Noms : Stefan Roman, Simo Hakim
Titre : TP3 Étude de corrélations

1 Conception d'une quasi-expérience

1.1 Énoncé de l'hypothèse

L'hypothèse est la suivante : « Les classes qui contiennent plus de 20 assertions sont plus complexes que celles contiennent moins de 20 assertions. » Cette hypothèse est une explication anticipée que nous cherchons à exposer la véracité. Pour ce faire, la variable étudier est le nombre d'assertions dans une classe et le comportement étudier est la complexité de cette classe. Ainsi, la collecte de données minimale comportera des valeurs mesurant le nombre d'assertions dans la classe et la complexité de la classe ou des méthodes de cette classe.

1.2 Choix d'étude

Le choix d'étude est un aspect crucial à discuter dans cette étude, notamment pour la formation d'un groupe de contrôle. Il est impossible d'en créer un pour des raisons pratiques, l'une étant la manipulation des tests pour les complexifier systématiquement pour créer un groupe de tests qui ont une complexité artificielle. Or un groupe de contrôle artificiel n'est pas désiré.

1.3 Définition des variables

Les variables à étudier porterons sur le nombre de tests dans une classe et la complexité de cette classe. Ces variables sont les suivantes : TLOC, WMC et TASSERT. Respectivement ces variables sont définies comme suit : le nombre de lignes de code exécutables dans la classe, la complexité pondérée de la classe et le nombre d'assertions dans la classe.

Pour valider ou invalider la conclusion les données seront aussi sujettes à des groupements en fonction de la valeur de TASSERT de chacune de classes afin de produire des valeurs appropriées pour l'hypothèse. Les trois groupes visés seront le groupe global contenant tous les classes, le groupement ayant que les classes contenantes au moins 20 assertions et le groupe de classes contenant moins de 20 assertions.

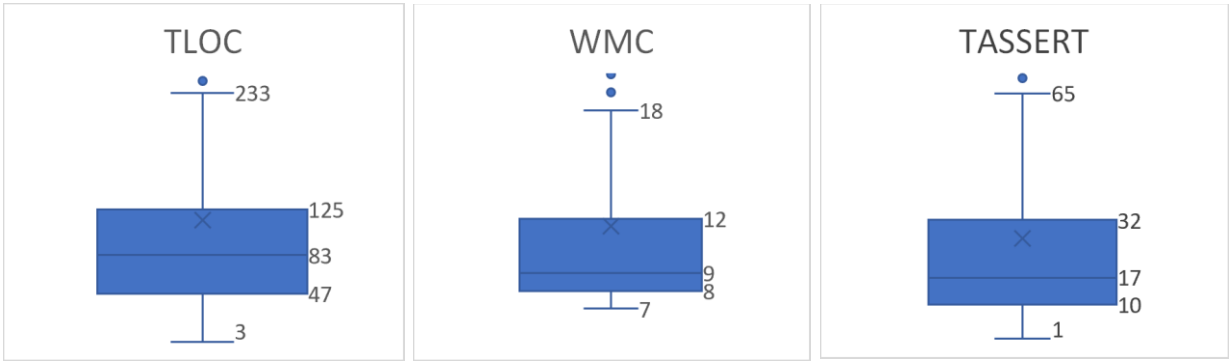
Il y a une relation principale à évaluer qui sera la relation du nombre d'assertions dans une classe et la complexité de cette classe. Cela pourra être accomplie avec un coefficient de corrélation pour soulever l'existence probable d'une corrélation entre ces deux valeurs. Ensuite, l'usage d'une droite de régression permettra de vérifier la nature de cette relation et son coefficient de proportionnalité.

2 Évaluation de l'hypothèse

2.1 Visualisation des données collectées

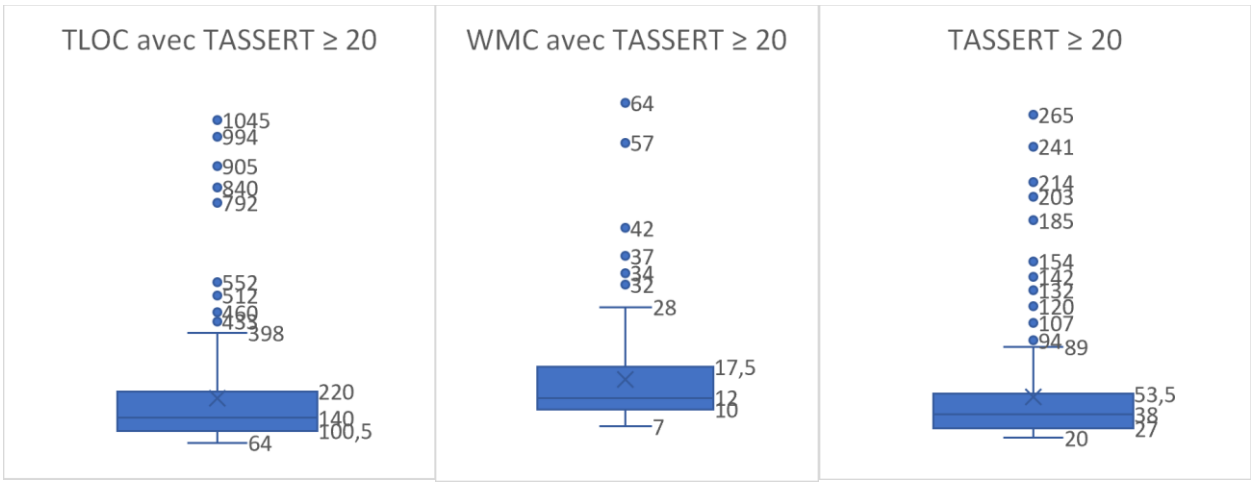
Une observation à faire en regardant ces valeurs est la médiane de WMC. En effet la métrique évaluant la complexité des méthodes de la classe nous permet d'établir un lien entre les classes contenant plus de 20 assertions et leur complexité.

2.1.1 Distribution de toutes les données collectées



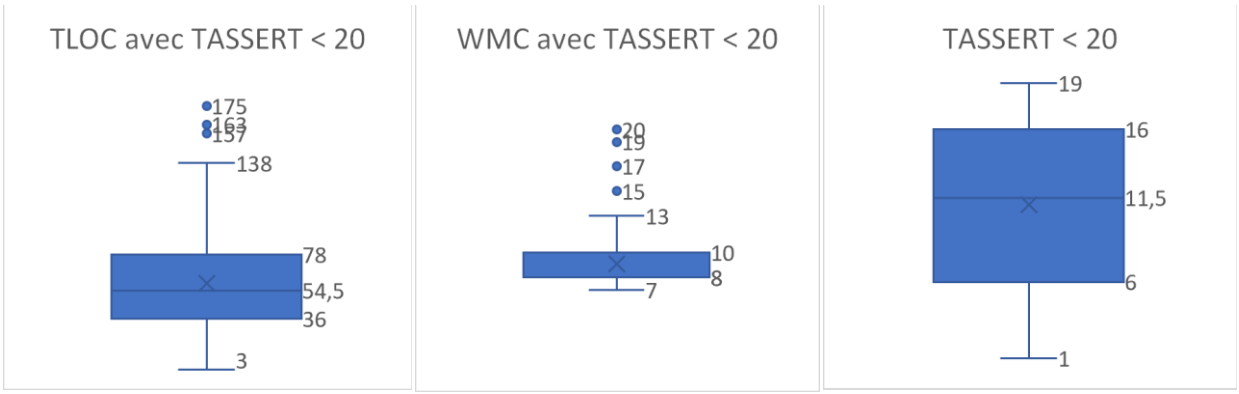
2.1.2 Distributions avec une quantité d'assertions supérieure ou égale à 20

Ces graphiques de distributions démontrent qu'il y a une grande partie de la population étudiée qui se retrouve dans les valeurs extrêmes. Cela indique que la distribution n'est pas exactement normale mais elle peut quand même être presque normale dépendamment du nombre de valeur dans chaque quartile.



2.1.3 Distributions avec une quantité d'assertions inférieure à 20

Dans ce cas il est observable que la distribution est mieux répartie et tends moins vers des valeurs extrêmes. Cependant il y a des points d'intérêts dans ces distributions. Notamment WMC qui se retrouve avec une médiane de 8 qui agit comme la limite du premier quartile.



2.2 Étude de corrélations

2.2.1 Mesure d'association

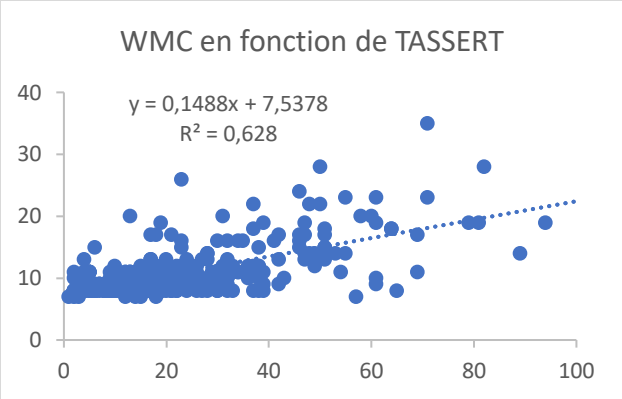
Utilisant le coefficient de corrélation de Pearson (r) nous permet d'établir l'existence d'une relation entre les deux variables principalement étudiées : TASSERT ET WMC. Dans l'ensemble complet étudié, cette valeur est de ~0.79 alors que dans l'ensemble contenant que les classes ayant 20 ou plus assertions elle est de ~0.75 et dans le restant c'est une valeur de ~0.22. Cela étant dit les classes avec TASSERT < 20 il est claire qu'il n'y a pas de corrélation selon cette mesure. Or il est visible que dans l'ensemble avec TASSERT > 20, cette valeur étant de ~0.75 nous permet d'assumer une corrélation linéaire positive satisfaisante entre TASSERT et WMC.

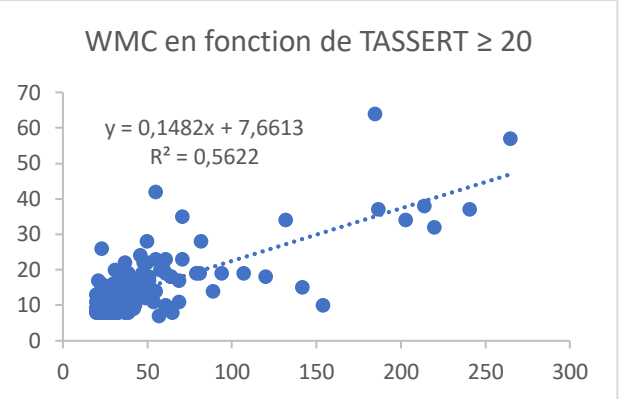
2.2.2 Corrélation de la complexité d'une classe par rapport au nombre d'assertions

En ayant trouvé une corrélation linéaire positive entre TASSERT et WMC nous pouvons créer des droites de régression linéaire pour évaluer la nature de cette proportionnalité.

Lorsque l'on évalue la complexité de la classe par rapport aux nombres d'assertions il est visible que la droite de régression est quasi-équivalente entre l'ensemble complet de données et l'ensemble de classes contenant 20 ou plus assertions. Ceci indique que la complexité de la classe est en effet proportionnelle au nombre d'assertions dans cette classe.

Le facteur d'influence étant de ~7.5 et de ~7.6 dans les deux graphiques indique aussi qu'il y a un seuil minimal de complexité indépendamment du nombre d'assertions dans la classe. Ceci n'est pas un problème pour l'évaluation de l'hypothèse qu'une classe de 20 assertions et plus est plus complexe qu'une classe contenant moins de 20 assertions. Cependant, ce phénomène est perçu dans la distribution des valeurs collectées montrant qu'il y a très peu de classes qui ont des WMC dans le premier quartile.





3 Conclusion

L'analyse des données a cherché à vérifier l'hypothèse selon laquelle les classes contenant plus de 20 assertions sont plus complexes que celles avec moins de 20 assertions. Corrélations indiquent une corrélation significative entre le nombre d'assertions (TASSERT) et la complexité des classes (WMC), surtout pour les classes de plus de 20 assertions.

Il faut tout d'abord éliciter la présence d'un facteur d'influence d'environ 7.5 qui est le seuil minimal de complexité peu importe le nombre d'assertions dans une classe. Cela peut être attribué au fait qu'une classe a probablement une structure qui assure cette complexité minime. Ensuite, pour arriver à la régression linéaire, la valeur r de Pearson a été calculé pour déterminer s'il pourrait y avoir une relation proportionnelle entre TASSERT et WMC. Ainsi le résultat ~0.79 sur l'ensemble général et ~0.75 sur l'ensemble ayant un TASSERT ≥ 20 suggère qu'une relation de proportionnalité linéaire existe. Tout en vient à la distribution des valeurs qui contient trop de valeurs dans les extrêmes et qui rendent la quasi-expérience peu fiable malgré sa validité.

Cela étant dit la quasi-expérience est validé par la distribution des valeurs de WMC lorsque les classes étudiées sont filtrés à TASSERT ≥ 20 étant plus haute que celle générale et celle où le filtrage est TASSERT < 20. Autrement dit, toutes les valeurs de quartiles sont supérieures dans le groupement contenant plus de 20 assertions.

La non-normalité des distributions pour TASSERT et WMC, ainsi que la présence de valeurs extrêmes, peuvent affecter significativement les analyses statistiques, notamment les tests de corrélation et de régression. Ces distributions non normales peuvent indiquer que les données sont influencées par des facteurs spécifiques au projet ou par des pratiques de codage et de test qui ne sont pas uniformément réparties.

Les résultats observés sont en fait valides pourtant leur distribution démontre que ces résultats ne peuvent être considérer fiables. S'il advient que cette expérience soit reproduite, certains autres facteurs contribuant à la complexité devraient être évalués car simplement le nombre d'assertions ne nous permet d'avoir une conclusion fiable sur laquelle on pourrait compter pour étudier d'autres ensembles de tests.

La non-normalité des distributions pour TASSERT et WMC, ainsi que la présence de valeurs extrêmes, peuvent affecter significativement les analyses statistiques, notamment les tests de corrélation et de régression. Ces distributions non normales peuvent indiquer que les données sont influencées par des facteurs spécifiques au projet ou par des pratiques de codage et de test qui ne sont pas uniformément réparties. De plus, la présence de valeurs extrêmes pourrait fausser les résultats globaux et masquer des tendances ou des relations subtiles entre les variables.