

**VERSION
PERMUTATIONNELLE DES
TESTS DE MONTE CARLO :
APPLICATION A L'ETUDE DE
L'INTERACTION A. GAMBIAE -
P. FALCIPARUM**

Par :

TOUSSILE F. WILSON

Maître ès Sciences

Dirigé par

Pr. Henri GWET

Maître de Conférences, ENSP de Yaoundé

et

Dr. Isabelle MORLAIS

Chercheur IRD, OCEAC de Yaoundé

Octobre 2006

Dédicaces

A ma mère
NGAFO Jeanne.

A mon père
TOUSSELE Jacques.

A mon grand-père
YEMELI Pierre.

A Maman
NKENGUE TSAGUE Julienne.

A ma fille
TOUSSILE Amandla.

Remerciements

Au **Pr. Henri GWET** et au **Dr. Isabelle MORLAIS**, pour avoir su raffermir ma passion pour la recherche, pour l'intérêt qu'ils portent à mes travaux, pour avoir toujours été disponibles pendant tout mon travail.

A tous les enseignants du Master de Statistique Appliquée, en particulier au **Pr. Jean-Christophe THALABARD** de l'Université Paris 5 René Descartes, au **Dr. Eugène-Patrice NDONG NGUEMA**, chargé de cours à l'ENSP de Yaoundé, qui n'ont pas hésité à m'apporter leur soutien, qui a été d'une importance certaine dans la réalisation de ce travail.

A mes **parents, mes frères et sœurs**, pour le soutien sans faille et la patience qu'ils ont su me montrer.

Aux amis et camarades, pour leur soutien inconditionnel qui m'a permis de venir à bout de certaines des difficultés rencontrées tout au long de la réalisation de ce travail. Je pense en particulier à Olaf KOUAMO, Pascal SIELENOU, Solange WHEGANG, Achille PEGOUE...

Je remercie toutes les personnes dont j'aurais oublié les noms et qui, de près ou de loin, ont contribué à la réalisation de ce travail.

Table des matières

Dédicaces	i
Remerciements	ii
Table des figures	iii
Liste des tableaux	iii
Abréviations	iv
Résumé	v
Abstract	vi
Introduction	1
1 Généralités sur la théorie des tests	4
1.1 Définitions	4
1.2 Choix d'un test	7
1.3 Tests paramétriques - Tests non paramétriques	8
1.4 Quelques exemples de tests de comparaison	8
1.4.1 Test de Fisher et test de Student	8
1.4.2 Tests de Wilcoxon et Mann-Whitney	9
2 Technique des tests de Monte Carlo	11
2.1 Rappels sur la fonction quantile	11
2.2 Principe des tests de Monte Carlo	13
2.3 Tests de Monte Carlo basés sur des statistiques de test continues	15
2.3.1 Résultats de base	15

2.3.2	Application aux tests	20
2.4	Cas général des tests de Monte Carlo	21
2.5	Algorithme des tests de Monte Carlo	25
2.6	Estimation du nombre de permutations N	26
2.7	Quelques statistiques de test	28
2.8	Combinaison des tests de MC standardisés	30
2.9	Comparaison	31
2.9.1	Distributions continues	31
2.9.2	Distributions discrètes	33
3	Applications	37
3.1	Quelques termes-clef	37
3.2	Cycle biologique du <i>Plasmodium</i>	39
3.3	Matériel et méthodes	40
3.3.1	Collecte des données	40
3.3.2	Description des données	42
3.3.3	Caractéristiques des données	43
3.4	Résultats et discussion	45
3.4.1	Comparaison des moyennes	45
3.4.2	Comparaison des distributions	47
4	Conclusion	51
5	Annexe : Code R de la procédure des tests de Monte Carlo	52
	Bibliographie	61

Table des figures

2.1	Tests de MC basés sur KS, KS classique, Wilcoxon et Student	32
2.2	Puissances dans le cas où l'écart-type varie.	33
3.1	Cycle biologique du <i>Plasmodium</i> [1].	40
3.2	Perte et amplification du parasite dans le moustique [2].	41
3.3	Prévalences dans les groupes <i>GFP</i>	43
3.4	Histogrammes des charges oocystiques.	44
3.5	Moyennes des charges oocystiques.	45
3.6	Prévalences par gène.	46

Liste des tableaux

1.1	Tableau récapitulatif : Décision vs Erreur dans un test.	5
2.1	Puissances et niveaux des tests lorsque les moyennes égales et les variances différentes.	35
2.2	Variances égales et moyennes différentes.	36
3.1	Effectif des moustiques par replicat et par gène.	43
3.2	Caractéristiques des nombres d'oocystes agglomérées.	44
3.3	Intervalles de confiance bootstrap des moyennes des charges parasitaires. . .	45
3.4	p -values des comparaisons des moyennes des groupes traités à celle du groupe contrôle, test de Kruskal-Wallis.	46
3.5	p -values des tests.	47
3.6	Comparaison des groupes de référence	48
3.7	p -values des tests sur les données du groupe <i>P234</i>	49
3.8	p -values des tests sur les données du groupe <i>P57</i>	49
3.9	p -values des tests de MC avec KS comme statistique de test.	49
3.10	p -values des tests de KS classique.	50
3.11	p -values des tests de Wilcoxon.	50

Abréviations

- OCEAC : Organisation de Coordination pour la lutte contre les Endémies en Afrique Centrale
- TAB : tableau
- FIG : figure
- *P.* : *Plasmodium*
- A. : Anophèle
- MC : Monte Carlo
- KS : Kolmogorov-Smirnov
- CM : Cramer-Von Mises
- \mathbb{N} : ensemble des nombres entiers naturels
- \mathbb{Z} : ensemble des nombres entiers relatifs
- \mathbb{R} : ensemble des nombres réels
- \mathbb{P} : probabilité
- $\langle x \rangle$: partie entière du nombre réel x
- ADN : Acide désoxyribonucléique
- ARN : Acide ribonucléique
- PRR : Pattern recognition receptors
- GFP : Green Fluorescent Protein
- APO1 : Apolipophorine I
- CTL4 : C-type lectin 4
- CTLMA2 : C-type lectin MA2
- LRIM1 : Leucine-rich repeat protein
- SRPN2 : Serpine 2
- FIG. : Figure
- TAB. : Tableau

Résumé

Dans ce mémoire, nous implémentons, sous le logiciel R, la version permutationnelle des tests de Monte Carlo pour l'égalité de deux distributions de probabilité inconnues. Nous nous sommes basés sur les travaux de *Jean-Marie DUFOUR* et *Abdeljelil FARHAT* [9], qui ont montré que cette procédure de test contrôle la taille des tests, et qu'elle permet de combiner plusieurs statistiques pour améliorer la puissance des tests individuels. La procédure des tests de Monte Carlo se présente comme une solution aux problèmes de tests d'homogénéité lorsque les conditions d'application d'autres tests non paramétriques ne sont pas remplies. Enfin, nous appliquons cette technique de test pour traiter les données issues d'une étude de l'effet de certains gènes *Anophèles gambiae* sur le cycle biologique du *Plasmodium falciparum*.

Mots clés : Tests d'homogénéité ; test de Monte Carlo ; test de permutations ; *Plasmodium falciparum*.

Abstract

In this thesis, using *R statistical software*, we implemented the permutational version of Monte Carlo tests to solve two-sample problem. We based ourselves on the works of *Jean-Marie DUFOUR* and *Abdeljelil FARHAT* [9], which showed that this procedure controls the size of tests and permits the combination of tests statistics to increase the power of individual tests. The procedure of Monte Carlo tests may serve as a solution to solve two-sample problem when conditions to use other tests are not met. We then use this test procedure to analyse data from a study of the effect of certain immunity genes of *A. gambiae* on the life cycle of *P. falciparum*.

Key words : two-sample problem, Monte Carlo tests, permutation tests, *P. falciparum*.

Introduction

Le paludisme continue de peser de façon importante dans la morbidité et la mortalité en régions tropicales, notamment dans les pays de l'Afrique au Sud du Sahara. Chaque année, trois cent à cinq cent millions de personnes contractent le paludisme dans le monde et près de deux millions en meurent, principalement des enfants [1]. Cette maladie est causée par un parasite appelé *Plasmodium* (abrégé « *P.* » ci-après). Il est transmis à l'Homme par des moustiques femelles du genre *Anophèles* (*A.*). Il existe près de 400 espèces d'Anophèles dont une quarantaine sont capables de transmettre le paludisme, et parmi celles-ci, seulement 15 sont des vecteurs d'importance majeure [3]. Les quatre espèces de *Plasmodium* qui causent le paludisme chez l'homme sont : *P. falciparum*, *P. vivax*, *P. malariae* et *P. ovale*. Parmi ces espèces, le *P. falciparum* est la plus importante dans la plupart des régions tropicales et est responsable de nombreux cas de maladie grave et de décès.

Plusieurs stratégies ont été mises en œuvre pour tenter de limiter la transmission du paludisme, notamment par le contrôle des populations de moustiques à l'aide d'insecticides. Si cette solution a été efficace dans les pays tempérés, le paludisme sévit toujours en Afrique (près de 90 % de décès liés au paludisme surviennent en Afrique [1]), et le problème risque de s'aggraver avec l'émergence, chez les moustiques, de résistances aux insecticides, et le réchauffement de la planète qui risque de favoriser l'extension de l'aire de répartition des espèces anophéliennes vectrices. D'autres tentatives de lutte contre la maladie se sont concentrées sur la prévention de la transmission à l'Homme par la vaccination. Cependant, à ce jour, aucun vaccin efficace n'a été mis au point, et les recherches dans cette voie se poursuivent. Le contrôle génétique des vecteurs du *Plasmodium* s'inscrit également dans les moyens de lutte envisageables. L'objectif de cette méthode est d'interrompre le cycle de transmission du parasite chez son vecteur, donc de rendre le moustique réfractaire au *Plasmodium*. Le contrôle gé-

nétique nécessite alors de comprendre les mécanismes génétiques associés à la compétence vectorielle.

Le *Plasmodium* a développé un cycle biologique complexe qui se déroule en trois phases : une chez le moustique appelée *cycle sporogonique*, et deux chez l'hôte vertébré qui sont les cycles *érythrocytaire* (dans les cellules sanguines) et *exo-érythrocytaire* (hors des cellules sanguines) (cf FIG. 3.2 page 40). A chacun de ces cycles, le parasite subit des phases de multiplications asexuées, palliant les pertes dues au système immunitaire de l'Homme ou du moustique. Certaines espèces de moustiques bloquent complètement le développement de certaines espèces de *Plasmodium*. Par ailleurs, des travaux sur l'expression de gènes de l'immunité suite à l'infection par le *Plasmodium* ou des bactéries ont montré que le moustique est capable de développer une réponse immunitaire contre certains agents pathogènes (Richman 97).

Chez le moustique, l'infection débute lorsque l'*Anophèle* ingère, lors de son repas de sang, des *gamétocytes*, formes sexuées du *Plasmodium*. Les *gamétocytes* mâles et femelles évoluent en *gamètes*. La fécondation aboutit alors à la formation d'un *ookinète*, forme mobile qui va traverser la paroi intestinale du vecteur pour s'enkyster au niveau de la paroi basale, se transformant ainsi en *oocyste*. Une phase d'intense multiplication s'opère à l'intérieur de l'oocyste et au bout d'environ 12 jours après le repas infectant, l'oocyste mûr libère ainsi des millions de *sporozoïtes*. Ceux-ci gagneront les glandes salivaires et seront inoculés à l'Homme lors du prochain repas de sang du moustique.

Depuis quelques années, des chercheurs mènent des études pour identifier les gènes directement impliqués dans la réponse immunitaire du moustique contre le *Plasmodium*. Les nouveaux outils de génomique permettent dorénavant d'étudier l'ensemble des gènes du moustique exprimés dans des conditions biologiques distinctes. Ces outils ont alors été employés, dans le système *P. berghei-A. gambiae*, pour identifier des gènes spécifiquement exprimés lors de l'infection du *Plasmodium* chez le moustique. Les gènes candidats, ainsi détectés, ont été ensuite étudiés par *ARN interférence* afin de vérifier leur rôle dans la transmission. L'*ARN interférence* consiste à supprimer l'expression d'un gène donné pour déterminer sa fonction.

Le laboratoire d'entomologie du paludisme de l'OCEAC (Organisation et Co-ordination pour la lutte contre les Endémies en Afrique Centrale) conduit des recherches sur l'effet de certains gènes candidats sur le développement du *P. falciparum* chez *Anophèles gambiae*. Parmi les gènes testés, la *Leucine-rich re-*

peat protein 1 (en abrégé «LRIM1») et deux *C-types lectin* *CTL4* et *CTLMA2* ont déjà fait l'objet d'étude par *ARN interférence* dans l'interaction *P. berghei-Anophèle gambiae* [4], le système modèle utilisé en laboratoire. Ces gènes candidats, ainsi que deux autres *Apolipoporphine 1 (APO1)* et *Serpine2 (SRPN2)* ont alors été étudiés à l'OCEAC sur le système naturel *P. falciparum-A. gambiae*.

Les données de ce type d'études présentent, d'une part un marqueur génotypique (le gène supprimé), et de l'autre un décompte de parasites sept (7) jours après l'infection expérimentale des moustiques, en plusieurs réplicats indépendants. L'effet d'un gène est déterminé par le test de l'hypothèse nulle \mathcal{H}_0 d'égalité des distributions de la charge parasitaire du groupe traité par ce gène et du groupe contrôle ayant reçu un placebo (*GFP*), contre l'alternative \mathcal{H}_1 que ces distributions sont différentes. Dans des études similaires, *Mike A. Osta et al* [4] utilisent les tests de Kolmogorov-Smirnov ou le test de *Student*, après avoir aggloméré les données des différents réplicats ; d'autres, comme *Michelle M. Riehle et al* [5] utilisent le test de *Wilcoxon-Mann-Whitney*. Le test de Student exige que les données soient issues d'une variable normale. Celui de *Kolmogorov-Smirnov* fait une hypothèse de continuité sur les données. Le test de *Wilcoxon-Mann-Whitney*, quant à lui, comme tous les tests basés plutôt sur les rangs des observations que sur les observations elles mêmes, est sensible aux répétitions dans les données.

D'autre part, le type de distribution de probabilité de la charge parasitaire chez les moustiques, qui est à valeurs dans \mathbb{N} , n'est pas connue à ce jour. L'hypothèse d'égalité des distributions des groupes traité et contrôle est alors non paramétrique. Tester cette hypothèse requiert une procédure statistique utilisable quelque soit le type de distribution de la variable d'intérêt, indépendamment du caractère continu ou discret de la distribution nulle, et qui conserve de bonnes propriétés (niveau et puissance) dans les cas d'échantillons de petite taille. A ce titre, nous allons montrer, dans un premier temps, en nous appuyant sur les travaux de *Jean-Marie DUFOUR* et *Abdeljelil FARHAT* [9], que la procédure des tests de Monte Carlo, utilisée conjointement avec la technique des tests de permutation, est bien adaptée à cette situation. Deuxièmement, nous implémentons cette procédure de test sous le logiciel R et montrons, par des simulations, que sa puissance peut être améliorée en combinant plusieurs statistiques standardisées de test. Nous appliquons enfin la procédure aux données issues de l'étude de l'effet de certains gènes sur le système immunitaire de l'*Anophèle gambiae* sur le développement du *P. falciparum*.

GÉNÉRALITÉS SUR LA THÉORIE DES TESTS

La démarche scientifique dans les sciences expérimentales est la suivante : on fait une hypothèse qui est plus ou moins plausible selon celui qui la fait. Ensuite, on définit une expérience permettant de tester cette hypothèse. Si les résultats de l'expérience ne mènent pas à rejeter l'hypothèse, on n'en conclut pas pour autant qu'elle est exacte ; mais on l'accepte provisoirement, en attendant par exemple qu'une expérience plus puissante permette de la remettre en question.

Les tests statistiques épousent parfaitement cette démarche scientifique. Un test statistique est donc un mécanisme qui permet de trancher entre deux hypothèses le plus souvent contraires l'une de l'autre, au vu des résultats d'un échantillon.

1.1 Définitions

Définition 1.1.1. *Hypothèse nulle*

C'est l'hypothèse dont on cherche à savoir si elle peut être rejetée grâce aux observations dont on dispose. Elle est généralement notée \mathcal{H}_0 .

Cette hypothèse doit être solidement établie et facile à être identifiée. C'est l'hypothèse en laquelle on a le plus confiance, ou à laquelle on tient particulièrement.

Définition 1.1.2. *Hypothèse alternative*

C'est l'hypothèse qui est en concurrence avec l'hypothèse nulle. Elle est généralement notée \mathcal{H}_1 .

L'hypothèse alternative \mathcal{H}_1 est, dans la plupart des cas, le contraire de l'hypothèse nulle \mathcal{H}_0 ; dans tous les cas, le postulat est qu'une seule des deux hypothèses est vraie.

Définition 1.1.3. Région de rejet de \mathcal{H}_0 ou région critique

La région de rejet de l'hypothèse nulle \mathcal{H}_0 d'un test, encore appelée « région critique », est l'ensemble W des valeurs expérimentales pour lesquelles \mathcal{H}_0 est rejetée.

Supposons que les hypothèses \mathcal{H}_0 et \mathcal{H}_1 concernent une variable aléatoire X . Notons \mathcal{P}_0 et \mathcal{P}_1 , les ensembles possibles des lois de probabilité de X sous \mathcal{H}_0 et \mathcal{H}_1 respectivement.

Définition 1.1.4. Risque de première espèce

Le risque de première espèce est la correspondance qui à tout élément $\mathbb{P}_0 \in \mathcal{P}_0$ associe :

$$\alpha_{\mathbb{P}_0} = \mathbb{P}_0(W),$$

soit la probabilité de rejeter \mathcal{H}_0 à tort lorsque \mathbb{P}_0 est la vraie loi de X .

Définition 1.1.5. Niveau d'un test

C'est la borne supérieure α , relativement à toutes les lois de probabilité $\mathbb{P}_0 \in \mathcal{P}_0$, du risque de première espèce; c'est-à-dire

$$\alpha = \sup_{\mathbb{P}_0 \in \mathcal{P}_0} \mathbb{P}_0(W).$$

Dans la pratique, on fixe le niveau des tests à 5%.

Définition 1.1.6. Risque de deuxième espèce et puissance d'un test

• Le risque de deuxième espèce est la correspondance qui à toute loi de probabilité $\mathbb{P}_1 \in \mathcal{P}_1$ associe :

$$\beta_{\mathbb{P}_1} = \mathbb{P}_1(W^c).$$

C'est la probabilité d'accepter à tort l'hypothèse nulle lorsque \mathbb{P}_1 est la vraie loi de X .

• La puissance π d'un test est la probabilité de rejeter avec raison l'hypothèse nulle \mathcal{H}_0 . On a alors

$$\pi_{\mathbb{P}_1} = 1 - \beta_{\mathbb{P}_1} = \mathbb{P}_1(W),$$

pour tout $\mathbb{P}_1 \in \mathcal{P}_1$. Elle dépend de l'hypothèse alternative.

Hypothèses	\mathcal{H}_0 acceptée	\mathcal{H}_0 rejetée
Si \mathcal{H}_0 est vraie	Décision correcte	Erreur de 1 ^{ere} espèce
Si \mathcal{H}_0 est fausse	Erreur de 2 ^{eme} espèce	Décision correcte

TAB. 1.1 – Tableau récapitulatif : Décision vs Erreur dans un test.

Définition 1.1.7. Statistique de test

La statistique de test est une variable aléatoire, fonction de l'échantillon, dont on connaît la loi sous l'hypothèse nulle \mathcal{H}_0 .

Définition 1.1.8. Valeur- p , probabilité critique ou niveau réel d'un test (p -value)

Soient T , la statistique de test, et \mathbb{P}_0 sa loi de probabilité sous \mathcal{H}_0 .

1. Pour un test bilatéral (rejet de \mathcal{H}_0 pour des valeurs trop à droite ou trop à gauche de T), la valeur- p d'une valeur observée T_{obs} de T est donnée par :

$$p(T_{obs}) = \begin{cases} 2\mathbb{P}_0(T \leq T_{obs}) & \text{si } \mathbb{P}_0(T \leq T_{obs}) < 0.5 \\ 2\mathbb{P}_0(T \geq T_{obs}) & \text{si } \mathbb{P}_0(T \leq T_{obs}) \geq 0.5. \end{cases}$$

2. Pour un test unilatéral à droite (rejet de \mathcal{H}_0 pour des valeurs de T trop à droite), la valeur- p d'une valeur observée T_{obs} de T est donnée par :

$$p(T_{obs}) = \mathbb{P}_0(T \geq T_{obs})$$

3. Pour un test unilatéral à gauche (rejet de \mathcal{H}_0 pour des valeurs de T trop à gauche), la valeur- p d'une valeur observée T_{obs} de T est donnée par :

$$p(T_{obs}) = \mathbb{P}_0(T \leq T_{obs}).$$

La valeur- p peut être perçue comme une mesure du poids de l'évidence d'un test statistique : plus elle est petite, plus l'évidence contre l'hypothèse nulle est forte. Au vu de la valeur observée T_{obs} de T , on aura tendance à effectuer un test unilatéral visant à décider si la valeur observée est trop grande ou trop petite. En pratique, pour une statistique de test T de loi de probabilité \mathbb{P}_0 sous \mathcal{H}_0 , on définit souvent la valeur- p d'une valeur T_{obs} observée par :

$$p(T_{obs}) = \min\{\mathbb{P}_0(T \leq T_{obs}), \mathbb{P}_0(T \geq T_{obs})\}.$$

La connaissance de la valeur- p rend inutile la détermination préalable de la région de rejet : on obtient un test au seuil α par la règle de rejet suivante :

$$\text{Rejet de } \mathcal{H}_0 \Leftrightarrow p(T_{obs}) \leq \alpha. \quad (1.1)$$

Dans le cas où T est continue, ceci revient à remplacer la statistique T par $F(T)$ ou $1 - F(T)$, où F désigne la fonction de répartition de T sous \mathcal{H}_0 . Rappelons alors le résultat classique :

Proposition 1.1.1. *Sous l'hypothèse \mathcal{H}_0 , lorsque F est continue, les statistiques $F(T)$ et $1 - F(T)$ suivent la loi uniforme $\mathcal{U}(0, 1)$.*

1.2 Choix d'un test

Soit X , une variable aléatoire à valeurs dans \mathcal{X} , dont on possède un échantillon indépendant de taille n , $(x_1, x_2, \dots, x_n) \in \mathcal{X}^n$.

Se poser un problème de test d'une hypothèse nulle \mathcal{H}_0 contre une alternative \mathcal{H}_1 concernant X revient à rechercher un mécanisme décisionnel qui, au vu du point $\underline{x} = (x_1, x_2, \dots, x_n)$ de \mathcal{X}^n , permet de répondre à la question : laquelle des hypothèses \mathcal{H}_0 ou \mathcal{H}_1 est vraie ?

L'ensemble D des décisions possibles se réduit à deux éléments : $D = \{d_0, d_1\}$, où

$$\begin{cases} d_0 & : \text{ne pas rejeter } \mathcal{H}_0 \\ d_1 & : \text{rejeter } \mathcal{H}_0. \end{cases}$$

Un test est donc défini par une règle de décision Φ , appelée *fonction de test*, définie de \mathcal{X}^n dans D , conduisant à la région critique $W = \Phi^{-1}(d_1)$. Il est fréquent de confondre *test*, *fonction de test* ϕ et *région critique* W .

Définition 1.2.1. Test préférable

Soient Φ et Φ' , deux tests d'une même hypothèse nulle.

On dit que Φ est préférable au sens large à Φ' et on note $\Phi \geq \Phi'$ si

$$\begin{cases} \alpha(\Phi) \leq \alpha(\Phi') \\ \beta(\Phi) \leq \beta(\Phi'), \end{cases}$$

où α et β sont les risques de première et de deuxième espèce respectivement.

Malheureusement, il n'y a aucune raison que les tests qui minimisent un des risques minimisent aussi l'autre [14]. Afin de résoudre ce problème, Neyman et Pearson (1933) proposent de traiter les deux risques de façon non symétrique et de limiter l'ensemble des tests possibles à la classe des tests Φ ayant un risque de 1^{ère} espèce au plus égal à un seuil α_1 fixé préalablement. Notons $\mathcal{C}(\alpha_1)$, cette classe :

$$\mathcal{C}(\alpha_1) = \{\Phi : \mathcal{X}^n \rightarrow \{d_0, d_1\} / \alpha(\Phi) \leq \alpha_1\}.$$

La procédure de Neyman et Pearson consiste donc à rechercher dans $\mathcal{C}(\alpha_1)$, un test Φ minimisant le risque de 2^e espèce β , c'est-à-dire maximisant la puissance π .

Remarque 1.2.1. Lorsque deux tests ont le même niveau, on préférera celui qui a la plus grande puissance.

1.3 Tests paramétriques - Tests non paramétriques

Les *tests paramétriques* sont construits à partir d'hypothèses sur les lois des variables étudiées. En d'autres termes, ces tests supposent de connaître la loi ou la famille de lois sous-jacentes. Lorsque la famille de lois sous-jacentes est inconnue, on optera pour des *tests non paramétriques* qui ne font aucune hypothèse sur la variable étudiée.

1.4 Quelques exemples de tests de comparaison

Les tests de comparaison ont pour but de vérifier si k (≥ 2) échantillons supposés indépendants sont issus de la même loi.

1.4.1 Test de Fisher et test de Student

Les tests de Fisher et de Student servent respectivement à comparer les variances et les moyennes de deux échantillons gaussiens indépendants.

Considérons deux échantillons indépendants (X_1, \dots, X_n) et (Y_1, \dots, Y_m) de lois $\mathcal{N}(\mu_x, \sigma_x)$ et $\mathcal{N}(\mu_y, \sigma_y)$ respectivement. Notons \bar{X} et \bar{Y} les moyennes empiriques, S_x^2 et S_y^2 les variances empiriques des deux échantillons respectivement. L'hypothèse nulle du test de Fisher est $\mathcal{H}_0 : \sigma_x^2 = \sigma_y^2$, et celle du test de Student est $\mathcal{H}_0 : \mu_x = \mu_y$. Les deux théorèmes suivants permettent d'effectuer ces tests.

Théorème 1.4.1. *Le rapport*

$$\frac{\frac{nS_x^2}{(n-1)\sigma_x^2}}{\frac{mS_y^2}{(m-1)\sigma_y^2}}$$

suit la loi $\mathcal{F}(n-1, m-1)$.

La statistique de Fisher est définie par :

$$T = \frac{\frac{nS_x^2}{n-1}}{\frac{mS_y^2}{m-1}}.$$

Remarque 1.4.1. *Sous \mathcal{H}_0 , T suit la loi de Fisher $\mathcal{F}(n-1, m-1)$. Pour un test bilatéral, on rejettera \mathcal{H}_0 lorsque la valeur observée T_{obs} de T est trop petite ou trop grande. Notons α , le seuil du test.*

$$\text{Rejet de } \mathcal{H}_0 \Leftrightarrow T \notin [\mathcal{Q}_{\mathcal{F}(n-1, m-1)}(\alpha/2), \mathcal{Q}_{\mathcal{F}(n-1, m-1)}(1 - \alpha/2)],$$

où $Q_{\mathcal{F}(n-1, m-1)}$ est la fonction quantile de la loi de Fisher $\mathcal{F}(n-1, m-1)$.

Théorème 1.4.2. Si $\sigma_x^2 = \sigma_y^2$, alors

$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{nS_x^2 + mS_y^2}{n+m-2}}}$$

suit la loi de Student $\mathcal{T}(n+m-2)$.

Ce résultat permet de tester l'égalité des moyennes des deux échantillons en comparant la statistique de Student

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{nS_x^2 + mS_y^2}{n+m-2}}}$$

aux quantiles de la loi de Student $\mathcal{T}(n+m-2)$.

1.4.2 Tests de Wilcoxon et Mann-Whitney

Supposons cette fois que nos deux échantillons précédents sont de lois F_x et F_y inconnues. Considérons l'hypothèse $\mathcal{H}_0 : F_x = F_y$, et supposons que $F_x(t) = F_y(t - \delta)$, où δ est le paramètre de translation. Alors \mathcal{H}_0 devient $\delta = 0$. L'idée du test de Wilcoxon est la suivante : si on rassemble les deux échantillons, et que l'on range les valeurs dans l'ordre croissant par exemple, l'alternance des X_i et des Y_j devrait être assez régulière sous \mathcal{H}_0 . On aura des doutes sur \mathcal{H}_0 si les Y_j sont plutôt plus grands que les X_i , ou plus petits, ou plus fréquents dans une certaine plage de valeurs. On commence donc par écrire les statistiques d'ordre de l'échantillon global en tirant des permutations pour les ex-æquo. On obtient ainsi une suite mélangée des X_i et des Y_j . On calcule ensuite la somme W_x des rangs des X_i : W_x est la statistique de Wilcoxon.

Sur un échantillon de taille $n+m$, il y a $(n+m)!$ ordres possibles sous \mathcal{H}_0 . Le nombre de rangements possibles des X_i est $C(n+m, n)$ (avec $C(n, k) = \frac{n!}{k!(n-k)!}$), et ils sont équiprobables sous \mathcal{H}_0 . Pour t entier allant de $C(n+1, 2)$ à $C(n+m+1, 2) - C(m+1, 2)$,

$$\mathbb{P}(W_x = t) = \frac{k_t}{C(n+m, n)},$$

où k_t est le nombre de n -uplets d'entiers r_1, r_2, \dots, r_n dont la somme vaut t et qui sont tels que $1 \leq r_1 < r_2 < \dots < r_n \leq n+m$

Pour des grandes valeurs de n et m on a le résultat d'approximation normale suivant :

Théorème 1.4.3. *Sous l'hypothèse \mathcal{H}_0 , la loi de*

$$\frac{W_x - \frac{1}{2}n(n+m+1)}{\sqrt{\frac{1}{12}nm(n+m+1)}}$$

converge vers la loi normale $\mathcal{N}(0,1)$ quand n et m tendent vers ∞ .

Le test de Mann-Whitney est équivalent au précédent. Sa statistique de test est :

$$U = \sum_{i=1}^n \sum_{j=1}^m 1_{X_i > Y_j} ;$$

U et W_x sont liées par la relation suivante :

$$U = W_x - \frac{1}{2}n(n+1).$$

Remarque 1.4.2. *Les tests de Wilcoxon et de Mann-Whitney, comme la plupart des tests basés sur les rangs des observations, ne sont pas adaptés aux données avec beaucoup de répétitions, en particulier dans le cas où il y a des répétitions dans les données, du fait de l'attribution aléatoire des rangs aux ex-aequo.*

TECHNIQUE DES TESTS DE MONTE CARLO

La méthode de test de Monte Carlo (MC) a été proposée initialement par Dwass (1957) [9]. Elle permet de construire des tests exacts à partir de statistiques dont les distributions, dans le cas d'échantillons finis, ne sont pas standard, mais peuvent être simulées. Un des principaux avantages de cette technique est le fait que, contrairement à d'autres méthodes conventionnelles de test basées sur des résultats asymptotiques, une inférence exacte sur des échantillons finis peut être obtenue.

2.1 Rappels sur la fonction quantile

Définition 2.1.1. Fonction quantile

La fonction quantile d'une fonction de répartition F définie sur \mathbb{R} est la fonction inverse généralisée $F^{-1} : [0, 1] \rightarrow \mathbb{R}$ définie par :

$$F^{-1}(p) = \begin{cases} \inf \{x / F(x) \geq p\} & \text{si } 0 < p < 1, \\ \inf \{x / F(x) > 0\} & \text{si } p = 0, \\ \sup \{x / F(x) < 1\} & \text{si } p = 1. \end{cases}$$

En général, F^{-1} prend ses valeurs dans $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$, et, pour être cohérent, on pose $F(-\infty) = 0$ et $F(+\infty) = 1$.

Remarque 2.1.1. Lorsque F est continue et strictement croissante sur $]F^{-1}(0), F^{-1}(1)[$, F^{-1} est la bijection réciproque de F définie de $]0, 1[$ vers $]F^{-1}(0), F^{-1}(1)[$.

Lemme 2.1.1. Pour tous $0 < p < 1$ et $x \in \mathbb{R}$,

- i) $F^{-1}(p) \leq x \Leftrightarrow p \leq F(x)$;
- ii) $F[F^{-1}(p)] \geq p$ avec égalité si F est continue en $F^{-1}(p)$;
- iii) $F_- [F^{-1}(p)] \leq p$, où $F_-(x) = \lim_{\varepsilon \downarrow 0} F(x - \varepsilon)$;

iv) $F^{-1}[F(x)] \leq x$ avec inégalité stricte si et seulement si F est constante sur un intervalle de la forme $[x - h, x]$, $h > 0$.

Preuve.

i) Soient $p \in]0, 1[$ et $x \in \mathbb{R}$. On a : $p \leq F(x)$ entraîne $F^{-1}(p) \leq x$ par définition. Supposons maintenant $F^{-1}(p) \leq x$. Pour tout $\varepsilon > 0$, il existe $x_\varepsilon \in \mathbb{R}$ tel que $p \leq F(x_\varepsilon)$ et $F^{-1}(p) \leq x_\varepsilon < F^{-1}(p) + \varepsilon$ (*). La 2^e inégalité de (*) implique que $F(x_\varepsilon) \leq F[F^{-1}(p) + \varepsilon]$, F étant croissante. On a alors $p \leq F[F^{-1}(p) + \varepsilon]$. Ceci étant vrai pour tout $\varepsilon > 0$, et F étant continue à droite, il vient $p \leq F[F^{-1}(p)]$. D'après l'hypothèse, $F[F^{-1}(p)] \leq F(x)$. Ce qui achève de montrer *i)*.

ii) Poser $x = F^{-1}(p)$ et appliquer *i)*, d'une part.

Pour la 2^e partie, supposons F continue en $F^{-1}(p)$. Soit $\varepsilon > 0$. On a $F^{-1}(p) > F^{-1}(p) - \varepsilon$, d'où d'après *i)*, $p > F[F^{-1}(p) - \varepsilon]$. En passant à la limite lorsque $\varepsilon \downarrow 0$, $p \geq F[F^{-1}(p)]$, F étant continue à gauche en $F^{-1}(p)$; d'où l'égalité.

iii) Dans la preuve de *ii)*, nous avons vu que $p > F[F^{-1}(p) - \varepsilon]$ pour tout $\varepsilon > 0$. On en déduit *iii)*.

iv) Poser $p = F(x)$ et appliquer *i)*, d'une part.

Supposons $F^{-1}[F(x)] < x$. Alors $F[F^{-1}(F(x))] \leq F(x)$. D'après *ii)*, $F[F^{-1}(F(x))] \geq F(x)$. D'où l'égalité $F[F^{-1}(F(x))] = F(x)$, et par suite F est constante sur $[F^{-1}(F(x)), x]$.

Supposons maintenant F constante sur un intervalle de la forme $[x - h, x]$, $h > 0$. Alors $F(x - h) = F(x)$. Or $F^{-1}[F(x)] = \inf\{y / F(y) \geq F(x)\}$. On a donc $x - h \in \{y / F(y) \geq F(x)\}$, ce qui implique $F^{-1}[F(x)] \leq x - h < x$ ■

Proposition 2.1.1. Soient F , une fonction de répartition sur \mathbb{R} et F^{-1} , sa fonction quantile. Soit U , une variable aléatoire de loi uniforme sur $[0, 1]$. La variable aléatoire $X = F^{-1}(U)$ a pour fonction de répartition F .

Preuve.

D'après *i)* du **Lemme 2.1.1**,

$$\begin{aligned} \mathbb{P}[F^{-1}(U) \leq x] &= \mathbb{P}[U \leq F(x)] \\ &= F(x). \end{aligned}$$

■

Définition 2.1.2. Fonction de répartition empirique

Soit (X_1, \dots, X_n) , un échantillon de loi μ sur \mathbb{R} et de fonction de répartition F .

On appelle fonction de répartition empirique associée à l'échantillon (X_1, \dots, X_n) , la fonction aléatoire $\widehat{F}_n : \mathbb{R} \rightarrow [0, 1]$ définie pour tout $x \in \mathbb{R}$ par :

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{[X_i \leq x]},$$

où

$$1_{[cond]} = \begin{cases} 1 & \text{si la condition cond est vérifiée} \\ 0 & \text{sinon.} \end{cases}$$

Définition 2.1.3. Fonction quantile empirique

On appelle fonction quantile empirique associée à un échantillon (X_1, \dots, X_n) , la fonction quantile de la fonction de répartition \widehat{F}_n . Elle est notée \widehat{F}_n^{-1} .

2.2 Principe des tests de Monte Carlo

Soient $X_{11}, X_{12}, \dots, X_{1n}$ et $X_{21}, X_{22}, \dots, X_{2m}$, deux échantillons indépendants et identiquement distribués chacun. Posons :

$$F_1(x) = \mathbb{P}(X_{1i} \leq x), \quad i = 1, \dots, n,$$

et

$$F_2(x) = \mathbb{P}(X_{2j} \leq x), \quad j = 1, \dots, m.$$

On n'impose aucune condition sur les fonctions de répartition F_1 et F_2 , lesquelles peuvent être continues ou non. Le problème est de tester l'hypothèse

$$\mathcal{H}_0 : F_1 = F_2 \tag{2.1}$$

contre

$$\mathcal{H}_1 : F_1 \neq F_2. \tag{2.2}$$

Soit T , une statistique de test. Notons T_0 , la valeur prise par T sur un jeu d'observations $(x_{11}, \dots, x_{1n}; x_{21}, \dots, x_{2m})$:

$$T_0 = T(x_{11}, \dots, x_{1n}; x_{21}, \dots, x_{2m}). \tag{2.3}$$

Note :

- Cette valeur observée T_0 est supposée fixée pour toute la suite.

Sans nuire à la généralité, plaçons nous dans le cadre d'un test unilatéral à droite, ce qui signifie que l'hypothèse nulle (2.1) est rejetée pour des "trop grandes" valeurs de T_0 . Soient

$$F(x) = \mathbb{P}(T \leq x / \mathcal{H}_0)$$

et

$$G(x) = \mathbb{P}(T \geq x / \mathcal{H}_0),$$

les fonctions de répartition et de survie de T sous \mathcal{H}_0 .

Lorsque F est connue et F^{-1} calculable par des méthodes analytiques, on rejette \mathcal{H}_0 si $T_0 > F^{-1}(1 - \alpha)$, où α est le seuil fixé du test.

Lorsque F n'est pas connue, la méthode des tests de Monte Carlo consiste à remplacer la distribution théorique F par un estimateur simulé \widehat{F}_N , tout en préservant le niveau du test, quelque soit le nombre de simulations N .

Dans toute la suite de ce chapitre, $T(N) = (T_1, \dots, T_N)$ est un échantillon indépendant et identiquement distribué (iid) de taille N de la statistique T sous l'hypothèse nulle \mathcal{H}_0 , et nous supposons que $T(N)$ est indépendant de T_0 , de sorte que (T_0, T_1, \dots, T_N) est un échantillon iid de taille $N + 1$ de T .

On considère l'estimateur de F suivant :

$$\widehat{F}_N(x) \equiv \widehat{F}_N(x; T(N)) = \frac{1}{N} \sum_{i=1}^N 1_{[T_i \leq x]}. \quad (2.4)$$

Posons :

$$\widehat{G}_N(x) \equiv \widehat{G}_N(x; T(N)) = \frac{1}{N} \sum_{i=1}^N 1_{[T_i \geq x]} \quad (2.5)$$

Pour un test unilatéral à droite, il est plus pertinent d'estimer la valeur- p $p(T_0) = \mathbb{P}_0(T \geq T_0)$ (cf **Définition 1.1.8**) par :

$$\begin{aligned} \widehat{p}_N(T_0) &= \frac{1}{N+1} \sum_{i=0}^N 1_{[T_i \geq T_0]} \\ &= \frac{1}{N+1} \left[1 + \sum_{i=1}^N 1_{[T_i \geq T_0]} \right], \end{aligned}$$

c'est-à-dire par

$$\widehat{p}_N(T_0) = \frac{N \cdot \widehat{G}_N(T_0) + 1}{N + 1}. \quad (2.6)$$

Et pour un test unilatéral à gauche, on estimera la valeur- p $p(T_0) = \mathbb{P}_0(T \leq T_0)$ par :

$$\begin{aligned}\widehat{p}_N(T_0) &= \frac{1}{N+1} \sum_{i=0}^N 1_{[T_i \leq T_0]} \\ &= \frac{1}{N+1} \left[1 + \sum_{i=1}^N 1_{[T_i \leq T_0]} \right],\end{aligned}$$

c'est-à-dire

$$\widehat{p}_N(T_0) = \frac{N \cdot \widehat{F}_N(T_0) + 1}{N+1}. \quad (2.7)$$

2.3 Tests de Monte Carlo basés sur des statistiques de test continues

2.3.1 Résultats de base

Définition 2.3.1. Rang de T_j

Si les T_0, T_1, \dots, T_N sont deux à deux distincts, le rang $R_j \equiv R_j(T(N))$ de T_j dans T_0, T_1, \dots, T_N est la position de T_j parmi les statistiques d'ordre $T_{(0)}, T_{(1)}, \dots, T_{(N)}$.

Si T_j est égale à d'autres observations, cette définition n'est plus valide. le rang R_j de T_j est défini comme la moyenne arithmétique de tous les indices i tels que $T_j = T_{(i)}$, ou comme

$$R_j = \sum_{i=0}^N 1_{[T_i \leq T_j]}. \quad (2.8)$$

Remarque 2.3.1. 1. La définition du rang (2.8) est vérifiée lorsque T_0, \dots, T_N sont deux à deux distincts.

2. On a, dans le cas de non répétitions :

$$\begin{aligned}R_0 &= 1 + N \cdot \widehat{F}_N(T_0; T(N)) \\ &= 1 + N \left[1 - \widehat{G}_N(T_0; T(N)) \right],\end{aligned}$$

ce qui permet d'écrire :

$$N \widehat{G}_N(T_0; T(N)) = 1 + N - R_0, \quad (2.9)$$

et, d'après (2.6) et (2.7), on estime les valeurs- p des tests unilatéraux à droite et gauche respectivement par :

$$\widehat{p}_N(T_0) = \frac{1 + N - R_0 + 1}{N + 1} \quad (2.10)$$

$$= 1 - \frac{R_0 - 1}{N + 1} \quad (2.11)$$

et

$$\widehat{p}_N(T_0) = \frac{R_0}{N + 1}. \quad (2.12)$$

La **Proposition 2.3.1** ci-dessous montre que l'estimateur \widehat{F}_N de F nous permet d'obtenir les valeurs critiques ou d'estimer la valeur- p de sorte à préserver la taille du test, indépendamment du nombre de simulations N . La preuve de cette proposition utilise le **Lemme 2.3.1** ci-dessous [9].

Définition 2.3.2. Variables aléatoires échangeables

Des variables aléatoires X_1, \dots, X_N sont dites échangeables si les vecteurs (X_1, \dots, X_N) et $(X_{\sigma(1)}, \dots, X_{\sigma(N)})$ suivent la même loi pour toute permutation σ des entiers $1, \dots, N$.

Définition 2.3.3. Nœud dans un échantillon

On dit qu'il y a un nœud dans un échantillon X_1, \dots, X_k , s'il existe au moins un couple (i, j) dans $\{1, \dots, k\}^2$, $i \neq j$, tel que

$$\mathbb{P}(X_i = X_j) > 0.$$

En plus de l'échantillon T_1, \dots, T_N de T , nous disposons de la valeur prise sur les observations par T , que nous avons notée T_0 . Nous avons ainsi un échantillon T_0, T_1, \dots, T_N de taille $N + 1$ de T .

Lemme 2.3.1. Distribution des rangs en l'absence de nœuds

Supposons l'échantillon T_0, \dots, T_N sans nœuds, c'est-à-dire que :

$$\mathbb{P}(T_i = T_j) = 0 \quad \text{si } i \neq j, \quad i \text{ et } j = 0, 1, \dots, N. \quad (2.13)$$

Sous \mathcal{H}_0 , on a, pour tout $j = 0, 1, \dots, N$,

$$\mathbb{P}\left(\frac{R_j}{N + 1} \leq x\right) = \frac{\langle x(N + 1) \rangle}{N + 1}, \quad \forall x \in [0, 1] \quad (2.14)$$

$$\mathbb{P}\left(\frac{R_j}{N+1} \geq x\right) = \begin{cases} 1 & \text{si } x \leq 0 \\ \frac{\langle(1-x)(N+1)\rangle + 1}{N+1} & \text{si } 0 < x \leq 1 \\ 0 & \text{si } x > 1, \end{cases} \quad (2.15)$$

où, rappelons le, $\langle x \rangle$ désigne la partie entière de x .

Preuve.

D'après l'hypothèse (2.13), les variables T_0, T_1, \dots, T_N sont presque sûrement distinctes deux à deux. On en déduit que le vecteur des rangs associés (R_0, R_1, \dots, R_N) est presque sûrement une permutation des entiers $1, 2, \dots, N+1$, d'une part. D'autre part, les variables aléatoires T_0, T_1, \dots, T_N sont échangeables sous \mathcal{H}_0 , ce qui implique que sous \mathcal{H}_0 , la probabilité d'obtenir un $(N+1)$ -uplet de rangs particuliers associés à un échantillon (T_0, T_1, \dots, T_N) est de $\frac{1}{(N+1)!}$. Pour tout $j = 0, 1, \dots, N$, on a :

$$\mathbb{P}(R_j = i) = \frac{N!}{(N+1)!} = \frac{1}{N+1}, \quad i = 1, 2, \dots, N+1,$$

$$\begin{aligned} \mathbb{P}[R_j \leq x(N+1)] &= \mathbb{P}[R_j \leq \langle x(N+1) \rangle] ; \\ &= \frac{\langle x(N+1) \rangle}{N+1}, \forall x \in [0; 1], \end{aligned}$$

c'est-à-dire (2.14).

Par ailleurs,

- pour $x \leq 0$, $\mathbb{P}[R_j \geq x(N+1)] = 1$;
- pour $x > 1$, $\mathbb{P}[R_j \geq x(N+1)] = 0$.

Pour conclure (2.15), il reste à monter que :

$$\forall 0 < x \leq 1, \quad \mathbb{P}\left(\frac{R_j}{N+1} \geq x\right) = \frac{\langle(1-x)(N+1)\rangle + 1}{N+1}.$$

Pour cela, remarquons que, pour tout réel r ,

$$\langle N - r \rangle = \begin{cases} N - r & \text{si } r \in \mathbb{Z}, \\ N - \langle r \rangle - 1 & \text{sinon.} \end{cases} \quad (2.16)$$

En effet, le cas $r \in \mathbb{Z}$ est trivial. Alors que, pour $r \notin \mathbb{Z}$, on a :

$$\begin{aligned} \langle r \rangle < r < \langle r \rangle + 1 &\Leftrightarrow -\langle r \rangle - 1 < -r < -\langle r \rangle \\ &\Leftrightarrow N - \langle r \rangle - 1 < N - r < N - \langle r \rangle, \end{aligned}$$

d'où $\langle N - r \rangle = N - \langle r \rangle - 1$.

Soit $x \in]0; 1]$.

• 1^{er} cas : $x(N + 1) \in \mathbb{N}$.

$$\begin{aligned} \mathbb{P}\left(\frac{R_j}{N+1} < x\right) &= \mathbb{P}[R_j < x(N+1)] \\ &= \frac{x(N+1) - 1}{N+1}. \end{aligned}$$

Alors

$$\begin{aligned} \mathbb{P}\left(\frac{R_j}{N+1} \geq x\right) &= 1 - \mathbb{P}\left(\frac{R_j}{N+1} < x\right) \\ &= 1 - \frac{x(N+1) - 1}{N+1} \\ &= \frac{N+1 - x(N+1) + 1}{N+1} \\ &= \frac{(1-x)(N+1) + 1}{N+1}, \end{aligned}$$

ce qui est bien (2.15) dans ce cas, car $(1-x)(N+1) \in \mathbb{Z}$.

• 2^e cas : $x(N+1) \notin \mathbb{N}$.

$$\begin{aligned} \mathbb{P}\left(\frac{R_j}{N+1} < x\right) &= \mathbb{P}[R_j < x(N+1)] \\ &= \frac{\langle x(N+1) \rangle}{N+1}. \end{aligned}$$

On a alors :

$$\begin{aligned} \mathbb{P}\left(\frac{R_j}{N+1} \geq x\right) &= 1 - \mathbb{P}\left(\frac{R_j}{N+1} < x\right) \\ &= 1 - \frac{\langle x(N+1) \rangle}{N+1} \\ &= \frac{N+1 - \langle x(N+1) \rangle}{N+1} \\ &= \frac{\langle (1-x)(N+1) \rangle + 1}{N+1} \text{ par (2.16),} \end{aligned}$$

ce qui achève de montrer (2.15) ■

Proposition 2.3.1. Validité des tests de Monte Carlo sous (2.13)

Considérons le vecteur des $N + 1$ variables aléatoires $(T_i)_{i=0}^N$. Supposons que ce vecteur n'ait pas de nœuds (cf (2.13)). Soient les fonctions \widehat{F}_N et \widehat{G}_N définies par (2.4) et (2.5).

Sous \mathcal{H}_0 , on a :

$$\mathbb{P} \left[\widehat{G}_N(T_0) \leq \alpha_1 \right] = \mathbb{P} \left[\widehat{F}_N(T_0) \geq 1 - \alpha_1 \right] = \frac{\langle \alpha_1 N \rangle + 1}{N + 1}, \quad \forall \alpha_1 \in [0, 1]; \quad (2.17)$$

$$\mathbb{P} \left[T_0 \geq \widehat{F}_N^{-1}(1 - \alpha_1) \right] = \frac{\langle \alpha_1 N \rangle + 1}{N + 1}, \quad \forall \alpha_1 \in]0, 1[; \quad (2.18)$$

$$\mathbb{P} \left[\widehat{p}_N(T_0) \leq \alpha \right] = \frac{\langle \alpha(N + 1) \rangle}{N + 1}, \quad \forall \alpha \in [0, 1]. \quad (2.19)$$

Preuve.

Supposons que $(T_i)_{i=0}^N$ vérifie (2.13). On a, presque sûrement :

$$\widehat{G}_N(T_0) = \frac{N + 1 - R_0}{N},$$

où $R_0 = \sum_{i=0}^N \mathbf{1}_{[T_i \leq T_0]}$ est le rang de T_0 parmi les $N + 1$ variables $(T_i)_{i=0}^N$ rangés dans l'ordre croissant (cf (2.9) de la **Remarque 2.3.1**). D'après (2.15) du **Lemme 2.3.1**, on a, $\forall \alpha_1 \in [0, 1]$:

$$\begin{aligned} \mathbb{P} \left[\widehat{G}_N(T_0) \leq \alpha_1 \right] &= \mathbb{P} \left[\frac{N + 1 - R_0}{N} \leq \alpha_1 \right] \\ &= \mathbb{P} \left[\frac{R_0}{N + 1} \geq \frac{(1 - \alpha_1)N + 1}{N + 1} \right] \\ &= \frac{\left\langle \left(1 - \frac{(1 - \alpha_1)N + 1}{N + 1}\right)(N + 1) \right\rangle + 1}{N + 1} \\ &= \frac{\langle \alpha_1 N \rangle + 1}{N + 1}, \end{aligned}$$

d'où (2.17).

D'autre part, $\widehat{F}_N(T_0) = 1 - \widehat{G}_N(T_0)$ presque sûrement, et, d'après *i*) du **Lemme 2.1.1**, $\widehat{F}_N(y) \geq p \Leftrightarrow y \geq \widehat{F}_N^{-1}(p)$ pour tous $y \in \mathbb{R}$ et $p \in]0, 1[$, ce qui permet d'avoir (2.18).

Notons enfin que :

$$\widehat{p}_N(T_0) \equiv \frac{N\widehat{G}_N(T_0) + 1}{N + 1} \leq \alpha \Leftrightarrow \widehat{G}_N(T_0) \leq \frac{\alpha(N + 1) - 1}{N}.$$

Or, $0 \leq \widehat{G}_N(T_0) \leq 1$ et, en utilisant (2.18),

$$\begin{aligned} \mathbb{P}[\widehat{p}_N(T_0) \leq \alpha] &= \mathbb{P}\left[\widehat{G}_N(T_0) \leq \frac{\alpha(N+1)-1}{N}\right] \\ &= \mathbb{P}\left\{T_0 \geq \widehat{F}_N^{-1}\left[1 - \frac{\alpha(N+1)-1}{N}\right]\right\}, \text{ par (2.17),} \\ &= \begin{cases} 0 & \text{si } \alpha < \frac{1}{N+1} \\ \frac{\langle \alpha(N+1) - 1 \rangle + 1}{N+1} = \frac{\langle \alpha(N+1) \rangle}{N+1} & \text{si } \frac{1}{N+1} \leq \alpha \leq 1 \\ 1 & \text{si } \alpha > 1, \end{cases} \end{aligned}$$

d'où (2.19) en observant que $\langle \alpha(N+1) \rangle = 0$ pour $0 \leq \alpha < \frac{1}{N+1}$ ■

2.3.2 Application aux tests

On a le résultat suivant :

Lemme 2.3.2. Si α_1 et α sont liés par $\alpha_1 = \alpha - \frac{1-\alpha}{N}$, on a

$$\left(\widehat{G}_N(T_0) \leq \alpha_1\right) = \left(\widehat{p}_N(T_0) \leq \alpha\right),$$

au sens de l'égalité de 2 ensembles.

Preuve. Il suffit de remarquer que $\widehat{p}_N(T_0) = \frac{N\widehat{G}_N(T_0) + 1}{N+1}$ ■

Soit α le niveau visé du test. Pour des raisons pratiques, nous choisissons N tel que $\alpha(N+1) \in \mathbb{N}$ et $\alpha_1 = \alpha - \frac{1-\alpha}{N}$. Il est clair que α_1 tend vers α lorsque N tend vers ∞ . Donc pour N très grand, les régions critiques $\left(\widehat{G}_N(T_0) \leq \alpha_1\right)$ et $\left(\widehat{G}_N(T_0) \leq \alpha\right)$ sont équivalentes. On peut aussi remarquer que

$$\begin{aligned} \frac{\langle \alpha_1 N \rangle + 1}{N+1} &= \frac{\langle \alpha(N+1) - 1 \rangle + 1}{N+1} \\ &= \alpha = \frac{\langle \alpha(N+1) \rangle}{N+1}; \end{aligned}$$

de sorte que sous ces contraintes supplémentaires sur α et N , la région critique randomisée $\left(\widehat{G}_N(T_0) \leq \alpha_1\right) = \left(\widehat{p}_N(T_0) \leq \alpha\right)$ a le même niveau α que la région critique non randomisée $\left(G(T_0) \equiv 1 - F(T_0) \leq \alpha\right)$.

Corollaire 2.3.1. *Si la statistique de test T est continue, α et N choisis de sorte que $\alpha(N+1) \in \mathbb{N}$, alors la région critique randomisée ($\widehat{p}_N(T_0) \leq \alpha$) a le même niveau α que celle non randomisée ($G(T_0) \equiv 1 - F(T_0) \leq \alpha$).*

Preuve. Il suffit de remarquer que si T est continue, l'hypothèse (2.13) est vérifiée ■

2.4 Cas général des tests de Monte Carlo

L'hypothèse d'absence de nœuds (2.13) joue un important rôle dans la preuve de la **Proposition 2.3.1**. On étend les résultats obtenus sous cette hypothèse en détruisant les éventuels nœuds par un procédé d'attribution de rangs aléatoires aux valeurs répétées de la statistique de test dans l'échantillon (T_0, \dots, T_N) .

Associons à chaque T_i , $i = 0, 1, \dots, N$ une variable aléatoire uniforme U_i telle que :

$$U_0, U_1, \dots, U_N \text{ soient iid} \quad (2.20)$$

et indépendantes des T_i , $i = 0, 1, \dots, N$.

Posons $U(N) = (U_i)_{i=1}^N$ et $T(N) = (T_i)_{i=1}^N$, et considérons les paires

$$Z_i = (T_i, U_i), \quad i = 0, 1, \dots, N. \quad (2.21)$$

Définition 2.4.1. Ordre lexicographique

$$(T_i, U_i) < (T_j, U_j) \Leftrightarrow [T_i < T_j \text{ ou } (T_i = T_j \text{ et } U_i < U_j)]. \quad (2.22)$$

On en déduit l'indicatrice

$$1_{[(x_1, u_1) < (x_2, u_2)]} = 1_{[x_1 < x_2]} + \delta_0(x_1 - x_2)1_{[u_1 < u_2]} \quad (2.23)$$

où δ_0 est la mesure de Dirac en 0. Le rang de T_j est alors défini comme celui du couple $Z_j = (T_j, U_j)$ par rapport à l'ordre lexicographique.

Posons :

$$\widetilde{R}_j \equiv \widetilde{R}_j [T(N), U(N)] = \sum_{i=0}^N 1_{[(T_i, U_i) \leq (T_j, U_j)]}, \quad j = 0, 1, \dots, N. \quad (2.24)$$

Remarque 2.4.1. Il découle de la continuité de la loi uniforme que les couples Z_j , $j = 0, 1, \dots, N$, sont presque sûrement deux à deux distincts, de sorte que le vecteur des rangs aléatoires $\left(\tilde{R}_j\right)_{j=0}^N$ est presque sûrement une permutation des entiers $1, 2, \dots, N + 1$. Les $\left(\tilde{R}_j\right)_{j=0}^N$ vérifient donc le **Lemme 2.3.1** lorsqu'on y remplace les T_i par les Z_i , et les R_j par les \tilde{R}_j .

On associe aux rangs \tilde{R}_j , $j = 1, 2, \dots, N$, la fonction empirique suivante :

$$\tilde{F}_N(x) \equiv \tilde{F}_N(x; U_0, T(N), U(N)) = \frac{1}{N} \sum_{i=1}^N 1_{[(T_i, U_i) \leq (x, U_0)]}. \quad (2.25)$$

Remarquons que :

$$\forall x \in \mathbb{R}, (T_i, U_i) \leq (x, U_0) \Rightarrow T_i \leq x.$$

Il s'ensuit :

$$\forall x \in \mathbb{R}, \tilde{F}_N(x) \leq \hat{F}_N(x).$$

Par ailleurs, par (2.23), presque sûrement :

$$\tilde{F}_N(x) = 1 - \hat{G}(x; T(N)) + T^{(N)}(x; U_0, T(N), U(N)) \quad (2.26)$$

où

$$T^{(N)}(x; U_0, T(N), U(N)) = \frac{1}{N} \sum_{i=1}^N \delta_0(T_i - x) 1_{[U_i \leq U_0]} \quad (2.27)$$

$$= \frac{1}{N} \sum_{i \in E_N(x)} 1_{[U_i \leq U_0]}, \quad (2.28)$$

avec

$$E_N(x) = \{i / T_i = x, 1 \leq i \leq N\}.$$

Remarque 2.4.2. La fonction $\tilde{F}_N(x)$ a toutes les propriétés d'une fonction de répartition, sauf la continuité à droite. En effet, en certains points, elle peut prendre des valeurs intermédiaires entre ses limites à gauche et à droite.

On peut aussi définir la fonction aléatoire de survie associée suivante :

$$\tilde{G}_N(x; U_0, T(N), U(N)) = \frac{1}{N} \sum_{i=1}^N 1_{[(T_i, U_i) \geq (x, U_0)]},$$

c'est-à-dire :

$$\tilde{G}_N(x; U_0, T(N), U(N)) = 1 - \hat{F}_N(x; T(N)) + \bar{T}^{(N)}(x; U_0, T(N), U(N)), \quad (2.29)$$

où

$$\bar{T}^{(N)}(x; U_0, T(N), U(N)) = \frac{1}{N} \sum_{i=1}^N \delta_0(T_i - x) 1_{[U_i \geq U_0]} \quad (2.30)$$

$$= \frac{1}{N} \sum_{i \in E_N(x)} 1_{[U_i \geq U_0]}. \quad (2.31)$$

Remarque 2.4.3. On peut déduire de (2.26) - (2.31) les inégalités suivantes :

$$1 - \hat{G}_N(x; T(N)) \leq \tilde{F}_N(x; U_0, T(N), U(N)) \leq \hat{F}_N(x; T(N)) \quad (2.32)$$

$$1 - \hat{F}_N(x; T(N)) \leq \tilde{G}_N(x; U_0, T(N), U(N)) \leq \hat{G}_N(x; T(N)) ; \quad (2.33)$$

et, lorsque $E_N(x)$ est vide, on a :

$$\tilde{G}_N(x; U_0, T(N), U(N)) = \hat{G}_N(x; T(N)) \quad (2.34)$$

$$= 1 - \hat{F}_N(x; T(N)) \quad (2.35)$$

$$= 1 - \tilde{F}_N(x; U_0, T(N), U(N)). \quad (2.36)$$

On estime alors la valeur- p par

$$\tilde{p}_N(T_0) = \frac{N\tilde{G}_N(T_0) + 1}{N + 1}.$$

On peut alors énoncer la proposition suivante :

Proposition 2.4.1. Validité des tests de MC dans le cas général

Sous l'hypothèse nulle \mathcal{H}_0 , on a, pour tout $\alpha_1 \in [0, 1]$:

$$\begin{aligned} \mathbb{P} \left[\hat{G}_N(T_0) \leq \alpha_1 \right] &\leq \mathbb{P} \left[\tilde{G}_N(T_0) \leq \alpha_1 \right] = \mathbb{P} \left[\tilde{F}_N(T_0) \geq 1 - \alpha_1 \right] \\ &= \frac{\langle \alpha_1 N \rangle + 1}{N + 1} \leq \mathbb{P} \left[\hat{F}_N(T_0) \geq 1 - \alpha_1 \right] ; \end{aligned} \quad (2.37)$$

$$\mathbb{P} \left[\hat{p}_N(T_0) \leq \alpha \right] \leq \mathbb{P} \left[\tilde{p}_N(T_0) \leq \alpha \right] = \frac{\langle \alpha(N + 1) \rangle}{N + 1}, \quad \forall \alpha \in [0, 1]. \quad (2.38)$$

En remarquant que $\widehat{G}_N(T_0) = \widetilde{G}_N(T_0)$ presque sûrement lorsqu'il n'y a pas de nœuds, on peut considérer la **Proposition 2.4.1** comme une généralisation de la **Proposition 2.3.1**.

Preuve.

Comme les paires $(T_i, U_i)_{i=0}^N$ sont presque sûrement distinctes deux à deux, on a :

$$\begin{aligned}\widetilde{G}_N(T_0) &= \frac{1}{N} \sum_{i=1}^N 1_{[(T_i, U_i) \geq (T_0, U_0)]} \\ &= 1 - \frac{1}{N} \sum_{i=1}^N 1_{[(T_i, U_i) \leq (T_0, U_0)]} \\ &= 1 - \frac{1}{N} \left\{ -1 + \sum_{i=0}^N 1_{[(T_i, U_i) \leq (T_0, U_0)]} \right\} \\ &= \frac{N+1 - \widetilde{R}_0}{N} \text{ presque sûrement,}\end{aligned}$$

où, \widetilde{R}_0 est le rang du couple (T_0, U_0) dans $((T_i, U_i))_{i=0}^N$ par rapport à l'ordre lexicographique. Il vient :

$$\begin{aligned}\mathbb{P} \left[\widetilde{G}_N(T_0) \leq \alpha_1 \right] &= \mathbb{P} \left[\frac{N+1 - \widetilde{R}_0}{N} \leq \alpha_1 \right] \\ &= \mathbb{P} \left[\frac{\widetilde{R}_0}{N+1} \geq \frac{(1 - \alpha_1)N + 1}{N+1} \right].\end{aligned}$$

D'autre part, pour tout $\alpha_1 \in [0, 1]$, on a $\frac{(1 - \alpha_1)N + 1}{N+1} \in \left[\frac{1}{N+1}, 1 \right]$. Il suffit d'appliquer (2.15) du **Lemme 2.3.1** et on a :

$$\mathbb{P} \left[\widetilde{G}_N(T_0) \leq \alpha_1 \right] = \frac{\langle \alpha_1 N \rangle + 1}{N+1}.$$

Comme les paires $((T_i, U_i))_{i=0}^N$ sont presque sûrement distinctes deux à deux, on a aussi

$$\mathbb{P} \left[\widetilde{G}_N(T_0) \leq \alpha_1 \right] = \mathbb{P} \left[\widetilde{F}_N(T_0) \geq 1 - \alpha_1 \right].$$

On applique ensuite les inégalités de la **Remarque 2.4.3** et il vient :

$$\begin{aligned}\mathbb{P} \left[\widehat{G}_N(T_0) \leq \alpha_1 \right] &\leq \mathbb{P} \left[\widetilde{G}_N(T_0) \leq \alpha_1 \right] = \mathbb{P} \left[\widetilde{F}_N(T_0) \geq 1 - \alpha_1 \right] \\ &\leq \mathbb{P} \left[\widehat{F}_N(T_0) \geq 1 - \alpha_1 \right]\end{aligned}$$

d'où (2.37).

Par ailleurs ,

$$\tilde{p}_N(T_0) \equiv \frac{N\tilde{G}_N(T_0) + 1}{N + 1} \leq \alpha \Leftrightarrow \tilde{G}_N(T_0) \leq \frac{\alpha(N + 1) - 1}{N}.$$

Et d'après (2.37),

$$\begin{aligned} \mathbb{P}[\tilde{p}_N(T_0) \leq \alpha] &= \mathbb{P}\left[\tilde{G}_N(T_0) \leq \frac{\alpha(N + 1) - 1}{N}\right] \\ &= \begin{cases} 0 & \text{si } \alpha < \frac{1}{N + 1} \\ \frac{\langle \alpha(N + 1) - 1 \rangle + 1}{N + 1} = \frac{\langle \alpha(N + 1) \rangle}{N + 1} & \text{si } \frac{1}{N + 1} \leq \alpha \leq 1 \end{cases} \end{aligned}$$

d'où (2.38) en observant que $\langle \alpha(N + 1) \rangle = 0$ si $0 \leq \alpha \leq \frac{1}{N + 1}$ ■

2.5 Algorithme des tests de Monte Carlo

Soit T , la statistique de test choisie de l'hypothèse nulle (2.1). La procédure du test de Monte Carlo se réalise en 4 étapes :

1. Calculer la valeur T_0 prise par la statistique de test T sur les observations ;
2. Réaliser N tirages indépendants T_1, T_2, \dots, T_N de T sous \mathcal{H}_0 ;
3. Si on note R_0 le rang (aléatoire en cas d'ex-æquo) de T_0 dans T_0, T_1, \dots, T_N , une estimation de la valeur- p associée au test de Monte Carlo est donnée par :
 - $\hat{p}_N(T_0) = 1 - \frac{R_0 - 1}{N + 1}$ pour un test unilatéral à droite (cf (2.6)) ;
 - $\hat{p}_N(T_0) = \frac{R_0}{N + 1}$ pour un test unilatéral à gauche (cf (2.7)).
4. Prendre une décision en fonction du seuil α choisi.

Dans la pratique, à l'étape 2, l'échantillon T_1, T_2, \dots, T_N de T s'obtient de la manière suivante :

- 2.1 Réaliser N permutations aléatoires des observations agglomérées

$$(x_{11}, x_{12}, \dots, x_{1n}; x_{21}, x_{22}, \dots, x_{2m}).$$

- 2.2 Prendre les n premiers éléments de chaque permutation comme premier échantillon, et les m derniers comme deuxième échantillon.
- 2.3 Calculer les valeurs T_1, T_2, \dots, T_N prises par T sur ces permutations.

Le code R de la version permutationnelle des tests de Monte Carlo se trouve en annexe à la fin du document. Nous avons conçu ce code à l'aide de la version 2.3.0 de R , qu'on peut télécharger gratuitement : www.r-project.org/.

R est un langage de programmation interactif interprété et orienté objet contenant une très large collection de méthodes statistiques et des facilités graphiques importantes.

2.6 Estimation du nombre de permutations N

Nous supposons toujours que la statistique T du test de l'hypothèse nulle (2.1) est telle que le test soit unilatéral à droite. Notons F la fonction de répartition de T sous \mathcal{H}_0 , et supposons T continue.

Lorsque F est calculable par des méthodes analytiques, l'hypothèse nulle \mathcal{H}_0 est rejetée au seuil α si et seulement si la valeur observée T_0 de T vérifie :

$$T_0 > F^{-1}(1 - \alpha).$$

Plaçons nous dans un cas où F est difficile à calculer, mais simulable. On peut alors obtenir un échantillon T_1, \dots, T_N i.i.d. de T sous l'hypothèse nulle \mathcal{H}_0 , ce qui permet d'estimer le p -quantile de F par $T_{\langle(Np)\rangle}$, où $\langle x \rangle$ désigne la partie entière de x , et $T_{(1)} \leq \dots \leq T_{(N)}$ les statistiques d'ordre associées à l'échantillon T_1, \dots, T_N . D'autre part, les variables aléatoires $U_i = F(T_i)$ sont i.i.d. et suivent la loi uniforme $\mathcal{U}(0, 1)$ (cf **Proposition 1.1.1** page 6). Notons aussi qu'on a la correspondance $U_{(i)} = F(T_{(i)})$.

Théorème 2.6.1. *Soit $X_{(1)}, \dots, X_{(N)}$, les statistiques d'ordre associées à un échantillon X_1, \dots, X_N i.i.d. d'une variable aléatoire réelle continue X , de fonction de répartition F_X , et de densité de probabilité f_X .*

Alors la densité de probabilité de $X_{(i)}$ est donnée par :

$$f_{X_{(i)}}(x) = \frac{N!}{(i-1)!(N-i)!} f_X(x) [F_X(x)]^{i-1} [1 - F_X(x)]^{N-i}. \quad (2.39)$$

Dans le cas particulier où X suit $\mathcal{U}(0, 1)$,

$$f_{X_{(i)}}(x) = \frac{N!}{(i-1)!(N-i)!} x^{i-1} (1-x)^{N-i} 1_{]0, 1[}(x). \quad (2.40)$$

i.e. $X_{(i)}$ suit la loi bêta de paramètre $a = i$ et $b = N - i + 1$.

Rappelons la densité de la loi bêta de paramètres a et b :

$$f(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} 1_{]0, 1[}(x),$$

où

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt.$$

Dans le cas où a et b sont des entiers naturels, ceci donne :

$$f(x; a, b) = \frac{(a+b-1)!}{(a-1)!(b-1)!} x^{a-1} (1-x)^{b-1} 1_{]0, 1[}(x).$$

Il vient que : $F(T_{(\langle pN \rangle)})$ suit la loi bêta $\beta(\langle pN \rangle, N - \langle pN \rangle + 1)$.

• **Valeur minimale de N pour que l'estimation $T_{(\langle pN \rangle)}$ vérifie l'inégalité (2.41) ci-dessous :**

Posons $\alpha = 5\%$ et $p = 1 - \alpha$, et déterminons le nombre minimal N de simulations de T , pour que

$$\mathbb{P}[F^{-1}(p - \varepsilon) \leq T_{(\langle pN \rangle)} \leq F^{-1}(p + \varepsilon)] \geq 0.95, \quad (2.41)$$

pour une valeur fixée de ε .

Remarquons que :

$$\begin{aligned} (2.41) &\Leftrightarrow \mathbb{P}[p - \varepsilon \leq F(T_{(\langle pN \rangle)}) \leq p + \varepsilon] \geq 0.95 \\ &\Leftrightarrow \mathbb{P}[p - \varepsilon \leq \beta(\langle pN \rangle, N - \langle pN \rangle + 1) \leq p + \varepsilon] \geq 0.95. \end{aligned}$$

L'estimation de N se fait de la manière suivante : pour différentes valeurs de N , nous calculons

$$P_r = \mathbb{P}[p - \varepsilon \leq \beta(\langle pN \rangle, N - \langle pN \rangle + 1) \leq p + \varepsilon],$$

et nous faisons croître N jusqu'à ce que $P_r \geq 0.95$ (cf la fonction *estim.N* du document **ANNEXE**).

Bradley [11] recommande de prendre $\varepsilon = 0.005$. Pour cette valeur de ε , il est nécessaire de choisir le nombre de simulations $N \geq 7\,400$. Nous pensons qu'on peut faire un compromis en prenant $\varepsilon = 0.0061$, de sorte que le nombre minimal nécessaire de permutations dans les tests de Monte Carlo soit $N = 5\,000$, ce qui permet de réduire le temps machine nécessaire pour effectuer un test de Monte Carlo basé sur des permutations aléatoires des observations entre les deux échantillons.

2.7 Quelques statistiques de test

Définition 2.7.1. Statistique de Kolmogorov-Smirnov (KS)

La statistique de Kolmogorov-Smirnov KS est définie par :

$$KS = \sup_x \left| \widehat{F}_{1n}(x) - \widehat{F}_{2m}(x) \right| \quad (2.42)$$

D'après Jean-Marie Dufour et Abdeljeh Farhat [9], la statistique KS est libre sous \mathcal{H}_0 lorsque les distributions des échantillons sont continues, mais sa distribution exacte ou asymptotique n'est pas standard. KS n'est pas libre dans le cas où les distributions sont discrètes. Il peut donc y avoir une perte de puissance du test si l'éventuelle nature discrète des distributions n'est pas prise en compte.

Définition 2.7.2. Cramer-Von Mises CM

La statistique de test de Cramer-Von Mises est définie par :

$$CM = \frac{mn}{(m+n)^2} \left\{ \sum_{i=1}^n \left[\widehat{F}_{1n}(X_{1i}) - \widehat{F}_{2m}(X_{1i}) \right]^2 + \sum_{j=1}^m \left[\widehat{F}_{1n}(X_{2j}) - \widehat{F}_{2m}(X_{2j}) \right]^2 \right\} \quad (2.43)$$

La distribution exacte ou asymptotique de CM sous \mathcal{H}_0 n'est pas standard [9].

Définition 2.7.3. Les statistiques de tests \widehat{L}_1 , \widehat{L}_2 et \widehat{L}_∞

Ces statistiques sont définies par :

$$\widehat{L}_1 = \sum_{i=1}^n \left| \widehat{f}_{1n}(X_{1i}) - \widehat{f}_{2m}(X_{1i}) \right| + \sum_{j=1}^m \left| \widehat{f}_{1n}(X_{2j}) - \widehat{f}_{2m}(X_{2j}) \right| \quad (2.44)$$

$$\widehat{L}_2 = \left\{ \sum_{i=1}^n \left[\widehat{f}_{1n}(X_{1i}) - \widehat{f}_{2m}(X_{1i}) \right]^2 + \sum_{j=1}^m \left[\widehat{f}_{1n}(X_{2j}) - \widehat{f}_{2m}(X_{2j}) \right]^2 \right\}^{\frac{1}{2}} \quad (2.45)$$

$$\widehat{L}_\infty = \sup_x \left| \widehat{f}_{1n}(x) - \widehat{f}_{2m}(x) \right| \quad (2.46)$$

$$= \max_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}} \left\{ \left| \widehat{f}_{1n}(X_{1i}) - \widehat{f}_{2m}(X_{1i}) \right|, \left| \widehat{f}_{1n}(X_{2j}) - \widehat{f}_{2m}(X_{2j}) \right| \right\} \quad (2.47)$$

où \widehat{f}_{1n} et \widehat{f}_{2m} sont les estimateurs par la méthode du noyau des densités de probabilité associées à \widehat{F}_1 et \widehat{F}_2 respectivement lorsque les distributions sont continues.

Remarque 2.7.1. \hat{f}_{1n} et \hat{f}_{2m} ne sont pas bien définis dans le cas de distributions discrètes. Lorsque les distributions ne sont pas continues, on les remplace par les fonctions de masse ; Et dans ce cas, les statistiques de test \hat{L}_1 , \hat{L}_2 et \hat{L}_∞ restent bien définies et peuvent toujours être utilisées comme statistiques de test.

Définition 2.7.4. Fonction de masse

On appelle «fonction de masse» associée à un échantillon X_1, X_2, \dots, X_n la fonction notée f_{masse} définie par $f_{masse}(x) = \frac{1}{n} \sum_{i=1}^n 1_{[x=X_i]}$, i.e. l'image par cette fonction d'un réel x est la proportion de ce réel parmi X_1, X_2, \dots, X_n .

Jean Marie DUFOUR et Abdeljeh Farhat [9] propose d'utiliser les estimateurs de densité par la méthode du noyau suivants :

$$\hat{f}_{1n}(x) = \frac{C_1}{n} \sum_{i=1}^n K [C_1(x - X_{1i})] , \quad \hat{f}_{2m}(x) = \frac{C_2}{m} \sum_{j=1}^m K [C_2(x - X_{2j})] \quad (2.48)$$

avec

$$C_1 = \begin{cases} \frac{n^{\frac{1}{5}}}{2S_1} & \text{si } S_1 \neq 0 \\ 1 & \text{si } S_1 = 0 \end{cases}$$

$$C_2 = \begin{cases} \frac{m^{\frac{1}{5}}}{2S_2} & \text{si } S_2 \neq 0 \\ 1 & \text{si } S_2 = 0 \end{cases}$$

où

$$S_1 = \left[\frac{1}{n-1} \sum_{i=1}^n (X_{1i} - \bar{X}_1)^2 \right]^{\frac{1}{2}}$$

$$S_2 = \left[\frac{1}{m-1} \sum_{j=1}^m (X_{2j} - \bar{X}_2)^2 \right]^{\frac{1}{2}}$$

$$K(x) = \begin{cases} \frac{1}{2} & \text{si } |x| \leq 1 \\ 0 & \text{sinon} \end{cases}$$

D'autres statistiques de test sont :

$$\hat{\theta}_1 = |\bar{X}_1 - \bar{X}_2| \quad (2.49)$$

$$\hat{\theta}_2 = |S_1^2 - S_2^2| \quad (2.50)$$

$$\hat{\theta}_3 = \left| \frac{1}{n} \sum_{i=1}^n \left(\frac{X_{1i} - \bar{X}_1}{S_1} \right)^3 - \frac{1}{m} \sum_{j=1}^m \left(\frac{X_{2j} - \bar{X}_2}{S_2} \right)^3 \right| \quad (2.51)$$

$$\hat{\theta}_4 = \left| \frac{1}{n} \sum_{i=1}^n \left(\frac{X_{1i} - \bar{X}_1}{S_1} \right)^4 - \frac{1}{m} \sum_{j=1}^m \left(\frac{X_{2j} - \bar{X}_2}{S_2} \right)^4 \right| \quad (2.52)$$

Par convention, si $S_k = 0$, on pose $\frac{X_{ki} - \bar{X}_k}{S_k} = 0$ pour tout i , parce que dans ce cas, $X_{k1} = X_{k2} = \dots$

2.8 Combinaison des tests de MC standardisés

Il s'agit de combiner des statistiques de test différentes dans l'espoir d'améliorer la puissance du test. La standardisation a pour but d'assurer la compatibilité entre différentes statistiques de test et consiste tout simplement à soustraire la moyenne empirique de chaque statistique et à diviser ensuite par l'écart-type. Formellement, soient $V = (T^{(1)}, T^{(2)}, \dots, T^{(k)})'$ un vecteur de k statistiques de test sélectionnées, $V_0 = (T_0^{(1)}, T_0^{(2)}, \dots, T_0^{(k)})'$ la valeur de V calculée à partir des observations $(x_{11}, x_{12}, \dots, x_{1n}; x_{21}, x_{22}, \dots, x_{2m})$, et $V_i = (T_i^{(1)}, T_i^{(2)}, \dots, T_i^{(k)})'$, $i = 1, 2, \dots, N$ celles calculées à partir de N permutations des observations.

Les statistiques standardisées sont données par :

$$\tilde{T}_i^{(j)} = \frac{T_i^{(j)} - \bar{T}^{(j)}}{\sigma_j} \quad j = 1, \dots, k, \quad i = 0, 1, \dots, N, \quad (2.53)$$

où

$$\begin{cases} \bar{T}^{(j)} = \frac{1}{N+1} \sum_{i=0}^N T_i^{(j)} \\ \sigma_j = \left\{ \frac{1}{N} \sum_{i=0}^N [T_i^{(j)} - \bar{T}^{(j)}]^2 \right\}^{\frac{1}{2}} \end{cases} \quad (2.54)$$

Nous supposons toujours que le test est unilatéral à droite. Pour la valeur V_0 du vecteur V sur les observations, et les valeurs simulées V_1, V_2, \dots, V_N , on peut calculer les statistiques combinées suivantes :

$$\mathcal{Q}(V_i) = \max_{1 \leq j \leq k} \left\{ \tilde{T}_i^{(j)} \right\} \quad i = 0, 1, \dots, N, \quad (2.55)$$

$$\mathcal{Q}_a(V_i) = \max_{1 \leq j \leq k} \left\{ \left| \tilde{T}_i^{(j)} \right| \right\} \quad i = 0, 1, \dots, N. \quad (2.56)$$

Le test combiné basé sur la statistique \mathcal{Q} rejette l'hypothèse nulle \mathcal{H}_0 lorsque le maximum des statistiques standardisées est "grand", alors que celui basé sur \mathcal{Q}_a rejette \mathcal{H}_0 lorsque le maximum des valeurs absolues des statistiques standardisées est "grand".

Notons

$$\mathcal{Q}_i \equiv \mathcal{Q}(V_i), \quad i = 0, 1, \dots, N$$

et

$$\mathcal{Q}_{ai} \equiv \mathcal{Q}_a(V_i), \quad i = 0, 1, \dots, N.$$

Les réels \mathcal{Q}_0 et \mathcal{Q}_{a0} représentent les valeurs des statistiques de test associées aux observations. Les réels \mathcal{Q}_i et \mathcal{Q}_{ai} , $i \neq 0$, peuvent être interprétés comme des valeurs basées sur des permutations aléatoires des observations. On peut alors remarquer que les

$$\mathcal{Q}_i, \quad i = 0, 1, \dots, N, \text{ sont échangeables sous } \mathcal{H}_0, \quad (2.57)$$

et

$$\mathcal{Q}_{ai}, \quad i = 0, 1, \dots, N, \text{ sont échangeables sous } \mathcal{H}_0. \quad (2.58)$$

On en déduit alors : On peut donc leur appliquer la théorie précédente sur les tests de Monte Carlo.

2.9 Comparaison

Il s'agit de comparer, par des simulations, les puissances des tests de Monte Carlo basés sur certaines des statistiques définies plus haut, à celles de certains tests classiques (Kolmogorov-Smirnov, Wilcoxon-Mann Withney, Student).

2.9.1 Distributions continues

Le premier échantillon est simulé sous la loi $\mathcal{N}(0, 1)$ et le second sous $\mathcal{N}(\mu, \sigma)$, μ ou σ étant destinées à varier selon qu'on considère une différence par translation (shift) ou par dispersion autour de la moyenne. La taille du premier échantillon est fixé à $n = 10$ et celui du second à $m = 12$. Nous avons fixé le nombre de permutations à $N = 5\,000$ pour les tests de Monte Carlo.

La construction de la courbe de puissance d'un test s'est faite de la manière suivante :

Différence par les paramètres de position

1. Nous avons fixé $\sigma = 1$, et fait varier μ dans l'intervalle $[0; 3]$ avec un pas de $h = 0.2$.
2. Pour chaque valeur de μ , nous avons réalisé 1 000 simulations des deux échantillons, ce qui nous a permis d'effectuer 1 000 tests.
3. La puissance est estimée, pour chaque valeur de μ , par la proportion des tests ayant rejeté l'hypothèse nulle (2.1) au seuil 5 %.

D'après la Figure 2.1, le test de Monte Carlo basé sur la statistique de Kolmogorov-Smirnov a sensiblement le même niveau et la même puissance que celui de Kolmogorov-Smirnov classique, dans une situation où les deux échantillons sont normaux de même variance. Nous remarquons aussi que le test de Student est le plus puissant dans cette situation, ce qui est tout à fait prévisible dans la mesure où les tests paramétriques sont en général plus puissants que leurs homologues non paramétriques, lorsque les conditions d'application sont remplies. Il faut aussi noter que le test de Wilcoxon est bien adapté aux cas où on peut faire une hypothèse de translation sur les observations d'un des échantillons par rapport à celles de l'autre.

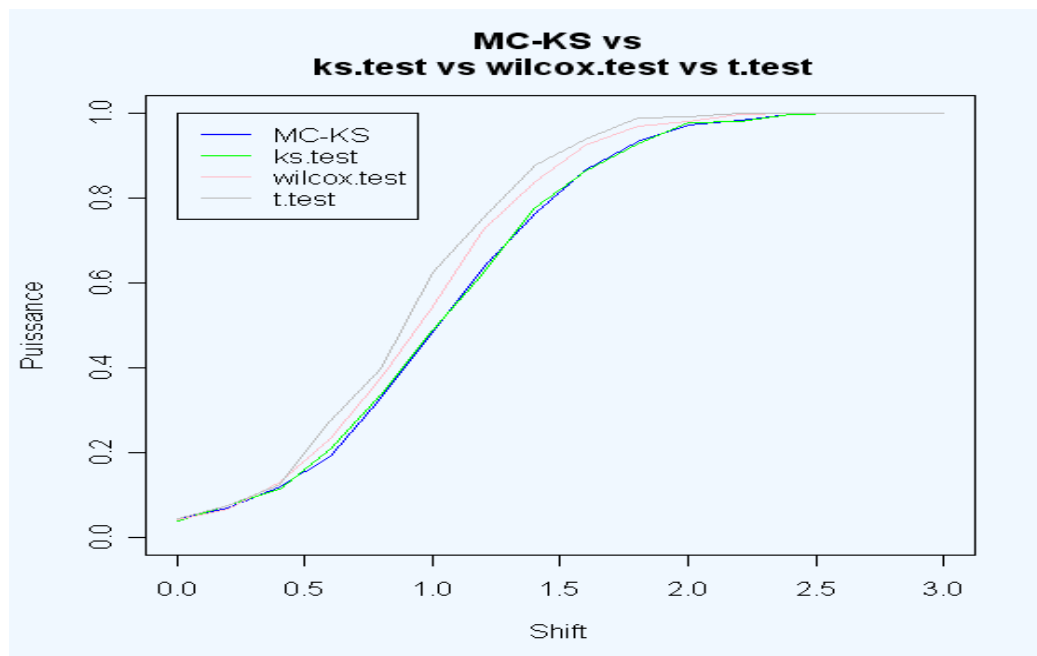


FIG. 2.1 – Tests de MC basés sur KS, KS classique, Wilcoxon et Student dans le cas d'un échantillon gaussien en fonction du shift sur la moyenne.

Différence par les paramètres de dispersion

1. Nous avons fixé $\mu = 0$, et fait varier σ dans l'intervalle $[1; 5]$ avec un pas de $h = 0.4$.
2. Comme dans le cas du shift, nous avons effectué 1 000 simulations des deux échantillons, ce qui nous a permis de réaliser 1 000 tests.
3. La puissance est estimée, pour chaque valeur de σ , par la proportion des tests ayant rejeté l'hypothèse nulle (2.1) au seuil 5 %.

Dans le cas où les deux échantillons ont la même moyenne et des variances différentes, les tests de Kolmogoro-Smirnov classique et de Monte Carlo basé sur la statistique de Kolmogorov-Smirnov sont nettement plus puissants que ceux de Wilcoxon et de Student (cf Figure 2.2). La faible puissance du test de Student s'explique par le fait qu'il est conçu pour tester l'égalité des moyennes, et celle de Wilcoxon au fait qu'une hypothèse de translation sur les données, dans ce cas précis, serait trop forte.

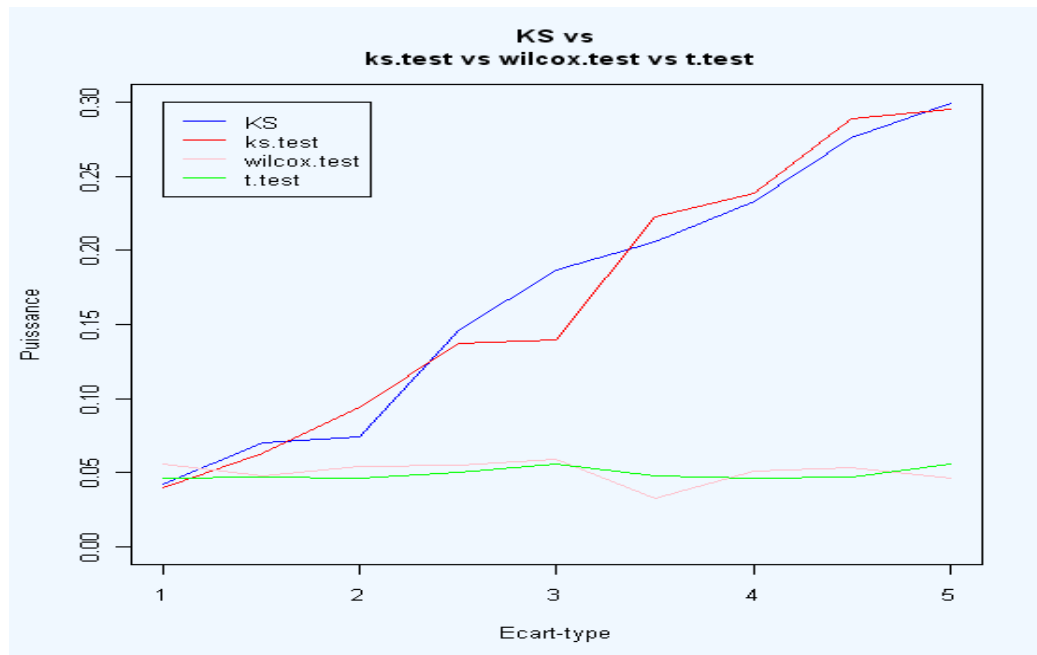


FIG. 2.2 – Puissances dans le cas où l'écart-type varie.

2.9.2 Distributions discrètes

Dans le cas des distributions discrètes, nous avons considéré les lois binomiales négatives $[Nbin(N, p)]$, binomiales $[B(n, p)]$, uniformes discrètes $[UD(n)]$, géo-

métriques [$Geo(p)$] et de Poisson [$P(\lambda)$]. Nous avons distingué trois cas :

1. Distributions ayant la même moyenne et des variances différentes : $UD(19)$, $B(20, 0.5)$, $Geo(0.1)$, $Nbin(8, 0.2)$ et $P(10)$. Ces distributions ont la même moyenne 10 et les variances 30, 5, 90, 2.5 et 10 respectivement.
2. Distributions ayant la même variance et des moyennes différentes : $UD(10)$, $B(30, 0.5)$, $Geo\left(\frac{\sqrt{34}-1}{16.5}\right)$, $Nbin\left(3, \frac{\sqrt{108}-3}{16.5}\right)$ et $P(8.25)$. Toutes ont pour variance 8.25, et leurs moyennes sont 5.5, 16.5, 3.42, 2.23 et 8.25 respectivement.
3. Distributions différentes par leurs moyennes et par leurs variances : $UD(10)$, $B(10, 0.1)$, $Geo(0.3)$, $Nbin(10, 0.2)$ et $P(5)$. Les moyennes de ces distributions sont 5.5, 1, 3.33, 50 et 5 respectivement, et leurs variances 8.25, 0.9, 7.78, 200 et 5 respectivement.

Le seuil des tests étant fixé à 5 %, nous avons estimé le niveau ou la puissance de chaque test par la proportion des tests significatifs pour 1 000 simulations. Le nombre de permutations était $N = 5\,000$.

Note : Les résultats des Tableaux 2.1 et 2.2 sont donnés en %.

Les premières simulations montrent que le niveau des tests de Monte Carlo est parfaitement contrôlé comme prévu par la théorie. D'après les résultats obtenus dans les Tableaux 2.1 et 2.2, il apparaît que la version permutationnelle des tests de Monte Carlo est mieux adaptée aux données discrètes, que les tests non paramétriques de Kolmogorov-Smirnov classique et de Wilcoxon-Mann Withney. La puissance de ces derniers est en plus très sensible aux répétitions dans les données. On note aussi une amélioration globale de la puissance lorsque le test de Monte Carlo est basé sur la combinaison des statistiques KS , CM , $\hat{\theta}_1$ et $\hat{\theta}_2$.

$F_1 = UD$

F_2	KS	CM	Combin	ks	wilc
UD	04.90	05.10	04.50	01.80	05.60
Bin	71.50	67.46	97	100	100
Geo	24.30	30.90	38.70	15	23.60
Nbin	100	100	100	100	100
Pois	23.80	20.24	64.70	10.90	05.10

 $F_1 = Bin$

F_2	KS	CM	Combin	ks	wilc
Bin	05.10	04.87	04.10	02	04.10
Geo	88.80	50.20	95.40	74.90	78.20
Nbin	100	100	100	100	100
Pois	07.10	06.50	22.20	22.00	21.15

 $F_1 = Geo$

F_2	KS	CM	Combin	ks	wilc
Geo	03.50	04.25	05.45	02.20	04.20
Nbin	100	100	100	100	100
Pois	65.70	60.75	60.57	43	34.10

 $F_1 = Nbin$

F_2	KS	CM	Combin	ks	wilc
Nbin	04.80	03.95	03.98	02.10	04.10
Pois	100	100	100	100	100

 $F_1 = Pois$

F_2	KS	CM	Combin	ks	wilc
Pois	04.80	04.20	04.66	04.10	04.70

TAB. 2.1 – Puissances et niveaux des tests lorsque les moyennes égales et les variances différentes. KS désigne le test de MC basé sur la statistique de KS, CM le test de MC basé sur la statistique de CM, Combin le test de MC basé sur la combinaison des statistiques KS, CM, $\hat{\theta}_1$ et $\hat{\theta}_2$, ks le test de KS classique et wilc le test de Wilcoxon.

$$F_1 = UD$$

F_2	KS	CM	Combin	ks	wilc
UD	04.50	04.53	03.10	02.20	06.00
Bin	97.20	98.20	97.58	92.20	99.30
Geo	89.20	75.90	78.23	77.00	94.00
Nbin	98.70	99.10	98.25	94.10	98.90
Pois	63.10	61.90	65.24	38.60	72.90

TAB. 2.2 – Variances égales et moyennes différentes.

APPLICATIONS : EFFET DE CERTAINS GÈNES SUR LE DÉVELOPPEMENT DE *P. FALCIPARUM* CHEZ *A.* *GAMBIAE*

Pour être transmis d'un hôte vertébré à l'autre, le *Plasmodium* doit compléter un cycle biologique complexe chez l'*Anophèle* femelle, et au cours de ce cycle, le parasite subit de nombreuses pertes. Des études récentes ont montré l'influence du système immunitaire du moustique sur le développement du parasite [2]. La réaction immunitaire est activée lorsque des molécules d'origine microbienne ou parasitaire sont détectées et reconnues comme étrangères. Ce stade de reconnaissance implique des récepteurs de reconnaissance de motifs de l'hôte, en anglais "pattern recognition receptors" (PRRs) [5]. Certains des gènes dont il est question ici font partie des PRRs. Il s'agit de LRIM1 (Leucine-rich repeat protein) et de deux "C-type lectin" CTL4 et CTLMA2. Ces PRRs ont déjà montré leur effet dans l'interaction *P. berghei-A. gambiae*, le gène LRIM1 inhibant le développement du parasite au contraire de CTL4 et CTLMA2 qui le favorisent [4]. Les autres gènes sont : *Serpine 2* (SRPN2) et l'*Apolipoporphine I* (APO1). La SRPN2 est une molécule qui régule les voies de signalisation de la réponse immunitaire et l'APO1 est une lipoprotéine également associée à la reconnaissance de motifs de l'agent pathogène.

3.1 Quelques termes-clef

Nous donnons dans cette section quelques termes-clef de la biologie moléculaire et leur définition.

Protéine

C'est l'un des quatre matériaux de base de tout organisme, avec les glucides, les lipides, les lipides et les acides nucléiques.

Acide désoxyribonucléique (ADN)

C'est le support biochimique de l'information génétique chez tous les êtres vivants (à l'exception de quelques virus qui utilisent l'*ARN*). L'*ADN* se présente le plus souvent sous forme de deux longs filaments (ou chaînes) torsadés l'un dans l'autre pour former une structure en double hélice.

Acide ribonucléique (*ARN*)

Dans les cellules, on distingue plusieurs types d'*ARN* suivant leur fonction. Les trois principaux types d'*ARN* sont : les *ARN messagers*, les *ARN de transfert* et les *ARN ribosomiaux*. L'*ARN* est un *acide nucléique* constitué d'une seule chaîne de nucléotides, de structure analogue à celle de l'*ADN*. L'*ARN* est produit par *transcription* de l'*ADN*.

ARN messenger (*ARNm*)

C'est une photocopie d'une séquence d'*ADN*. Il sert à transférer l'information génétique de son lieu de stockage (le *chromosome*) jusqu'au lieu de synthèse des protéines (les *ribosomes*).

ARN ribosomal (*ARNr*)

C'est le constituant principal des ribosomes, la machinerie cellulaire où a lieu la *traduction* en *protéines* de l'information contenue dans les *ARNm*.

ARN de transfert (*ARNt*)

Les *ARNt* sont des petits *ARN* responsables du transport des acides aminés jusqu'aux *ribosomes* lors de la *traduction* des *ARNm* : chaque *ARNt* transporte un acide aminé, de façon spécifique.

Code génétique

C'est le système de correspondance permettant de traduire une séquence d'acide nucléique en protéine.

Fonctions d'une protéine

Les fonctions remplies par les protéines sont très variées et permettent de classer les protéines :

- les *protéines de structure* qui sont comparables à des briques cellulaires ;
- les *protéines de transport* qui sont chargées du transport d'autres molécules dans la cellule ou entre les cellules d'un organisme ;
- les *enzymes* qui permettent d'accélérer les réactions chimiques nécessaires à la vie ;
- les *protéines de l'immunité* ou *anticorps* qui contribuent à la défense de l'organisme.

Gène

C'est un fragment d'*ADN* portant les informations nécessaires à la fabrication d'une ou de plusieurs protéines.

RNAi ou ARNi

Afin de pouvoir déterminer la fonction d'un gène dans les interactions avec le *Plasmodium*, on procède à l'inactivation de l'expression de ce gène par *interférence ARN* noté *ARNi* ou *RNAi*. Le principe consiste à inoculer de l'*ARN* double brin (*ARNdb*) dans le corps du moustique ; cet *ARNdb* va être synthétisé dans l'organisme, les petits fragments *siRNA*, iront alors spécifiquement s'hybrider sur les *ARNm* du gène à inhiber, empêchant ainsi sa traduction.

3.2 Cycle biologique du *Plasmodium*

Le *Plasmodium* a un cycle biologique qui se déroule en trois phases :

- 1- le cycle **sporogonique** (FIG. 3.2) qui se déroule chez le moustique Anophèle ;
- 2- le cycle **exo-érythrocytaire** qui se déroule hors des cellules sanguines d'un hôte vertébré et
- 3- le cycle **érythrocytaire** qui se déroule dans les cellules sanguines d'un hôte vertébré.

Si le moustique ingère des *gamétocytes* de *P. falciparum* lors d'un repas sanguin, ils forment des *gamètes* dans l'estomac de l'insecte. Les gamètes mâle et femelle, issus des gamétocytes mâles et femelles respectivement, s'unissent pour

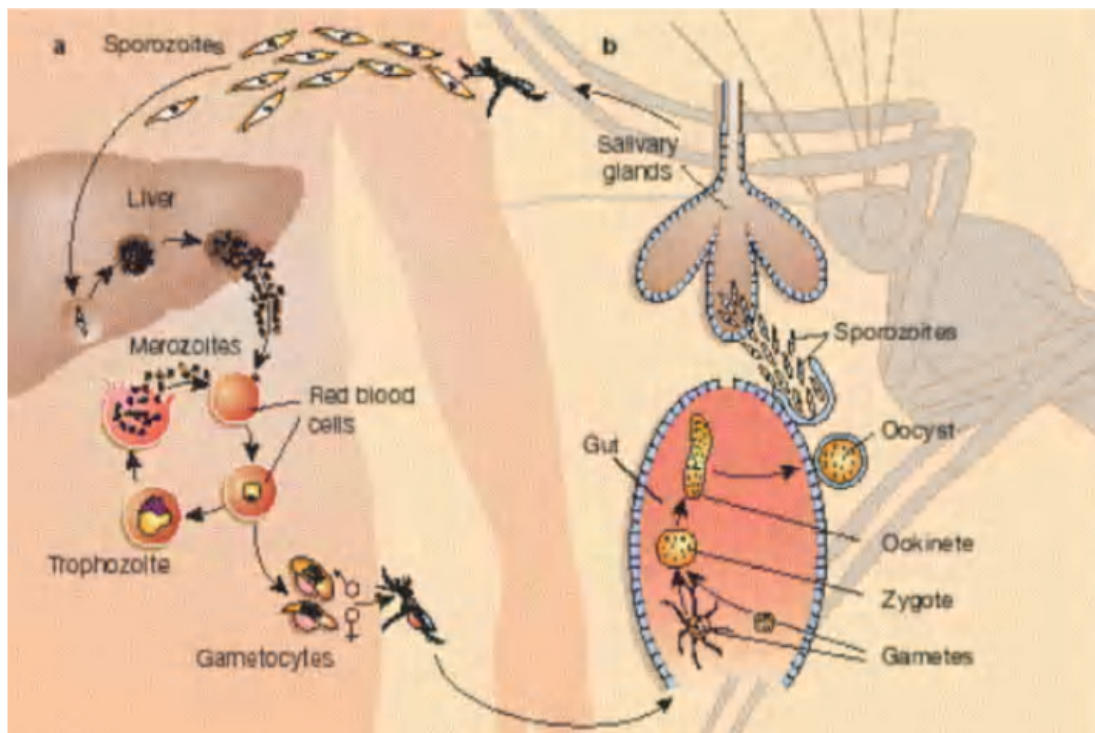


FIG. 3.1 – Cycle biologique du *Plasmodium* [1].

former un zygote mobile appelé *ookinete*. L'*ookinete* pénètre la paroi de l'estomac du moustique et devient un *oocyste* sphérique. A l'intérieur de l'*oocyste*, le noyau se divise à répétition, un grand nombre de *sporozoïtes* est formé et l'*oocyste* grossit. Quand les *sporozoïtes* sont complètement développés, l'*oocyste* se rompt, les libérant dans la cavité générale du corps du moustique. Ils migrent alors vers les glandes salivaires, et peuvent ainsi être transmis à un hôte vertébré lors du prochain repas sanguin du moustique (FIG. 3.2). Le parasite subit de nombreuses pertes pendant le cycle sporogonique, en particulier lors de la transition *ookinètes*-*oocystes* (FIG. 3.2).

3.3 Matériel et méthodes

3.3.1 Collecte des données

Les moustiques utilisés dans cette étude sont des *Anophèles gambiae* femelles provenant de l'insectarium du laboratoire d'entomologie du Paludisme de l'OCEAC. Cette colonie d'*A. gambiae* a été mise en place en 1988 à partir d'échantillons prélevés dans les banlieues de la ville de Yaoundé. Les moustiques sont adaptés à se nourrir à travers une membrane de parafilm, et ils sont main-

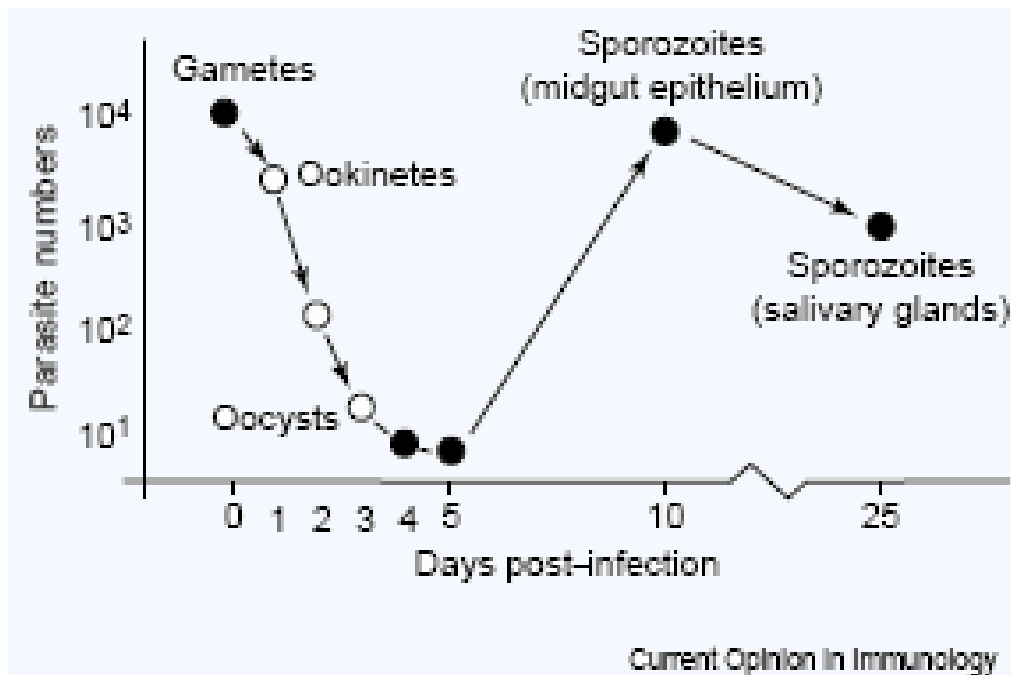


FIG. 3.2 – Perte et amplification du parasite dans le moustique [2].

tenus dans les conditions standards de température et d'humidité. La souche de Yaoundé est représentative du système de transmission locale et naturelle du paludisme dans la région Sud du Cameroun.

L'inactivation de l'expression des gènes concernés par l'étude a été réalisée par injection d'acide ribonucléique double brins (*ARN-db*) dans le thorax des femelles âgées d'un jour [7]. Sept replicats ont été réalisés et, dans chacun d'eux, 60 à 100 femelles ont été injectés avec l'*ARN-db* de chaque gène à tester ou de l'*ARN-db GFP* (Green Fluorescent Protein) pour servir de référence. En effet, le *GFP* est une molécule neutre fluorescente qui permet de vérifier si l'expérimentation fonctionne correctement. L'injection de *GFP* au groupe de référence permet de mettre tous les moustiques dans les mêmes conditions pour éliminer l'effet possible dû au fait d'injecter une aiguille dans le moustique. La réduction effective de l'expression génétique a été vérifiée 4 jours après l'injection en quantifiant par *RT-PCR* l'*ARNm* chez les moustiques ayant reçu de l'*ARNdb*. La *réaction en chaîne par polymérase* (*PCR* en anglais pour *Polymerase Chain Reaction*), est une méthode de biologie moléculaire permettant d'amplifier le nombre de copies d'une séquence spécifique d'*ADN*, même si la quantité initiale est très faible. La méthode de *PCR* est également employée pour doser des *ARNm* et on parle alors de *transcription inverse-PCR* (*RT-PCR*).

Pour l'infection expérimentale, les porteurs de gamétocytes ont été recrutés parmi les enfants de 5 à 12 ans dans les écoles de Mfou, une localité située à une trentaine de kilomètres de la ville de Yaoundé, dans une zone endémique du paludisme. Les enfants présentant une gamétocytemie positive étaient pris comme volontaires après que leurs tuteurs légaux aient signé des papiers de consentement. Ces procédures de recrutement ont été approuvées par le comité de révision et d'éthique du Cameroun et de l'OMS.

Quatre jours après l'injection des *ARNdb*, tous les moustiques femelles d'un même replicat ont été nourris du sang d'un même porteur de gamétocytes. Pour limiter les effets du facteur humain, le serum du sang des porteurs a été remplacé par celui d'un donneur non-immunisé. Les moustiques privés de nourriture pendant 16 heures ont été alors nourris pendant 15 minutes, les moustiques non gorgés étaient retirés de l'étude. Sept jours après le repas de sang, les moustiques ont été dissequés et les oocystes contenus dans l'estomac comptés. L'expérience était considérée comme réussie lorsque la prévalence des moustiques infectés (au moins un oocyste) était d'au moins 50 % dans le groupe contrôle (*GFP*).

3.3.2 Description des données

Variables

Les données de cette étude présentent, d'un côté, le gène inhibé, et de l'autre, un décompte d'oocystes 7 jours après l'infection expérimentale des moustiques, soit :

1. Genes : *GFP*, *APO1*, *CTL4*, *CTLMA2*, *LRIM1* et *SRPN2*.
2. Nbre.ooc : C'est la variable indiquant le nombre d'oocystes de chaque moustique 7 jours après l'infection par des gamétocytes.

Les effectifs

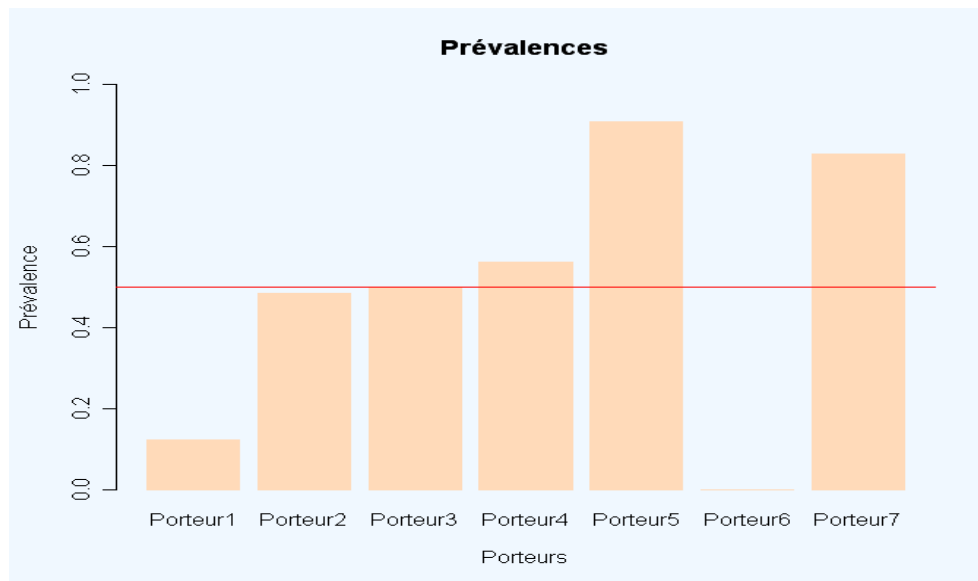
On avait un total de 976 moustiques gorgés répartis comme l'indique le Tableau 3.1.

Rappelons que l'infection était considérée comme réussie lorsque la prévalence dans le groupe contrôle était au moins de 50 %. De ce fait, on a exclu de l'étude les moustiques ayant été nourris par le sang des porteurs 1 et 6 (FIG 3.3). En plus des observations associées aux porteurs 3, 4, 5 et 7 dont les prévalences

Porteurs	<i>GFP</i>	<i>CTL4</i>	<i>CTLMA2</i>	<i>APO1</i>	<i>LRIM1</i>	<i>SRPN2</i>	Totaux
<i>Porteur1</i>	41	28	29	18	30	27	173
<i>Porteur2</i>	35	34	20	23	13	13	138
<i>Porteur3</i>	18	31	13	31	13	11	117
<i>Porteur4</i>	32	18	30	24	25	20	149
<i>Porteur5</i>	22	27	24	26	14	26	139
<i>Porteur6</i>	30	0	0	0	0	0	30
<i>Porteur7</i>	93	0	0	74	0	63	230
Totaux	271	138	116	196	95	160	976

TAB. 3.1 – Effectif des moustiques par replicat et par gène.

dans les groupes *GFP* étaient $\geq 50\%$, on a décidé de conserver les observations associées au porteur 2, avec une prévalence dans le groupe *GFP* d'environ 49 %. Finalement, le nombre d'observations retenues pour l'étude était de 773.

FIG. 3.3 – Prévalences dans les groupes *GFP*.

3.3.3 Caractéristiques des données

Les moustiques associés au porteur 7 ont été divisés en 3 groupes, contrairement à ceux associés aux autres porteurs qui ont été divisés en 6 groupes. De ce

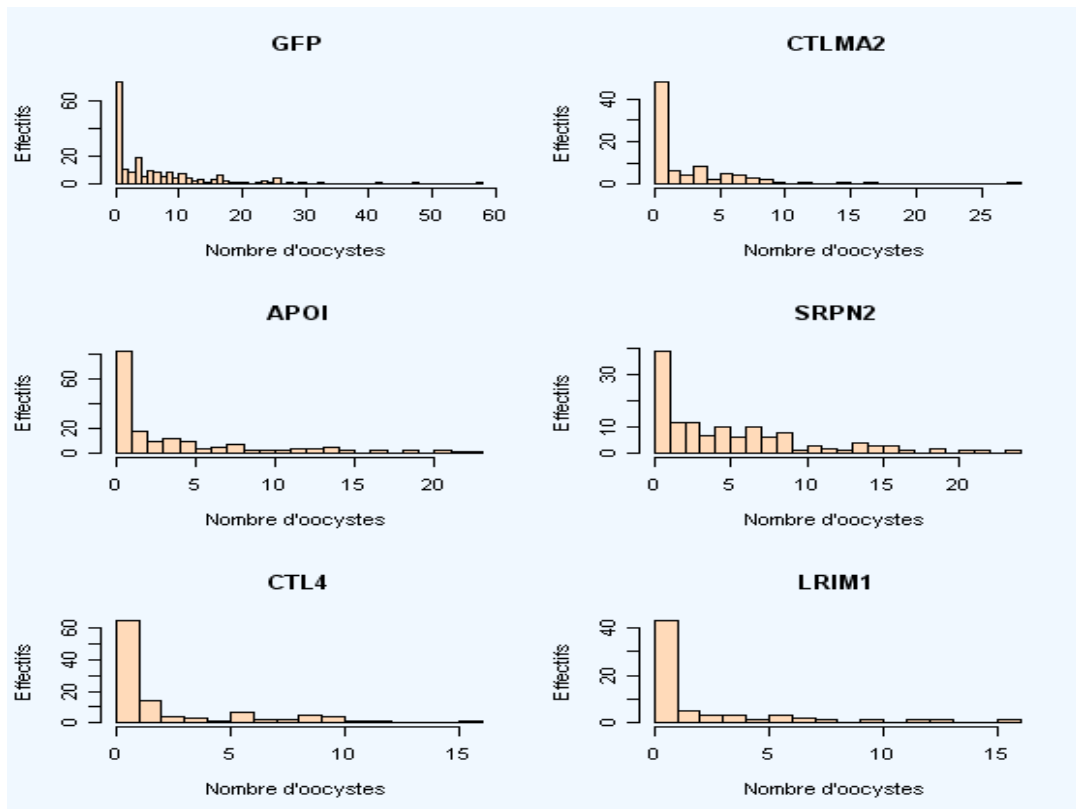


FIG. 3.4 – Histogrammes des charges oocystiques.

	<i>Min</i>	<i>Med</i>	<i>Mean</i>	<i>Max</i>	<i>SD</i>	<i>n</i>	<i>N>0</i>	<i>Prev</i>
<i>GFP</i>	0	4	6.710	58	8.95	200	141	0.70
<i>APO1</i>	0	2	4.191	23	5.46	178	114	0.64
<i>SRPN2</i>	0	4	5.466	24	5.47	133	106	0.79
	0	3	5.509	58	7.111	511	361	0.70
<i>GFP</i>	0	1	2.953	17	3.777	107	141	0.705
<i>CTL4</i>	0	1	2.455	16	3.484	110	62	0.563
<i>CTLMA2</i>	0	1	2.989	28	4.504	87	51	0.586
<i>LRIM</i>	0	1	2.200	16	3.446	65	38	0.584
	0	1	2.680	28	3.822	369	215	0.582

TAB. 3.2 – Caractéristiques des nombres d'oocystes agglomérées : *Min* représente le nombre minimal d'oocystes, *Med* la médiane, *Mean* la moyenne, *Max* le nombre maximal d'oocystes, *SD* l'écart-type, *n* le nombre de moustiques, *N>0* le nombre d'estomacs positifs et *Prev* la prévalence.

fait, nous avons aggloméré d'une part toutes les données pour l'effet des gènes *APO1*, *SRPN2*, et d'autre part nous n'avons pas considéré le groupe *GFP* du porteur 7 pour l'effet des gènes *CTL4*, *CTLMA2* et *LRIM1*.

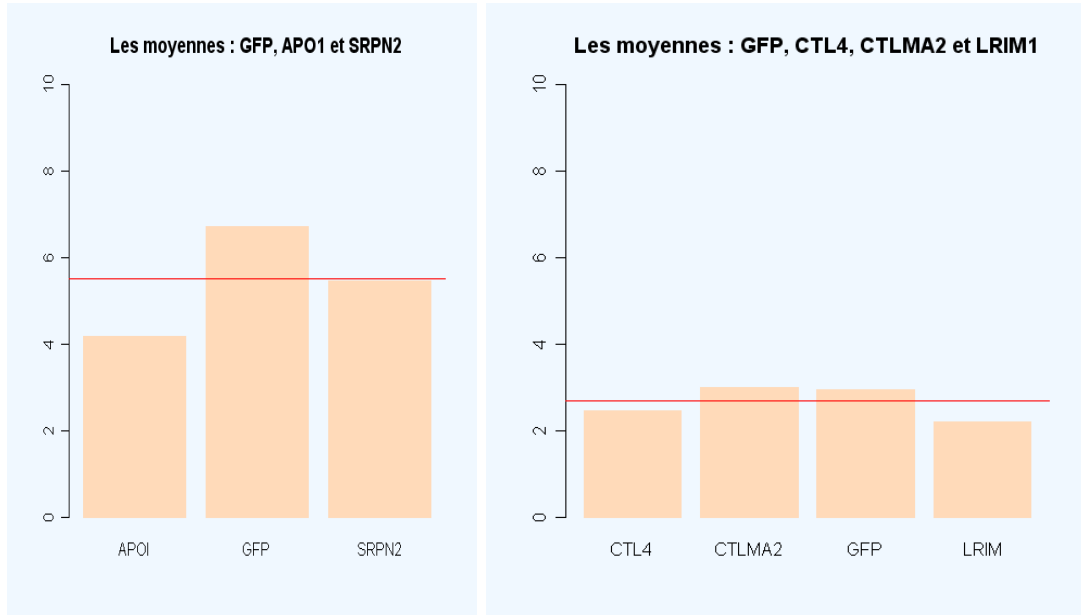


FIG. 3.5 – Moyennes des charges oocystiques.

3.4 Résultats et discussion

3.4.1 Comparaison des moyennes

	Mean	IC95%
<i>GFP</i>	6.71	[5.46; 7.96]
<i>APO1</i>	4.19	[3.38; 5.01]
<i>SRPN2</i>	5.46	[4.54; 6.39]
	5.50	[4.88; 6.13]
<i>GFP</i>	2.95	[2.24; 3.65]
<i>CTL4</i>	2.45	[1.80; 3.11]
<i>CTLMA2</i>	2.98	[2.04; 3.94]
<i>LRIM1</i>	2.20	[1.36; 3.04]
	2.68	[2.29; 3.07]

TAB. 3.3 – Intervalles de confiance bootstrap des moyennes des charges parasitaires.

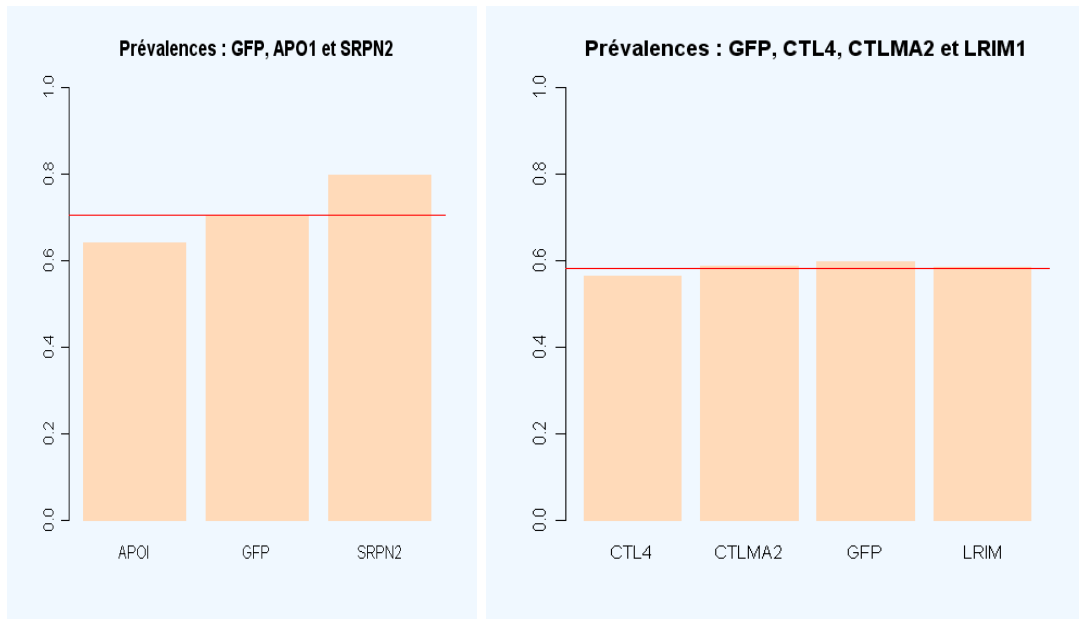


FIG. 3.6 – Prévalences par gène.

Les intervalles de confiance à 95 % des moyennes du Tableau 3.3 sont calculés par la méthode *bootstrap*, avec $B = 10\,000$ rééchantillonnages.

Le Tableau 3.3 montre que seul l'intervalle de confiance à 95 % de la moyenne du groupe *APO1* est disjoint de celui du groupe de référence *GFP*. Les intervalles de confiance à 95 % des autres groupes ont une intersection non vide avec celui du groupe contrôle *GFP*. On en déduit que, la moyenne de la charge parasitaire dans le groupe *APO1* est significativement inférieure à celle du groupe de référence *GFP*, et que celles des autres groupes ne sont pas significativement différentes de celle du groupe *GFP*. Ces résultats sont confirmés par les tests de Kruskal-Wallis (cf Tableau 3.4). On en déduit qu'en moyenne, les gènes *APO1* favorisent le développement du *P. falciparum* chez l'*A. gambiae*, alors que les gènes *CTL4*, *CTLMA2*, *LRIM1* et *SRPN2* n'ont pas d'effet significatif sur le développement du *P. falciparum* chez l'*A. gambiae*.

Gènes	<i>APO1</i>	<i>CTL4</i>	<i>CTLMA2</i>	<i>LRIM1</i>	<i>SRPN2</i>
p -value	0.010	0.341	0.788	0.252	0.800

TAB. 3.4 – p -values des comparaisons des moyennes des groupes traités à celle du groupe contrôle, test de Kruskal-Wallis.

Notons que nous n'avons pas utilisé le test de Student parce que le test de Shapiro a rejeté l'hypothèse de normalité de la charge parasitaire, avec une p -value

hautement significative ($< 2.2 \times 10^{-16}$). Le type de loi de la charge oocystique n'étant pas connu, nous avons opté pour un test non paramétrique.

3.4.2 Comparaison des distributions

Comme nous l'avons dit plus haut, le type de distribution de la charge parasitaire chez les *A. gambiae* infectés au *P. falciparum* n'est pas connu. Il faut alors envisager une procédure non paramétrique pour tester l'égalité des distributions. Les tests de rangs comme ceux de Wilcoxon et de Mann-Whitney, sont non paramétriques, mais du fait des valeurs répétées dans les observations, il peut y avoir une distorsion dans le calcul de la p -value, entraînant une perte de puissance. Celui de Komogorov-Smirnov classique est plutôt adapté aux variables continues ; or une hypothèse de continuité sur le nombre d'oocystes nous semble agressive. Nous pensons donc que la procédure de test de Monte Carlo, utilisé conjointement avec la technique des tests de permutation, est bien adapté à cette situation. Les p -values des tests de MC calculées dans la suite du document sont plus fiables que celles des tests de Kolmogorov-Smirnov classique et de Wilcoxon.

Pour effectuer les tests d'homogénéité de chacun des groupes traités et du groupe de référence, nous avons fixé le nombre de simulations N à 4 999, de sorte que pour un seuil de $\alpha = 5 \%$, le niveau théorique est égal à 5 % (cf **Proposition 2.4.1**).

Données agglomérées

Dans un premier temps, nous avons aggloméré les données des replicats. Les p -values sont consignées dans le Tableau 3.5.

	MC-KS	MC-CM	ks	wilcox
<i>APO1</i>	0.018	0.013	0.043	0.010
<i>CTL4</i>	0.167	0.331	0.415	0.000
<i>CTLMA2</i>	0.998	0.790	1	0.789
<i>LRIM1</i>	0.155	0.238	0.403	0.252
<i>SRPN2</i>	0.303	0.752	0.508	0.801

TAB. 3.5 – p -values des tests : 'MC-KS' et 'MC-CM' désignent les tests de MC basés sur les statistiques de test KS et CM respectivement (voir **ANNEXE**), 'ks' et 'Wilcox' les tests classiques de Kolmogorov-Smirnov et Wilcoxon respectivement.

Les résultats obtenus confirment ceux du test de Kruskal-Wallis des comparaisons des moyennes.

Données non agglomérées

La question qu'on se pose naturellement est celle de savoir si les distributions de probabilité de la charge parasitaire des replicats sont suffisamment semblables pour justifier l'agglomération des données, comme l'ont fait Mike A. Osta et al dans une étude similaire [4]. Pour répondre à cette question, nous avons effectué des comparaisons des distributions de probabilité du nombre d'*oocystes* des groupes de référence *GFP* associés aux différents porteurs. Comme l'indique le Tableau 3.6, l'agglomération de toutes les données n'est pas un bon choix. En effet, les tests d'homogénéité indiquent une différence significative entre certains groupes de référence. Néanmoins, d'après le Tableau 3.6, il est raisonnable de réunir les observations associées aux porteurs 2, 3 et 4 d'une part, et celles associées aux porteurs 5 et 7 d'autre part.

	MC KS	ks	Wilcox
<i>Porteur2-Porteur3</i>	0.053	0.093	0.189
<i>Porteur2-Porteur4</i>	0.088	0.356	0.179
<i>Porteur2-Porteur5</i>	0.000	0.000	0.000
<i>Porteur2-Porteur7</i>	0.000	0.000	0.000
<i>Porteur3-Porteur4</i>	0.179	0.398	0.572
<i>Porteur3-Porteur5</i>	0.038	0.072	0.021
<i>Porteur3-Porteur7</i>	0.004	0.014	0.001
<i>Porteur4-Porteur5</i>	0.000	0.000	0.000
<i>Porteur4-Porteur7</i>	0.000	0.000	0.000
<i>Porteur5-Porteur7</i>	0.118	0.204	0.166

TAB. 3.6 – Comparaison des groupes de référence

Notons *P234*, les données agglomérées associées aux porteurs 2, 3 et 4, et *P57*, celles associées aux porteurs 5 et 7. D'après les résultats obtenus dans les Tableaux 3.7 et 3.8, au seuil de 5 %, aucun des gènes dont il est question ici, n'influence de manière significative le développement du *P. falciparum* chez les *A. gambiae* qui ont été infectés par le sang des porteurs 2, 3 et 4. Par contre, dans le groupe des moustiques infectés par les *gamétocytes* provenant des porteurs 5 et 7, l'*Apolipophorine 1 (APO1)*, *C-type lectin 4 (CTL4)*, *C-type lectin MA2 (CTLMA2)* et la *Leucine-rich repeat protein (LRIM1)* ont un effet significatif sur le développement du *P. falciparum*, alors que la *Serpine 2 (SRPN2)* n'a pas d'effet significatif. Cette différence peut être due à des facteurs liés aux conditions expérimentales, au niveau de maturation des *gamétocytes*, au ratio des sexes des *gamétocytes* ou à d'autres caractéristiques des porteurs.

D'autres tests effectués sur les observations associées aux différents porteurs montrent que celles associées au *Porteur5* ont une influence significative sur les résultats (cf Tableau 3.9). Nous pensons que d'autres études devraient être menées avec plus de porteurs de *gamétocytes*, afin de déterminer les éventuels facteurs liés aux porteurs (gamétocytémie, niveau de maturation des gamétocytes, . . .), qui peuvent influencer le cycle sporogonique du *Plasmodium*.

	MC-KS	MC-CM	ks	Wilcox
<i>APO1</i>	0.238	0.133	0.612	0.158
<i>CTL4</i>	0.551	0.868	0.915	0.991
<i>CTLMA2</i>	0.428	0.375	0.791	0.315
<i>LRIM1</i>	0.762	0.930	0.994	0.888
<i>SRPN2</i>	0.621	0.454	0.952	0.388

TAB. 3.7 – p -values des tests sur les données du groupe P234.

	MC-KS	MC-CM	ks	Wilcox
<i>APO1</i>	0.006	0.002	0.012	0.003
<i>CTL4</i>	0.003	0.001	0.008	0.001
<i>CTLMA2</i>	0.009	0.001	0.023	0.001
<i>LRIM1</i>	0.004	0.001	0.009	0.002
<i>SRPN2</i>	0.068	0.054	0.134	0.060

TAB. 3.8 – p -values des tests sur les données du groupe P57.

Test de MC KS

<i>GFP</i> vs	<i>APO1</i>	<i>CTL4</i>	<i>CTLMA2</i>	<i>LRIM1</i>	<i>SRPN2</i>
<i>Porteur2</i>	0.054	0.350	0.331	0.551	0.870
<i>Porteur3</i>	0.560	0.426	0.459	0.687	0.325
<i>Porteur4</i>	0.528	0.146	0.049	0.736	0.246
<i>Porteur5</i>	0.000	0.003	0.007	0.001	0.057
<i>Porteur7</i>	0.146	-	-	-	0.054

TAB. 3.9 – p -values des tests de MC avec KS comme statistique de test.

Remarque 3.4.1. Les p -values des tests classiques de KS et de Wilcoxon sont calculées avec des erreurs dues aux violations des conditions d'application. Le logiciel R affiche le message suivant : « impossible de calculer les p -values correctes avec des *ex-aequo* ».

Test de KS classique

<i>GFP</i> vs	<i>APO1</i>	<i>CTL4</i>	<i>CTLMA2</i>	<i>LRIM1</i>	<i>SRPN2</i>
<i>Porteur2</i>	0.270	0.877	0.772	0.996	0.999
<i>Porteur3</i>	0.949	0.698	0.798	0.942	0.494
<i>Porteur4</i>	0.932	0.398	0.105	0.990	0.561
<i>Porteur5</i>	0.003	0.008	0.023	0.009	0.003
<i>Porteur7</i>	0.255	-	-	-	0.116

TAB. 3.10 – p -values des tests de KS classique.

Test de Wilcoxon

<i>GFP</i> vs	<i>APO1</i>	<i>CTL4</i>	<i>CTLMA2</i>	<i>LRIM1</i>	<i>SRPN2</i>
<i>Porteur2</i>	0.036	0.180	0.155	0.400	0.484
<i>Porteur3</i>	0.400	0.475	0.933	0.966	0.444
<i>Porteur4</i>	0.710	0.991	0.033	0.677	0.544
<i>Porteur5</i>	0.003	0.001	0.001	0.002	0.003
<i>Porteur7</i>	0.054	-	-	-	0.515

TAB. 3.11 – p -values des tests de Wilcoxon.

CONCLUSION

La procédure des tests de Monte Carlo, utilisée conjointement avec la technique des tests de permutation, permet, à partir de statistiques dont les distributions exactes ou asymptotiques sont analytiquement difficiles à calculer mais peuvent être simulées, de construire des tests d'homogénéité de deux échantillons vérifiant quelques propriétés intéressantes. En effet, ces tests sont exacts quelque soit le nombre N de simulations choisi de sorte que $\alpha(N+1) \in \mathbb{N}$, pour un niveau α fixé, et ils sont applicables quelque soit la nature continue ou discrète de la distribution de probabilité de la variable d'intérêt. Cette procédure de test permet aussi de combiner plusieurs statistiques différentes pour améliorer la puissance. Nous avons montré, par des simulations, que cette procédure de tests est comparable à d'autres tests classiques, et qu'elle peut être une solution adaptée aux cas où les conditions d'application d'autres tests non paramétriques d'homogénéité ne sont pas vérifiées. On peut déplorer le fait qu'elle soit gourmande en temps d'exécution, mais il est parfois nécessaire de se donner le temps d'obtenir des résultats fiables.

Après avoir implémenté la procédure des tests de Monte Carlo sous le logiciel R, nous l'avons appliquée aux données réelles d'une étude menée par l'OCEAC en vu d'évaluer l'effet de certains gènes de l'immunité de *A. gambiae* sur le développement de *P. falciparum*. Notons que les conditions d'application (continuité, absence de nœuds,...) d'autres tests non paramétriques ne sont pas vérifiées par les données issues de cette étude.

La procédure des tests de Monte Carlo est randomisée dans le sens où le résultat dépend de simulations auxiliaires. Il se pose alors le problème de la convergence vers 0, quand le nombre de permutations N tend vers ∞ , de la probabilité que le résultat de cette procédure de test soit différent de celui du test fondamental, d'une part. D'autre part, au cas où elle converge, quelle serait sa vitesse de convergence ?

ANNEXE : CODE R DE LA PROCÉDURE DES TESTS DE MONTE CARLO

```
#      I- VERSION PERMUTATIONNELLE DES TESTS DE MONTE CARLO

#      I.1- ALGORITHME DES TESTS NON COMBINES

#      Nous notons T la statistique du test et nous ne considérons que
#      les tests unilatéraux à droite ou à gauche.

# 1- Calculer la valeur T0 prise par la statistique de test sur les
#      observations;
# 2- Tirer N permutations indépendantes et aléatoires des observations;
# 3- Calculer les valeurs T1,...,TN prises par T sur ces N permutations;
# 4- Déterminer le rang R0 de T0 dans T0, T1,..., TN;
# 5- Estimer la valeur-p par  $1-(R0-1)/(N+1)$  pour un test unilatéral
#      à droite ou par  $R0/(N+1)$  pour un test unilatéral à gauche.

#      Pour des échantillons de petites tailles, on considère toutes les
#      permutations possibles des données entre les deux échantillons.

#      LES ARGUMENTS DE LA FONCTION DU TEST

# 1- x et y : les observations des deux groupes dont on veut tester
#      l'homogénéité;
# 2- N : le nombre de permutations;
# 3- type : indique le caractère continu ou discret de la variable
#      d'intérêt;
# 4- alternative : indique si le test est unilatéral à droite ou à gauche.
```

```

#          I.2- FONCTION DU TEST NON COMBINE DE MONTE CARLO

# Charger le package 'gtools'
library(gtools)

MC.t=function(x,y,T,N=5000,type='continu',alternative='droite')
{
  if((type!="continu")&(type!="discret"))
    stop("Argument type incorrect! continu ou discret.")

  if((alternative!="droite")&(alternative!="gauche"))
    stop("Argument alternative incorrect! droite ou gauche.")

  v=c(x,y)
  n=length(x) ; m=length(y)
  nl=n+1 ; nbre=n+m
  Nl=choose(nbre,n)

  # Gestion des arguments des statistiques de test
  #car toutes n'ont pas le même nombre d'arguments
  if(length(formals(T))==2)
    formals(T)=alist(x=,y=,type=)
  T0=T(x,y,type)
  vect_T=c()          # Vecteur contenant les valeurs de T
  if(Nl<=5000)
    {
      vect_T=apply(combinations(nbre,n,set=FALSE),1,function(a)
        {
          T(v[a],v[-a],type)
        })
    }
  else {
    vect_T=sapply(1:N,function(i)
      {
        vl=sample(v,nbre,replace=FALSE)
        return(T(vl[1:n],vl[nl:nbre],type))
      })
  }
  # Rang (éventuellement aléatoire) de T0 dans l'ensemble {T0,T1,...,TN}
  R=rank(c(T0,vect_T),ties.method='random')
  RO=R[1]
  M=length(R)
  p_value=ifelse(alternative=='droite',1-((RO-1)/M),RO/M)

  reponse=list(alternative,M-1,T0,p.value=p_value)

  return(reponse)
}

# Affichage des résultats plus convivial

MC.test=function(x,y,T,N=5000,type='continu',alternative='droite')
{
  # Gestion de l'affichage

  reponse=MC.t(x,y,T,N,type,alternative)

```

```

cat(" ", "\n")
cat("          VERSION PERMUTATIONNELLE DES TESTS DE MONTE CARLO : " ,
    "\n" )
cat("          -----", "\n")
cat(" ", "\n")
cat("          Test unilatéral à ", reponse[[1]], "\n")
cat(" ", "\n")
cat("          - Nombre de simulations N = ", reponse[[2]], "\n")
cat(" ", "\n")
cat("          - Valeur observée de la statistique de test T0 = "
    , reponse[[3]], "\n")
cat(" ", "\n")
cat("          - p.value = ", reponse[[4]])
cat(" ", "\n")
}

#          I.3- TESTS COMBINES DE MONTE CARLO

library(gtools)

Combin.MC.t=function(x,y,V=list(KS,CM,tetal,teta2),N=5000,
    type='continu',alternative='droite',absolu='FALSE')
{
  if((type!="continu")&(type!="discret"))
    stop("Argument type incorrect! type=c('continu','discret')!")
  if((absolu!='TRUE')&(absolu!='FALSE'))
    stop("Argument absolu=c('TRUE','FALSE')!")
  if((alternative!="droite")&(alternative!="gauche"))
    stop("Argument alternative incorrect! droite ou gauche.")

  if(length(V)==1)
  {
    MC.t(x,y,T=V[[1]],N,type,alternative)
  }
  else
  {
    k=length(V) # Nbre de statistiques de test
    v=c(x,y)
    n=length(x) ; m=length(y)
    nl=n+1 ; nbre=n+m
    Nl=choose(nbre,n)

    if(Nl<=5000)
    {
      mat_T=apply(combinations(nbre,n,set=FALSE),1,function(a)
        {
          xnew=v[a] ; ynew=v[-a]
          u=sapply(1:k,function(i)
            {
              # Gestion des arguments des statistiques de tests
              # car toutes n'ont pas le même nombre d'arguments

              if(length(formals(V[[i]]))==2)
                formals(V[[i]])=alist(x=,y=,type=)
              V[[i]](xnew,ynew,type)
            })
          return(u)
        })
    }
  }
}

```

```

    })
  }
else {
  mat_T=sapply(1:N,function(i)
  {
    vl=sample(v,nbre,replace=FALSE)
    xnew=vl[1:n] ; ynew=vl[nl:nbre]
    u=sapply(1:k,function(j)
    {
      # Gestion des arguments des statistiques de tests
      # car toutes n'ont pas le même nombre d'arguments

      if(length(formals(V[[j]]))==2)
        formals(V[[j]])=alist(x=,y=,type=)
      V[[j]](xnew,ynew,type)
    })
    return(u)
  })

  mat_T=t(mat_T)

  # Ajout de V0
  V0=sapply(1:k,function(i)
  {
    if(length(formals(V[[i]]))==2)
      formals(V[[i]])=alist(x=,y=,type=)
    V[[i]](x,y,type)
  })
  mat_T=rbind(V0,mat_T)

  # Standardisation des statistiques
  standard=sapply(mat_T,2,function(u) {
    return((u-mean(u))/sd(u))
  })
  if(absolu=='TRUE') Q=apply(abs(standard),1,max)
  else Q=apply(standard,1,max)

  R=rank(Q,ties.method='random')
  M=length(R)
  RO=R[1]

  p_value=ifelse(alternative=='droite',1-((RO-1)/M),RO/M)

  reponse=list(alternative,M-1,V0,p.value=p_value)

  return(reponse)
}

# Affichage des résultats

Combin.MC.test=function(x,y,V=list(KS,CM,tetal,teta2),N=5000,
  type='continu', alternative='droite',absolu='FALSE')
{
  if(length(V)==1) reponse=MC.t(x,y,T=V[[1]],N,type,alternative)
  else reponse=Combin.MC.t(x,y,V,N,type,alternative,absolu)
}

```

```

cat(" ", "\n")
cat("          VERSION PERMUTATIONNELLE DES TESTS DE MONTE CARLO : ",
      "\n" )
cat("          -----", "\n")
cat(" ", "\n")
cat("          Test unilatéral à ", reponse[[1]], "\n")
cat(" ", "\n")
cat("          - Nombre de simulations N = ", reponse[[2]], "\n")
cat(" ", "\n")
cat("          - Valeurs observées des statistiques de test VO = ",
      reponse[[3]], "\n")
cat(" ", "\n")
cat("          - p.value = ", reponse[[4]])
cat(" ", "\n")
}

#          II- STATISTIQUES DE TEST

#          II.1- KOLMOGOROV SMIRNOW KS

KS=function(x,y)
{
  if(!(is.vector(x))&! (is.vector(y)))
    stop("Les arguments doivent être des vecteurs!")
  if((length(x)==0) | (length(y)==0))
    stop("Un des vecteurs est vide!")
  z=c(x,y)
  KS=max(abs(ecdf(x)(z)-ecdf(y)(z)))
  return(KS)
}

KS_plus=function(x,y)
{
  if(!(is.vector(x))&! (is.vector(y)))
    stop("Les arguments doivent être des vecteurs!")
  if((length(x)==0) | (length(y)==0))
    stop("Un des vecteurs est vide!")
  z=c(x,y)
  K_plus=max(ecdf(x)(z)-ecdf(y)(z))
  return(K_plus)
}

#          II.2- CRAMER-VON-MISES

CM=function(x,y)
{
  if(!(is.vector(x))&! (is.vector(y)))
    stop("Les arguments doivent être des vecteurs!")
  if((length(x)==0) | (length(y)==0))
    stop("Un des vecteurs est vide!")
  n=length(x) ; m=length(y)
  CM=((n*m)/(n+m)^2)*(sum((ecdf(x)(x)-ecdf(y)(x))^2)+
    sum(ecdf(x)(y)-ecdf(y)(y))^2)
  return(CM)
}

```

II.3- MOMENT D'ORDRE 1

```
teta1=function(x,y)
{
  if(!(is.vector(x))&! (is.vector(y)))
    stop("Les arguments doivent être des vecteurs!")
  if((length(x)==0) | (length(y)==0))
    stop("Un des vecteurs est vide!")
  m1=abs(mean(x)-mean(y))
  return(m1)
}
```

II.4- MOMENT D'ORDRE 2

```
teta2=function(x,y)
{
  if(!(is.vector(x))&! (is.vector(y)))
    stop("Les arguments doivent être des vecteurs")
  if((length(x)==0) | (length(y)==0))
    stop("Un des vecteurs est vide!")
  m2=abs(var(x)-var(y))
  return(m2)
}
```

II.4- MOMENT D'ORDRE 3

```
teta3=function(x,y)
{
  if(!(is.vector(x))&! (is.vector(y)))
    stop("Les arguments doivent être des vecteurs")
  if((length(x)==0) | (length(y)==0))
    stop("Un des vecteurs est vide!")
  moy_x=mean(x)
  moy_y=mean(y)
  m3_x=ifelse(var(x)==0,0,mean(((x-moy_x)/sd(x))^3))
  m3_y=ifelse(var(y)==0,0,mean(((y-moy_y)/sd(y))^3))
  m3=abs(m3_x-m3_y)
  return(m3)
}
```

II.5- MOMENT D'ORDRE 4

```
teta4=function(x,y)
{
  if(!(is.vector(x))&! (is.vector(y)))
    stop("Les arguments doivent être des vecteurs!")
  if((length(x)==0) | (length(y)==0))
    stop("Un des vecteurs est vide!")
  moy_x=mean(x)
  moy_y=mean(y)
  m4_x=ifelse(var(x)==0,0,mean(((x-moy_x)/sd(x))^4))
  m4_y=ifelse(var(y)==0,0,mean(((y-moy_y)/sd(y))^4))
  m4=abs(m4_x-m4_y)
  return(m4)
}
```

```

#      II.6- DISTANCES L1, L2, Linf

#      Dans le cas des distributions discrètes, les estimateurs
# de densités de probabilité sont remplacées par les fonctions de masse.

coeff_c=function(x)
{
  if(!(is.vector(x))) stop("L'argument doit être un vecteur!")
  if(length(x)==0) stop("La longueur du vecteur doit être >= 1 !")
  n=length(x)
  s=sqrt(var(x))
  coeff=ifelse(s==0,1,(n^(1/5))/(2*s))
  return(coeff)
}

#      Noyau

noyau_K=function(x)
{
  K=sapply(x,function(u){0.5*(abs(u)<=1)})
  return(K)
}

#      Estimateur de la densité de probabilité: méthode du noyau
densite_est=function(v)
{
  #Pour un vecteur v donné, cette fonction estime la
  #densité par la méthode du noyau
  if(!(is.vector(v))) stop("L'argument doit être un vecteur!")
  if(length(v)==0) stop("La longueur du vecteur doit être >= 1 !")
  c=coeff_c(v)
  est_dens=function(x)
  {
    d_x=sapply(x,function(u){c*mean(noyau_K(c*(u-v)))})
    return(d_x)
  }
  return(est_dens)
}

#      Fonction de masse

f_mass=function(ech)
{
  #Cette fonction calcule et renvoie la fonction de masse
  #d'un échantillon en un point
  if(!(is.vector(ech)))
    stop("L'argument doit être un vecteur!")
  if(length(ech)==0)
    stop("La longueur du vecteur doit être >= 1 !")
  masse_ech=function(x)
  {
    masse_x=sapply(x,function(u){mean(ech==u)}) #
    return(masse_x)
  }
  return(masse_ech)
}

```

```

#           II.6.1- Distance L1

L1=function(x,y,type="continu")
{
  #L'argument "type" indique si les distributions sont continues
  #ou discrètes. Les deux valeurs possibles de cet argument sont:
  #"continu" et "discret"
  if((type!="continu")&(type!="discret"))
    stop("Argument type incorrect! continu ou discret.")

  #Choix de l'estimateur en fonction de la nature continue ou non
  #des distributions.
  if(type=="continu") densite=densite_est else densite=f_mass

  L1=sum(abs(densite(x)(x)-densite(y)(x)))+
    sum(abs(densite(x)(y)-densite(y)(y)))
  return(L1)
}

#           II.6.2- Distance L2

L2=function(x,y,type="continu")
{
  #L'argument type indique si les distributions sont continues
  #ou discrètes. Les deux valeurs possibles de cet argument sont:
  #"continu" et "discret"
  if((type!="continu")&(type!="discret"))
    stop("Argument type incorrect! continu ou discret.")

  #Choix de l'estimateur en fonction de la nature continue
  #ou non des distributions.
  if(type=="continu") densite=densite_est else densite=f_mass

  L2=sqrt(sum((abs(densite(x)(x)-densite(y)(x)))^2)+
    sum((abs(densite(x)(y)-densite(y)(y)))^2))
  return(L2)
}

#           II.6.3- Distance Linf

Linf=function(x,y,type="continu")
{
  #L'argument "type" indique si les distributions sont continues
  #ou discrètes. Les deux valeurs possibles de cet argument sont:
  #"continu" et "discret"
  if((type!="continu")&(type!="discret"))
    stop("Argument type incorrect! continu ou discret.")

  #Choix de l'estimateur en fonction de la nature continue
  #ou non des distributions.
  if(type=="continu") densite=densite_est else densite=f_mass

  Linf=max(c(abs(densite(x)(x)-densite(y)(x)),
    abs(densite(x)(y)-densite(y)(y))))
  return(Linf)
}

```

```

#           III- ESTIMATION DU NOMBRE DE PERMUTATIONS ALEATOIRES

# Cette fonction estime le nombre N de simulations nécessaire pour
# que l'estimation du quantile d'ordre p appartienne à l'intervalle
# [Q(binf); Q(bsup)], où Q est la fonction quantile de la fonction de
# répartition de T sous l'hypothèse Ho.

estim.N=function(epsil=0.005,p=0.95,N0=4000,P=0.95)
{
  N=N0
  binf=p-epsil ; bsup=p+epsil
  repeat
  {
    a=floor(p*(N)) ; b=N-a+1
    Pr=pbeta(bsup,a,b)-pbeta(binf,a,b)
    if(Pr>=P) break
    N=N+50
  }
  if(Pr>=P)
  {
    cat("      Le nombre minimal de permutations est N =",N,"\n")
    cat("      La probabilité que l'estimation du quantile d'ordre
",p,"\n")
    cat("      appartienne à l'intervalle [Q(",binf,")
Q(",bsup,")] est Pr =",Pr,"","\n")
    cat("      où Q est la fonction quantile de la fonction de
répartition de la","\n")
    cat("      statistique de test T sous Ho","\n")
  }
  else
  {
    cat("La probabilité critique n'est pas atteinte pour N =",N,"\n")
  }
}

```

Bibliographie

- [1] Rapport sur le paludisme en Afrique 2003, OMS/UNISEF, 2003.
- [2] Stéphanie Blandin and Elena A Levashina (2004) : Mosquito immune responses against malaria parasites. Science 2004, 16-20.
- [3] Entomologie du paludisme et contrôle des vecteurs, OMS juillet 2003.
- [4] Mike A. Osta et al. (2004) : Effects of Mosquito Genes on *Plasmodium* Development SCIENCE, VOL 303, 2030-2032.
- [5] Michelle M. Riehle et al. (2006) : Natural Malaria Infection in *Anopheles gambiae* is Regulated by a Single Genomic Control Region, SCIENCE, Vol 321 ; 577-579 (www.sciencemag.org).
- [6] Bertrand Boisson et al. (2006) : Gene silencing in mosquito salivary glands by RNAi FEBS Letters 580(2006), 1988-1992.
- [7] Stéphanie Blandin et al. (2002) : Reverse genetics in the mosquito *Anopheles gambiae* : targeted disruption of Defensing gene EMBO reports vol. 3|no.9|pp 852-856|2002.
- [8] Paul R. BURTON, Martin D. Torbin, John L. Happer (2005) : Key concepts in genetic epidemiology, Series, Vol 366 ; 941-951.
- [9] Jean-Marie DUFOUR et Abdeljelil FARHAT (2001) : Exact nonparametric two-sample homogeneity tests for possibly discrete distributions.
- [10] Jean-Marie Dufour, Abdeljelil Farhat, Lynda Khalaf (2005) : Tests multiples simulés et tests de normalité basés sur plusieurs moments dans les modèles de régression. ISSN 1198-8177.
- [11] Bradley, J.V. 1978. Robustness ? British Journal of Mathematical and Statistical Psychology.
- [12] Bradley Efron et Robert J. Tibshirani : An introduction to the Bootstrap.
- [13] A.W Van Der Vaart : Asymptotic Statistics ; Cambridge University Press 1998 ; 265-289.
- [14] Philippe TASSI : Méthodes statistiques ; 2^e édition Janvier 1989.