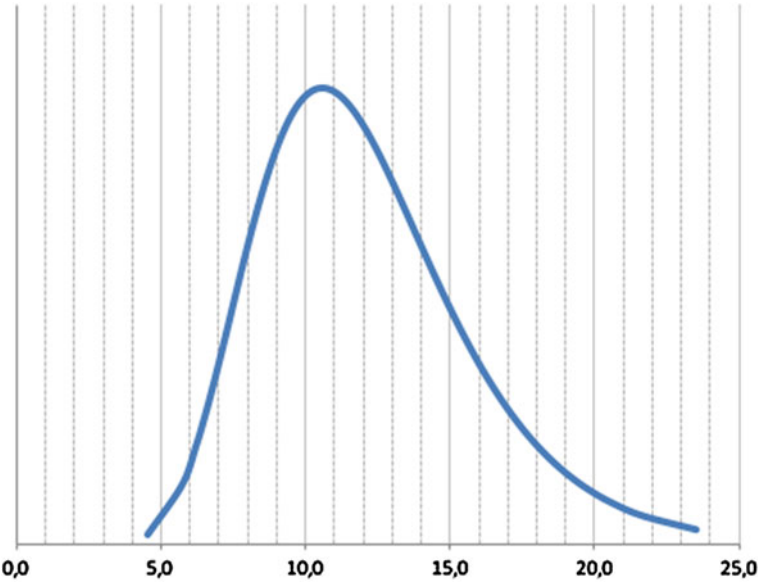

Fig. 6.1 Probability distribution (density) of time usage based on a p20 value of nine hours and a p80 value of 15 hours

the model may give you a different (hopefully better) point estimate of the most likely time based on your input of historical data.

Based on the time prediction and the two pX values from Table 6.1, we can find any pX prediction in the cumulative distribution (Fig. 6.2). The p5 and p95 predictions, for example, can be found to be seven hours and 19 hours, respectively, implying that the 90% confidence interval should be seven to 19 hours. As can also be seen in Fig. 6.2, the two pX predictions we gave as input, the p20 and p80 predictions, still have the values of nine and 15 hours.

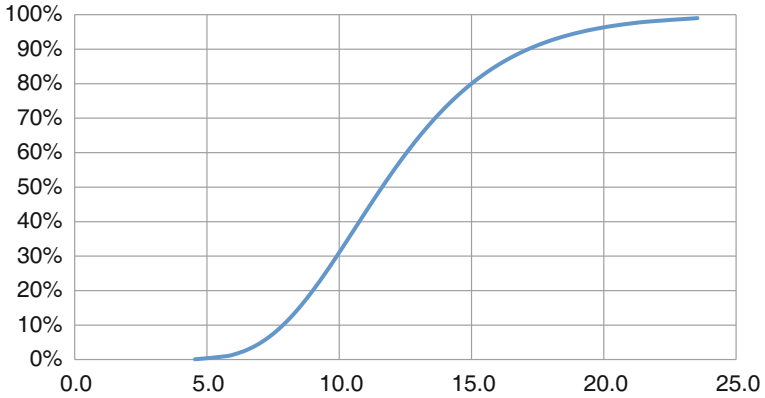

Fig. 6.2 Cumulative (pX) distribution of time usage based on a p20 value of nine hours and a p80 value of 15 hours

Assuming that the input of historical prediction error (or one's judgement of it) is relevant for the task to be predicted and not far off from the actual prediction errors, you can obtain a great deal of other useful information from such a probability distribution. Figure 6.2 suggests, for example, that the task is only about 30% likely to be finished after 10 work hours and 50% likely that it will take 11.6 work hours or less and that you can be pretty sure (around 96%) that the task will be finished within 20 hours.

The prediction intervals and the pX predictions obtained by the method explained here have their limitations. The method is dependent on correctly remembered or correctly recorded time prediction errors of a larger set of previously completed tasks, and it relies on the assumption that the task to be predicted is similar in prediction complexity to the previous tasks. Still, we think that the method will be better at providing realistic time prediction uncertainty information than typical, unaided predictions of minimum–maximum time usage for given confidence levels.

Take home message: We propose a method (with an associated tool) for assessing time prediction uncertainty. The method is based on an alternative way of asking for confidence intervals (that does not require a predefined confidence level) and requests historical information about previous time prediction accuracy. Based on this input, the tool can derive a full distribution of potential outcomes.

6.2.2 Learning from Accuracy Feedback

Repeated feedback about the accuracy of time prediction tasks may increase realism and decrease the level of overconfidence of time prediction intervals. In one study, software professionals predicted the most likely time usage and provided the 90%

time prediction intervals of 30 tasks [11]. The tasks had previously been completed by other developers, so we were able to give the participants information regarding actual time usage after they predicted each task. On average, the 90% confidence intervals given by the participants included only 64% of the actual values for the first 10 tasks, 70% for the next 10 tasks, and 81% for the last 10 tasks. In other words, the realism improved. But even after 20 tasks with accuracy feedback, and even when the participants were not personally involved in the task execution, there was still a bias towards too narrow confidence intervals.

There may be less improvement from feedback when one is trying to learn in a context with greater personal involvement. This was demonstrated in a study of software developers sequentially predicting the time usage, assessing the time prediction uncertainty, and completing five software development tasks [12]. The developers' uncertainty assessments involved a judgement of the probabilities that the actual time usage would fall within 90–110%, 60–150%, and 50–200% of their predicted time usage. Another group, also software developers, received data on the predicted and actual time usages for the first three tasks for one of the developers and then provided uncertainty assessments for the remaining two tasks by the same developer. These developers did not themselves complete any work and were, consequently, less personally involved. The first group, who performed the work, were strongly overconfident, whereas those not involved in the task execution gave more realistic (actually, highly accurate) confidence levels.

Several other studies document that, even after many cycles of predicting time usage, assessing time usage uncertainty, completing tasks, and receiving feedback about actual time usage, people remain overconfident [13]. The lack of learning from experience is especially evident when people are instructed to give high confidence intervals, such as 98% confidence intervals.

Take home message: People tend to remain overconfident in the accuracy of their time predictions even after extensive accuracy feedback, especially when they are personally involved in the task execution.

References

1. Jørgensen M, Gruschke TM (2009) The impact of lessons-learned sessions on effort estimation and uncertainty assessments. IEEE Trans Software Eng 35(3):368–383
2. Connolly T, Dean D (1997) Decomposed versus holistic estimates of effort required for software writing tasks. Manage Sci 43(7):1029–1045
3. Jørgensen M, Teigen KH, Moløkken-Østvold K (2004) Better sure than safe? Over-confidence in judgment based software development effort prediction intervals. J Syst Softw 70(1):79–93
4. Teigen KH, Jørgensen M (2005) When 90% confidence intervals are 50% certain: on the credibility of credible intervals. Appl Cogn Psychol 19(4):455–475
5. Jørgensen M (2014) The ignorance of confidence levels in minimum–maximum software development effort intervals. Lect Notes on Softw Eng 2(4):327–340
6. Jørgensen M (2004) Realism in assessment of effort estimation uncertainty: it matters how you ask. IEEE Trans Softw Eng 30(4):209–217

7. Løhre E, Teigen KH (2017) Probabilities associated with precise and vague forecasts. J Behav Decis Making. Advance online publication. https://doi.org/10.1002/bdm.2021
8. Jørgensen M, Faugli B, Gruschke T (2007) Characteristics of software engineers with optimistic predictions. J Syst Softw 80(9):1472–1482
9. Cook JD (2010) Determining distribution parameters from quantiles. www.johndcook.com/quantiles_parameters.pdf. Accessed May 2017
10. Budzier A (2014) Theorizing outliers: explaining variation in IT project performance. Doctoral dissertation, University of Oxford
11. Jørgensen M, Teigen KH (2002) Uncertainty intervals versus interval uncertainty: an alternative method for eliciting effort prediction intervals in software development projects. In: International Conference on Project Management (ProMAC). Singapore, pp 343–352
12. Gruschke TM, Jørgensen M (2008) The role of outcome feedback in improving the uncertainty assessment of software development effort estimates. ACM Trans Softw Eng Methodol 17(4):1–35
13. Gruschke T, Jørgensen M (2005) Assessing uncertainty of software development effort estimates: the learning from outcome feedback. In: Software metrics, 2005. 11th IEEE international symposium. IEEE, p 4