Measurements are essential in supporting studies on the effectiveness of software engineering tool and techniques as any other engineering discipline. Although the maximum available accuracy of a measure may depend on the measuring instrument, the quality of the measure depends on the integrity of the measuring process. In this poster we present an Empirical Investigation of the applicability of Benford’s Law (Benford, 1938) and Digital Statistics (Nigrine, 1995) in the context of software engineering metrics analysis and process validation.

Consider a list of numbers generated in the course of an investigation into financial statements, stock market values or Lines of Code (LOC) metrics extracted from 100 public domain industrial Java Projects. In a counter-intuitive manner, Benford’s Law states that the values in those lists of numbers, subject to certain exceptions, can be expected to have the number $D_1$ as the first non-zero digit with a probability: $P(D_1) = \log(1 + 1 / D_1)$

This distribution is counter-intuitive for at least two reasons. First it would seem “obvious” that the numbers drawn from a list generated from widely different arbitrary processes would have roughly equally probabilities for the digits 1 and 9 to be first digits. This is not normally the case. If the list of numbers does not have artificial limits, or include invented numbers such as postal codes, then approximately 30% of the numbers will have 1 as their first digit, but only 5% will have 9 as their first digit. Deviations from the expected Benford Distribution indicate the presence of some special characteristic of the data. The second, more theoretically challenging, problem is: What is the underlying property associated with so many widely different processes which generates lists of numbers that follow Benford’s Law?

We have conducted an empirical investigation to determine under what circumstances various software metrics follow Benford’s Law, and whether any special characteristics, or irregularities, in the data can be uncovered if the data are found not to follow the law. The more tricky problem of understanding why the list of metrics might follow Benford’s Law is left to another study.

Lists were from three software metrics extracted from 100 public domain industrial Java Projects. These metrics were Lines of Code (LOC), Fan-Out (FO) and McCabe Cyclomatic Complexity (MCC). Given that a Benford’s Law analysis requires a list of considerable length, the data were divided into two groups. The first group was from projects containing more than 100 files. This was intended as the “control group” and what was expected to follow Benford’s Law if that Law was applicable for the analysis of software engineering metrics. To study the sensitivity of the digital analysis technique to project size, projects with a smaller number of files were compared to the control group.

The empirical results indicate that the first digits of numbers in lists of LOC metrics extracted from the projects closely followed the probabilities predicted by Benford’s Law than an “equal probability of occurrence” suggested by intuitive reasoning. This was shown using both qualitative and quantitative measures. The FO and MCC lists did not follow the standard Benford’s Law as well as did the LOC metrics. This is because the FO and MCC lists contain a significant number of numbers less than 10 and follow a different first digit distribution. Further investigation of the digital analysis technique to project size, projects with a smaller number of files were compared to the control group.

The empirical results indicate that the first digits of numbers in lists of LOC metrics extracted from the projects closely followed the probabilities predicted by Benford’s Law than an “equal probability of occurrence” suggested by intuitive reasoning. This was shown using both qualitative and quantitative measures. The FO and MCC metrics did not follow the standard Benford’s Law as well as did the LOC metrics. This is because the FO and MCC lists contain a significant number of numbers less than 10 and follow a different first digit distribution. Further investigation of the digital analysis technique is necessary to evaluate the applicability of Benford’s Law in the total context of Software Metrics.

Benford, F. *Proceedings of the American Philosophical Society*, (March 1938), 551-571.