On the Statistical Distribution of Object-Oriented System Properties

Israel Herraiz, Polytechnical University of Madrid, Spain
Daniel Rodriguez, University of Alcala, Spain
Rachel Harrison, Oxford Brookes University, UK
Outline

Introduction
- Power laws
- Lognormal
- Double Pareto

Experimental Work
- Distribution of the OO Metrics

Conclusions and future work
Power laws are everywhere

Termed the signature of human activity

Power laws found in different fields

- Including software

- And of course, Object-Oriented Systems

- However, some CK metrics are not a power law, but lognormal
The lognormal – power law controversy

Some properties that appear to be a power law can sometimes be described by a lognormal

• And the other way around

It has happened in every field where power laws have been identified

OO Metrics - Power laws or lognormal?

• Mitzenmacher stated that in many of the controversies between power and lognormal distributions, there is a missing third party:
  • The Double Pareto Distribution
  • A mixture between power law and lognormal distributions
Object-Oriented Metrics

OO Metrics - Power laws or lognormal?

Mitzenmacher stated that in many of the controversies between power and lognormal distributions, there is a missing third party:

• The Double Pareto Distribution
• A mixture between power law and lognormal distributions
Can OO metrics be described using a Double Pareto distribution?

We measured 69 Java projects of the Qualitas Corpus repository

- Data for each project includes source code, JAR files (compiled versions of the source code), documentation and meta-data about the project, with some basic metrics and a classification of the project.
- http://qualitascorpus.com/

Used all Chidamber & Kemerer metrics

- ckjm tool — Chidamber and Kemerer Java Metrics
- http://www.spinellis.gr/sw/ckjm/

Fitted power law and lognormal distributions to different ranges of the data

- Used the procedure recommended by Clauset et al.
- http://tuvalu.santafe.edu/~aaronc/powerlaws/

Procedure is easily replicable with all necessary scripts and data available from the first author Web page:

- http://mat.caminos.upm.es/~iht/wetsom2012/
Some results

Typical pattern of a Double Pareto distribution

- Lognormal
- Powerlaw
The case of WMC

This metric has been found to be both good described by lognormal and power law

But lognormal alone is not a perfect fit
Some metrics are very well fitted by Power laws

- Not a trace of Double Pareto
In summary

We have three groups of metrics

- Double Pareto
  - WMC, CBO, RFC
- Power laws
  - NOC, LCOM
- Not fitted by either double Pareto or Power laws
  - DIT - This metric is generally composed of low values for this type of analysis
Is it really a Double Pareto?

One reviewers concern about our fitting procedure

- We think the shape of the graph is very clear and typical of a double Pareto
- In all the cases, a lognormal alone deviated for very large values of the metric
- It only affects the value of the parameters of the distribution, not the fact that the distribution describes the data
- We have now developed a new fitting procedure
  - We will obtain some new results soon using this new procedure
So what?

Why is interesting to know the distribution of the CK metrics suite?

Effort prediction in agile methods

- Traditionally based on size metrics (LOC)
- Some modern techniques use CK metrics for effort prediction

Use CK metrics as predictors but discarded: NOC and LCOM

- Interestingly, the power law metrics. Why interestingly?
  - They only use those metrics that are subject to the “controversy”
  - That is, those which are in the border between power law, lognormal and double Pareto
Effort prediction models that use double Pareto metrics but assume that the metric is lognormal will systematically underestimate the value of effort.
Summary and Future Work

Some metrics are power law, some others lognormal or double Pareto

Some controversial metrics are being used for effort prediction models

Knowing the distribution can help devise new models and improve existing ones

But if the metric is assumed to be lognormal and it is not, the model will systematically underestimate

New release of the **Qualitas Corpus**
- Repeat with newer releases of the same case studies
- Repeat with more case studies

Fitting procedure
- We have developed a new fitting procedure
- Compare with our current approach
- Set once for all the double Pareto question

Results useful for effort prediction model builders
- That is, both practitioners and researchers