

Christian Lungescu (20079725)

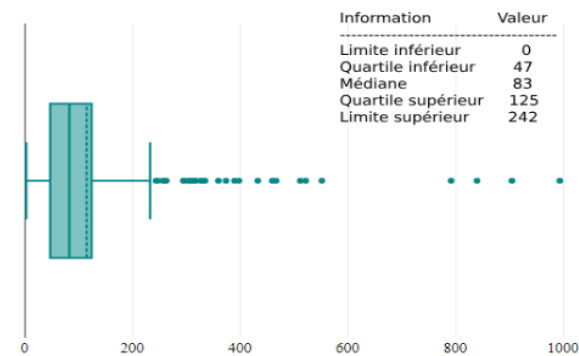
Malik Abada (20173066)

TP3 du cours IFT-3913

Tache 1 : Visualisation des métriques de l'échantillon en créant des boîtes à moustaches et description des distributions.

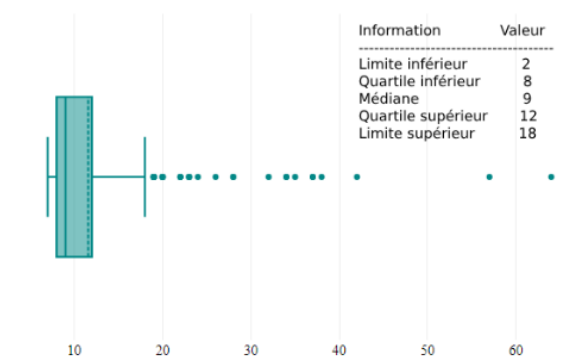
Notes pour la tache 1 : Nos valeurs de « boîte à moustache » ont été calculé à l'aide d'un programme qu'on a écrit pour ce but (disponible dans le projet GitHub et la remise). Tandis que le diagramme a été généré par un le programme en ligne «Datatab» en fonction du data set.

Métrique "TLOC"



Description de base : En analysant les données de l'image de gauche, on remarque ici que 50% des classes "test" ont un nombre de lignes (TLOC) entre 47 et 125 ("range" entre le quartile inférieur et supérieur). On remarque aussi que ~ 50% des classes "test" ont un nombre de ligne inférieur à 83 et que l'autre ~50% a un nombre de ligne supérieur à 83.

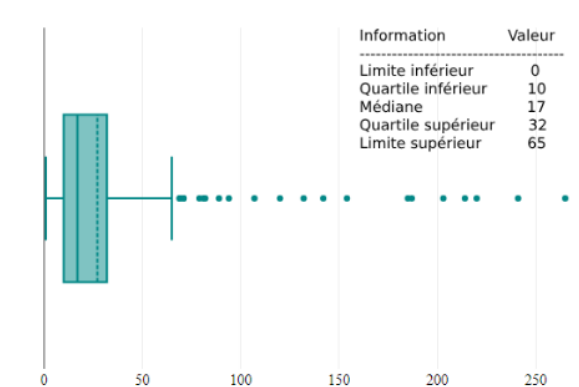
Métrique "WMC"



Description de base : En analysant les données de l'image de gauche, on remarque ici que 50% des classes "test" ont un WMC entre 8 et 12 ("range" entre le quartile inférieur et supérieur). On remarque aussi que ~ 50% des classes "test" ont un WMC inférieur à 9 et que l'autre ~50% ont un WMC supérieur à 9.

Remarques : On peut aussi voir que qu'il y au moins 25% des classes « test » qui ont un WMC de 8 ou 9 , vu que le « quartile inférieur » est égal a 8 est la mediane est de 9. Ce qui est interessant vu que ce que ces 2 valeurs (8 et 9) representent le quart des informations dans notre ensemble de données WMC tandis qu'il ya plus 60 valeurs possibles dans l'ensemble d'après le schema de gauche(16 valeurs possibles si l'on rejette les données à l'exterieur des limites inferieur et supérieur).

Métrique "TASSERT"



Description de base : En analysant les données de l'image de gauche, on remarque ici que 50% des classes "test" ont un nombre de assertions JUnit entre 10 et 32 ("range" entre le quartile inférieur et supérieur). On remarque aussi que ~50% des classes "test" ont un nombre de assertions JUnit inférieur à 17 et que l'autre ~50% à un nombre d'assertions supérieur à 17.

Tache 2 : Étude des corrélations entre TLOC et TASSERT, WMC et TASSERT ; calcul des droites de régression et des coefficients de corrélation qui ont du sens.

Tout d'abord, il faut déterminer si les ensembles de chaque métrique est normalement distribué. Pour cela on va se servir du "P value" obtenu par différents tests de normalité. Comme on peut le voir dans le tableau ci-dessous, tous les "P values" trouvées avec nos tests sont < 0.05 donc on peut rejeter l'hypothèse nulle et assumer que les « sets » ne sont pas normalement distribués. Pour calculer ces "P values", nous avons juste copié chacun des sets dans le calculateur en ligne « <https://www.statkingdom.com/> » (section fit test).

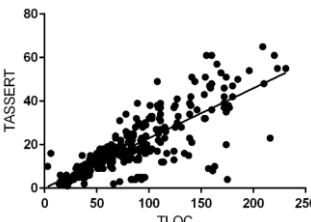
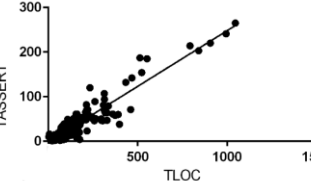
Métrique \ Nom du test	"P value" avec Shapiro-Wilk	"P value" avec Kolmogorov
TLOC	~ 0	$2.539e-7$
WMC	~ 0	~ 0
TASSERT	~ 0	~ 0

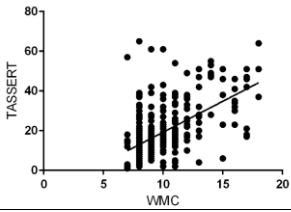
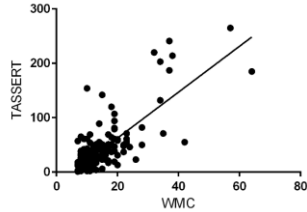
Maintenant qu'on a déterminé que les « sets » ne sont probablement pas distribués normalement, on sait que qu'il faut calculer le coefficient de corrélation de rang (p) avec la « méthode de Spearman » entre les métriques TLOC et TASSERT, WMC et TASSERT. Voici un tableau contenant les coefficients de corrélation avec le filtrage de points extrêmes qui se trouvent à l'extérieur des limites inférieures et supérieures calculées dans les boîtes à moustache de la tache 1, et sans filtrage de points extrêmes.

Données avec filtrage ? Pair de métriques étudiées	Données avec filtrage	Données sans filtrage
TLOC ET TASSERT	$p = 0.7761$ Fichier de données utilisées = « tlocTassertFilteredData.csv »	$p = 0.8346$ Fichier de données utilisées = « tlocTassertUnfilteredData.csv »
WMC ET TASSERT	$p = 0.4930$ Fichier de données utilisées = « wmcTassertFilteredData.csv »	$p = 0.61$ Fichier de données utilisées = « wmcTassertUnfilteredData.csv »

D'après les coefficients de corrélations (p) du tableau ci-dessus, on constate que tous paires de métriques étudiées (filtrées ou pas) sont corrélés positivement. On voit aussi qu'il y a une plus forte corrélation lorsqu'on ne filtre pas les données. Cela pourrait être expliqué par le fait que les données supplémentaires raffinent la droite de régression.

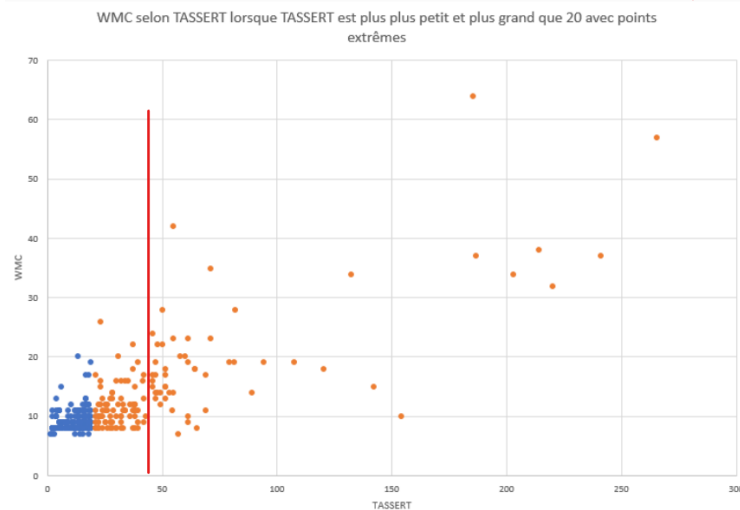
Ci-dessous on retrouve les équations des droites de régressions avec leur représentation graphique :

Informations Pair de métriques étudiées	Représentation graphique	Équation
TLOC et TASSERT avec filtrage de données		$y = 0.2294x + 0.06986$
TLOC et TASSERT sans filtrage de données		$y = 0.2498x - 1.568$

Wmc et TASSERT avec filtrage de données		$y = 3.116x - 11.94$
WMC et TASSERT sans filtrage de données		$y = 4.221x - 21.70$

T3. Nous voulons évaluer l'hypothèse : « les classes qui contiennent plus de 20 assertions sont plus complexes que celles qui contiennent moins de 20 assertions ». Décrivez la conception d'une quasi-expérience qui vous permettra de le faire. Ensuite, évaluez l'hypothèse, discutez les résultats et décrivez vos conclusions.

<p>Choix d'étude : Étude de cas</p> <p>Raison du choix d'étude : On a choisi de faire une étude de cas car on veut étudier un cas spécifique, le cas spécifique est le niveau de complexité d'une classe lorsqu'elle a plus ou moins 20 assertions (en utilisant la métrique du nombre d'assertions JUnit (TASSERT)).</p>
<p>Hypothèse : L'hypothèse est déjà donnée dans la l'énoncé de la tâche, la voici : « les classes qui contiennent plus de 20 assertions sont plus complexes que celles qui contiennent moins de 20 assertions ».</p>
<p>Variables : La variable indépendante est TASSERT, le nombre d'assertions JUnit d'une classe et la variable dépendante est WMC, la somme des complexités cyclomatiques des méthodes d'une classe.</p>
<p>Interprétation et généralisation des résultats: Tout d'abord, on a séparé les données des classes en 2 groupes. Le 1^{er} groupe contient les données dont le nombre d'assertions est inférieur à 20 (TASSERT < 20) et le 2^{ème} groupe contient les données ou le nombre d'assertions est supérieur à 20 (TASSERT > 20). Dans le graphe ci-dessous, on remarque qu'il est difficile de voir une corrélation positive pour un TASSERT < ~45 (les données à gauche de la ligne rouge). Par contre, pour les données dont les ou le TASSERT > ~45 (données à droite de la ligne rouge), on remarque une certaine corrélation positive.</p>



Cependant, ce que l'on cherche c'est de savoir si **en général** les classes ayant plus de 20 assertions sont plus complexes que celles qui ont moins de 20 assertions. On peut utiliser le test de Mann-Whitney U pour comparer la probabilité d'obtenir une valeur plus élevée du groupe 1 avec la probabilité d'obtenir une valeur plus élevée du groupe 2.

Supposons qu'on a les hypothèses suivantes :

- **Hypothèse nulle (H0)** : La complexité cyclomatique médiane des classes avec moins de 20 assertions est égale à la complexité cyclomatique médiane des classes avec 20 assertions ou plus.
- **Hypothèse alternative (H1)** : La complexité cyclomatique médiane des classes avec moins de 20 assertions est différente de la complexité cyclomatique médiane des classes avec 20 assertions ou plus.

On obtient une p-value de $3.271557e-25$ lorsqu'on applique le Mann-Whitney U test avec les valeurs des deux groupes WMC. On voit que le p-value est inférieur 0.05 donc on peut rejeter l'hypothèse nulle et assumer qu'il y a une différence de complexité entre le 1er groupe de données et le 2iemes groupe. Sachant que la relation entre TASSERT et WMC est corrélé positivement, on peut déduire que la complexité du groupe deuxième groupe (>20 assertions) est plus élevée que celle du premier groupe.

Cette quasi-expérience est restreinte à un seul code: jfreechart. Ce projet est populaire, le code-source est large, il y a beaucoup de collaborateurs, on pourrait donc supposer que jfreechart est assez représentatif des projets Java en général. Notre résultat est donc possiblement généralisable.

Validité :

Q : Les résultats de l'étude peuvent-ils être généralisables à des cas qui ne sont pas inclus dans l'étude ?

R : La mesure de nombre d'assertion/tassert (variable indépendante) peut ne pas modéliser adéquatement la complexité (WMC) parce qu'il y a d'autres variables qui peut influencer la valeur du WMC, parce exemple le nombre de boucles, if-statement, etc.

Q : Les changements dans la variable dépendante peuvent-elles être raisonnablement attribués aux changements dans les variables indépendantes ?

R : Puisque c'est une quasi-expérience, il est possible que les résultats soient influencés par d'autres facteurs. Par exemple, on pourrait imaginer que le style de codage des développeurs de jfreechart fasse en sorte que leurs classes plus complexes utilisent un nombre plus élevé d'assertions.

Q : Les résultats de l'étude peuvent-ils être généralisables à des cas qui ne sont pas inclus dans l'étude ?

R : Cette quasi-expérience est restreinte à un seul code: jfreechart. Ce projet est populaire, le code-source est large, il y a beaucoup de collaborateurs, on pourrait donc supposer que jfreechart est assez représentatif des projets Java en général. Notre résultat est donc possiblement généralisable.