*Evolution – Lehman – Stats Reference*

\*see HP 4 in the table, should be just combine the two stated in one?

\*I can get comments - https://api.github.com/repos/jquery/jquery/issues/2916/comments

Project selection -> set selection criteria (Have this already chosen) -> 500 initial projects (50 per language) -> random choice of 100? 200?

Data Structure -> organize into weeks -> remove first year (outliers, look only at mature projects evolution etc.)

|  |  |  |
| --- | --- | --- |
| *Hypothesis* | *Measure/Metric* | *Test* |
| If a projects commits increase, stargazers will also increase and vice versa. (law 1 and 6) | Each projects commits and stargazers are organized into weekly counts/total per week.  P1-C = {20,22,11,40}  P1-S = {12,22,26,32}  … | 1) Get the mean of each projects stargazers and commits.  P1-C-Mean = 30  P1-S-Mean = 27  ….  2) Perform correlation (pearson/spearman) on the commits/stargazers vectors of means. |
| Total lines of code increases as software evolves (law 2) | LOC | 1) get growth rate between weekly interval of LOC  2) get mean growth rate for the project  3) find out the % of projects have a positive mean growth rate |
| Issues, additions and deletions over time will be normally distributed (law 3) | Issues, additions, deletions counts as weekly intervals | 1) Apply shapiro wilks to each of the 3 vectors for the project.  2) get the p number  3) state how many projects for each of the three metrics are in the 0.05 threshold of significance |
| As software evolves changes to lines of code should not fluctuate (law 4)  Total lines of code in the software increases incrementally at a constant rate (law 5) | LOC – organized into weeks as a vector with the amount of LOC in that interval | 1) get the variance for each project LOC  2) then do a mean of the variances |
| Project issues will increase as code churn decreases (law 6) | Count of issues and LOC as weekly intervals | Same process as the first HP, but looking for no correlation? (ie not positive) |
| As the number of issue comments increases the number of issues should decrease (law 8) | Count of issues, count of issues comments organized into weeks. | Same as the above but difference metrics/measures. |

Other stuff

If we try to get issues in bulk it only returns open issues by default – therefore it is easier to traverse one at a time that way I can ignore pull requests as well.