## **Correlation between Code Coverage and Post-Release Defect Density**

***We have started our analysis with a hypothesis that, the projects with low code coverage contains more bugs***. Code coverage gives us an idea of the thoroughness of testing by providing information about the amount of code that is tested. So that, increase in code coverage is likely to lead to a decrease in post-release bugs. For computing the Correlation between the statement coverage and branch coverage with post release defect density, we have calculated the spearmen correlation coefficient for the statement coverage and the number of bugs as well as for the branch coverage and the number of bugs as Shown in the Table-1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Project Name** | **Statement Coverage** | **Branch Coverage** | **Number of Bugs** |
| Apache Commons Collections | 51% | 82% | 5 |
| Apache Commons Configuration | 89% | 83% | 11 |
| Apache Commons DbUtils | 64% | 77% | 6 |

Table-1: Showing Statement Coverage, Branch Coverage, and Number of Bugs.

The spearman coefficient for statement coverage and post release defect density is 1 i.e. strong correlation is observed from all projects. The spearman coefficient for branch coverage and post release defect density is 0.5 i.e. medium correlation is observed.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Projects** | **Versions** | **Statement Coverage** | **Branch Coverage** | **Number of bugs** | **Spearman Coefficient**  **(M1 & M6)** | **Spearman Coefficient**  **(M2 & M6)** |
| **Apache Commons DbUtils** | 1.1  1.2  1.4  1.6  1.7 | 56%  63%  79%  57%  64% | 51%  53%  66%  64%  77% | 4  4  2  4  6 | *rs* = -0.22361 | *rs* = 0.22361 |
| **Apache Commons Configuration** | 1.8  2.0  2.2  2.4  2.6 | 77%  68%  86%  79%  85% | 70%  63%  56%  77%  81% | 16  11  7  2  4 | *rs* = -0.5 | *rs* = -0.5 |
| **Apache Commons Collections** | 2.0  3.0  3.2  4.1  4.4 | 65%  76%  86%  69%  51% | 49%  68%  81%  77%  82% | 4  15  55  16  5 | *rs* = 0.8 | *rs* = 0.4 |

For collections, both correlation values are positive. For configurations, both correlation values are negative.

***Calculating Spearman Correlation***

Spearman’s Rank correlation coefficient is one of the most-prominent technique which can be used to find out the strength and correlation between two variables. We have calculated the spearman correlation coefficient using the following method.

***Method used to calculate the Spearman correlation***

* Create a table from your data and get the ordered pairs of two variables.
* Rank the two data sets. Ranking is achieved by giving the ranking '1' to the biggest number in a column, '2' to the second biggest value and so on. The smallest value in the column will get the lowest ranking. This should be done for both sets of measurements or the variables used to find the correlation for.
* Tied scores are given the mean (average) rank.
* Find the difference in the ranks (d).
* Square the differences (d²) To remove negative values and then sum them
* Calculate the coefficient (***Rs***) using the formula mentioned below.

When written in mathematical notation the Spearman Rank formula looks like this:



Here,

ρ= Spearman rank correlation

di= the difference between the ranks of corresponding variables

n= number of observations

We have used the formula of spearmen correlation coefficient by considering the statement coverage values as X- values and the number of bugs as Y- values.