Modulated power-law behaviour in Stirling's approximation

Les Hatton
CISM, University of Kingston*

September 15, 2009

Abstract

In a previous paper, an intimate link between power-law distribution of component sizes and defect growth in maturing software systems, independently of their representation language, was revealed by the use of a variational method built on statistical mechanical arguments. This argument used Stirling's approximation which limits its relevance to larger component sizes. Although power-law behaviour was indeed confirmed at these larger component sizes for numerous systems, the current paper seeks to broaden Stirling's approximation to see if it corresponds with the departures from power-law observed for smaller component sizes.

`Keywords: Defects, Macroscopic system behaviour, Component size distribution, Power-law`

Following [1], we consider a software system of N lines of code split up into M components of n_i, i = 1, .., M lines each. Then the number of ways of organising this system is given by:-

$$W = \frac{N!}{n_1! n_2! .. n_M!} \tag{1}$$

where

$$N = \sum_{i=1}^{M} n_i \tag{2}$$

*L.Hatton@kingston.ac.uk, lesh@oakcomp.co.uk

Such a model includes all possible orderings of different commands of course but as will be seen later, the introduction of a constraint related to observed defect distributions in deployed systems in which commands must follow some logical order related to their specifications will restrict the solution set to those observed in the real world. This is closely analogous to the physical constraints applied when using this combinatorial argument with a physical system such as a gas, [4].

Taking the natural log of equation (1), Stirling's approximation can then be used. However, this assumes both N and n_i are large. Normally of course, N is large but to describe smaller components we have to relax the constraint on n_i a little. We can do this by using the fact that the error term in Stirling's approximation is of the order of 1/2 log(n) and so the conventional form of Stirling's approximation

$$logn! \simeq nlogn - n \tag{3}$$

can be replaced by a better approximation for smaller n

$$logn! \simeq (n + \frac{1}{2})logn - n \tag{4}$$

This is equivalent to averaging the step function corresponding to log 1 + log 2 + .. + log n = log n! and log 1 + log 2 + .. + log (n-1) = log (n-1)! = log n! - log n. These two step functions straddle the curve y = log x whose integral is n log n - n.

Now taking the log of equation (1), gives

$$logW = NlogN - \sum_{i=1}^{M} n_i log(n_i) \tag{5}$$

In physical systems, the most likely combination is found by maximising equation (5) subject to additional constraints. The first is obviously equation (2). The second constraint is usually of the form

$$U = \sum_{i=1}^{M} n_i \varepsilon_i \tag{6}$$

where ε_i is the energy of the ith component, determined by quantum mechanical arguments and U is therefore the total energy. In other words, they are externally imposed. Using the method of Lagrangian multipliers and the

form of Stirling's approximation in equation (4), the following variational is to be maximised

$$logW = (N+\frac{1}{2})logN - \sum_{i=1}^{M}(n_i+\frac{1}{2})log(n_i) + \gamma\{N - \sum_{i=1}^{M} n_i\} + \beta\{U - \sum_{i=1}^{M} n_i\varepsilon_i\} \tag{7}$$

where γ and β are the multipliers. Setting $\delta(logW) = 0$ leads to

$$0 = -\sum_{i=1}^{M} \delta n_i\{log(n_i) + \alpha + \beta\varepsilon_i + \frac{1}{2n_i}\} \tag{8}$$

where $\alpha = 1 + \gamma$. This must be true for all variations δn_i and so

$$log(n_i) = -\alpha - \beta\varepsilon_i - \frac{1}{2n_i} \tag{9}$$

Using equation 2 to replace α, this can be manipulated into the most likely, i.e. the equilibrium distribution

$$n_i = \frac{Ne^{-\beta\varepsilon_i}e^{-\frac{1}{2n_i}}}{\sum_{i=1}^{M} e^{-\beta\varepsilon_i}} \tag{10}$$

Following [3] and defining $p_i = \frac{n_i}{N}$ and referring to equation (6), p_i can be interpreted as the probability that a component gets a share of U equal to ε_i. Manipulating equation (10) then yields

$$p_i = \frac{e^{-\beta\varepsilon_i}e^{-\frac{1}{2n_i}}}{\sum_{i=1}^{M} e^{-\beta\varepsilon_i}} \tag{11}$$

For $\varepsilon_i \sim logn_i$ as described in [1], this can be written as

$$p_i = \frac{n_i^{-\beta a_1}e^{-\frac{1}{2n_i}}}{Q'(\beta)} \tag{12}$$

where the denominator

$$Q'(\beta) = \sum_{i=1}^{M} n_i^{-\beta a_1}e^{-\frac{1}{2n_i}} \tag{13}$$

is not a function of n_i.

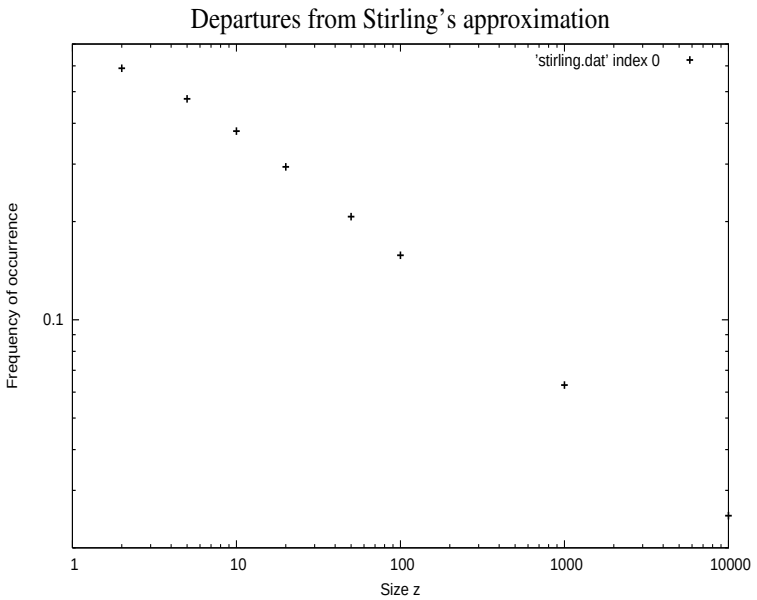

Figure 1: Showing the effect of the exponential modulating term on power-law behaviour for small component sizes

It can be seen therefore that improving Stirling's approximation for smaller n_i leads to an exponential modulating term on power-law behaviour whose effects appear as a slow roll-off for smaller n_i as can be seen in Figure 1. There is some evidence for this slow roll-off in size distributions for smaller components of real systems as shown by Figure 2 which shows function size distributions in various medium-sized systems written in C. It should be noted however that Mitzenmacher [2] also gives a model for this which leads to lognormal bodies with power-law tails.

References

[1] L. Hatton. Power-law distributions of component sizes in general software systems. *IEEE Transactions on Software Engineering*, July/August 2009.

[2] Michael Mitzenmacher. Dynamic models for file sizes and double pareto distributions. *Internet Mathematics*, 1(3):305–333, 2003.

[3] P.K. Rawlings, D. Reguera, and H. Reiss. Entropic basis of the pareto law. *Physica A*, 343:643–652, July 2004.

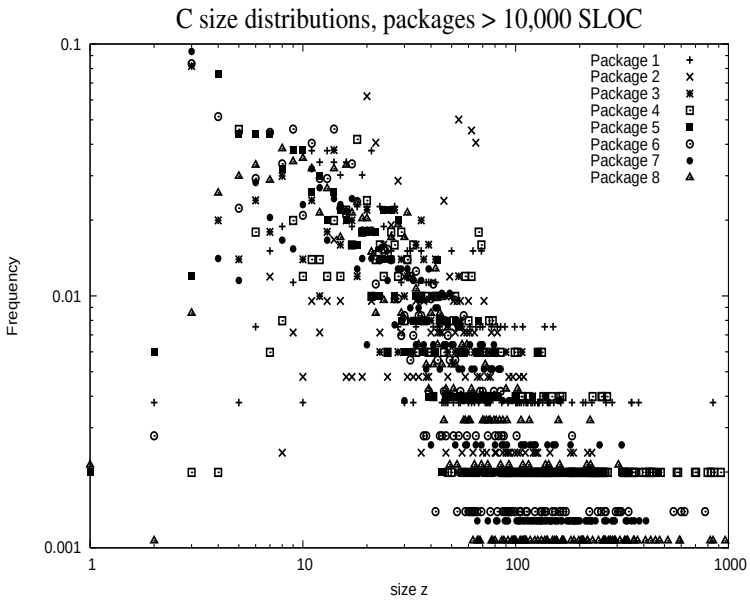

Figure 2: The distribution of function sizes in C systems in the range 10,000-43,700 lines ranked in order of size

[4] A. Sommerfeld. *Thermodynamics and Statistical Mechanics.* Academic Press, 1956.