On the Distribution of Software Faults

Hongyu Zhang, *Member, IEEE*

Abstract—The Pareto principle is often used to describe how faults in large software systems are distributed over modules. A recent paper by Andersson and Runeson again confirmed the Pareto principle of fault distribution. In this paper, we show that the distribution of software faults can be more precisely described as the Weibull distribution.

Index Terms—Software fault distribution, empirical research, replication.

1 Introduction

In the May 2007 issue of this journal, a paper entitled "A Replicated Quantitative Analysis of Fault Distributions in Complex Software Systems" [1] described a replication of Fenton and Ohlsson's quantitative study [3] of faults in a complex software system. One of the hypotheses it confirmed was the widespread belief that the number of faults in large software systems follows the Pareto principle [4]. We have replicated the studies of [1] and [3], and discovered that the distribution of faults over modules can be better modeled using the Weibull probability distribution function.

The Pareto principle is named after the economist Vilfredo Pareto, who proposed a model to describe the distribution of wealth among individuals. The idea is sometimes expressed more simply as the Pareto principle or the "20-80 rule" which says that 20% of the population owns 80% of the wealth. Formally, the cumulative distribution function (CDF) of the Pareto distribution [2] can be defined as: $P(x) = 1 - \left(\frac{\gamma}{x}\right)^{\beta} \quad (\gamma > 0, \beta > 0)$.

The Weibull distribution, developed by the physicist Waloddi Weibull, is one of the most widely used probability distributions in the reliability engineering discipline [6]. The CDF of the Weibull distribution can be formally defined as: $P(x) = 1 - \exp\left(-\left(\frac{x}{\gamma}\right)^{\beta}\right) \quad (\gamma > 0, \beta > 0)$.

In this correspondence, we show that the Weibull distribution fits the actual data better and therefore, it is more appropriate to describe the distribution of the software faults (including both prerelease and postrelease faults) as the Weibull distribution.

2 The Weibull Distribution of Software Faults Over Modules

In this research, we replicate the studies described in [1] and [3] using the public Eclipse data collected by the University of Saarland [7]. Eclipse is a widely used integrated development platform for creating Java, C++ and web applications. The public Eclipse datasets contain measurement and fault data for Eclipse versions 2.0, 2.1 and 3.0, which are collected from Eclipse's bug databases and version achieves. The original datasets contain data at file and package levels. In this correspondence, we use the package level data to present our findings as the granularity level of "package" is more similar to the level of "modules" used in [1] (i.e., a collection of files). However, we should note that our findings apply to the file level data as well.

TABLE 1
The Eclipse Datasets (package level)

Project	Lines of Code	# modules	# pre-release faults	# post-release faults
Eclipse 2.0	796K	376	4152	2049
Eclipse 2.1	985K	433	2007	1394
Eclipse 3.0	800K	431	3312	2151

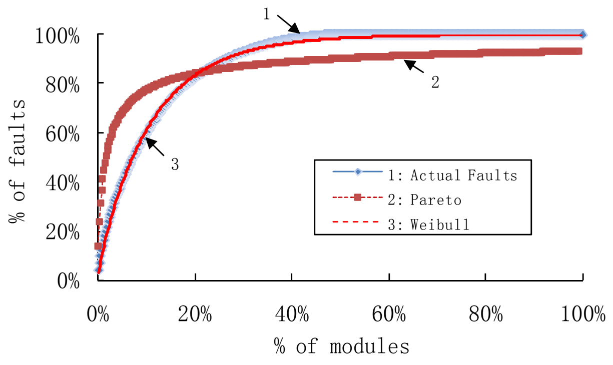

Fig. 1. The distribution of prerelease faults in Eclipse 2.1. It shows the percentage of the accumulated number of faults when the modules are ordered by decreasing number of faults.

For each Eclipse project, we analyze the distribution of its faults across the modules (in this example, the packages). As in the cases in the original studies [1, 3], we also

- *H. Zhang is with the School of Software, Tsinghua University, Beijing 100084, China. E-mail: hongyu@tsinghua.edu.cn.*

Manuscript received (insert date of submission if desired). Please note that all acknowledgments should be placed at the end of the paper, before the bibliography.

find that the distribution is highly skewed - that a small number of modules accounts for most of the faults.

Further analysis shows that the faults follow the Weibull distribution instead of the Pareto distribution. Using statistical packages such as the SPSS, we are able to perform non-linear regression analysis and derive the parameters for each distribution. Fig. 1 shows the fitted curves of Weibull and Pareto distributions for the prerelease faults in Eclipse 2.1. Clearly the Weibull distribution fits the actual data better. To statistically compare the goodness-of-fit of these two distributions, we compute the coefficient of determination (R^2) and the Standard Error of Estimate (S_e) [5]. The R^2 statistic measures the percentage of variations that can be explained by the model. Its value is between 0 and 1, with higher value indicating a better fit. Therefore, R^2 can be seen as an index of the relative goodness-of-fit of a sample regression curve. In Fig. 1, the Weibull distribution has the R^2 value 0.998, meaning that the model accounts for 99.8% of the variations, whereas the R^2 value for the Pareto distribution is only 0.684. S_e is a measure of the absolute prediction error and is computed as $S_e = \sqrt{\frac{\sum (y - y')^2}{n-2}}$, where y and y′ are the actual and predicted values, respectively. The larger S_e indicates the larger prediction error. In Fig. 1, the Weibull distribution has the S_e value 0.01, whereas the Pareto distribution has the S_e value 0.11. Therefore, we conclude that the Weibull distribution is a better fitting distribution.

In the same way, we analyze all projects in the Eclipse datasets and for both prerelease and postrelease faults. The results show that the Weibull distribution can better describe the distribution of faults (Table 2).

TABLE 2
The Weibull Distribution of Eclipse Faults

		Weibull Distribution				Pareto Distribution			
		γ	β	R^2	S_e	γ	β	R^2	S_e
Pre-release faults	Eclipse 2.0	0.109	0.918	0.999	0.01	0.007	0.560	0.700	0.11
	Eclipse 2.1	0.108	0.939	0.998	0.01	0.007	0.544	0.684	0.11
	Eclipse 3.0	0.099	0.889	0.998	0.01	0.006	0.552	0.701	0.10
Post-release faults	Eclipse 2.0	0.125	0.918	0.998	0.01	0.008	0.541	0.701	0.12
	Eclipse 2.1	0.119	0.871	0.994	0.02	0.007	0.519	0.685	0.11
	Eclipse 3.0	0.091	0.733	0.993	0.01	0.006	0.535	0.730	0.09

3 Conclusion

We perform a replicated study of [1] and [3], and find that the Weibull distribution describes the actual fault data well. We suggest using Weibull distribution to precisely model the fault distribution over modules, instead of the commonly used term "Pareto Principle".

Acknowledgment

We thank Carina Andersson, the author of [1], who confirmed our findings using the datasets described in [1]. We gratefully acknowledge her support for this work. We also thank the reviewers for the valuable comments.

References

[1] C. Andersson and P. Runeson, A Replicated Quantitative Analysis of Fault Distributions in Complex Software Systems, *IEEE Trans. Software Eng.*, 33 (5), pp. 273-286, May 2007.

[2] R. Cooper and A. Weekes, *Data, Models, and Statistical Analysis*, Oxford : Philip Allan Pub., 1983.

[3] N. Fenton and N. Ohlsson, Quantitative Analysis of Faults and Failures in a Complex Software System, *IEEE Trans. Software Eng.*, 26 (8), pp. 797-814, August 2000.

[4] J.M. Juran and F.M. Gryna, Jr., *Quality Control Handbook*, fourth ed., McGraw-Hill, 1988.

[5] G. Keller and B.Warrack, *Statistics for management and Economics*, Duxbury, 1999.

[6] R. Ramakumar, *Enginnering Reliability: fundamentals and applications*, Prentice-Hall, 1993.

[7] T. Zimmermann, R. Premraj and A. Zeller, Predicting Defects for Eclipse, *Proc. of the Third International Workshop on Predictor Models in Software Engineering*, Minneapolis, USA, May 2007. The Eclipse data is available at http://www.st.cs.uni-sb.de/softevo/.