Don't trust on strangers – How to assess the quality of a research dataset?

Israel Herraiz
<israel.herraiz@upm.es>
Technical University of Madrid

Daniel Rodríguez
<daniel.rodriguez@uah.es>
University of Alcala, Madrid

Replicability is an issue in the empirical software engineering community. Reusable datasets helps achieve those desirable features. However, what happens when we trust on third-parties datasets? How can we assess the quality of the data? How can we be sure it is not a poisoned candy?

We propose two preliminary techniques:

- Check how duplicates affect (above all, in unbalanced datasets)
- “Bug-normality” and “size-normality” tests, based on verified statistical properties of bugs and software

Preliminary results about the quality of some datasets exctracted from the PROMISE repository. Note the high duplicity in the data, and how the defect rate is affected by duplicity.

DS	Data	% Dup	% Dup (NB)	% Defect	% Defect No dup
CM1	498	11	11	10	11
JM1	10885	18	19	19	23
KC1	2109	**43**	**43**	**15**	**26**
KC2	522	**28**	**29**	20	28
KC3	458	**29**	**29**	9	13
MC1	9466	**79**	**79**	0.7	2
MC2	161	1	1	32	33
MW1	403	5	6	8	8
PC1	1109	14	15	7	7
PC2	5589	**75**	**75**	0.4	1.6
PC3	1563	8	8	10	11
PC4	1458	8	8	12	13
PC5	17186	**89**	**89**	**3**	**26**

“Bug-normality” and “Size-normality” tests

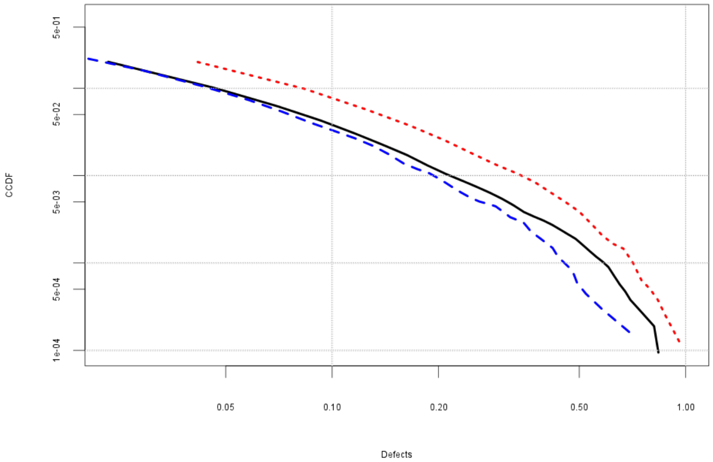
Does it look like verified defects data? Does it look like other software systems? Is the distribution the same as other data?

Kolmogorov-Smirnov: if two samples come from the same statistical distribution, the vertical distance between the two Distribution Functions is below a certain threshold.

Normalize data, plot distribution function, compare against verified defect data (Kolmogorov-Smirnov test)

This is similar to the Alberg diagrams, proposed by Ohlsson and Alberg. TSE v.22, n. 12. (1996)

Comparison of pre-release defects in Eclipse. Thre three CCDF are very close and within a certain threshold given by the KS test.



Assumptions and further research

Do all software defects data have the same distribution? Do all software systems show the same statistical distributions?

Other tests (i.e. Cramer-Von Mises) to compare the distribution of two samples?

What are those distributions? How can we use to asses the quality of research datasets?