Homework 3

David Östling
dostl@kth.se
DA2210

December 27, 2021

Two research studies

1a) The result of the programs clearly points towards that commenting the code gives an advantage. However this does not necessarily have to be the case as the students do not have to share the same knowledge skills. If group A contains a student with better programming skills (who might have also faced a similar problem before or similar) and B does not then the commenting does not really matter too much in the end. So no, we can not draw the conclusion but the probability that commenting code helps has increased.

1b) In order to improve the study we could (for example) make sure that all students in each group have the exact same knowledge and education. This can be done by grouping them by grades or similar, as long as all people involved in the study are equally good at programming and that either all people have faced a similar problem before or that no one has (this can be done by conducting a short survey before the test).

2a) This yet again comes down to how good the students from each group are at the programming languages and programming in general. If it was given that they have the same exact skills (which is not given in the information) then it could conclude that in this particular scenario (the problem that each group was faced with) Haskell is faster than java. However even if that would have been the case here it does not necessarily mean that Haskell is always faster. Hence we can not draw an exact conclusion here.

2b) In order to improve this study we would first of all *(as in the previous task 1b)* need to make sure that all students in each group share the same skills. The study could also benefit from creating more programs (covering more scenarios) to test on (the more the

better) so that we can understand which of the programming languages are faster overall and not just for one task.

Zipf's law

a) Chosen dataset: "frequencies of family names, e.g., in Sweden"

b) The log to log graph

Here is the plotted diagram, it was achieved by following the provided instructions [1].

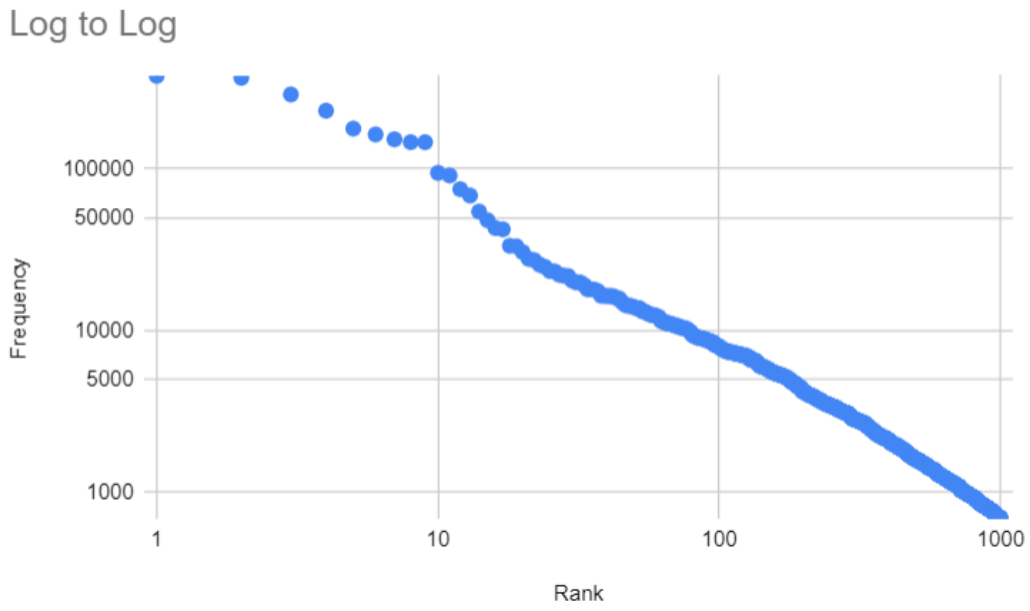

Figure 1: The log to log graph

c) Now we simply perform a linear regression and use the following formula: $Frequency = Cn^a$ which yields the following values; [C = 855883, a = -1,02]. Then we simply plot this in the format below and we receive a corresponding line which we then print within the log to log diagram as follows:

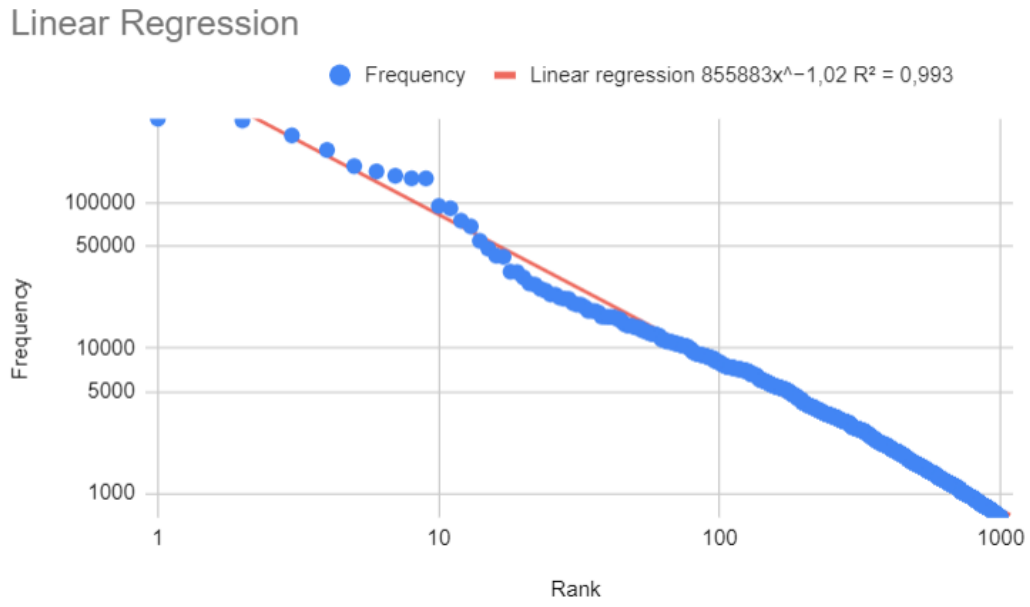

Figure 2: A plot of the linear regression *(where C = 855883 and a = -1,02)*

I believe that my results agree with the law [1].

d) While I could not find a research study containing my particular dataset I found another one containing chinese names. I resorted to studying the graph and it did show some similarities to mine [3]. While it is unknown if there is a correlation between the frequencies- and ranks of Swedish and Chinese surnames it did show that my graph probably is on the right track (as it was not completely off).

e) Zipf's law is considered an empirical law [1] meaning that it is a scientific law that can be proven or disproven by using experiments/observations [2]. Thus it is not a hypothesis, conjecture nor a theorem.

References

[1] Wikipedia (2021) "*Zipf's law*" Available: https://en.wikipedia.org/wiki/Zipf%27s_law

[2] Ruby, Jane. "*Origins of Scientific "Law"*." Laws of Nature. De Gruyter, 2011. 289-315.

[3] Statistical distribution of Chinese names - Scientific Figure on ResearchGate. Available: https://www.researchgate.net/figure/colour-online-The-Zipf-plot-of-surname-distributions-of-our-sample-circles-and-the_fig1_258306874 [accessed 3 Oct, 2021]