Modèles de mutation : étude probabiliste et estimation paramétrique

Adrien Mazoyer

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Mathématiques Appliquées**

Arrêté ministériel : 7 août 2006

Présentée par

Adrien Mazoyer

Thèse dirigée par **Bernard Ycart**

préparée au sein du **Laboratoire Jean Kuntzmann**
et de **l'École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Modèles de mutation

Étude probabiliste et estimation paramétrique

Thèse soutenue publiquement le **4 juillet 2017**,
devant le jury composé de :

Mme Adeline Samson Leclercq
Professeure, Université Grenoble Alpes, Présidente
M. Philip J. Gerrish
Directeur de recherche, Georgia Institute of Technology, Rapporteur
Mme Sylvie Méléard
Professeure, École Polytechnique , Rapporteur
Mme Agnès Hamon
Maître de Conférence, Université Grenoble Alpes, Examinatrice
M. Guillaume Martin
Chargé de recherche, ISEM Montpellier, Examinateur
M. Bernard Ycart
Professeur, Université Grenoble Alpes, Directeur de thèse

Modèles de mutation

Étude probabiliste et estimation paramétrique

Résumé : Les modèles de mutations décrivent le processus d'apparitions rares et aléatoires de mutations au cours de la croissance d'une population de cellules. Les échantillons obtenus sont constitués de nombres finaux de cellules mutantes, qui peuvent être couplés avec des nombres totaux de cellules ou un nombre moyen de cellules en fin d'expérience. La loi du nombre final de mutantes est une loi à queue lourde : de grands décomptes, appelés "jackpots", apparaissent fréquemment dans les données. Une construction générale des modèles se décompose en trois niveaux. Le premier niveau est l'apparition de mutations aléatoires au cours d'un processus de croissance de population. En pratique, les divisions cellulaires sont très nombreuses, et la probabilité qu'une de ces divisions conduise à une mutation est faible, ce qui justifie une approximation poissonnienne pour le nombre de mutations survenant pendant un temps d'observation donné. Le second niveau est celui des durées de développement des clones issus de cellules mutantes. Du fait de la croissance exponentielle, la majeure partie des mutations ont lieu à la fin du processus, et les durées de développement sont alors indépendantes et exponentiellement distribuées. Le troisième niveau concerne le nombre de cellules qu'un clone issu d'une cellule mutante atteint pendant une durée de développement donnée. La loi de ce nombre dépend principalement de la loi des instants de division des mutantes. Le modèle classique, dit de Luria-Delbrück, suppose que les développements cellulaires des cellules normales aussi bien que mutantes s'effectue selon un processus de Yule. On peut dans ce cas expliciter la loi du nombre final de mutantes. Elle dépend de deux paramètres, qui sont le nombre moyen de mutations et le paramètre de fitness (rapport des taux de croissance des deux types de cellules). Le problème statistique consiste à estimer ces deux paramètres au vu d'un échantillon de nombres finaux de mutantes. Il peut être résolu par maximisation de la vraisemblance, ou bien par une méthode basée sur la fonction génératrice. Diviser l'estimation du nombre moyen de mutations par le nombre total de cellules permet alors d'estimer la probabilité d'apparition d'une mutation au cours d'une division cellulaire. l'estimation de cette probabilité est d'une importance cruciale dans plusieurs domaines de la médecine et de biologie : rechute de cancer, résistance aux antibiotiques de Mycobacterium Tuberculosis, etc. La difficulté provient de ce que les hypothèses de modélisation sous lesquelles la distribution du nombre final de mutants est explicite sont irréalistes. Or estimer les paramètres d'un modèle quand la réalité en suit un autre conduit nécessairement à un biais d'estimation. Il est donc nécessaire de disposer de méthodes d'estimation robustes pour lesquelles le biais, en particulier sur la probabilité de mutation, reste le moins sensible possible aux hypothèses de modélisation. Cette thèse contient une étude probabiliste et statistique de modèles de mutations prenant en compte les sources de biais suivantes : durées de vie non exponentielles, morts cellulaires, variabilité du nombre final de cellules, durées de vie non-exponentielles et non-identiquement distribuées, dilution de la population initiale. Des études par simulation des méthodes considérées sont effectuées afin de proposer, selon les caractéristiques du modèle, l'estimation la plus fiable possible. Ces méthodes ont également été appliquées à des jeux de données réelles, afin de comparer les résultats avec les estimations obtenues sous les modèles classiques. Un package R a été implémenté en

collaboration avec Rémy Drouilhet et Stéphane Despréaux et est disponible sur le CRAN. Ce package contient les différents résultats obtenus au cours de ce travail. Il contient des fonctions dédiées aux modèles de mutations, ainsi qu'à l'estimation des paramètres. Les applications ont été en partie développées pour le Labex TOUCAN (Toulouse Cancer).

Mots clefs : Modèles de mutation. Loi de Luria-Delbrück. Analyse de fluctuations. Processus de branchement. Processus inhomogène.

Abstract : Mutation models are probabilistic descriptions of the growth of a population of cells, where mutations occur randomly during the process. Data are samples of integers, interpreted as final numbers of mutant cells. These numbers may be coupled with final numbers of cells (mutant and non mutant) or a mean final number of cells. The frequent appearance in the data of very large mutant counts, usually called "jackpots", evidences heavy-tailed probability distributions. Any mutation model can be interpreted as the result of three ingredients. The first ingredient deals with the number of mutations occuring with small probability among a large number of cell divisions. Due to the law of small numbers, the number of mutations approximately follows a Poisson distribution. The second ingredient models the developing duration of the clone stemming from each mutation. Due to exponential growth, most mutations occur close to the end of the experiment. Thus the developing time of a random clone has exponential distribution. The last ingredient represents the number of mutant cells that any clone developing for a given time will produce. This number depends mainly on the distribution of division times of mutants. One of the most often used mutation models is the Luria-Delbrück model. In this model, division times of mutant cells are supposed to be exponentially distributed. Thus a clone develops according to a Yule process and its size at any given time follows a geometric distribution. This approach leads to a family of probability distributions which depend on the expected number of mutations and the relative fitness (the ratio of the growth rate of normal cells to that of mutants). The statistical purpose of mutation models is the estimation of these parameters. The probability for a mutant cell to appear upon any given cell division is estimated dividing the mean number of mutations by the mean final number of cells. Given samples of final mutant counts, it is possible to build estimators maximizing the likelihood, or using the probability generating function. Computing robust estimates is of crucial importance in medical applications, like cancer tumor relapse or multidrug resistance of Mycobacterium Tuberculosis for instance. The problem with classical mutation models, is that they are based on quite unrealistic assumptions : constant final number of cells, no cell deaths, exponential distribution of lifetimes, or time homogeneity. Using a model for estimation, when the data have been generated by another one, necessarily induces a bias on estimates. Several sources of bias has been partially dealt with until now : non-exponential lifetimes, cell deaths, fluctuations of the final count of cells, dependence of the lifetimes, plating efficiency. The time homogeneity remains untreated. This thesis contains probabilistic and statistical study of mutation models taking into account the following bias sources : non-exponential and non-identical lifetimes, cell deaths, fluctuations of the final count of cells, plating efficiency. Simulation studies have been performed in order to propose robust estimation methods, whatever the modeling assumptions. The methods have also been applied to real data sets, to compare the results with the estimates obtained under classical models. An R package based on the different results obtained in this work has been implemented (joint work with Rémy Drouilhet and Stéphane Despréaux) and is available on the CRAN. It includes functions dedicated to the mutation models and parameter estimation. The applications have been

partially developed for the Labex TOUCAN (Toulouse Cancer).

Keywords : Mutation models. Luria-Delbrück distribution. Fluctuation analysis. Branching processes. Inhomogeneous processes.

Remerciements

Je considérais initialement que la rédaction de remerciements n'était qu'une simple formalité. Mais au fur et à mesure de leur écriture, et avec un peu de recul, je réalise que cette étape clôt définitivement ces 3 années. Ce manuscrit n'aurait jamais pu voir le jour sans l'intervention de nombreuses personnes que je me dois de mentionner ici.

Mes premiers remerciements sont évidemment pour mon directeur de thèse Bernard Ycart. Dire que cette expérience n'aurait pas été la même si j'avais été encadré par une autre personne est assez trivial. Le fait est que l'aboutissement de ces 3 années est en très grande partie due à tes grandes qualités d'encadrant, que ce soit humaines ou scientifiques. Tu as réussi à m'inculquer le goût de l'enseignement et de la recherche. J'espère un jour réussir à transmettre à mon tour ce que tu m'as appris.

Je tiens à remercier Philip J. Gerrish et Sylvie Méléard d'avoir accepté de prendre le temps de lire et évaluer cette thèse, ainsi que Adeline Leclercq Samson, Agnès Hamon et Guillaume Martin pour leur participation à mon jury de soutenance.

Par rapport à d'autres laboratoires de mathématiques, le Laboratoire Jean Kuntzmann sort un peu du commun et ce grâce à beaucoup de monde. Merci aux équipes administratives pour leur réactivité et leur compétence, dont en particulier Juana Dos-Santos, Héléne Baum ainsi que Laurence Wazné et Catherine Laiolo pour votre bonne humeur ! Merci aux équipes informatiques (dont Frédéric Audra, Bruno Rusconi, Patrice Navarro et Stéphane Despréaux qui a également collaboré à l'élaboration du package R) de nous permettre de travailler dans de très bonnes conditions. Je pense également à plusieurs membres permanents du département Probabilité et Statistique du LJK : Sana Louhichi qui a fait germer dans mon esprit l'idée de faire une thèse alors que je n'étais qu'étudiant de master 1, et qui par la suite m'a présenté Bernard pour mon stage de master 2, point de départ de cette aventure ; Rémy Drouilhet, qui a été d'une grande aide lors de l'optimisation du package R, avec qui j'ai pu découvrir sous un nouvel angle l'enseignement des statistiques et d'autres aspects de la vie moins scientifiques ; Adeline Leclerq Samson, ancienne chef de l'équipe SVH, pour m'avoir permis d'encadrer des TPs avec elle ainsi que pour m'avoir donner un énorme coup de pouce pour décrocher ce stage postdoctoral à Montréal ; Jean-François Coeurjolly, ancien du LJK, et avec qui j'aurai justement le plaisir d'effectuer ce stage.

Mais il y a eu surtout d'autres thésards, sans qui la vie au laboratoire aurait été bien différente ! En commençant par Achmad, co-bureau avec qui j'ai partagé cette expérience

depuis le début de la thèse, merci pour ta gentillesse, à dans quelques semaines pour ta propre soutenance et bonne chance au Danemark ! Bonne chance également à Modibo - continue sur ta lancée ! - et Aude - félicitations pour le(la) petit(e) ! - qui nous ont rejoint il y a maintenant un an. En parallèle, un noyau dur d'autres doctorants m'a permis de me rappeler que dans le fond, nous sommes encore étudiants : Jean-Baptiste, Charles, Nelson, Alexandre, Kévin, Margaux, Rémi, Meriem. Ces dernières années en votre compagnie n'ont vraiment pas été anodines. Il est rassurant de voir que la relève est là avec les générations suivantes : Lionel, Arnaud, Émilie, Victor, David, Benoît, Sophie, Arthur, Reda. Merci pour tous ces moments ensemble. Cette ambiance va vraiment me manquer.

Le fait d'avoir des proches complètement extérieurs au milieu de la thèse a également eu son importance. J'ai une première pensée pour les Forains que j'ai rencontré en arrivant à Grenoble : Rachel, Sarah, Robin, Baptiste, Tanguy, Astrid, Hélène, Alice, Léa M et R et Clémentine. Papy Adri aurait eu du mal à garder la tête hors de l'eau sans ces longues discussions à base d'orge et ces parties de baseball. Merci à mes amis de longue date de la licence de Mathématiques : Corentin, Rémi et Simon. Alors que nous avions jeté l'éponge après une année scolairement ratée, c'est vous qui m'avez poussé un an plus tard à retenter ma chance. Je n'écrirais pas ces mots si vous ne l'aviez pas fait, et je ne serais jamais assez reconnaissant pour cela. Merci à mes autres vieux amis de Dijon : Antoine, Lisa, Julie C, Gaël, Nicolas, Romain. Je peux enfin vous dire que oui, c'est fait ! De manière plus générale, merci à tous ceux qui m'ont suivi et soutenu depuis que nous nous connaissons : Fabien, Julie B, Amande, Thibaud M, Jules, Sarah FT, Marion, Léa S, Julie P, Justine B, François, Sébastien, Perceval, Ophélie, Maud, Chantal, Philippe, Isabelle et Daniel.

Je me dois de remercier très particulièrement Charly et Justine S. Votre présence ces derniers mois depuis votre retour à Grenoble a été bien plus important que je ne peux l'exprimer. Vous avez été toujours été là jusque dans l'organisation de la soutenance, et je ne pourrais jamais assez vous remercier pour ça. Plus que les montagnes, l'ambiance du laboratoire et la vie grenobloise, c'est bien vous qui allez le plus me manquer une fois parti de Grenoble ! Mention spéciale pour Charly qui a continuellement tenté de connaitre le titre de ma thèse et d'être capable d'expliquer le sujet de ma thèse. Comprendre et s'intéresser à ce que l'on nous raconte c'est une chose, tenter de le retransmettre est bien plus ardue !

Mes derniers remerciements vont à ma famille. Merci les frangines, Orianne et Clara, pour les hauts et bas que nous avons partagés depuis notre enfance. Merci à ma grand-mère Rosine d'avoir fait le déplacement pour partager le moment de la soutenance, cela signifie beaucoup pour moi. J'ai par ailleurs une pensée pour feu mon grand-père Bernard, qui aurait suivi ces 3 années avec beaucoup d'attention et d'enthousiasme. Enfin, merci à ma petite mère Dominique, qui m'a toujours poussé à tout simplement faire ce qu'il me plaisait peu importe le temps que cela prenait. Merci de m'avoir supporté (dans tous les sens du terme) toutes ces années.

Table des matières

1 Introduction **1**
- 1.1 Description des modèles de mutations . . . 1
- 1.2 Le problème statistique . . . 4
- 1.3 Travail proposé . . . 6

2 État de l'art **9**
- 2.1 Motivations biologiques . . . 9
  - 2.1.1 L'expérience de Luria et Delbrück . . . 9
  - 2.1.2 Mise en évidence de biais via les données expérimentales . . . 10
- 2.2 Modélisations probabilistes . . . 11
  - 2.2.1 Approches déterministes . . . 13
  - 2.2.2 Approches stochastiques . . . 15
- 2.3 Estimation paramétrique . . . 25
  - 2.3.1 Méthode P0 . . . 30
  - 2.3.2 Méthode du Maximum de Vraisemblance (ML) . . . 33
  - 2.3.3 Méthode basée sur la fonction génératrice (GF) . . . 37

3 Modèles de mutations avec dépendance en âge **41**
- 3.1 Hypothèses et modèle . . . 42
  - 3.1.1 Justification biologique et intuitions . . . 42
  - 3.1.2 Définition du modèle . . . 44
- 3.2 Étude de la fonction génératrice . . . 48
  - 3.2.1 Écriture en temps fini des fonctions génératrices . . . 50
  - 3.2.2 Étude analytique de l'asymptotique . . . 58
  - 3.2.3 Étude « simplifiée » de l'asymptotique . . . 63
- 3.3 Cas particuliers . . . 71
  - 3.3.1 Modèle de Haldane . . . 71
  - 3.3.2 Taux instantanés proportionnels . . . 73
  - 3.3.3 Prise en compte de la dilution . . . 76

4 Test de fluctuation - Étude par simulation **80**
4.1 Extensions des estimateurs classiques . . . 81
4.1.1 Méthode P0 . . . 81
4.1.2 Méthode ML . . . 84
4.1.3 Méthode GF . . . 91
4.2 **flan** : un package R pour l'ANalyse de FLuctuation . . . 93
4.2.1 Fonctionnalités et interface . . . 94
4.2.2 Implémentation . . . 95
4.3 Étude par simulation des méthodes d'estimation . . . 96
4.3.1 Comparaison des méthodes d'estimation . . . 96
4.3.2 Étude des biais d'estimation . . . 100

5 Applications sur des jeux de données réelles **117**
5.1 Données de Boe et al. . . . 117
5.2 Données de David . . . 119
5.3 Données de Werngren et Hoffner . . . 127

6 Perspectives **141**
6.1 Modélisation . . . 141
6.2 Estimation . . . 142

A Résultats sur les processus de comptage **144**

Table des figures

1.1 Croissance d'un clone . . . 2
1.2 Croissance d'un clone avec apparition de mutations . . . 2

2.1 Croissance de deux clones suivant des modèles différents . . . 20
2.2 Croissance d'un clone avec des morts cellulaires . . . 21
2.3 Comparaison de différentes méthodes d'estimation du paramètre m . . . 27

3.1 Comparaison d'une croissance logistique et d'une croissance exponentielle . 44

4.1 Comparaison des trois méthodes d'estimation en terme de MSE sous le modèle LD . . . 98
4.2 Comparaison des trois méthodes d'estimation en terme de MSE sous le modèle H . . . 99
4.3 Estimations par la méthode GF en prenant en compte ou non les morts cellulaires ($\delta = 0.05$) . . . 101
4.4 Estimations par la méthode GF en prenant en compte ou non les morts cellulaires ($\delta = 0.2$) . . . 102
4.5 Estimations par la méthode ML en prenant en compte ou non les fluctuations du nombre final de cellules . . . 104
4.6 Estimations par la méthode GF sous les modèles LD et H sur des données simulées sous un modèle MM . . . 105
4.7 Estimations par la méthode GF sous les modèles LD et H sur des données simulées sous un modèle LD . . . 107
4.8 Estimations par la méthode GF sous les modèles LD et H sur des données simulées sous un modèle H . . . 108
4.9 Estimations par la méthode ML sous les modèles LD et H sur des données simulées sous un modèle MM . . . 109
4.10 Estimations par la méthode GF en prenant en compte ou non la dilution ($\zeta = 0.2$) . . . 111
4.11 Estimations par la méthode GF en prenant en compte ou non la dilution ($\zeta = 0.05$). . . . 112

4.12 Estimations par la méthode GF sous les modèles LD, H et LDI sur des données simulées sous un modèle LDI ($\eta_{\mu,\infty} = \log(100)$) 114
4.13 Estimations par la méthode GF sous les modèles LD, H et LDI sur des données simulées sous un modèle LDI ($\eta_{\mu,\infty} = \log(1000)$) 115

Liste des tableaux

1.1 Notations principales 8

2.1 Notations du chapitre 2 11
2.2 Notations des différents modèles de mutations 24
2.3 Comparaison de différentes méthodes d'estimation du paramètre π 29
2.4 Résumé du chapitre 2 40

3.1 Notations du chapitre 3 43
3.2 Notations des différents modèles de mutations 79

5.1 Estimations de m avec les données de Boe et al. 119
5.2 Estimations de m et ρ avec les données de Boe et al. 120
5.3 Estimations de π avec les données de David sous le modèle LD 122
5.4 Estimations de π avec les données de David sous le modèle H 123
5.5 Estimations de π et ρ avec les données de David sous le modèle LD en prenant en compte la dilution 125
5.6 Estimations de π et ρ avec les données de David sous le modèle H en prenant en compte la dilution 126
5.7 Estimations de π et ρ avec les données de David sous les modèles LDF et HF 128
5.8 Estimations de π avec les données de Werngren et Hoffner sous le modèle LD 130
5.9 Estimations de π avec les données de Werngren et Hoffner sous le modèle H 131
5.10 Estimations de π et ρ avec les données de Werngren et Hoffner sous le modèle LD en prenant en compte la dilution 132
5.11 Estimations de π et ρ avec les données de Werngren et Hoffner sous le modèle H en prenant en compte la dilution 133
5.12 p-valeurs des tests de comparaison (deux à deux) des probabilités de mutation des différentes souches des données de Werngren et Hoffner avec la méthode GF sous le modèle LD 137

5.13 p-valeurs des tests de comparaison (deux à deux) des paramètres de fitness des différentes souches des données de WERNGREN et HOFFNER avec la méthode GF sous le modèle LD 138
5.14 p-valeurs des tests de comparaison (deux à deux) des probabilités de mutation des différentes souches des données de WERNGREN et HOFFNER avec la méthode GF sous le modèle H 139
5.15 p-valeurs des tests de comparaison (deux à deux) des paramètres de fitness des différentes souches des données de WERNGREN et HOFFNER avec la méthode GF sous le modèle H 140

Chapitre 1

Introduction

Cette thèse propose une étude probabiliste et statistique des modèles de mutations, qui décrivent le processus d'apparitions rares et aléatoires de mutations au cours de la croissance d'une population de cellules. Nous nous intéresserons en particulier à l'extension des modèles existants au cas où les processus de croissance mis en jeu ne sont plus homogènes par rapport au temps.

Cette introduction s'organise en trois parties. La partie 1.1 est consacrée à la description du modèle et au problème d'une modélisation réaliste. La partie suivante expose le problème statistique suivant : obtenir une estimation dite « robuste » de la probabilité d'apparition d'une mutation au cours d'une division cellulaire, étant donné un échantillon de décomptes de cellules mutantes. Cette partie introduit également le principal obstacle à la résolution de ce problème qu'est l'existence de biais d'estimation dus aux hypothèses irréalistes sous lesquelles sont construits les estimateurs existants. Nous terminons cette introduction par la description du travail proposé dans cette thèse dans la partie 1.3.

1.1 Description des modèles de mutations

Les modèles de mutations font intervenir des notions de dynamique des populations, domaine qui s'intéresse aux variations de la taille d'une population au cours du temps. La modélisation probabiliste la plus connue dans ce domaine est le processus de Galton-Watson (voir entre autres [33, 6, 46]). Dans notre cas, la population considérée est composée de cellules et se développe par une succession de mitoses : chaque cellule, après une certaine durée, se divise en deux cellules filles, qui elles aussi se diviseront en deux cellules, etc. D'une cellule va donc descendre un ensemble de cellules, appelé *clone*. La croissance d'une population peut être schématisée sous la forme d'un arbre binaire où une cellule est représentée par une branche dont la taille correspond à la durée de développement de la cellule. Les embranchements correspondent alors aux événements de divisions. La Figure 1.1 est un exemple d'une telle représentation, dans le cas où les durées de développements des cellules sont toutes égales à une même constante $a = 1$. Dans cet exemple,

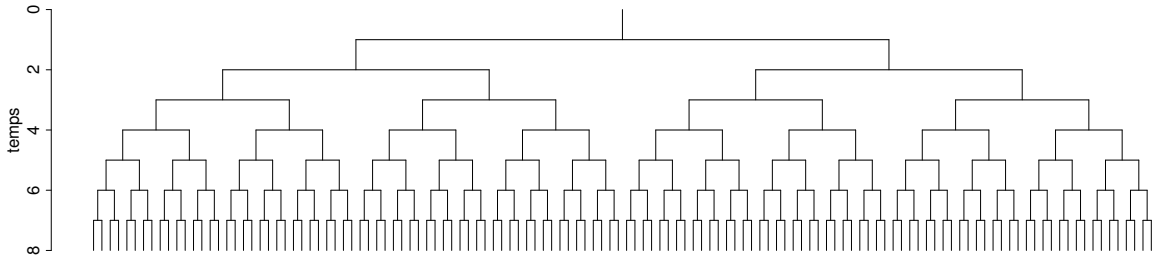

FIGURE 1.1 – **Croissance d'un clone.** Les durées de vie de toutes les cellules sont identiques et égales à $a = 1$.

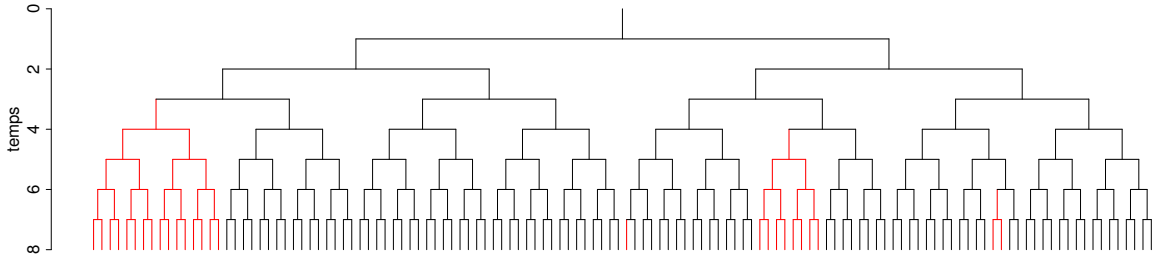

FIGURE 1.2 – **Croissance d'un clone avec l'apparition de mutations.** La durée de vie de toutes les cellules sont identiques et égales à $a = 1$. Les branches rouges correspondent aux cellules mutantes.

le clone croît jusqu'à l'instant 8 et contient donc 128 cellules.

Deux cellules filles sont censées posséder un génome identique à celui de la cellule mère. Cependant, une des deux filles peut, avec une faible probabilité, être sensiblement différente de la cellule mère. Un tel événement est appelé *mutation*. La cellule *mutante* va se développer selon le même mécanisme que les cellules *non-mutantes* (ou *normales*), donnant ainsi deux populations de cellules distinctes. La Figure 1.2 représente le même processus que la Figure 1.1 mais cette fois avec l'apparition de mutations au cours de la croissance du clone. Les mutantes sont représentées par les branches rouges. Dans cet exemple, il y a eu 5 événements de mutations, et la population est constituée à l'instant 8 de 27 cellules mutantes et 101 non-mutantes. Ces processus sont également caractérisés par le contexte asymptotique suivant : une très faible probabilité d'apparition d'une mutation lors d'une division cellulaire (en pratique de l'ordre de 10^{-11} à 10^{-9}), un très grand nombre de divisions cellulaires (en pratique de l'ordre de 10^6 à 10^8), mais un nombre restreint d'événements de mutation (en pratique de l'ordre de quelques unités). La description

probabiliste de ces processus d'apparitions rares et aléatoires de mutations au cours de la croissance d'une population de cellules constitue les *modèles de mutations*. Des études de ce genre de modèle étaient déjà présentes au début du siècle dernier par Yule [100], mais l'un des plus connus a été proposé par Luria et Delbrück [60], et porte depuis leur nom. Dans cette approche, les clones mutants et non-mutants se développent tous selon une croissance déterministe et exponentielle de taux unitaire, et la seule variable aléatoire est le nombre de mutations apparaissant durant l'expérience. D'autres approches dans lesquelles les clones mutants se développent selon des processus discrets et déterministes ont par la suite été proposées, dont en particulier le modèle de Lea et Coulson [56] (dont une formulation rigoureuse a été décrite par Armitage [4, sec 2.1.1]) ou encore la formulation de Haldane exposée par Sarkar [81]. Toutes ces formulations avaient l'avantage de donner des expressions explicites des probabilités de la loi du nombre final de mutantes. Cependant, comme le nombre de mutations est très faible par rapport au nombre de cellules normales, l'approximation déterministe du développement de chaque clone mutant n'est pas adaptée. Le premier modèle de mutation dans lequel les clones mutants se développent selon des processus stochastiques fut décrit par Lea et Coulson [56, p. 24] : chaque cellule mutante se divise au bout d'une durée exponentiellement distribuée. Ils proposèrent ainsi deux formulations différentes tout en obtenant la même expression de la loi du nombre final de mutantes. Bartlett [8, part. 4.31] retrouva analytiquement leur résultat dans le cas d'une formulation entièrement stochastique où les clones mutants et non-mutants se développent selon le même processus que celui proposé par Lea et Coulson.

Dans toutes ces formulations, le taux de croissance des cellules normales est supposé identique à celui des mutantes. La prise en compte de taux de croissance différents a été étudiée par de nombreux auteurs pour les formulations déterministes, parmi lesquels figurent Koch [47, p. 137], Stewart et al. [86] et Jones [38], ainsi que pour les modèles stochastiques (voir entre autres [4, part. 2.3.], [8, p. 134], [82, 38, 35, 101]). Ces formulations font ainsi intervenir le rapport du taux de croissance des cellules normales sur celui des mutantes, communément appelé *fitness* : une fitness plus petite que l'unité indique que les mutantes se développent plus rapidement que les non-mutantes, et inversement. La formulation de Lea et Coulson a été longtemps considérée comme la plus appropriée, et l'est encore de nos jours [96, 37], bien qu'il existe des justifications empiriques d'une fitness différente de 1 [26, 12]. D'autres données expérimentales apparues avant et après l'article de Lea et Coulson mirent en évidence d'autres sources de biais dans ce modèle et furent le sujet d'études mathématiques par la suite, comme les hypothèses de distribution exponentielle des durées de vie des cellules (voir les données de [40, 72], et les études de [43, 50, 97]), ou d'absence de morts cellulaires (voir les données de [83, 24], et les études de [3, part. 3.1], [18, 48, 98]). Nous reviendrons sur d'autres sources de biais dans le chapitre 2, où nous donnerons à la fois des exemples de données et les extensions appropriées.

L'ensemble des formulations décrites ci-dessus peuvent être interprétées comme la

composition des trois ingrédients suivants [31] : le nombre de mutations apparues durant la croissance de la population de cellules ; les durées de développement de chacun des clones issus de ces mutations ; la taille qu'un clone atteint après une durée de développement donnée. Selon les hypothèses de modélisation, il est parfois possible d'expliciter la distribution du nombre de mutantes observées en fin d'expérience. Dans ces cas, la loi obtenue est une loi à queue lourde dépendant du nombre moyen de mutations et du paramètre de queue qu'est la fitness. Selon la formulation, d'autres paramètres sont également impliqués, comme par exemple la probabilité de mort d'une cellule mutante.

Notons que ces modèles n'appartiennent pas au domaine de la génétique des population [23]. Ce dernier s'intéresse à l'apparition ou à la variation de fréquence d'allèles dans une population en faisant intervenir des interactions entre les individus. Cette remarque est d'autant plus importante que nous introduisons dans cette thèse une notion de fitness, dont la signification est sensiblement différente de celle donnée en génétique des population (voir entre autres le cours de Etheridge [22, Def. 5.1]). Il est cependant intéressant de noter certaines similarités comme le contexte asymptotique (mutations rares, grandes populations), l'enjeu statistique d'une estimation robuste du taux de mutation (à ne pas confondre avec la probabilité de mutation décrite dans cette thèse) ou encore l'évolution des modèles mathématiques directement reliée à l'amélioration des techniques de mesure, comme le séquençage ADN, et donc des données disponibles.

1.2 Le problème statistique

Les données disponibles sont des échantillons de nombres finaux de mutantes, qui peuvent être couplés avec des nombres totaux de cellules ou un nombre moyen de cellules en fin d'expérience. L'analyse de ces données constitue le domaine des *tests de fluctuation*, dont l'objectif statistique principal est l'estimation de la probabilité d'apparition d'une mutation lors d'une division cellulaire. Il est d'une importance cruciale dans plusieurs domaines de la médecine et de biologie, tels que l'oncologie (risque de rechute, croissance d'une tumeur), ou la microbiologie (résistance aux antibiotiques), d'estimer avec précision cette grandeur. Une estimation est la réalisation d'un estimateur, qui est une variable aléatoire dépendant de l'échantillon considéré. Il est donc nécessaire de disposer d'estimateurs robustes, c'est-à-dire qui sont peu sensible aux hypothèses de modélisation. Un « bon » estimateur vérifie deux propriétés : il doit être consistant, et sa loi asymptotique doit pouvoir être explicitée. Cela permet de quantifier ses fluctuations, et ainsi de construire des intervalles de confiance. Dans leur article pionnier, Luria et Delbrück ont présenté deux premières méthodes d'estimation du nombre moyen de mutations. La première méthode utilise le fait que la probabilité de n'observer aucune mutante est directement reliée au nombre moyen de mutations [60, eq. (5)] et est appelée méthode P0. Il s'agit d'un estimateur qui possède des propriétés de consistance et normalité asymptotique, et peut donc être considéré comme un « bon » estimateur. La deuxième méthode proposée dans

l'article, dite des moyennes, s'appuie sur une relation entre le nombre moyen de mutantes, la taille de l'échantillon et le nombre final de cellules [60, eq. (8)]. Cette méthode ne devrait pas être employée : l'estimateur ainsi construit n'a pas d'espérance, et n'est donc pas consistant. Dans les deux cas, l'estimation de la probabilité de mutation est déduite en divisant celle du nombre de mutations par le nombre total de cellules. De nombreuses autres méthodes ont été proposées par la suite : elles ont par exemple été exposées par Rosche et Foster [79] ou encore Foster [25]. La majorité de ces méthodes tentent de contourner l'aspect queue lourde en considérant des estimateurs construits à partir de la médiane du nombre de mutantes. Certaines de ces méthodes donnent de bons résultats en pratique, mais les propriétés citées ci-dessus ne sont pas vérifiées ou vérifiables. En effet, la médiane empirique est un bon estimateur de la médiane, mais uniquement dans le cas d'une loi continue. Des méthodes d'estimation avec de bonnes propriétés ont également été proposées, et devraient être préférées. Une première possibilité est l'utilisation du Maximum de Vraisemblance [57, chap. 6]. Les probabilités du nombre de mutantes ainsi que leurs dérivées par rapport au nombre moyen de mutations et la fitness peuvent en effet être explicitées via des algorithmes (voir entre autres [75, 103, 31]). Cependant, l'optimisation de la vraisemblance se heurte en pratique des obstacles de stabilité et peut être numériquement très coûteuse. Ces difficultés sont directement liées à la valeur maximale de l'échantillon. En pratique, il est possible de réduire ces effets de queue via différentes méthodes qui ont été exposées par Wilcox [95, part. 2.2]. L'une d'elles, appelée *winsorisation*, consiste à remplacer n'importe quelle valeur de l'échantillon qui dépasse une certaine borne par cette même borne. Une autre alternative basée sur l'utilisation de la fonction génératrice a été proposée par Hamon et Ycart [31, part. 4]. Cette méthode donne des estimateurs du nombre de mutations et de la fitness comparable au Maximum de Vraisemblance en terme de précision, mais beaucoup plus stable en pratique. Les estimateurs construits avec ces méthodes sont consistants, et asymptotiquement normaux. Nous nous intéresserons dans cette thèse uniquement à ces deux méthodes, ainsi que la méthode P0 de Luria et Delbrück qui possède également ces propriétés. Nous définirons plus précisément ces méthodes dans le chapitre 2. Estimer les paramètres d'un modèle, quand la réalité en suit un autre, conduit nécessairement à un biais d'estimation. De même que les modèles de mutations ont beaucoup évolué depuis la première approche de Luria et Delbrück, les méthodes d'estimation doivent également être ajustées afin que les estimateurs obtenus soient le plus robuste possible. Une des sources de biais les plus mentionnées jusqu'à présent est la prise en compte de la *dilution*, également appelée *plating efficiency* [86, 84, 38, 2]. D'autres sources de biais et leurs conséquences statistiques, comme les hypothèses de durées de vie exponentielles [97] ou des morts cellulaires [98], ont également été traitées. Les extensions existantes des méthodes d'estimations citées ci-dessus seront exposées dans le chapitre 2, et nous mettrons en évidence certaines sources de biais à l'aide d'études de simulation dans le chapitre 4.

1.3 Travail proposé

Dans cette thèse, nous commencerons par étudier une source de biais peu traitée jusqu'à présent : l'homogénéité en temps des processus de croissance. Une telle hypothèse ne donne que deux possibilités en terme d'évolution de la population : soit elle suit une croissance exponentielle (et tend donc vers l'infini), soit elle s'éteint [6, 36]. Or, en pratique, une croissance de type logistique est observée, comme par exemple dans l'étude de la croissance d'une tumeur effectuée par Laird [54]. Cela est dû au fait que l'expérience a lieu dans un environnement où la quantité de ressources ou de place est limitée. Parmi les modélisations de ce genre de croissance non-exponentielle, figurent des approches déterministes telles que celles proposées par Verhulst [90, 91]. Une description à l'aide d'équation différentielle stochastique des modèles de type logistique a été donnée par Allen [1, part. 9.4.2]. Des études mathématiques sur des modèles avec des croissances logistiques quelconques ont également été menées par de nombreux auteurs [88, 89, 73, 55]. Ce type de croissance a également été étudié dans des contextes différents, comme celui des modèles avec immigration [64] ou en « dynamique adaptative » [14]. Plus généralement, Kendall [42] a exposé des processus de naissance et de mort non Markovien avec lesquels il est possible d'ajuster la croissance d'une population à une fonction du temps quelconque « appropriée ». Une approche des modèles de mutations prenant en compte le taux de divisions décroissants avait été proposée par Stewart et al. [86], et une formulation discrète ne faisant aucune hypothèse sur le type de croissance des clones mutants a été décrite par Houchmandzadeh [34]. Cependant, il n'existe pas de résultat sur la loi du nombre final de mutantes, ou d'étude statistique du biais d'estimation lorsque les durées de vie des cellules ne sont plus identiquement distribuées. Nous proposons dans cette thèse un modèle de mutation dans lequel les processus de croissance ne sont plus homogènes dans le temps : nous supposons que l'instant de division de chaque cellule, quel que soit son type, dépend de sa date de naissance (c'est-à-dire l'instant de division de sa cellule mère). Il est alors possible d'ajuster la croissance d'un clone à n'importe quelle fonction du temps. Cette approche permet également de généraliser les extensions existantes, comme par exemple la prise en compte des morts cellulaires. L'étude probabiliste des modèles de mutation inhomogènes en temps a été menée selon deux approches, présentées dans le chapitre 3. La première aborde le problème analytiquement et généralise la démonstration proposée par Bartlett [8, p. 155] pour le cas où toutes les cellules ont des durées de vie *i.i.d.* et exponentielles. Cette première étude a fait l'objet d'un article soumis [66]. La deuxième approche, dite *simplifiée* permet de démontrer que n'importe quel modèle de mutation inhomogène peut se décomposer selon trois ingrédients :

1. l'apparition de mutations aléatoires au cours d'un processus de croissance de population. Les divisions cellulaires sont très nombreuses, et la probabilité de mutation est faible, ce qui justifie une approximation poissonnienne pour le nombre de mutations survenues ;
2. les instants de mutations. Le processus d'apparition des mutations étant équivalent

en loi à un Processus de Poisson Non-Homogène donné, il est possible d'en déduire la loi asymptotique des instants de mutations ;

3. le nombre de cellules qu'un clone issu d'une cellule mutante née à un instant donné, atteint au bout d'une certaine durée.

Ainsi, la décomposition en trois ingrédients décrite dans la partie 1.1 est généralisée au cas inhomogène en temps. De plus, cette décomposition permet de construire un algorithme de simulation rapide, indispensable pour l'étude empirique des méthodes d'estimations. En effet, les méthodes d'estimation d'intérêt sont étendues à cette nouvelle formulation, tout en conservant les précédents ajustements effectués et leurs propriétés de consistance et de normalité asymptotique. L'approche simplifiée et les méthodes statistiques feront l'objet d'un article qui est en cours de rédaction.

Une autre partie importante de ce travail porte sur la reprise des précédentes études de sources de biais (prise en compte des morts cellulaires pour le modèle de Haldane, des fluctuations des nombres finaux pour toutes les méthodes d'estimation, etc.). En particulier, nous reprenons dans la section 3.3 l'étude de la dilution. Nous formalisons rigoureusement le problème, ce qui nous permet de construire des estimateurs consistants et asymptotiquement normaux qui sont plus robustes que la correction de Stewart et al. [86]. Nous constaterons d'ailleurs que cette dernière ne devrait pas être employée systématiquement. La dilution étant une manipulation largement répandue dans les tests de fluctuations, une étude statistique de cette source de biais fera l'objet d'une publication.

Le package R **flan** (pour **FL**uctation **AN**alysis) a été développé au cours de ce travail et est également présenté ici. Il fournit des fonctions dédiées à la loi du nombre final de mutantes et à l'inférence statistique selon différentes méthodes et à partir de différents types d'échantillons. Le package est d'ores et déjà disponible sur le CRAN [68] et sur GitHub [69] et un article lui a été dédié [70]. Ce dernier a été récemment publié dans le R Journal. Il a également été utilisé dans des études [67] effectuées sur les données de Werngren et Hoffner [94].

Cette thèse est composée de cinq chapitres. Le chapitre 2 décrit dans un premier temps les motivations biologiques à l'origine de cette thèse et des formulations de modèles de mutations en général, ainsi qu'un état de l'art un peu plus exhaustif que ce soit sur les modélisations probabilistes et les résultats associés, que sur le problème de l'estimation de la probabilité de mutation. Le chapitre 3 est ensuite dédié à la partie probabiliste de ce travail : après avoir donné une formalisation rigoureuse du modèle, nous exposons les différents résultats obtenus pour la loi du nombre final de mutantes. La partie statistique est ensuite traitée dans le chapitre 4 : les extensions des méthodes d'estimation y sont présentées, ainsi que le package **flan**. Les différentes sources de biais d'estimation sont mises en évidence via des études de simulations à grande échelle avec **flan**. Le package est également appliqué à des données réelles dans le chapitre 5. Enfin nous étudierons quelques pistes de réflexions suite à ce travail dans le chapitre 6. Tout au long de cette thèse, nous nous efforçons de conserver les mêmes notations. La table 1.1 référence toutes les notations principales qui resteront valables jusqu'à la fin de cette thèse. Lorsque des notations

$N(t)$	Nombre de cellules normales vivantes à un instant t
N	Nombre final de cellules normales
$M(t)$	Nombre de cellules mutantes vivantes à un instant t
M	Nombre final de cellules mutantes
$\widetilde{M}(s,t)$	Taille à un instant t du clone démarré par une cellule mutante née à un instant s
$Z(t)$	Nombre de mutations apparues dans un intervalle de temps $[0\,;t]$
Z	Nombre total de mutations apparues durant l'expérience
$m(t)$	Nombre moyen de mutations dans un intervalle de temps $[0\,;t]$
m	Nombre moyen de mutations apparaissant au cours de l'expérience
π	Probabilité d'apparition d'une mutation lors d'une division cellulaire
n	Nombre initial de cellules normales
τ	Instant de fin d'expérience

TABLE 1.1 – **Notations principales**

supplémentaires seront nécessaires, nous les préciserons dans le chapitre concerné.

Chapitre 2

État de l'art

Les modèles de mutations et les tests de fluctuations sont apparus avec les résultats de l'expérience menée par Luria et Delbrück [60] en 1943. Depuis, les formulations et les méthodes d'estimation ont beaucoup évolué. Ce chapitre constitue un historique non-exhaustif de ces évolutions. Nous commençons par décrire dans la partie 2.1 l'expérience de Luria et Delbrück et donnerons également des résultats empiriques justifiant les évolutions du modèle. Les différentes modélisations probabilistes sont ensuite exposées dans la partie 2.2. Nous nous intéresserons finalement aux méthodes d'estimations, ainsi qu'aux ajustements nécessaires en fonction des différentes sources de biais identifiées jusqu'à présent, dans la partie 2.3.

2.1 Motivations biologiques

2.1.1 L'expérience de Luria et Delbrück

L'expérience de Salvador Luria et Max Delbrück est l'une des plus connues dans le domaine des tests de fluctuations. Le but de l'expérience était de déterminer si la mutation d'une bactérie était induite par l'environnement (théorie lamarckienne) ou avait lieu de manière spontanée au cours de la croissance de la population (théorie darwinienne). Pour cela, ils laissèrent croître durant plusieurs générations et de manière indépendante plusieurs cultures de *Escherichia Coli*. Après avoir extrait la même quantité dans chacune des cultures, ils exposèrent ces échantillons à un virus, et observèrent en fin d'expérience le nombre de bactéries mutantes, c'est-à-dire résistantes au virus. La principale caractéristique des données obtenues est une forte fluctuation, avec une apparition fréquente de grandes valeurs, surnommées par la suite *jackpots*. Ces observations tendaient donc à montrer que les mutations étaient survenues avant l'exposition au virus, et allaient donc dans le sens d'une évolution darwinienne. De fait, ils émirent l'hypothèse que le nombre moyen de mutations apparues était proportionnel au nombre de divisions cellulaires, et non au nombre de bactéries exposées. À partir de cette hypothèse, ils proposèrent une

première modélisation du processus d'apparition de mutations au cours du développement d'une culture de cellules, et construisirent les premiers estimateurs du nombre moyen de mutations, et donc de la probabilité de mutation. L'expérience de Luria et Delbrück ayant été par la suite adaptée à d'autres types de cellules, elle est souvent considérée comme pionnière dans le domaine des tests de fluctuations.

2.1.2 Mise en évidence de biais via les données expérimentales

Une formulation mathématique ne peut être validée que si elle est cohérente avec les données expérimentales. Ainsi, lorsque des données mettent en évidence des biais à prendre en compte, il est nécessaire d'affiner les modèles mathématiques. En l'occurrence, les premiers modèles de mutations et les différentes méthodes d'estimation ont été construits sous des hypothèses irréalistes, et ont depuis subi de nombreuses extensions. Nous nous intéresserons à ces modifications dans les deux parties suivantes, et donnons pour le moment quelques justifications empiriques.

La présence de jackpots dans les observations de Luria et Delbrück sont typiques d'une loi à queue lourde. Il s'avère que le paramètre de fitness, qui était initialement fixé à 1, est en fait le paramètre de queue. Rappelons que ce paramètre est le rapport du taux de croissance des cellules normales sur celui des mutantes. Ainsi, plus ce paramètre est faible, plus les cellules mutantes se développent rapidement en comparaison des non-mutantes et plus les jackpots sont élevés. Des justifications empiriques d'une fitness différente de 1 peuvent être trouvées dans certaines études où elle a été mesurée [26, 12], mais elle est encore malgré tout ignorée dans d'autres expériences, (voir entre autres [96, 37]). Les modèles de mutations classiques supposent également que les cellules se divisent forcément en deux. En réalité, chaque cellule peut avec probabilité non-nulle mourir au lieu de se reproduire. Bien que la probabilité de mort semble être relativement faible en pratique [83, 24], il en découle néanmoins un biais d'estimation. Comme précisé plus tôt, le paramètre d'intérêt est la probabilité de mutation. Cependant, les méthodes utilisées permettent d'estimer le nombre moyen de mutations. L'estimation de la probabilité de mutation est ensuite calculée en divisant celle du nombre moyen de mutations par le nombre final de cellules. L'approche utilisée la majeure partie du temps est de relever le nombre final de cellules dans chaque culture, puis de ne considérer que la moyenne de ces décomptes [41, p. 1062]. Or, il est difficile en pratique d'éviter d'obtenir des nombres finaux de cellules sans fluctuations, et ce même en appliquant des contrôles expérimentaux précis [49]. Bien qu'il n'y ait que peu de données de nombres finaux de cellules ou avec des informations sur ces fluctuations (par ex. [17, 94]) ces variations devraient être prises en compte.

Les dernières sources de biais que nous allons citer concernent les processus de croissance des clones : les durées de développement des cellules sont supposées indépendantes et identiquement distribuées selon une loi exponentielle. Bien que l'hypothèse d'une distribution exponentielle facilite l'approche mathématique du problème, il a été remarqué très tôt, avant même l'expérience de Luria et Delbrück, que cela ne correspondait pas à la

ν	Taux de croissance des cellules normales
μ	Taux de croissance des cellules mutantes
$\rho = \dfrac{\nu}{\mu}$	Paramètre de fitness

TABLE 2.1 – **Notations du chapitre 2**

réalité [40, 78]. Des données montrant que cette hypothèse n'est pas réaliste ont également été exposées par MURPHY et al. [72]. Des preuves de corrélations entre une cellule et ses descendantes, ainsi qu'entre deux cellules issues de la même division ont également été mises en évidence expérimentalement par WANG et al. [92]. Cependant, nous constaterons dans la partie suivante que même si ces deux hypothèses sont prises en compte, la croissance générale reste exponentielle, du fait de l'homogénéité des processus de croissance. Or, une population de cellules se développe dans un environnement qui contient une quantité finie de ressources, et la croissance de la population observée a une allure logistique [54]. Nous nous intéressons dans cette thèse à cette dernière source de biais.

2.2 Modélisations probabilistes

Dans cette partie, nous décrivons quelques formulations mathématiques des modèles de mutations qui ont été proposées jusqu'à présent. Une étude en détail des principales formulations proposées durant la seconde moitié du XX$^{\text{e}}$ siècle a été effectuée par ZHENG [101]. Nous ne ferons donc que décrire brièvement ces approches, et nous attarderons un peu plus sur les formulations plus récentes. Dans cette partie, les notations exposées dans la table 1.1 seront employées, auxquelles nous ajoutons celles de la table 2.1. Tout au long de cette thèse, la plupart des résultats seront exprimés en terme de fonction génératrice. Notons U le disque unité ouvert de $\mathbb{C}$. Si N est une variable aléatoire discrète, sa fonction génératrice G_N est la fonction à valeurs dans $\mathbb{R}^+$ définie pour tout $z \in U$ par

$$\begin{aligned} G_N(z) &:= \mathbb{E}\left[z^N\right] \\ &= \sum_{k \geqslant 0} \mathbb{P}[N = k] z^k \,. \end{aligned}$$

Les propriétés de base utilisées ici sont les suivantes :

1. Si N et M sont deux variables aléatoires discrètes indépendantes, alors la fonction génératrice de la somme $N + M$ est donnée par le produit des fonctions génératrices de N et M, c'est-à-dire pour tout $z \in U$:

$$G_{N+M}(z) = G_N(z) G_M(z) \,.$$

2. La convergence faible des mesures est équivalente à la convergence simple des fonctions génératrices.

Il est également nécessaire de préciser que les approches qui suivent se placent dans un certain contexte asymptotique. Cependant les premières approches que nous présentons ne définissent pas ou peu ce cadre. Nous ne le définirons que plus tard dans cette partie, tout en gardant les intuitions suivantes : la probabilité de mutation π est de l'ordre de 10^{-11} à 10^{-9}, le nombre final de cellules normales N (et a fortiori le nombre de divisions cellulaires) est de l'ordre de 10^8 à 10^9 et le nombre m de mutations survenues durant le processus est de l'ordre de l'unité.

Les hypothèses basiques de modélisation sont les suivantes :

— la population initiale est composée de n cellules normales et aucune mutante ;

— lorsqu'une cellule normale termine son développement, elle est remplacée par :

  — une cellule normale et une cellule mutante avec probabilité π ;

  — deux cellules normales avec probabilité $1 - \pi$;

— lorsqu'une cellule mutante termine son développement, elle est remplacée par deux cellules mutantes ;

— les événements de mutations sont indépendants des processus de croissance.

— les cellules se développent indépendamment les unes des autres.

La dernière hypothèse d'indépendance implique que deux clones distincts (c'est-à-dire issus de deux cellules mères distinctes) se développent indépendamment l'un de l'autre. Considérer la taille d'une population initialement composée de n cellules revient donc à considérer la somme de n populations indépendantes, chacune initialement composée d'une unique cellule.

Nous allons à présent nous intéresser à deux classes principales de formulations différentes. La première regroupe les modélisations dans lesquelles les clones (mutants ou non) croissent selon une fonction déterministe. Ainsi, le seul processus aléatoire intervenant dans ces modèles est celui des événements de mutations. La deuxième classe contient toutes les formulations dans lesquelles les clones d'au moins un des deux types de cellules (mutants ou non) se développent selon des processus aléatoires. Dans cette partie, nous considérerons les formulations appartenant à la première classe comme des approches déterministes, tandis que celles appartenant à la deuxième classe seront désignées comme étant stochastiques.

Afin de classifier explicitement les différentes modélisations qui vont suivre, nous allons utiliser dans cette introduction une notation directement inspirée de celle proposée par Kendall [44, p. 61]. Cette notation s'exprimera sous la forme de deux lettres et suivra les règles suivantes :

1. La première lettre indiquera le modèle de croissance des cellules normales ;
2. La deuxième lettre indiquera le modèle de croissance des cellules mutantes ;

3. La liste des lettres possibles est la suivante :
   - D : modèle déterministe classique (croissance exponentielle) ;
   - DD : modèle déterministe discret ;
   - DG : modèle déterministe quelconque (croissance non exponentielle) ;
   - M : modèle markovien (durées de vie *i.i.d.* et exponentielles) ;
   - MD : modèle markovien discret ;
   - G : modèle plus généraliste (par exemple durées de vie *i.i.d.* mais selon une loi non-exponentielle).

À titre d'exemple, la notation D/M indique que la population des cellules normales croît suivant une fonction exponentielle déterministe tandis que les cellules mutantes se développent selon un processus stochastique dans lequel les durées de vie des cellules sont distribuées indépendamment et identiquement selon une loi exponentielle.

Les premières approches que nous allons décrire dans la sous-partie 2.2.1 sont des formulations déterministes où seule l'apparition des mutations se comporte aléatoirement. Nous n'en donnerons qu'une brève description. La sous-partie 2.2.2 est dédiée à différentes formulations stochastiques et sera plus détaillée que la première.

2.2.1 Approches déterministes

Le premier modèle de mutation à proprement parler est la formulation D/D décrite par Luria et Delbrück [60]. Dans cette approche, le nombre de cellules normales $N(t)$ croît selon un taux déterministe ν. Ce taux est directement relié au choix de l'unité de temps. Les auteurs ont fixé cette unité de temps à un cycle de division d'une bactérie divisé par $\log(2)$, de sorte que ν vaut 1. Ainsi :

$$N(t) = N(0)\mathrm{e}^t\,. \tag{2.1}$$

La croissance de la population de cellules mutantes est également exponentielle selon un taux lui aussi unitaire. La taille à un instant t d'un clone initié par une mutante née à un instant s est ainsi donnée par

$$\widetilde{M}(s,t) = \mathrm{e}^{t-s}\,. \tag{2.2}$$

En supposant que le nombre de mutantes est négligeable par rapport au nombre total de cellules, le nombre moyen de mutations apparues au cours d'un intervalle de temps $[0\,;t]$ est donc donné par $m(t) = \pi(N(t)-n)$. Sous l'hypothèse que la probabilité de mutation π est faible et que le nombre de divisions $N(t)-n$ est grand, le nombre de mutations $Z(t)$ suit alors une loi de Poisson de paramètre $m(t)$ (Loi des Petits Nombres [5, p. 321]).

L'espérance ainsi que la variance du nombre de mutantes $M(t)$ peuvent alors être explicitées. Une généralisation au cas où le taux de croissance des mutantes μ est différent de 1 a été proposée plus tard par Koch [47, p. 137]. Comme le remarqueront Luria et Delbrück, il est difficile avec cette formulation d'obtenir une forme explicite pour la densité

du nombre de mutantes $M(t)$. Par la suite d'autres auteurs tels que Bailey [7, sec. 10.4] ou Crump et Hoel [16, p. 243] mettront en avant le fait que la variable $M(t)$ varie via des processus continus (croissance des clones mutants) et discrets (apparition des mutations), créant ainsi des obstacles analytiques dans l'étude mathématique. Un premier moyen naturel de contourner cette difficulté consiste à modéliser la croissance d'un clone selon un processus déterministe discret. Une formulation D/DD a ainsi été proposée par Lea et Coulson [56, p. 269]. Dans cet article, l'échelle de temps classique est remplacée par la subdivision de l'intervalle $[1\,;N]$ suivante :

$$I_k = \left[\frac{N}{2^{k+1}}\,;\frac{N}{2^k}\right] .$$

En conséquence, une mutation apparue dans l'intervalle de temps I_k aura le temps de se diviser $k-1$ fois et la taille finale du clone issu de cette mutation sera ainsi 2^{k-1}. Une formalisation plus rigoureuse du modèle a été donnée par Armitage [4, part. 2.1.1.]. Si cette approche est considérée comme la réponse à la question de l'expression des probabilités du nombre final de mutantes dans la population, des calculs non publiés de Haldane furent exposés à la fin du XX$^{\text{e}}$ siècle par Sarkar [81] (et plus récemment par Zheng [105]). Nous y reviendrons dans la prochaine sous-partie. À partir de ce raisonnement, les auteurs ont obtenu une forme explicite pour la fonction génératrice du nombre final de mutante M en fonction du nombre moyen de mutations m :

$$\psi(z) = \exp\left\{-m\left(1-\sum_{i\geqslant 1}\frac{z^i}{i(i+1)}\right)\right\} ,$$

plus généralement écrite sous la forme suivante :

$$\psi(z) = (1-z)^{m\frac{1-z}{z}} . \tag{2.3}$$

L'expression (2.3) a d'ailleurs été généralisée au cas où les taux ν et μ ne sont plus unitaires par Stewart et al. [86] :

$$\psi(z) = \exp\left\{-m\left(1-\sum_{i\geqslant 1}\left(\frac{1}{i^\rho}-\frac{1}{(i+1)^\rho}\right)z^i\right)\right\} . \tag{2.4}$$

Une généralisation au cas DG/DD est proposée par Houchmandzadeh [34, part. II.]. L'approche du problème y est similaire à celle énoncée par Lea et Coulson. Les résultats sont exposés en termes de fonction génératrice des cumulants, pour une fitness $\rho = \frac{\nu}{\mu}$ égale à 1 ou non. Cependant, pour les mêmes raisons que celles exposées pour le modèle de Luria-Delbrück, seules la moyenne et la variance de M y sont explicitées.

La prochaine sous-partie s'intéresse aux formulations du problème dans lesquelles les processus de croissance ne sont plus déterministes.

2.2.2 Approches stochastiques

Une première modélisation stochastique D/M du problème où les durées de vie des mutantes sont exponentiellement *i.i.d.* selon un taux unitaire est décrite par Lea et Coulson [56, p. 266]. La population de cellules normales se développe toujours selon (2.1). Sous cette hypothèse, Lea et Coulson retrouvent la même fonction génératrice que (2.3). Une généralisation au cas où la croissance des non-mutantes n'est plus exponentielle (modèle DG/MD) a été récemment proposée Houchmandzadeh [34] :

$$\psi(z) = \left(1 - \left(1 - \frac{n}{N}\right) z\right)^{\pi N \frac{1-z}{z}} . \tag{2.5}$$

L'expression (2.5) avait déjà été exposée par Zheng [101, eq. 65], mais dans un contexte différent : il s'agissait d'une étude en temps fini du cas où la croissance des cellules non-mutantes est exponentielle.

Comme le prouvera dans un premier temps Bartlett [8, part. 4.31], (2.3) est également exact pour une formalisation M/M dans laquelle les durées de vie des non-mutantes et des mutantes sont *i.i.d.* selon une loi exponentielle de même taux ν. La démonstration exposée s'appuie uniquement sur des calculs élémentaires et la résolution d'une équation différentielle vérifiée par la fonction génératrice double du couple des nombres de non-mutantes et de mutantes. En se plaçant dans le cadre asymptotique classique et en supposant que τ est également grand, il obtint la fonction génératrice (2.3).

Par la suite, une généralisation de cette formulation dans laquelle :

— les durées de vie des cellules normales sont *i.i.d.* selon une loi exponentielle de taux ν ;

— les durées de vie des mutantes sont *i.i.d.* selon une loi exponentielle de taux μ ;

fut proposée par Armitage [4, part. 2.3.]. Cependant, il n'obtint pas de forme explicite pour la fonction génératrice du nombre final de mutantes. Il est en fait possible d'adapter la démonstration de Bartlett [8, part. 4.31] au cas où $\nu \neq \mu$. Comme ces outils seront réutilisés dans la suite de cette partie et de cette thèse, nous allons donner un peu plus de détails sur cette généralisation.

Théorème 2.2.1. *Soient $\pi = (\pi_n)_{n\in\mathbb{N}}$ et $\tau = (\tau_n)_{n\in\mathbb{N}}$ deux suites de réels positifs, et une constante $m > 0$ telles que :*

$$\lim_{n\to+\infty} \pi_n = 0 , \quad \lim_{n\to+\infty} \tau_n = +\infty , \quad \lim_{n\to+\infty} \pi_n n \mathrm{e}^{\nu\tau_n} = m .$$

Lorsque n tend vers l'infini, la fonction génératrice du nombre de mutantes au temps τ_n dans les clones issus de n cellules normales nées au temps 0 converge vers la fonction génératrice suivante :

$$\psi(z) = \exp(-m(1 - h(z))) , \tag{2.6}$$

avec

$$h(z) = \rho \int_0^1 \frac{z v^\rho}{1 - z(1-v)} \mathrm{d}v . \tag{2.7}$$

Il s'agit du théorème 2.2.3 proposé par Hamon et Ycart [31, Théo. 1.1] appliqué au cas M/M. Nous reviendrons sur ce résultat par la suite. Les hypothèses sur les trois limites de ce théorème formalisent le contexte asymptotique décrit au début de ce chapitre. En particulier, la troisième limite traduit le fait que le nombre de mutations reste fini. La fonction génératrice h représente la taille finale de n'importe quel clone mutant. En appliquant le changement de variable $v = \mathrm{e}^{-\mu u}$ dans (2.7), nous obtenons l'expression suivante :

$$h(z) = \int_0^\infty \frac{z\mathrm{e}^{-\mu u}}{1 - z(1 - \mathrm{e}^{-\mu u})} \nu \mathrm{e}^{-\nu u} \mathrm{d}u \,.$$

Il s'agit de la fonction génératrice d'un mélange exponentiel de loi géométriques de paramètre $\mathrm{e}^{-\mu s}$ (c'est-à-dire un processus de Yule de paramètre μ). Cette interprétation est expliquée dans un cadre plus général à à la suite du théorème 2.2.3. La preuve analytique du théorème 2.2.1 nécessite tout d'abord d'exprimer la fonction génératrice du nombres de mutantes pour un temps fini.

Proposition 2.2.1. *La fonction génératrice du nombre à un instant t de mutantes dans le clone issu d'une cellule normale est donnée par*

$$\begin{aligned}\psi(z,t) =& \mathrm{e}^{n\nu t} \left(1 - z + z\mathrm{e}^{\mu t}\right)^{\pi\rho} \\ & \times \left\{1 - \nu(1-\pi) \int_0^t \mathrm{e}^{-\nu s} \left(1 - z + z\mathrm{e}^{\mu s}\right)^{-\pi\rho} \mathrm{d}s \right\}^{-1} . \end{aligned} \tag{2.8}$$

La démonstration n'est pas détaillée ici : une généralisation de cette démonstration au cas non-homogène en temps est en effet donnée dans le chapitre 3, et son application à la fonction génératrice (2.8) en est un cas particulier.

Si $\rho = 1$, la partie intégrale de (2.8) peut être calculée et nous retrouvons le résultat obtenu par Lea et Coulson, ainsi que Bartlett. Ainsi la fonction génératrice du nombre à un instant t de mutantes dans les clones issus de n cellules normales est donnée par

$$\begin{aligned}\psi_n(z,t) =& \psi(z,t)^n \\ =& \mathrm{e}^{n\nu t} \left(1 - z + z\mathrm{e}^{\mu t}\right)^{n\pi\rho} \\ & \times \left\{1 - \nu(1-\pi) \int_0^t \mathrm{e}^{-\nu s} \left(1 - z + z\mathrm{e}^{\mu s}\right)^{-\pi\rho} \mathrm{d}s \right\}^{-n} . \end{aligned} \tag{2.9}$$

La démonstration du théorème 2.2.1 consiste en l'étude asymptotique des deux termes multiplicateurs de (2.8), afin de prouver la convergence de (2.9) pour tout $z \in U$. Comme il s'agit d'une démonstration très calculatoire, nous n'en donnerons que le cheminement :

Démonstration du théorème 2.2.1. La preuve est constituée des quatre étapes suivantes :

1. Déduire du développement en série entière de

$$(1+x)^{\pi\rho} = \exp(\pi\rho \log(1+x)) \,,$$

le lemme suivant :

Lemme 2.2.1. *Pour tout* $\pi \in]-1;1[$, $x > -1$ *:*

$$|(1+x)^{\rho\pi} - (1 + \pi\rho\log(1+x)| \leqslant \pi^2(1+|x|)^{\rho}\,.$$

2. Utiliser le lemme 2.2.1 afin de prouver que le premier terme de (2.8) s'écrit :

$$\mathrm{e}^{\nu t}\left(1 - z + z\mathrm{e}^{\mu t}\right)^{\pi\rho} = \mathrm{e}^{\nu t}(1 + \pi\rho\log(1-z) + \pi^2 C_1 + \pi\mathrm{e}^{\nu t}C_2)\,, \tag{2.10}$$

où C_1 et C_2 sont uniformément bornées par rapport à π et t.

3. Utiliser le lemme 2.2.1 afin de prouver que la partie intégrale de (2.8) s'écrit :

$$\begin{aligned}\int_0^t \mathrm{e}^{-\nu s}\left(1 - z + z\mathrm{e}^{\mu s}\right)^{-\pi\rho} = &\frac{1}{\nu}(1 - \mathrm{e}^{-\nu t} + \pi h(z)) \\ &+ \pi^2 C_3 + \pi\mathrm{e}^{-(\mu+\nu)t}C_4 + \pi\mathrm{e}^{-\nu t}C_5\,,\end{aligned} \tag{2.11}$$

où C_3, C_4 et C_5 sont uniformément bornées par rapport à π et t. Cette partie est la plus longue et fastidieuse des trois étapes.

4. Injecter (2.10) et (2.11) $\psi(z,\tau_n)^n$ avec π_n dans (2.9) afin d'obtenir :

$$\psi(z,\tau_n) = \left(1 + \pi_n\mathrm{e}^{\nu\tau_n}(1 - h(z)) + o(\pi_n,\tau_n)\right)^{-n}\,,$$

où $o(\pi_n,\tau_n)$ est telle que

$$\lim_{n\to+\infty} n o(\pi_n,\tau_n) = 0\,,$$

dans le cadre asymptotique de l'énoncé du théorème 2.2.1. D'où le résultat.

□

Comme mentionné plus haut, le théorème 2.2.1 a été proposé par Hamon et Ycart [31] pour une formulation G/M dans laquelle les durées de vie des non-mutantes sont *i.i.d.* selon une loi quelconque. Cette extension fait intervenir quelques notions liées aux processus de Bellman-Harris (voir entre autres [33, 6]), que nous allons brièvement décrire ici. Nous donnerons les énoncés dans le contexte suivant :

— la population est initialement constituée d'un unique individu ;

— les durées de vie des individus sont *i.i.d.* selon une fonction de répartition G ;

— lorsqu'un individu termine sa vie, il donne naissance à un nombre aléatoire K d'individus identiques.

De plus, la population ne s'éteint pas avec une probabilité positive, c'est-à-dire $\mathbb{E}[K]$ est supérieur à 1 (cas supercritique). La taille de la population à un instant t sera notée par $Y(t)$. La première notion concerne le *paramètre malthusien*, également appelé *taux de croissance*.

Définition 2.2.1 (Paramètre malthusien)**.** *Le paramètre malthusien ν associé à la fonction de répartition G est définie comme étant l'unique racine de*

$$\mathbb{E}[K]\int_0^{+\infty} \mathrm{e}^{-\nu s}\mathrm{d}G(s) = 1\,.$$

Comme exemple trivial, si G est la fonction de répartition de la loi exponentielle de paramètre λ et $\mathbb{E}[K] = 2$, alors $\nu = \lambda$. Le principal résultat relié au paramètre malthusien est le théorème de Harris [33, Théo. 17.1, p.142] suivant :

Théorème 2.2.2. *Soit la constante C définie par*

$$C = \frac{\mathbb{E}[K] - 1}{\nu\mathbb{E}[K]^2\displaystyle\int_0^{+\infty} s\mathrm{e}^{-\nu s}\mathrm{d}G(s)}\,, \tag{2.12}$$

alors :

$$\lim_{t\to+\infty} \mathbb{E}[Y(t)]\mathrm{e}^{-\nu t} = C\,.$$

En d'autres termes, tant que les durées de vie des non-mutantes sont *i.i.d.*, la croissance de la population est asymptotiquement exponentielle selon un taux directement défini par la loi des durées de vie. Notons qu'il a été montré par Louhichi et Ycart [58, Théo. 3.1] que même dans le cas où la dépendance mère-fille est prise en compte, la croissance de la population est asymptotiquement exponentielle selon un taux de croissance déduit du modèle de croissance. Comme exemple général, si $\mathbb{E}[K] = 2$, la constante C est donnée par

$$C = \left(4\nu\int_0^{+\infty} s\mathrm{e}^{-\nu s}\mathrm{d}G(s)\right)^{-1}\,. \tag{2.13}$$

De plus, si nous considérons encore une fois le cas classique où G est la fonction de répartition de la loi exponentielle de paramètre λ et $\mathbb{E}[K] = 2$, alors il est facile de constater que $C = 1$.

Replaçons nous à nouveau dans le contexte des modèles de mutation. Nous considérons la formulation G/M suivante :

- les durées de vie des cellules normales sont *i.i.d.* selon une fonction de répartition quelconque G ;
- les durées de vie des cellules mutantes sont *i.i.d.* selon une loi exponentielle de taux μ ;

À partir de maintenant, le taux ν sera le paramètre malthusien associé à la fonction de répartition G. De plus l'échelle de temps est telle que le taux μ vaut 1 et $\nu = \rho$, c'est-à-dire :

$$2\int_0^{+\infty} \mathrm{e}^{-\rho s}\mathrm{d}G(s) = 1\,.$$

La généralisation du théorème 2.2.1 est donnée par Hamon et Ycart [31, Théo. 1.1] :

Théorème 2.2.3. *Soient $\pi = (\pi_n)_{n\in\mathbb{N}}$ et $\tau = (\tau_n)_{n\in\mathbb{N}}$ deux suites de réels positifs, et une constante $m > 0$ telles que :*

$$\lim_{n\to+\infty} \pi_n = 0\,, \qquad \lim_{n\to+\infty} \tau_n = +\infty\,, \qquad \lim_{n\to+\infty} \pi_n n C \mathrm{e}^{\rho\tau_n} = m\,,$$

où la constante C est définie par (2.13). Alors le résultat du théorème 2.2.1 est toujours valide.

Ainsi, la loi asymptotique du nombre final de mutantes ne dépend de la loi G qu'à travers son paramètre malthusien ρ. Les auteurs donnent une démonstration non-analytique qui se basent sur la description d'un modèle de mutation comme la composition des trois ingrédients suivants :

(A_1) le nombre de mutations converge en loi vers la loi de Poisson de paramètre m ;

(A_2) la loi jointe des durées de développement d'un nombre donné k de clones mutants converge vers la loi du produit de k variables exponentiellement *i.i.d.* de paramètre ρ ;

(A_3) la taille à un instant t d'un clone mutant suit la loi géométrique de paramètre e^{-t}.

Si l'affirmation (A_3) est un résultat que nous avons déjà mentionné plus tôt dans la preuve de la proposition 2.2.1, les deux autres ingrédients demandent une manipulation des différentes asymptotiques qui entrent en jeu dans le modèle. Leurs démonstrations peuvent être trouvées dans [31, p. 1255–1256]. Notons que cette décomposition est valable pour les prochaines formulations : seule l'affirmation (A_3), qui dépend directement du processus de croissance des mutantes, change d'une formulation à l'autre. Notons également que que la loi de Yule [100] associée à (2.7) n'a pas forcément d'espérance : elle n'est définie que lorsque ρ est strictement supérieur à 1.

Par la suite, d'autres lois de durées de vie furent ajustées à partir de données réelles : par exemple Kendall [43] ajusta les paramètres d'une loi Gamma aux données de Kelly et Rahn [40] (voir également [7, sec. 10.3]), et Kubitschek [50] ajusta les paramètres d'une loi log-Normale à ses propres données obtenues quelques années auparavant. Le fait est qu'il est compliqué de décider quelle loi est la plus adaptée pour modéliser des durées de vie cellulaire, car de nombreuses familles de lois peuvent facilement s'ajuster à des jeux de données [45, 97]. Une généralisation du théorème ci-dessus au cas G/G fut proposée par Ycart [97]. Des lois quelconques pour les durées de vie des cellules mutantes sont donc considérées, en introduisant le paramètre malthusien associé. La Figure 2.1 représente deux clones qui croissent selon des durées de vie cellulaires différentes. La formulation ainsi considérée est la suivante :

— les durées de vie des cellules normales sont *i.i.d.* selon une fonction de répartition quelconque G ;

— les durées de vie des cellules mutantes sont *i.i.d.* selon une fonction de répartition quelconque F ;

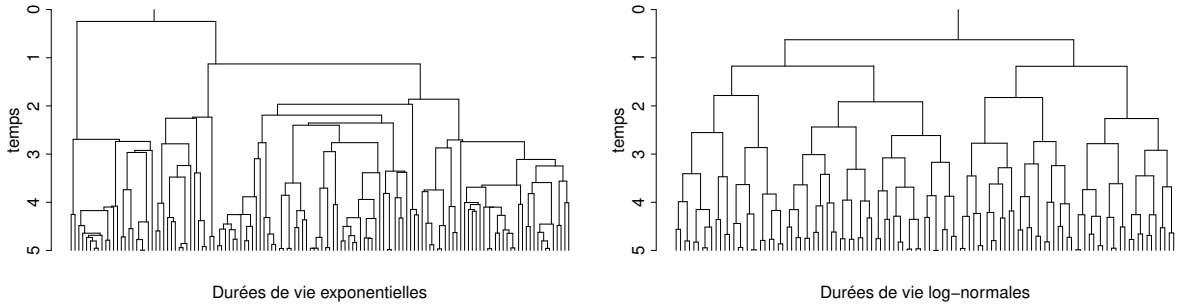

FIGURE 2.1 – **Croissance de deux clones suivant des modèles différents**. Le clone de gauche croît selon des durées de vie exponentielles. Le clone de droite croît selon des durées de vie de loi log-normale.

Le taux μ jouera donc le rôle de paramètre malthusien de la fonction de répartition F. Comme précédemment, l'échelle de temps est supposée telle que

$$2\int_0^{+\infty} \mathrm{e}^{-\rho s}\mathrm{d}G(s) = 2\int_0^{+\infty} \mathrm{e}^{-s}\mathrm{d}F(s) = 1\,.$$

On peut alors réécrire le théorème ainsi :

Théorème 2.2.4. *Soient $\pi = (\pi_n)_{n\in\mathbb{N}}$ et $\tau = (\tau_n)_{n\in\mathbb{N}}$ deux suites de réels positifs, et une constante $m > 0$ telles que :*

$$\lim_{n\to+\infty} \pi_n = 0\,, \quad \lim_{n\to+\infty} \tau_n = +\infty\,, \quad \lim_{n\to+\infty} \pi_n n C\mathrm{e}^{\rho\tau_n} = m\,,$$

où la constante C est définie par (2.13). Lorsque n tend vers l'infini, la fonction génératrice du nombre à un instant τ_n de mutantes dans les clones issus de n cellules normales converge vers la fonction génératrice (2.6) avec :

$$h(z) = \int_0^{+\infty} \widetilde{\psi}(z,s)\rho\mathrm{e}^{-\rho s}\mathrm{d}s\,. \tag{2.14}$$

L'affirmation (A_3) n'est alors explicite que lorsque l'expression de $\widetilde{\psi}(z,s)$ l'est également. Jusqu'à présent une forme explicite de $\widetilde{\psi}(z,s)$ n'est disponible que dans deux cas : le cas exponentiel considéré jusqu'à présent, et le cas où les durées de vie des mutantes sont constantes, c'est-à-dire la formulation d'Haldane mentionnée dans la sous-partie précédente. La fonction génératrice $\widetilde{\psi}$ a été explicitée par B. Ycart pour ce dernier cas, et la fonction génératrice h est alors donnée par [97, eq. (5)] :

$$h(z) = (1-2^{-\rho})\sum_{k\geqslant 0} 2^{-k\rho} z^{2^n}\,. \tag{2.15}$$

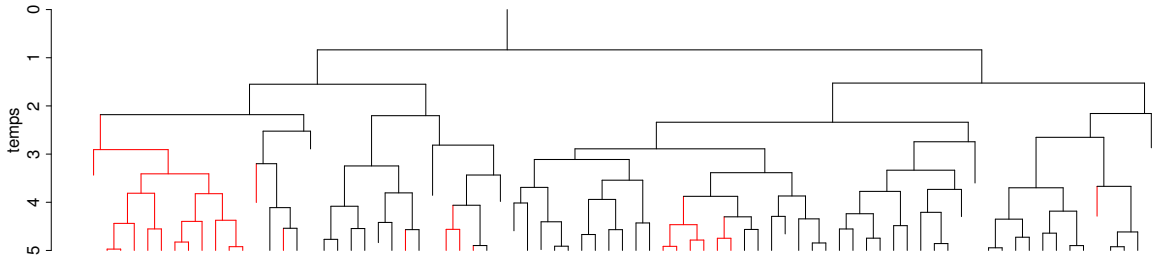

FIGURE 2.2 – **Croissance d'un clone avec des morts cellulaires**. Les durées de vie de toutes les cellules sont log-normales. Les branches rouges correspondent aux cellules mutantes.

Précisons qu'il y est supposé que la durée de vie d'une cellule mutante vaut $\log(2)$ afin de conserver un paramètre malthusien égal à 1. La simulation rapide n'est également possible que dans ces deux cas : simuler un modèle de mutation avec des durées de vie quelconques requiert de simuler le processus itérativement.

La dernière formulation que nous allons présenter prend en compte les morts des cellules. La Figure 2.2 représente sous forme d'arbre la croissance d'un clone issu d'une cellule normale au cours de laquelle certaines cellules meurent. Nous y observons en particulier des mutations qui donnent un clone de taille nulle (par exemple, la 2^{e} mutation en partant de la gauche).

Différents auteurs se sont penchés sur la question. Une modélisation M/M ainsi qu'un algorithme calculant la loi du nombre final de mutantes avaient déjà été proposés par TAN [87]. L'influence du paramètre de mort sur le modèle a également été étudiée par ANGERER [3, part. 3], DEWANJI et al. [18] ou encore KOMAROVA et al. [48]. Nous allons nous intéresser ici à la fomulation G/M proposée par YCART [98], qui se résume aux hypothèses suivantes :

- les durées de vie des cellules normales sont *i.i.d.* selon une fonction de répartition quelconque G ;
- les durées de vie des cellules mutantes sont *i.i.d.* selon une loi exponentielle de taux μ ;
- lorsqu'une cellule normale termine son développement, elle est remplacée par :
  - aucune cellule avec probabilité γ (mort) ;
  - une cellule normale et une cellule mutante avec probabilité π ;
  - deux cellules normales identiques avec probabilité $1 - \gamma - \pi$;
- lorsqu'une cellule mutante termine son développement, elle est remplacée par :
  - aucune cellule avec probabilité δ (mort) ;
  - deux cellules mutantes identiques avec probabilité $1 - \delta$;

— les événements de divisions, de mutations et de morts sont indépendants entre eux.

La constante γ intervient uniquement dans la définition du paramètre malthusien ν de la fonction de répartition G :

$$2(1-\gamma)\int_0^{+\infty} \mathrm{e}^{-\nu s}\mathrm{d}G(s) = 1\,,$$

ainsi que dans le calcul de la constante C :

$$C = \left(4\nu\frac{(1-\gamma)^2}{1-2\gamma}\int_0^{+\infty} s\mathrm{e}^{-\nu s}\mathrm{d}G(s)\right)^{-1}. \tag{2.16}$$

À titre d'exemple, si les durées de vie des non-mutantes sont *i.i.d.* selon une loi exponentielle de paramètre λ, alors le paramètre malthusien associé est $\nu = (1-2\gamma)\lambda$, et $C = 1$.

Dans ce contexte la fitness ρ est définie par le ratio de ν sur $(1-2\delta)\mu$. Nous nous plaçons dans le cas où l'échelle de temps est telle que $\mu = \frac{1}{1-2\delta}$ et que $\nu = \rho$, c'est-à-dire :

$$2(1-\gamma)\int_0^{+\infty} \mathrm{e}^{-\rho s}\mathrm{d}G(s) = 1\,.$$

Théorème 2.2.5. *Soient $\pi = (\pi_n)_{n\in\mathbb{N}}$ et $\tau = (\tau_n)_{n\in\mathbb{N}}$ deux suites de réels positifs, et une constante $m > 0$ telles que :*

$$\lim_{n\to+\infty}\pi_n = 0\,,\quad \lim_{n\to+\infty}\tau_n = +\infty\,,\quad \lim_{n\to+\infty}\pi_n n\frac{1-2\gamma}{1-\gamma}C\mathrm{e}^{\rho\tau_n} = m\,,$$

où la constante C est définie par (2.16).

Lorsque n tend vers l'infini, la fonction génératrice du nombre de mutantes au temps τ_n dans les clones issus de n cellules normales nées au temps 0 converge vers la fonction génératrice (2.6) avec :

$$h(z) = \rho\int_0^1 \frac{\delta(1-z)+v((1-\delta)z-\delta)}{(1-\delta)(1-z)+v((1-\delta)z-\delta)}v^{\rho-1}\mathrm{d}v\,, \tag{2.17}$$

L'expression de (2.17) vient du fait que la taille à un instant t d'un clone mutant débuté à l'instant 0 est à présent caractérisée par la fonction génératrice suivante (voir [6, p. 109]) :

$$\widetilde{\psi}(z,t) = \frac{\delta(1-z)+\mathrm{e}^{-t}((1-\delta)z-\delta)}{(1-\delta)(1-z)+\mathrm{e}^{-t}((1-\delta)z-\delta)}\,.$$

La même généralisation peut d'ailleurs être effectuée pour la formulation G/G du théorème 2.2.6. Tout comme pour la formulation G/G sans morts, nous supposerons que l'échelle de temps est telle que

$$2(1-\gamma)\int_0^{+\infty}\mathrm{e}^{-\rho s}\mathrm{d}G(s) = 2(1-\delta)\int_0^{+\infty}\mathrm{e}^{-s}\mathrm{d}F(s) = 1\,.$$

Le théorème de convergence s'écrit alors comme suit :

Théorème 2.2.6. *Soient $\pi = (\pi_n)_{n\in\mathbb{N}}$ et $\tau = (\tau_n)_{n\in\mathbb{N}}$ deux suites de réels positifs, et une constante $m > 0$ telles que :*

$$\lim_{n\to+\infty} \pi_n = 0\,, \quad \lim_{n\to+\infty} \tau_n = +\infty\,, \quad \lim_{n\to+\infty} \pi_n n \frac{1-2\gamma}{1-\gamma} C\mathrm{e}^{\rho\tau_n} = m\,.$$

où la constante C est définie par (2.16).

Lorsque n tend vers l'infini, la fonction génératrice du nombre de mutantes au temps τ_n dans les clones issus de n cellules normales nées au temps 0 converge vers la fonction génératrice (2.6) où la fonction h s'écrit sous la même forme que (2.14) :

$$h(z) = \int_0^{+\infty} \widetilde{\psi}(z,s)\rho\mathrm{e}^{-\rho s}\mathrm{d}s\,. \tag{2.18}$$

À nouveau, l'affirmation (A_3) n'est explicite que pour les cas « extrêmes » des durées de vie mutantes exponentielles ou constantes. Nous reviendrons sur le cas constant avec $\delta > 0$ dans le chapitre 3.

Dans toutes les formulations décrites ci-dessus, la probabilité de mutation π est ensuite interprétée comme le rapport m par le nombre final de cellules N. Ces approches considèrent que ce nombre est constant : la grande majorité des données de tests de fluctuation disponibles ne contiennent pas d'échantillons doubles de type (M, N) (c'est-à-dire des couples (nombre final de mutantes – nombre final de cellules)). Quelques exceptions telles que les données de David [17] existent. Cependant, même en appliquant des contrôles expérimentaux précis, il est difficile en pratique d'obtenir des nombres finaux de cellules sans fluctuations [49]. Des modèles de mutations avec fluctuations des nombres finaux de cellules ont été proposés par Angerer [3] et Komarova et al. [48]. Les conséquences statistiques que nous décrirons dans la partie 2.3 ont été étudiées par Ycart et Veziris [99]. Le nombre N est à présent considéré comme une variable aléatoire, avec une certaine fonction de répartition K définie sur $[0\,;\infty[$. En reprenant l'expression (2.6), la fonction génératrice conditionnelle du nombre final de mutantes sachant que $N = k$ peut-être définie par

$$\psi\left(z \,|\, N = k\right) = \exp\left(-\pi k(1 - h(z))\right)\,.$$

En d'autres termes, la loi conditionnelle du nombre final de mutantes sachant que $N = k$ est la loi définie par la fonction génératrice du théorème 2.2.6. Supposons que K est connue. Sa transformée de Laplace, notée $\mathcal{L}$, est donnée par

$$\mathcal{L}(z) = \mathbb{E}\left[\mathrm{e}^{-zN}\right] = \int_0^{\infty} \mathrm{e}^{-zk}\mathrm{d}K(k)\,, \tag{2.19}$$

La fonction génératrice du nombre final de mutantes est alors donnée par

$$\psi(z) = \int_0^{\infty} \psi\left(z \,|\, N = k\right)\mathrm{d}K(k) = \mathcal{L}\left(\pi(1 - h(z))\right)\,. \tag{2.20}$$

$MM(m, \rho, \delta, F)$	Modèles de mutations où les durées de vie mutantes sont *i.i.d.* de fonction de répartition F (théorème 2.2.6)
$LD(m, \rho, \delta)$	Modèles MM où les durées de vie des mutantes sont *i.i.d.* et exponentielles (modèles de Luria-Delbrück)
$H(m, \rho, \delta)$	Modèles MM où les durées de vie des mutantes sont constantes (modèles de Haldane)

TABLE 2.2 – **Notations des différents modèles de mutations**. Les quantités m, ρ et δ dénotent respectivement le nombre moyen de mutations, le paramètre de fitness, et la probabilité de mort d'une cellule mutante.

Dans le cas où N est constant, nous retrouvons l'expression (2.6) avec $m = \pi N$.

En pratique, il arrive que les cellules sont trop nombreuses et difficilement dénombrables. Dans ce cas, une proportion ζ de cellules est extraite de la population totale. Chaque cellule initialement présente a donc une probabilité ζ d'être présente dans la population « diluée ». Formellement, le nombre final $M^{(\zeta)}$ de mutantes dans une population obtenue après une dilution avec une plating efficiency ζ suit une loi binomiale de paramètres M et ζ. Sa fonction génératrice est donc donnée par

$$\begin{aligned}\psi^{(\zeta)}(z) = \mathbb{E}\left[z^{M^{(\zeta)}}\right] &= \mathbb{E}\left[\mathbb{E}\left[z^{M^{(\zeta)}} \mid M\right]\right] \\ &= \mathbb{E}\left[(1-\zeta+\zeta z)^M\right] \\ &= \psi(1-\zeta+\zeta z)\,.\end{aligned}$$

Le calcul des probabilités de $M^{(\zeta)}$ a été explicité par différents auteurs [84, 28, 104], mais uniquement pour le cas où $\rho = 1$, $\delta = 0$ et lorsque les durées de vie sont exponentiellement distribuées. La prise en compte du paramètre ζ sera étendue dans la section 3.3.3. Dans la suite de ce chapitre, nous utiliserons les notations de la table 2.2 pour désigner les différents modèles de mutations considérés. Chacun de ces modèles peut être étendu au cas où le nombre final de cellules suit une loi de fonction de répartition K. Nous désignerons ces modèles selon le schéma suivant :

— les modèles de mutation MM avec un nombre final de cellules aléatoire seront notés $MMF(\pi, \rho, \delta, F, K)$: la réalisation de cette loi est un couple (M, N) tel que, conditionnellement à N, M suit la loi $MM(m, \rho, \delta, F)$;

— les modèles de mutation LD avec un nombre final de cellules aléatoire seront notés $LDF(\pi, \rho, \delta, K)$: la réalisation de cette loi est un couple (M, N) tel que, conditionnellement à N, M suit la loi $LD(\pi N, \rho, \delta)$;

— les modèles de mutation H avec un nombre final de cellules aléatoire seront notés $HF(\pi, \rho, \delta, K)$: la réalisation de cette loi est un couple (M, N) tel que, conditionnellement à N, M suit la loi $H(\pi N, \rho, \delta)$;

Le modèle proposé par Luria et Delbrück a ainsi beaucoup évolué depuis sa création, prenant en compte de plus en plus d'hypothèses. Des estimateurs de la probabilité de mutation π et de la fitness ρ de plus en plus robustes ont ainsi pu être déduits de ces différentes extensions.

2.3 Estimation paramétrique

Le principal objectif des tests de fluctuation est d'estimer à partir d'un jeu de données de décomptes finaux de mutants la probabilité individuelle de mutation π. L'approche classique consiste à estimer d'abord le nombre moyen de mutations m puis de diviser l'estimation obtenue par le nombre total de cellules. Comme nous l'avons mentionné dans l'introduction, un estimateur $\hat{\theta}_n$ d'une valeur théorique θ construit sur un échantillon de taille n ne devrait être considéré en pratique que s'il est consistant :

$$\lim_{n\to+\infty} \hat{\theta}_n = \theta\,,$$

où la limite considérée peut être en probabilité (consistance *faible*) ou presque sûre (consistance *forte*). De plus, ses fluctuations doivent pouvoir être quantifiées ce qui revient à étudier la convergence en loi de $\sqrt{n}\left(\hat{\theta}_n - \theta\right)$. Dans le cas des estimateurs auxquels nous allons nous intéresser, cette variable tend en loi vers une loi Normale centrée.

Luria et Delbrück ont initialement proposé deux estimateurs [Eqs. (5) et (8)][60]. Le premier est construit à partir de l'estimateur de la probabilité de n'observer aucune mutation, et hérite ainsi des ses propriétés de consistance et de normalité asymptotique. La deuxième méthode d'estimation s'appuie sur une relation entre le nombre moyen de mutations et le nombre moyen de mutantes. Or, comme nous l'avons mentionné dans la partie 2.2, l'espérance du nombre de mutantes n'existe pas lorsque la fitness vaut 1. L'estimateur ainsi construit n'est donc pas consistant, et ne devrait pas être utilisé. Lea et Coulson ont eux aussi proposé une méthode en relation avec leur modélisation, appelé méthode de la médiane, basée sur une approximation de la loi du nombre de mutantes [56, eq. (25)]. Cependant, les propriétés de consistance et de normalité asymptotique de la médiane empirique ne sont vraies que dans le cas d'une loi continue. Ce n'est plus le cas pour une loi discrète à queue lourde. De plus, cette méthode ne donne des résultats pertinents que pour des valeurs théoriques de m comprises entre 4 et 15, ignorant ainsi une grande partie des valeurs typiques de m, plus faibles que 4. Par la suite, de nombreuses méthodes basées sur la médiane ou les quantiles ont également été proposées. Ces méthodes ont en grande partie été présentées par Foster [25]. La prise en compte de la fitness dans un modèle stochastique n'ayant pas été proposée avant Armitage [4], les premières méthodes d'estimation, comme celles de Luria-Delbrück ou de Lea-Coulson, ne proposent pas d'estimateurs pour ρ.

Nous nous intéressons dans cette thèse à deux autres méthodes d'estimation. La probabilité d'avoir k mutantes dans une population ainsi que ses dérivées par rapport à m

et ρ peuvent être calculées explicitement ou via des algorithmes dans le cas des formulations LD et H (pour $\delta = 0$) (voir entre autres [103, 31, 97]). De ce fait, la méthode du Maximum de Vraisemblance peut être utilisée pour l'estimation de m et ρ. Elle peut également être appliquée avec la méthode P0 afin d'estimer la fitness après avoir estimé m. Cependant, la méthode du Maximum de Vraisemblance a des limites pratiques à prendre en compte. Le calcul des probabilités fait en effet intervenir des opérations lourdes et est potentiellement instable numériquement [31]. Une méthode alternative basée sur l'utilisation de la fonction génératrice (2.6) a été proposée par Hamon et Ycart [31, part. 4]. Les estimations de m et ρ obtenues sont comparables en terme de précision à celles obtenues par le Maximum de Vraisemblance, et leur calcul est en pratique plus stable. Les estimateurs obtenus par ces trois méthodes sont asymptotiquement sans biais et normaux. Il est alors possible de construire des intervalles de confiance et de calculer des p-valeurs, afin d'effectuer des tests statistiques sur un ou deux échantillons.

Afin de mettre en évidence le fait que les méthodes basées sur la médiane ou les quantiles ne devraient pas être employées, nous avons effectué des études de simulations. La Figure 2.3 expose les résultats obtenus pour les sept estimateurs suivants :

- P0 : l'estimateur P0 de Luria et Delbrück décrit dans la sous-partie 2.3.1 ;
- GF : l'estimateur basé sur la fonction génératrice ψ décrit dans la sous-partie 2.3.3 ;
- ML : l'estimateur du Maximum de Vraisemblance décrit dans la sous-partie 2.3.2 ;
- LC : l'estimateur par la médiane de Lea et Coulson [56, eq. (25)] ;
- JM : l'estimateur par la médiane de Jones et al. [39, eq. (6)] ;
- KQ : l'estimateur par les quartiles de Koch [47, eqs. (3)-(5)] ;
- AC : l'estimateur par accumulation des clones de Luria [59] (voir également ([25, eqs. (14)-(15)]).

Nous avons simulé 10^4 échantillons de taille 100 selon la loi $LD(m, 1, 0)$, où le paramètre m prend ses valeurs dans $(0.5, 1, 2, 4)$. Nous avons ensuite estimé le paramètre m via les méthodes décrites ci-dessus, et observé la distribution de $\hat{m}/m$ via des boxplots obtenus grâce au logiciel R [77, 71]. Pour chaque figure, les lignes rouges correspondent à la valeur théorique. Les lignes bleues représentent les biais relatifs de plus ou moins 10%. Nous pouvons tirer de ces résultats visuels plusieurs observations. Tout d'abord, les méthodes GF et ML donnent de bons résultats quelle que soit la valeur théorique de m : pour ces 2 méthodes, au moins la moitié des estimations a un biais relatif inférieur à 10%. La méthode P0 donne également de bons résultats, mais perd en robustesse lorsque m augmente. Notons par ailleurs que lorsque $m = 4$, la probabilité qu'il n'y ait aucune mutante est très faible. De fait, la méthode P0 n'est pas tout le temps applicable dans ce cas. Nous avons affecté 0 au rapport $\hat{m}/m$ lorsque c'est le cas (d'où le grand nombre de valeurs aberrantes dans le boxplot correspondant). Les méthodes basées sur la médiane ou les quartiles du nombre de mutantes donnent au contraire de bonnes estimations lorsque m est grand. Cependant, nous pouvons clairement constater que ces méthodes ne devraient en aucun

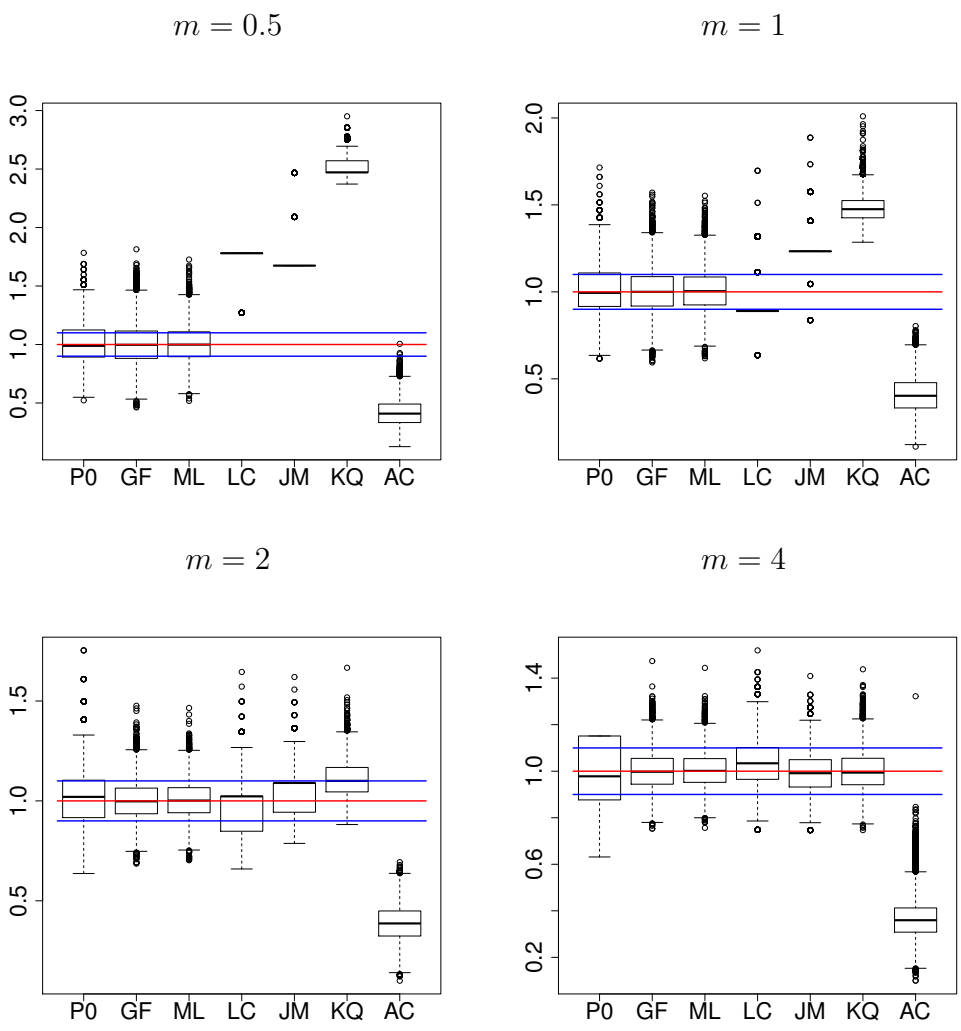

FIGURE 2.3 – **Comparaison de différentes méthodes d'estimation du paramètre m.** Pour chaque valeur du paramètre $m = (0.5, 1, 2, 4)$, 10^4 échantillons de taille 100 de la loi $LD(m, 1, 0)$ ont été simulés. Chaque boxplot représente la distribution des 10^4 rapports $\hat{m}/m$ pour chacune des méthodes considérées.

cas être utilisée de manière générale. En effet ces méthodes perdent toute pertinence pour des valeurs de m proches de l'unité : par exemple, les estimations obtenues par les méthodes LC et JM sont presque déterministes et ne sont plus centrées en m.

Rappelons que le vrai paramètre d'intérêt est la probabilité de mutation π. Dans le cas des méthodes P0, GF et ML, l'estimation de la probabilité de mutation est calculée en divisant celle du nombre de mutations par le nombre moyen de cellules. Nous avons mentionné plus tôt la méthode des moyennes de Luria et Delbrück [60, eq. (8)] (que nous noterons LDM) où π est estimée en calculant la racine d'une équation reliant le nombre moyen de mutantes et le nombre final de cellules. Nous avons effectué la même expérience que ci-dessus, afin d'observer les biais obtenus avec les méthodes P0, GF, ML et LDM. La méthode LDM nécessitant la recherche de la racine d'une fonction croissante, il faut donc spécifier un intervalle de recherche de taille finie. La méthode GF donnant des estimations précises, nous avons fixé l'intervalle de recherche pour l'estimation par LDM à $[0.01 \times \hat{\pi}_{GF} ; 100 \times \hat{\pi}_{GF}]$. La table 2.3 contient le résumé de ces résultats, obtenus via la commande `summary`. Il est en effet impossible de tracer un boxplot lisible : comme nous pouvons le constater, la variance de la méthode LDM est très grande en comparaison de celle des trois autres méthodes. Nous pouvons également remarquer que les estimations par la méthode LDM ne sont pas centrées en la valeur théorique. Même la médiane des estimations est relativement éloignée de la vraie valeur de π. Ces observations vont dans le sens de Rosche et Foster [79] : l'utilisation de cette méthode n'est pas recommandée, bien qu'elle apparaisse dans certaines études récentes (par exemple [94]). En conséquence, les trois seules méthodes que nous considérerons dans cette thèse et que nous qualifierons d'intérêt sont les méthodes P0, GF et ML.

Comme nous avons pu le constater dans la partie 2.2, le modèle initial de Luria et Delbrück a été soumis à de nombreuses transformations afin de se rapprocher de plus en plus de la réalité. Il est alors nécessaire d'ajuster les méthodes d'estimation existantes. Par exemple, certains modèles de mutations font intervenir deux autres paramètres : la fitness ρ et la probabilité de mort δ. La majorité des estimateurs de m dépendent donc plus ou moins fortement de ces deux autres paramètres. Leur valeur étant en pratique inconnues, il est nécessaire de construire des estimateurs pour ρ et δ. Comme nous allons le voir, l'estimation de ρ ne pose pas de réel problème, mais celle de δ est plus délicate. Nous nous concentrerons donc sur l'estimation des paramètres m, π et ρ.

Comme nous l'avons mentionné plus haut, l'ajustement des méthodes d'estimation est nécessaire afin de supprimer les biais d'estimation identifiés grâce à l'ajustement des modèles de mutation. Nous nous intéressons dans cette thèse aux sources de biais suivantes :

1. certaines cellules meurent sans se diviser dans le processus, ce qui n'est pas pris en compte dans le modèle d'estimation. Les conséquences statistiques ont été traitées par Ycart [98] pour les trois méthodes appliquées aux modèles LD. Cette étude est étendue aux modèles H dans cette thèse ;
2. le nombre de cellules obtenu en fin d'expérience est aléatoire, mais est considéré constant lors de l'estimation. Cette source de biais a été étudiée par Ycart et

$m = 0.5$

	P0	GF	ML	LDM
Minimum	0.4969	0.4525	0.4839	0.4004
Premier quartile	0.8926	0.8851	0.8973	0.9648
Médiane	0.9886	0.9974	1.0019	1.2805
Moyenne	1.0077	1.0049	1.0062	2.1454
Dernier quartile	1.1242	1.1169	1.1093	1.8589
Maximum	1.8326	1.8643	1.7287	182.3220

$m = 1$

	P0	GF	ML	LDM
Minimum	0.5978	0.5255	0.5838	0.4452
Premier quartile	0.9163	0.9123	0.9185	0.9647
Médiane	0.9943	0.9958	0.9971	1.2422
Moyenne	1.0078	1.0011	1.0030	2.0966
Dernier quartile	1.0788	1.0844	1.0812	1.7669
Maximum	1.5606	1.5312	1.5104	127.472

$m = 2$

	P0	GF	ML	LDM
Minimum	0.6189	0.6744	0.6938	0.5226
Premier quartile	0.9163	0.9371	0.9041	0.9646
Médiane	1.0201	0.9976	1.0013	1.12159
Moyenne	1.0176	1.0012	1.0048	2.1193
Dernier quartile	1.1036	1.0605	1.0659	1.6787
Maximum	1.9560	1.3820	1.3700	2059.1089

$m = 4$

	P0	GF	ML	LDM
Minimum	0.6314	0.7459	0.7639	0.5879
Premier quartile	0.8766	0.9433	0.9512	0.9670
Médiane	0.9780	0.9965	1.0004	1.1835
Moyenne	$+\infty$	1.0004	1.0033	1.7693
Dernier quartile	1.1513	1.0545	1.0526	1.6041
Maximum	$+\infty$	1.3727	1.3050	110.0355

TABLE 2.3 – **Comparaison de différentes méthodes d'estimation du paramètre π.** Pour chaque valeur du paramètre $m = (0.5, 1, 2, 4)$, 10^4 échantillons de taille 100 de la loi $LDF(\pi, 1, 0, K)$ ont été simulés, où $\pi = m/\kappa$ (avec $\kappa = 10^9$) et K est la fonction de répartition de la mesure de Dirac localisée en κ. Chaque colonne donne les principaux indicateurs statistiques des 10^4 rapports $\hat{\pi}/\pi$ pour chacune des méthodes considérées.

Veziris [99] pour la méthode P0 appliquée aux modèles MM et MMF (sans morts), ainsi que pour la méthode ML appliquée aux modèles LDF et HF. Nous traiterons également dans cette thèse le cas $\delta > 0$ pour les trois méthodes appliquées aux modèles LD, H, LDF et HF ;

3. la prise en compte de la dilution : que ce soit pour des raisons techniques ou expérimentales, les données sont généralement obtenues à partir d'un extrait de la culture totale. Un paramètre de dilution égal à ζ signifie que seule une proportion ζ de la culture totale est observée. Ainsi, chaque mutante présente dans la culture initiale sera observée dans l'extrait avec probabilité ζ. Différentes approches ont été exposées dans le cas des modèles $LD(m, 1, 0)$ [28, 104]. La correction de l'estimation du nombre moyen de mutations la plus utilisée est celle proposée par Stewart et al. [86, eq. (41)]. Nous constaterons dans la section 4.1.1 que cette correction n'est également applicable que pour les modèles $LD(m, 1, 0)$. Nous traiterons dans cette thèse le cas plus général où $\rho \neq 1$ et $\delta > 0$, pour les trois méthodes.
4. la durée de vie des cellules n'est en réalité pas exponentielle, mais l'est dans le modèle d'estimation. Les biais d'estimation induits ont été étudiés par Ycart [97] lorsque $\delta = 0$;
5. le processus réel de croissance n'est pas homogène, mais l'est dans le modèle. L'étude de cette source de biais est l'objectif principal de cette thèse.

Nous exposons à présent les trois méthodes d'estimation d'intérêt que sont les méthodes P0, ML (du Maximum de Vraisemblance) et GF (basée sur la fonction génératrice) dans les sous-parties qui suivent. Il y a deux types d'échantillons possibles :

$T1$ n réalisations indépendantes $(M_i)_{i=1,\dots,n}$ d'une variable M selon un même modèle de mutation de type MM, avec en plus la moyenne et l'écart-type du nombre final de cellules ;

$T2$ n réalisations indépendantes $\big((M_i, N_i)\big)_{i=1,\dots,n}$ d'un couple de variables (M, N) selon un même modèle de mutation de type MMF, où aucune hypothèse particulière n'est faite sur la loi des nombres finaux de cellules.

La méthode P0 (sous-partie 2.3.1) ne nécessite aucune hypothèse sur la durée de vie des cellules mutantes. Il est par contre nécessaire pour les méthodes ML (sous-partie 2.3.2) et GF (sous-partie 2.3.3) de considérer les cas où la loi du nombre de mutantes est explicite, c'est-à-dire les modèles LD et H pour les échantillons de type $T1$, et les modèles LDF et HF pour les échantillons de type $T2$. Enfin, nous noterons $(p_k)_{k\in\mathbb{N}}$ les probabilités du nombre de mutantes :

$$p_k = \mathbb{P}\left[M = k\right] .$$

2.3.1 Méthode P0

La version initiale de la méthode P0 de Luria-Delbrück s'appuie sur le fait que la probabilité de n'avoir aucune mutante est donnée par $p_0 = \mathrm{e}^{-m}$ (lorsque $\delta = 0$). Ainsi m

peut être estimé par

$$\hat{m}_0 = -\log\left(\hat{p}_0\right) , \tag{2.21}$$

où $\hat{p}_0$ est la proportion de cultures ne contenant pas de mutante :

$$\hat{p}_0 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{M_i=0} .$$

Par définition, $\hat{p}_0$ est un estimateur asymptotiquement sans biais et normal de P0. Par la Δ-méthode [93, p. 79], $\hat{m}_0$ est également asymptotiquement sans biais et normal, et de variance asymptotique :

$$v_{\hat{m}_0} = \frac{1-\hat{p}_0}{n\hat{p}_0} . \tag{2.22}$$

Bien entendu, il est impossible d'utiliser cette méthode lorsque l'échantillon ne contient pas de zéro.

Une extension de la méthode P0 au cas où $\delta > 0$ est décrite par Ycart [98], sous le nom d'estimateur FP (du Point Fixe). En supposant que $\delta < 1/2$, un point fixe de la fonction génératrice $\widetilde{\psi}(\cdot, t)$ est la probabilité d'extinction d'un clone mutant (voir [6, Théo. 1., chap. I]) donnée par

$$\delta_* = \frac{\delta}{1-\delta} .$$

Ainsi, δ_* est également un point fixe de la fonction génératrice (2.18), et $\psi(\delta_*) = \mathrm{e}^{-m(1-\delta_*)}$. Un estimateur asymptotiquement sans biais et normal de m est alors donné par

$$\hat{m}_0 = \frac{-\log(\hat{\psi}_n(\delta_*))}{1-\delta_*} , \tag{2.23}$$

où $\hat{\psi}_n$ est la fonction génératrice empirique du nombre de mutantes :

$$\hat{\psi}_n(z) = \frac{1}{n}\sum_{i=1}^{n} z^{M_i} . \tag{2.24}$$

Pour tout $z \in]-1\,;1[$, $\hat{\psi}_n(z)$ est un estimateur asymptotiquement sans biais et normal de $\widetilde{\psi}(z)$. Par la Δ-méthode, $\hat{m}_0$ est toujours asymptotiquement sans biais et normal, et de variance asymptotique :

$$v_{\hat{m}_0} = \frac{1}{n(1-\delta_*)^2}\left(\frac{\psi(\delta_*^2)}{\psi(\delta_*)^2} - 1\right) . \tag{2.25}$$

L'estimateur proposé par Luria et Delbrück peut alors être retrouvé dans le cas $\delta = 0$.

L'estimation de la probabilité de mutation π est ensuite obtenue en divisant l'estimation de m par la moyenne du nombre final de cellules (fournie pour les échantillons de

type $T1$, calculée empiriquement pour les échantillons de type $T2$). Les fluctuations du nombre final de cellules N n'apparaissent donc pas dans le calcul, et il a été montré par Ycart et Veziris [99] que l'estimation de π est biaisée. Selon le type d'échantillon, il est possible de réduire le biais induit (échantillons de type $T1$) ou de construire un estimateur asymptotiquement sans biais de π (échantillons de type $T2$). Considérons l'estimateur de $\hat{m}_0$ défini par (2.21). Si N est aléatoire, $\hat{m}_0$ est en fait un estimateur asymptotiquement sans biais de :

$$-\log\Big(\mathcal{L}[\pi]\Big)\,,$$

où $\mathcal{L}$ est la transformée de Laplace (2.19). Par l'inégalité de Jensen :

$$-\log\Big(\mathcal{L}[\pi]\Big) < \pi\mathbb{E}[N]\,.$$

Ainsi $\hat{m}_0$ sous-estime m, et $\hat{\pi}_0$ sous-estime également π. Si la loi de N est connue, et en supposant que la fonction réciproque $\mathcal{L}^{-1}$ de $\mathcal{L}$ existe, un estimateur asymptotiquement sans biais et normal de π peut être obtenu en appliquant $\mathcal{L}^{-1}$ à $\hat{p}_0$. Dans le cas des échantillons de type $T1$, seules des estimations de la moyenne et la variance de N sont disponibles. À défaut de supprimer le biais d'estimation de π, il est possible de le réduire en manipulant le développement suivant :

$$\mathcal{L}[\pi] = 1 - \mathbb{E}[N]\pi + \frac{\mathbb{E}\left[N^2\right]}{2}\pi^2 + \ldots \tag{2.26}$$

Une approximation du biais sur π peut alors être identifiée, et un estimateur asymptotiquement « sans biais » de π est alors obtenu :

$$\hat{\pi}_0 = \frac{\hat{m}_0}{\mathbb{E}[N]}\left(1 + \frac{C^2}{2}\hat{m}_0\right)\,, \tag{2.27}$$

où C est le *coefficient de variation* de N, c'est-à-dire :

$$C = \frac{\sqrt{\mathbb{V}\mathrm{ar}[N]}}{\mathbb{E}[N]}\,. \tag{2.28}$$

La variance asymptotique de $\hat{\pi}_0$ est donnée par la Δ-méthode :

$$v_{\hat{\pi}_0} = \Big(1 + \hat{m}_0 C^2\Big)^2 \frac{v_{\hat{m}_0}}{\mathbb{E}[N]^2}\,. \tag{2.29}$$

Nous reviendrons sur cette méthode de réduction du biais dans le chapitre 4 afin de prendre en compte le cas $\delta > 0$.

Considérons à présent les échantillons de type $T2$. À partir des observations (M_i, N_i), définissons les couples (X_i, N_i), où $X_i = \mathbb{1}_{M_i=0}$. Ces nouveaux couples sont ainsi des

réalisations indépendantes du couple de variables (X, N) telle que la loi conditionnelle de X sachant N est définie par

$$\mathbb{P}\left[X = 1 | N = k\right] = \mathrm{e}^{-\pi k}\,.$$

La log-vraisemblance de l'échantillon $\big((X_i, N_i)\big)_{i=1,\dots,n}$ est alors définie par

$$\ell(\pi) = \sum_{i=1}^{n} -\pi N_i X_i + (1 - X_i)\log\left(1 - \mathrm{e}^{-\pi N_i}\right)\,. \tag{2.30}$$

S'il existe, le maximum $\hat{\pi}_{ML0}$ de ℓ est un estimateur asymptotiquement sans biais de π et de variance asymptotique [57, Corollaire 3.11., chapitre 6] :

$$v_{\hat{\pi}_{ML0}} = \left(\sum_{i=1}^{n}\left(-N_i X_i + \frac{(1 - X_i)N_i}{\mathrm{e}^{\pi N_i - 1}}\right)^2\right)^{-1}\,. \tag{2.31}$$

Rappelons que cet estimateur ne dépend pas de la loi de N.

Les estimateurs de la méthode P0 ne dépendent pas des hypothèses de modélisation des cellules mutantes. De fait, si une estimation de la fitness ρ est souhaitée, elle doit être effectuée ultérieurement par Maximum de Vraisemblance en fixant $m = \hat{m}_0$.

2.3.2 Méthode du Maximum de Vraisemblance (ML)

Des algorithmes permettant de calculer les probabilités des modèles LD et $H(m, \rho, 0)$ et leurs dérivées par rapport à m et ρ, (voir entre autres [103, 31, 97]), l'estimation par le Maximum de Vraisemblance semble être un choix évident. Notons $(q_k)_{k\in\mathbb{N}}$ les probabilités associées à la fonction génératrice h :

$$h(z) = \sum_{k \geq \geqslant 0} q_k z^k\,,$$

pour tout $z \in U$. À partir de (2.18) nous pouvons définir les q_k pour tout $k \geqslant 0$:

$$q_k = \int_0^{+\infty} \widetilde{p}_k(s)\rho \mathrm{e}^{-\rho s}\mathrm{d}s\,,$$

où la suite $(\widetilde{p}_k(s))_{k\in\mathbb{N}}$ correspond aux probabilités associées à la fonction génératrice $\widetilde{\psi}(\cdot, s)$. Comme mentionné précédemment, les probabilités q_k ne peuvent être explicitées que pour les formulations LD et H. Dans le premier cas, elles ont été données par Ycart [98, part. 3] :

$$q_0 = \int_0^1 \frac{\delta(1 - v)}{1 - \delta - \delta v}\rho v^{\rho - 1}\mathrm{d}v\,,$$

et pour tout $k > 0$:

$$q_k = (1-\delta)^{k-1}(1-2\delta)^2 \int_0^1 \frac{(1-v)^{k-1}}{(1-\delta-\delta v)^{k+1}} \rho v^\rho \mathrm{d}v \,.$$

Remarquons que si δ est nul, les q_k sont les probabilités de la loi de Yule de paramètre ρ, et pour tout $k > 0$:

$$q_k = \rho B(\rho+1, k) \,,$$

où B est la fonction Beta :

$$B(x,y) = \int_0^{+\infty} v^{x-1}(1-v)^{y-1}\mathrm{d}v \,.$$

À partir de (2.15), les probabilités q_k pour le modèle H lorsque $\delta = 0$ peuvent également être explicitées :

$$q_k = \begin{cases} \dfrac{1-2^{-\rho}}{k^\rho} & \text{si } \exists\, i \in \mathbb{N} \text{ t.q. } k = 2^i \,, \\ 0 \quad \text{sinon} \,. \end{cases}$$

L'expression des q_k dans le cas où $\delta > 0$ sera donnée dans le chapitre 3. Il est par la suite possible de calculer les dérivées des q_k par rapport à ρ. Leurs expressions peuvent être trouvées par exemple dans [31, 98].

À partir des expressions des q_k, il est ensuite possible d'exprimer la suite des probabilités $(p_k)_{k\in\mathbb{N}}$ via un algorithme récursif présenté par Embrechts et Hawkes [21] pour les mélanges poissonniens :

$$\begin{aligned} p_0(m,\rho) &= \mathrm{e}^{-m(1-q_0)} \,, \\ p_k(m,\rho) &= \frac{m}{k}\sum_{i=1}^{k} i q_i p_{k-i} \quad \forall\, k > 0 \,. \end{aligned} \tag{2.32}$$

La dépendance des p_k en m et ρ sera par la suite sous-entendue pour plus de lisibilité. La démonstration de cet algorithme s'effectue en remarquant que la dérivée de (2.6) par rapport à z s'écrit sous les deux formes suivantes :

$$\begin{aligned} \frac{\partial \psi(z)}{\partial z} &= m \frac{\partial h(z)}{\partial z} \psi(z) \\ &= m \left(\sum_{i\geqslant 1} i q_i z^{i-1} \right) \left(\sum_{j \geqslant 0} p_j z^j \right) \,, \\ &= m \sum_{i,j\geqslant 1} i q_i p_j z^{i+j-1} \,, \end{aligned}$$

et :

$$\frac{\partial \psi(z)}{\partial z} = \sum_{k\geqslant 1} k p_k z^{k-1} \,.$$

L'algorithme (2.32) est finalement obtenu par identification des coefficients des deux expressions. Des algorithmes similaires permettent de calculer les dérivées des p_k par rapport aux paramètres m et ρ :

$$\begin{array}{rclcrcl} \dfrac{\partial p_0}{\partial m} & = & -(1-q_0)p_0\,, & & \dfrac{\partial p_0}{\partial \rho} & = & m\dfrac{\partial p_0}{\partial \rho}q_0\,, \\ \dfrac{\partial p_k}{\partial m} & = & \displaystyle\sum_{i=0}^{k} q_i p_{k-i} - p_k\,, & & \dfrac{\partial p_k}{\partial \rho} & = & \displaystyle m\sum_{i=0}^{k}\frac{\partial q_i}{\partial \rho}p_{k-i} \quad \forall\, k>0\,. \end{array} \tag{2.33}$$

Pour retrouver ces expressions, il suffit d'effectuer le même raisonnement que pour (2.32). Par exemple, la dérivée de ψ par rapport à m s'écrit :

$$\begin{aligned} \frac{\partial \psi}{\partial m}(z) &= -\,(1-h(z))\psi(z) \\ &= -\left[\sum_{j\geqslant 0} p_j z^j - \left(\sum_{i\geqslant 0} q_i z^i\right)\left(\sum_{j\geqslant 0} p_j z^j\right)\right] \\ &= -\left[\sum_{j\geqslant 0} p_j z^j - \sum_{i,j\geqslant 0} q_i p_j z^{i+j}\right]\,, \end{aligned}$$

Mais également sous la forme suivante :

$$\frac{\partial \psi}{\partial m}(z) = \sum_{k\geqslant 0}\frac{\partial p_k}{\partial m} z^k\,.$$

Ainsi pour tout $k>0$:

$$\frac{\partial p_k}{\partial m} = -\left(p_k - \sum_{\substack{i,j\geqslant 0 \\ i+j=k}} q_i p_j\right) = \sum_{i=1}^{k} q_i p_{k-i} - p_k\,. \tag{2.34}$$

Pour retrouver ces expressions, il suffit d'effectuer le même raisonnement que pour (2.32), mais en dérivant cette fois (2.6) par rapport à m ou ρ selon les dérivées souhaitées. À partir des algorithmes (2.32), il est alors possible de calculer la log-vraisemblance d'un échantillon de type $T1$:

$$\begin{aligned} \ell(m,\rho) &= \sum_{i=1}^{n} \log(p_{M_i}) \\ &= \sum_{i=0}^{\max_j M_j}\left[\log(p_i)\sum_{k=1}^{n}\mathbb{1}_{X_k=j}\right]\,. \end{aligned} \tag{2.35}$$

Le couple d'estimateurs $(\hat{m}_{ML}, \hat{\rho}_{ML})$ obtenu en maximisant la log-vraisemblance ℓ est asymptotiquement sans biais et normal [57, Théo. 5.1., chapitre 6]. Les variances asymptotiques respectives de $\hat{m}_{ML}$ et $\hat{\rho}_{ML}$ sont données par

$$v_{\hat{m}_{ML}} = \frac{\mathcal{I}_{2,2}}{\det(I)} \quad \text{et} \quad v_{\hat{\rho}_{ML}} = \frac{\mathcal{I}_{1,1}}{\det(I)}\,,$$

où $\mathcal{I} = (\mathcal{I}_{i,j})_{i,j\in\{1,2\}}$ est la matrice d'information suivante :

$$\mathcal{I} = \sum_{j=0}^{\max M_i} \left[\begin{pmatrix} \left(\frac{\partial p_j}{\partial m}\frac{1}{p_j}\right)^2 & \frac{\partial p_j}{\partial m}\frac{\partial p_j}{\partial \rho}\frac{1}{p_j^2} \\ \frac{\partial p_j}{\partial m}\frac{\partial p_j}{\partial \rho}\frac{1}{p_j^2} & \left(\frac{\partial p_j}{\partial \rho}\frac{1}{p_j}\right)^2 \end{pmatrix} \sum_{i=1}^{n} \mathbb{1}_{M_i=j} \right] .$$

L'estimation de π est ensuite obtenue en divisant $\hat{m}_{ML}$ par le nombre moyen de cellules. La prise en compte des fluctuations du nombre final de cellules sera exposée dans le chapitre 4.

Dans le cas d'un échantillon de type $T2$, la log-vraisemblance s'exprime cette fois en fonction de π et ρ [99] :

$$\ell(\pi, \rho) = \sum_{i=1}^{n} \log\left(p_{M_i|N_i}(\pi, \rho)\right)\,, \tag{2.36}$$

où, pour tout $k \geqslant 0$, $p_{k|j}$ correspond à (2.32) avec $m = \pi j$:

$$p_{k|j}(\pi, \rho) = p_k(\pi j, \rho)\,.$$

La dépendance des $p_{k|j}$ en π et ρ sera sous-entendue par la suite. Les variances asymptotiques respectives de $\hat{\pi}_{ML}$ et $\hat{\rho}_{ML}$ sont données par

$$v_{\hat{\pi}_{ML}} = \frac{\mathcal{I}_{2,2}}{\det(I)} \quad \text{et} \quad v_{\hat{\rho}_{ML}} = \frac{\mathcal{I}_{1,1}}{\det(I)}\,, \tag{2.37}$$

où la matrice $\mathcal{I}$ est la matrice d'information suivante :

$$\mathcal{I} = \sum_{i=1}^{n} \left[\begin{pmatrix} \left(\frac{\partial p_{M_i}}{\partial m}\frac{N_i}{p_{M_i}}\right)^2 & \frac{\partial p_{M_i}}{\partial m}\frac{\partial p_{M_i}}{\partial \rho}\frac{N_i}{p_{M_i}^2} \\ \frac{\partial p_{M_i}}{\partial m}\frac{\partial p_{M_i}}{\partial \rho}\frac{N_i}{p_{M_i}^2} & \left(\frac{\partial p_{M_i}}{\partial \rho}\frac{1}{p_{M_i}}\right)^2 \end{pmatrix} \right] . \tag{2.38}$$

Bien que l'utilisation du Maximum de Vraisemblance ait été initalement recommandée par plusieurs auteurs pour le cas $\delta = 0$ (voir entre autres [61, 85, 103]), cette méthode

implique en pratique de lourds calculs et peut-être numériquement instable [31]. En l'occurrence, les convolutions nécessaires aux calculs de (2.35) et de ses dérivées requièrent de sommer des valeurs très faibles entre elles. L'instabilité numérique s'aggrave à mesure que le nombre de jackpots ainsi que leurs valeurs sont élevées, c'est-à-dire lorsque m est élevé et/ou ρ est faible. Le temps de calcul est par ailleurs directement impacté par la valeur maximale de l'échantillon, ainsi que par le choix du modèle : (2.36) est en pratique plus coûteuse que (2.35). Dans le cas d'une formulation LD, il est possible de calculer des équivalents des expressions q_k pour de grandes valeurs de k [98, eq. (3.4)], afin d'accélérer le calcul de (2.35). En pratique, il est également possible de réduire ces effets de queue via différentes méthodes [95, part. 2.2]. L'une d'elles, appelée *winsorisation*, consiste en remplacer n'importe quelle valeur de l'échantillon qui dépasse une certaine borne par cette même borne. Plus cette borne sera faible, plus il y aura de décomptes de mutants égaux, et moins les estimations obtenues par la méthode ML seront pertinentes. Cette méthode n'est donc pas adaptée aux modèles de mutations avec de gros jackpots.

2.3.3 Méthode basée sur la fonction génératrice (GF)

La dernière méthode présentée ici utilise la fonction génératrice (2.6) pour estimer m. Elle a été proposée par Hamon et Ycart [31, part. 4], mais l'estimation du paramètre d'une composée poissonnienne via sa fonction génératrice avait déjà été exposée auparavant [80, 63]. Les estimateurs de m et ρ dans [31] correspondent à la formulation LD sans mort. Cependant, cette méthode ne dépend que du fait d'avoir ou non des formes explicites de h et de ses dérivées par rapport à ρ, et est donc applicable aux autres formulations (voir entre autres [98] pour la formulation LD avec morts). Considérons un échantillon de type quelconque.

Soient z_1, z_2, z_3 dans $]0\,;1[$, avec $z_1 \neq z_2$. Les estimateurs par la méthode GF de m et ρ sont définis par

$$\hat{m}_{GF}(z_3) = \frac{\log\left(\hat{\psi}_n(z_3)\right)}{h_{\hat{\rho}_{GF}(z_1,z_2)}(z_3) - 1} \quad \text{et} \quad \hat{\rho}_{GF}(z_1, z_2) = g^{-1}(\hat{y}_n)\,,$$

où h_x est la fonction génératrice (2.18) avec $\rho = x$, $\hat{\psi}_n$ est la fonction empirique (2.24) et :

$$g(x) = \frac{h_x(z_1) - 1}{h_x(z_2) - 1} \quad \text{et} \quad \hat{y}_n = \frac{\log\left(\hat{\psi}_n(z_1)\right)}{\log\left(\hat{\psi}_n(z_2)\right)}\,.$$

Par le théorème (3.4) de Rémillard et Theodorescu [80] et la Δ-méthode, le couple d'estimateurs $(\hat{m}_{GF}, \hat{\rho}_{GF})$ est fortement consistant et asymptotiquement normal. De plus, sa matrice de covariance asymptotique peut être explicitée [31, Prop. 4.1.] :

Proposition 2.3.1. *Soient z_1, z_2, z_3 dans $]0\,;1[$, avec $z_1 \neq z_2$. Soit $C = \big(c(z_i, z_j)\big)_{i,j=1,2,3}$ la matrice de covariance asymptotique du vecteur aléatoire :*

$$\sqrt{n}\left(\left(\hat{\psi}_n(z_1), \hat{\psi}_n(z_2), \hat{\psi}_n(z_3)\right) - \left(\psi(z_1), \psi(z_2), \psi(z_3)\right)\right),$$

c'est-à-dire

$$c(z_i, z_j) = \psi(z_i z_j) - \psi(z_i)\psi(z_j)\,.$$

Soit la matrice $A = (a_{i,j})_{\substack{i=1,2,3\\ j=1,2}}$ suivante :

$$a_{1,1} = \frac{m a_{1,2}}{h(z_3)-1}\frac{\partial h(z_3)}{\partial \rho}\,; \quad a_{1,2} = \frac{h(z_2)-1}{m\psi(z_1)\left(\frac{\partial h(z_1)}{\partial \rho}(h(z_2)-1) - \frac{\partial h(z_2)}{\partial \rho}(h(z_1)-1)\right)}\,;$$
$$a_{2,1} = \frac{m a_{2,2}}{h(z_3)-1}\frac{\partial h(z_3)}{\partial \rho}\,; \quad a_{2,2} = \frac{h(z_1)-1}{m\psi(z_2)\left(\frac{\partial h(z_2)}{\partial \rho}(h(z_1)-1) - \frac{\partial h(z_1)}{\partial \rho}(h(z_2)-1)\right)}\,;$$
$$a_{3,1} = \frac{1}{\psi(z_3)(h(z_3)-1)}\,; \quad a_{3,2} = 0\,.$$

Le vecteur aléatoire :

$$\sqrt{n}\Big((\hat{m}_{GF}, \hat{\rho}_{GF}) - (m, \rho)\Big)$$

converge en loi vers la loi Normale bivariée centrée et de matrice de covariance $A^t C A$.

Les paramètres z_1, z_2 et z_3 sont arbitraires. En théorie, ils devraient être choisis de sorte que les variances asymptotiques de la proposition 2.3.1 soient minimales. Comme ces variances dépendent également des valeurs réelles de m et ρ qui sont inconnues, il n'est pas possible en pratique de choisir rapidement les valeurs optimales de z_1, z_2 et z_3. Cependant, les fluctuations des variances en fonction de ces trois paramètres sont relativement faibles. Leur valeur ont été fixées par Hamon et Ycart [31, p.1262] à partir de résultats numériques obtenus après simulations. La méthode GF est comparable à la méthode ML en terme de précision, mais est beaucoup plus stable numériquement. De plus, cette méthode ne nécessitant qu'une recherche d'un point fixe d'une fonction croissante, elle est très rapide. Elle peut donc être utilisée en pratique pour initialiser l'optimisation de (2.35) avec une valeur proche de l'optimum.

En pratique, il est possible que le paramètre ρ ne puisse être estimé. La recherche d'un point fixe implique de définir un intervalle de taille fini. Dans le cas où l'échantillon ne contient pas de jackpot, il est alors possible que cet intervalle de recherche ne contienne pas la valeur réelle de la fitness. Il n'est cependant pas judicieux de vouloir adapter un modèle de mutation à un échantillon sans jackpot.

Contrairement aux deux autres méthodes, il n'existe pas de cas spécial pour les échantillons de type $T2$. Pour les deux types d'échantillons, l'estimation de π est déduite en divisant $\hat{m}_{GF}$ par la moyenne du nombre final de cellules (fournie pour les échantillons de type $T1$, calculée empiriquement dans le cas d'un échantillons de type $T2$).

Les trois méthodes d'estimations P0, ML et GF seront généralisées dans le chapitre 4. Elles y seront également comparées à travers des études de simulation.

Il y a ainsi beaucoup de variations possibles dans un modèle de mutation. Pour chacune de ces variations, il est nécessaire de répondre aux quatre problèmes suivants :

1. le calcul de la loi du nombre de mutantes ;
2. l'estimation des paramètres d'intérêt, dont en particulier la probabilité de mutation ;
3. la construction d'algorithme de simulation rapide afin d'éprouver la robustesse des estimateurs construits ;
4. l'implémentation sous forme par exemple de scripts R des trois items précédents.

La table 2.4 donne un résumé de ce chapitre : pour chacune des hypothèses de modélisations mentionnées, les références précédemment citées sont listées selon les quatre problématiques exposées ci-dessus. Les chapitres ou parties de cette thèse traitant de ces points sont également précisées. Précisons que les références exposées dans la colonne « Simulation » proposent des algorithmes de simulation rapide, qui permettent d'obtenir directement un nombre final de mutantes. En effet, l'implémentation de l'évolution complète d'une population de cellules (comme nous l'avons fait pour les arbres exposés précédemment) est l'approche la plus simple mais également la plus coûteuse. Les références listées dans la colonne « Implémentation » sont des articles décrivant soit des outils informatiques dédiés aux modèles de mutation, soit des études utilisant ces outils.

Hypothèses de modélisation	Calcul de la loi	Simulation	Estimation	Implémentation
Durées de vie *i.i.d.* exponentielles	[60, 56, 8, 101, 31]	[31]	[60, 56, 61, 85] [103, 25, 31]	[102, 30, 31] Part. 4.2
Durées de vie constantes	[86, 81, 105, 97] Sous-part 3.3.1	[86, 81, 105, 97] Sous-part 3.3.1	[105, 97] Part. 4.1	[105, 97] Part. 4.2
Durées de vie *i.i.d.* quelconques	[97] Chap. 3	[97] Chap. 3	[97] Part. 4.1	[97] Part. 4.2
Durées de vie inhomogènes	Chap. 3	Chap. 3	Part. 4.1	Part. 4.2
Taux de croissance des mutantes et des cellules normales différents	[8, 47, 103, 31] Chap. 3	[31] Chap. 3	[103, 31] Part. 4.1	[31, 29] Part. 4.2
Morts cellulaires	[87, 3, 18, 48, 98] Chap. 3	[87, 98] Chap. 3	[48, 98] Part. 4.1	[98] Part. 4.2
Fluctuations des nombres finaux	[3, 48, 99]	[99]	[99] Part. 4.1	[99] Part. 4.2
Dilution	[86, 2] Sous-part. 3.3.3	Sous-part. 3.3.3	[86, 39] Part. 4.1	[29] Part. 4.2

TABLE 2.4 – **Résumé du chapitre 2**. Pour chacune des hypothèses de modélisations mentionnées dans le chapitre 2, les principales références citées sont listées selon les quatre problématiques à traiter.

Chapitre 3

Modèles de mutations avec dépendance en âge

Ce chapitre est dédié à la partie probabiliste de cette thèse. Nous y décrivons des modèles de mutations dans lesquels les processus de croissance des populations sont inhomogènes : l'instant de division d'une cellule, quelle que soit sa nature, dépend de sa date de naissance (c'est-à-dire de l'instant de division de sa mère). Sous certaines hypothèses concernant la loi des instants de divisions des cellules, la fonction génératrice du nombre de cellules mutantes à un instant donné peut être explicitée. Il est alors possible d'en déduire, dans le contexte asymptotique habituel, la convergence de la loi du nombre final de mutantes lorsque le nombre initial de cellules non-mutantes tend vers l'infini. Nous obtenons alors une famille de lois de probabilités qui dépendent du nombre moyen de mutations, de la probabilité de mort des cellules mutantes, du taux de division instantané des cellules normales et de celui des cellules mutantes. Nous démontrons ce résultat via deux approches différentes. La première preuve est analytique et généralise la démonstration de Bartlett [8, part. 4.31]. Elle est déduite de l'écriture de l'équation intégrale de Bellman et Harris [9, eq. (2)] dans le cas des processus de branchement inhomogènes en temps. La convergence en loi du nombre final de mutantes peut également être obtenue en démontrant la décomposition du modèle selon les trois niveaux suivants :

1. l'apparition de mutations aléatoires au cours d'un processus de croissance de population. Les divisions cellulaires sont très nombreuses, et la probabilité de mutation est faible, ce qui justifie une approximation poissonnienne pour le nombre de mutations survenues ;
2. les instants de mutations. Nous montrerons que la loi jointe des instants de mutations est équivalente à celle des instants d'occurrence d'un certain processus de Poisson inhomogène ;
3. le nombre de cellules qu'un clone issu d'une cellule mutante née à un instant donné, atteint au bout d'une certaine durée.

La décomposition en trois niveaux introduite par Hamon et Ycart [31] est ainsi généralisée aux modèles de mutation inhomogènes en temps. Cette approche dite *simplifiée* permet d'expliciter les probabilités du nombre final de mutantes et d'obtenir des algorithmes de simulation rapide de modèles de mutations. Nous considérons plusieurs exemples particuliers, tels que le modèle de Haldane avec morts cellulaires ainsi que le cas où les taux instantanés de divisions des non-mutantes et celui des mutantes existent et sont proportionnels. Ce dernier cas est par ailleurs une généralisation du modèle de Luria-Delbrück au cas où les durées de vies ne sont plus forcément exponentielles, mais répondent à d'autres conditions. De plus, sous des hypothèses supplémentaires, il est possible de modéliser le développement d'une population selon une fonction croissante quelconque et bornée.

Nous commencerons par justifier et décrire le modèle probabiliste dans la partie 3.1. Les résultats sur la fonction génératrice du nombre final de mutantes sont ensuite exposés dans la partie 3.2. Nous aborderons le problème selon les deux approches mentionnées ci-dessus. Enfin nous nous intéresserons dans la partie 3.3 aux formulations particulières que sont le modèle de Haldane, ainsi que le cas où les fonctions $F_\nu(s,\cdot)$ et $F_\mu(s,\cdot)$ admettent des taux instantanés de division qui sont proportionnels. Nous généralisons ainsi la notion de fitness présentée dans le chapitre précédent. Nous donnons pour ce cas particulier l'expression des probabilités du nombre final de mutantes et un algorithme de simulation rapide. Nous donnons également une formulation du modèle prenant en compte une éventuelle dilution de la culture. Nous utiliserons dans ce chapitre les notations de la table 1.1 complétées par celles de la table 3.1.

3.1 Hypothèses et modèle

3.1.1 Justification biologique et intuitions

Dans les modèles de mutation classiques, les cellules ont des durées de vie indépendantes et identiquement distribuées. La population se développe alors selon une croissance exponentielle. En pratique, la quantité de ressources disponible dans l'environnement où évoluent les cellules est limitée. Considérer que la croissance d'une population de cellules est exponentielle revient alors à dire que la croissance s'arrête net dès que la capacité maximale d'accueil est atteinte (courbe noire sur la figure 3.1). Or, les observations empiriques montrent que le nombre de cellules croît selon une courbe logistique (courbe bleue sur la figure 3.1) [54] : le nombre de divisions cellulaires augmente exponentiellement jusqu'à un certain instant (point d'inflexion), à partir duquel il diminue jusqu'à devenir asymptotiquement nul. La modélisation déterministe la plus connue de ce type de croissance a été décrite par Verhulst [90, 91]. Les fonctions logistiques sont caractérisées par le fait que les deux asymptotiques sont approchées de manière symétrique par rapport au point d'inflexion. Les fonctions de Gompertz permettent également de décrire des croissances ralentissant à mesure que la taille de la population s'approche de la capacité d'accueil, mais pour lesquelles il n'y a pas de symétrie entre les deux asymptotiques.

$\boldsymbol{N}(s,t)$	Couple du nombre de cellules normales et du nombre de mutantes vivantes à un instant t dans un clone issu d'une cellule normale née à un instant s
$Z_n(t)$	Nombre de mutations apparues dans un intervalle de temps $[0\,;\,t]$ dans une population homogène initialement constituée de n cellules normales
$\left(T_i^{(n)}\right)_{i\geqslant 1}$	Suite croissante des instants de mutation dans une population homogène initialement constituée de n cellules normales
$\boldsymbol{T^{(n)}}$	Vecteurs des k premiers instants de mutation dans une population homogène initialement constituée de n cellules normales
γ	Probabilité de mort d'une cellule normale
δ	Probabilité de mort d'une cellule mutante
$F_\nu(s,\cdot)$	Fonction de répartition de l'instant de fin de développement d'une cellule normale née à un instant s
$F_\mu(s,\cdot)$	Fonction de répartition de l'instant de fin de développement d'une cellule mutante née à un instant s

TABLE 3.1 – **Notations du chapitre 3**

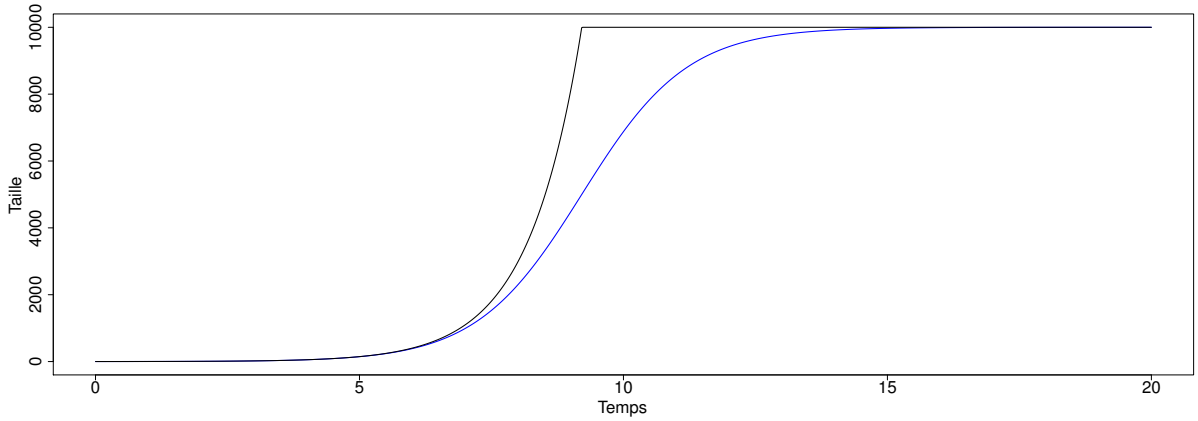

FIGURE 3.1 – **Comparaison d'une croissance logistique et d'une croissance exponentielle.**

Nous nous intéressons dans cette thèse à l'extension des modèles de mutation classiques au cas où la croissance n'est plus exponentielle. Nous allons donner ici quelques intuitions quant aux conséquences auxquelles nous nous attendons. Tout d'abord, le nombre moyen de divisions de cellules normales étant le même dans les deux croissances, le nombre moyen de mutations devrait rester équivalent. La différence réside dans les instants d'apparition des mutations : ces événements ont plus de chances de survenir lorsque le nombre de divisions cellulaires est plus grand. Dans le cas d'une croissance exponentielle, les mutations ont donc plus de chances d'arriver en fin d'expérience. C'est d'ailleurs l'argument justifiant l'ingrédient (A_2) dans la décomposition des modèles de mutation classiques [31]. Dans le cas d'une croissance logistique, le nombre de divisions est plus grand autour du point d'inflexion, c'est-à-dire plus tôt dans l'expérience. De fait, la durée de développement d'un clone devrait être plus longue que dans le cas d'une croissance exponentielle. Dans les modèles décrits dans le chapitre précédent, la durée de développement d'un clone est paramétrée par la fitness : plus elle est faible, plus la durée de développement est grande.

3.1.2 Définition du modèle

Nous donnons à présent une description du modèle probabiliste sous forme d'un processus indexé par un arbre (voir entre autres [10, 76]). Soit $\mathbb{T}$ l'arbre binaire complet de taille infinie et 0 sa racine. Les nœuds de $\mathbb{T}$ représentent ici les cellules. Si v est un nœud de $\mathbb{T}$, le nombre de nœuds entre la racine 0 et v sera noté $|v|$. L'arbre binaire constitué du nœud v et de ses descendants sera noté $\mathbb{T}_v$. Si u et v sont deux nœuds distincts de $\mathbb{T}$ alors : $u \preccurlyeq v$ est la relation d'ordre vérifiée si u est sur le chemin entre 0 et v ; $u \wedge v$ est

l'ancêtre commun le plus récent de u et v, ce qui correspond formellement à :

$$u \wedge v = \underset{w \in \bigwedge(u,v)}{\arg\max}(|w|)\,,$$

où

$$\begin{aligned}\bigwedge(u,v) &= \left\{w \in \mathbb{T} \text{ t.q. } w \preccurlyeq u \text{ et } w \preccurlyeq v\right\} \\ &= \left\{w \in \mathbb{T} \text{ t.q. } u, v \in \mathbb{T}_w\right\},\end{aligned}$$

désigne l'ensemble des ancêtres communs à u et v. Pour alléger la rédaction, la mention « le plus récent » sera dorénavant sous-entendue. Notons que si $u \preccurlyeq v$, alors $u \wedge v = u$. La mère d'une cellule v est notée $\widetilde{v}$: c'est la cellule telle que $\widetilde{v} \preccurlyeq v$ et $|\widetilde{w}| = |v| - 1$. La taille de l'arête reliant une cellule v à sa mère $\widetilde{v}$ représente la durée de vie de la cellule $\widetilde{v}$. Chaque cellule v_0 différente de 0 a une sœur v_1 : elles satisfont $\widetilde{v_0} = \widetilde{v_1} = v_0 \wedge v_1$.

Le développement d'un clone issu d'une unique cellule présente au temps 0 sera modélisé selon un processus aléatoire $(C_v)_{v\in\mathbb{T}}$ indexé par l'arbre binaire $\mathbb{T}$. Pour tout $v \in \mathbb{T}$, C_v est un couple (B_v, T_v) où B_v décrit la nature de v :

- $B_v = 0$ si la cellule v est morte ;
- $B_v = 1$ si la cellule v est une non-mutante ;
- $B_v = 2$ si la cellule v est une mutante ;

et T_v correspond à l'instant où v finit son développement. Plus généralement, T_v sera appelé *instant final* ou *de fin de développement* de v. Notons $\overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{+\infty\}$ la demi-droite des réels positifs étendue, et $\mathcal{B}(\overline{\mathbb{R}}_+)$ sa tribu borélienne. D'après les hypothèses précédentes, le processus aléatoire $(C_v)_{v\in\mathbb{T}}$ est défini sur l'espace mesurable $(\Omega, \mathcal{A})$, avec $\Omega = \{0,1,2\} \times \overline{\mathbb{R}}_+$ et la tribu associée $\mathcal{A} = \mathcal{P}(\{0,1,2\}) \times \mathcal{B}(\overline{\mathbb{R}}_+)$. Le fait que T_v puisse être infini sera expliqué plus tard dans cette sous-partie. Définissons à présent la loi de probabilité de cet espace. Rappelons que la date de naissance de la racine est fixée à 0. Supposons que sa nature B_0 est connue. Précisons également que, si une cellule meurt, alors ses descendantes n'existent pas. En d'autres termes, si $B_v = 0$ alors pour tout nœud $w \in \mathbb{T}_v$, la variable B_w vaut 0.

Considérons une cellule $v_0 \neq 0$ et sa sœur v_1. Leur nature B_{v_0} et B_{v_1} dépendent uniquement de la nature de leur mère $B_{\widetilde{v_0}}$:

- si $B_{\widetilde{v_0}} = 0$, alors $B_{v_0} = B_{v_1} = 0$ avec probabilité 1 ;
- si $B_{\widetilde{v_0}} = 1$, alors :
  - $B_{v_0} = 1$ et $B_{v_1} = 0$ avec probabilité $\pi/2$;
  - $B_{v_0} = 0$ et $B_{v_1} = 1$ avec probabilité $\pi/2$;
  - $B_{v_0} = B_{v_1} = 0$ avec probabilité γ ;
  - $B_{v_0} = B_{v_1} = 1$ avec probabilité $1 - \pi - \gamma$;

— si $B_{\widetilde{v_0}} = 2$, alors :
  — $B_{v_0} = B_{v_1} = 0$ avec probabilité δ ;
  — $B_{v_0} = B_{v_1} = 2$ avec probabilité $1 - \delta$;

En d'autres termes, lorsqu'elle a terminé son développement, une cellule normale produit une cellule normale et une mutante avec probabilité π (*mutation*), meurt avec probabilité γ ou donne naissance à deux non-mutantes avec probabilité $1 - \pi - \gamma$. Une cellule mutante, lorsqu'elle a terminé son développement, meurt avec probabilité δ ou donne naissance à deux mutantes avec probabilité $1-\delta$. Les événements de mort et de mutation ne dépendent donc pas de l'instant final de la cellule.

Pour toute cellule v, son instant final T_v dépend de sa nature B_v et de sa date de naissance, c'est-à-dire de l'instant de division de sa mère $T_{\widetilde{v}}$: si $B_v = 1$ et $T_{\widetilde{v}} = s$, la fonction de répartition de T_v est $F_\nu(s, \cdot)$; si $B_v = 2$ et $T_{\widetilde{v}} = s$, la fonction de répartition de T_v est $F_\mu(s, \cdot)$. Ces deux fonctions vérifient $F_\nu(s, t) = 0$ et $F_\mu(s, t) = 0$ pour $t \leqslant s$. De plus, l'instant final d'une cellule (quelque soit sa nature) née en à un temps donné peut être infini avec une probabilité non-nulle. Cette hypothèse se justifie par la majoration en pratique du nombre total de cellules par la capacité d'accueil, c'est-à-dire le nombre maximal de cellules pouvant survivre dans le milieu considéré. À mesure que le nombre de cellules s'approche de cette borne, et donc que les ressources disponibles baissent, la croissance de la population diminue. En d'autres termes, certaines cellules ne produisent pas de descendante avant la fin du processus de croissance, ce que l'on interprète ici comme un instant final infini. Cette hypothèse requiert la notion de mesure de sous-probabilité sur un espace mesurable $(\Omega\,,\,\mathcal{A})$, c'est-à-dire une mesure ς^* sur $(\Omega\,,\,\mathcal{A})$ telle que $\varsigma^*(\Omega) \leqslant 1$ [74, p. 170]. Considérons une telle mesure ς^* sur $(\mathbb{R}_+\,,\,\mathcal{B}(\mathbb{R}_+))$. Alors la limite lorsque x tend vers l'infini de sa fonction de répartition $F^*(x)$ est plus petite que 1. Pour tout $a, b \in \mathbb{R}_+$, définissons la fonction $\tilde{\varsigma}$ par

$$\tilde{\varsigma}(]a\,;b[) = \varsigma^*(]a\,;b[)\,; \quad \tilde{\varsigma}([0\,;b[) = \varsigma^*([0\,;b[)\,;$$

et

$$\tilde{\varsigma}(]a\,;+\infty]) = \varsigma^*(]a\,;+\infty[)\,.$$

Comme les ensembles de la forme $]a\,;b[$, $[0\,;b[$, ou $]a\,;+\infty]$ engendrent les boréliens $\mathcal{B}(\overline{\mathbb{R}}_+)$, il est possible par le théorème de Carathéodory de prolonger $\tilde{\varsigma}$ sur $\mathcal{B}(\overline{\mathbb{R}}_+)$. À partir de la mesure $\tilde{\varsigma}$, il est ensuite possible de définir une mesure de probabilité ς pour tout $A \in \mathcal{B}(\overline{\mathbb{R}}_+)$:

$$\varsigma(A) = \tilde{\varsigma}(A) + (1 - \tilde{\varsigma}(\mathbb{R}_+))\,\mathbb{1}_{A \in \mathcal{B}(\overline{\mathbb{R}}_+) \setminus \mathcal{B}(\mathbb{R}_+)}$$

de fonction de répartition :

$$F(x) = F^*(x)\mathbb{1}_{x \in [0\,;+\infty)} + \mathbb{1}_{x=+\infty}\,. \tag{3.1}$$

Notons que si ς^* est bien une mesure de probabilité sur $\mathbb{R}_+$, alors ς est également une mesure de probabilité sur $\mathbb{R}_+$.

Le processus aléatoire $(C_v)_{v\in\mathbb{T}} = (B_v, T_v)_{v\in\mathbb{T}}$ ainsi défini décrit le développement d'un clone issu d'une cellule née à l'instant 0. Cette définition peut être étendue à un clone issu d'une unique cellule u née à un instant $s > 0$. Le développement de ce clone consiste en un nouveau processus aléatoire $(C_v)_{v\in\mathbb{T}_u}$ indexé par l'arbre binaire $\mathbb{T}_u$, avec les mêmes hypothèses de modélisation que ci-dessus, conditionnellement à C_u.

Il ne reste plus qu'à définir les dépendances entre les cellules. Soit une cellule $v_0 \neq 0$ et sa sœur v_1. Les instants finaux T_{v_0} et T_{v_1} sont supposés indépendants, conditionnellement à $C_{\widetilde{v_0}}$. En conséquence, les clones $(C_w)_{w\in\mathbb{T}_{v_0}}$ et $(C_w)_{w\in\mathbb{T}_{v_1}}$ sont également indépendants, conditionnellement à $C_{\widetilde{v_0}}$. Considérons à présent deux cellules $u \neq 0$ et $v \neq 0$ et leur ancêtre commun $u\wedge v$. Supposons que cet ancêtre commun n'est ni u ni v. Une des deux filles de $u\wedge v$ se situe entre 0 et u, et l'autre se situe entre 0 et v. Ainsi, d'après les hypothèses que nous venons de poser, les instants finaux T_u et T_v sont indépendants, conditionnellement à $C_{u\wedge v}$. En conséquence, les clones $(C_w)_{w\in\mathbb{T}_u}$ et $(C_w)_{w\in\mathbb{T}_v}$ sont également indépendants, conditionnellement à $C_{u\wedge v}$.

À partir de maintenant, la racine 0 est une cellule normale, c'est-à-dire $B_0 = 1$. Les hypothèses de modélisation peuvent finalement être résumées ainsi :

- à l'instant 0, une unique cellule normale est présente ;
- l'instant final d'une cellule dépend de sa nature et de sa date de naissance ;
- l'instant final d'une cellule normale née à un instant s est une variable aléatoire de fonction de répartition $F_\nu(s, \cdot)$ définie sur $\overline{\mathbb{R}}_+$;
- lorsqu'une cellule normale a terminé son développement :
  - une cellule normale et une mutante sont produites avec probabilité π ;
  - la cellule meurt avec probabilité γ ;
  - deux cellules normales sont produites avec probabilité $1 - \gamma - \pi$;
- l'instant final d'une cellule mutante née à un instant s est une variable aléatoire de fonction de répartition $F_\mu(s, \cdot)$ définie sur $\overline{\mathbb{R}}_+$;
- lorsqu'une cellule mutante a terminé son développement :
  - la cellule meurt avec probabilité δ ;
  - deux cellules mutantes sont produites avec probabilité $1 - \delta$;
- pour chaque cellule, les événements de mort ou de mutation ne dépendent pas de son instant final ;
- deux cellules, quelles que soient leurs natures, sont indépendantes conditionnellement à leur ancêtre commun ;
- deux clones sont indépendants conditionnellement à l'ancêtre commun des cellules qui ont démarré ces clones.

3.2 Étude de la fonction génératrice

Cette partie est dédiée aux résultats obtenus concernant la loi du nombre final de mutantes. Tous les résultats sont exprimés en termes de fonctions génératrices. Nous considérons la fonction génératrice de $\boldsymbol{N}(s,t)$, définie par

$$\varphi(y,z,s,t) = \sum_{i,j\geqslant 0} y^i z^j \mathbb{P}\left[\boldsymbol{N}(s,t) = (i,j)\right] ,$$

ainsi que la fonction génératrice de $\widetilde{M}(s,t)$:

$$\widetilde{\psi}(z,s,t) = \sum_{j\geqslant 0} z^j \mathbb{P}\left[\widetilde{M}(s,t) = j\right] .$$

Nous commençons par étendre l'équation intégrale (2) de Bellman et Harris [9] au cas général des processus de branchement multitypes avec des durées de vie non identiquement distribuées. Cette généralisation est prouvée selon des arguments similaires à ceux exposés pour le cas *i.i.d.* par Kimmel et Axelrod [46, chap. 5]. Nous obtenons ainsi l'équation (BHM). Des descriptions de processus où les durées de développement sont *i.i.d.* peuvent être trouvées dans [6, chap. 5] ou plus récemment dans [46, chap. 6]. Considérerons donc un processus de branchement multitype avec un nombre l de types de cellules, dont les hypothèses de modélisation sont les suivantes :

- l'instant final d'une cellule de type i née à un instant s est une variable aléatoire de fonction de répartition $F_i(s,\cdot)$ définie sur $\overline{\mathbb{R}}_+$ et telle que $F_i(s,t) = 0$ si $t \leqslant s$;
- lorsqu'une cellule de type i a terminé son développement, elle produit un nombre aléatoire $K_{i,j}$ de cellules de type j ;
- deux cellules sont indépendantes conditionnellement à leur ancêtre commun ;
- deux clones sont indépendants conditionnellement à l'ancêtre commun des cellules qui ont démarré ces clones.

Pour tout $1 \leqslant i \leqslant l$, notons χ_i la fonction génératrice du vecteur $(K_{i,j})_{j=1,\dots,l}$ définie pour tout $\boldsymbol{z} = (z_j)_{j=1,\dots,l} \in [0\,;1]^l$ par

$$\chi_i(\boldsymbol{z}) = \sum_{k_1,\dots,k_l\geqslant 0} \left[\mathbb{P}\left[K_{i,1} = k_1, \dots, K_{i,l} = k_l\right] \prod_{j=1}^{l} z_j^{k_j}\right] .$$

Notons $X_{i,j}(s,t)$ le nombre au temps t de cellules de type j dans le clone démarré par une cellule de type i née à un instant s. Nous cherchons à exprimer la loi jointe du vecteur $\boldsymbol{X}_i(s,t) = (X_{i,j}(s,t))_{j=1,\dots,l}$. Supposons dans un premier temps que la cellule initiale termine son développement à un instant $u > s$. Pour tout temps d'observation t, il y a alors deux possibilités : soit $t < u$, auquel cas le clone issu de la cellule initiale

est uniquement constitué de cette dernière ; soit $t \geqslant u$, auquel cas le nombre de cellules de type j est donné par la somme des cellules de ce type dans les clones démarrés par les cellules produites par cette première division. Ainsi, si $X_{i,j}(s,t|u)$ est le nombre au temps t de cellules de type j dans le clone démarré en s sachant que la cellule initiale (de type i) termine son développement à un instant u, nous obtenons l'équation suivante :

$$X_{i,j}(s,t|u) = \left(\sum_{k=1}^{l} \sum_{h=1}^{K_{i,k}} X_{k,j}^{(h)}(u,t) \right) \mathbb{1}_{t \geqslant u} + \mathbb{1}_{t<u} \mathbb{1}_{i=j} \,,$$

où les $X_{i,j}^{(h)}(u,t)$ sont des copies *i.i.d.* de la variable $X_{i,j}(u,t)$. Rappelons que les clones se développent indépendamment entre eux. En conséquence, si $\phi_i(\boldsymbol{z},s,t|u)$ et $\phi_i(\boldsymbol{z},s,t)$ sont les fonctions génératrices des vecteurs $\boldsymbol{X}_i(s,t|u) = (X_{i,j}(s,t|u))_{j=1,\dots,l}$ et $\boldsymbol{X}_i(s,t)$:

$$\phi_i(\boldsymbol{z},s,t|u) = \chi_i\left[\phi_1(\boldsymbol{z},u,t),\dots,\phi_l(\boldsymbol{z},u,t)\right] \mathbb{1}_{t \geqslant u} + z_i \mathbb{1}_{t<u} \,.$$

En intégrant par rapport à $F_i(s,\cdot)$, nous obtenons l'équation intégrale suivante :

$$\phi_i(\boldsymbol{z},s,t) = \int_s^t \chi_i\left[\phi_1(\boldsymbol{z},u,t),\dots,\phi_l(\boldsymbol{z},u,t)\right] \mathrm{d}F_i(s,u) + z_i(1-F_i(s,t)) \,. \tag{BHM}$$

L'équation (BHM) traduit l'intuition suivante. Pour tout temps $t > s$, il y a deux possibilités : soit la cellule termine son développement après l'instant t avec probabilité $(1-F_i(s,t))$. Le clone n'est alors composé que de la cellule initiale, et $\phi_i(\boldsymbol{z},s,t) = z_i$ (deuxième terme de (BHM)) ; soit l'instant où la cellule termine son développement en un temps u (dans un intervalle de temps $[u\,;u+\mathrm{d}u]$ où $s<u\leqslant t$) avec probabilité $\mathrm{d}F_i(s,u)$. Dans ce cas, chaque cellule de type j produite va, indépendamment des autres cellules, démarrer son propre clone. La taille au temps t de ce clone sera alors donnée par la fonction génératrice $\phi_j(\boldsymbol{z},u,t)$. Le nombre de cellules pour chaque type étant donné par la fonction génératrice χ_i, nous obtenons ainsi la partie intégrale de (BHM).

Nous allons appliquer cette équation dans la sous-partie 3.2.1 aux processus de branchements binaires qui nous intéressent dans cette thèse : sans et avec mutations. Dans les deux cas, l'équation intégrale (BHM) peut s'écrire sous forme d'une équation différentielle vérifiée par les fonctions génératrices concernées, sous certaines conditions devant être vérifiées par F_ν ou F_μ que nous introduirons. Les solutions générales de ces équations peuvent ensuite être explicitées. Les deux sous-parties suivantes sont ensuite dédiées à l'étude asymptotique des modèles avec dépendance en âge, sous les conditions que nous venons de mentionner.

Une première approche analytique est décrite dans la sous-partie 3.2.2. Nous appliquons l'équation (BHM) à la fonction génératrice φ, et obtenons une équation différentielle dont la solution générale peut être explicitée lorsque $\gamma = 0$. Nous pouvons ensuite en déduire l'expression de la loi du nombre final de mutantes à un instant t donné dans un clone démarré avec une unique cellule non-mutante, via sa fonction génératrice $\psi(\cdot,t)$.

Les clones se développant indépendamment les uns des autres, il est ensuite possible de considérer le cas où n cellules normales sont initialement présentes : la fonction génératrice du nombre final de mutantes dans n clones, chacun démarré par une unique cellule normale est alors donnée par $\psi(\cdot,t)^n$. Une étude analytique des différents termes de $\psi(\cdot,\tau_n)$ mène enfin au résultat principal (Théorème 3.2.1). Les probabilités du nombre final de mutantes peuvent alors être déduites. La démonstration de Bartlett [8, part. 4.31] est ainsi généralisée au cas où les instants de divisions des cellules dépendent de leur date de naissance.

La deuxième démarche exposée dans la sous-partie 3.2.3 généralise la décomposition proposée par Hamon et Ycart [31]. Nous montrons ainsi que tout modèle de mutation peut être décomposé selon trois ingrédients : le nombre aléatoire de mutations apparues durant la croissance de la population ; l'instant d'apparition de chacune des mutations ; la taille d'un clone mutant démarré à un instant donné. Le résultat obtenu est identique à celui obtenu via la démonstration analytique. Cette approche dite « simplifiée » présente cependant plusieurs avantages. Tout d'abord, le cas où le paramètre de mort γ est non-nul peut être considéré. De plus, chaque grandeur intervenant dans les résultats possède une interprétation rigoureuse. Enfin, il est possible de déduire de cette décomposition un algorithme de simulation rapide du nombre final de mutantes.

3.2.1 Écriture en temps fini des fonctions génératrices

Nous allons à présent appliquer les équations construites précédemment aux processus de branchement qui nous intéressent dans cette thèse.

Considérons tout d'abord un processus de branchement binaire avec un seul type de cellules. Les hypothèses de modélisation sont les suivantes :

- l'instant final d'une cellule née à un instant s est une variable aléatoire de fonction de répartition $F(s,\cdot)$ définie sur $\overline{\mathbb{R}}_+$ et telle que $F(s,t)=0$ si $t \leqslant s$;
- lorsqu'une cellule a terminé son développement :
  - elle meurt avec probabilité β ;
  - deux nouvelles cellules sont produites avec probabilité $1-\beta$;
- deux cellules sont indépendantes conditionnellement à leur ancêtre commun ;
- deux clones sont indépendants conditionnellement à l'ancêtre commun des cellules qui ont démarré ces clones.

Nous noterons $\phi(z,s,t)$ la fonction génératrice de la taille à un instant t d'un clone démarré par une unique cellule née à un instant s. La fonction génératrice du nombre de descendantes d'une cellule est donnée par

$$\chi(z) = \beta + (1-\beta)z^2 \,.$$

En appliquant (BHM), nous obtenons l'équation suivante :

$$\phi(z,s,t) = \int_s^t \beta + (1-\beta)\phi(z,u,t)^2 \mathrm{d}F(s,u) + z(1-F(s,t))\,. \tag{3.2}$$

Jusqu'à présent, nous n'avons fait aucune hypothèse sur la fonction F, excepté son domaine de définition et le fait que $F(s,t) = 0$ si $t \leqslant s$. Afin de pouvoir résoudre (3.2), nous allons préciser d'autres conditions devant être vérifiées par F. Nous dirons qu'une fonction F définie sur $\mathbb{R}_+ \times \overline{\mathbb{R}}_+$ et à valeurs dans $[0\,;1]$ vérifie $(\mathcal{H})$ s'il existe une fonction notée F^* définie sur $\mathbb{R}_+^2$ et à valeurs dans $[0\,;1]$ telle que

$(\mathcal{H}_1)$ $\lim\limits_{t\to+\infty} F^*(s,t) \leqslant 1$ pour tout $s \in \mathbb{R}_+$ et $F^*(s,t) = 0$ si $t \leqslant s$;

$(\mathcal{H}_2)$ la fonction F^* est dérivable en s et en t pour tout couple (s,t) tel que $0 < s < t$, croissante en t et décroissante en s ;

$(\mathcal{H}_3)$ pour tout $s \geqslant 0$, $F(s,\cdot)$ est déduite de $F^*(s,\cdot)$ par (3.1) ;

$(\mathcal{H}_4)$ la fonction η définie pour tout $(s,t) \in \mathbb{R}_+^2$ par

$$\eta(s,t) = -\log\left(1 - F^*(s,t)\right)\,,$$

vérifie pour tout $t \geqslant s$:

$$\eta(s,t) = \eta(0,t) - \eta(0,s)\,.$$

Pour tout $s \geqslant 0$, $F^*(s,\cdot)$ est donc la fonction de répartition d'une mesure de sous-probabilité sur $\mathbb{R}_+$. Ainsi la fonction η est par définition positive, dérivable par rapport à s et t, croissante en t et décroissante en s. De plus, pour tout $(s,t) \in \mathbb{R}_+^2$:

$$\eta(s,t) \leqslant \lim_{t\to+\infty} \eta(0,t)\,.$$

La limite ci-dessus existe et est finie si et seulement si la limite pour tout $s \geqslant 0$ de $F^*(s,t)$ lorsque t tend vers l'infini est strictement inférieure à 1. Il existe une fonction λ qui est continue, définie sur $\mathbb{R}_+$ et positive telle que

$$\eta(s,t) = \int_s^t \lambda(u)\mathrm{d}u\,.$$

Les fonctions λ et η peuvent être respectivement interprétées comme le taux de risque instantané et le taux de risque cumulé sur un intervalle de temps $[s\,;t]$. La fonction de répartition $F(s,\cdot)$ peut ainsi être définie sur $\overline{\mathbb{R}}_+$ par

$$F(s,t) = \begin{cases} \left(1 - \exp\left(-\displaystyle\int_s^t \lambda(u)\mathrm{d}u\right)\right)\mathbb{1}_{s\leqslant t} & \text{si } t < +\infty\,, \\ 1 & \text{si } t = +\infty\,. \end{cases} \tag{3.3}$$

Réciproquement, toute fonction de répartition définie par (3.3), à partir d'une fonction λ qui est continue, définie sur $\mathbb{R}_+$ et positive, vérifie cette propriété. En particulier, si la limite de λ en $+\infty$ est nulle, alors la limite de $\eta(0,t)$ lorsque t tend vers l'infini est finie. Alors, pour tout $s \geqslant 0$, $F^*(s,\cdot)$ est la fonction de répartition d'une mesure de sous-probabilité au sens strict : sa limite en $+\infty$ est strictement inférieure à 1. Notons que si $T(s)$ est une variable de fonction de répartition $F(s,\cdot)$, telle que F vérifie $(\mathcal{H})$, alors :

$$\begin{aligned}\mathbb{P}[T(s) > u+t | T(s) > t] &= \frac{\mathrm{e}^{-\eta(s,u+t)}}{\mathrm{e}^{-\eta(s,t)}} \\ &= \mathrm{e}^{-\eta(t,u+t)} \\ &= \mathbb{P}[T(t) > u+t]\,.\end{aligned}$$

Il s'agit d'une propriété semblable à la propriété « sans mémoire » de la loi exponentielle.

Bien que la condition $(\mathcal{H}_4)$ paraisse contraignante, nous verrons par la suite que nous pouvons dans ce cas ajuster la croissance moyenne de la population à n'importe quelle fonction strictement positive, continue et croissante sur $\mathbb{R}_+$. Notons par ailleurs que l'assertion $(\mathcal{H}_4)$ est équivalente à :

$$F^*(s,t) = \frac{F^*(0,t) - F^*(0,s)}{1 - F^*(0,s)}\,.$$

Pour toute fonction de répartition G définie sur $\mathbb{R}_+$, la fonction de répartition $F^*(s,\cdot)$ définie pour tout $s \in \mathbb{R}_+$ par

$$F^*(s,t) = \frac{G(t) - G(s)}{1 - G(s)}\,.$$

vérifie donc l'hypothèse $(\mathcal{H}_4)$.

Remplaçons à présent F par son expression dans (3.2) :

$$\phi(z,s,t) = \int_s^t \left[\beta + (1-\beta)\phi(z,u,t)^2\right] \lambda(u)\mathrm{e}^{-\eta(s,u)}\mathrm{d}u + z\mathrm{e}^{-\eta(s,t)}\,.$$

Par la propriété d'additivité $(\mathcal{H}_4)$:

$$\phi(z,s,t)\mathrm{e}^{-\eta(0,s)} = \int_s^t \left[\beta + (1-\beta)\phi(z,u,t)^2\right] \lambda(u)\mathrm{e}^{-\eta(0,u)}\mathrm{d}u + z\mathrm{e}^{-\eta(0,t)}\,.$$

Puis en dérivant l'équation par rapport à s et en la multipliant par $\mathrm{e}^{\eta(0,s)}$, nous obtenons l'équation de Riccati suivante :

$$\frac{\partial\phi(z,s,t)}{\partial s} = -\lambda(s)\left[\beta - \phi(z,s,t) + (1-\beta)\phi(z,s,t)^2\right]\,, \tag{3.4}$$

avec la condition initiale $\phi(z,t,t) = z$. ici, la solution générale peut être explicitée sans conditions sur les coefficients : il suffit de remarquer que 1 est solution particulière de (3.4) [32].

Proposition 3.2.1. *La solution générale de l'équation de Riccati* (3.4) *est donnée par*

$$\phi(z,s,t) = \frac{\beta(1-z) + \mathrm{e}^{-\eta^*(s,t)}((1-\beta)z-\beta)}{(1-\beta)(1-z) + \mathrm{e}^{-\eta^*(s,t)}((1-\beta)z-\beta)}\,, \tag{3.5}$$

où :

$$\eta^*(s,t) = (1-2\beta)\eta(s,t)\,.$$

Démonstration de la proposition 3.2.1. Comme 1 est solution particulière, la solution générale de (3.4) est donnée par

$$\phi(z,s,t) = 1 + \frac{\exp\left(\displaystyle\int_s^t \lambda(u)\,(-1+2(1-\beta))\,\mathrm{d}u\right)}{C - \displaystyle\int_s^t \lambda(u)\exp\left(\int_u^t \lambda(v)\,(-1+2(1-\beta))\,\mathrm{d}v\right)\mathrm{d}u}\,,$$

où C est une constante. Comme $\phi(z,t,t) = z$, la constante C est donnée par $C = (z-1)^{-1}$. D'où :

$$\begin{aligned}
\phi(z,s,t) &= 1 + \frac{(z-1)\mathrm{e}^{\eta^*(s,t)}}{1-(z-1)\displaystyle\int_s^t \lambda(u)(1-\beta)\mathrm{e}^{\eta^*(u,t)}\mathrm{d}u} \\
&= 1 + \frac{(z-1)\mathrm{e}^{\eta^*(s,t)}}{1-(z-1)\frac{1-\beta}{1-2\beta}\left(\mathrm{e}^{\eta^*(s,t)}-1\right)} \\
&= 1 - \frac{(1-2\beta)(1-z)\mathrm{e}^{\eta^*(s,t)}}{(1-\beta)(1-z)\mathrm{e}^{\eta^*(s,t)} + (1-\beta)z - \beta} \\
&= \frac{\beta(1-z) + \mathrm{e}^{-\eta^*(s,t)}((1-\beta)z-\beta)}{(1-\beta)(1-z) + \mathrm{e}^{-\eta^*(s,t)}((1-\beta)z-\beta)}\,.
\end{aligned}$$

□

Observons que si λ est une constante, alors $\eta(s,t) = \lambda(t-s)$ et la Proposition 3.2.1 correspond à l'exemple donné dans [6, p. 109]. De plus, si λ est une constante et si β est nul, alors (3.4) est une équation de Bernoulli, dont la solution est la fonction génératrice de la loi géométrique de paramètre $\mathrm{e}^{-\lambda(t-s)}$: il s'agit du résultat classique dont la formule est donnée par Yule [100, p. 35] ou encore dans [6, p. 109]. À partir de la proposition 3.2.1, la loi de probabilité de la taille à un instant donné d'un clone démarré en un temps s peut être explicitée.

Proposition 3.2.2. *Notons $(r_k(s,t))_{k\in\mathbb{N}}$ la suite des probabilités associées à la fonction génératrice $\phi(\cdot,s,t)$. Alors :*

$$r_0(s,t) = \frac{\beta(1-\mathrm{e}^{-\eta^*(s,t)})}{1-\beta-\beta\mathrm{e}^{-\eta^*(s,t)}}\,,$$

et pour tout $k \geqslant 1$,

$$r_k(s,t) = (1-r_0(s,t))P(s,t)(1-P(s,t))^{k-1}\,,$$

où :

$$P(s,t) = \frac{(1-2\beta)\mathrm{e}^{-\eta^*(s,t)}}{1-\beta-\beta\mathrm{e}^{-\eta^*(s,t)}}\,,$$

et :

$$\eta^*(s,t) = (1-2\beta)\eta(s,t)\,.$$

En d'autres termes, une variable aléatoire avec $\phi(\cdot,s,t)$ comme fonction génératrice est le mélange aléatoire suivant : soit 0 avec probabilité $r_0(s,t)$, soit la variable géométrique de paramètre $P(s,t)$. De plus, la probabilité de survie à un instant t d'un clone démarré à un instant s est donnée par

$$1-r_0(s,t) = \frac{1-2\beta}{1-\beta-\beta\mathrm{e}^{-\eta^*(s,t)}}\,.$$

En particulier, si $\eta(0,t)$ tend vers l'infini lorsque t tend vers l'infini, alors :

$$\lim_{t\to+\infty} r_0(0,t) = \frac{\beta}{1-\beta}\,,$$

c'est-à-dire la probabilité d'extinction d'un processus de branchement binaire où les durées de vie sont homogènes en temps.

Pour démontrer la proposition 3.2.2, il suffit d'écrire la fonction comme une fraction rationnelle en z et d'en déduire son écriture sous forme de série entière (par rapport à la variable z).

Démonstration de la proposition 3.2.2. La fonction ϕ peut être écrite sous la forme suivante :

$$\phi(z,s,t) = \frac{n_0(s,t)+zn_1(s,t)}{d_0(s,t)+zd_1(s,t)}\,,$$

où

$$n_0(s,t) = \beta\left(1-\mathrm{e}^{-\eta^*(s,t)}\right)\,,\quad n_1(s,t) = -\left(\beta-\mathrm{e}^{-\eta^*(s,t)}(1-\beta)\right)\,,$$

et

$$d_0(s,t) = 1-\beta-\beta\mathrm{e}^{-\eta^*(s,t)}\,,\quad d_1(s,t) = -(1-\beta)\left(1-\mathrm{e}^{-\eta^*(s,t)}\right)\,.$$

Et les probabilités $r_k(s,t)$ peuvent ainsi être définies de manière récursive :

$$r_0(s,t) = \frac{n_0(s,t)}{d_0(s,t)} = \frac{\beta\left(1-\mathrm{e}^{-\eta^*(s,t)}\right)}{1-\beta-\beta\mathrm{e}^{-\eta^*(s,t)}}\,,$$

et

$$\begin{aligned} r_1(s,t) &= \frac{n_1(s,t)}{d_0(s,t)} - \frac{d_1(s,t)}{d_0(s,t)} r_0(s,t) \\ &= \frac{(1-2\beta)^2\mathrm{e}^{-\eta^*(s,t)}}{d_0(s,t)^2} \\ &= (1-r_0(s,t))\,\frac{(1-2\beta)\mathrm{e}^{-\eta^*(s,t)}}{d_0(s,t)}\,. \end{aligned}$$

Soient

$$P(s,t) = \frac{(1-2\beta)\mathrm{e}^{-\eta^*(s,t)}}{d_0(s,t)}\,.$$

Alors pour $k \geqslant 2$:

$$\begin{aligned} r_k(s,t) &= -\frac{d_1(s,t)}{d_0(s,t)} r_{k-1}(s,t) \\ &= \left(-\frac{d_1(s,t)}{d_0(s,t)}\right)^{k-1} (1-r_0(s))P(s,t) \\ &= (1-r_0(s,t))P(s,t)(1-P(s,t))^{k-1}\,. \end{aligned}$$

□

Remarquons que si λ est une constante, la Proposition 3.2.2 correspond à l'équation (3.1) dans [98].

Considérons à présent une fonction f strictement positive, dérivable et croissante sur $\mathbb{R}_+$. Si la fonction de répartition $F^*(s,t)$ est définie pour tout $(s,t) \in \mathbb{R}_+^2$ par

$$F^*(s,t) = 1 - \frac{f(s)}{f(t)}\,, \tag{3.6}$$

alors les fonctions η et λ sont données par

$$\eta(s,t) = \log\left(\frac{f(t)}{f(s)}\right)\,,$$

et

$$\lambda(u) = \frac{f'(u)}{f(u)}\,.$$

L'espérance de la taille à l'instant t d'un clone démarré à un instant s est alors donnée par

$$\left(\frac{f(t)}{f(s)}\right)^{1-2\beta}.$$

En particulier, si $\beta = 0$, il est possible d'ajuster la trajectoire moyenne à une fonction f strictement positive, continue et croissante sur $\mathbb{R}_+$. Remarquons également que l'exemple donné dans [6, p. 109] est encore retrouvé lorsque la fonction f est définie par $f(t) = \mathrm{e}^{\lambda t}$, où λ est une constante positive.

Nous allons à présent considérer le processus de croissance d'une population de cellules au cours de laquelle des mutations apparaissent aléatoirement. Les hypothèses de modélisation seront celles décrites à la fin de la sous-partie 3.1.2. Considérons donc une cellule non-mutante née à un instant s. La fonction génératrice double du nombre de descendantes d'une cellule non-mutante est définie par

$$\chi(y, z) = \gamma + \pi yz + (1 - \pi - \gamma)y^2 .$$

Par simple application de (BHM), la fonction génératrice $\varphi(y, z, s, t)$ vérifie donc l'équation suivante :

$$\begin{aligned}\varphi(y, z, s, t) = \int_s^t \gamma + \pi\varphi(y, z, u, t)\widetilde{\psi}(z, u, t) + (1 - \pi - \gamma)\varphi(y, z, u, t)^2 \mathrm{d}F_\nu(s, u) \\ + y(1 - F_\nu(s, t)) .\end{aligned} \tag{3.7}$$

Aucune hypothèse n'est faite pour le moment sur F_μ. Cependant, nous supposerons dans certains cas ou exemples que F_μ vérifie $(\mathcal{H})$, auquel cas nous noterons η_μ la fonction définie par l'assertion $(\mathcal{H}_4)$ et λ_μ le taux de division instantané associé. Supposons à présent que la fonction de répartition F_ν vérifie $(\mathcal{H})$. Notons η_ν la fonction définie par l'assertion $(\mathcal{H}_4)$, et λ_ν le taux de division instantané associé. Remplaçons F_ν par son expression dans (3.7) :

$$\begin{aligned}\varphi(y, z, s, t) = \int_s^t \left[\gamma + \pi\varphi(y, z, u, t)\widetilde{\psi}(z, u, t) + (1 - \pi - \gamma)\varphi(y, z, u, t)^2\right] \lambda_\nu(u)\mathrm{e}^{-\eta_\nu(s,u)}\mathrm{d}u \\ + y\mathrm{e}^{-\eta_\nu(s,t)} .\end{aligned}$$

Le même raisonnement que pour l'obtention de (3.4) nous donne l'équation de Riccati suivante :

$$\begin{aligned}\frac{\partial\varphi(y, z, s, t)}{\partial s} = -\lambda_\nu(u) \Big[\gamma - (1 - \pi\widetilde{\psi}(z, s, t))\varphi(y, z, s, t) \\ + (1 - \pi - \gamma)\varphi(y, z, s, t)^2\Big] ,\end{aligned} \tag{3.8}$$

avec la condition initiale $\varphi(y, z, t, t) = y$. Selon leurs coefficients, les équations de Riccati peuvent avoir des solutions explicites [51, 32]. Tout d'abord, (3.8) peut être résolue lorsque $\gamma = 0$: l'équation devient alors une équation de Bernoulli d'ordre 2 et il est possible d'expliciter $\varphi(y, z, s, t)$.

Proposition 3.2.3. *Supposons $\gamma = 0$. La solution générale de (3.8) est donnée par*

$$\varphi(y,z,s,t) = \mathrm{e}^{\pi\Upsilon(z,s,t)-\eta_\nu(s,t)}\left\{\frac{1}{y} - (1-\pi)\int_s^t \lambda_\nu(u)\mathrm{e}^{\pi\Upsilon(z,u,t)-\eta_\nu(u,t)}\mathrm{d}u\right\}^{-1}, \tag{3.9}$$

où

$$\Upsilon(z,s,t) = \int_s^t \lambda_\nu(u)\widetilde{\psi}(z,u,t)\mathrm{d}u\,. \tag{3.10}$$

La solution est obtenue en effectuant le changement de variable $\phi = 1/\varphi$. La fonction (3.10) n'a pas de signification particulière et n'est introduite que pour alléger la rédaction. Nous pouvons malgré tout remarquer que $\Upsilon(1,s,t) = \eta_\nu(s,t)$. Si F_μ vérifie $(\mathcal{H})$ telle que $\lambda_\mu(t) = \lambda_\nu(t)$ pour tout $t \in \mathbb{R}_+$, l'équation de Riccati (3.8) peut également être résolue lorsque $\gamma = \delta$. Dans ce cas $\widetilde{\psi}$ est une solution particulière, et la solution générale de (3.8) peut en être déduite. Il s'agit du cas où les mutantes et les non-mutantes suivent exactement la même dynamique (même probabilité de mort et même loi des instants finaux). Il n'y a donc pas de distinction entre les deux types, ce qui semble peu pertinent en pratique. Une solution explicite peut également être obtenue si $1-\pi-\gamma = 0$. Cependant, γ doit être strictement inférieur à 0.5 (processus supercritique), et π est en pratique de l'ordre de 10^{-11} à 10^{-9}. Ce cas est donc irréaliste et ne sera pas étudié ici.

La conséquence directe de la proposition 3.2.3, est l'expression de la fonction génératrice du nombre de mutantes dans une culture démarrée avec une unique cellule non-mutante au temps 0.

Corollaire 3.2.1. *Supposons que $\gamma = 0$. La fonction génératrice du nombre de mutantes au temps t dans une culture démarrée avec une unique cellule non-mutante au temps 0 est donnée par*

$$\psi(z,t) = \mathrm{e}^{\pi\Upsilon(z,0,t)-\eta_\nu(0,t)}\left\{1 - (1-\pi)\int_0^t \lambda_\nu(u)\mathrm{e}^{\pi\Upsilon(z,u,t)-\eta_\nu(u,t)}\mathrm{d}u\right\}^{-1}. \tag{3.11}$$

Dans la suite de cette sous-partie et dans la suivante, le paramètre γ sera supposé nul. Observons qu'aucune hypothèse sur la loi des instants finaux des mutantes n'est nécessaire. En particulier, F_μ peut ne pas vérifier $(\mathcal{H})$. Tant que la fonction génératrice $\widetilde{\psi}$ est connue, la loi du décompte de mutantes à un instant donné peut être explicitée. Nous pourrions considérer comme exemple le modèle de Haldane décrit dans la sous-partie 3.3.1. Nous nous intéresserons également dans la sous-partie 3.3.2 au cas où F_μ vérifie $(\mathcal{H})$ telle que pour tout $u \in \mathbb{R}_+$:

$$\lambda_\nu(u) = (1-2\delta)\rho\lambda_\mu(u)\,,$$

où ρ est une constante positive. La quantité ρ est une généralisation de la notion de fitness décrite plus tôt dans le chapitre 2. Nous commenterons plus en détail cette hypothèse dans

la partie 3.3. Dans ce cas, (3.10) peut être explicitée :

$$\begin{aligned}\Upsilon(z,s,t) &= \frac{\rho(1-2\delta)}{1-\delta}\log\left[\frac{(1-2\delta)\mathrm{e}^{(1-\delta)\eta_\mu(s,t)}}{(1-\delta)\left((1-z)\mathrm{e}^{(1-2\delta)\eta_\mu(s,t)}+z\right)-\delta}\right]\\ &= \eta_\nu(s,t)+\frac{\rho(1-2\delta)}{1-\delta}\log\left[\frac{(1-2\delta)}{(1-\delta)\left((1-z)\mathrm{e}^{(1-2\delta)\eta_\mu(s,t)}+z\right)-\delta}\right].\end{aligned}$$

En particulier, considérons le cas où F_μ est définie par (3.6), avec $\delta = 0$. Alors :

$$\begin{aligned}\psi(z,t) = &\left(\frac{f(t)}{f(0)}\right)^{-\rho}\left(1-z+z\frac{f(0)}{f(t)}\right)^{\pi\rho}\\ &\times\left\{1-(1-\pi)\int_{\frac{f(0)}{f(t)}}^{1}\rho w^{\rho-1}\left(1-z+zw\right)^{-\pi\rho}\mathrm{d}w\right\}^{-1}.\end{aligned}\tag{3.12}$$

Par exemple, si f est définie pour tout $t \geqslant 0$ par $f(t) = \mathrm{e}^t$, (3.12) est donnée par

$$\begin{aligned}\psi(z,t) = &\,\mathrm{e}^{-\rho t}\left(1-z+z\mathrm{e}^{-t}\right)^{-\pi\rho}\\ &\times\left\{1-(1-\pi)\int_0^t\rho\mathrm{e}^{-\rho u}\left(1-z+z\mathrm{e}^{-u}\right)^{-\pi\rho}\mathrm{d}u\right\}^{-1},\end{aligned}$$

ce qui correspond à l'inverse de l'équation (10) dans [8, p. 116]. D'autres types de fonctions telles que les fonctions logistiques ou de Gompertz peuvent également être considérées et injectées dans (3.12).

3.2.2 Étude analytique de l'asymptotique

Notons à présent $\psi_n(z,t)$ la fonction génératrice du nombre final de mutantes lorsque la culture contient initialement n cellules non-mutantes. Comme les clones se développent indépendamment les uns des autres, $\psi_n(z,t)$ correspond à la n-ième puissance de (3.11). Nous cherchons à présent à expliciter la limite de $\psi_n(z,\tau_n)$ lorsque le nombre initial de cellules non-mutantes n tend vers l'infini. Rappelons que F_ν vérifie $\mathcal{H}$. Notons η_ν et λ_ν les taux de division cumulé (défini par ($\mathcal{H}_4$)) et instantané associés à F_ν, ainsi que :

$$\eta_{\nu,\infty} = \lim_{t\to\infty}\eta_\nu(0,t)\,.$$

Rappelons que cette limite peut être finie ou non. Nous commençons par définir la fonction suivante :

$$h(z,t) = \frac{1}{1-\mathrm{e}^{-\eta_\nu(0,t)}}\int_0^t\widetilde{\psi}(z,u,t)\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,t)}\mathrm{d}u\,.$$

Il s'agit pour l'instant d'une notation pour simplifier la rédaction. Cependant, nous montrerons dans la sous-partie suivante que les instants de mutation sont distribués selon la

densité $\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,t)}\mathbb{1}_{[0\,;t]}$. La fonction h sera alors interprétée comme la fonction génératrice de la taille à un instant t d'un clone quelconque. Nous présentons maintenant l'un des résultats principaux de cette thèse.

Théorème 3.2.1. *Soient $\pi = (\pi_n)_{n\in\mathbb{N}}$ et $\tau = (\tau_n)_{n\in\mathbb{N}}$ deux suites de réels positifs, et une constante $\alpha \geqslant 0$ telles que :*

$$\lim_{n\to+\infty} \pi_n = 0\,, \qquad \lim_{n\to+\infty} \tau_n = +\infty\,, \qquad \lim_{n\to+\infty} \pi_n n\mathrm{e}^{\eta_\nu(0,\tau_n)} = \alpha\,. \tag{3.13}$$

Supposons également que la limite

$$\Upsilon(z,0,t)\mathrm{e}^{-\eta_\nu(0,t)} \tag{3.14}$$

existe et est finie. Lorsque n tend vers l'infini, la fonction génératrice du nombre de mutantes au temps τ_n dans les clones issus de n cellules normales converge vers la fonction génératrice :

$$\psi(z) = \exp\left(-m(1-h(z))\right)\,, \tag{3.15}$$

avec

$$\begin{aligned} h(z) &= \lim_{n\to+\infty} h(z,\tau_n) \\ &= \frac{1}{1-\mathrm{e}^{-\eta_{\nu,\infty}}} \lim_{t\to+\infty} \int_0^t \widetilde{\psi}(z,u,t)\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,t)}\mathrm{d}u\,, \end{aligned}$$

et

$$m = \alpha\left(1-\mathrm{e}^{-\eta_{\nu,\infty}}\right)\,.$$

Notons que (3.15) est la fonction génératrice d'une composée poissonnienne de paramètre m. Nous montrerons dans la sous-partie suivante que le nombre d'occasions de mutation est presque sûrement équivalent à $n\left(\mathrm{e}^{\eta_\nu(0,\tau_n)}-1\right)$ lorsque n tend vers l'infini. Ainsi, la quantité m correspond au nombre moyen de mutations.

Une preuve analytique dans le cas où les durées de vie des cellules sont exponentiellement *i.i.d.* a été donnée dans [8, part. 4.31]. Cette approche est adaptée pour démontrer le théorème 3.2.1. L'outil principal de cette démonstration est le lemme suivant.

Lemme 3.2.1. *Pour tous $\pi \in [0\,;1[$, $z \in U$, $t \in \mathbb{R}_+$, et $s \in [0\,;t]$, l'inégalité suivante est vérifiée :*

$$\left|\mathrm{e}^{\pm\pi\Upsilon(z,s,t)} - (1 \pm \Upsilon(z,s,t))\right| \leqslant \pi^2\mathrm{e}^{\Upsilon(z,0,t)}\,.$$

L'inégalité ci-dessus est obtenue grâce à un développement en série entière de $\mathrm{e}^{\pm\pi\Upsilon(z,s,t)}$.

Démonstration du lemme 3.2.1. À partir du développement suivant :

$$\mathrm{e}^{\pm\pi\Upsilon(z,s,t)} = \sum_{k\geqslant 0} \frac{(\pm\pi\Upsilon(z,s,t))^k}{k!}\,,$$

il vient :

$$\begin{aligned}\left|\mathrm{e}^{\pm\pi\Upsilon(z,s,t)} - (1 \pm \pi\Upsilon(z,s,t))\right| &\leqslant \sum_{k\geqslant 2} |\pm\pi|^k \frac{\Upsilon(z,s,t)^k}{k!} \\ &\leqslant \pi^2 \mathrm{e}^{\Upsilon(z,s,t)} \\ &\leqslant \pi^2 \mathrm{e}^{\Upsilon(z,0,t)}\,.\end{aligned}$$

□

Nous donnons à présent la preuve du théorème 3.2.1.

Démonstration du théorème 3.2.1. Définissons tout d'abord les deux fonctions suivantes :

$$f_1(z,u,\tau_n,\pi_n) = \mathrm{e}^{\pi_n\Upsilon(z,u,\tau_n)} - (1+\pi_n\Upsilon(z,u,\tau_n)) \quad \text{et} \quad f_2(z,\tau_n,\pi_n) = f_1(z,0,\tau_n,\pi_n)\,.$$

D'après le Lemme 3.2.1 :

$$|f_1(z,u,\tau_n,\pi_n)| \leqslant \pi_n^2 \mathrm{e}^{\Upsilon(z,0,\tau_n)} \quad \text{et} \quad |f_2(z,\tau_n,\pi_n)| \leqslant \pi_n^2 \mathrm{e}^{\Upsilon(z,0,\tau_n)}\,. \tag{3.16}$$

Afin d'alléger la rédaction, les arguments des fonctions f_1 et f_2 ne seront plus précisés. Le second facteur de (3.11) est :

$$1 - (1-\pi_n)\int_0^{\tau_n} \lambda_\nu(u)\mathrm{e}^{\pi_n\Upsilon(z,u,\tau_n)}\mathrm{e}^{-\eta_\nu(u,\tau_n)}\mathrm{d}u\,.$$

Il s'écrit donc :

$$\begin{aligned}&1 - (1-\pi_n)\left[\int_0^{\tau_n} \lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,\tau_n)} f_1 \mathrm{d}u + \left(1-\mathrm{e}^{-\eta_\nu(0,\tau_n)}\right)\right. \\ &\qquad\left. +\pi_n\int_0^{\tau_n}\Upsilon(z,u,\tau_n)\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,\tau_n)}\mathrm{d}u\right] \\ =\,&\mathrm{e}^{-\eta_\nu(0,\tau_n)} - \int_0^{\tau_n}\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,\tau_n)} f_1 \mathrm{d}u \\ &+\pi_n\left[1-\mathrm{e}^{-\eta_\nu(0,\tau_n)} - \int_0^{\tau_n}\left(\Upsilon(z,u,\tau_n)+f_1\right)\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,\tau_n)}\mathrm{d}u\right] \\ &+\pi_n^2\int_0^{\tau_n}\Upsilon(z,u,\tau_n)\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,\tau_n)}\mathrm{d}u\,.\end{aligned}$$

Soit :

$$\begin{aligned}f_3(z,\tau_n) &= \int_0^{\tau_n}\Upsilon(z,u,\tau_n)\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,\tau_n)}\mathrm{d}u \\ &= \left(1-\mathrm{e}^{-\eta_\nu(0,\tau_n)}\right)h(z,\tau_n) - \mathrm{e}^{-\eta_\nu(0,\tau_n)}\Upsilon(z,0,\tau_n)\,,\end{aligned}$$

et

$$f_4(z,\tau_n,\pi_n) = \int_0^{\tau_n} \lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,\tau_n)} f_1(z,u,\tau_n,\pi_n)\mathrm{d}u\,.$$

D'après (3.14), la limite de $f_3(z,\tau_n)$ lorsque n tend vers l'infini existe et est finie. Considérons à présent le terme suivant :

$$\mathrm{e}^{-\pi_n\Upsilon(z,0,\tau_n)}\left\{1-(1-\pi_n)\int_0^{\tau_n}\lambda_\nu(u)\mathrm{e}^{\pi_n\Upsilon(z,u,\tau_n)}\mathrm{e}^{-\eta_\nu(u,\tau_n)}\mathrm{d}u\right\}.$$

Nous pouvons le réécrire ainsi :

$$\begin{aligned}
&(f_2+1-\pi_n\Upsilon(z,0,\tau_n))\left\{\mathrm{e}^{-\eta_\nu(0,\tau_n)}-f_4(z,\tau_n,\pi_n)\right.\\
&\qquad\left.+\pi_n\left(1-\mathrm{e}^{-\eta_\nu(0,\tau_n)}-f_3(z,\tau_n)+f_4(z,\tau_n,\pi_n)\right)+\pi_n^2 f_3(z,\tau_n)\right\}\\
&=(f_2+1)\left(\mathrm{e}^{-\eta_\nu(0,\tau_n)}-f_4(z,\tau_n,\pi_n)\right)\\
&\quad+\pi_n\left\{(f_2+1)\left(1-\mathrm{e}^{-\eta_\nu(0,\tau_n)}-f_3(z,\tau_n)+f_4(z,\tau_n,\pi_n)\right)\right.\\
&\qquad\left.-\Upsilon(z,0,\tau_n)\left(\mathrm{e}^{-\eta_\nu(0,\tau_n)}-f_4(z,\tau_n,\pi_n)\right)\right\}\\
&\quad+\pi_n^2\left\{(f_2+1)f_3(z,\tau_n)-\Upsilon(z,0,\tau_n)\left(1-\mathrm{e}^{-\eta_\nu(0,\tau_n)}-f_3(z,\tau_n)+f_4(z,\tau_n,\pi_n)\right)\right\}\\
&\quad-\pi_n^3\Upsilon(z,0,\tau_n)f_3(z,\tau_n)\,.
\end{aligned}$$

En multipliant par $\mathrm{e}^{\eta_\nu(0,\tau_n)}$, il vient :

$$\begin{aligned}
\frac{1}{\psi(z,\tau_n)}&=(f_2+1)\left(1-\mathrm{e}^{\eta_\nu(0,\tau_n)}f_4(z,\tau_n,\pi_n)\right)\\
&\quad+\pi_n\left\{(f_2+1)\left(\mathrm{e}^{\eta_\nu(0,\tau_n)}-1-\mathrm{e}^{\eta_\nu(0,\tau_n)}f_3(z,\tau_n)+\mathrm{e}^{\eta_\nu(0,\tau_n)}f_4(z,\tau_n,\pi_n)\right)\right.\\
&\qquad\left.-\Upsilon(z,0,\tau_n)\left(1-\mathrm{e}^{\eta_\nu(0,\tau_n)}f_4(z,\tau_n,\pi_n)\right)\right\}\\
&\quad+\pi_n^2\left\{(f_2+1)\mathrm{e}^{\eta_\nu(0,\tau_n)}f_3(z,\tau_n)\right.\\
&\qquad\left.-\Upsilon(z,0,\tau_n)\left(\mathrm{e}^{\eta_\nu(0,\tau_n)}-1-\mathrm{e}^{\eta_\nu(0,\tau_n)}f_3(z,\tau_n)+\mathrm{e}^{\eta_\nu(0,\tau_n)}f_4(z,\tau_n,\pi_n)\right)\right\}\\
&\quad-\pi_n^3\Upsilon(z,0,\tau_n)\mathrm{e}^{\eta_\nu(0,\tau_n)}f_3(z,\tau_n)\,.
\end{aligned}$$

Notons à présent que d'après l'inégalité vérifiée par f_1 dans (3.16) :

$$\begin{aligned}
f_4(z,\tau_n,\pi_n)&\leqslant\pi_n^2\mathrm{e}^{\Upsilon(z,0,\tau_n)}\int_0^{\tau_n}\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,\tau_n)}\mathrm{d}u\\
&\leqslant\pi_n^2\mathrm{e}^{\Upsilon(z,0,\tau_n)}\left(1-\mathrm{e}^{-\eta_\nu(0,\tau_n)}\right)\\
&\leqslant\pi_n^2\mathrm{e}^{\eta_\nu(0,\tau_n)}\,.
\end{aligned}$$

En effet, $\widetilde{\psi}$ étant une fonction génératrice, il vient pour $z\in U$ et $(s,t)\in\mathbb{R}_+^2$:

$$\Upsilon(z,s,t)=\int_s^t\lambda_\nu(u)\widetilde{\psi}(z,u,t)\mathrm{d}u\leqslant\int_s^t\lambda_\nu(u)\mathrm{d}u=\eta_\nu(s,t)\,.$$

Ainsi, comme $n\pi_n e^{\eta_\nu(0,\tau_n)}$ tend vers α lorsque n tend vers l'infini, $\pi_n e^{\eta_\nu(0,\tau_n)}$ tend vers 0. Donc :

$$\lim_{n\to+\infty} n e^{\eta_\nu(0,\tau_n)} f_4(z,\tau_n,\pi_n) = 0\,.$$

Notons $o(\pi_n,\tau_n)$ toute fonction telle que

$$\lim_{n\to+\infty} n o(\pi_n,\tau_n) = 0\,.$$

Alors :

$$\begin{aligned}\frac{1}{\psi(z,\tau_n)} &= f_2 + 1 + \pi_n\left\{(f_2+1)\left(e^{\eta_\nu(0,\tau_n)}(1-f_3(z,\tau_n))-1\right) - \Upsilon(z,0,\tau_n)\right\}\\ &\quad + \pi_n^2\left\{(f_2+1)e^{\eta_\nu(0,\tau_n)} f_3(z,\tau_n) - \Upsilon(z,0,\tau_n)\left(e^{\eta_\nu(0,\tau_n)}(1-f_3(z,\tau_n))-1\right)\right\}\\ &\quad - \pi_n^3\Upsilon(z,0,\tau_n)e^{\eta_\nu(0,\tau_n)} f_3(z,\tau_n)\\ &\quad + o(\pi_n,\tau_n)\\ &= 1 + \pi_n\left(e^{\eta_\nu(0,\tau_n)}\left(1-\left(1-e^{-\eta_\nu(0,\tau_n)}\right)h(z,\tau_n)\right)-1\right) + o(\pi_n,\tau_n)\end{aligned}$$

Soit $(\psi_n)_{n\in\mathbb{N}}$ la suite des fonctions définies par $\psi_n(z,\tau_n) = \psi(z,\tau_n)^n$. Alors :

$$\begin{aligned}\psi_n(z,\tau_n) = \exp\Big\{ &- n\log\Big[1+\pi_n\left(e^{\eta_\nu(0,\tau_n)}\left(1-\left(1-e^{-\eta_\nu(0,\tau_n)}\right)h(z,\tau_n)\right)-1\right)\\ &+ o(\pi_n,\tau_n)\Big]\Big\}\,.\end{aligned}$$

Puis, $\pi_n n$ étant équivalent à $\alpha e^{-\eta_{\nu,\infty}}$:

$$\lim_{n\to+\infty}\psi_n(z,\tau_n) = \exp\left(-m(1-h(z))\right)\,.$$

□

Notons que le théorème 3.2.1 reste vrai que F_μ vérifie $(\mathcal{H})$ ou non. À titre d'exemple, nous considérerons le modèle de Haldane dans la partie suivante.

Le théorème 3.2.1 donne ainsi une forme explicite de la loi asymptotique du décompte final de mutantes, dans le cas où les dynamiques en jeu sont les plus générales possibles. Comme nous le verrons dans la partie 3.3, il est ensuite possible de considérer le cas où il existe une constante $\rho > 0$ vérifiant (3.33). Cependant, il y a quelques points non satisfaisants. Tout d'abord, il n'y a pas d'interprétation rigoureuse des différentes grandeurs, si ce n'est par analogie avec les cas où les durées de vie des cellules sont identiquement distribuées. De plus, nous cherchons à construire de nouveaux estimateurs robustes de la probabilité de mutation π. Puisqu'il est possible d'exprimer les probabilités et donc leur dérivée par rapport à m, nous devrions pouvoir estimer ce dernier par maximisation de la vraisemblance. Or, afin de vérifier empiriquement la robustesse de cet estimateur, nous avons besoin de pouvoir simuler rapidement de nombreux échantillons de grande taille

de décomptes finaux de mutants. L'approche analytique que nous venons de proposer ne permet pas de construire un algorithme de simulation rapide et n'est donc pas entièrement satisfaisante.

3.2.3 Étude « simplifiée » de l'asymptotique

Nous proposons dans cette sous-partie une approche différente du problème. La décomposition d'un modèle de mutation proposée par Hamon et Ycart [31] est généralisée ici, et nous en déduisons un autre théorème de convergence du nombre final de mutantes. Dans cette nouvelle approche, le paramètre de mort γ n'est plus forcément nul. Les hypothèses de modélisation sont les mêmes que dans la sous-partie précédente : la fonction de répartition F_ν vérifie $(\mathcal{H})$, et nous notons η_ν la fonction définie par l'assertion $(\mathcal{H}_4)$, $\eta_{\nu,\infty}$ la limite en l'infini de $\eta_\nu(0,t)$ et λ_ν le taux de division instantané associé. Aucune hypothèse n'est faite sur F_μ pour le moment.

Nous commençons par exprimer la convergence en loi du nombre de mutations, et des instants de mutations, en considérant tout d'abord le cas où γ est nul.

Proposition 3.2.4. *Supposons $\gamma = 0$. Soient $\pi = (\pi_n)_{n\in\mathbb{N}}$ et $\tau = (\tau_n)_{n\in\mathbb{N}}$ deux suites de réels positifs, et une constante $\alpha \geqslant 0$ telles que :*

$$\lim_{n\to+\infty} \pi_n = 0\,, \qquad \lim_{n\to+\infty} \tau_n = +\infty\,, \qquad \lim_{n\to+\infty} \pi_n n \mathrm{e}^{\eta_\nu(0,\tau_n)} = \alpha\,. \tag{3.17}$$

Alors :

$\left(\mathcal{A}_1^{(0)}\right)$ *lorsque n tend vers l'infini, le nombre total de mutations $Z_n(\tau_n)$ converge en loi vers la loi de Poisson de paramètre :*

$$m = \alpha\left(1 - \mathrm{e}^{-\eta_{\nu,\infty}}\right)\,,$$

$\left(\mathcal{A}_2^{(0)}\right)$ *Lorsque n tend vers l'infini, le vecteur $\left(T_1^{(n)}, \ldots, T_k^{(n)}\right)$ d'un nombre fixé k d'instants de mutation dans un intervalle de temps $[0\,;t]$ tend en loi vers la statistique d'ordre d'un échantillon de taille k de la loi de densité :*

$$\frac{\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,t)}}{1-\mathrm{e}^{-\eta_\nu(0,t)}}\mathbb{1}_{u\in[0\,;t]}\,,$$

c'est-à-dire la densité $\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,t)}$ conditionnée à l'intervalle $[0\,;t]$.

En particulier, d'après l'assertion $(\mathcal{A}_2^{(0)})$, la fonction génératrice de la taille à un instant t d'un clone mutant est donnée par la fonction $h(z,t)$ suivante :

$$h(z,t) = \int_0^t \widetilde{\psi}(z,u,t)\frac{\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,t)}}{1-\mathrm{e}^{-\eta_\nu(0,t)}}\mathrm{d}u\,, \tag{3.18}$$

Démonstration de la proposition 3.2.4.

Assertion ($\mathcal{A}_1^{(0)}$) Considérons le processus de branchement binaire **sans mutation** démarrant avec un unique individu et pour lequel une cellule née à un instant s se divise à un temps aléatoire de fonction de répartition $F_\nu(s,\cdot)$. Pour tout $t \geqslant 0$, considérons la suite $(N_n(t))_{n\in\mathbb{N}}$ définie pour tout $n > 0$ par

$$N_n(t) = \sum_{i=1}^{n} N_1^{(i)}(t)\,, \tag{3.19}$$

où les $N_1^{(i)}(t)$ sont des copies *i.i.d.* de $N_1(t)$. Pour tout $n > 0$, $N_n(t)$ correspond donc au nombre total de cellules vivantes au temps t dans n copies du processus. D'après la proposition 3.2.1, l'espérance de $N_1(t)$ est donnée par

$$\mathbb{E}[N_1(t)] = \mathrm{e}^{\eta_\nu(0,t)}\,.$$

Soit $\varepsilon > 0$. Par la Loi des Grands Nombres, il existe pour tout $t > 0$ un entier $n_0(t)$ tel que pour tout $n > n_0(t)$:

$$\mathbb{P}\left[\left|\frac{N_n(t)}{n\mathrm{e}^{\eta_\nu(0,t)}} - 1\right| < \left(1 - \mathrm{e}^{-\eta_\nu(0,t)}\right)\varepsilon\right] = 1\,. \tag{3.20}$$

Comme :

$$\begin{aligned}\frac{N_n(t) - n}{n\left(\mathrm{e}^{\eta_\nu(0,t)} - 1\right)} - 1 &= \frac{1}{1 - \mathrm{e}^{-\eta_\nu(0,t)}}\left(\frac{N_n(t)}{n\mathrm{e}^{\eta_\nu(0,t)}} - \mathrm{e}^{-\eta_\nu(0,t)}\right) - 1\\ &= \frac{1}{1 - \mathrm{e}^{-\eta_\nu(0,t)}}\left(\frac{N_n(t)}{n\mathrm{e}^{\eta_\nu(0,t)}} - 1\right)\,,\end{aligned}$$

il vient que pour tout $n > n_0(t)$:

$$\mathbb{P}\left[\left|\frac{N_n(t) - n}{n\left(\mathrm{e}^{\eta_\nu(0,t)} - 1\right)} - 1\right| < \varepsilon\right] = 1\,. \tag{3.21}$$

Le nombre de divisions cellulaires ayant lieu dans l'intervalle de temps $[0\,;t]$ est donc presque sûrement équivalent à $n\left(\mathrm{e}^{\eta_\nu(0,t)} - 1\right)$. Soit $\vartheta > 0$. Comme τ_n tend vers l'infini, il existe $n_1 \in \mathbb{N}$ tel que pour tout $n > n_1$:

$$\tau_n > \vartheta\,. \tag{3.22}$$

En conséquence, d'après (3.21) et (3.22), pour tout $n > \max(n_0(\vartheta), n_1)$:

$$\mathbb{P}\left[\left|\frac{N_n(\tau_n) - n}{n\left(\mathrm{e}^{\eta_\nu(0,\tau_n)} - 1\right)} - 1\right| < \varepsilon\right] = 1\,. \tag{3.23}$$

Le nombre total de divisions $N_n(\tau_n) - n$ est presque sûrement équivalent à $n\left(\mathrm{e}^{\eta_\nu(0,\tau_n)} - 1\right)$. Les mutantes se développant selon une dynamique différente, le nombre de divisions de cellules normales dans un modèle de mutation n'a pas la même distribution. Nous allons à présent montrer que cette différence reste négligeable. Marquons dans les n copies avec probabilité π_n et de manière indépendante les divisions où ont lieu une mutation. Soit $(X_n(t))_{n\in\mathbb{N}}$ la suite du nombre de divisions marquées au temps $t \geqslant 0$ dans n copies. Comme $\pi_n n \mathrm{e}^{\eta_\nu(0,\tau_n)}$ tend vers α lorsque n tend vers l'infini, $X_n(\tau_n)$ converge en loi vers la loi de Poisson de paramètre $m = \alpha(1 - \mathrm{e}^{-\eta_{\nu,\infty}})$. Le nombre de divisions marquées reste donc borné en probabilité. Considérons à présent que les clones issus des divisions marquées sont des clones mutants. Les divisions marquées dans ces clones seront alors ignorées. Autrement dit, le nombre de mutations $Z_n(\tau_n)$ est inférieur au nombre de divisions marquées $X_n(\tau_n)$ avec probabilité 1. De plus, comme $X_n(\tau_n)$ est borné en probabilité, la différence entre ces deux nombres l'est également. En d'autres termes, pour tout $\varepsilon > 0$, il existe un entier n_2 tel que pour tout $n > n_2$:

$$\mathbb{P}\left[|X_n(\tau_n) - Z_n(\tau_n)| \geqslant \varepsilon\right] = 0\,.$$

Les nombres $X_n(\tau_n)$ et $Z_n(\tau_n)$ sont donc équivalent en probabilité. Ainsi, la loi de $Z_n(\tau_n)$ converge vers la loi de Poisson de paramètre m lorsque n tend vers l'infini. D'où ($\mathcal{A}_1^{(0)}$).

Assertion ($\mathcal{A}_2^{(0)}$) D'après (3.21), le nombre d'occasions de mutation dans un intervalle de temps $[0\,;t]$ est équivalent en probabilité à $n\left(\mathrm{e}^{\eta_\nu(0,t)} - 1\right)$, lui même équivalent à sa partie entière lorsque n tend vers l'infini. Alors pour tout $k \in \mathbb{N}$ et pour tout $t \in \mathbb{R}_+$:

$$\lim_{n\to+\infty} \frac{\mathbb{P}\left[Z_n(t) = k\right]}{\iota_n(k,t)} = 1\,,$$

avec

$$\iota_n(k,t) = \binom{\left\lfloor n\left(\mathrm{e}^{\eta_\nu(0,t)} - 1\right)\right\rfloor}{k} \pi_n^k (1-\pi_n)^{\left\lfloor n\left(\mathrm{e}^{\eta_\nu(0,t)}-1\right)\right\rfloor - k}\,,$$

où $\lfloor x \rfloor$ est l'unique entier relatif vérifiant pour tout $x \in \mathbb{R}$:

$$\lfloor x \rfloor \leqslant x < \lfloor x \rfloor + 1\,,$$

c'est-à-dire la partie entière de x. Alors :

$$\begin{aligned}
\iota_n(k,t) &= \frac{\left\lfloor n\left(\mathrm{e}^{\eta_\nu(0,t)} - 1\right)\right\rfloor \left(\left\lfloor n\left(\mathrm{e}^{\eta_\nu(0,t)} - 1\right)\right\rfloor - 1\right), \ldots, \left(\left\lfloor n\left(\mathrm{e}^{\eta_\nu(0,t)} - 1\right)\right\rfloor - k + 1\right)}{k!} \\
&\qquad \times \left(\frac{\pi_n}{1-\pi_n}\right)^k \exp\left(\left\lfloor n\left(\mathrm{e}^{\eta_\nu(0,t)} - 1\right)\right\rfloor \log(1-\pi_n)\right) \\
&\underset{n\to+\infty}{\sim} \frac{\left(\pi_n \left\lfloor n\left(\mathrm{e}^{\eta_\nu(0,t)} - 1\right)\right\rfloor\right)^k}{k!} \exp\left(-\pi_n \left\lfloor n\left(\mathrm{e}^{\eta_\nu(0,t)} - 1\right)\right\rfloor\right) \\
&\underset{n\to+\infty}{\sim} \frac{\left(\pi_n n\left(\mathrm{e}^{\eta_\nu(0,t)} - 1\right)\right)^k}{k!} \exp\left(-\pi_n n\left(\mathrm{e}^{\eta_\nu(0,t)} - 1\right)\right)\,.
\end{aligned}$$

De plus, par construction de $\{X_n(t)\}_{t\geqslant 0}$, le processus $\{Z_n(t)\}_{t\geqslant 0}$ vérifie les propriétés suivantes :

— $\{Z_n(t)\}_{t\geqslant 0}$ est simple, c'est-à-dire pour tout $t \in \mathbb{R}_+$:

$$\lim_{\Delta t\to 0} \frac{1}{\Delta t}\mathbb{P}\left[Z_n(t+\Delta t) - Z_n(t) > 1\right] = 0\,;$$

— $\{Z_n(t)\}_{t\geqslant 0}$ est à accroissements indépendants ;

— $Z_n(0) = 0$ avec probabilité 1.

Notons à présent ξ_n l'intensité du processus $\{Z_n(t)\}_{t\geqslant 0}$ définie pour tout $t \in \mathbb{R}_+$ par

$$\xi_n(t) = \lim_{\Delta t\to 0} \frac{1}{\Delta t}\mathbb{P}\left[Z_n(t+\Delta t) - Z_n(t) = 1\right]\,.$$

Si $\boldsymbol{t} = (t_1, \ldots, t_k) \in \mathbb{R}_+^k$, nous noterons $\boldsymbol{T^{(n)}} = \boldsymbol{t}$ l'événement $\left(T_1^{(n)} = t_1, \ldots, T_k^{(n)} = t_k\right)$. Pour tout $k \in \mathbb{N}$ et tout $\boldsymbol{t} \in \mathbb{R}_+^k$, la loi de l'instant $T_{k+1}^{(n)}$ de la $(k+1)^{\text{e}}$ mutation conditionnée à $\boldsymbol{T^{(n)}} = \boldsymbol{t}$ est alors donnée pour tout $t \in \mathbb{R}_+$ par (proposition (A.0.2)) :

$$f_{\left(T_{k+1}^{(n)} \,|\, \boldsymbol{T^{(n)}}=\boldsymbol{t}\right)}(t) = \left(\xi_n(t)\exp\left(-\int_{t_k}^{t}\xi_n(u)\mathrm{d}u\right)\right)\mathbb{1}_{0<t_1<\cdots<t_k<t}\,,$$

et la loi jointe des k premiers instants de mutations $\boldsymbol{T^{(n)}}$ par (proposition (A.0.2)) :

$$f_{\boldsymbol{T^{(n)}}}(\boldsymbol{t}) = \left(\prod_{i=1}^{k}\xi_n(t_i)\right)\exp\left(-\sum_{i=1}^{k}\int_{t_{i-1}}^{t_i}\xi_n(u)\mathrm{d}u\right)\mathbb{1}_{0<t_1<\cdots<t_k}\,.$$

Considérons à présent $\{Y_n(t)\}_{t\geqslant 0}$ le processus de Poisson inhomogène d'espérance :

$$m_n(t) = n\pi_n\left(\mathrm{e}^{\eta_\nu(0,t)} - 1\right)\,.$$

Pour tout $k \in \mathbb{N}$, notons $\boldsymbol{S^{(n)}} = \left(S_1^{(n)}, \ldots, S_k^{(n)}\right)$ les instants des k premières occurrences du processus $\{Y_n(t)\}_{t\geqslant 0}$. Alors, pour tout $k \in \mathbb{N}$, tout $t \in \mathbb{R}_+$ et tout $\boldsymbol{t} \in \mathbb{R}_+^k$, nous avons les trois équivalences suivantes :

$$\lim_{n\to\infty}\frac{\mathbb{P}\left[Z_n(t) = k\right]}{\mathbb{P}\left[Y_n(t) = k\right]} = 1\,, \tag{3.24}$$

et :

$$\lim_{n\to\infty}\frac{f_{\boldsymbol{T^{(n)}}}(\boldsymbol{t})}{f_{\boldsymbol{S^{(n)}}}(\boldsymbol{t})} = 1\,, \tag{3.25}$$

et finalement :

$$\lim_{n\to\infty}\frac{f_{\left(T_{k+1}^{(n)} \,|\, \boldsymbol{T^{(n)}}=\boldsymbol{t}\right)}(t)}{f_{\left(S_{k+1}^{(n)} \,|\, \boldsymbol{S^{(n)}}=\boldsymbol{t}\right)}(t)} = 1\,. \tag{3.26}$$

Puis, pour tout $k \in \mathbb{N}$ et tout $t \in \mathbb{R}_+$, la loi du vecteur $\boldsymbol{T^{(n)}}$ conditionnée à $Z_n(t) = k$ s'écrit pour tout $\boldsymbol{t} \in \mathbb{R}_+^k$:

$$\begin{aligned} f_{\left(\boldsymbol{T^{(n)}} \,|\, Z_n(t)=k\right)}(\boldsymbol{t}) &= \frac{f_{\boldsymbol{T^{(n)}}}(\boldsymbol{t})\mathbb{P}\left[Z_n(t) = k \,\middle|\, \boldsymbol{T^{(n)}} = \boldsymbol{t}\right]}{\mathbb{P}\left[Z_n(t) = k\right]} \\ &= \frac{f_{\boldsymbol{T^{(n)}}}(\boldsymbol{t})\mathbb{P}\left[T_{k+1}^{(n)} > t \,\middle|\, \boldsymbol{T^{(n)}} = \boldsymbol{t}\right]}{\mathbb{P}\left[Z_n(t) = k\right]} \end{aligned}$$

Ainsi, nous déduisons à partir des trois équivalences (3.24), (3.25) et (3.26) que pour tout $k \in \mathbb{N}$:

$$\lim_{n\to\infty} \frac{f_{\left(\boldsymbol{T^{(n)}} \,|\, Z_n(t)=k\right)}(\boldsymbol{t})}{f_{\left(\boldsymbol{S^{(n)}} \,|\, Y_n(t)=k\right)}(\boldsymbol{t})} = 1\,.$$

Or, sachant que $Y_n(t) = k$, le vecteur $\boldsymbol{S^{(n)}}$ a même loi que k variables indépendantes triées par ordre croissant, où chacune de ces variables suit la loi suivante (voir proposition A.0.5) :

$$\begin{aligned} \frac{m_n'(u)}{m_n(t)}\mathbb{1}_{u\in[0\,;\,t]} &= \frac{\lambda_\nu(u)\mathrm{e}^{\eta_\nu(0,u)}}{\mathrm{e}^{\eta_\nu(0,t)-1}}\mathbb{1}_{u\in[0\,;\,t]} \\ &= \frac{\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,t)}}{1-\mathrm{e}^{-\eta_\nu(0,t)}}\mathbb{1}_{u\in[0\,;\,t]}\,. \end{aligned}$$

D'où $(\mathcal{A}_2^{(0)})$. □

Remarquons que l'assertion $(\mathcal{A}_2^{(0)})$ concerne les instants de mutations, tandis que les résultats les plus proches dans les modèles homogènes sont dédiés aux durées de développement des clones mutants. En l'occurrence, les théorèmes 2.1 et 3.1 de Kuzcek [52] permettent de montrer que la loi jointe d'un nombre fixé k de durées de développement de clones mutants converge en loi vers le produit de k copies indépendantes de loi exponentielle. Précisons que dans le cadre de cette thèse, nous n'avons pas cherché à généraliser les résultats de Kuzcek [52, 53], mais qu'il s'agit d'une des principales perspectives.

La conséquence directe de la proposition 3.2.5 est l'écriture de la loi asymptotique du décompte final de mutantes.

Théorème 3.2.2. *Supposons que $\gamma = 0$. Soient $\pi = (\pi_n)_{n\in\mathbb{N}}$ et $\tau = (\tau_n)_{n\in\mathbb{N}}$ deux suites de réels positifs, et une constante $\alpha \geqslant 0$ satisfaisant le cadre asymptotique (3.17).*

Lorsque n tend vers l'infini, la fonction génératrice du nombre de mutantes au temps τ_n dans les clones issus de n cellules normales nées au temps 0 converge vers la fonction génératrice (3.15) où $h(z)$ est définie par

$$\begin{aligned} h(z) &= \lim_{n\to+\infty} h(z,\tau_n) \\ &= \frac{1}{1-\mathrm{e}^{-\eta_{\nu,\infty}}} \lim_{t\to+\infty} \int_0^t \widetilde{\psi}(z,u,t)\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,t)}\mathrm{d}u\,, \end{aligned} \tag{3.27}$$

Le théorème 3.2.1 est ainsi retrouvé, sans la condition (3.14). Une première conséquence de ce résultat est l'interprétation de n'importe quel modèle de mutation comme la composition des trois ingrédients suivants :

1. le nombre aléatoire de mutations apparues durant la croissance de la population. Par la Loi des Petits Nombres, le nombre de mutations suit asymptotiquement la loi de Poisson d'espérance $m = \alpha\,(1 - \mathrm{e}^{-\eta_{\nu,\infty}})$;
2. chaque mutation apparaît à un instant aléatoire. L'instant d'une mutation quelconque suit asymptotiquement la loi de densité $\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,t)}$ tronquée sur l'intervalle de temps $[0\,;t]$;
3. un clone mutant démarré à un instant s se développe selon un certain processus. La taille de ce clone à un instant t est donnée par la fonction génératrice $\widetilde{\psi}(z,s,t)$.

Nous nous intéressons à présent à la simulation des instants de mutations. Commençons par considérer un processus de Poisson inhomogène $\{Y(t)\}_{t\geqslant 0}$ d'espérance $\Lambda(t)$. Notons $\left(\widetilde{T}_i\right)_{i\in\mathbb{N}}$ la suite croissante des instants d'occurrence de $\{Y(t)\}_{t\geqslant 0}$. La loi conditionnée à $Y(t) = k$ du vecteur $\left(\frac{\Lambda(\widetilde{T}_1)}{\Lambda(t)},\dots,\frac{\Lambda(\widetilde{T}_k)}{\Lambda(t)}\right)$ est la même que la statistique d'ordre d'un échantillon de taille k de la loi uniforme sur $[0\,;1]$ (voir proposition A.0.6). L'assertion $(\mathcal{A}_2^{(0)})$ permet alors de déterminer la loi asymptotique conditionnelle des instants de mutations.

Corollaire 3.2.2. *Sachant que $Z_n(t) = k$, le vecteur $\left(\frac{\mathrm{e}^{\eta_\nu(0,T_1)}-1}{\mathrm{e}^{\eta_\nu(0,t)}-1},\dots,\frac{\mathrm{e}^{\eta_\nu(0,T_k)}-1}{\mathrm{e}^{\eta_\nu(0,t)}-1}\right)$ converge en loi lorsque n tend vers l'infini vers la loi de la statistique d'ordre d'un échantillon de taille k de la loi uniforme sur $[0\,;1]$.*

Ainsi, un vecteur de k instants croissants de mutation $(T_i)_{i=1,\dots,k}$ dans un intervalle de temps $[0\,;t]$ peut être simulé par la démarche suivante :

1. simuler k variables uniformes $U_1,\dots,U_k$ et les trier par ordre croissant ;
2. appliquer pour tout $i = 1,\dots,k$:

$$T_i = \eta_{\nu,0}^{-1}\left[\log\left(U_i\left(\mathrm{e}^{\eta_\nu(0,t)} - 1\right) + 1\right)\right]\,, \tag{3.28}$$

où pour tout s, $\eta_{\nu,s}^{-1}$ est la fonction vérifiant :

$$\eta_\nu\left(s,\eta_{\nu,s}^{-1}(u)\right) = u \quad \text{et} \quad \eta_{\nu,s}^{-1}(\eta_\nu(s,t)) = t\,.$$

Supposons à présent que $\gamma \geqslant 0$. Il est possible de généraliser la proposition 3.2.4.

Proposition 3.2.5. *Soient $\pi = (\pi_n)_{n\in\mathbb{N}}$ et $\tau = (\tau_n)_{n\in\mathbb{N}}$ deux suites de réels positifs, et une constante $\alpha \geqslant 0$ telles que :*

$$\lim_{n\to+\infty}\pi_n = 0\,,\quad \lim_{n\to+\infty}\tau_n = +\infty\,,\quad \lim_{n\to+\infty}\pi_n n\omega(\tau_n)\mathrm{e}^{\eta_\nu^*(0,\tau_n)} = \alpha \tag{3.29}$$

où

$$\eta_\nu^*(s,t) = (1-2\gamma)\eta_\nu(s,t)\,,$$

et

$$\omega(t) = \frac{1-2\gamma}{1-\gamma-\gamma \mathrm{e}^{-\eta_\nu^*(0,t)}}\,,$$

est la probabilité de non-extinction avant l'instant t d'un clone issu d'une cellule normale née au temps 0. Alors :

$\left(\mathcal{A}_1^{(\gamma)}\right)$ *lorsque n tend vers l'infini, le nombre total de mutations $Z_n(\tau_n)$ converge en loi vers la loi de Poisson de paramètre :*

$$m = \alpha\left(1 - \left(\omega_\infty \mathrm{e}^{\eta_{\nu,\infty}^*}\right)^{-1}\right)\,,$$

où

$$\eta_{\nu,\infty}^* = \lim_{t\to+\infty} \eta_\nu^*(0,t)\,, \qquad \textit{et} \qquad \omega_\infty = \lim_{t\to+\infty} \omega(t)\,;$$

$\left(\mathcal{A}_2^{(\gamma)}\right)$ *lorsque n tend vers l'infini, le vecteur $\left(T_1^{(n)},\dots,T_k^{(n)}\right)$ d'un nombre fixé k d'instants de mutation dans un intervalle de temps $[0\,;t]$ tend en loi vers la statistique d'ordre d'un échantillon de taille k de la loi de densité :*

$$\frac{\left(\omega'(u) + \lambda_\nu(u)(1-2\gamma)\omega(u)\right)\mathrm{e}^{-\eta_\nu^*(u,t)}}{\omega(t) - \mathrm{e}^{-\eta_\nu^*(0,t)}} \mathbb{1}_{u\in[0\,;t]}\,.$$

Démonstration de la proposition 3.2.5.

Assertion ($\mathcal{A}_1^{(\gamma)}$) Considérons à nouveau le processus sans mutation : un unique individu est présent au temps 0, et une cellule née au temps s meurt (avec probabilité γ) ou se divise (avec probabilité $1-\gamma$) à un instant aléatoire de fonction de répartition $F_\nu(s,\cdot)$. Pour tout $t \geqslant 0$, considérons la suite $(N_n(t))_{n\in\mathbb{N}}$ du nombre total de cellules vivantes au temps dans n copies de ce processus, définie pour tout $n > 0$ par (3.19). Chacune des n copies peut avec probabilité $1-\omega(t)$ s'éteindre avant l'instant t. Par le même raisonnement que dans la démonstration de ($\mathcal{A}_1^{(0)}$), le nombre de divisions dans un intervalle de temps $[0\,;t]$ dans la proportion $\omega(t)$ de copies survivantes est asymptotiquement équivalent à $n\left(\omega(t)\mathrm{e}^{\eta_\nu^*(0,t)} - 1\right)$. De plus, le nombre de divisions cellulaires ayant lieu dans la proportion $1-\omega(t)$ de clones qui s'éteignent est borné et peut être négligé face au nombre de divisions ayant lieu dans les clones survivants. Le nombre total de divisions dans les n copies est donc presque sûrement équivalent à $n\left(\omega(\tau_n)\mathrm{e}^{\eta_\nu^*(0,\tau_n)} - 1\right)$.

Nous marquons avec probabilité π_n chaque division cellulaire, et considérons à nouveau la suite $(X_n(t))_{n\in\mathbb{N}}$ du nombre de divisions marquées dans l'intervalle de temps $[0\,;t]$ dans les n copies. Par un raisonnement analogue à la preuve de ($\mathcal{A}_1^{(0)}$), le nombre de mutations $Z_n(\tau_n)$ converge en loi vers la loi de Poisson de paramètre $m = \alpha\left(1 - \left(\omega_\infty \mathrm{e}^{\eta_{\nu,\infty}^*}\right)^{-1}\right)$.

Assertion ($\mathcal{A}_2^{(\gamma)}$) Le nombre d'occasions de mutations dans un intervalle de temps $[0\,;t]$ est équivalent en probabilité à $n\left(\omega(t)\mathrm{e}^{\eta_\nu^*(0,t)}-1\right)$. Par conséquent :

$$\lim_{n\to+\infty}\frac{\mathbb{P}\left[Z_n(t)=k\right]}{\iota_n(k,t)}=1\,,$$

où

$$\iota_n(k,t)=\binom{\left\lfloor n\left(\omega(t)\mathrm{e}^{\eta_\nu^*(0,t)}-1\right)\right\rfloor}{k}\pi_n^k(1-\pi_n)^{\left\lfloor n\left(\omega(t)\mathrm{e}^{\eta_\nu^*(0,t)}-1\right)\right\rfloor-k}\,.$$

Notons ensuite que :

$$\begin{aligned}\iota_n(k,t)&\underset{n\to+\infty}{\sim}\frac{\left(\pi_n\left\lfloor n\left(\omega(t)\mathrm{e}^{\eta_\nu^*(0,t)}-1\right)\right\rfloor\right)^k}{k!}\exp\left(-\pi_n\left\lfloor n\left(\omega(t)\mathrm{e}^{\eta_\nu^*(0,t)}-1\right)\right\rfloor\right)\\&\underset{n\to+\infty}{\sim}\frac{\left(\pi_n n\left(\omega(t)\mathrm{e}^{\eta_\nu^*(0,t)}-1\right)\right)^k}{k!}\exp\left(-\pi_n n\left(\omega(t)\mathrm{e}^{\eta_\nu^*(0,t)}-1\right)\right)\,.\end{aligned}$$

Pour tout $t\geqslant 0$, la loi de $Z_n(t)$ est donc équivalente lorsque n tend vers l'infini à la loi de Poisson de paramètre :

$$m_n(t)=n\pi_n\left(\omega(t)\mathrm{e}^{\eta_\nu^*(0,t)}-1\right)\,.$$

Puis par un raisonnement analogue à la démonstration de ($\mathcal{A}_2^{(0)}$), nous en déduisons l'assertion ($\mathcal{A}_2^{(\gamma)}$). □

Ainsi, si l'expression de $\widetilde{\psi}$ est connue, la loi asymptotique du nombre total de mutantes peut être explicitée via sa fonction génératrice ψ.

Théorème 3.2.3. *Soient $\pi=(\pi_n)_{n\in\mathbb{N}}$ et $\tau=(\tau_n)_{n\in\mathbb{N}}$ deux suites de réels positifs, et une constante $\alpha\geqslant 0$ satisfaisant le cadre asymptotique (3.29).*

Lorsque n tend vers l'infini, la fonction génératrice du nombre de mutantes au temps τ_n dans les clones issus de n cellules normales nées au temps 0 converge vers la fonction génératrice (3.15) avec :

$$m=\alpha\left(1-\left(\omega_\infty\mathrm{e}^{\eta_{\nu,\infty}^*}\right)^{-1}\right)\,,$$

et

$$h(z)=\frac{1}{\omega_\infty-\mathrm{e}^{-\eta_{\nu,\infty}^*}}\lim_{t\to+\infty}\int_0^t\widetilde{\psi}(z,u,t)\left(\omega'(u)+\lambda_\nu(u)(1-2\gamma)\omega(u)\right)\mathrm{e}^{-\eta_\nu^*(u,t)}\mathrm{d}u\,.\qquad(3.30)$$

Tout comme dans le cas où $\gamma=0$, il est ainsi possible de décomposer n'importe quel modèle de mutation prenant en compte les morts des cellules non-mutantes :

1. le nombre aléatoire de mutations apparues durant la croissance de la population. Par la Loi des Petits Nombres, le nombre de mutations suit asymptotiquement la loi de Poisson d'espérance $m=\alpha\left(1-\left(\omega_\infty\mathrm{e}^{\eta_{\nu,\infty}^*}\right)^{-1}\right)$;

2. chaque mutation apparaît à un instant aléatoire. L'instant d'une mutation quelconque suit asymptotiquement la loi de densité $\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,t)}$ tronquée sur l'intervalle de temps $[0\,;t]$;
3. un clone mutant démarré à un instant s se développe selon un certain processus. La taille de ce clone à un instant t est donnée par la fonction génératrice $\widetilde{\psi}(z,s,t)$.

3.3 Cas particuliers

Nous commençons par étudier dans cette dernière partie deux cas particuliers des modèles décrits dans la partie précédente. Dans un premier temps, nous étendons dans la sous-partie 3.3.1 la formulation de Haldane [81] au cas où $\delta > 0$ via les outils exposés dans la partie précédente. Nous étudions ensuite dans la sous-partie 3.3.2 le cas particulier où les fonctions de répartition F_ν et F_μ vérifient $(\mathcal{H})$ telles que les taux instantanés associés λ_ν et λ_μ sont proportionnels. Nous généralisons ainsi la notion de fitness. Nous nous intéresserons finalement à la prise en compte de la dilution dans la dernière sous-partie.

3.3.1 Modèle de Haldane

Dans ce modèle, les durées de vie des cellules normales sont exponentielles et *i.i.d.* tandis que celles des mutantes sont égales à une constante a. La fonction de répartition $F_\nu(s,\cdot)$ est donc définie pour tout $t \geqslant s$ par

$$F_\nu(s,t) = \left(1 - \mathrm{e}^{-\lambda(t-s)}\right) \mathbb{1}_{t\geqslant s}\,,$$

où λ est une constante positive. Le taux de division cumulé η_ν est ainsi défini par

$$\eta_\nu(s,t) = \lambda(t-s)\,,$$

et $\lambda_\nu(u) = \lambda$ pour tout $u \in \mathbb{R}_+$. La fonction de répartition $F_\mu(s,\cdot)$ est définie pour tout $t \geqslant s$ par

$$F_\mu(s,t) = \begin{cases} 1 & \text{si} \quad t \geqslant s+a\,, \\ 0 & \text{sinon}\,. \end{cases}$$

Remarquons que la condition $(\mathcal{H}_2)$ n'est pas satisfaite, et la fonction de répartition F_μ ne vérifie donc pas $(\mathcal{H})$. Nous pouvons cependant identifier la fonction génératrice $\widetilde{\psi}$. Considérons une mutante née à un instant s, et notons $b_i(z)$ la fonction génératrice de la taille du clone issu de cette cellule dans l'intervalle de temps $[s+ia\,;s+(i+1)a[$. Alors $b_0(z) = z$, et pour tout $i > 0$,

$$b_i(z) = \delta + (1-\delta)\left(b_{i-1}(z)\right)^2\,.$$

Et donc la fonction génératrice de la taille au temps t d'un clone démarré au temps s est donnée par

$$\widetilde{\psi}(z,s,t) = \sum_{i\geqslant 0} b_i(z)\mathbb{1}_{t\in[s+ia\,;\,s+(i+1)a[}\,.$$

La fonction génératrice (3.18) est alors donnée par

$$\begin{aligned} h(z,t) &= \frac{1}{1-\mathrm{e}^{-\lambda_\nu t}}\int_0^t \widetilde{\psi}(z,u,t)\lambda_\nu \mathrm{e}^{-\lambda_\nu(t-u)}\mathrm{d}u \\ &= \frac{1}{1-\mathrm{e}^{-\lambda_\nu t}}\sum_{i\geqslant 0} b_i(z)\int_0^t \mathbb{1}_{w\in[ia\,;\,(i+1)a[}\lambda_\nu \mathrm{e}^{-\lambda_\nu w}\mathrm{d}w \\ &= \frac{1}{1-\mathrm{e}^{-\lambda_\nu t}}\sum_{i\geqslant 0} b_i(z)\left(\left(\mathrm{e}^{-\lambda_\nu ia}-\mathrm{e}^{-\lambda_\nu t}\right)\mathbb{1}_{t\in[ia\,;\,(i+1)a[}\right. \\ &\qquad \left.+\left(\mathrm{e}^{-\lambda_\nu ia}-\mathrm{e}^{-\lambda_\nu (i+1)a}\right)\mathbb{1}_{t\in[(i+1)a\,;\,+\infty[}\right)\,. \end{aligned}$$

D'où la limite lorsque t tend vers l'infini de $h(z,t)$:

$$h(z) = \sum_{i\geqslant 0} b_i(z)\mathrm{e}^{-\lambda_\nu ia}\left(1-\mathrm{e}^{-\lambda_\nu a}\right)\,.$$

Observons que pour $\delta = 0$ et $a = \log(2)$, nous retrouvons l'équation (5) de Ycart [97]. Puis nous pouvons appliquer le théorème 3.2.2 afin d'expliciter la fonction génératrice asymptotique du décompte final de mutantes. Nous allons à présent expliciter les probabilités $(p_k)_{k\in\mathbb{N}}$ du nombre de mutantes. Considérons à nouveau une mutante née à un instant s, et notons $\left(r_k^{(i)}\right)_{k\in\mathbb{N}}$ les probabilités de la taille du clone issu de cette cellule dans l'intervalle de temps $[s+ia\,;s+(i+1)a[$. En d'autres termes :

$$b_i(z) = \sum_{k\geqslant 0} r_k^{(i)} z^k\,,$$

pour tout z dans le disque unité. Puis :

$$\begin{aligned} h(z) &= \sum_{i\geqslant 0} \mathrm{e}^{-ia\lambda_\nu}(1-\mathrm{e}^{-a\lambda_\nu})\sum_{k\geqslant 0} r_k^{(i)} z^k \\ &= \sum_{k\geqslant 0} z^k \sum_{i\geqslant 0} \mathrm{e}^{-ia\lambda_\nu}(1-\mathrm{e}^{-a\lambda_\nu})r_k^{(i)}\,. \end{aligned}$$

Nous en déduisons l'expression des probabilités $(q_k)_{k\geqslant 0}$ associées à fonction génératrice h :

$$q_k = \sum_{i\geqslant 0} \mathrm{e}^{-ia\lambda_\nu}(1-\mathrm{e}^{-a\lambda_\nu})r_k^{(i)}\,.$$

Ainsi, si les probabilités $\left(r_k^{(i)}\right)_{k\geqslant 0}$ peuvent être explicitées pour tout $i \geqslant 0$, il est possible d'en déduire les probabilités $(q_k)_{k\geqslant 0}$. En pratique, il est possible d'identifier pour tout $i, k \geqslant 0$ les $r_k^{(i)}$ par transformée de Fourier rapide. L'identification de la suite des probabilités $(p_k)_{n\in\mathbb{N}}$ en fonction de la suite des probabilités $(q_k)_{n\in\mathbb{N}}$ s'effectue ensuite en appliquant l'algorithme de Embrechts et Hawkes [21] (voir (2.32)), c'est-à-dire :

$$p_0 = \mathrm{e}^{-m(1-q_0)}\,, \tag{3.31}$$

et pour tout $k > 0$:

$$p_k = \frac{m}{k}\sum_{i=1}^{k} i q_i p_{k-i}\,. \tag{3.32}$$

3.3.2 Taux instantanés proportionnels

Dans cette partie, nous supposons que $\gamma = 0$ et que la fonction F_μ vérifie $(\mathcal{H})$. Nous notons η_μ la fonction définie par $(\mathcal{H}_4)$, $\eta_{\mu,\infty}$ la limite en l'infini de $\eta_\mu(0,t)$ et λ_μ le taux de division instantané associé. Nous considérons ici le cas particulier où il existe une relation de proportionnalité entre les taux λ_ν et λ_μ de la forme suivante :

$$\lambda_\nu(u) = (1-2\delta)\rho\lambda_\mu(u)\,, \tag{3.33}$$

pour tout $u \in \mathbb{R}_+$, avec $\rho > 0$. L'hypothèse de taux instantanés proportionnels a du sens : en analyse de survie, cette classe de modèle est connue sous le nom de « modèle à risque proportionnel » ou « régression de Cox » [15]. Cette hypothèse généralise la notion de fitness définie dans les modèles de mutations homogènes, et nous conserverons cette appellation. Par application du théorème 3.2.2, et par le changement de variable $v = \mathrm{e}^{-\eta_\mu^*(u,t)}$, nous obtenons le résultat suivant.

Théorème 3.3.1. *Supposons que $\gamma = 0$, et qu'il existe une constante $\rho > 0$ telle que (3.33) soit vérifiée. Soient $\pi = (\pi_n)_{n\in\mathbb{N}}$ et $\tau = (\tau_n)_{n\in\mathbb{N}}$ deux suites de réels positifs, et une constante $\alpha \geqslant 0$ satisfaisant le cadre asymptotique (3.17).*

Lorsque n tend vers l'infini, la fonction génératrice du nombre de mutantes au temps τ_n dans les clones issus de n cellules normales nées au temps 0 converge vers la fonction génératrice (3.15) avec :

$$h(z) = \frac{1}{1-\mathrm{e}^{-\rho\eta_{\mu,\infty}^*}}\int_{\mathrm{e}^{-\eta_{\mu,\infty}^*}}^{1} \frac{\delta(1-z)+v((1-\delta)z-\delta)}{(1-\delta)(1-z)+v((1-\delta)z-\delta)}\rho v^{\rho-1}\mathrm{d}v\,, \tag{3.34}$$

où pour tout $(s,t) \in \mathbb{R}_+^2$:

$$\eta_\mu^*(s,t) = (1-2\delta)\eta_\mu(s,t)\,,$$

et

$$\eta_{\mu,\infty}^* = (1-2\delta)\eta_{\mu,\infty}\,.$$

Comme cas particulier de ce résultat, la loi de Luria-Delbrück avec morts cellulaires [98] est retrouvée lorsque la limite $\eta_{\mu,\infty}$ est infinie.

Corollaire 3.3.1. *Supposons que les hypothèses du théorème 3.3.1 sont respectées, et que $\eta_{\mu,\infty} = +\infty$. Lorsque n tend vers l'infini, la fonction génératrice du nombre de mutantes au temps τ_n dans les clones issus de n cellules normales nées au temps 0 converge vers la fonction génératrice (3.15) où $h(z)$ est définie par (2.17).*

En d'autres termes, la loi de Luria-Delbrück avec morts cellulaires peut être étendue au cas où les instants de divisions des cellules normales et mutantes ne sont pas distribués exponentiellement, tant que les fonctions de répartition $F_\nu(s,\cdot)$ et $F_\mu(s,\cdot)$ vérifient $(\mathcal{H})$ et sont issues de mesures de probabilité sur $\mathbb{R}_+$.

Nous allons à présent donner les conséquences probabilistes de ce résultat, à savoir la forme explicite des probabilités du nombre final de mutantes, ainsi qu'un algorithme de simulation rapide.

La fonction génératrice (3.18) peut s'écrire sous la forme :

$$h(z,t) = \sum_{k\geqslant 0} q_k(t) z^k \,,$$

où, pour tout $t \geqslant 0$, la suite $(q_k(t))_{k\in\mathbb{N}}$ est ici définie pour tout $k \geqslant 0$ par

$$q_k(t) = \int_0^t \widetilde{p}_k(u,t) \frac{\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,t)}}{1-\mathrm{e}^{-\eta_\nu(0,t)}}\mathrm{d}u \,,$$

et la suite des $(\widetilde{p}_k(s,t))_{k\in\mathbb{N}}$ est la suite des probabilités de la variable $\widetilde{M}(s,t)$ (voir proposition 3.2.2). Alors :

$$q_0(t) = \int_0^t \frac{\delta\left(1-\mathrm{e}^{-\eta_\mu^*(u,t)}\right)}{1-\delta-\delta\mathrm{e}^{-\eta_\mu^*(u,t)}}\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,t)}\mathrm{d}u \,,$$

et

$$q_k(t) = \int_0^t \left(1 - \frac{\delta\left(1-\mathrm{e}^{-\eta_\mu^*(u,t)}\right)}{1-\delta-\delta\mathrm{e}^{-\eta_\mu^*(u,t)}}\right) P(u,t)(1-P(u,t))^{k-1}\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,t)}\mathrm{d}u \,,$$

où :

$$P(s,t) = \frac{(1-2\delta)\mathrm{e}^{-\eta_\mu^*(u,t)}}{1-\delta-\delta\mathrm{e}^{-\eta_\mu^*(u,t)}} \,.$$

En particulier, si nous revenons à présent dans le cadre du théorème 3.3.1, nous obtenons :

$$q_0(t) = \frac{1}{1-\mathrm{e}^{-\rho\eta_\mu^*(0,t)}} \int_{\mathrm{e}^{-\eta_\mu^*(0,t)}}^1 \frac{\delta-\delta v}{1-\delta-\delta v}\rho v^{\rho-1}\mathrm{d}v \,,$$

et pour tout $k > 0$:

$$q_k(t) = \frac{(1-2\delta)^2(1-\delta)^{k-1}}{1-\mathrm{e}^{-\rho\eta_\mu^*(0,t)}} \int_{\mathrm{e}^{-\eta_\mu^*(0,t)}}^1 \frac{(1-v)^{k-1}}{(1-\delta-\delta v)^{k+1}} \rho v^\rho \mathrm{d}v \,.$$

Ainsi, la fonction génératrice (3.34) s'écrit :

$$h(z) = \sum_{k\geqslant 0} q_k z^k \,,$$

où pour tout $k \geqslant 0$:

$$q_k = \lim_{t\to+\infty} q_k(t) \,.$$

En d'autres termes :

$$q_0 = \frac{1}{1-\mathrm{e}^{-\rho\eta_{\mu,\infty}^*}} \int_{\mathrm{e}^{-\eta_{\mu,\infty}^*}}^1 \frac{\delta-\delta v}{1-\delta-\delta v} \rho v^{\rho-1} \mathrm{d}v \,, \tag{3.35}$$

et pour tout $k > 0$:

$$q_k = \frac{(1-2\delta)^2(1-\delta)^{k-1}}{1-\mathrm{e}^{-\rho\eta_{\mu,\infty}^*}} \int_{\mathrm{e}^{-\eta_{\mu,\infty}^*}}^1 \frac{(1-v)^{k-1}}{(1-\delta-\delta v)^{k+1}} \rho v^\rho \mathrm{d}v \,. \tag{3.36}$$

L'identification de la suite des probabilités $(p_k)_{n\in\mathbb{N}}$ en fonction de la suite des probabilités $(q_k)_{n\in\mathbb{N}}$ s'effectue ensuite par l'algorithme rappelé dans la sous-partie précédente (voir eqs. (3.31) et (3.32)). Bien entendu, les probabilités des modèles $LD(m,\rho,\delta)$ (voir [98]) sont retrouvées lorsque $\eta_{\mu,\infty}$ est infini.

Remarquons que si $\delta = 0$, il vient que pour tout $k > 0$:

$$\begin{aligned} q_k &= \frac{1}{1-\mathrm{e}^{-\rho\eta_{\mu,\infty}}} \int_{\mathrm{e}^{-\eta_{\mu,\infty}}}^1 (1-v)^{k-1} \rho v^\rho \mathrm{d}v \\ &= \frac{\rho}{1-\mathrm{e}^{-\eta_{\mu,\infty}}} \left(B(\rho+1,k) - B_{\mathrm{e}^{-\eta_{\mu,\infty}}}(\rho+1,k) \right) \,, \end{aligned}$$

où $B_w(x,y)$ est la fonction Beta incomplète définie pour tout $x \in]0\,;1[$ par

$$B_w(x,y) = \int_0^w v^{x-1}(1-v)^{y-1} \mathrm{d}v \,,$$

et $B(x,y)$ est la fonction Beta complète, c'est-à-dire $B(x,y) = B_1(x,y)$. Nous retrouvons donc les probabilités des modèles $LD(m,\rho,0)$ [31] lorsque $\eta_{\mu,\infty}$ est infini.

Reprenons à présent l'algorithme de simulation présenté dans la sous-section 3.2.3. Il suffit de réécrire (3.28) en prenant en compte (3.33) :

$$T_i = \eta_{\mu,0}^{-1} \left[\frac{1}{\rho(1-2\delta)} \log\left(U_i \left(\mathrm{e}^{\rho\eta_\mu^*(0,t)} - 1 \right) + 1 \right) \right] \,, \tag{3.37}$$

où pour tout s, $\eta_{\mu,s}^{-1}$ est la fonction vérifiant :

$$\eta_\mu\left(s, \eta_{\mu,s}^{-1}(u)\right) = u \quad \text{et} \quad \eta_{\mu,s}^{-1}(\eta_\mu(s,t)) = t\,.$$

Notons que cette fonction existe bien grâce à ($\mathcal{H}_4$). Par exemple, dans le cas où F_μ est définie par (3.6), la fonction η_μ est alors donnée par

$$\eta_\mu(s,t) = \log\left(\frac{f(t)}{f(s)}\right)\mathbb{1}_{t\geqslant s}\,,$$

et pour tout $u \in \mathbb{R}_+$:

$$\eta_{\mu,s}^{-1}(u) = f^{-1}\left(f(s)\mathrm{e}^u\right)\,.$$

Ainsi, le nombre final de mutations peut être simulé à partir de l'algorithme ci-dessous :

1. simuler le nombre k de mutations selon la loi de Poisson de paramètre m ;
2. pour chacune des k mutations, simuler le développement du clone correspondant :
   (a) simuler l'instant de mutation s (appliquer (3.37) avec $t = +\infty$) ;
   (b) calculer $p_0(s,\infty)$ et $P(s,\infty)$ (Proposition 3.2.2), et simuler la taille finale du clone :
   — avec probabilité $p_0(s,\infty)$, le clone s'éteint (taille nulle) ;
   — avec probabilité $1 - p_0(s,\infty)$, la taille finale du clone est simulée selon la loi géométrique de paramètre $P(s,\infty)$.
3. sommer les k tailles finales des clones.

Notons que cet algorithme ne peut pas être appliqué si $\eta_{\mu,\infty} = +\infty$: le calcul de (3.37) nécessiterait dans ce cas d'identifier $\eta_{\mu,0}^{-1}(+\infty)$, qui n'est alors plus défini. Cependant, par le corollaire 3.3.1, la loi des modèles $LD(m,\rho,\delta)$ est retrouvée lorsque $\eta_{\mu,\infty}$ est infini. La simulation de nombres finaux de mutantes peut alors être effectuée via l'algorithme exposé par Ycart [98].

3.3.3 Prise en compte de la dilution

Supposons que seule une proportion ζ de la population totale est observée. Comme nous l'avons rappelé dans la section 2.2.2, le nombre final $M^{(\zeta)}$ de cellules mutantes présentes dans la sous-population obtenue après dilution suit une loi binomiale de paramètres M et ζ. Ainsi, la fonction génératrice $\psi^{(\zeta)}$ de $M^{(\zeta)}$ est donnée pour tout $z \in U$ par

$$\begin{aligned}\psi^{(\zeta)}(z) = \mathbb{E}\left[z^{M^{(\zeta)}}\right] &= \mathbb{E}\left[\mathbb{E}\left[z^{M^{(\zeta)}} \mid M\right]\right] \\ &= \mathbb{E}\left[(1-\zeta+\zeta z)^M\right] \\ &= \psi(1-\zeta+\zeta z)\,.\end{aligned}$$

De manière générale (c'est-à-dire dans le contexte du théorème 3.2.2), nous pouvons écrire pour tout $z \in U$:

$$\psi^{(\zeta)}(z) = \exp\left(-m\left(1 - h^{(\zeta)}(z)\right)\right), \tag{3.38}$$

$$h^{(\zeta)}(z) = \frac{1}{1 - \mathrm{e}^{-\eta_{\nu,\infty}}} \lim_{t\to\infty} \int_0^t \widetilde{\psi}^{(\zeta)}(z, u, t)\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,t)}\mathrm{d}u\,, \tag{3.39}$$

et pour tout $(s,t) \in \mathbb{R}_+^2$:

$$\widetilde{\psi}^{(\zeta)}(z, s, t) = \widetilde{\psi}(1 - \zeta + \zeta z, s, t)\,.$$

Considérons à présent le cas particulier du théorème 3.3.1. Alors :

$$\widetilde{\psi}^{(\zeta)}(z, s, t) = \frac{\delta(1 - (1 - \zeta + \zeta z)) + \mathrm{e}^{-\eta_\mu^*(s,t)}((1-\delta)(1 - \zeta + \zeta z) - \delta)}{(1-\delta)(1 - (1 - \zeta + \zeta z)) + \mathrm{e}^{-\eta_\mu^*(s,t)}((1-\delta)(1 - \zeta + \zeta z) - \delta)}\,.$$

Notons $\left(\widetilde{p}_k^{(\zeta)}(s,t)\right)_{k\in\mathbb{N}}$ les probabilités associées à la fonction génératrice $\widetilde{\psi}^{(\zeta)}(\cdot, s, t)$. Il est possible de les expliciter en exprimant $\widetilde{\psi}^{(\zeta)}(\cdot, s, t)$ comme une fraction rationnelle en z :

$$\begin{aligned}\widetilde{\psi}^{(\zeta)}(z, s, t) &= \frac{\delta\left(1 - \mathrm{e}^{-\eta_\mu^*(s,t)}\right) - (1-\zeta)\left(\delta - \mathrm{e}^{-\eta_\mu^*(s,t)}(1-\delta)\right) - z\zeta\left(\delta - \mathrm{e}^{-\eta_\mu^*(s,t)}(1-\delta)\right)}{1 - \delta - \delta\mathrm{e}^{-\eta_\mu^*(s,t)} - (1-\zeta)(1-\delta)\left(1 - \mathrm{e}^{-\eta_\mu^*(s,t)}\right) - z\zeta(1-\delta)\left(1 - \mathrm{e}^{-\eta_\mu^*(s,t)}\right)} \\ &= \frac{n_0(s,t) + z n_1(s,t)}{d_0(s,t) + z d_1(s,t)}\,,\end{aligned}$$

où

$$\begin{aligned} n_0(s,t) &= \delta\left(1 - \mathrm{e}^{-\eta_\mu^*(s,t)}\right) - (1-\zeta)\left(\delta - \mathrm{e}^{-\eta_\mu^*(s,t)}(1-\delta)\right), \\ n_1(s,t) &= -\zeta\left(\delta - \mathrm{e}^{-\eta_\mu^*(s,t)}(1-\delta)\right), \\ d_0(s,t) &= 1 - \delta - \delta\mathrm{e}^{-\eta_\mu^*(s,t)} - (1-\zeta)(1-\delta)\left(1 - \mathrm{e}^{-\eta_\mu^*(s,t)}\right), \end{aligned}$$

et

$$d_1(s,t) = -\zeta(1-\delta)\left(1 - \mathrm{e}^{-\eta_\mu^*(s,t)}\right).$$

Les probabilités $\widetilde{p}_k^{(\zeta)}(s,t)$ peuvent ainsi être définies de manière récursive :

$$\widetilde{p}_0^{(\zeta)}(s,t) = \frac{n_0(s,t)}{d_0(s,t)}\,, \qquad \widetilde{p}_1^{(\zeta)}(s,t) = \frac{n_1(s,t)}{d_0(s,t)} - \frac{d_1(s,t)}{d_0(s,t)}\widetilde{p}_0^{(\zeta)}(s,t)\,,$$

et pour $k \geqslant 2$:

$$\widetilde{p}_k^{(\zeta)}(s,t) = -\frac{d_1(s,t)}{d_0(s,t)}\widetilde{p}_{k-1}^{(\zeta)}(s,t)\,.$$

Puis, par le même raisonnement que dans la sous-partie 3.3.2, il est possible d'exprimer les probabilités $\left(q_k^{(\zeta)}\right)_{k\in\mathbb{N}}$ associées à la fonction génératrice $h^{(\zeta)}$ en fonction des probabilités $\left(\widetilde{p}_k^{(\zeta)}\right)_{k\in\mathbb{N}}$:

$$q_k^{(\zeta)} = \int_0^t \widetilde{p}_k^{(\zeta)}(u,t)\frac{\lambda_\nu(u)\mathrm{e}^{-\eta_\nu(u,t)}}{1-\mathrm{e}^{-\eta_\nu(0,t)}}\mathrm{d}u\,.$$

Finalement, les probabilités $\left(p_k^{(\zeta)}\right)_{k\in\mathbb{N}}$ de la variable $M^{(\zeta)}$ sont calculées grâce au même algorithme que précédemment, i.e. :

$$p_0^{(\zeta)} = \mathrm{e}^{-m(1-q_0^{(\zeta)})}\,, \tag{3.40}$$

et pour tout $k > 0$:

$$p_k^{(\zeta)} = \frac{m}{k}\sum_{i=1}^{k} i q_i^{(\zeta)} p_{k-i}^{(\zeta)}\,. \tag{3.41}$$

Dans la suite de cette thèse, nous utiliserons les notations de la table 3.2 pour désigner les différents modèles de mutations possibles (ζ représente pour toutes ces notations le paramètre de dilution). Chacun de ces modèles peut être étendu au cas où le nombre final de cellules suit une loi de fonction de répartition K (voir fin de la sous-partie 2.2.2). Nous désignerons ces modèles selon le schéma suivant :

— les modèles de mutations MMI avec un nombre final de mutation aléatoire seront notés $MMIF(\pi,\gamma,\delta,F_\nu,F_\mu,\zeta,K)$: la réalisation de cette loi est un couple (M,N) tel que, conditionnellement à N, M suit la loi $MMI(\pi N,\gamma,\delta,F_\nu,F_\mu,\zeta)$;

— les modèles de mutations LDI avec un nombre final de mutation aléatoire seront notés $LDIF(\pi,\rho,\delta,\eta_{\mu,\infty},\zeta,K)$: la réalisation de cette loi est un couple (M,N) tel que, conditionnellement à N, M suit la loi $LDI(\pi N,\rho,\delta,\eta_{\mu,\infty},\zeta)$;

et ainsi de suite.

$MMI(m, \gamma, \delta, F_\nu, F_\mu, \zeta)$	Modèles de mutations inhomogènes où l'instant de division d'une cellule normale née à un temps s(resp. mutantes) a pour fonction de répartition $F_\nu(s, \cdot)$ (resp. $F_\mu(s, \cdot)$ et où F_ν vérifie $(\mathcal{H})$ (théorème 3.2.3)
$LDI(m, \rho, \delta, \eta_{\mu,\infty}, \zeta)$	Modèles MMI où F_μ vérifie $(\mathcal{H})$ tel que les taux de divisions des mutantes est proportionnel à celui des non-mutantes (théorème 3.3.1)
$MM(m, \rho, \delta, F, \zeta)$	Modèles de mutations homogènes où les durées de vie des mutantes sont *i.i.d.* de fonction de répartition F (théorème 2.2.6)
$LD(m, \rho, \delta, \zeta)$	Modèles LDI où $\eta_{\mu,\infty}$ est infini (corollaire 3.3.1)
	Modèles MM où les durées de vies des mutantes sont *i.i.d.* et exponentielles
$H(m, \rho, \delta, \zeta)$	Modèles MM où les durées de vie des mutantes sont constantes (modèles de Haldane)

TABLE 3.2 – **Notations des différents modèles de mutations**. Les quantités m, ρ , γ, δ et ζ dénotent respectivement le nombre moyen de mutations, le paramètre de fitness, la probabilité de mort d'une cellule normale, la probabilité de mort d'une cellule mutante et le paramètre de dilution.

Chapitre 4

Test de fluctuation - Étude par simulation

Nous allons maintenant tenter de répondre au problème statistique posé au début de cette thèse : construire des estimateurs robustes pour les paramètres d'intérêt que sont la probabilité de mutation π et le paramètre de fitness ρ. Les méthodes d'estimation P0, ML et GF décrites dans la partie 2.3 sont basées sur l'hypothèse que les processus de croissance des clones sont homogènes en temps. De plus les méthodes ML et GF ne peuvent être appliquées qu'en supposant que les durées de vie des mutantes sont soit constantes, soit distribuées selon une loi exponentielle. Toutes ces hypothèses ne correspondent cependant pas à la réalité [40, 78, 72, 92]. Or, estimer les paramètres d'un modèle, quand la réalité en suit un autre, conduit nécessairement à un biais d'estimation. Nous avons construit dans le chapitre précédent de nouveaux modèles de mutations, qui sont plus proches de la réalité que les modèles homogènes considérés jusqu'à présent.

Dans ce chapitre, nous allons nous placer dans les contextes décrits dans la partie 3.3. Nous ne nous intéresserons qu'à l'estimation de m et ρ et supposerons que la probabilité de mort δ et la limite du taux de divisions cumulés $\eta_{\mu,\infty}$ des mutantes sont connues. S'il est possible en théorie d'en construire des estimateurs, par exemple par le Maximum de Vraisemblance, il est très compliqué de les mettre en pratique. Comme cela était déjà le cas pour les formulations homogènes en temps [98], les variations en fonction de δ des modèles de mutations sont trop faibles pour que ce paramètre puisse être estimé en pratique. Or il est apparemment possible de n'identifier que l'ordre de grandeur de δ via des méthodes de mesures [83, 24]. Il s'avère que c'est également le cas pour $\eta_{\mu,\infty}$. En effet, nous pouvons constater que ces deux paramètres sont directement reliés :

$$\eta_{\mu,\infty} = \lim_{t\to\infty} \frac{1}{1-2\delta} \log\left(\mathbb{E}\left[\widetilde{M}(0,t)\right]\right) .$$

En d'autres termes, même s'il était possible d'estimer la taille finale d'un clone mutant démarré en 0, il faudrait connaître au moins une estimation de δ afin de pouvoir estimer $\eta_{\mu,\infty}$.

Les estimateurs de m et ρ que nous proposons sont issus des extensions des méthodes P0, ML et GF au cas où les modèles de croissance ne sont plus homogènes par rapport au temps. Plus précisément, nous généralisons ces méthodes afin de prendre en compte toutes les sources de biais mentionnées auparavant dans la partie 2.3. En particulier, nous étudierons de plus près le problème de la dilution. La prise en compte du paramètre de dilution s'effectue généralement via la correction proposée par STEWART et al. [86, eq. (41)]. Nous constaterons dans ce chapitre que cette correction n'est applicable que pour la méthode P0 pour les modèles $LD(m, 1, 0, \zeta)$. Nous étudierons également d'autres cas ignorés jusqu'à présent, comme par exemple la prise en compte des fluctuations des décomptes totaux pour la méthode GF. Les différentes extensions présentées ici permettent de construire de nouveaux estimateurs des quantités m et ρ qui conservent leurs propriétés de consistance et de normalité asymptotique.

Les trois méthodes d'estimation et leurs extensions ont été implémentées dans le package **flan** (disponible sur le CRAN [68] et sur GitHub [69]) pour le logiciel R [77, 71] qui a été développé durant cette thèse. Ce package contient des fonctions dédiées à la loi du décompte final de mutantes, ainsi que des outils d'inférence statistique (estimation et test statistiques). À partir de **flan**, nous avons été en mesure d'effectuer des études de simulations afin de comparer les trois méthodes et d'observer empiriquement les différents biais d'estimation décrits précédemment.

Nous commencerons par exposer les extensions des estimateurs de m, π et ρ construits via les méthodes P0, ML et GF dans la partie 4.1. Le package **flan** est ensuite présenté dans la partie 4.2. La partie 4.3 est finalement dédiée à l'étude par simulations de la robustesse des méthodes et des biais d'estimation via le package **flan**.

4.1 Extensions des estimateurs classiques

De même que dans la partie 2.3, nous considérerons les deux types d'échantillons $T1$ et $T2$, que nous rappelons ci-dessous :

$T1$ n réalisations indépendantes $(M_i)_{i=1,\dots,n}$ d'une variable M selon un même modèle de mutation, avec en plus les grandeurs statistiques suivantes : la moyenne et l'écart-type du nombre final de cellules ;

$T2$ n réalisations indépendantes $\big((M_i, N_i)\big)_{i=1,\dots,n}$ d'un couple de variables (M, N) selon un même modèle de mutation où le nombre final de cellules est aléatoire et de loi inconnue.

4.1.1 Méthode P0

Considérons tout d'abord que nous disposons d'un échantillon de type $T1$, supposé distribué selon un modèle de mutation MMI avec $\gamma = 0$. La construction des estimateurs pour ces modèles est identique lorsque $\zeta = 1$ à celle présentée dans la sous-partie 2.3.1.

Nous nous contenterons ici d'étudier la prise en compte d'une éventuelle dilution de la population, c'est-à-dire lorsque $\zeta < 1$. De manière générale, la probabilité d'avoir un décompte de mutantes nul après dilution est donnée par

$$\begin{aligned} p_0^{(\zeta)} &= \exp\left(-m\left(1-q_0^{(\zeta)}\right)\right) \\ &= \exp\left(-m\left(1-h(1-\zeta)\right)\right) . \end{aligned}$$

Rappelons que dans le cas d'une formulation $LD(m,1,0)$, la fonction génératrice $h(z)$ est donnée par (voir [56] et [8, part. 4.31])

$$h(z) = 1 + \frac{1-z}{z}\log(1-z)\,,$$

et donc

$$p_0^{(\zeta)} = \exp\left(-m\frac{\zeta}{1-\zeta}\log(\zeta)\right) .$$

Dans ce cas cas, si $\hat{m}_0$ est défini par (2.21), un estimateur sans biais de m est alors donné par

$$\hat{m}_0^{(\zeta)} = \frac{-\hat{m}_0(1-\zeta)}{\zeta\log(\zeta)} .$$

Nous retrouvons bien la formule de Stewart et al. [86, eq. (41)]. Cependant, cette correction n'est applicable que dans ce cas très précis. Par exemple, dans le cas d'une formulation $LD(m,\rho,0)$ avec $\rho \neq 1$, la probabilité qu'il n'y ait aucune mutante après dilution est égale à

$$p_0^{(\zeta)} = \exp\left(-m\left(1-\rho\int_0^1 \frac{(1-\zeta)v^\rho}{1-(1-\zeta)(1-v)}\mathrm{d}v\right)\right) .$$

Estimer m nécessite alors de connaître ρ. Plus généralement, cette méthode n'est plus applicable telle quelle lorsque $\zeta < 1$.

Il est cependant possible d'étendre l'estimateur (2.23) aux modèles MMI et MM lorsque $\zeta < 1$. Remarquons que pour tout $\delta < 1/2$, la probabilité d'extinction

$$\delta_* = \frac{\delta}{1-\delta}\,,$$

est toujours un point fixe de la fonction génératrice $\widetilde{\psi}(\cdot,s,t)$ (même lorsque $\eta_{\mu,\infty} < \infty$ dans le cas des modèles LDI), et donc de la fonction h. Ainsi :

$$\psi^{(\zeta)}\left(\delta_*^{(\zeta)}\right) = \exp\left(-m\left(1-\delta_*\right)\right) ,$$

où

$$\delta_*^{(\zeta)} = \frac{\delta_* - (1-\zeta)}{\zeta} . \tag{4.1}$$

Dans ce cas, un estimateur asymptotiquement sans biais et normal de m est alors donné par

$$\hat{m}_0 = \frac{-\log\left(\hat{\psi}_n\left(\delta_*^{(\zeta)}\right)\right)}{1-\delta_*},$$

où $\hat{\psi}_n$ est la fonction génératrice empirique du nombre de mutantes :

$$\hat{\psi}_n(z) = \frac{1}{n}\sum_{i=1}^{n} z^{M_i}.$$

Par la Δ-méthode [93, p. 79], la variance asymptotique de $\hat{m}_0$ est donnée par

$$v_{\hat{m}_0} = \frac{1}{n(1-\delta_*)^2}\left(\frac{\psi\left(\delta_*^{(\zeta)^2}\right)}{\psi\left(\delta_*^{(\zeta)}\right)^2} - 1\right).$$

Notons cependant que l'estimateur $\hat{m}_0$ n'a de sens que si $\delta_*^{(\zeta)}$ appartient au disque unité U. En particulier, si $\delta = 0$, alors le paramètre de dilution ζ ne peut être inférieur à 0.5.

L'estimation de la probabilité de mutation π est ensuite calculée en divisant par le nombre moyen de cellules. Cependant, nous avons vu dans la sous-partie 2.3 qu'il est nécessaire de prendre en compte d'éventuelles fluctuations du nombre final de cellules. Nous avions exposé la correction (2.27) proposée par Ycart et Veziris [99]. En supposant que $\zeta = 1$, nous allons généraliser cette correction au cas où $\delta > 0$. Rappelons que si le nombre final N de cellules est aléatoire, alors la loi du nombre final de mutantes M conditionnée à N est un modèle de mutation avec $m = \pi N$. Ainsi :

$$\mathbb{E}\left[{\delta_*}^M \mid N = k\right] = \mathrm{e}^{-\pi k(1-\delta_*)},$$

et donc :

$$\psi(\delta_*) = \mathcal{L}(\pi(1-\delta_*)),$$

où $\mathcal{L}$ est la transformée de Laplace (2.19) par rapport à la variable N. Ainsi, l'estimateur $\hat{m}_0$ défini par (2.23) est un estimateur sans biais de

$$\frac{-\log\left(\mathcal{L}\left(\pi\left(1-\delta_*\right)\right)\right)}{1-\delta_*}.$$

Par l'inégalité de Jensen, $\hat{m}_0$ sous-estime m, et $\hat{\pi}_0 = \hat{m}_0/\mathbb{E}[N]$ sous-estime également π. Si la loi de N est connue, et en supposant que la fonction réciproque $\mathcal{L}^{-1}$ de $\mathcal{L}$ existe, un estimateur asymptotiquement sans biais et normal de π est alors donné par

$$\frac{\mathcal{L}^{-1}\left(\hat{\psi}_n\left(\delta_*\right)\right)}{1-\delta_*}.$$

Par la Δ-méthode, sa variance asymptotique est égale à

$$\frac{\psi\left(\delta_*^2\right) - \psi(\delta_*)^2}{\Big((1-\delta_*)\mathcal{L}'\left(\pi(1-\delta_*)\right)\Big)^2}.$$

Or, la loi de N étant inconnue, nous devons nous contenter de réduire le biais d'estimation à partir du développement (2.26). Nous obtenons l'approximation au premier ordre suivante :

$$\begin{aligned}\frac{-\log\left(\mathcal{L}(\pi(1-\delta_*)\right)}{\mathbb{E}[N](1-\delta_*)} &= \pi - \frac{\pi^2 C^2 \mathbb{E}\left[N\right]\left((1-\delta_*)\right.}{2} \\ &= \pi\left(1 - \frac{mC^2\mathbb{E}\left[N\right]\left((1-\delta_*)\right.}{2}\right),\end{aligned}$$

où C est le coefficient de variation de N défini par (2.28). En conséquence, un estimateur asymptotiquement « sans biais » de π est alors donné par

$$\hat{\pi}_0 = \frac{\hat{m}_0}{\mathbb{E}\left[N\right]}\left(1 + \frac{\hat{m}_0 C^2\left(1-\delta_*\right)}{2}\right).$$

La variance asymptotique de $\hat{\pi}_0$ est donnée par

$$v_{\hat{\pi}_0} = \Big(1 + mC^2\left(1-\delta_*\right)\Big)^2 \frac{v_{\hat{m}_0}}{\mathbb{E}[N]^2}.$$

4.1.2 Méthode ML

Considérons que nous disposons d'un échantillon de type $T1$, supposé distribué selon un modèle de mutation LDI. Comme les probabilités $(p_k)_{k\in\mathbb{N}}$ et leurs dérivées partielles par rapport à m et ρ sont explicites, il est possible d'estimer m et ρ par le Maximum de Vraisemblance. Considérons d'abord que $\zeta = 1$. Cette méthode requiert le calcul des probabilités $(q_k)_{k\in\mathbb{N}}$ définies dans la sous-partie 3.3.2. Il est par ailleurs possible de calculer un équivalent de q_k pour de grandes valeurs de k. En effet, pour tout $k > 0$, (3.36) peut s'écrire :

$$\begin{aligned}q_k &= \frac{1}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}\left(\frac{1-2\delta}{1-\delta}\right)^2\left(\int_0^1 \frac{(1-v)^{k-1}}{\left(1-\frac{\delta}{1-\delta}v\right)^{k+1}}\rho v^\rho \mathrm{d}v - \int_0^{\mathrm{e}^{-\eta^*_{\mu,\infty}}} \frac{(1-v)^{k-1}}{\left(1-\frac{\delta}{1-\delta}v\right)^{k+1}}\rho v^\rho \mathrm{d}v\right) \\ &= \frac{1}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}\left(\frac{1-2\delta}{1-\delta}\right)^2\left(k^{-\rho-1}\int_0^k \frac{\left(1-\frac{w}{k}\right)^{k-1}}{\left(1-\frac{\delta}{1-\delta}\frac{w}{k}\right)^{k+1}}\rho w^\rho \mathrm{d}w\right. \\ &\qquad \left. - \int_0^{\mathrm{e}^{-\eta^*_{\mu,\infty}}} \frac{(1-v)^{k-1}}{(1-\frac{\delta}{1-\delta}v)^{k+1}}\rho v^\rho \mathrm{d}v\right).\end{aligned}$$

Et nous obtenons ainsi l'équivalent suivant :

$$q_k \underset{k\to\infty}{\sim} \frac{\rho}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}\left(\frac{1-2\delta}{1-\delta}\right)^2\left(\frac{\Gamma(\rho+1)}{k^{\rho+1}}\left(\frac{1-2\delta}{1-\delta}\right)^{-\rho-1} - \int_0^{\mathrm{e}^{-\eta^*_{\mu,\infty}}} \frac{(1-v)^{k-1}}{(1-\frac{\delta}{1-\delta}v)^{k+1}}v^\rho \mathrm{d}v\right), \tag{4.2}$$

où Γ est la fonction Gamma définie pour tout $x>0$ par

$$\Gamma(x) = \int_0^{+\infty} v^{x-1}\mathrm{e}^{-v}\mathrm{d}v\,.$$

Écrivons à présent les dérivées par rapport à ρ des probabilités $(q_k)_{k\in\mathbb{N}}$:

$$\frac{\partial q_0}{\partial\rho} = \frac{1}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}\int_{\mathrm{e}^{-\mu^*_\infty}}^1 \frac{\delta-\delta v}{1-\delta-\delta v}\rho v^{\rho-1}\log(v)\mathrm{d}v + \left(\frac{1}{\rho} - \frac{\mu^*_\infty \mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}\right)q_0\,,$$

et pour tout $k>0$:

$$\frac{\partial q_k}{\partial\rho} = \frac{1}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}\int_{\mathrm{e}^{-\mu^*_\infty}}^1 \frac{(1-v)^{k-1}}{(1-\delta-\delta v)^{k+1}}\rho v^{\rho}\log(v)\mathrm{d}v + \left(\frac{1}{\rho} - \frac{\mu^*_\infty \mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}\right)q_k\,.$$

D'après (4.2), un équivalent pour des grandes valeurs de k peut s'écrire :

$$\begin{aligned}\frac{\partial q_k}{\partial\rho} =& \frac{\rho}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}\left[\left(\Gamma(\rho+1)k^{-\rho-1}\left(\frac{1-2\delta}{1-\delta}\right)^{1-\rho}\right)\left(\digamma(\rho+1)-\log\left(\frac{k(1-2\delta)}{1-\delta}\right)\right)\right.\\ &\left. - \left(\frac{1-2\delta}{1-\delta}\right)^2\int_0^{\mathrm{e}^{-\eta^*_{\mu,\infty}}}\frac{(1-v)^{k-1}}{\left(1-\frac{\delta}{1-\delta}v\right)^{k+1}}\log(v)v^\rho\mathrm{d}v\right] + \left(\frac{1}{\rho} - \frac{\eta^*_{\mu,\infty}\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}\right)q_k\,,\end{aligned}$$

où $\digamma$ est la fonction Digamma définie pour tout $x>0$ par

$$\digamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}\,.$$

Remarquons également que pour $\delta=0$:

$$\begin{aligned}\frac{\partial q_k}{\partial\rho} =& \frac{\rho}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}\left[B(\rho+1,k)(\digamma(\rho+1)+\digamma(\rho+1+k)) - \int_0^{\mathrm{e}^{-\eta^*_{\mu,\infty}}}\log(v)v^\rho(1-v)^{k-1}\mathrm{d}v\right]\\ &+\left(\frac{1}{\rho} - \frac{\eta^*_{\mu,\infty}\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}\right)q_k\,.\end{aligned}$$

Puis, nous pouvons alors en déduire le gradient des probabilités $(p_k)_{k\in\mathbb{N}}$. Tout d'abord :

$$\frac{\partial p_0}{\partial m} = -(1-q_0)p_0\,, \quad \text{et} \quad \frac{\partial p_0}{\partial \rho} = -\left[\frac{\partial m}{\partial \rho}(1-q_0) - m\frac{\partial q_0}{\partial \rho}\right]p_0$$
$$= -\,m\left[\frac{\mu_\infty^* \mathrm{e}^{-\rho\eta_{\mu,\infty}^*}}{1-\mathrm{e}^{-\rho\eta_{\mu,\infty}^*}}(1-q_0) - \frac{\partial q_0}{\partial \rho}\right]p_0\,.$$

La dérivée par rapport à ρ de ψ est donnée par

$$\begin{aligned}
\frac{\partial \psi}{\partial \rho}(z) &= -\left[\frac{\partial m}{\partial \rho}(1-h(z)) - m\frac{\partial h}{\partial \rho}(z)\right]\psi(z) \\
&= -\,m\left[\frac{\eta_{\mu,\infty}^* \mathrm{e}^{-\rho\eta_{\mu,\infty}^*}}{1-\mathrm{e}^{-\rho\eta_{\mu,\infty}^*}}(1-h(z)) - \frac{\partial h}{\partial \rho}(z)\right]\psi(z) \\
&= -\,m\left[\frac{\eta_{\mu,\infty}^* \mathrm{e}^{-\rho\eta_{\mu,\infty}^*}}{1-\mathrm{e}^{-\rho\eta_{\mu,\infty}^*}}\left(1-\sum_{i\geqslant 0} z^i q_i\right) - \left(\sum_{i\geqslant 0}\frac{\partial q_i}{\partial \rho}z^i\right)\right]\left(\sum_{j\geqslant 0} p_j z^j\right) \\
&= -\,m\left[\frac{\eta_{\mu,\infty}^* \mathrm{e}^{-\rho\eta_{\mu,\infty}^*}}{1-\mathrm{e}^{-\rho\eta_{\mu,\infty}^*}} - \sum_{i\geqslant 0} z^i\left(\frac{\mu_\infty^* \mathrm{e}^{-\rho\eta_{\mu,\infty}^*}}{1-\mathrm{e}^{-\rho\eta_{\mu,\infty}^*}}q_i + \frac{\partial q_i}{\partial \rho}\right)\right]\left(\sum_{j\geqslant 0} p_j z^j\right)
\end{aligned}$$

Et donc pour tout $k > 0$:

$$\frac{\partial p_k}{\partial \rho} = -m\left[\frac{\eta_{\mu,\infty}^* \mathrm{e}^{-\rho\eta_{\mu,\infty}^*}}{1-\mathrm{e}^{-\rho\eta_{\mu,\infty}^*}}p_k - \sum_{i=1}^{k} p_{k-i}\left(\frac{\mu_\infty^* \mathrm{e}^{-\rho\eta_{\mu,\infty}^*}}{1-\mathrm{e}^{-\rho\eta_{\mu,\infty}^*}}q_i + \frac{\partial q_i}{\partial \rho}\right)\right]\,. \tag{4.3}$$

Pour tout $k > 0$, la dérivée par rapport à m de p_k est calculée via (2.34) :

$$\frac{\partial p_k}{\partial m} = \sum_{i=1}^{k} q_i p_{k-i} - p_k\,.$$

Définissons à présent la log-vraisemblance d'un échantillon de type $T1$ par

$$\begin{aligned}
\ell(m,\rho) &= \sum_{i=1}^{n} \log(p_{M_i}) \\
&= \sum_{i=0}^{\max_j M_j}\left[\log(p_i)\sum_{k=1}^{n}\mathbb{1}_{X_k=j}\right]\,,
\end{aligned} \tag{4.4}$$

Le couple d'estimateurs $(\hat{m}_{ML}, \hat{\rho}_{ML})$ obtenu en maximisant la log-vraisemblance ℓ est asymptotiquement sans biais et normal [57, Théo. 5.1., chapitre 6]. Les variances asymptotiques respectives de $\hat{m}_{ML}$ et $\hat{\rho}_{ML}$ sont données par

$$v_{\hat{m}_{ML}} = \frac{\mathcal{I}_{2,2}}{\det(I)} \quad \text{et} \quad v_{\hat{\rho}_{ML}} = \frac{\mathcal{I}_{1,1}}{\det(I)}\,, \tag{4.5}$$

où $\mathcal{I} = (\mathcal{I}_{i,j})_{i,j\in\{1,2\}}$ est la matrice d'information suivante :

$$\mathcal{I} = \sum_{j=0}^{\max M_i} \left[\begin{pmatrix} \left(\frac{\partial p_j}{\partial m}\frac{1}{p_j}\right)^2 & \frac{\partial p_j}{\partial m}\frac{\partial p_j}{\partial \rho}\frac{1}{p_j^2} \\ \frac{\partial p_j}{\partial m}\frac{\partial p_j}{\partial \rho}\frac{1}{p_j^2} & \left(\frac{\partial p_j}{\partial \rho}\frac{1}{p_j}\right)^2 \end{pmatrix} \sum_{i=1}^{n} \mathbb{1}_{M_i=j} \right] . \tag{4.6}$$

L'estimation de π est ensuite obtenue en divisant $\hat{m}_{ML}$ par le nombre moyen de cellules η. La prise en compte des fluctuations du nombre final de cellules sera exposée dans le chapitre 4. À partir des algorithmes (3.32), (2.34) et (4.3), la log-vraisemblance et son gradient peuvent être calculés de manière itérative. Cependant, ces formules doivent être appliquées à des vecteurs de taille égale au maximum de l'échantillon. Ainsi, les difficultés pratiques mises en avant par Hamon et Ycart [31], sont toujours présentes : l'optimisation peut être très longue et numériquement instable. Il est toujours possible d'utiliser des méthodes comme la *winsorisation* (voir [95, part. 2.2.]) afin de réduire ces effets de queue. La winsorisation consiste en remplacer les décomptes excédant une certaine borne par cette dernière : s'il y a beaucoup de valeurs qui dépassent cette borne (ce qui peut être le cas lorsque m est élevé ou que ρ est petit), l'estimation par la méthode ML ne sera pas pertinente.

Bien que l'optimisation de la vraisemblance par rapport à δ reste compliquée en pratique [98], les dérivées par rapport à δ des probabilités $(q_k)_{k\in\mathbb{N}}$ et $(p_k)_{k\in\mathbb{N}}$ peuvent être calculées, par l'algorithme suivant :

$$\begin{aligned} \frac{\partial q_0}{\partial \delta} =& \frac{2\rho\eta_{\mu,\infty}\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}} \left(q_0 - \frac{\delta\left(1-\mathrm{e}^{\eta^*_{\mu,\infty}}\right)\rho\mathrm{e}^{(\rho-1)\eta^*_{\mu,\infty}}}{1-\delta-\delta\mathrm{e}^{\eta^*_{\mu,\infty}}} \right) \\ &+ \frac{1}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}} \int_{\mathrm{e}^{-\eta^*_{\mu,\infty}}}^{1} \frac{(1-v)\rho v^{\rho-1}}{(1-\delta-\delta v)^2}\mathrm{d}v \,, \end{aligned}$$

et :

$$\begin{aligned} \frac{\partial p_0}{\partial \delta} &= -\left[\frac{\partial m}{\partial \delta}(1-q_0) - m\frac{\partial q_0}{\partial \delta}\right] p_0 \\ &= -m\left[\frac{2\rho\eta_{\mu,\infty}\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}(1-q_0) - \frac{\partial q_0}{\partial \delta}\right] p_0 \,. \end{aligned}$$

Puis, pour tout $k > 0$:

$$\begin{aligned}\frac{\partial q_k}{\partial \delta} =& \frac{1}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}\Bigg\{ q_k\left(\frac{-2\delta(1-\delta)-(1-2\delta)(k-1)}{(1-2\delta)(1-\delta)} + 2\rho\eta^*_{\mu,\infty}\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}\right)\\ &+ (1-2\delta)(1-\delta)^{k-1}\Bigg[-\frac{\left(1-\mathrm{e}^{-\eta^*_{\mu,\infty}}\right)^{k-1}}{1-\delta-\delta\mathrm{e}^{-\eta^*_{\mu,\infty}}} 2\eta^*_{\mu,\infty}\rho\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}\\ &+ \int_{\mathrm{e}^{-\eta^*_{\mu,\infty}}}^1 \frac{(1-v)^{k-1}(1+v)(k+1)}{(1-\delta-\delta v)^{k+1}}\rho v^\rho \mathrm{d}v\Bigg]\Bigg\}.\end{aligned}$$

La dérivée par rapport à δ de ψ étant donnée par

$$\begin{aligned}\frac{\partial \psi}{\partial \delta}(z) =& -\left[\frac{\partial m}{\partial \delta}(1-h(z)) - m\frac{\partial h}{\partial \delta}(z)\right]\psi(z)\\ =& -m\left[\frac{2\rho\eta_{\mu,\infty}\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}(1-h(z)) - \frac{\partial h}{\partial \delta}(z)\right]\psi(z)\\ =& -m\left[\frac{2\rho\eta_{\mu,\infty}\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}} - \sum_{i\geqslant 0} z^i\left(\frac{\partial q_i}{\partial \delta} + \frac{2\rho\eta_{\mu,\infty}\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}q_i\right)\right]\left(\sum_{j\geqslant 0} p_j z^j\right),\end{aligned}$$

Nous pouvons ensuite en déduire pour tout $k > 0$:

$$\frac{\partial p_k}{\partial \delta} = -m\left[\frac{2\rho\eta_{\mu,\infty}\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}p_k - \sum_{i=1}^{k} p_{k-i}\left(\frac{\partial q_i}{\partial \delta} + \frac{2\eta^*_{\mu,\infty}\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}q_i\right)\right].$$

Les dérivées par rapport à $\eta^*_{\mu,\infty}$ des probabilités $(q_k)_{k\in\mathbb{N}}$ et $(p_k)_{k\in\mathbb{N}}$ peuvent également être calculées par itération :

$$\frac{\partial q_0}{\partial \eta^*_{\mu,\infty}} = \frac{\rho\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}\left(-q_0 + \frac{\delta\left(1-\mathrm{e}^{-\mu^*_\infty}\right)}{1-\delta-\delta\mathrm{e}^{-\mu^*_\infty}}\right),$$

d'où :

$$\begin{aligned}\frac{\partial p_0}{\partial \eta^*_{\mu,\infty}} =& -\left[\frac{\partial m}{\partial \eta^*_{\mu,\infty}}(1-q_0) - m\frac{\partial q_0}{\partial \eta^*_{\mu,\infty}}\right]p_0\\ =& m\left[\frac{\rho\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}(1-q_0) + \frac{\partial q_0}{\partial \eta^*_{\mu,\infty}}\right]p_0.\end{aligned}$$

Puis, pour tout $k > 0$:

$$\frac{\partial q_k}{\partial \eta^*_{\mu,\infty}} = \frac{\rho\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}\left(-q_k + \frac{\mathrm{e}^{-\eta^*_{\mu,\infty}}\left(1-\mathrm{e}^{-\eta^*_{\mu,\infty}}\right)^{k-1}}{\left(1-\delta-\delta\mathrm{e}^{-\eta^*_{\mu,\infty}}\right)^{k+1}}\right).$$

En écrivant la dérivée par rapport à $\eta^*_{\mu,\infty}$ de ψ sous la forme suivante :

$$\begin{aligned}\frac{\partial \psi}{\partial \eta^*_{\mu,\infty}}(z) =& -\left[\frac{\partial m}{\partial \eta^*_{\mu,\infty}}(1-h(z)) - m\frac{\partial h(z)}{\partial \eta^*_{\mu,\infty}}\right]\psi(z)\\ =& m\left[\frac{\rho \mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}(1-h(z)) + \frac{\partial h(z)}{\partial \eta^*_{\mu,\infty}}\right]\psi(z)\\ =& m\left[\frac{\rho \mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}} + \sum_{i\geqslant 0} z^i\left(\frac{\partial q_i}{\partial \eta^*_{\mu,\infty}} - \frac{\rho \mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}q_i\right)\right]\left(\sum_{j\geqslant 0} p_j z^j\right),\end{aligned}$$

nous en déduisons pour tout $k > 0$:

$$\frac{\partial p_k}{\partial \eta^*_{\mu,\infty}} = m\left[\frac{\rho \mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}p_k + \sum_{i=1}^{k} p_{k-i}\left(\frac{\partial q_i}{\partial \eta^*_{\mu,\infty}} - \frac{\rho \mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}q_i\right)\right].$$

Il est également possible de considérer les modèles LDI avec $\zeta < 1$. Considérons à présent les modèles H avec $\zeta = 1$. Pour tout $k \geqslant 0$:

$$\begin{aligned}\frac{\partial q_k}{\partial \rho} &= \sum_{i\geqslant 0} r_k^{(i)}\left(-ia\mathrm{e}^{-na\rho} + (i+1)a\mathrm{e}^{-(i+1)a\rho}\right)\\ &= \sum_{i\geqslant 0} r_k^{(n)} a\mathrm{e}^{-ia\rho}\left((i+1)\mathrm{e}^{-a\rho} - n\right),\end{aligned}$$

et

$$\begin{aligned}\frac{\partial q_k}{\partial \delta} &= \sum_{i\geqslant 0}\frac{\partial r_k^{(i)}}{\partial \delta}\left(\mathrm{e}^{-ia\rho} - \mathrm{e}^{-(i+1)a\rho}\right) + r_k^{(i)}\left(\frac{i\rho}{(1-\delta)}\mathrm{e}^{-ia\rho} - \frac{(i+1)\rho}{(1-\delta)}\mathrm{e}^{-(i+1)a\rho}\right)\\ &= \left(1-\mathrm{e}^{-a\rho}\right)\sum_{i\geqslant 0}\frac{\partial r_k^{(i)}}{\partial \delta}\mathrm{e}^{-ia\rho} + 2\rho\mathrm{e}^{-a}\sum_{i\geqslant 0} r_k^{(i)}\mathrm{e}^{-ia\rho}\left(i-(i+1)\mathrm{e}^{-a\rho}\right)\end{aligned}$$

Le calcul pour tout $i \geqslant 1$ des dérivées $\left(\frac{\partial a_k^{(i)}}{\partial \delta}\right)_{k\in\mathbb{N}}$ se fait de la même manière que pour le calcul des probabilités $\left(r_k^{(i)}\right)_{k\in\mathbb{N}}$. En effet pour tout z dans le disque U :

$$\frac{\partial b_i}{\partial \delta}(z) = \sum_{k\geqslant 0}\frac{\partial r_k^{(i)}}{\partial \delta}z^k = 1 - (b_{i-1}(z))^2 + 2(1-\delta)\frac{\partial b_{i-1}}{\partial \delta}(z)b_{i-1}(z),$$

et

$$\frac{\partial b_0}{\partial \delta}(z) = 0.$$

La suite de polynômes $\left(\frac{\partial b_i}{\partial \delta}\right)_{i\geqslant 0}$ peut ainsi être construite, et nous pouvons en déduire pour tout $i \geqslant 0$, la suite $\left(\frac{\partial r_n^{(i)}}{\partial \delta}\right)_{k\geqslant 0}$. De même que pour l'identification des probabilités $\left(r_k^{(i)}\right)_{k\in\mathbb{N}}$, il est possible d'utiliser la transformée de Fourier rapide afin d'identifier les $\left(\frac{\partial r_n^{(i)}}{\partial \delta}\right)_{k\geqslant 0}$ comme coefficients de carrés de polynômes. Cependant, chaque polynôme b_i ayant 2^i coefficients, le calcul de ces derniers devient en pratique rapidement coûteux. En conséquence, l'utilisation de la méthode ML est déconseillée pour estimer m et/ou ρ lorsqu'un modèle H avec $\delta > 0$ est considéré. De plus, jusqu'à présent nous n'avons pas été en mesure d'expliciter les probabilités des modèles H lorsque $\zeta < 1$.

Pour finir, considérons le cas où le nombres final est aléatoire. Si nous disposons d'un échantillon de type $T2$, distribué selon un modèle $LDIF$ où HF, il est possible d'estimer le couple (π, ρ) en maximisant la log-vraisemblance (2.36). Si l'échantillon considéré est de type $T1$ avec un coefficient de variation du nombre final de cellules non-nul, nous devons trouver une correction prenant en compte ces fluctuations. Rappelons que dans ce cas :

$$\psi(z) = \mathcal{L}\left(\pi(1-h(z))\right) ,$$

où $\mathcal{L}$ est la transformée de Laplace (2.19). Nous pouvons mettre en relation le paramètre m d'une formulation MMI avec les paramètres π et K d'un modèle $MMIF$:

$$m = \frac{-\log\left(\mathcal{L}\left(\pi(1-h(z))\right)\right)}{1-h(z)} .$$

Un estimateur sans biais de π peut alors être donné par

$$\hat{\pi}_{ML} = \frac{\mathcal{L}^{-1}\left(\mathrm{e}^{-\hat{m}_{ML}(1-h(z))}\right)}{1-h(z)} .$$

Par la Δ-méthode, sa variance est donnée par

$$v_{\hat{\pi}_{ML}} = v_{\hat{m}_{ML}} \left(\frac{\mathrm{e}^{-m(1-h(z))}}{\mathcal{L}'(\pi(1-h(z)))}\right)^2 .$$

Encore une fois, la transformée $\mathcal{L}$ est inconnue. À partir du développement (2.26), nous pouvons écrire l'approximation suivante pour tout $z \in U$:

$$\frac{-\log\left(\mathcal{L}\left(\pi(1-h(z))\right)\right)}{\mathbb{E}[N](1-h(z))} \approx \pi\left(1 - \frac{m(1-h(z))C^2}{2}\right) . \tag{4.7}$$

En conséquence, un estimateur approximativement sans biais et normal de π est donné par

$$\hat{\pi}_{ML} = \frac{\hat{m}_{ML}}{\mathbb{E}[N]} \left(1 + \frac{\hat{m}_{GF}\,(1-h(z))C^2}{2}\right). \tag{4.8}$$

Sa variance asymptotique s'écrit :

$$v_{\hat{\pi}_{ML}} = \left(1 + m\,(1-h(z))C^2\right)^2 v_{\hat{m}_{ML}}\,.$$

Remarquons que la définition de cette correction dépend directement du paramètre z. En théorie, cette quantité doit être choisie de manière à minimiser la variance $v_{\hat{\pi}_{ML}}$. Or cette variance dépend de la valeur théorique m. Afin de sélectionner à l'avance la valeur de z, nous avons suivi le même processus que pour la sélection des paramètres personnalisés pour la méthode GF dans [31, part. 4.].

4.1.3 Méthode GF

Nous donnons à présent différentes extensions de la méthode GF exposée par Hamon et Ycart [31]. Nous commençons par écrire cette méthode pour les échantillons de type $T1$. Nous nous plaçons d'abord dans le cas des modèles LDI avec $\zeta = 1$. Cette méthode nécessite de calculer la fonction (3.34), ainsi que sa dérivée par rapport à ρ. Elles peuvent être implémentées sous la forme suivante :

$$h(z) = \delta_* + \frac{z_*(1-\delta_*)}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}} \int_{\mathrm{e}^{-\eta^*_{\mu,\infty}}}^1 \frac{\rho v^\rho}{1+z_* v}\mathrm{d}v\,,$$

et

$$\frac{\partial h(z)}{\partial \rho} = \frac{z_*(1-\delta_*)}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}} \left\{ \left[1 - \frac{\rho\eta^*_{\mu,\infty}\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho\eta^*_{\mu,\infty}}}\right] \int_{\mathrm{e}^{-\eta^*_{\mu,\infty}}}^1 \frac{v^\rho}{1+z_* v}\mathrm{d}v + \int_{\mathrm{e}^{-\eta^*_{\mu,\infty}}}^1 \frac{\rho v^\rho}{1+z_* v}\log(v)\mathrm{d}v \right\},$$

où les constantes δ_* et z_* sont définies par

$$\delta_* = \frac{\delta}{1-\delta}\,, \quad \text{et} \quad z_* = \frac{z-\delta_*}{1-z}\,.$$

Considérons z_1, z_2, z_3 dans $]0\,;1[$. Les expressions des estimateurs de m et ρ restent identiques à celles données dans la sous-partie 2.3.3.

$$\hat{m}_{GF}(z_3) = \frac{\log\left(\hat{\psi}_n(z_3)\right)}{h_{\hat{\rho}_{GF}(z_1,z_2)}(z_3) - 1} \quad \text{et} \quad \hat{\rho}_{GF}(z_1,z_2) = g^{-1}(\hat{y}_n)\,, \tag{4.9}$$

où h_x est la fonction (3.34) avec $\rho = x$, et

$$g(x) = \frac{h_x(z_1)-1}{h_x(z_2)-1} \quad \text{et} \quad \hat{y}_n = \frac{\log\left(\hat{\psi}_n(z_1)\right)}{\log\left(\hat{\psi}_n(z_2)\right)}\,.$$

Puis, en appliquant le théorème (3.4) de Rémillard et Theodorescu [80] et la Δ-méthode, nous pouvons prouver que le couple d'estimateurs $(\hat{m}_{GF}, \hat{\rho}_{GF})$ conserve les propriétés de la proposition 2.3.1 : il est fortement consistant, asymptotiquement normal et sa matrice de covariance asymptotique peut être explicitée.

Considérons à présent que $\zeta \leqslant 1$. Les expressions des estimateurs de m et ρ s'écrivent alors :

$$\hat{m}_{GF}(z_3) = \frac{\log\left(\hat{\psi}_n(z_3)\right)}{h^{(\zeta)}_{\hat{\rho}_{GF}(z_1,z_2)}(z_3) - 1} \quad \text{et} \quad \hat{\rho}_{GF}(z_1, z_2) = g^{-1}(\hat{y}_n)\,, \tag{4.10}$$

où $h^{(\zeta)}_x$ est la fonction (3.39) avec $\rho = x$, et

$$g(x) = \frac{h^{(\zeta)}_x(z_1) - 1}{h^{(\zeta)}_x(z_2) - 1} \quad \text{et} \quad \hat{y}_n = \frac{\log\left(\hat{\psi}_n(z_1)\right)}{\log\left(\hat{\psi}_n(z_2)\right)}\,.$$

Encore une fois, le couple d'estimateurs $(\hat{m}_{GF}, \hat{\rho}_{GF})$ conserve les propriétés de la proposition 2.3.1.

Proposition 4.1.1. *Soient z_1, z_2, z_3 dans $]0\,;1[$, avec $z_1 \neq z_2$. Soit $C = \big(c(z_i, z_j)\big)_{i,j=1,2,3}$ la matrice de covariance asymptotique du vecteur aléatoire :*

$$\sqrt{n}\left(\left(\hat{\psi}_n(z_1), \hat{\psi}_n(z_2), \hat{\psi}_n(z_3)\right) - \left(\psi^{(\zeta)}(z_1), \psi^{(\zeta)}(z_2), \psi^{(\zeta)}(z_3)\right)\right)\,,$$

c'est-à-dire

$$\begin{aligned} c(z_i, z_j) &= \psi^{(\zeta)}(z_i z_j) - \psi^{(\zeta)}(z_i)\psi^{(\zeta)}(z_j) \\ &= \psi(1 - \zeta + \zeta z_i z_j) - \psi(1 - \zeta + \zeta z_i)\psi(1 - \zeta + \zeta z_j)\,. \end{aligned}$$

Soit la matrice $A = (a_{i,j})_{\substack{i=1,2,3\\ j=1,2}}$ suivante :

$$\begin{aligned} a_{1,1} &= \frac{m a_{1,2}}{h^{(\zeta)}(z_3) - 1}\frac{\partial h^{(\zeta)}(z_3)}{\partial \rho}\,; \\ a_{1,2} &= \frac{h^{(\zeta)}(z_2) - 1}{m\psi^{(\zeta)}(z_1)\left(\frac{\partial h^{(\zeta)}(z_1)}{\partial \rho}(h^{(\zeta)}(z_2) - 1) - \frac{\partial h^{(\zeta)}(z_2)}{\partial \rho}(h^{(\zeta)}(z_1) - 1)\right)}\,; \\ a_{2,1} &= \frac{m a_{2,2}}{h^{(\zeta)}(z_3) - 1}\frac{\partial h^{(\zeta)}(z_3)}{\partial \rho}\,; \\ a_{2,2} &= \frac{h^{(\zeta)}(z_1) - 1}{m\psi^{(\zeta)}(z_2)\left(\frac{\partial h^{(\zeta)}(z_2)}{\partial \rho}(h^{(\zeta)}(z_1) - 1) - \frac{\partial h^{(\zeta)}(z_1)}{\partial \rho}(h^{(\zeta)}(z_2) - 1)\right)}\,; \\ a_{3,1} &= \frac{1}{\psi^{(\zeta)}(z_3)(h^{(\zeta)}(z_3) - 1)}\,; \\ a_{3,2} &= 0\,. \end{aligned}$$

Le vecteur aléatoire

$$\sqrt{n}\Big((\hat{m}_{GF}, \hat{\rho}_{GF}) - (m, \rho)\Big)$$

converge en loi vers la loi Normale bivariée centrée et de matrice de covariance $A^t C A$.

Notons que cette prise en compte de la dilution est également valable pour les modèles de Haldane H.

Rappelons que jusqu'à présent l'estimation de la probabilité de mutation π pour la méthode GF était ensuite calculée en divisant $\hat{m}_{GF}$ par le nombre moyen de cellules. Nous allons exposer ici une correction permettant de réduire le biais d'estimation sur π. Cet ajustement est similaire à ceux présentés pour la méthode P0 dans les parties 2.3.1 et 4.1.1. Tel qu'il est défini par (4.9), $\hat{m}_{GF}$ est un estimateur asymptotiquement sans biais et normal de :

$$\frac{-\log\left(\mathcal{L}\left(\pi\left(1-h(z_3)\right)\right)\right)}{1-h(z_3)}\,.$$

Par l'inégalité de Jensen, $\hat{m}_{GF}/\mathbb{E}\left[N\right]$ sous-estime π. Lorsque la fonction $\mathcal{L}^{-1}$ est connue, un estimateur asymptotiquement sans biais et normal de π est donné par

$$\hat{\pi}_{GF} = \frac{\mathcal{L}^{-1}\left(\hat{\psi}_n(z_3)\right)}{1-h(z_3)}\,,$$

et sa variance asymptotique par

$$v_{\hat{\pi}_{GF}} = \frac{\psi\left(z_3^2\right) - \psi(z_3)^2}{\Big((1-h(z_3))\mathcal{L}'\left(\pi(1-h(z_3))\right)\Big)^2}\,.$$

Rappelons qu'en pratique, la loi de N est inconnue. Cependant, en considérant à nouveau l'approximation (4.7), nous pouvons déduire de $\hat{m}_{GF}$ un estimateur approximativement sans biais et normal de π :

$$\hat{\pi}_{GF} = \frac{\hat{m}_{GF}}{\mathbb{E}\left[N\right]}\left(1 + \frac{\hat{m}_{GF}\left(1-h(z_3)\right)C^2}{2}\right)\,.$$

Sa variance asymptotique s'écrit :

$$v_{\hat{\pi}_{GF}} = \Big(1 + m\left(1-h(z_3)\right)C^2\Big)^2 v_{\hat{m}_{GF}}\,.$$

4.2 flan : un package R pour l'ANalyse de FLuctuation

Différents outils informatique ont déjà été développés autour de la problématique des tests de fluctuations [102, 30, 29]. Ces outils ont l'avantage d'être ergonomiques, mais sont incomplets, dans le sens où les hypothèses de modélisations pouvant être prises en compte

sont assez limitées. Le package pour le logiciel R **flan** a été développé dans le cadre de cette thèse. Il est disponible sur le CRAN [68] et sur GitHub [69]. Il fournit des outils pour l'analyse de fluctuation, et inclut en particulier les méthodes d'estimations et leurs extensions exposées dans la partie précédente. Nous présentons ici les différentes fonctions accessibles dans le package, et donnons également quelques détails d'implémentation.

4.2.1 Fonctionnalités et interface

Le package **flan** contient tout d'abord des fonctions dédiées à la loi du décompte final de mutantes. Les fonctions `dflan`, `pflan`, `qflan` calculent la densité, les probabilités et les quantiles des lois LD, H. La fonction `rflan` permet de simuler des échantillons de type $T1$ distribués selon les lois LD, H, et MM (où F est la loi log-normale ou gamma). Il est également possible de simuler des échantillons de type $T2$ pour lesquels les nombres finaux simulés selon une loi log-normale ajustée à une moyenne et un coefficient de variation renseignés par l'utilisateur. Ces fonctions ont été implémentées selon le même principe que les fonctions de R dédiées aux lois classiques. La fonction graphique `draw.clone` permet de représenter sous la forme d'un arbre binaire la croissance jusqu'à un temps fini d'une population issue d'une cellule normale. L'utilisation principale du package réside cependant dans les fonctions d'inférence statistique. La fonction `mutestim` calculant des estimations des paramètres m, π et/ou ρ à partir d'échantillons de type $T1$ ou $T2$. Cette fonction renvoie également les estimations des écart-types des estimateurs employés. Les trois méthodes d'estimations et leurs extensions décrites dans la partie précédente sont disponibles. La fonction `flan.test` permet de construire des tests statistiques sur un ou deux échantillons, en s'appuyant sur les propriétés de normalité asymptotique des estimateurs construits dans la partie précédente. Cette fonction a été implémentée selon le même principe que la fonction `t.test` de R.

Comme nous l'avons mentionné dans les parties 2.3 et 4.1, chacune des trois méthodes considérées dans cette thèse a ses limites pratiques. Si les arguments de la fonction `mutestim` ne respectent pas ces limitations, des messages d'erreur ou des avertissements seront renvoyés. Pour commencer, si $\delta = 0$, la méthode P0 ne peut être utilisée si l'échantillon ne contient aucun décompte nul. De plus, si $\zeta < 1$, cette méthode ne peut être employée si (4.1) n'est pas dans le disque unité U.

L'utilisation de la méthode ML nécessite généralement de seuiller l'échantillon, par winsorisation. Il faut alors prendre en compte deux faits lors du choix de la borne :

1. Si la valeur minimale de l'échantillon est plus grand que la borne, tous les éléments de l'échantillon seront remplacés par cette borne ;
2. Si la borne est trop grande, la maximisation de la log-vraisemblance peut être très longue ;

La valeur de la borne est fixée par défaut à 1024 dans la fonction `mutestim`.

Pour finir, la méthode GF n'a pas réellement de limite d'utilisation, même dans les cas extrêmes où les estimateurs de la méthode ML échouent (grandes valeurs de m et/ou

faibles valeurs de ρ). Cependant, l'estimation de ρ nécessite de résoudre une équation (voir eq.(4.9)). En théorie, cette équation est résoluble sur $\mathbb{R}_+$. En pratique, l'intervalle de recherche de la racine est borné. Donc s'il n'y a pas de dilution et que l'échantillon ne contient pas de jackpots, en d'autres termes lorsque $\zeta = 1$ et ρ est très grand, la solution de l'équation peut ne pas être dans l'intervalle de recherche. Dans ce cas, la fonction envoie un avertissement et fixe l'estimation de ρ à 1, et l'estimation de l'écart-type à 0. Remarquons qu'il est possible qu'un échantillon ne contienne pas de jackpots du fait d'un paramètre de dilution ζ très faible, auquel cas l'obstacle disparaît. L'intervalle de recherche est fixé à $[0.01\,;100]$ dans la fonction `mutestim`. De plus, l'initialisation de la méthode ML est effectuée avec la méthode GF. La recherche du maximum de vraisemblance est effectuée sur l'intervalle $[0.1\times\hat{\theta}_{GF}\,;10\times\hat{\theta}_{GF}]$, où $\hat{\theta}_{GF}$ est l'estimation du ou des paramètres d'intérêt. En conséquence si la méthode GF ne permet pas d'estimer la fitness ρ, il n'y aucune chance d'y parvenir avec la méthode ML. Un avertissement est également envoyé si l'initialisation par la méthode GF échoue. Ceci dit, si la méthode GF ne parvient pas à estimer la fitness, le fait que l'échantillon est distribué selon un des modèles de mutation décrit dans cette thèse devrait être remis en cause.

4.2.2 Implémentation

Nous allons à présent décrire brièvement l'implémentation de **flan**. Les calculs présents dans la majeure partie des fonctions font intervenir des boucles. En conséquence, le package a été codé principalement en C++ à l'aide des modules du package **Rcpp** [19, chap. 7]. Ce package permet d'exposer facilement des fonctions et des classes C++ à R. Les principales classes C++ sont les suivantes :

- `FLAN_Sim` : classe pour la simulation du décompte final de mutantes, couplé ou non à un nombre final de cellules ;
- `FLAN_SimClone` : classe pour la simulation de la taille d'un clone, en fonction du modèle de croissance choisi ;
- `FLAN_MutationModel` : classe pour le calcul des fonctions descriptives (probabilités, fonction génératrice,...) pour les modèles LDI et H ;
- `FLAN_Clone` : classe pour le calcul des fonctions descriptives pour la loi de la taille d'un clone, en fonction du modèle de croissance choisi.

L'interface de **Rcpp** permet également d'importer n'importe quelle fonction R dans le code C++. En particulier, cela permet de simuler un modèle de croissance inhomogène dans le temps en laissant le choix à l'utilisateur de choisir la croissance moyenne de la population. Il est également possible d'importer des fonctions R prédéfinies, évitant ainsi de ne pas employer de bibliothèque C. Ceci permet d'alléger la taille du package une fois installé et de faciliter son installation. Par exemple, le calcul numérique d'intégrale se fait habituellement via les librairies C **integration** et **alglib**, qui permettent de calculer des intégrales avec une précision proche de celle de la machine. Il est également

possible d'importer simplement la fonction `integrate` de R, qui est déjà implémentée en C. Cependant, ces importations augmentent le temps de calcul, ainsi que la mémoire consommée. Cette solution n'est donc pas satisfaisante. Une alternative consiste en l'utilisation du package **RcppGSL** [19, chap. 11], afin de créer une interface entre **Rcpp** et la librairie **gsl**. Cette dernière inclut entre autres de nombreuses méthodes d'intégration. Le lien avec la librairie **gsl** se fait automatiquement grâce à **RcppGSL** du moment qu'elle a été correctement installée. Il s'agit donc d'un bon compromis entre facilité d'installation et réduction des temps de calcul. De même, le calcul des probabilités des modèles H lorsque $\delta > 0$ nécessite de calculer les coefficients de carrés de polynômes de hauts degrés. Le package R **polynom** pourrait être utilisé, mais les fonctions proposées sont trop limitées en terme de degrés et provoquent des problèmes de mémoire. Une bonne alternative est la transformée de Fourier rapide, disponible en R via la fonction `fft`. Cependant, pour les mêmes raisons que pour la fonction `integrate`, nous préférons utiliser le package **RcppArmadillo** (voir [19, chap.10] et [20]), qui permet de faire facilement le lien entre **Rcpp** et la bibliothèque C++ **armadillo**. Cette librairie fournit des fonctions dédiées à l'algèbre linéaire, et en particulier, une transformée de Fourier performante. Pour finir, l'optimisation d'une fonction en R se fait habituellement à l'aide de la fonction `optim`. Nous avons décidé d'utiliser à la place la fonction `lbfgsb3` fournie par le package éponyme, implémentée en Fortran et bien plus performante que `optim`.

4.3 Étude par simulation des méthodes d'estimation

Comme mentionné au début de cette thèse, il est très important de construire des estimateurs qui soient le plus robuste possible malgré les variations des modèles. Les trois méthodes présentées dans cette thèse donnent des estimateurs dotés de bonnes propriétés asymptotiques, mais nous avons également constaté dans les parties 2.3 et 4.1 qu'ils possédaient tous leurs limites en pratique. De plus, leurs variances asymptotiques ne sont pas équivalentes. Par exemple, l'estimateur de m par la méthode P0 perd beaucoup en précision lorsque m augmente. À l'aide du package **flan** présenté dans la partie précédente, nous avons effectué des simulations de Monte Carlo afin de comparer ces trois méthodes en fonction des paramètres réels, et d'observer les différents biais d'estimation étudiés dans cette thèse.

4.3.1 Comparaison des méthodes d'estimation

Les Figures 4.1 et 4.2 illustrent les différences entre les trois méthodes d'estimation en fonction des valeurs réelles de m et ρ. Les méthodes sont comparées par rapport à leur erreur quadratique (MSE) relative définie par

$$\sqrt{\left(1-\frac{\hat{m}}{m}\right)^2+\left(1-\frac{\hat{\rho}}{\rho}\right)^2}. \tag{4.11}$$

Nous utilisons un code RGB : rouge pour la méthode GF, vert pour la méthode P0, bleu pour la méthode ML. Nous avons effectué les simulations en choisissant 20 valeurs de m allant de 0.5 à 10, et autant de valeurs de ρ allant de 0.2 à 5. Pour chacun des 400 couples (m, ρ) ainsi obtenus, nous avons appliqué la procédure suivante :

1. simuler 10^4 échantillons de taille 100 de la loi $LD(m, \rho, 0)$;
2. pour chaque échantillon, calculer les estimations de (m, ρ) selon les trois méthodes ;
3. à partir des 10^4 estimations, calculer les MSEs (4.11) de chaque méthode ;
4. pour chaque méthode, donner une couleur RGB selon les MSEs :
   - si le MSE est plus petit que 0.05, donner 1 à la couleur correspondante ;
   - si le MSE est plus grand que 1, donner 0 à la couleur correspondante ;
   - sinon, donner 1 moins le MSE à la couleur correspondante.

La carte a été tracée selon une échelle logarithmique en base 5 pour l'axe des ordonnées (valeurs de ρ).

La Figure 4.1 peut être séparée selon les quatre parties suivantes :

- Pour $(m, \rho) \in]0.5\,;3[\times]0.2\,;2.5[$, la couleur obtenue est grise : les trois méthodes sont plus ou moins équivalentes.
- Pour $(m, \rho) \in]3\,;10[\times]0.2\,;3.5[$, la couleur obtenue est magenta : les méthodes GF et ML sont équivalentes. Pour ces valeurs, la méthode P0 donne des estimations avec de grandes erreurs quadratiques, ou ne peut tout simplement pas être utilisé faute de zéro dans l'échantillon.
- Pour des valeurs faibles de ρ, la couleur obtenue est principalement rouge : la méthode GF est la seule méthode donnant des estimations pertinentes. Du fait de la winsorisation, les méthodes ML et P0 (cette dernière utilise le Maximum de Vraisemblance pour estimer ρ) donnent des estimations avec de grandes erreurs quadratiques.
- Pour des valeurs grandes de ρ, la couleur obtenue tend vers le noir : les trois méthodes donnent des estimations avec de grandes erreurs quadratiques. C'est particulièrement le cas pour $\rho \in]3.5\,;5[$: les jackpots sont alors très faibles, voire inexistants. La méthode GF ne peut estimer ρ, et l'estimation de m est fortement biaisée. En conséquence, les estimations via la méthode ML vont être également biaisées. La présence de vert dans le haut de la carte indique cependant que la méthode P0 donne des bonnes estimations : cela est dû à l'indépendance de la méthode vis-à-vis de ρ. Dans le cas où l'échantillon ne contient pas de jackpot, le fait qu'un modèle de mutation n'est pas adapté devrait donc être considéré.

La Figure 4.2 ressemble beaucoup à la Figure 4.1. Remarquons cependant que les estimations de la méthode GF obtenues pour de grandes valeurs de ρ semblent être meilleures en utilisant un modèle H qu'un modèle LD. De plus, les trois méthodes semblent être équivalentes même pour $\rho \in]0.2\,;2[$, tant que $m \leqslant 2$.

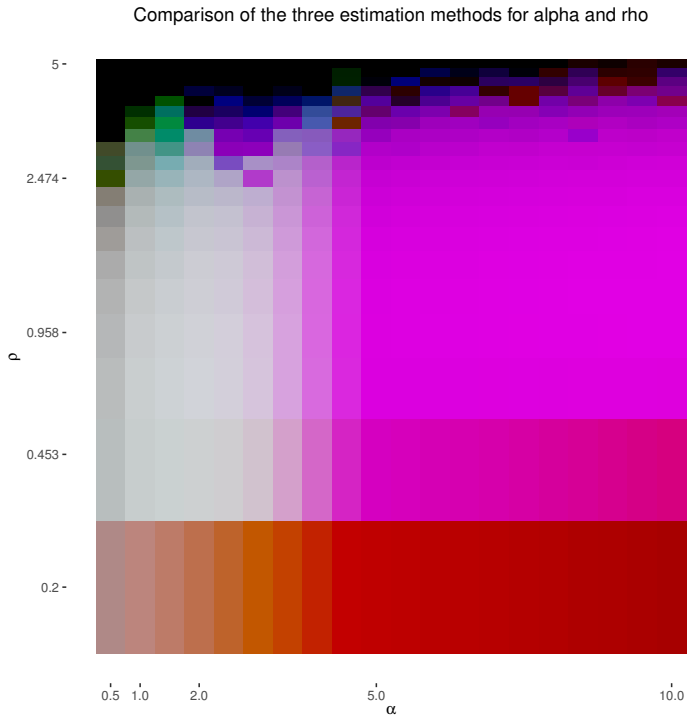

FIGURE 4.1 – **Comparaison des trois méthodes d'estimation en terme de MSE sous le modèle *LD*.** Pour chacun des 400 couples de paramètres $m = (0.5, \ldots, 10)$ (abscisses) et $\rho = (0.2, \ldots, 5)$ (ordonnées, échelle $\log_5$), 10^4 échantillons de taille 100 de la loi $LD(m, \rho, 0)$ ont été simulés. Les estimations de m et ρ ont été calculées avec les trois méthodes. Chaque méthode est caractérisée par une couleur : rouge pour la méthode GF, vert pour la méthode P0, bleu pour la méthode ML. L'intensité de la couleur d'une méthode est caractérisée par son MSE (4.11).

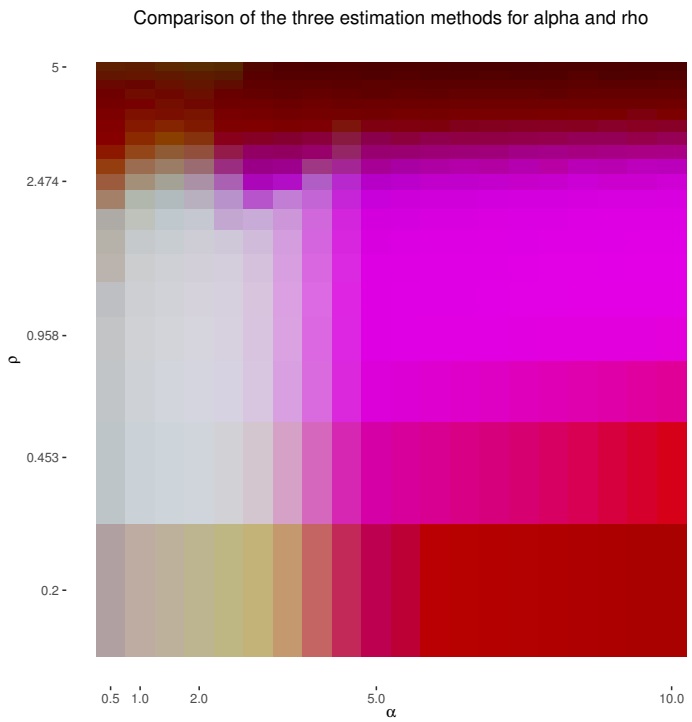

FIGURE 4.2 – **Comparaison des trois méthodes d'estimation en terme de MSE sous le modèle $\boldsymbol{H}$.** Pour chacun des 400 couples de paramètres $m = (0.5, \ldots, 10)$ (abscisses) et $\rho = (0.2, \ldots, 5)$ (ordonnées, échelle $\log_5$), 10^4 échantillons de taille 100 de la loi $H(m, \rho, 0)$ ont été simulés. Les estimations de m et ρ ont été calculées avec les trois méthodes. Chaque méthode est caractérisée par une couleur : rouge pour la méthode GF, vert pour la méthode P0, bleu pour la méthode ML. L'intensité de la couleur d'une méthode est caractérisée par son MSE (4.11).

4.3.2 Étude des biais d'estimation

Les estimateurs exposés dans la partie 2.3 sont construits sous des hypothèses non réalistes. En conséquence, des biais d'estimations vont forcément apparaître. Les biais existants sont ici illustrés par des études de simulations effectuées avec le package R **flan** (partie 4.2) selon le schéma suivant :

1. simuler 10^4 échantillons de taille 100, selon un modèle donné ;
2. pour chaque échantillon, estimer m et éventuellement ρ ;
3. observer la distribution de $\hat{\theta}/\theta$, où $\hat{\theta}$ est l'estimateur de la valeur réelle θ.

Les distributions des estimations sont représentées via des boxplots. Pour chaque figure, les lignes rouges correspondent à la valeur théorique. Nous considérerons qu'une quantité est correctement estimée si au moins la moitié des estimations ont un biais relatif inférieur à 10% (représenté par les lignes bleues). Nous avons constaté dans la sous-partie précédente que la méthode GF était en terme de MSE soit la meilleure, soit équivalente à au moins une des deux autres méthodes. Comme il s'agit également de la méthode la moins coûteuse des trois, les biais d'estimations seront illustrés principalement avec cette méthode.

Nous avons exposé dans la partie 2.2 l'écriture du modèle MM avec morts cellulaires. Bien que la probabilité de mort d'une cellule semble être faible en pratique [83, 24], il y a inévitablement un biais sur l'estimation de m. En effet, comme peut l'illustrer dans un premier temps la Figure 2.2, une proportion non-négligeable n'apparaît pas en fin d'expérience : le nombre moyen de mutations sera alors sous-estimé. La Figure 4.4 illustre l'influence du paramètre de mort δ sur les estimations de m et ρ. Les estimations sont calculées de deux manières différentes :

1. par la méthode GF avec $\delta = 0$ (boxplots de gauche) ;
2. par la méthode GF avec la valeur réelle de δ (boxplots de droite).

Nous constatons bien la présence d'un biais négatif lorsque les morts cellulaires sont ignorées. Nous pouvons cependant remarquer que pour $\delta = 0.05$ (Figure 4.3), la majorité des biais relatifs reste inférieur à 10%, que le paramètre de mort soit pris en compte ou non. Ce fait illustre par ailleurs la difficulté d'obtenir une estimation correcte du paramètre de mort (voir partie 2.3) : les estimations obtenues pour $\delta = 0.05$ sont très proches de celles obtenues lorsque $\delta = 0$. Lorsque δ augmente, le biais augmente également (Figure 4.4). Les résultats visuels montrent que le biais relatif d'estimation de m peut facilement dépasser 0.90 pour des valeurs de δ plus élevées. Rappelons qu'en théorie la valeur de δ est plus petite que 0.5, afin d'assurer que les clones ont une probabilité non-nulle de ne pas s'éteindre, cette quantité reste en pratique inférieure à 0.3. Quant à l'estimation de ρ, le biais d'estimation est très léger pour de grandes valeurs de m. L'erreur devient plus prononcée pour de plus faibles valeurs de m, et les plus grandes valeurs de ρ. Cependant, même en prenant en compte la valeur théorique de δ, l'estimation de ρ dans ce cas peut être délicate en pratique du fait de l'absence possible de jackpots (voir sous-

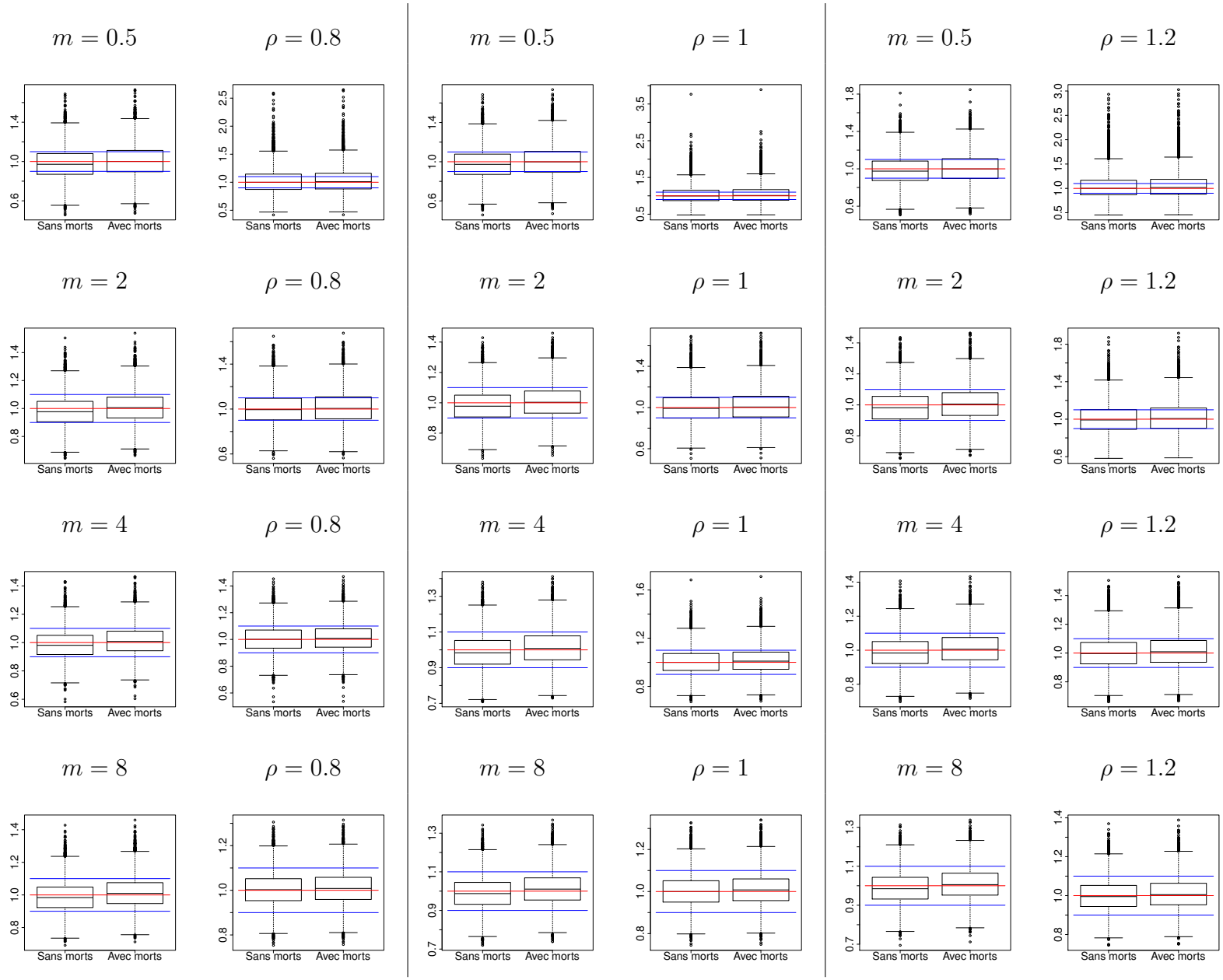

FIGURE 4.3 – **Estimations par la méthode GF en prenant en compte ou non les morts cellulaires ($\delta = 0.05$).** Pour chacun des 12 couples de paramètres $m = (0.5, 2, 4, 8)$ (colonnes) et $\rho = (0.8, 1, 1.2)$ (lignes), 10^4 échantillons de taille 100 de la loi $LD(m, \rho, \delta, 1)$ ont été simulés, avec $\delta = 0.05$. Pour chaque colonne, les deux premiers boxplots représentent la distribution des 10^4 rapports $\hat{m}_{GF}/m$ obtenus avec $\delta = 0$ (gauche), et avec la valeur théorique de δ (droite) ; les deux derniers boxplots représentent la distribution des 10^4 rapports $\hat{\rho}_{GF}/\rho$ obtenus avec $\delta = 0$ (gauche), et avec la valeur théorique de δ (droite).

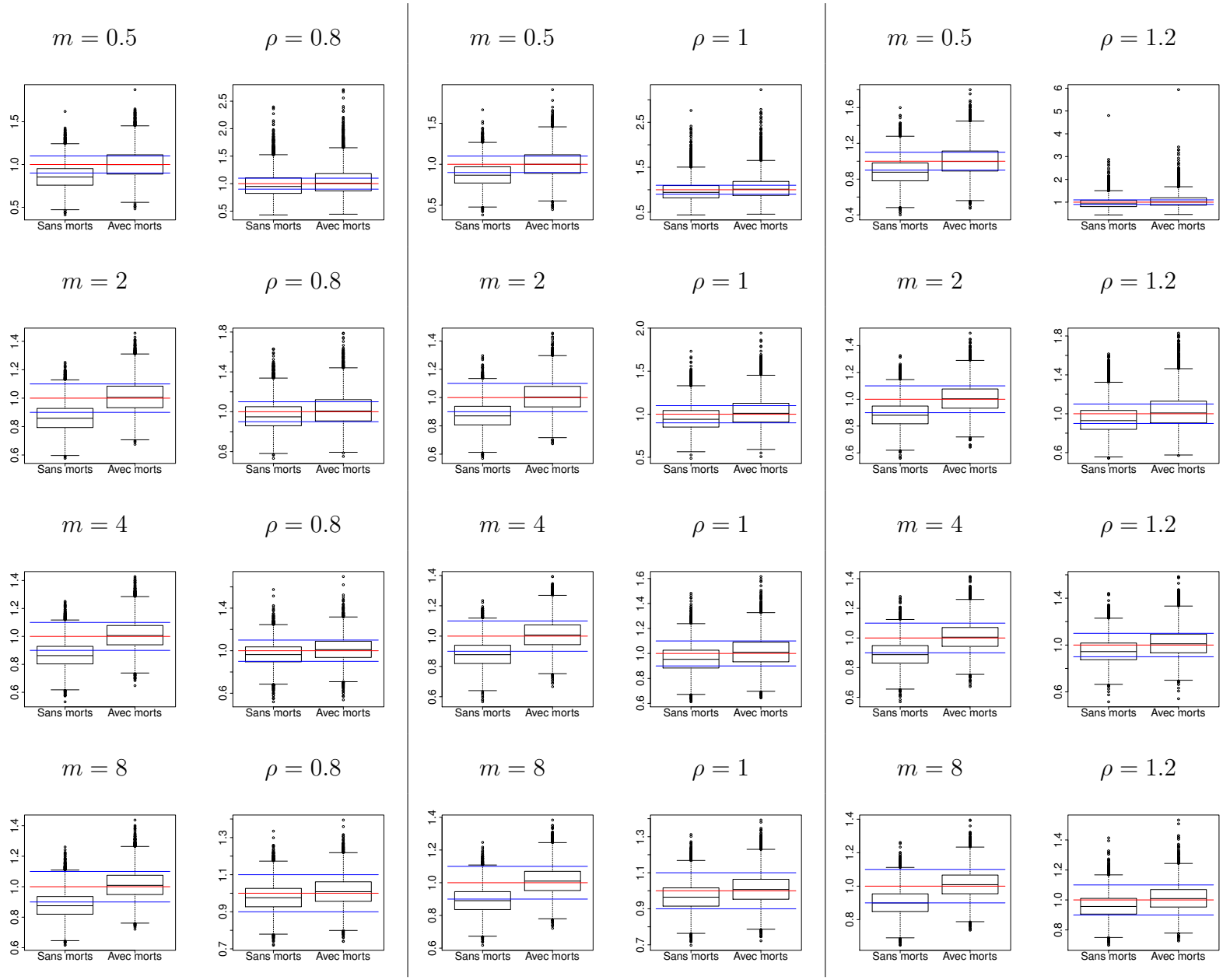

FIGURE 4.4 – **Estimations par la méthode GF en prenant en compte ou non les morts cellulaires ($\delta = 0.2$).** Pour chacun des 12 couples de paramètres $m = (0.5, 2, 4, 8)$ (colonnes) et $\rho = (0.8, 1, 1.2)$ (lignes), 10^4 échantillons de taille 100 de la loi $LD(m, \rho, \delta, 1)$ ont été simulés, avec $\delta = 0.2$. Pour chaque colonne, les deux premiers boxplots représentent la distribution des 10^4 rapports $\hat{m}_{GF}/m$ obtenus avec $\delta = 0$ (gauche), et avec la valeur théorique de δ (droite) ; les deux derniers boxplots représentent la distribution des 10^4 rapports $\hat{\rho}_{GF}/\rho$ obtenus avec $\delta = 0$ (gauche), et avec la valeur théorique de δ (droite).

partie 4.2.1), d'où la présence de valeurs aberrantes présentes (par exemple pour $m = 0.5$ et $\rho = 1.2$).

Nous avons décrit dans la partie 2.2 les modèles MMF, qui prennent en compte les fluctuations du nombre final de cellules N. Comme nous l'avons constaté dans les parties 2.3 et 4.1, ignorer ces fluctuations conduit inévitablement à une sous-estimation de la probabilité de mutation π, quelle que soit la méthode employée. La Figure 4.5 illustre l'influence du coefficient de variation C du nombre final de cellules sur l'estimation de π par la méthode ML. Elle est calculée à partir d'échantillons de type $T2$ distribués selon un modèle LDF. Pour chaque échantillon, trois estimations de π sont calculées :

1. en divisant $\hat{m}_{ML}$ par la moyenne empirique du nombre final de cellules et sans prendre en compte leurs fluctuations (boxplots centraux) ;
2. en appliquant la correction (4.8) (boxplots de droite).
3. en calculant directement $\hat{\pi}_{ML}$ à partir des couples (nombre final de mutantes – nombre final de cellules) (boxplots de gauche) ;

Nous constatons bien la présence d'un biais négatif lorsque les fluctuations sont ignorées. La correction (4.8) semble donner de bons résultats pour des valeurs faibles de m et/ou des valeurs faibles de C : la majorité des estimations obtenues ont un biais relatif inférieur à 10%. Cependant, l'efficacité de la réduction du biais semble diminuer à mesure que π et C augmentent. En particulier, le biais induit par la correction semble dépasser le biais initial pour des valeurs plus élevées de π. Cela pourrait être amélioré avec une meilleure approximation de $\mathcal{L}$, ce qui requiert de connaître des moments de N d'ordre plus élevé. Il est également possible de chercher à améliorer l'estimation du coefficient de variation. En effet, C est ici estimé par le rapport de l'écart-type empirique sur la moyenne empirique du nombre final de cellule. Or cette estimation est mauvaise en terme d'erreur quadratique [13]. Bien entendu, la meilleure estimation reste celle obtenue en appliquant la méthode ML avec le modèle LDF : une grande majorité des estimations ont un biais relatif inférieur à 10%, indépendamment de la valeur de π et C.

Nous avons également exposé dans la partie 2.2 l'écriture du modèle MM dans le cas où les durées de vie des mutantes ne sont plus *i.i.d.* selon une loi exponentielle. La loi du nombre final de mutantes n'est explicite que pour les modèles LD et H, et ce sont donc ces deux modèles qui sont utilisés pour estimer les paramètres m et ρ. En conséquence, si un autre modèle de croissance est utilisé lors de la simulation, les estimations seront biaisées. La Figure 4.6 illustre l'influence du modèle de croissance sur des estimations de m et ρ. Les échantillons sont simulés selon des modèles $MM(m, \rho, 0, F, 1)$, où F est la fonction de répartition de la loi Log-Normale ajustées aux données de Kelly et Rahn [40]. Pour chaque échantillon les estimations de m et ρ sont calculées :

1. par la méthode GF en supposant que l'échantillon est distribué selon une loi LD (boxplots de gauche) ;
2. par la méthode GF en supposant que l'échantillon est distribué selon une loi H (boxplots de droite) ;

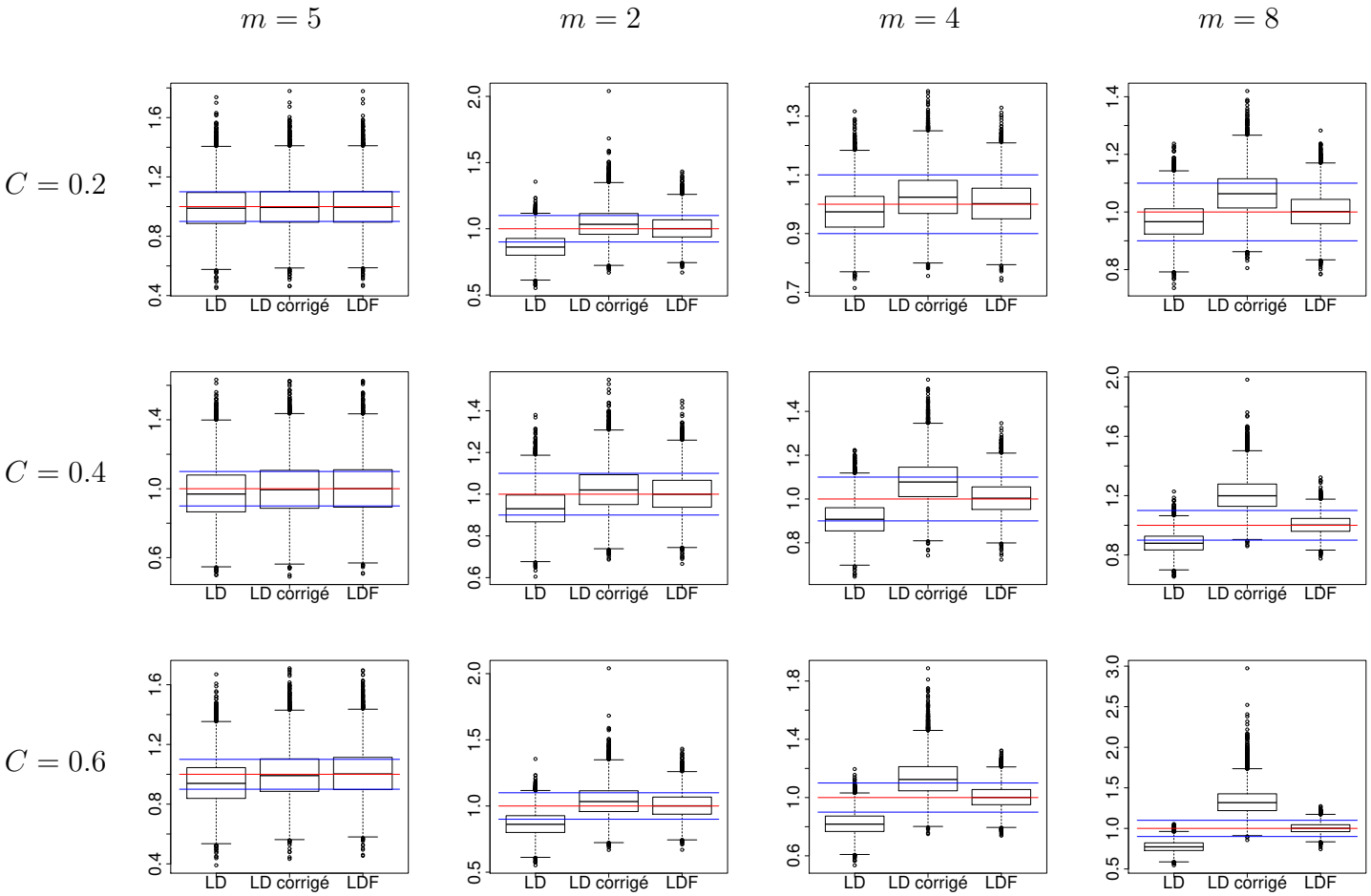

FIGURE 4.5 – **Estimations par la méthode ML en prenant en compte ou non les fluctuations du nombre final de cellules.** Pour chacun des 12 couples de paramètres $m = (0.5, 2, 4, 8)$ (colonnes) et $C = (0.2, 0.4, 0.6)$ (lignes), 10^4 échantillons de taille 100 de la loi $LDF(\pi, 1, 0, K, 1)$ ont été simulés, où $\pi = m/\kappa$ (avec $\kappa = 10^9$) et K est la fonction de répartition de la loi Log-Normale ajustée à la moyenne κ et au coefficient de variation C. Chaque boxplot représente la distribution des 10^4 rapports $\hat{\pi}_{ML}/\pi$ obtenus avec $C = 0$ (gauche), avec la correction (4.8) (centre) et avec le modèle LDF (droite). Dans les trois cas, la fitness ρ est fixée à 1 pour la simulation des échantillons et l'estimation de π.

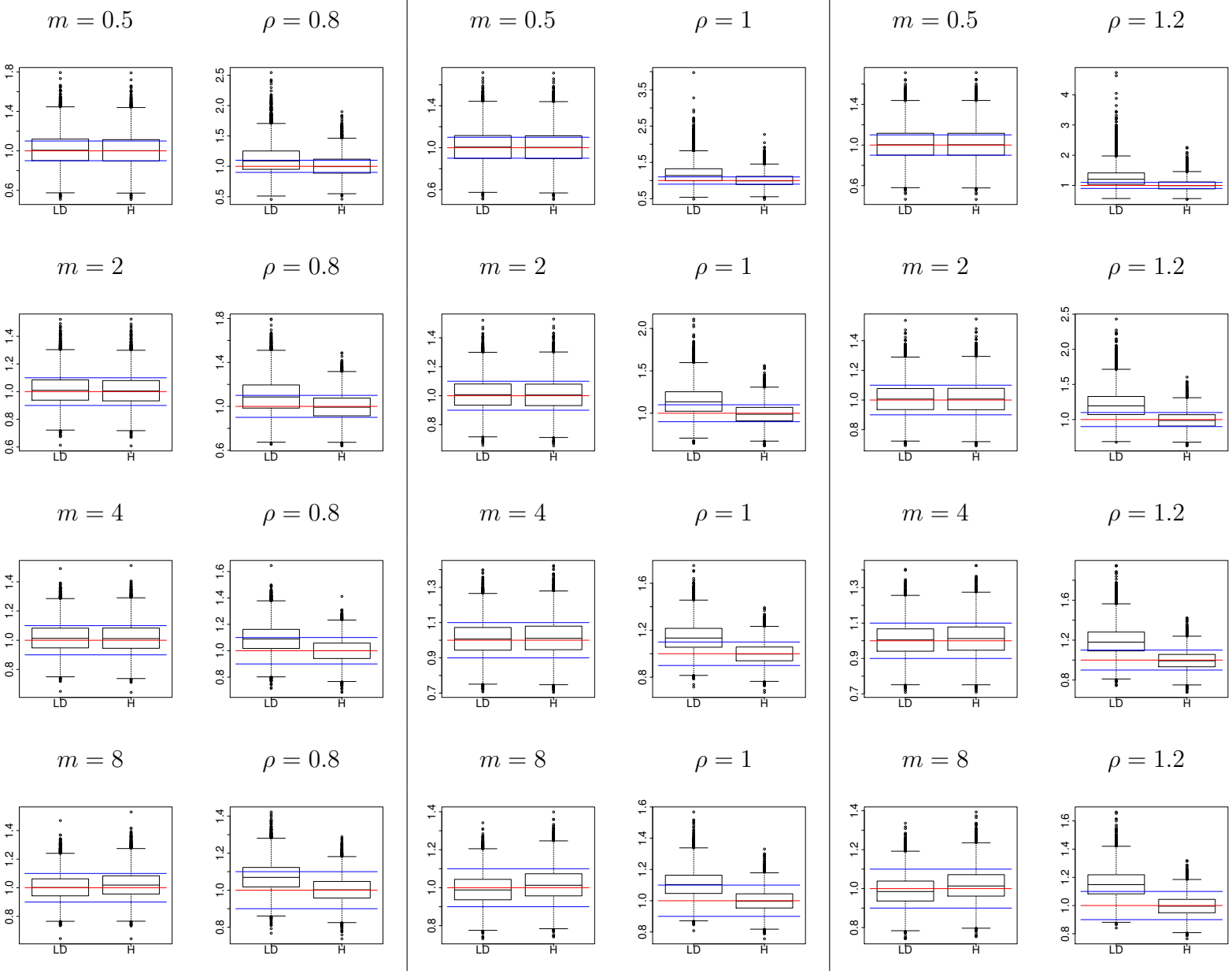

FIGURE 4.6 – **Estimations par la méthode GF sous les modèles *LD* et *H* sur des données simulées sous un modèle *MM*.** Pour chacun des 12 couples de paramètres $m = (0.5, 2, 4, 8)$ (colonnes) et $\rho = (0.8, 1, 1.2)$ (lignes), 10^4 échantillons de taille 100 de la loi $MM(m, \rho, 0, F, 1)$ ont été simulés, où F est la fonction de répartition de la loi Log-Normale ajustée aux données de KELLY et RAHN [40]. Pour chaque colonne, les deux premiers boxplots représentent la cellules des 10^4 rapports $\hat{m}_{GF}/m$ obtenus avec le modèle LD (gauche) et le modèle H (droite) ; les deux derniers boxplots représentent la cellules des 10^4 rapports $\hat{\rho}_{GF}/\rho$ obtenus avec le modèle LD (gauche) et le modèle H (droite).

Nous pouvons constater avec ces résultats visuels que seule l'estimation de m ne semble pas affectée par le choix du modèle de croissance. En effet, les biais d'estimations de m sont en majorité inférieurs à 10%, que ce soit en utilisant le modèle LD ou le modèle H. Seule l'estimation de ρ semble être sensible au choix du modèle : le modèle LD surestime ρ, avec une plus grande variance des estimations. La variance semble augmenter lorsque m augmente. Le modèle H estime très correctement le paramètre de fitness ρ. À partir de ces seuls résultats visuels, nous pourrions considérer que le modèle H devrait être systématiquement employé. Pour mettre cette possibilité à l'épreuve, nous avons effectué la même expérience mais en simulant cette fois des échantillons distribués selon des modèles LD. Les résultats sont illustrés par la Figure 4.7. Nous constatons alors que le modèle H sous-estime systématiquement le paramètre ρ. Le biais négatif ainsi induit augmente avec m. Cependant, la variance des estimations obtenues avec le modèle H est plus faible que celle des estimations obtenues avec le modèle LD. La Figure 4.8 illustre la même expérience, mais en simulant cette fois des échantillons distribués selon des modèles H. Nous constatons que les observations pour la Figure 4.6 sont toujours valables : le modèle LD estime correctement le paramètre m mais surestime ρ. À partir de ces trois figures, nous pourrions interpréter les modèles LD et H comme des cas extrêmes dans le sens où :

— les deux modèles estiment correctement m ;
— le modèle LD surestime systématiquement ρ lorsque le modèle de croissance est différent ;
— le modèle H sous-estime ρ lorsque le modèle de croissance est exponentiel ;
— le modèle H estime correctement ρ lorsque le modèle de croissance n'est pas exponentiel ;
— la variance du modèle LD est systématiquement plus grande que celle du modèle H.

Cependant, si les estimations de m et ρ sont calculées par la méthode ML, les observations sont différentes. La figure 4.9 expose les résultats obtenus dans ce cas. Nous constatons tout d'abord qu'avec le modèle LD, le comportement des estimations par la méthode ML est le même que par la méthode GF : l'estimation de m est correcte, tandis qu'il y a une surestimation de ρ. Il est par ailleurs intéressant de remarquer que les ordres de grandeurs des biais sont similaires à ceux de la méthode GF : l'initialisation par la méthode GF pour les modèles LD semble bien pertinente. Nous observons cependant un comportement très différent lorsque le modèle H est considéré. En effet, les estimations de ρ ne sont correctes que lorsque $m \geqslant 4$, et il est compliqué de mettre en évidence un comportement spécifique des estimations de m. Cela pourrait mettre en évidence le fait que le modèle H est plus sensible à la winsorisation que le modèle LD, or cette manipulation n'a un réel impact sur un échantillon que lorsque m est élevé et/ou ρ est faible.

Jusqu'à présent, si la population observée était le résultat de la dilution d'une population plus grande, l'estimation de m était corrigée en appliquant l'équation (41) de Stewart et al. [86]. Nous avons montré dans la sous-partie 4.1.1 que cette correction

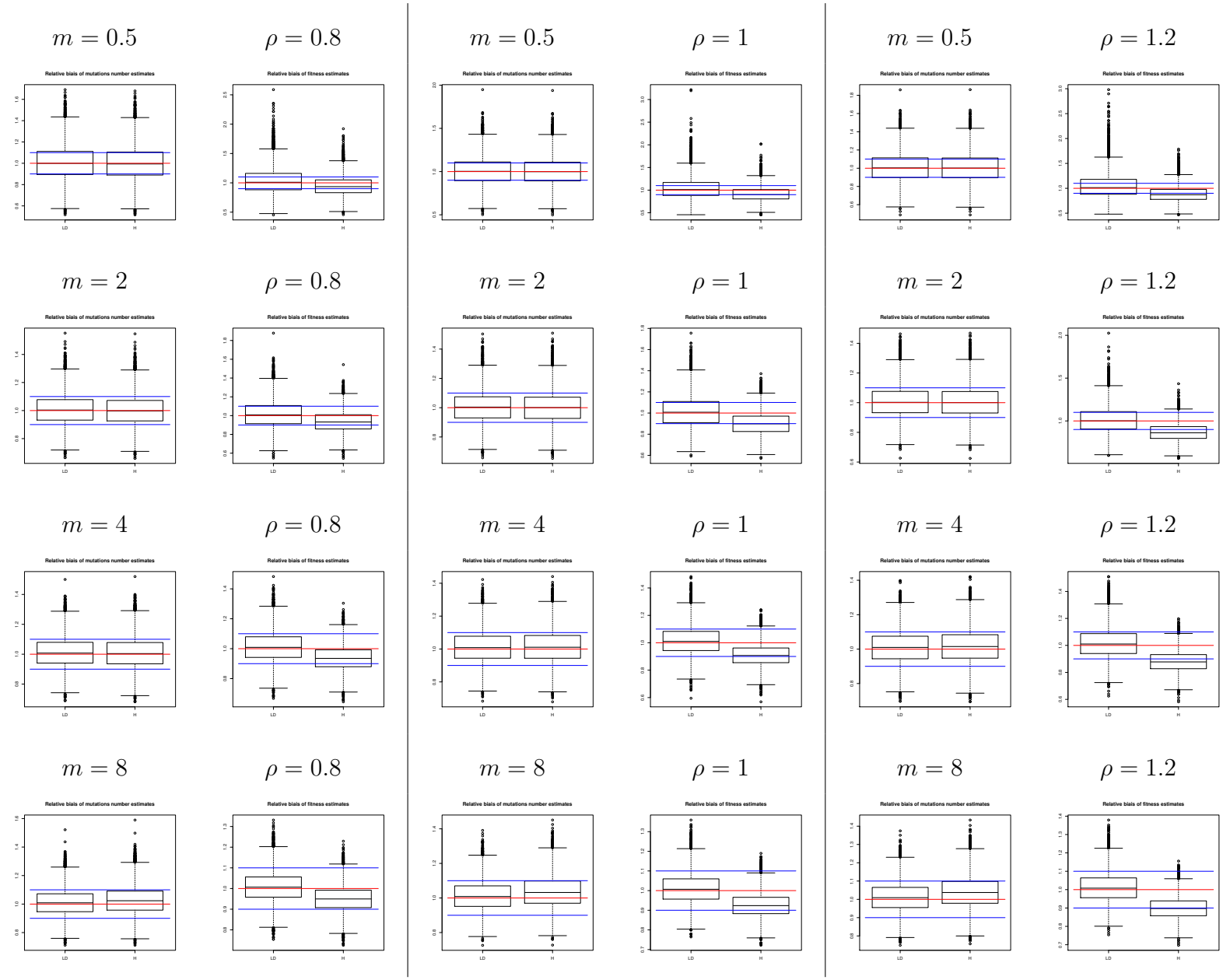

FIGURE 4.7 – **Estimations par la méthode GF sous les modèles *LD* et *H* sur des données simulées sous un modèle *LD*.** Pour chacun des 12 couples de paramètres $m = (0.5, 2, 4, 8)$ (colonnes) et $\rho = (0.8, 1, 1.2)$ (lignes), 10^4 échantillons de taille 100 de la loi $LD(m, \rho, 0, 1)$ ont été simulés. Pour chaque colonne, les deux premiers boxplots représentent la cellules des 10^4 rapports $\hat{m}_{GF}/m$ obtenus avec le modèle LD (gauche) et le modèle H (droite) ; les deux derniers boxplots représentent la cellules des 10^4 rapports $\hat{\rho}_{GF}/\rho$ obtenus avec le modèle LD (gauche) et le modèle H (droite).

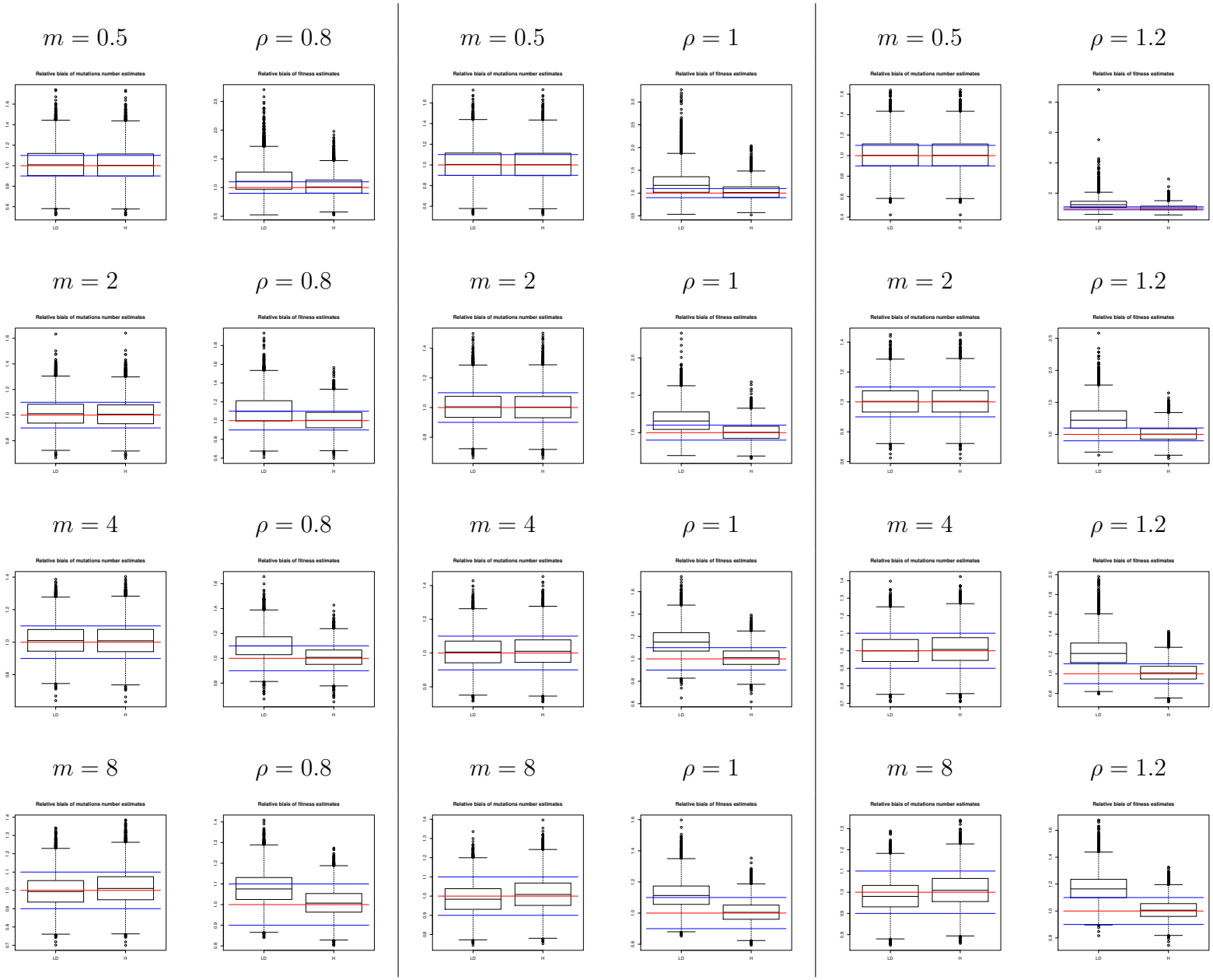

FIGURE 4.8 – **Estimations par la méthode GF sous les modèles *LD* et *H* sur des données simulées sous un modèle *H*.** Pour chacun des 12 couples de paramètres $m = (0.5, 2, 4, 8)$ (colonnes) et $\rho = (0.8, 1, 1.2)$ (lignes), 10^4 échantillons de taille 100 de la loi $H(m, \rho, 0, 1)$ ont été simulés. Pour chaque colonne, les deux premiers boxplots représentent la cellules des 10^4 rapports $\hat{m}_{GF}/m$ obtenus avec le modèle LD (gauche) et le modèle H (droite) ; les deux derniers boxplots représentent la cellules des 10^4 rapports $\hat{\rho}_{GF}/\rho$ obtenus avec le modèle LD (gauche) et le modèle H (droite).

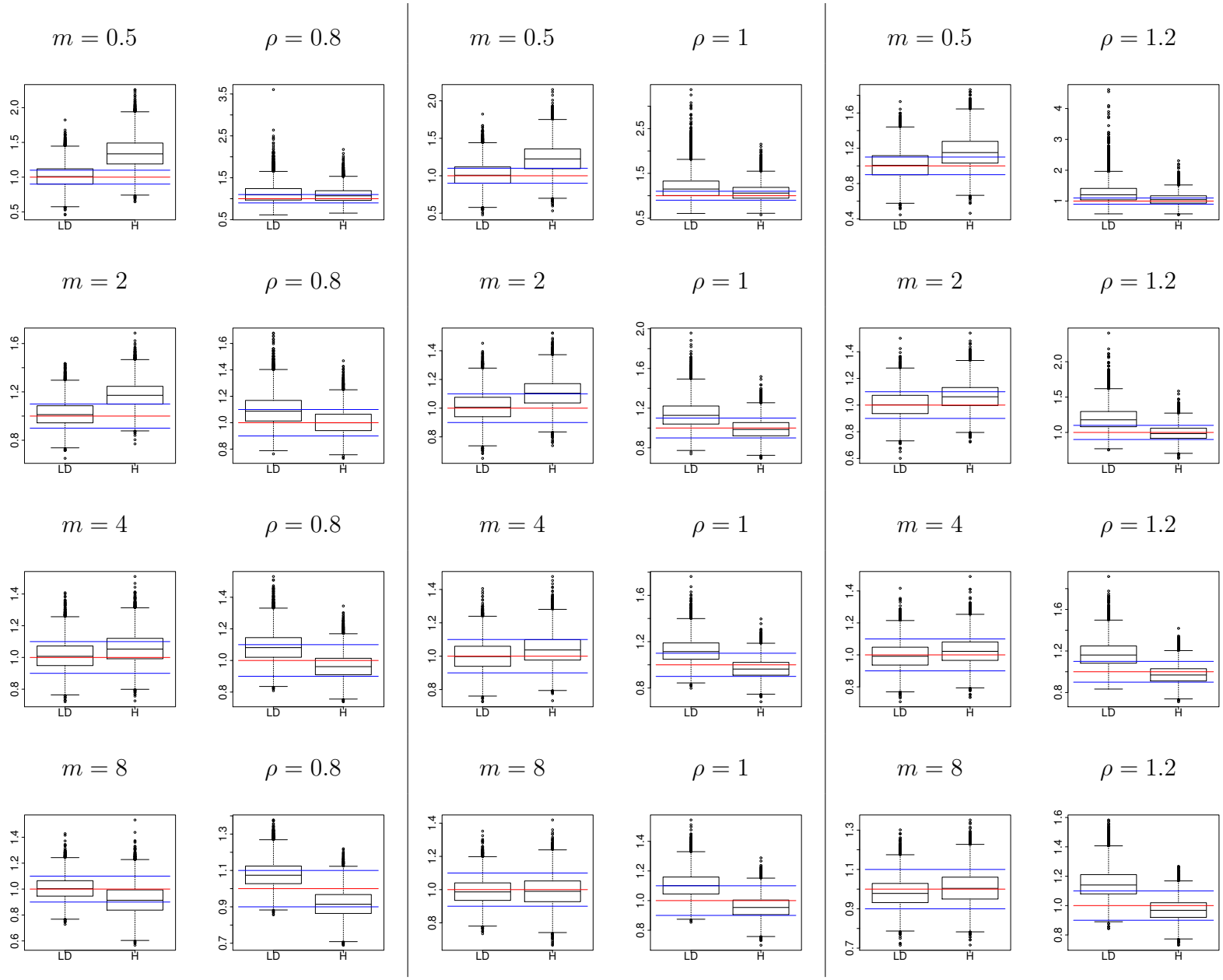

FIGURE 4.9 – **Estimations par la méthode ML sous les modèles *LD* et *H* sur des données simulées sous un modèle *MM*.** Pour chacun des 12 couples de paramètres $m = (0.5, 2, 4, 8)$ (colonnes) et $\rho = (0.8, 1, 1.2)$ (lignes), 10^4 échantillons de taille 100 de la loi $MM(m, \rho, 0, F, 1)$ ont été simulés, où F est la fonction de répartition de la loi Log-Normale ajustée aux données de KELLY et RAHN [40]. Pour chaque colonne, les deux premiers boxplots représentent la cellules des 10^4 rapports $\hat{m}_{GF}/m$ obtenus avec le modèle LD (gauche) et le modèle H (droite) ; les deux derniers boxplots représentent la cellules des 10^4 rapports $\hat{\rho}_{GF}/\rho$ obtenus avec le modèle LD (gauche) et le modèle H (droite).

ne pouvait être utilisée que dans le cas très particulier des modèles $LD(m, 1, 0, \zeta)$. Les Figures 4.11 et 4.10 illustrent l'influence du paramètre de dilution sur les estimations de m et ρ. Les estimations de m sont calculées de trois manières différentes :

1. par la méthode GF avec $\zeta = 1$ (boxplots de gauche) ;
2. en appliquant la correction de Stewart et al. [86] (boxplots centraux) ;
3. par l'extension de la méthode GF présentée dans la sous-partie 4.1.3 (boxplots de droite).

Les estimations de ρ sont calculées de deux manières différentes :

1. par la méthode GF avec $\zeta = 1$ (boxplots de gauche) ;
2. par l'extension de la méthode GF présentée dans la sous-partie 4.1.3 (boxplots de droite).

De manière générale, ne pas prendre en compte la dilution cause forcément une sous-estimation du nombre de mutations m. Ces figures illustrent également le fait que la correction de Stewart et al. [86] n'est pertinente que lorsque $\rho = 1$. Notons cependant que la variance des estimations obtenues par cette correction est plus faible que celle des estimations obtenues par la méthode GF. Cependant, elle donne une sur-estimation de m lorsque $\rho < 1$, et une sous-estimation de m pour $\rho > 1$. Ignorer le paramètre de dilution conduit également à une sur-estimation de la fitness ρ. De plus, il est plus compliqué pour la méthode GF d'estimer ρ, car les jackpots sont moins forts ou moins présents (en particulier lorsque m est faible). De fait, il y a plus de valeurs aberrantes lorsque l'on fixe ζ à 1. Les écarts observés sont d'autant plus grands que le paramètre ζ est faible.

Nous allons à présent nous intéresser aux biais induits lorsque les estimations sont calculées sous des modèles LD ou H alors que les échantillons sont simulés selon des modèles LDI. Nous supposerons que $\delta = 0$ et $\zeta = 1$. Plaçons nous dans le cas où la fonction de répartition F_μ est définie par (3.6), où f est définie pour tout $t \geqslant 0$ par

$$f(t) = \frac{f_\infty}{1 + \left(\frac{f_\infty}{f_0} - 1\right) \mathrm{e}^{-t}}, \tag{4.12}$$

où f_0 et f_∞ correspondent au nombre initial de cellules et à la taille maximale que la population peut atteindre. Alors :

$$\eta_\mu(s,t) = \log\left(\frac{f(t)}{f(s)}\right),$$

et

$$\eta_{\mu,\infty} = \log\left(\frac{f_\infty}{f_0}\right).$$

Notons que le choix de la fonction f n'a d'importance que pour la simulation. Seul le paramètre $\eta_{\mu,\infty}$ importe dans le cas d'un modèle LDI (théorème 3.3.1). Les figures 4.12

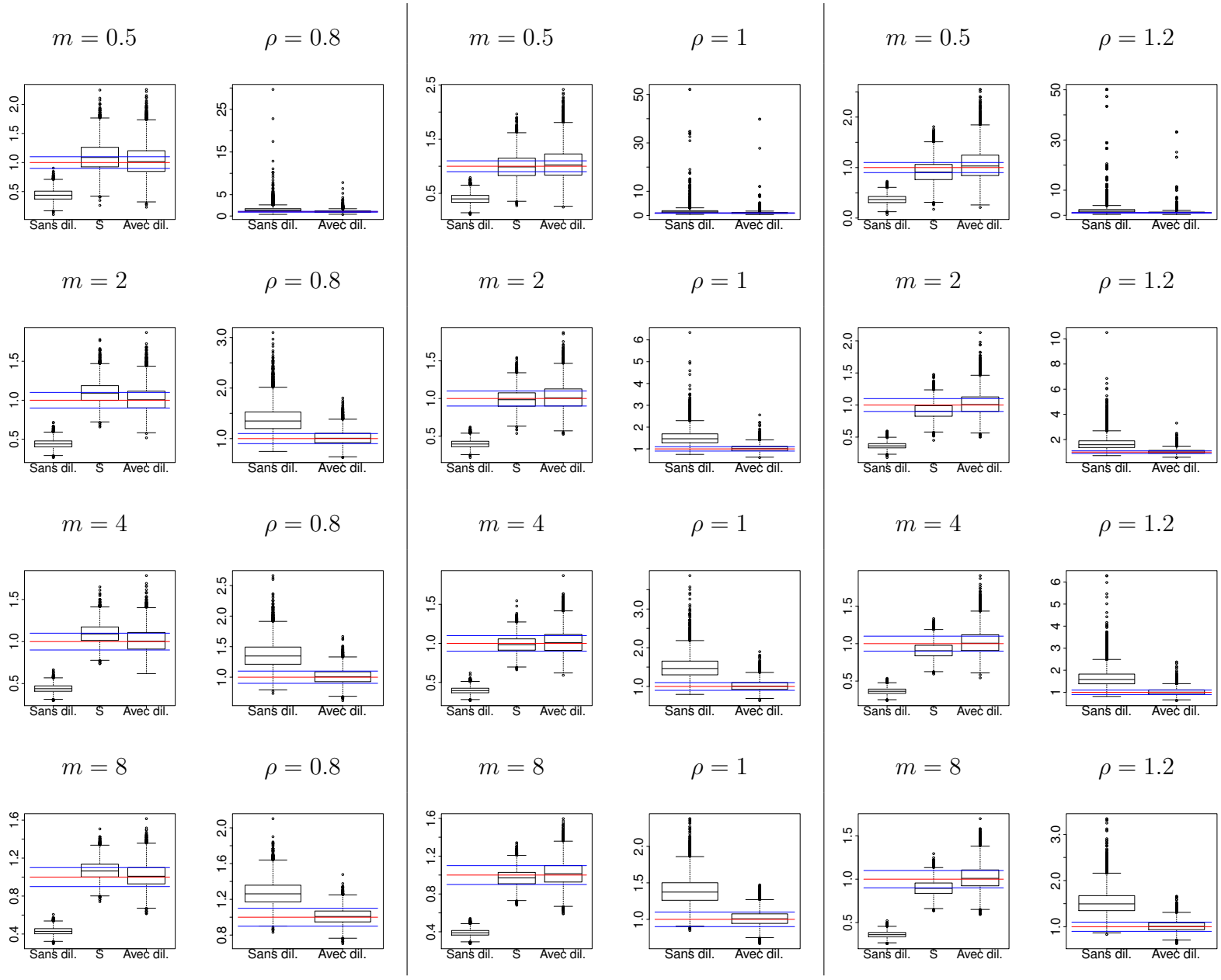

FIGURE 4.10 – **Estimations par la méthode GF en prenant en compte ou non la dilution ($\zeta = 0.2$).** Pour chacun des 12 couples de paramètres $m = (0.5, 2, 4, 8)$ (colonnes) et $\rho = (0.8, 1, 1.2)$ (lignes), 10^4 échantillons de taille 100 de la loi $LD(m, \rho, 0, \zeta)$ ont été simulés, avec $\zeta = 0.2$. Pour chaque colonne, les trois premiers boxplots représentent la cellules des 10^4 rapports $\hat{m}_{GF}/m$ obtenus avec $\zeta = 1$ (gauche), en appliquant la correction de STEWART et al. [86] (centre), et par l'extension de la méthode GF (droite) ; les deux derniers boxplots représentent la cellules des 10^4 rapports $\hat{\rho}_{GF}/\rho$ obtenus avec $\zeta = 1$ (gauche), et via l'extension de la méthode GF (droite).

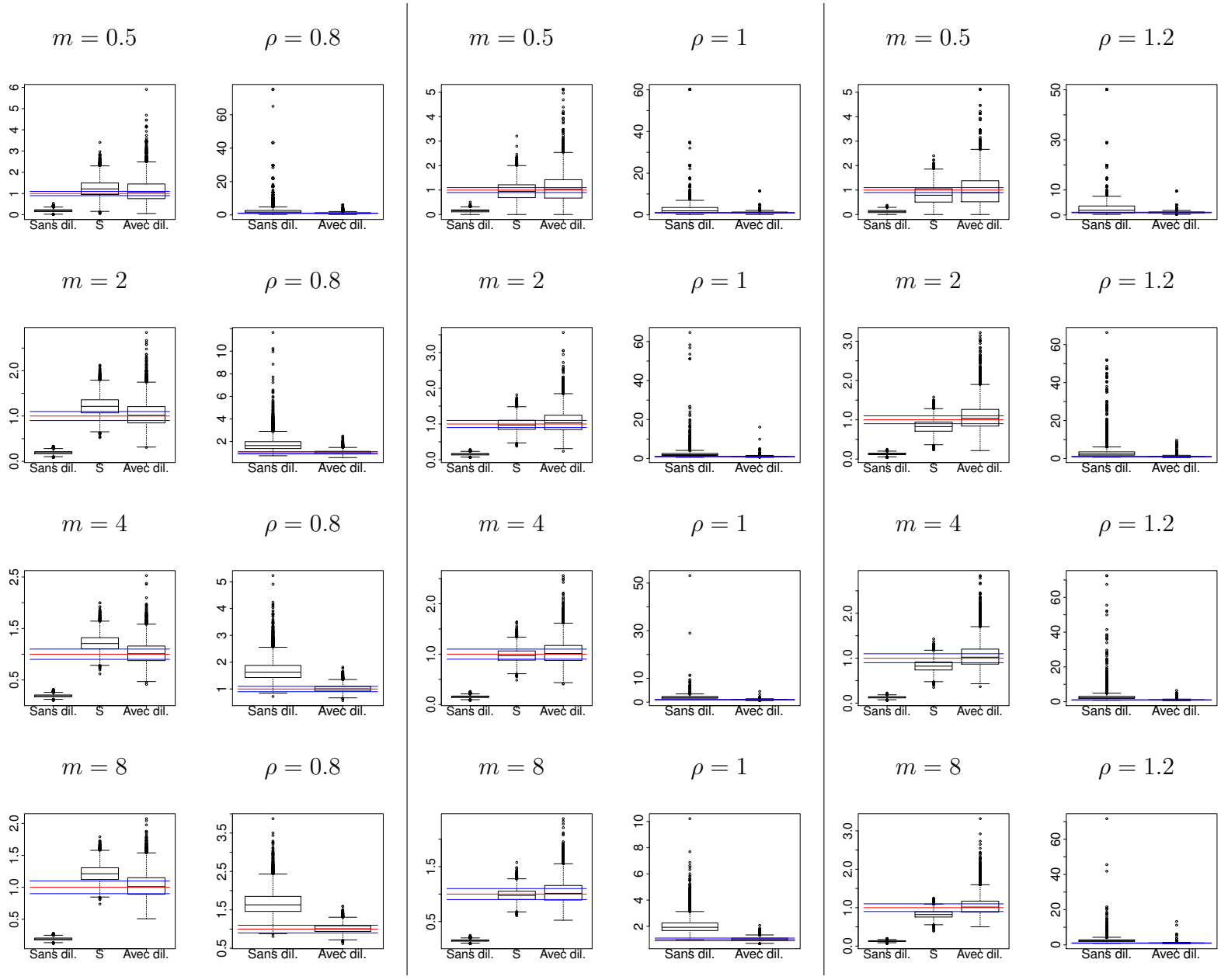

FIGURE 4.11 – **Estimations par la méthode GF en prenant en compte ou non la dilution ($\zeta = 0.05$).** Pour chacun des 12 couples de paramètres $m = (0.5, 2, 4, 8)$ (colonnes) et $\rho = (0.8, 1, 1.2)$ (lignes), 10^4 échantillons de taille 100 de la loi $LD(m, \rho, 0, \zeta)$ ont été simulés, avec $\zeta = 0.05$. Pour chaque colonne, les trois premiers boxplots représentent la cellules des 10^4 rapports $\hat{m}_{GF}/m$ obtenus avec $\zeta = 1$ (gauche), en appliquant la correction de STEWART et al. [86] (centre), et par l'extension de la méthode GF (droite) ; les deux derniers boxplots représentent la cellules des 10^4 rapports $\hat{\rho}_{GF}/\rho$ obtenus avec $\zeta = 1$ (gauche), et via l'extension de la méthode GF (droite).

et 4.13 illustrent les biais d'estimations de m et ρ selon la valeur de $\eta_{\mu,\infty}$. Les estimations de m et ρ sont calculées de trois manières différentes :

1. par la méthode GF avec le modèle LD (boxplots de gauche) ;
2. par la méthode GF avec le modèle H (boxplots centraux) ;
3. par la méthode GF avec le modèle LDI (boxplots de droite).

Nous pouvons constater avec ces résultats visuels que les modèles LD et H estiment correctement m, tant que sa valeur réelle n'est pas trop grande : pour les valeurs $m = 4$, le paramètre est légèrement surestimé par les deux modèles. Le biais d'estimation est d'autant plus fort que les décomptes de mutantes sont élevés et nombreux, par exemple lorsque $m = 8$ et $\rho = 0.8$. Nous observons ensuite que le modèle LD surestime le paramètre ρ. Le biais est d'autant plus fort que m est grand et ρ est faible. En l'occurrence, la distinction entre le modèle LD et le modèle LDI est légère lorsque ρ est plus grand que 1 : voir par exemple les cas $m \leqslant 2$ et $\rho = 1.2$. Cette similarité est d'autant plus remarquable que dans le cas opposé, comme lorsque $m = 8$ et $\rho = 0.8$, un quart seulement des estimations de ρ ont un biais relatif inférieur à 20%. Le comportement des estimations de ρ obtenues par le modèle H est beaucoup plus compliqué à cerner.

La même expérience a été effectuée avec $\eta_{\mu,\infty} = \log(10^4)$: dans ce cas il n'y a pas de distinction entre les modèles LD et LDI. De fait, nous retrouvons le comportement observé avec la figure 4.7 en ce qui concerne le modèle H. En pratique, le nombre initial de cellules est de l'ordre de 10^3 à 10^5 et le nombre final de cellules est de l'ordre de 10^8 à 10^9. Au vu de ces ordres de grandeurs, supposer en pratique que la croissance est exponentielle au lieu d'être logistique induit un biais négligeable, à condition que l'hypothèse (3.33) soit respectée.

À partir de ces expériences, nous pouvons classer les sources de biais par ordre « d'impact » en fonction des erreurs d'estimations observées dans ce chapitre :

1. Prise en compte des fluctuations des nombres finaux : les données de type $T2$ sont en réalité peu répandues. En conséquence, dans la majorité des cas, la correction (4.8) sera utilisée. Or nous avons constaté que pour des valeurs élevées du coefficient de variation C ou de la valeur réelle de π, cette correction peut mener à une surestimation de π avec un biais du même ordre que sans correction (Figure 4.5).
2. Prise en compte de la dilution : le biais causé en ignorant simplement la dilution peut être très grand (Figures 4.10 et 4.11). De manière générale, il est préférable d'utiliser la méthode GF pour estimer m et ρ. Il est cependant envisageable, lorsque l'estimation de ρ est très proche de 1, d'estimer m par la méthode GF en ignorant la dilution puis d'appliquer la correction de Stewart et al. [86] : cette dernière donne dans ce cas des estimations plus précises que la méthode GF dans le sens où la variance des estimations est bien plus faible.
3. Prise en compte des morts cellulaires : en pratique, le paramètre de mort δ est inférieur à 0.3. Nous avons constaté que pour des valeurs proches de ce seuil, le biais

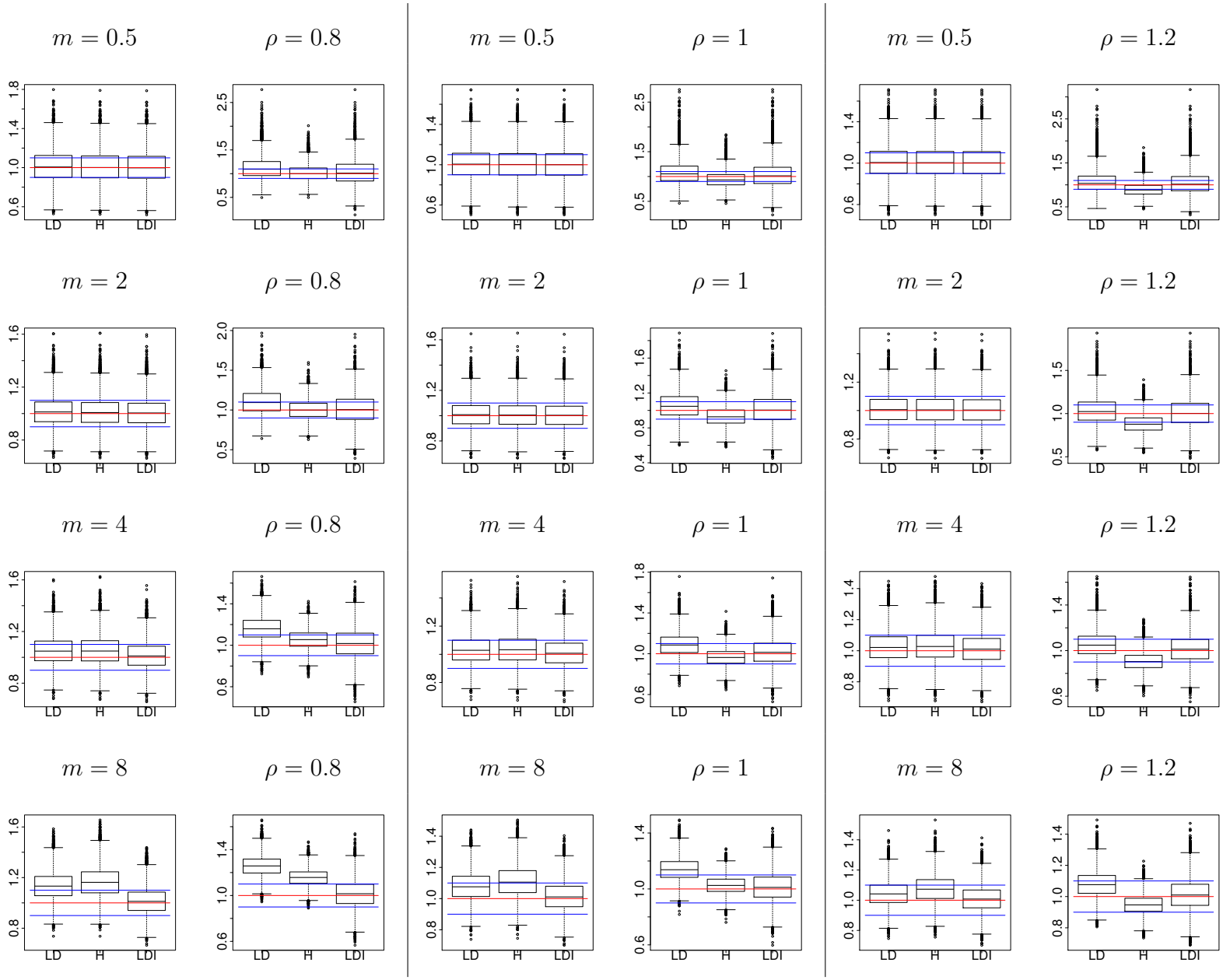

FIGURE 4.12 – **Estimations par la méthode GF sous les modèles *LD*, *H* et *LDI* sur des données simulées sous un modèle *LDI* ($\eta_{\mu,\infty} = \log(100)$).** Pour chacun des 12 couples de paramètres $m = (0.5, 2, 4, 8)$ (colonnes) et $\rho = (0.8, 1, 1.2)$ (lignes), 10^4 échantillons de taille 100 de la loi $LDI(m, \rho, 0, \eta_{\mu,\infty}, 1)$ ont été simulés, où F est défini par (3.6) telle que $\eta_{\mu,\infty} = \log(100)$. Pour chaque colonne, les trois premiers boxplots représentent la cellules des 10^4 rapports $\hat{m}_{GF}/m$ obtenus avec le modèle LD (gauche), H (centre), et LDI (droite) ; les trois derniers boxplots représentent la cellules des 10^4 rapports $\hat{\rho}_{GF}/\rho$ obtenus avec le modèle LD (gauche), H (centre), et LDI (droite).

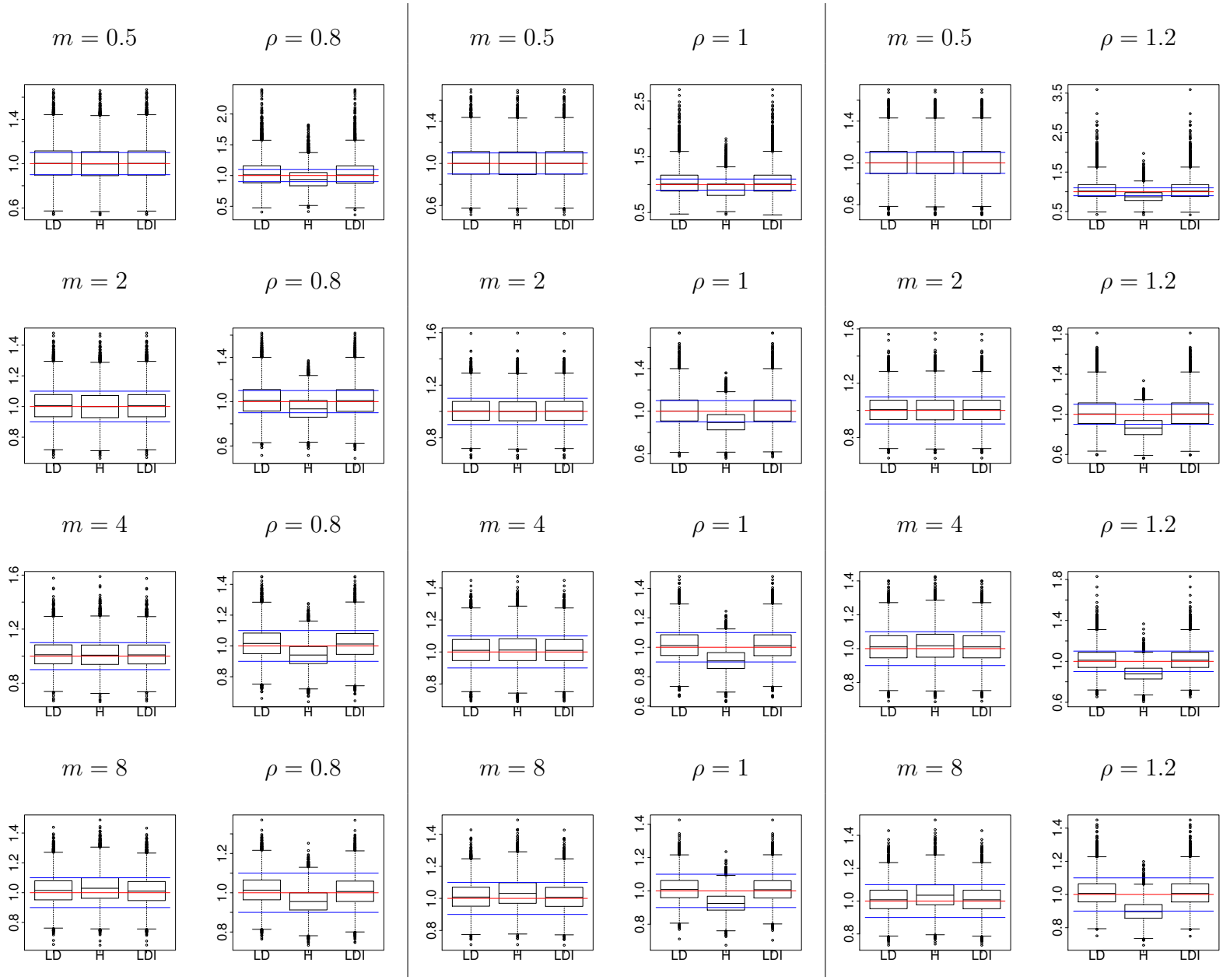

FIGURE 4.13 – **Estimations par la méthode GF sous les modèles *LD*, *H* et *LDI* sur des données simulées sous un modèle *LDI* ($\eta_{\mu,\infty} = \log(10^4)$).** Pour chacun des 12 couples de paramètres $m = (0.5, 2, 4, 8)$ (colonnes) et $\rho = (0.8, 1, 1.2)$ (lignes), 10^4 échantillons de taille 100 de la loi $LDI(m, \rho, 0, \eta_{\mu,\infty}, 1)$ ont été simulés, où F est défini par (3.6) telle que $\eta_{\mu,\infty} = \log(10^4)$. Pour chaque colonne, les trois premiers boxplots représentent la cellules des 10^4 rapports $\hat{m}_{GF}/m$ obtenus avec le modèle LD (gauche), H (centre), et LDI (droite) ; les trois derniers boxplots représentent la cellules des 10^4 rapports $\hat{\rho}_{GF}/\rho$ obtenus avec le modèle LD (gauche), H (centre), et LDI (droite).

d'estimation peut être non négligeable (Figures 4.3 et 4.4). La difficulté réside dans le fait que la valeur de δ est en pratique inconnue, et qu'aucune méthode ne permet pour le moment d'en calculer des estimations fiables. La marche à suivre serait alors de calculer les estimations de m et ρ en faisant varier δ.

4. Choix du modèle de croissance : Les simulations tendent à montrer que le modèle LD surestime systématiquement ρ si les durées des vies des cellules mutantes sont *i.i.d.* mais non exponentielles, tandis que les estimations obtenues avec le modèle H sont correctes, tant que les durées des vies des cellules mutantes sont *i.i.d.* et non exponentielles (Figures 4.6, 4.7 et 4.8). Cependant, en pratique, la croissance de la population suit une courbe de type logistique, comme par exemple (4.12). Étant donnés les ordres de grandeur qui sont en jeu en pratique, les résultats illustrés par la figure 4.13 sont plus pertinents que ceux de la figure 4.12 : les simulations montrent que le modèle LD devrait donner en pratique de meilleurs résultats que le modèle H. Cependant, cela nécessite que l'hypothèse de proportionnalité (3.33) soit vérifiée.

Chapitre 5

Applications sur des jeux de données réelles

Nous allons à présent appliquer les méthodes d'estimation décrites dans le chapitre 4 à des jeux de données réelles. Chaque partie de ce chapitre est dédiée à l'étude d'un jeu de données en particulier. Certaines données sont issues de publications dans lesquelles sont exposés des résultats de tests de fluctuations. Nous reprendrons ces tests afin de comparer les nouvelles estimations avec celles des auteurs. Dans ce cas, nous comparerons dans un premier temps les estimations obtenues par les auteurs, à celles obtenues avec les trois méthodes P0, GF et ML sous les mêmes hypothèses de modélisation qu'eux. Nous ferons varier ensuite ces hypothèses, et calculerons à nouveau les estimations souhaitées avec les méthodes d'intérêt. Nous éprouverons ensuite la pertinence des estimations initiales à l'aide d'intervalles de confiance de niveau 95%. Dans chacune des parties, des extraits des scripts R utilisés seront donnés afin d'illustrer le fonctionnement du package **flan**.

Rappelons que les intervalles de confiance associés aux méthodes considérées dans cette thèse sont calculés en se basant sur la propriété de normalité asymptotique des estimateurs considérés. Si la plupart des échantillons considérés ici font une taille suffisamment grande pour exploiter cette propriété, ce n'est pas le cas pour certains d'entre eux et le calcul de certains intervalles pourrait être remis en cause.

5.1 Données de Boe et al.

Nous nous intéressons tout d'abord aux données publiées par Boe et al. [11, tab. 4]. Cet article porte sur l'étude de l'apparition d'une résistance chez *Escherichia Coli* à un certain antibiotique. Le jeu de données constitué de 23 échantillons de décomptes finaux de mutantes est disponible dans le package **flan** sous le nom de la variable `boeal`. Dans cet article, les auteurs ont décidé de rassembler les 23 échantillons, et ont effectué leurs tests de fluctuations sur l'échantillon résultant. Les différentes estimations de m calculées par les auteurs sont répertoriées dans leur table 3. Ils utilisent la méthode P0, et une méthode

du maximum de vraisemblance. Notons que la vraisemblance de l'échantillon y est calculée en imposant le fait qu'il y ait 8 classes (voir p. 2783). Les estimations sont calculées selon les modèles LD et H. Il n'y a pas d'autres hypothèses de modélisation : en d'autres termes la fitness ρ et le paramètre de dilution ζ valent 1 et le paramètre de mort δ vaut 0. La table 5.1 compare les estimations obtenues par les auteurs avec celles des méthodes GF et ML sous les mêmes conditions de modélisation. Pour chaque méthode, les intervalles de confiance de niveau 95% sont également fournis. Ci-dessous apparaît par exemple la partie du script R calculant l'estimation et l'intervalle de confiance correspondant à la méthode GF sous le modèle H :

```
# Extraction des échantillons
B <- unlist(boeal)

# Calcul de l'estimation du nombre de mutations
# et de son intervalle de confiance de niveau 0.95 par la méthode GF
# sous le modèle "H" et en supposant une fitness égale à 1
GFtest <- flan.test(mc = B, fitness = 1, method = "GF", model = "H")

# Extraction de l'estimation
mut.est <- GFtest$estimate

## mutations
##   0.83238

# Extraction de l'intervalle de confiance
mut.IC <- GFtest$conf.int

##      mutations
## bInf 0.7658623
## bSup 0.8988978
## attr(,"conf.level")
## [1] 0.95
```

En supposant que $\rho = 1$, l'estimation de m obtenue par les auteurs pour le modèle LD est incluse dans les intervalles de confiance des trois méthodes. Il est également remarquable que l'estimation des auteurs soit très proche de celles obtenues par la méthode ML. Au contraire, l'estimation de m obtenue par le modèle H n'est incluse que dans l'intervalle de confiance de la méthode GF (qui est l'intervalle le plus grand des trois). Par ailleurs, l'estimation par la méthode ML sous le modèle H est très différente de celle obtenue par les auteurs. Le modèle H semble donc être plus sensible que le modèle LD au fait de séparer les décomptes de mutantes en 8 classes. La table 5.2 compare les estimations obtenues par les auteurs avec celles des méthodes P0, GF et ML, en estimant

	LD		***H***	
	Estimations de m	**IC (95%) de m**	**Estimations de m**	**IC (95%) de m**
Auteurs	0.737	/	0.770	/
P0	0.710	[0.650 ; 0.750]	0.710	[0.650 ; 0.750]
ML	0.737	[0.680 ; 0.794]	0.941	[0.892 ; 0.989]
GF	0.781	[0.716 ; 0.846]	0.832	[0.766 ; 0.899]

TABLE 5.1 – **Estimations de m avec les données de BOE et al. [11, tab. 4]**. Les estimations de m par les méthodes P0, ML et GF sont calculées sous les modèles LD et H pour une fitness ρ et un paramètre de dilution ζ fixés à 1 et un paramètre de mort δ nul. Pour chaque méthode, les intervalles de confiance de niveau 95% de m sont également donnés.

cette fois-ci le paramètre ρ. Les estimations obtenues pour un paramètre de mort δ égal à 0.1 sont également exposées (3 dernières lignes). Le script R utilisé est le même que précédemment en ajoutant dans la fonction `flan.test` le paramètre `death = 0.1` et en retirant `fitness = 1`.

Nous constatons dans un premier temps que la valeur $\rho = 1$ n'est comprise dans aucun des intervalles de confiance, quelle que soit la méthode considérée ou le modèle sous lequel on se place. En particulier, nous avons constaté dans la sous-section 4.3.2 que la méthode GF sous le modèle H estime correctement ρ tant que les durées de vie des mutantes sont *i.i.d.* et non-exponentielles. Donc, si cette hypothèse sur les durées de vie est vérifiée, il n'y a aucune raison de considérer que la fitness vaut 1. Si au contraire les durées de vie sont bien exponentielles, les intervalles de confiance obtenus par la méthode GF sous le modèle LD nous permettent également d'affirmer que la fitness est différente de 1. Nous constatons également que l'estimation de m sous le modèle LD obtenue par les auteurs est incluse dans tous les intervalles de confiance, quelle que soit la méthode considérée. Cette observation concorde avec celles de la figure 4.1 : les trois méthodes sont équivalentes pour des faibles valeurs de m.

5.2 Données de DAVID

Nous nous intéressons à présent aux données publiées par DAVID [17]. Cet article porte sur l'étude de l'apparition d'une résistance chez *Mycobacterium tuberculosis* à un certain antibiotique. Deux jeux sont disponibles dans cet article. Le premier (table 1) est constitué de 10 échantillons de taille variable de décomptes de mutantes. Précisons que dans l'article, l'auteur a décider de classer les décomptes selon des plages (par exemple, nombre de décomptes compris entre 11 et 20). Nous avons arbitrairement choisi de remplacer ces

	LD			***H***		
	Estimations de m et ρ	**IC (95%) de m**	**IC (95%) de ρ**	**Estimations de m et ρ**	**IC (95%) de m**	**IC (95%) de ρ**
Auteurs	(0.737,1)	/	/	(0.770,1)	/	/
P0	(0.710,0.837)	[0.650 ; 0.750]	[0.758 ; 0.916]	(0.710,0.822)	[0.650 ; 0.750]	[0.771 ; 0.874]
ML	(0.714,0.838)	[0.655 ; 0.772]	[0.757 ; 0.919]	(0.920,0.845)	[0.869 ; 0.971]	[0.791 ; 0.900]
GF	(0.711,0.821)	[0.652 ; 0.770]	[0.736 ; 0.906]	(0.707,0.758)	[0.648 ; 0.766]	[0.691 ; 0.824]
P0 ($\delta = 0.1$)	(0.756,0.863)	[0.695 ; 0.821]	[0.781 ; 0.945]	(0.756,0.822)	[0.695 ; 0.821]	[0.841 ; 0.916]
ML ($\delta = 0.1$)	(0.760,0.855)	[0.696 ; 0.823]	[0.772 ; 0.938]	(0.851,0.885)	[0.792 ; 0.909]	[0.848 ; 0.923]
GF ($\delta = 0.1$)	(0.757,0.842)	[0.695 ; 0.819]	[0.753 ; 0.931]	(0.755,0.787)	[0.693 ; 0.817]	[0.716 ; 0.857]

Table 5.2 – **Estimations de m avec les données de Boe et al. [11, tab.4]**. Les estimations de m et ρ par les méthodes P0, ML et GF sont calculées sous les modèles LD et H pour un paramètre de dilution $\zeta = 1$ et un paramètre de mort $\delta = 0$ (lignes 2 à 4) et $\delta = 0.1$ (lignes 5 à 7). Pour chaque méthode, les intervalles de confiance de niveau 95% de m et ρ sont également donnés.

intervalles par la médiane des bornes (par exemple, 15 pour la plage 11-20). Chaque échantillon est associé à un nombre moyen de cellules. Le deuxième jeu de données (table 2) est un échantillon de taille 10 de couples (décompte de mutantes - nombre total de cellules). Toutes ces données sont disponibles dans le package **flan** dans la variable `david`. L'auteur calcule les estimations de π par la méthode des moyennes de Luria-Delbrück [60, eq. (8)]. Nous avons illustré dans la partie 2.3 le fait que cet estimateur est mauvais et ne devrait pas être utilisé. Aucune hypothèse de modélisation n'est faite : en d'autres termes, la fitness ρ et le paramètre de dilution valent 1 et le paramètre de mort δ est nul. Considérons d'abord le premier jeu de données. Les tables 5.3 et 5.4 comparent les estimations obtenues par l'auteur avec celles des méthodes P0, GF et ML sous les modèles LD et H. Pour chaque méthode, les intervalles de confiance de niveau 95% sont également donnés. Ci-dessous la partie du script R calculant les estimations de la probabilité de mutation pour la méthode GF sous le modèle H :

```
# Extraction des échantillons
D <- david[1:10]

# Calcul des estimations de la probabilité de mutation
# et de son intervalle de confiance de niveau 0.95 par la méthode GF
# sous le modèle "H" avec un paramètre de fitness égal à 1
GFtest <- lapply(D, function(d) flan.test(mc = d$mc, mfn = d$mfn,
                                          fitness = 1,
                                          method = "GF", model = "H"))

# Extraction des estimations
mut.est <- sapply(GFtest, function(res) res$estimate)

##   D1.mutprob   D2.mutprob   D3.mutprob   D4.mutprob   D5.mutprob
## 1.985275e-09 3.163999e-09 1.075708e-09 2.296974e-09 2.572064e-10
##   D6.mutprob   D7.mutprob   D8.mutprob   D9.mutprob  D10.mutprob
## 1.476104e-09 1.643759e-11 1.773973e-11 6.458300e-10 8.276786e-10

# Extraction de l'intervalle de confiance
mut.IC <- sapply(GFtest, function(res) res$conf.int)

##                D1           D2           D3           D4           D5
## bInf 1.233091e-09 2.427310e-09 6.128485e-10 1.871075e-09 4.589897e-11
## bSup 2.737459e-09 3.900687e-09 1.538568e-09 2.722874e-09 4.685138e-10
##                D6            D7            D8           D9           D10
## bInf 1.171687e-09 -3.594182e-11 -8.028789e-12 3.000204e-10 6.265837e-10
## bSup 1.780521e-09  6.881700e-11  4.350825e-11 9.916397e-10 1.028774e-09
```

	Auteur	P0		ML		GF	
Antibio.	$\hat{\pi}(\times 10^{-8})$	$\hat{\pi}(\times 10^{-8})$	**IC (95%) de** $\pi(\times 10^{-8})$	$\hat{\pi}(\times 10^{-8})$	**IC (95%) de** $\pi(\times 10^{-8})$	$\hat{\pi}(\times 10^{-8})$	**IC (95%) de** $\pi(\times 10^{-8})$
Ison.	1.84	1.85	[0.67 ; 3.03]	1.74	[1.07 ; 2.42]	1.78	[1.06 ; 2.50]
	3.50	0.943	[0.514 ; 1.37]	2.62	[2.27 ; 2.98]	2.68	[2.01 ; 3.35]
	1.70	1.17	[0.527 ; 1.82]	1.00	[0.545 ; 1.47]	1.01	[0.555 ; 1.46]
	3.20	0.746	[0.451 ; 1.04]	2.05	[1.81 ; 2.29]	2.09	[1.67 ; 2.50]
Strept.	0.900	0.493	[0.194 ; 0.792]	0.376	[0.0857 ; 0.666]	0.241	[0.035 ; 0.447]
	5.00	0.591	[0.375 ; 0.807]	1.89	[1.68 ; 2.09]	1.38	[1.08 ; 1.69]
Rifam.	0.180	0.0317	[0 ; 0.0937]	0.0309	[0 ; 0.0929]	0.0154	[0 ; 0.0665]
	0.0270	0.0289	[0 ; 0.0572]	0.028	[0 ; 0.0563]	0.0166	[0 ; 0.0418]
Etham.	7.00	0.315	[0.0944 ; 0.536]	0.381	[0.165 ; 0.598]	0.606	[0.267 ; 0.944]
	13.0	0.356	[0.224 ; 0.488]	0.649	[0.522 ; 0.777]	0.776	[0.578 ; 0.975]

Table 5.3 – **Estimations de π avec les données de David [17, tab.1] sous le modèle *LD*.** Les estimations de π par les méthodes P0, ML et GF sont calculées en divisant celles de m obtenues sous le modèle LD par le nombre final de cellules. La fitness ρ et le paramètre de dilution sont fixés à 1 et le paramètre de mort δ est nul. Pour chaque méthode, les intervalles de confiance de niveau 95% de π sont également donnés.

Remarquons que nous aurions pu reprendre la classification des décomptes de mutantes et calculer les estimations par maximum de vraisemblance selon la même approche que Boe et al. [11]. Notons par ailleurs que les données de David [17] correspondent aux nombres de mutantes contenues dans 0.1 mL de solution, alors que les cultures sont contenues dans un volume de 2 mL. L'auteur raisonne alors en « nombre moyen de mutantes par mL » et « nombre total de cellules par mL » lors de l'estimation de la probabilité de mutation par la méthode des moyennes. En particulier, les nombres moyens de mutantes sont obtenues en multipliant ceux obtenus avec les échantillons par 10. Nous savons d'ores et déjà qu'il ne s'agit pas de la manipulation appropriée afin de prendre en compte la dilution (voir partie 3.3.3). Notre but initial dans cette étude étant de montrer que la méthode des moyennes donne ici des estimations parfois aberrantes, nous nous contenterons dans cette première comparaison de multiplier également nos estimations par 10. Nous prendrons en compte la dilution avec la méthode GF par la suite. Ces deux tableaux reflètent

	Auteur	**P0**		**ML**		**GF**	
Antibio.	$\hat{\pi}(\times 10^{-8})$	$\hat{\pi}(\times 10^{-8})$	**IC (95%) de** $\pi(\times 10^{-8})$	$\hat{\pi}(\times 10^{-8})$	**IC (95%) de** $\pi(\times 10^{-8})$	$\hat{\pi}(\times 10^{-8})$	**IC (95%) de** $\pi(\times 10^{-8})$
Ison.	1.84	1.85	[0.67 ; 3.03]	2.29	[1.80 ; 2.78]	1.99	[1.23 ; 2.74]
	3.50	0.943	[0.514 ; 1.37]	2.48	[2.01 ; 2.86]	3.16	[2.43 ; 3.90]
	1.70	1.17	[0.527 ; 1.82]	1.46	[1.15 ; 1.78]	1.08	[0.613 ; 1.54]
	3.20	0.746	[0.451 ; 1.04]	2.12	[1.88 ; 2.37]	2.30	[1.87 ; 2.72]
Strept.	0.900	0.493	[0.194 ; 0.792]	0.362	[0.066 ; 0.658]	0.257	[0.0459 ; 0.469]
	5.00	0.591	[0.375 ; 0.807]	2.07	[1.87 ; 2.27]	1.48	[1.17 ; 1.78]
Rifam.	0.180	0.0317	[0 ; 0.0937]	0.0309	[0 ; 0.0929]	0.0164	[0 ; 0.0688]
	0.0270	0.0289	[0 ; 0.0572]	0.0279	[0 ; 0.0562]	0.0177	[0 ; 0.0435]
Etham.	7.00	0.315	[0.0944 ; 0.536]	1.14	[0.956 ; 1.33]	0.646	[0.300 ; 0.992]
	13.0	0.356	[0.224 ; 0.488]	1.36	[1.25 ; 1.48]	0.828	[0.627 ; 1.03]

TABLE 5.4 – **Estimations de π avec les données de DAVID [17, tab.1] sous le modèle H.** Les estimations de π par les méthodes P0, ML et GF sont calculées en divisant celles de m obtenues sous le modèle H par le nombre final de cellules. La fitness ρ et le paramètre de dilution ζ sont fixés à 1 et le paramètre de mort δ est nul. Pour chaque méthode, les intervalles de confiance de niveau 95% de π sont également donnés.

très bien les résultats empiriques de la table 2.3 au début de cette thèse : la méthode des moyennes peut donner des estimations correctes (1ère et 8e lignes), mais est biaisée dans la majeure partie des cas, avec parfois des valeurs complètement aberrantes (comme par exemple la dernière ligne). Parmi toutes les estimations de l'auteur, seules 3 sont incluses dans les intervalles de confiance correspondants (1ère, 3e et 8e lignes). Précisons que l'une des conclusions principales de l'étude menée par l'auteur était un fort taux de d'apparition de résistance à l'Ethambutol.

Nous allons à présent prendre en compte le fait que la fitness ρ est inconnue, et qu'il y a une dilution de paramètre $\zeta = 0.05$. Nous avons tout d'abord recalculé les estimations par la méthode des moyennes, selon le même raisonnement que l'auteur en convertissant le « nombre moyen de mutantes par mL » (resp. le « nombre total de cellules par mL ») en « nombre moyen de mutantes pour 2 mL » (resp. en « nombre total de cellules pour 2 mL »). Ces estimations sont comparées à celles obtenues avec la méthode GF dans les tables 5.5 et 5.6. Notons également que comme l'estimation par la méthode GF donne le nombre moyen de mutations survenues dans la population entière, c'est-à-dire celle contenue dans les 2 mL de solution, il faut également multiplier par 2 le nombre final de cellules dans cette méthode :

```
# Extraction des données et ajout des nombres moyens de mutantes
# fournis par le auteurs
D <- list(D1 = c(mmc = 190, david$D1), D2 = c(mmc = 907, david$D2),
          D3 = c(mmc = 170, david$D3), D4 = c(mmc = 862, david$D4),
          D5 = c(mmc = 80, david$D5), D6 = c(mmc = 1305, david$D6),
          D7 = c(mmc = 0.5, david$D7), D8 = c(mmc = 3.1, david$D8),
          D9 = c(mmc = 843, david$D9), D10 = c(mmc = 4040, david$D10))

# Conversion des données en
# « nombre moyen de cellules dans 2 mL de solution »
D <- lapply(D, function(d) {
        d$mmc <- 2*d$mmc ; d$mfn <- 2*d$mfn
        d
})
# Calcul des estimations de la probabilité de mutation
# et de son intervalle de confiance de niveau 0.95 par la méthode GF
# sous le modèle "H" avec une dilution de paramètre 0.05
GFtest <- lapply(D, function(d) flan.test(mc = d$mc, mfn = d$mfn,
                                          plateff = 0.05,
                                          method = "GF", model = "H"))
```

Avant de commenter les résultats, remarquons que la fitness ρ du 7e échantillon n'a pu être estimée et a été fixée par la fonction `mutestim` à 1. En théorie, cela signifie que la

	Auteur		**GF sous modèle LD avec $\zeta = 0.05$**			
Antibio.	**Estimations de $\pi(\times 10^{-8})$**	**Estimations recalculées de $\pi(\times 10^{-8})$**	**Estimations de $\pi(\times 10^{-8})$**	**IC (95%) de $\pi(\times 10^{-8})$**	**Estimations de ρ**	**IC (95%) de ρ**
Ison.	1.84	1.68	0.432	[0.143 ; 0.721]	0.719	[0.471 ; 0.967]
	3.50	3.21	0.781	[0.461 ; 1.1]	0.789	[0.649 ; 0.93]
	1.70	1.52	0.249	[0.0468 ; 0.451]	0.72	[0.383 ; 1.06]
	3.20	3.04	0.0803	[0.0327 ; 0.128]	0.222	[0.106 ; 0.339]
Strept.	0.900	0.791	0.456	[0 ; 3.15]	16.3	[0 ; 1480]
	5.00	4.39	0.0364	[0.0194 ; 0.0534]	0.0595	[0 ; 0.130]
Rifam.	0.0180	0.0140	0.00667	[0 ; 0.0255]	/	/
	0.0270	0.0234	0.0148	[0 ; 0.0358]	1.69	[0 ; 3.87]
Etham.	7.00	6.26	0.0236	[0.00299 ; 0.0442]	0.151	[0 ; 0.323]
	13.0	12.2	0.0236	[0.0127 ; 0.0346]	0.0929	[0.0157 ; 0.170]

TABLE 5.5 – **Estimations de π et ρ avec les données de DAVID [17, tab.1] sous le modèle LD en prenant en compte la dilution**. Les estimations par la méthode des moyennes ont été recalculées pour la culture totale (2^{e} colonne). Les estimations de m et ρ sont calculées par la méthode GF sous le modèle LD avec un paramètre de dilution $\zeta = 0.05$ et un paramètre de mort nul. Les estimations de π sont obtenues en divisant celles de m par le nombre final de cellules. Pour chaque méthode, les intervalles de confiance de niveau 95% de π et ρ sont également donnés.

	Auteur		GF sous modèle H avec $\zeta = 0.05$			
Antibio.	**Estimations de $\pi(\times 10^{-8})$**	**Estimations recalculées de $\pi(\times 10^{-8})$**	**Estimations de $\pi(\times 10^{-8})$**	**IC (95%) de $\pi(\times 10^{-8})$**	**Estimations de ρ**	**IC (95%) de ρ**
Ison.	1.84	1.68	0.452	[0.136 ; 0.767]	0.700	[0.470 ; 0.930]
	3.50	3.21	0.85	[0.483 ; 1.22]	0.774	[0.642 ; 0.906]
	1.70	1.52	0.256	[0.0388 ; 0.474]	0.694	[0.388 ; 1.00]
	3.20	3.04	0.0788	[0.032 ; 0.126]	0.222	[0.106 ; 0.337]
Strept.	0.900	0.791	0.457	[0 ; 2.9]	4.18	[0 ; 121]
	5.00	4.39	0.0361	[0.0195 ; 0.0526]	0.0596	[0 ; 0.130]
Rifam.	0.0180	0.0140	0.00796	[0 ; 0.0297]	/	/
	0.0270	0.0234	0.0161	[0 ; 0.0399]	1.46	[0 ; 3.02]
Etham.	7.00	6.26	0.0232	[0.00305 ; 0.0433]	0.151	[0 ; 0.322]
	13.0	12.2	0.0233	[0.0126 ; 0.034]	0.0930	[0.0160 ; 0.170]

Table 5.6 – **Estimations de π et ρ avec les données de David [17, tab.1] sous le modèle H en prenant en compte la dilution**. Les estimations par la méthode des moyennes ont été recalculées pour la culture totale (2^{e} colonne). Les estimations de m et ρ sont calculées par la méthode GF sous le modèle H avec un paramètre de dilution $\zeta = 0.05$ et un paramètre de mort nul. Les estimations de π sont obtenues en divisant celles de m par le nombre final de cellules. Pour chaque méthode, les intervalles de confiance de niveau 95% de π et ρ sont également donnés.

fitness est plus grande que 100 (borne maximale fixée pour la recherche de ρ). L'estimation de m qui en découle n'est donc pas pertinente. Seules les 5^e et 8^e estimations des auteurs appartiennent aux intervalles de confiance correspondants. Notons qu'en ce qui concerne, la 5^e ligne la taille de l'intervalle de confiance de π est très grande en comparaison des autres, ce qui est clairement relié à l'estimation très élevée de la fitness. En exagérant le trait, nous pourrions ainsi affirmer que seule la 8^e estimation calculée par David semble correcte.

Considérons à présent l'échantillon de la table 2 de David [17]. Avant toute chose, précisons que nous n'avons pas été en mesure de retrouver l'estimation donnée par l'auteur. Il apparaît en effet que l'estimation calculée par la méthode des moyennes est très différente. Nous avons donc commencé par recalculer l'estimation de π par cette méthode. Nous avons ensuite estimé les paramètres π et ρ par les méthodes P0, ML et GF sans tenir compte dans un premier temps de la dilution de paramètre $\zeta = 0.02$ effectuée. Celle-ci a ensuite été prise en compte par la méthode GF, et nous avons également appliqué la manipulation effectuée par l'auteur pour recalculer une estimation par la méthode des moyennes « prenant en compte » la dilution. Rappelons que contrairement à la méthode des moyennes, nos estimations prennent en compte les fluctuations des nombres finaux de cellules soit en inférant directement à partir des couples (décompte de mutantes - nombre final de cellule) (méthodes P0 et ML), soit en appliquant un coefficient de correction calculé à partir du coefficient de variation des nombres finaux de cellules (méthode GF). Dans tous les cas, le paramètre de mort δ est fixé à 0. Les estimations ainsi obtenues ainsi que les intervalles de confiance de niveau 95% associés sont exposés dans la table 5.7. Notons tout d'abord que l'échantillon étant de petite taille (10 mesures seulement), les intervalles de confiance sont larges et manquent donc de pertinence. Tout ce que nous pouvons remarquer à partir de ces résultats est le fait que les estimations de π recalculées par la méthode des moyennes sont relativement éloignées de celles obtenues par les méthodes P0, ML et GF (au minimum 50% plus élevées). Encore une fois, au vu de la taille de l'échantillon, il est compliqué de pouvoir réellement inférer sur ces résultats.

5.3 Données de Werngren et Hoffner

Nous considérons dans cette section les jeux de données publiées par Werngren et Hoffner [94]. De même que dans la partie précédente, cet article s'intéresse à l'apparition de résistance à différents antibiotiques chez les souches possédant le génotype *Beijing* de *Mycobacterium tuberculosis*. Le jeu de données est contenue dans la table 1 de [94]. Il est constitué de 13 échantillons de décomptes de mutantes (un pour chaque souche) de taille variant entre 23 et 25. Chaque échantillon est associé à un nombre moyen de cellules. L'ensemble de ces données est compris dans le package **flan** dans la variable `werhoff`. Comme dans la partie précédente, les auteurs calculent les estimations de π par la méthode des moyennes de Luria-Delbrück [60, eq. (8)]. De plus, les données exposées

	Estimations de $\pi(\times 10^{-10})$ et ρ	**IC (95%) de $\pi(\times 10^{-10})$**	**IC (95%) de ρ**	**Estimations de $\pi(\times 10^{-10})$ et ρ**	**IC (95%) de $\pi(\times 10^{-10})$**	**IC (95%) de ρ**
Auteur	(7.53,1)	⁄	⁄	⁄	⁄	⁄
Auteur (recalculée)	(1.16,1)	⁄	⁄	⁄	⁄	⁄
Auteur (recalculée pour 5 mL)	(0.819,1)	⁄	⁄	⁄	⁄	⁄
⁄	***LDF***			***HF***		
P0	(2.12,2.08)	[1.15 ; 3.08]	[0 ; 7.20]	(2.12,1.38)	[1.15 ; 3.08]	[0 ; 3.23]
ML	(1.98,2.05)	[0 ; 4.08]	[0 ; 7.13]	(1.81,1.37)	[0 ; 3.91]	[0 ; 3.19]
GF	(2.11,2.19)	[0 ; 4.48]	[0 ; 6.24]	(2.11,1.59)	[0 ; 4.49]	[0 ; 3.59]
GF ($\zeta = 0.02$)	(0.542,0.993)	[0 ; 1.58]	[0 ; 2.00]	(0.591,0.948)	[0 ; 1.79]	[0.0547 ; 1.84]

TABLE 5.7 – **Estimations de π et ρ avec les données de DAVID [17, tab. 2] sous les modèles *LDF* et *HF***. L'estimation par la méthode des moyennes a été recalculée pour 1 mL (2e ligne) et pour la culture totale (3e ligne). Les estimations de π et ρ par les méthodes P0 (4e ligne) et ML (5e ligne) sont calculées sous les modèles LDF et HF directement avec les couples (décompte de mutantes-décompte total). Les estimations de m et ρ par la méthode GF sont calculées sous les modèles LD et H pour un paramètre de dilution $\zeta = 0$ (6e ligne) et $\zeta = 0.02$ (7e ligne). Les estimations de π par P0 et GF sont obtenues en divisant celles de m par la moyenne des décomptes finaux. Le paramètre de mort δ est supposé nul. Les intervalles de confiance de niveau 95% de π et ρ sont également donnés.

correspondent aux nombres de mutantes contenues dans 1 mL de solution, alors que les cultures sont contenues dans un volume de 5 mL. Aucune hypothèse de modélisation n'est faite : la fitness ρ et le paramètre de dilution valent 1 et le paramètre de mort δ est nul. Les tables 5.8 et 5.9 comparent les estimations obtenues par l'auteur avec celles des méthodes P0, ML et GF sous les modèles LD et H. Pour chaque méthode, les intervalles de confiance de niveau 95% sont également donnés. Les estimations obtenues par les auteurs sont moins éloignés de celles calculées avec les méthodes P0, ML et GF que dans la partie précédente : seule la méthode P0 exclut certaines estimations des auteurs (2^{e}, 4^{e}, 5^{e} et 9^{e} souches. Notons d'ailleurs que dans deux cas ($1^{\text{ère}}$ et 7^{e} souches), l'échantillon ne contient pas de décomptes nuls et la méthode P0 ne peut être utilisée.

Nous allons à présent prendre en compte le fait que la fitness ρ est inconnue, et qu'il y a une dilution de paramètre $\zeta = 0.2$. Comme pour le jeu de données précédent, nous avons recalculé les estimations de π par la méthode des moyennes, en convertissant le « nombre moyen de mutantes par mL » (resp. le « nombre total de cellules par mL ») en « nombre moyen de mutantes pour 5 mL » (resp. en « nombre total de cellules pour 5 mL »). Ces estimations sont comparées à celles obtenues avec la méthode GF dans les tables 5.10 et 5.11. Rappelons que comme l'estimation par la méthode GF donne le nombre moyen de mutations survenues dans la population entière, c'est-à-dire celle contenue dans les 5 mL de solution, il faut également multiplier par 5 le nombre final de cellules dans cette méthode.

Notons que la fitness n'a pu être estimée pour le 11^{e} échantillon, et que les estimations obtenues pour la $1^{\text{ère}}$ et la 7^{e} souches sont assez aberrantes. Les estimations obtenues par les auteurs sont très éloignées de celles calculées par la méthode GF. Cependant, du fait de la faible taille des échantillons, la taille des intervalles de confiance est telle que la majorité des estimations des auteurs sont comprises dedans.

Dans cet article, les auteurs regroupent également les souches par génotype (*Beijing* et *non-Beijing*) afin de vérifier l'existence d'une souche avec un taux de mutations plus élevé que les autres. Sans effectuer de test d'hypothèse, ils affirment alors qu'il n'y a pas de différences notables entre les *Beijing* et les *non-Beijing*. Notons π_{nB} la probabilité de mutation pour une souche *non-Beijing* et π_B celle pour une souche *Beijing*. En effectuant le test statistique suivant :

$$H_0 \,:\, \pi_{nB} = \pi_B \text{ contre } H_1 \,:\, \pi_{nB} \neq \pi_B \,,$$

en considérant les estimations obtenues par les auteurs, nous ne pouvons effectivement pas affirmer qu'il y a une différence significative entre les deux probabilités de mutation (p-valeur égale à 0.566) :

```
# Estimations exposées par les auteurs
auth <- c(0.86, 2.4, 0.96, 1.1, 0.65, 1.5, 1.4, 1.3, 0.79, 1.0, 0.94,
          1.5, 1.2)*1e-8
```

	Auteurs	**P0**		**ML**		**GF**	
Souches	$\hat{\pi}(\times 10^{-8})$	$\hat{\pi}(\times 10^{-8})$	**IC (95%) de** $\pi(\times 10^{-8})$	$\hat{\pi}(\times 10^{-8})$	**IC (95%) de** $\pi(\times 10^{-8})$	$\hat{\pi}(\times 10^{-8})$	**IC (95%) de** $\pi(\times 10^{-8})$
H37Rv	0.860	/	/	1.54	[0.484 ; 2.60]	0.998	[0.548 ; 1.45]
E865/94	2.40	6.27	[2.44 ; 10.1]	3.99	[1.79 ; 6.19]	3.03	[1.60 ; 4.46]
E729/94	0.960	1.94	[0.92 ; 2.97]	1.45	[0.759 ; 2.15]	1.16	[0.638 ; 1.69]
E740/94	1.10	2.53	[1.20 ; 3.86]	1.93	[0.868 ; 2.99]	1.35	[0.687 ; 2.01]
E1221/94	0.650	1.33	[0.662 ; 1.99]	0.921	[0.447 ; 1.40]	0.761	[0.401 ; 1.12]
E1449/94	1.50	3.18	[1.26 ; 5.10]	2.45	[1.22 ; 3.67]	1.8	[0.975 ; 2.62]
Harl.	1.40	/	/	2.53	[0.898 ; 4.16]	1.72	[0.933 ; 2.50]
E26/95	1.30	2.29	[1.17 ; 3.41]	1.73	[0.846 ; 2.60]	1.51	[0.823 ; 2.20]
E80/95	0.790	2.10	[0.997 ; 3.21]	1.41	[0.648 ; 2.17]	1.00	[0.500 ; 1.51]
E55/94	1.00	1.83	[0.888 ; 2.77]	1.49	[0.639 ; 2.35]	1.21	[0.500 ; 1.91]
E26/94	0.940	3.16	[1.50 ; 4.82]	1.76	[0.565 ; 2.96]	1.10	[0.461 ; 1.74]
E3942/94	1.50	2.81	[1.33 ; 4.28]	2.31	[1.28 ; 3.33]	1.90	[1.09 ; 2.72]
E47/94	1.20	1.59	[0.811 ; 2.36]	1.48	[0.849 ; 2.11]	1.46	[0.815 ; 2.10]

TABLE 5.8 – **Estimations de π avec les données de WERNGREN et HOFFNER [94, tab.1] sous le modèle *LD*.** Les estimations de π par les méthodes P0, ML et GF sont calculées en divisant celles de m obtenues sous le modèle LD par le nombre final de cellules. La fitness ρ et le paramètre de dilution sont fixés à 1 et le paramètre de mort δ est nul. Pour chaque méthode, les intervalles de confiance de niveau 95% de π sont également donnés.

	Auteurs	**P0**		**ML**		**GF**	
Souches	$\hat{\pi}(\times 10^{-8})$	$\hat{\pi}(\times 10^{-8})$	**IC (95%) de** $\pi(\times 10^{-8})$	$\hat{\pi}(\times 10^{-8})$	**IC (95%) de** $\pi(\times 10^{-8})$	$\hat{\pi}(\times 10^{-8})$	**IC (95%) de** $\pi(\times 10^{-8})$
H37Rv	0.860	/	/	1.65	[0.689 ; 2.61]	1.14	[0.659 ; 1.62]
E865/94	2.40	6.27	[2.44 ; 10.1]	4.12	[2.14 ; 6.09]	3.33	[1.85 ; 4.81]
E729/94	0.960	1.94	[0.920 ; 2.97]	1.57	[0.967 ; 2.18]	1.28	[0.732 ; 1.83]
E740/94	1.10	2.53	[1.20 ; 3.86]	2.04	[1.02 ; 3.07]	1.49	[0.802 ; 2.19]
E1221/94	0.650	1.33	[0.662 ; 1.99]	0.933	[0.491 ; 1.37]	0.824	[0.453 ; 1.19]
E1449/94	1.50	3.18	[1.26 ; 5.10]	2.65	[1.62 ; 3.68]	2.01	[1.14 ; 2.87]
Harl.	1.40	/	/	2.69	[1.30 ; 4.08]	1.92	[1.09 ; 2.75]
E26/95	1.30	2.29	[1.17 ; 3.41]	1.87	[1.10 ; 2.65]	1.61	[0.909 ; 2.32]
E80/95	0.790	2.10	[0.997 ; 3.21]	1.58	[0.898 ; 2.27]	1.10	[0.578 ; 1.63]
E55/94	1.00	1.83	[0.888 ; 2.77]	1.76	[0.988 ; 2.54]	1.29	[0.564 ; 2.01]
E26/94	0.940	3.16	[1.50 ; 4.82]	1.74	[0.61 ; 2.86]	1.21	[0.543 ; 1.87]
E3942/94	1.50	2.81	[1.33 ; 4.28]	2.36	[1.38 ; 3.34]	2.09	[1.25 ; 2.94]
E47/94	1.20	1.59	[0.811 ; 2.36]	1.81	[1.28 ; 2.35]	1.55	[0.898 ; 2.21]

TABLE 5.9 – **Estimations de π avec les données de WERNGREN et HOFFNER [94, tab.1] sous le modèle H.** Les estimations de π par les méthodes P0, ML et GF sont calculées en divisant celles de m obtenues sous le modèle H par le nombre final de cellules. La fitness ρ et le paramètre de dilution sont fixés à 1 et le paramètre de mort δ est nul. Pour chaque méthode, les intervalles de confiance de niveau 95% de π sont également donnés.

	Auteurs		**GF sous modèle *LD* avec $\zeta = 0.2$**			
Souches	**Estimations de $\pi(\times 10^{-8})$**	**Estimations recalculées de $\pi(\times 10^{-8})$**	**Estimations de $\pi(\times 10^{-8})$**	**IC (95%) de $\pi(\times 10^{-8})$**	**Estimations de ρ**	**IC (95%) de ρ**
H37Rv	0.860	0.172	2.95	[0 ; 6.50]	8.59	[0 ; 87.0]
E865/94	2.40	0.482	4.05	[1.50 ; 6.59]	1.99	[0.433 ; 3.54]
E729/94	0.960	0.193	1.17	[0.552 ; 1.79]	1.45	[0.737 ; 2.17]
E740/94	1.10	0.223	2.11	[0.817 ; 3.40]	2.32	[0.233 ; 4.40]
E1221/94	0.650	0.128	0.739	[0.311 ; 1.17]	1.52	[0.636 ; 2.39]
E1449/94	1.50	0.300	2.35	[1.16 ; 3.55]	1.73	[0.815 ; 2.64]
Harl.	1.40	0.278	4.28	[0 ; 9.90]	8.01	[0 ; 81.9]
E26/95	1.30	0.256	1.29	[0.502 ; 2.07]	1.42	[0.555 ; 2.28]
E80/95	0.790	0.161	1.42	[0.549 ; 2.30]	2.16	[0.346 ; 3.98]
E55/94	1.00	0.202	1.12	[0.39 ; 1.85]	1.57	[0.478 ; 2.66]
E26/94	0.940	0.188	0.701	[0.338 ; 1.06]	/	/
E3942/94	1.50	0.302	1.75	[0.837 ; 2.66]	1.34	[0.725 ; 1.96]
E47/94	1.20	0.252	0.756	[0.320 ; 1.19]	0.874	[0.450 ; 1.30]

TABLE 5.10 – **Estimations de π et ρ avec les données de WERNGREN et HOFFNER [94, tab.1] sous le modèle *LD* en prenant en compte la dilution**. Les estimations par la méthode des moyennes ont été recalculées pour la culture totale (2e colonne). Les estimations de m et ρ sont calculées par la méthode GF sous le modèle LD avec un paramètre de dilution $\zeta = 0.2$ et un paramètre de mort nul. Les estimations de π sont obtenues en divisant celles de m par le nombre final de cellules. Pour chaque méthode, les intervalles de confiance de niveau 95% de π et ρ sont également donnés.

	Auteurs		**GF sous modèle H avec $\zeta = 0.2$**			
Souches	**Estimations de $\pi(\times 10^{-8})$**	**Estimations recalculées de $\pi(\times 10^{-8})$**	**Estimations de $\pi(\times 10^{-8})$**	**IC (95%) de $\pi(\times 10^{-8})$**	**Estimations de ρ**	**IC (95%) de ρ**
H37Rv	0.860	0.172	2.98	[0 ; 6.07]	3.37	[0 ; 14.6]
E865/94	2.40	0.482	4.24	[1.59 ; 6.89]	1.59	[0.696 ; 2.49]
E729/94	0.960	0.193	1.22	[0.560 ; 1.88]	1.26	[0.753 ; 1.77]
E740/94	1.10	0.223	2.21	[0.890 ; 3.53]	1.77	[0.701 ; 2.85]
E1221/94	0.650	0.128	0.767	[0.312 ; 1.22]	1.29	[0.692 ; 1.89]
E1449/94	1.50	0.300	2.49	[1.21 ; 3.77]	1.45	[0.856 ; 2.05]
Harl.	1.40	0.278	4.32	[0 ; 9.31]	3.26	[0 ; 14.7]
E26/95	1.30	0.256	1.33	[0.494 ; 2.16]	1.22	[0.619 ; 1.82]
E80/95	0.790	0.161	1.49	[0.584 ; 2.40]	1.69	[0.70 ; 2.68]
E55/94	1.00	0.202	1.16	[0.384 ; 1.93]	1.32	[0.593 ; 2.05]
E26/94	0.940	0.188	0.820	[0.416 ; 1.22]	⁄	⁄
E3942/94	1.50	0.302	1.82	[0.846 ; 2.80]	1.18	[0.729 ; 1.63]
E47/94	1.20	0.252	0.763	[0.309 ; 1.22]	0.810	[0.458 ; 1.16]

TABLE 5.11 – **Estimations de π et ρ avec les données de WERNGREN et HOFFNER [94, tab.1] sous le modèle H en prenant en compte la dilution**. Les estimations par la méthode des moyennes ont été recalculées pour la culture totale (2e colonne). Les estimations de m et ρ sont calculées par la méthode GF sous le modèle H avec un paramètre de dilution $\zeta = 0.2$ et un paramètre de mort nul. Les estimations de π sont obtenues en divisant celles de m par le nombre final de cellules. Pour chaque méthode, les intervalles de confiance de niveau 95% de π et ρ sont également donnés.

```
# Estimations concernant les souches non-Beijing
NB <- auth[1:7]
# Estimations concernant les souches Beijing
B <- auth[8:13]

# Test d'hypothèse : y a-t-il une différence significative
# entre les probabilité de mutation des deux souches ?
t.test(NB, B, alternative = "two.sided")$p.value

## [1] 0.5662851
```

Si nous considérons à présent les estimations obtenues avec la méthode GF et que nous effectuons le test unilatéral suivant :

$$H_0 \,:\, \pi_{nB} = \pi_B \text{ contre } H_1 \,:\, \pi_{nB} > \pi_B \,,$$

nous pouvons alors affirmer que la probabilité de mutation des souches de type *non-Beijing* est significativement plus grande que celle des souches de type *Beijing* (p-valeur égale à 0.0192) :

```
# Extraction des échantillons
W <- werhoff$samples

# Volume de chaque échantillon : 1mL
# Volume total de chaque culture : 5mL
# Donc dilution de paramètre 0.2
dil <- 0.2

# Calcul des estimations par la méthode GF
# Rappel : l'estimation de la fitness échoue pour le 11e échantillon
GFest <- sapply(W, function(w) mutestim(mc = w$mc,
                                        mfn = w$mfn/dil, plateff = dil,
                                        method = "GF")$mutprob)

# Estimations concernant les souches non-Beijing
NB <- GFest[1:7]
# Estimations concernant les souches Beijing
B <- GFest[8:13]

# Test d'hypothèse : la probabilité de mutation des souches non-Beijing
# est-elle significativement plus grande que celle des souches Beijing ?
t.test(NB, B, alternative = "greater")$p.value
```

```
## [1] 0.0191879
```

Il est également possible de considérer les souches, indépendamment du fait qu'elles appartiennent à la classe *Beijing* ou non, et d'effectuer des tests d'hypothèses pour deux échantillons. La table 5.12 expose les p-valeurs obtenues pour le test statistique :

$$H_0 \,:\, \pi_i = \pi_j \text{ contre } H_1 \,:\, \pi_i > \pi_j \,, \tag{5.1}$$

comparant la probabilité de mutation de la i^{e} souche (lignes) avec celle de la j^{e} souche (colonnes). De la même façon, la table 5.13 contient les p-valeurs du test :

$$H_0 \,:\, \rho_i = \rho_j \text{ contre } H_1 \,:\, \rho_i > \rho_j \,. \tag{5.2}$$

comparant le paramètre de fitness de la i^{e} souche (lignes) avec celui de la j^{e} souche (colonnes). Dans les deux tests, le modèle de croissance LD est considéré, le paramètre de mort δ est supposé égal à 0, et la dilution est prise en compte. Le risque de première espèce est fixé à 0.05.

```
nsamples <- length(W)  # Nombre d'échantillon

# Table des p-valeurs des tests de comparaison
pval.mutprob <- matrix(0.5, nrow = nsamples, ncol = nsamples)
pval.fitness <- matrix(0.5, nrow = nsamples, ncol = nsamples)

for(i in 1:nsamples){
# Données concernant la souche n°i
    s1 <- W[[i]] ; mc1 <- s1$mc ; mfn1 <- s1$mfn/dil
    for(j in 1:nsamples){
        if(i !=  j) {
# Données concernant la souche n°j
            s2 <- W[[j]] ; mc2 <- s2$mc ; mfn2 <- s2$mfn/dil

# Hypothèses de test éprouvées :
# H0 : la probabilité de mutation de la souche s1 est égale
#        à celle de la souche s2
# contre
# H1 : la probabilité de mutation de la souche s1 est plus grande
#        que celle de la souche s2
# ET
# H0 : le paramètre de fitness de la souche s1 est égal
#        à celui de la souche s2
# contre
```

```
# H1 : le paramètre de fitness de la souche s1 est plus grand
#        que celui de la souche s2
           res <- flan.test(mc = list(mc1, mc2), mfn = c(mfn1, mfn2),
                            mutprob0 = 0, fitness0 = 0,
                            plateff = dil,
                            alternative = "greater",
                            method = "GF", model = "LD")
# Extraction de la p-valeur pour la probabilité de mutation
           pval.mutprob[i,j] <- res$p.value[1]
# Extraction de la  p.value pour le paramètre de fitness
           pval.fitness[i,j] <- res$p.value[2]
        }
    }
}
```

Notons tout d'abord qu'il n'y a pas de différence significative de fitness entre les souches, quel que soit leur type (excepté entre les souches $E1449/94$ et $E47/94$. De plus, les tests de comparaison de fitness n'ont pas été effectué pour la souche $E26/94$, la méthode GF n'étant pas en mesure d'estimer sa fitness. Rappelons par ailleurs que nous avions remarqué quelques valeurs de fitness aberrantes dans les tableaux précédents. En ce qui concerne la probabilité de mutation, nous obtenons des résultats plus précis que précédemment : la probabilité de mutation de la souche $E865/94$ est significativement plus grande que celle des souches de type *Beijing*. Cependant, d'autres souches de type non-*Beijing* ne sont pas significativement différentes des souches *Beijing*, comme par exemple la souche $E729/94$ ou la souche Harlingen. Nous avons validé précédemment l'affirmation : « les souches non-*Beijing* ont une probabilité de mutation plus forte que les souches *Beijing* ». Au vu des résultats, cette décision semble être principalement causée par les souches $E865/94$ et $E1449/94$. Par ailleurs, des différences significatives sont également observées entre différentes souches de même type, par exemple entre les souches $E1449/94$ et $E1221/94$ qui sont toutes les deux de type *Beijing*.

Les tables 5.14 et 5.15 reprennent les mêmes tests, sous l'hypothèse d'un modèle de croissance H. Les conclusions sont identiques en terme de probabilité de mutation. Nous pouvons observer que les p-valeurs obtenues pour les tests comparant les paramètres de fitness sont plus faibles avec un modèle H. En l'occurrence, la fitness de la souche $E740/94$ serait également significativement plus grande que celle de la souche $E47/94$.

		non-*Beijing*							*Beijing*					
		H37Rv	**E865/94**	**E729/94**	**E740/94**	**E1221/94**	**E1449/94**	**Harl.**	**E26/95**	**E80/95**	**E55/94**	**E26/94**	**E3942/94**	**E47/94**
non-*Beijing*	**H37Rv**	⁄	0.689	0.166	0.331	0.113	0.377	0.653	0.185	0.207	0.161	0.108	0.261	0.115
	E865/94	0.311	⁄	**0.0156**	0.0917	**0.00599**	0.119	0.530	**0.0211**	**0.0281**	**0.0151**	**0.00537**	**0.0479**	**0.00624**
	E729/94	0.834	0.984	⁄	0.901	0.131	0.958	0.860	0.592	0.68	0.461	0.100	0.849	0.142
	E740/94	0.669	0.908	0.099	⁄	**0.0242**	0.607	0.770	0.143	0.195	0.0959	**0.0199**	0.328	**0.0259**
	E1221/94	0.887	0.994	0.869	0.976	⁄	0.994	0.891	0.885	0.916	0.812	0.448	0.975	0.522
	E1449/94	0.623	0.881	**0.0421**	0.393	**0.00632**	⁄	0.745	0.0721	0.110	**0.0423**	**0.00475**	0.216	**0.00692**
	Harl.	0.347	0.470	0.140	0.23	0.109	0.255	⁄	0.151	0.163	0.137	0.107	0.192	0.110
Beijing	**E26/95**	0.815	0.979	0.408	0.857	0.115	0.928	0.849	⁄	0.590	0.380	0.0921	0.774	0.123
	E80/95	0.793	0.972	0.320	0.805	0.0838	0.890	0.837	0.410	⁄	0.301	0.0672	0.693	0.0901
	E55/94	0.839	0.985	0.539	0.904	0.188	0.958	0.863	0.620	0.699	⁄	0.157	0.854	0.200
	E26/94	0.892	0.995	0.900	0.980	0.552	0.995	0.893	0.908	0.933	0.843	⁄	0.982	0.575
	E3942/94	0.739	0.952	0.151	0.672	**0.0246**	0.784	0.808	0.226	0.307	0.146	**0.0182**	⁄	**0.0270**
	E47/94	0.885	0.994	0.858	0.974	0.478	0.993	0.890	0.877	0.910	0.800	0.425	0.973	⁄

TABLE 5.12 – ***p*-valeurs des tests de comparaison (deux à deux) des probabilités de mutation des différentes souches des données de WERNGREN et HOFFNER [94, tab. 1] avec la méthode GF sous le modèle *LD***. L'hypothèse éprouvée est la suivante : « la probabilité de mutation de la i^{e} souche (lignes) est significativement plus grande que celle de la j^{e} souche (colonnes) ». Le paramètre de dilution est fixé à $\zeta = 0.2$ et le paramètre de mort est supposé nul.

		non-*Beijing*							*Beijing*						
		H37Rv	**E865/94**	**E729/94**	**E740/94**	**E1221/94**	**E1449/94**	**Harl.**	**E26/95**	**E80/95**	**E55/94**	**E26/94**	**E3942/94**	**E47/94**
non-*Beijing*	**H37Rv**	/	0.434	0.429	0.438	0.430	0.432	0.496	0.429	0.436	0.430	/	0.428	0.423
	E865/94	0.566	/	0.271	0.598	0.302	0.388	0.563	0.266	0.558	0.334	/	0.224	0.0878
	E729/94	0.571	0.729	/	0.778	0.542	0.677	0.569	0.476	0.762	0.570	/	0.407	0.0860
	E740/94	0.562	0.402	0.222	/	0.244	0.306	0.560	0.218	0.457	0.267	/	0.190	0.0918
	E1221/94	0.570	0.698	0.458	0.756	/	0.628	0.568	0.439	0.736	0.531	/	0.375	0.0988
	E1449/94	0.568	0.612	0.323	0.694	0.372	/	0.566	0.316	0.663	0.415	/	0.246	**0.0482**
	Harl.	0.504	0.437	0.431	0.44	0.432	0.434	/	0.431	0.438	0.432	/	0.43	0.425
Beijing	**E26/95**	0.571	0.734	0.524	0.782	0.561	0.684	0.569	/	0.766	0.584	/	0.442	0.133
	E80/95	0.564	0.442	0.238	0.543	0.264	0.337	0.562	0.234	/	0.292	/	0.200	0.0877
	E55/94	0.570	0.666	0.430	0.733	0.469	0.585	0.568	0.416	0.708	/	/	0.360	0.122
	E26/94	/	/	/	/	/	/	/	/	/	/	/	/	/
	E3942/94	0.572	0.776	0.593	0.810	0.625	0.754	0.570	0.558	0.8	0.64	/	/	0.110
	E47/94	0.577	0.912	0.914	0.908	0.901	0.952	0.575	0.867	0.912	0.878	/	0.890	/

TABLE 5.13 – **p-valeurs des tests de comparaison (deux à deux) des paramètres de fitness des différentes souches des données de WERNGREN et HOFFNER [94, tab.1] avec la méthode GF sous le modèle *LD***. L'hypothèse éprouvée est la suivante : « le paramètre de fitness de la i^{e} souche (lignes) est significativement plus grand que celui de la j^{e} souche (colonnes) ». Le paramètre de dilution est fixé à $\zeta = 0.2$ et le paramètre de mort est supposé nul.

		non-*Beijing*							*Beijing*					
		H37Rv	**E865/94**	**E729/94**	**E740/94**	**E1221/94**	**E1449/94**	**Harl.**	**E26/95**	**E80/95**	**E55/94**	**E26/94**	**E3942/94**	**E47/94**
non-*Beijing*	**H37Rv**	⟋	0.728	0.138	0.327	0.0824	0.387	0.673	0.156	0.182	0.131	0.0871	0.242	0.0821
	E865/94	0.272	⟋	**0.0153**	0.0900	**0.00574**	0.122	0.512	**0.0200**	**0.0273**	**0.0144**	**0.00626**	**0.0472**	**0.00569**
	E729/94	0.862	0.985	⟋	0.905	0.133	0.958	0.886	0.576	0.681	0.45	0.154	0.841	0.131
	E740/94	0.673	0.91	0.0946	⟋	**0.0214**	0.617	0.789	0.133	0.189	0.0885	**0.0241**	0.323	**0.021**
	E1221/94	0.918	0.994	0.867	0.979	⟋	0.994	0.918	0.876	0.919	0.803	0.568	0.973	0.495
	E1449/94	0.613	0.878	**0.0420**	0.383	**0.00639**	⟋	0.757	0.0673	0.105	**0.0401**	**0.00728**	0.209	**0.00627**
	Harl.	0.327	0.488	0.114	0.211	0.082	0.243	⟋	0.123	0.137	0.109	0.085	0.168	0.0818
Beijing	**E26/95**	0.844	0.98	0.424	0.867	0.124	0.933	0.877	⟋	0.603	0.385	0.142	0.777	0.122
	E80/95	0.818	0.973	0.319	0.811	0.0809	0.895	0.863	0.397	⟋	0.292	0.0925	0.688	0.0797
	E55/94	0.869	0.986	0.55	0.912	0.197	0.96	0.891	0.615	0.708	⟋	0.224	0.853	0.194
	E26/94	0.913	0.994	0.846	0.976	0.432	0.993	0.915	0.858	0.908	0.776	⟋	0.969	0.427
	E3942/94	0.758	0.953	0.159	0.677	**0.0274**	0.791	0.832	0.223	0.312	0.147	**0.0314**	⟋	**0.0269**
	E47/94	0.918	0.994	0.869	0.979	0.505	0.994	0.918	0.878	0.920	0.806	0.573	0.973	⟋

TABLE 5.14 – ***p*-valeurs des tests de comparaison (deux à deux) des probabilités de mutation des différentes souches des données de WERNGREN et HOFFNER [94, tab.1] avec la méthode GF sous le modèle *H***. L'hypothèse éprouvée est la suivante : « la probabilité de mutation de la i^{e} souche (lignes) est significativement plus grande que celle de la j^{e} souche (colonnes) ». Le paramètre de dilution est fixé à $\zeta = 0.2$ et le paramètre de mort est supposé nul.

		non-*Beijing*							*Beijing*					
		H37Rv	**E865/94**	**E729/94**	**E740/94**	**E1221/94**	**E1449/94**	**Harl.**	**E26/95**	**E80/95**	**E55/94**	**E26/94**	**E3942/94**	**E47/94**
non-*Beijing*	**H37Rv**	/	0.379	0.356	0.391	0.359	0.369	0.495	0.354	0.385	0.360	/	0.351	0.327
	E865/94	0.621	/	0.264	0.600	0.293	0.400	0.613	0.249	0.556	0.322	/	0.211	0.0556
	E729/94	0.644	0.736	/	0.801	0.531	0.684	0.635	0.459	0.775	0.552	/	0.409	0.0763
	E740/94	0.609	0.400	0.199	/	0.221	0.304	0.601	0.188	0.454	0.246	/	0.159	**0.0471**
	E1221/94	0.641	0.707	0.469	0.779	/	0.645	0.633	0.433	0.749	0.523	/	0.386	0.0869
	E1449/94	0.631	0.600	0.316	0.696	0.355	/	0.622	0.294	0.655	0.391	/	0.239	**0.0345**
	Harl.	0.505	0.387	0.365	0.399	0.367	0.378	/	0.363	0.394	0.369	/	0.360	0.337
Beijing	**E26/95**	0.646	0.751	0.541	0.812	0.567	0.706	0.637	/	0.787	0.583	/	0.460	0.124
	E80/95	0.615	0.444	0.225	0.546	0.251	0.345	0.606	0.213	/	0.278	/	0.180	0.0503
	E55/94	0.640	0.678	0.448	0.754	0.477	0.609	0.631	0.417	0.722	/	/	0.376	0.108
	E26/94	/	/	/	/	/	/	/	/	/	/	/	/	/
	E3942/94	0.649	0.789	0.591	0.841	0.614	0.761	0.640	0.54	0.820	0.624	/	/	0.102
	E47/94	0.673	0.944	0.924	0.953	0.913	0.966	0.663	0.876	0.950	0.892	/	0.898	/

TABLE 5.15 – **p-valeurs des tests de comparaison (deux à deux) des paramètres de fitness des différentes souches des données de WERNGREN et HOFFNER [94, tab.1] avec la méthode GF sous le modèle $\boldsymbol{H}$.** L'hypothèse éprouvée est la suivante : « le paramètre de fitness de la i^{e} souche (lignes) est significativement plus grand que celui de la j^{e} souche (colonnes) ». Le paramètre de dilution est fixé à $\zeta = 0.2$ et le paramètre de mort est supposé nul.

Chapitre 6

Perspectives

Nous avons exposé dans cette thèse une étude probabiliste et statistique des modèles de mutations. Nous allons à présent considérer séparément ces deux parties afin de donner quelques voies de réflexion pour chacune d'entre elles.

6.1 Modélisation

— Nous avons proposé dans la section 3.3.2 une nouvelle approche de la fitness : définie initialement comme le rapport entre le paramètre malthusien des cellules normales et celui des mutantes, nous considérons ici qu'il s'agit du rapport entre le taux instantané de division des cellules normales et celui des mutantes. Il s'agit d'une hypothèse généralement utilisée en analyse de survie, dans les modèles de régression de Cox [15]. Dans le cas où les durées de vie sont exponentiellement distribuées, le taux instantané de division et le paramètre malthusien sont identiques, et les deux définitions sont donc équivalentes. Mais de manière générale, le taux instantané de division associé à une loi quelconque n'est pas homogène en temps. Quel lien existe-t-il entre le paramètre malthusien tel qu'il est défini par Harris [33] et le taux instantané de division, et peut-on en déduire un résultat analogue au théorème 17.1 du même auteur ?

— Avec cette nouvelle approche, nous avons donc pu retrouver les résultats classiques du cas où les durées de vie sont *i.i.d.* et exponentielles. Nous avons également retrouvé le modèle de Haldane [81, 97]. Il serait à présent intéressant de considérer le cas général où les fonctions de répartition F_ν et F_μ vérifient $(\mathcal{H})$ mais sans la moindre relation de proportionnalité. Par exemple, que peut-on écrire lorsque F_ν et/ou F_μ sont définies à partir de deux fonctions de répartition G_ν et G_μ définies sur $\mathbb{R}_+$ par

$$F_\nu(s,t) = \frac{G_\nu(t) - G_\nu(s)}{1 - G_\nu(s)}\,,$$

et

$$F_\mu(s,t) = \frac{G_\mu(t) - G_\mu(s)}{1 - G_\mu(s)}\,,$$

sans aucune relation de proportionnalité ? En particulier, que se passe-t-il si seule G_μ est quelconque alors que G_ν est la fonction de répartition d'une loi exponentielle (comment retrouve-t-on le résultat de Hamon et Ycart [31] sans la notion de fitness ?).

— La décomposition initiale des modèles de croissance [31] fait intervenir les durées de vie des cellules normales via les résultats de convergence de Kuzcek [52, 53]. Peut-on généraliser ces résultats aux modèle inhomogènes en temps, et ainsi retrouver un ingrédient basé sur les durées de vie au lieu des instants de divisions ?

— Autres : peut-on calculer les probabilités du modèle H lorsqu'il y a dilution (c'est-à-dire $\zeta < 1$) ? Est-ce que la prise en compte des morts des cellules normales (i.e. $\gamma > 0$) peut apporter quelque chose ?

6.2 Estimation

— Nous avons mentionné dans la partie 2.3 des méthodes d'estimation basées sur la médiane ou d'autres quantiles du nombres de mutantes. Le fait est que ces méthodes se basaient sur les quantiles empiriques, qui n'ont pas de bonnes propriétés asymptotiques lorsque l'on considère une variable discrète. Cependant, des estimateurs robustes pour ces quantiles ont été proposés par Ma et al. [62]. Une réécriture complète de la méthode de Lea et Coulson [56], de Jones et al. [39] ou encore de Koch [47] est alors envisageable.

— Un autre problème d'estimation reste encore à résoudre : comment estimer la probabilité de mort δ ? Le fait est que fixer $\delta = 0$ cause une sous-estimation de m et de ρ, et que le biais ainsi induit peut être non-négligeable (voir figures 4.4). Nous avons vu qu'il est en théorie possible d'estimer δ par la méthode ML. Cela nous semble également possible en pratique, mais du moment que l'initialisation des paramètres d'intérêt soit déjà proche de la solution (comme nous le faisons déjà pour l'estimation de m et ρ). L'utilisation de méthodes de type ABC (Approximate Bayesian Computation) est-elle pertinente ? Une autre piste possible serait d'ajouter une relation (par exemple linéaire) entre la probabilité de mutation π et le paramètre de mort δ. Cette approche pourrait être pertinente en pratique , dans le sens où certaines mutations peuvent entraîner ou augmenter les chances de mort de la cellule. L'ajout d'une relation de proportionnalité entre π et δ pourrait éventuellement améliorer l'identification du modèle.

— Nous avons illustré dans la section 4.3.2 le fait que les modèles LD et H semblent être des cas « extrêmes » en ce qui concerne le modèle de croissance des clones mutants. Comment exprimer rigoureusement et démontrer cette assertion ?

Piste : définissons l'ensemble $\mathcal{F}_m$ suivant :

$$\mathcal{F}_m = \left\{ f : \mathbb{R}_+ \longmapsto \mathbb{R}_+ \text{ t.q. } \int_0^\infty f(x)\mathrm{d}x = 1 \text{ et } \int_0^\infty x f(x)\mathrm{d}x = m \right\},$$

c'est-à-dire l'ensemble des densités de loi sur $\mathbb{R}_+$ d'espérance égale à m.
Soit X une v.a. de densité f, telle que $f\log(f) \in L^1(\mathbb{R})$. L'entropie de la v.a. X est définie par la fonction H :

$$H(f) = -\int_{-\infty}^{+\infty} f(x)\log(f(x))\mathrm{d}x\,.$$

L'entropie H est maximisée sur $\mathcal{F}_m$ par la densité de la loi exponentielle de paramètre $\frac{1}{m}$. Elle est minimisée sur le même ensemble par la densité de la loi de Dirac localisée en m. Définissons à présent l'ensemble $\mathcal{F}_{\nu,m}$ suivant :

$$\mathcal{F}_{\nu,m} = \left\{ f : \mathbb{R}_+ \longmapsto \mathbb{R}_+ \text{ t.q. } \int_0^\infty f(x)\mathrm{d}x = 1 \text{ et } \int_0^\infty \mathrm{e}^{-\nu x} f(x)\mathrm{d}x = m \right\},$$

c'est-à-dire l'ensemble des densités de loi sur $\mathbb{R}_+$ de paramètre malthusien égale à ν (dans un processus de branchements où l'espérance du nombre de descendants de chaque individu vaut m). Quelle densité de loi maximise H dans l'ensemble $\mathcal{F}_{\nu,m}$ (pas la loi exponentielle apriori) ? Peut-on trouver une grandeur statistique pertinente maximisée (ou minimisée) sur $\mathcal{F}_{\nu,m}$ par la loi exponentielle ?

— Le package **flan** fournit des outils d'inférence statistique. Cependant, de nombreux paramètres peuvent être ajustés et différents types de données peuvent être employées. Afin de rendre le package accessible aux utilisateurs non formés au langage R, une application Shiny est actuellement en cours de développement [65].

Annexe A

Résultats sur les processus de comptage

Nous exposons ici les propriétés sur les processus de comptage utilisées dans la partie 3.2.3. Ces résultats sont issus des chapitres 2 et 3 de [27]. Leur démonstration sont également rappelées. Tout au long de cette annexe, nous considérerons une suite $(T_i)_{i\in\mathbb{N}}$ d'instants d'occurrence d'un certain événement (par exemple une panne d'un composant ou dans notre cas une mutation). Nous noterons $\{N(t)\}_{t\geqslant 0}$ le processus ponctuel de comptage associé, défini pour tout $t \in \mathbb{R}_+$ par

$$N(t) = \max\{i \in \mathbb{N}\,;\, T_i \leqslant t\}\ .$$

$N(t)$ est donc le nombre cumulé d'événements survenus entre 0 et t. Nous supposerons également que les hypothèses suivantes sont vérifiées :

1. $T_0 = 0$ et $N(0) = 0$ avec probabilité 1 ;
2. Les trajectoires de $\{N(t)\}_{t\geqslant 0}$ sont *cadlag* et croissantes ;
3. le processus de comptage $\{N(t)\}_{t\geqslant 0}$ est simple, c'est-à-dire :

$$\lim_{\Delta t\to 0} \frac{1}{\Delta t}\mathbb{P}\left[N(t+\Delta t) - N(t) > 1\right] = 0\,.$$

En d'autres termes, il ne peut y avoir plus d'une occurrence à la fois.

Considérons à présent les deux grandeurs suivantes :

1. L'intensité du processus $\{N(t)\}_{t\geqslant 0}$ définie par

$$\begin{aligned}\xi(t, N(t)) &\equiv \xi(t, N(t), T_1, \ldots, T_{N(t)}) \\ &= \lim_{\Delta t\to 0} \frac{1}{\Delta t}\mathbb{P}\left[N(t+\Delta t) - N(t) = 1 \,|\, N(t), T_1, \ldots, T_{N(t)}\right]\ . \end{aligned} \tag{A.1}$$

Cette intensité exprime la probabilité de l'occurrence d'un événement juste après un instant t, sachant le nombre d'événements avant t et leur instant d'occurrence respectif. L'intensité ξ est ainsi une variable aléatoire dépendant de $N(t), T_1, \ldots, T_{N(t)}$. Pour plus de lisibilité, nous nous contenterons parfois de noter $\xi(t, N(t))$.

2. L'intensité de comptage conditionnelle définie par

$$\tilde{\xi}(t,n) = \mathbb{E}[\xi(t,N(t)) \,|\, N(t)=n] \tag{A.2}$$

Contrairement à $\xi(t,N(t))$, la quantité $\tilde{\xi}(t,n)$ est déterministe. Elle exprime la probabilité de l'occurrence d'un événement juste après un instant t, sachant que n ont déjà été observés avant l'instant t.

À partir des deux grandeurs définies ci-dessus, il est possible d'exprimer la loi du nombre d'occurrences avant un instant donné, ainsi que celle des instants d'occurrences d'un nombre fixés d'événements.

Proposition A.0.1. *Pour tout $t \in \mathbb{R}_+$, la loi du nombre $N(t)$ d'occurrences observées avant l'instant t est donnée par*

$$\mathbb{P}\left[N(t)=0\right] = \exp\left(-\int_0^t \tilde{\xi}(u,0)\mathrm{d}u\right), \tag{A.3}$$

et pour tout $n>0$,

$$\mathbb{P}\left[N(t)=n\right] = \int_{0<t_1<\cdots<t_n} \left[\prod_{i=1}^{n} \tilde{\xi}(t_i,i-1)\right] \exp\left(-\sum_{i=0}^{n}\left(\int_{t_i}^{t_{i+1}} \tilde{\xi}(u,i)\mathrm{d}u\right)\right) \mathrm{d}t_1 \ldots \mathrm{d}t_n\,, \tag{A.4}$$

avec la convention $t_0=0$ et $t_{k+1}=t$.

Démonstration de la proposition A.0.1. Par définition de $\tilde{\xi}(t,n)$, il vient pour tout $t\in\mathbb{R}_+$ et tout $n\in\mathbb{N}$:

$$\mathbb{P}\left[N(t+\Delta t)=n+1 \,|\, N(t)=n\right] = \tilde{\xi}(t,n)\Delta t + o(\Delta t)\,.$$

En particulier, comme le processus est simple :

$$\mathbb{P}\left[N(t+\Delta t)=0 \,|\, N(t)=0\right] = 1-\tilde{\xi}(t,0)\Delta t + o(\Delta t)\,.$$

Pour tout $n\in\mathbb{N}$ et tout $t\in\mathbb{R}_+$, notons $P_n(t)=\mathbb{P}\left[N(t)=n\right]$. Alors

$$\frac{P_0(t+\Delta t)}{P_0(t)} = 1-\tilde{\xi}(t,0)\Delta t + o(\Delta t)\,,$$

et donc

$$\begin{aligned}
\frac{\mathrm{d}P_0(t)}{\mathrm{d}t} &= \lim_{\Delta t\to 0} \frac{P_0(t+\Delta t)-P_0(t)}{\Delta t} \\
&= P_0(t) \lim_{\Delta t\to 0}\left(-\tilde{\xi}(t,0)+\frac{o(\Delta t)}{\Delta t}\right) \\
&= -P_0(t)\tilde{\xi}(t,0)\,.
\end{aligned}$$

En d'autres termes :

$$\log\left(P_0(t)\right) = -\int_0^t \tilde{\xi}(u,0)\mathrm{d}u + C\,,$$

avec C une constante. Comme $N_0 = 0$ avec probabilité 1, il vient que $C = 0$ et donc

$$\mathbb{P}\left[N(t) = 0\right] = \exp\left(-\int_0^t \tilde{\xi}(u,0)\mathrm{d}u\right)\,.$$

Soit $n > 0$. Comme le processus est simple, pour tous $t, \Delta t \in \mathbb{R}_+$:

$$\begin{aligned} P_n(t+\Delta t) = &\left(1 - \tilde{\xi}(t,0)\Delta t + o(\Delta t)\right) P_n(t) \\ &+ \left(\tilde{\xi}(t,0)\Delta t + o(\Delta t)\right) P_{n-1}(t) \\ &+ o(\Delta t)\,. \end{aligned}$$

Nous obtenons donc l'équation différentielle suivante :

$$\frac{\mathrm{d}P_n(t)}{\mathrm{d}t} = -\tilde{\xi}(t,n)P_n(t) + \tilde{\xi}(t,n-1)P_{n-1}(t)\,.$$

Posons $Q_n(t) = P_n(t)\exp\left(\int_0^t \tilde{\xi}(u,n)\mathrm{d}u\right)$, alors :

$$\frac{\mathrm{d}Q_n(t)}{\mathrm{d}t} = \exp\left(\int_0^t \tilde{\xi}(u,n)\mathrm{d}u\right)\tilde{\xi}(t,n-1)P_{n-1}(t)\,.$$

Donc

$$Q_n(t) = \int_0^t \left[\tilde{\xi}(u,n-1)P_{n-1}(u)\exp\left(\int_0^u \tilde{\xi}(s,n)\mathrm{d}s\right)\right]\mathrm{d}u\,.$$

Finalement nous obtenons l'expression de $p_n(t)$ en fonction de $p_{n-1}(t)$ suivante :

$$P_n(t) = \int_0^t \tilde{\xi}(t,n-1)p_{n-1}(u)\exp\left(\int_u^t \tilde{\xi}(s,n)\mathrm{d}s\right)\mathrm{d}u\,.$$

Répétons la procédure pour obtenir l'expression $P_{n-1}(t)$ en fonction de $P_{n-2}(t)$. En injectant cette relation dans l'équation ci-dessus il vient alors :

$$\begin{aligned} P_n(t) = \int_0^t \int_0^{t_n} &\tilde{\xi}(t_n,n-1)\tilde{\xi}(t_{n-1},n-2)P_{n-2}(t_{n-1}) \\ &\times \exp\left(-\int_{t_{n-1}}^{t_n}\tilde{\xi}(u,n-1)\mathrm{d}u - \int_{t_n}^t \tilde{\xi}(u,n)\mathrm{d}u\right)\mathrm{d}t_{n-1}\mathrm{d}t_n\,. \end{aligned}$$

En recommençant l'opération, il vient :

$$P_n(t) = \int_0^t \int_0^{t_n} \cdots \int_0^{t_2} \left[\prod_{i=1}^n \tilde{\xi}(t_i, i-1)\right] P_0(t_1)\exp\left(-\sum_{i=1}^n \int_{t_i}^{t_{i+1}} \tilde{\xi}(u,i)\mathrm{d}u\right)\mathrm{d}t_1 \ldots \mathrm{d}t_n$$

D'où (A.4). □

Intéressons nous à présent à la loi des instants d'occurrences. Pour alléger la rédaction, si $\boldsymbol{t^{(n)}} = (t_1, \dots, t_n) \in \mathbb{R}_+^n$, nous noterons $\boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}$ l'événement $(T_1 = t_1, \dots, T_n = t_n)$.

Proposition A.0.2. *La loi de l'instant T_1 de la première occurrence a pour densité*

$$f_{T_1}(t) = \xi(t,0) \exp\left(- \int_0^t \xi(u,0) \mathrm{d}u \right) . \tag{A.5}$$

La loi de l'instant T_{n+1} conditionnée à $\boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}$ a pour densité

$$f_{\left(T_{n+1} | \boldsymbol{T^{(n)}}\right) = \boldsymbol{t^{(n)}}}(t) = \xi(t, n, \boldsymbol{t^{(n)}}) \exp\left(- \int_{t_n}^t \xi(u, n, \boldsymbol{t^{(n)}}) \mathrm{d}u \right) \mathbb{1}_{0<t_1<\dots<t_n} . \tag{A.6}$$

La loi jointe du vecteur $\boldsymbol{T^{(n)}}$ a pour densité :

$$f_{\boldsymbol{T^{(n)}}}(\boldsymbol{t^{(n)}}) = \left[\prod_{i=1}^n \xi(t_i, i-1, \boldsymbol{t^{(i-1)}}) \right] \exp\left(- \sum_{i=1}^n \int_{t_{i-1}}^{t_i} \xi(u, i-1, \boldsymbol{t^{(i-1)}}) \mathrm{d}u \right) \mathbb{1}_{0<t_1<\dots<t_n} . \tag{A.7}$$

Démonstration de la proposition A.0.2. La loi de T_1 s'obtient en remarquant tout d'abord que

$$\mathbb{P}\left[T_1 > t\right] = \mathbb{P}\left[N(t) = 0\right] = \exp\left(- \int_0^t \tilde{\xi}(u,0) \right) .$$

Puis

$$\begin{aligned}
\tilde{\xi}(t,0) &= \mathbb{E}\left[\xi(t, N(t)) \,|\, N(t)\right] \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{P}\left[N(t+\Delta t) - N(t) = 1 \,|\, N(t) = 0\right] \\
&= \xi(t,0)
\end{aligned}$$

D'où (A.5).

Soit $n > 0$ et $\boldsymbol{t^{(n)}} \in \mathbb{R}_+^n$ tel que $0 < t_1 < \dots < t_n < t$. Par définition :

$$\begin{aligned}
\xi(t, n, \boldsymbol{t^{(n)}}) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{P}\left[N(t+\Delta t) - N(t) = 1 \,|\, N(t) = n, \boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}\right] \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{P}\left[t < T_{n+1} \leqslant t + \Delta t \,|\, N(t) = n, \boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}\right] \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{\mathbb{P}\left[\{t < T_{n+1} \leqslant t + \Delta t\} \cap \{N(t) = n\} \,|\, \boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}\right]}{\mathbb{P}\left[N(t) = n \,|\, \boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}\right]}
\end{aligned}$$

Remarquons que conditionnellement à $\boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}$

$$\{N(t) = n\} \Leftrightarrow \{T_{n+1} > t\} .$$

Ainsi :

$$\xi(t, n, \boldsymbol{t^{(n)}}) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{\mathbb{P}\left[t < T_{n+1} \leqslant t + \Delta t \,|\, \boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}\right]}{\mathbb{P}\left[T_{n+1} > t \,|\, \boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}\right]}$$

$$= \frac{f_{T_{n+1}|\boldsymbol{T^{(n)}}=\boldsymbol{t^{(n)}}}(t)}{\mathbb{P}\left[T_{n+1} > t \,|\, \boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}\right]}$$

$$= -\frac{\mathrm{d}}{\mathrm{d}t} \log\left(\mathbb{P}\left[T_{n+1} > t \,|\, \boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}\right]\right) .$$

Il vient donc que

$$\mathbb{P}\left[T_{n+1} > t \,|\, \boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}\right] = 1 - \exp\left(-\int_0^t \xi(u, n, \boldsymbol{t^{(n)}})\mathrm{d}u + C\right) ,$$

avec C une constante. Puisque $\mathbb{P}\left[T_{n+1} \geqslant t_n \,|\, \boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}\right] = 0$, nous en déduisons que

$$\mathbb{P}\left[T_{n+1} > t \,|\, \boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}\right] = 1 - \exp\left(-\int_{t_n}^t \xi(u, n, \boldsymbol{t^{(n)}})\mathrm{d}u\right) .$$

En dérivant par rapport à t nous obtenons finalement (A.6).

Pour retrouver la loi jointe de $\boldsymbol{T^{(n)}}$, il suffit d'écrire que

$$\begin{aligned} f_{\boldsymbol{T^{(n)}}}(\boldsymbol{t^{(n)}}) &= f_{T_n|\boldsymbol{T^{(n-1)}}=\boldsymbol{t^{(n-1)}}}(t_n) f_{\boldsymbol{T^{(n-1)}}}(\boldsymbol{t^{(n-1)}}) \\ &= \dots \\ &= f_{T_1}(t_1) \prod_{i=2}^{n} f_{T_n|\boldsymbol{T^{(n-1)}}=\boldsymbol{t^{(n-1)}}}(t_i) , \end{aligned}$$

pour en déduire (A.7). □

Considérons à présent le cas où le processus $\{N(t)\}_{t \geqslant 0}$ est un processus de Poisson inhomogène en temps. Ce cas est caractérisé par le fait que l'intensité (A.1) est une fonction déterministe du temps :

$$\xi(t, N(t), T_1, \dots, T_{N(t)}) = \xi(t) .$$

La première conséquence est l'indépendance des accroissements. En effet, l'intensité ne dépend pas de $T_1, \dots, T_{N(t)}$ et peut donc s'écrire

$$\xi(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{P}\left[N(t + \Delta t) - N(t) = 1\right] .$$

De fait, pour $s_1 < t_1 < s_2 < t_2$, les variables $N(t_1) - N(s_1)$ et $N(t_2) - N(s_2)$ sont indépendantes. L'intensité de comptage conditionnelle (A.2) s'écrit quant à elle

$$\tilde{\xi}(t, N(t)) = \mathbb{E}\left[\xi(t) \,|\, N(t)\right] = \xi(t)\,.$$

En appliquant la proposition A.0.1 nous pouvons écrire la loi du nombre d'occurrences survenues avant un instant donné.

Proposition A.0.3. *Pour tout $t \in \mathbb{R}_+$, $N(t)$ suit la loi de Poisson d'espérance*

$$m(t) = \int_0^t \xi(u)\mathrm{d}u\,.$$

Démonstration de la proposition A.0.3. Pour commencer, d'après (A.3) :

$$\begin{aligned}\mathbb{P}\left[N(t) = 0\right] &= \exp\left(-\int_0^t \xi(u)\mathrm{d}u\right)\\ &= \mathrm{e}^{-m(t)}\,.\end{aligned}$$

Puis pour $n > 0$, d'après (A.4) :

$$\begin{aligned}\mathbb{P}\left[N(t) = n\right] &= \int_{0<t_1<\cdots<t_n} \left[\prod_{i=1}^n \xi(t_i)\right] \exp\left(-\sum_{i=0}^n \left(\int_{t_i}^{t_{i+1}} \xi(u)\mathrm{d}u\right)\right) \mathrm{d}t_1 \ldots \mathrm{d}t_n\\ &= \exp\left(-\int_0^t \xi(u)\mathrm{d}u\right) \int_{0<t_1<\cdots<t_n} \left[\prod_{i=1}^n \xi(t_i)\right] \mathrm{d}t_1 \ldots \mathrm{d}t_n\,.\end{aligned}$$

Il suffit ensuite de remarquer que

$$\int_0^{t_2} \xi(t_1)\mathrm{d}t_1 = m(t_2)\,;$$

$$\int_0^{t_3} m(t_2)\xi(t_2)\mathrm{d}t_2 = \frac{m(t_3)^2}{2}\,;$$

$$\int_0^{t_4} \frac{m(t_3)^2}{2}\xi(t_3)\mathrm{d}t_3 = \frac{m(t_3)^2}{3!}\,;$$

et ainsi de suite, afin d'obtenir finalement

$$\mathbb{P}\left[N(t) = n\right] = \mathrm{e}^{-m(t)}\frac{m(t)^n}{n!}\,.$$

□

L'application directe de la proposition A.0.2 donne les lois des instants de d'occurrence d'un nombre donné d'événements.

Proposition A.0.4. *La loi de l'instant T_1 de la première occurrence a pour densité*

$$f_{T_1}(t) = \xi(t)\mathrm{e}^{-m(t)} \,. \tag{A.8}$$

La loi de l'instant T_{n+1} conditionnée à $\boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}$ a pour densité

$$f_{\left(T_{n+1}|\boldsymbol{T^{(n)}}=\boldsymbol{t^{(n)}}\right)}(t) = \xi(t)\mathrm{e}^{m(t_n)-m(t)}\mathbb{1}_{0<t_1<\cdots<t_n} \,. \tag{A.9}$$

La loi jointe du vecteur $\boldsymbol{T^{(n)}}$ a pour densité

$$f_{\boldsymbol{T^{(n)}}}(\boldsymbol{t^{(n)}}) = \left[\prod_{i=1}^{n}\xi(t_i)\right]\mathrm{e}^{-m(t_n)}\mathbb{1}_{t_1<\cdots<t_n} \,. \tag{A.10}$$

Une propriétés très utile des processus de Poisson inhomogènes concerne la loi conditionnée à $N(t) = n$ du vecteur $\boldsymbol{T^{(n)}} = (T_1, \ldots, T_n)$.

Proposition A.0.5. *la loi conditionnée à $N(t) = n$ du vecteur $\boldsymbol{T^{(n)}}$ a même loi que la statistique d'ordre d'un échantillon de taille n de la loi de densité $\dfrac{\xi(u)}{m(t)}\mathbb{1}_{u\in[0\,;\,t]}$, c'est-à-dire :*

$$f_{\left(\boldsymbol{T^{(n)}}\,|\,N(t)=n\right)}(\boldsymbol{t^{(n)}}) = n!\left[\prod_{i=1}^{n}\frac{\xi(t_i)}{m(t)}\mathbb{1}_{t_i\in[0\,;\,t]}\right]\mathbb{1}_{0<t_1<\cdots<t_n} \,.$$

Démonstration de la proposition A.0.5. Par la formule de Bayes :

$$\begin{aligned}
f_{\left(\boldsymbol{T^{(n)}}\,|\,N(t)=n\right)}(\boldsymbol{t^{(n)}}) &= \frac{\mathbb{P}\left[N(t)=n\,|\,\boldsymbol{T^{(n)}}=\boldsymbol{t^{(n)}}\right]f_{\boldsymbol{T^{(n)}}}(\boldsymbol{t^{(n)}})}{\mathbb{P}\left[N(t)=n\right]} \\
&= \frac{\mathbb{P}\left[T_{n+1}>t\,|\,\boldsymbol{T^{(n)}}=\boldsymbol{t^{(n)}}\right]f_{\boldsymbol{T^{(n)}}}(\boldsymbol{t^{(n)}})}{\mathbb{P}\left[N(t)=n\right]} \\
&= \frac{\mathrm{e}^{m(t_n)-m(t)}\left[\prod_{i=1}^{n}\xi(t_i)\right]\mathrm{e}^{-m(t_n)}\mathbb{1}_{0<t_1<\cdots<t_n}}{\mathrm{e}^{-m(t)}\dfrac{m(t)^n}{n!}} \\
&= n!\left[\prod_{i=1}^{n}\frac{\xi(t_i)}{m(t)}\mathbb{1}_{t_i\in[0\,;\,t]}\right]\mathbb{1}_{t_1<\cdots<t_n} \,.
\end{aligned}$$

□

Cette propriété peut également s'écrire sous la forme suivante.

Proposition A.0.6. *La loi conditionnée à $N(t) = n$ du vecteur $\left(\dfrac{m(T_1)}{m(t)}, \dots, \dfrac{m(T_n)}{m(t)}\right)$ est la même que la statistique d'ordre d'un échantillon de taille n de la loi uniforme sur $[0\,;1]$.*

Démonstration de la proposition A.0.6. D'après la proposition A.0.5, la loi conditionnée à $N(t) = n$ du vecteur $\boldsymbol{T^{(n)}}$ a même loi que la statistique d'ordre d'un échantillon de taille n de la loi de densité $\dfrac{\xi(u)}{m(t)}\mathbb{1}_{u\in[0\,;t]}$. La fonction de répartition de cette loi s'écrit

$$F(u) = \frac{m(u)}{m(t)}\mathbb{1}_{0\leqslant u\leqslant t} + \mathbb{1}_{u>t}\,.$$

Ainsi, le vecteur $(F(T_1), \dots, F(T_n))$ a conditionnellement à $N(t) = n$ même loi que la statistique d'ordre d'un échantillon de taille n de la loi uniforme sur $[0\,;1]$. □

Bibliographie

[1] L.J.S. Allen. *Stochastic processes with applications to Biology*. 2e. Chapman et Hall/CRC, 2010.

[2] W.P. Angerer. « A note on the evaluation of fluctuation experiments ». *Mutation Research* 479 (2001), p. 207–224.

[3] W.P. Angerer. « An explicit representation of the Luria-Delbrück distribution ». *J. Math. Biol.* 42.2 (2001), p. 145–174.

[4] P. Armitage. « The statistical theory of bacterial populations subject to mutation ». *J. R. Statist. Soc. B* 14 (1952), p. 1–40.

[5] K.B. Athreya et S.N. Lahiri. *Mesure Theory and Probability Theory*. New York : Springer, 2006.

[6] K.B. Athreya et P.E. Ney. *Branching Processes*. Berlin Heidelberg : Springer, 1972.

[7] N.T.J. Bailey. *The Elements of Stochastic Processes with Applications to the Natural Sciences*. New York : Wiley, 1964.

[8] M.S. Bartlett. *An Introduction to Stochastic Processes, with Special Reference to Methods and applications*. 3e. Cambridge University Press, 1978.

[9] R. Bellman et T. Harris. « On age-dependent binary branching processes ». *Ann. Math.* 55.2 (1952), p. 280–295.

[10] I. Benjamini et Y. Peres. « Markov chains indexed by trees ». *Ann. Probab.* 22.1 (1994), p. 219–243.

[11] L. Boe, T. Tolker-Nielsen, K.M. Eegholm, H. Spliid et A. Vrang. « Fluctuation analysis of mutations to nalidixic acid resistance in *Escherichia Coli* ». *J. Bacteriol.* 176.10 (1994), p. 2781–2787.

[12] S. Borrell, Y. Teo, F. Giardina, E.M. Streicher, M. Klopper, J. Feldmann, B. Müller, T.C. Victor et S. Gagneux. « Epistasis between antibiotic resistance mutations drives the evolution of extensively drug-resistant tuberculosis ». *Evol. Med. Public Health.* 2013.1 (2013), p. 65–74. doi : 10.1093/emph/eot003.

[13] R. Breunig. « An almost unbiased estimator of the coefficient of variation ». *Econ. Lett.* 70.1 (2001), p. 15–19.

[14] N. Champagnat et A. Lambert. « Adaptive dynamics in logistic branching populations ». *Banach Center Publ.* 80 (2008), p. 235–244.

[15] D.R. Cox. « Regression Models and Life-Tables ». *J. R. Stat. Soc. (Series B)* 34.2 (1972), p. 187–220.

[16] K.S. Crump et D.G. Hoel. « Mathematical models for estimating mutation rates in cell populations ». *Biometrika* 61 (1974), p. 237.

[17] H.L. David. « Probability distribution of drug-resistant mutants in unselected populations of *Mycobacterium tuberculosis* ». *Appl. Microbiol.* 20.5 (1970), p. 810–814.

[18] A. Dewanji, E.G. Luebeck et S.H. Moolgavkar. « A generalized Luria-Delbrück model ». *Math. Biosci.* 197.2 (2005), p. 140–152.

[19] D. Eddelbuettel. *Seamless R and C++ Integration with Rcpp*. New York : Springer, 2013.

[20] D. Eddelbuettel et C. Sanderson. « RcppArmadillo : Accelerating R with high-performance C++ linear algebra ». *Comput. Stat. Data An.* 70 (2014), p. 1054–1063.

[21] P. Embrechts et J. Hawkes. « A limit theorem for tails of discrete infinitely divisible laws with applications to fluctuation theory ». *J. Austral. Math. Soc. Series A* 32 (1982), p. 412–422.

[22] A. Etheridge. *Some Mathematical Models from Population Genetics*. Cours donné à la 39e École d'Éte de Saint-Flour, 2009. Heidelberg : Springer, 2011.

[23] W.J. Ewens. *Mathematical Population Genetics*. 2e. Berlin : Springer, 2004.

[24] F. Fontaine, E.J. Stewart, A.B. Lindner et F. Taddei. « Mutations in two global regulators lower individual mortality in *Escherichia Coli* ». *Mol. Microbio.* 67.1 (2008), p. 2–14.

[25] P.L. Foster. « Methods for Determining Spontaneous Mutation Rates ». *Method. Enzymol.* 409 (2006), p. 195–213.

[26] S. Gagneux, C.D. Long, P.M. Small, T. Van, G.K. Schoolnik et B.J M. Bohannan. « The competitive cost of antibiotic resistance in *Mycobacterium tuberculosis* ». *Science* 312 (2006), p. 1944–1946.

[27] O. Gaudoin. « Fiabilité des Systèmes Réparables ». Notes de cours, http://www-ljk.imag.fr/membres/Olivier.Gaudoin/FiRep.pdf.

[28] P. Gerrish. « A simple formula for obtaining markedly improved mutation rate estimates ». *Genetics* 180.3 (2008), p. 1773–1778.

[29] A. Gillet-Markowska, G. Louvel et G. Fisher. « bz-rates : a web-tool to estimate mutation rates from fluctuation analysis ». *G3* 5.11 (2015), p. 2323–2327.

[30] B.M. Hall, C.X. Ma, P. Liang et K.K. Singh. « Fluctuation AnaLysis CalculatOR : a web tool for the determination of mutation rate using Luria–Delbrück fluctuation analysis ». *Bioinformatics* 25.12 (2009), p. 1564–1565.

[31] A. Hamon et B. Ycart. « Statistics for the Luria-Delbrück distribution ». *Elect. J. Statist.* 6 (2012), p. 1251–1272.

[32] T. Harko, F.S.N. Lobo et M.K. Mak. « Analytical Solutions of the Riccati Equation with Coefficients Satisfying Integral or Differential Conditions with Arbitrary Functions ». *Univ. J. Appl. Math.* 2.2 (2014), p. 109–118.

[33] T.E. Harris. *The Theory of Branching Processes*. Berlin : Springer, 1963.

[34] B. Houchmandzadeh. « General formulation of Luria-Delbrück distribution of the number of mutants ». *Physical Review E : Statistical, Nonlinear, and Soft Matter Physics* 92 (2015), p. 012719.

[35] G. Jaeger et S. Sarkar. « On the distribution of bacterial mutants : the effects of differential fitness of mutants and non-mutants ». *Genetica* 96 (1995), p. 217–223.

[36] P. Jagers. « Stabilities and instabilities in population dynamics ». *J. Appl. Probab.* 29 (1992), p. 770–780.

[37] L.W. Jean, M.T. Suchorolski, J. Jeon et E.G. Luebeck. « Multiscale estimation of cell kinetics ». *Comput. Math. Meth. Med.* 11.3 (2010), p. 239–254. doi : 10.1080/17486700903535922.

[38] M.E. Jones. « Luria-Delbrück fluctuation experiments ; accounting simultaneously for plating efficiency and differential growth rate ». *J. Theo. Biol.* 166.3 (1994), p. 355–363.

[39] M.E. Jones, S.M. Thomas et A. Rogers. « Luria-Delbrück Fluctuation Experiments : Design and Analysis ». *Genetics* 136 (1994), p. 1209–1216.

[40] C.D. Kelly et O. Rahn. « The growth rate of individual bacterial cells ». *J. Bacteriol.* 23.2 (1932), p. 147–153.

[41] W.S. Kendal et P. Frost. « Pitfalls and practice of Luria-Delbrück fluctuation analysis : a review ». *Cancer research* 48.5 (1988), p. 1060–1065.

[42] D.G. Kendall. « On the generalized “birth-and-death" process ». *ANN. Math. Statist.* 19 (1948), p. 1–15.

[43] D.G. Kendall. « On the role of variable generation time in the development of a stochastic birth process ». *Biometrika* 35 (1948), p. 316–330.

[44] D.G. Kendall. « Les processus stochastiques de croissance en biologie ». *Ann. IHP* 13.1 (1952), p. 43–108.

[45] D.G. Kendall. « On the Choice of a Mathematical Model to Represent Normal Bacterial Growth ». *Journal of the Royal Statistical Society* 14.1 (1952), p. 41–44.

[46] M. Kimmel et D.E. Axelrod. *Branching Processes in Biology*. New York : Springer, 2002.

[47] A.L. Koch. « Mutation and growth rates from Luria-Delbrück fluctuation tests ». *Mutation Res.* 95 (1982), p. 129.

[48] N.L. Komarova, L. Wu et P. Baldi. « The fixed-size Luria-Delbrück model with a nonzero death rate ». *Math. Biosci.* 210.1 (2007), p. 253–290.

[49] K.P. Koutsoumanis et A. Lianou. « Stochasticity in colonial growth dynamics of individual bacterial cells ». *Appl. Environ. Microbiol.* 79.7 (2013), p. 2294–2301.

[50] H.E. Kubitschek. « The distribution of cell generation times ». *Cell Tissue Kinet.* 4 (1971), p. 113–122.

[51] V. Kucera. « A review of the matrix Riccati Equation ». *Kybernetika* 9.1 (1973), p. 42–61.

[52] T. Kuzcek. « Almost sure limit results for the supercritical Bellman-Harris process ». *J. Appl. Probab.* 19.3 (1982), p. 668–674.

[53] T. Kuzcek. « On the convergence of the empiric age distribution for one dimensional supercritical age dependent branching processes ». *Ann. Probab.* 10.1 (1982), p. 252–258.

[54] A.K. Laird. « Dynamics of tumor growth ». *Brit. J. Cancer* 18 (1964), p. 490–502.

[55] A. Lambert. « The branching process with logistic growth ». *Ann. Appl. Probab.* 15.2 (2005), p. 1506–1535.

[56] D.E. Lea et C.A. Coulson. « The distribution of the number of mutants in bacterial populations ». *J. Genet.* 49.3 (1949), p. 264–285.

[57] E.L. Lehmann et G. Casella. *Theory of Point Estimation*. 2e. Springer Texts in Statistics. New York : Springer, 2003.

[58] S. Louhichi et B. Ycart. « Exponential growth of bifurcating processes with ancestral dependence ». *Adv. Appl. Probab.* 47.2 (2015), p. 545–564.

[59] S.E. Luria. « The frequency distribution of spontaneous bacteriophage mutants as evidence for the exponential rate of phage reproduction ». *Cold Spring Harbor Symp. Quant. Biol.* 16 (1951), p. 463–470.

[60] S.E. Luria et M. Delbrück. « Mutations of bacteria from virus sensitivity to virus resistance ». *Genetics* 28.6 (1943), p. 491–511.

[61] W.T. Ma, G.v.H. Sandri et S. Sarkar. « Analysis of the Luria-Delbrück distribution using discrete convolution powers ». *J. Appl. Probab.* 29.2 (1992), p. 255–267.

[62] Y. Ma, M.G. Genton et E. Parzen. « Asymptotic properties of sample quantiles of discrete distributions ». *Ann. Inst. Stat. Math.* 63 (2011), p. 227–243.

[63] M. Marcheselli, A. Baccini et L. Barabesi. « Parameter estimation for the discrete stable family ». *Commun. Statist. Theory Methods* 37.6-7 (2008), p. 815–830.

[64] J.H. Matis et T.R. Kiffe. « Effects of immigration on some stochastic logistic model : A cumulant truncation analysis ». *Theoret. Popul. Biol.* 56 (1999), p. 139–161.

[65] A. Mazoyer. « Fluctuation analysis application using flan ». (in progress). 2017. URL : https://github.com/AdriMaz/ShinyFlan/.

[66] A. Mazoyer. « Time inhomogeneous mutation models with birth-date dependence ». (revised). 2017. URL : https://hal.archives-ouvertes.fr/hal-01415995.

[67] A. Mazoyer, B. Ycart et N. Veziris. « Correction : Unbiased Estimation of Mutation Rates under Fluctuating Final Counts ». *PLOS ONE* 12.3 (2017), p. 1–4. URL : http://dx.doi.org/10.1371%2Fjournal.pone.0173143.

[68] A. Mazoyer, R. Drouilhet, S. Despréaux et B. Ycart. « flan : An R package for inference on mutation models ». version 0.5. 2017. URL : https://cran.r-project.org/package=flan.

[69] A. Mazoyer, R. Drouilhet, S. Despréaux et B. Ycart. « flan : An R package for inference on mutation models ». 2017. URL : https://github.com/AdriMaz/flan/.

[70] A. Mazoyer, R. Drouilhet, S. Despréaux et B. Ycart. « flan : An R package for inference on mutation models ». *The R Journal* (2017). URL : https://journal.r-project.org/archive/2017/RJ-2017-029/index.html.

[71] P.L. de Micheaux, R. Drouilhet et B. Liquet. *The R Software : Fundamentals of Programming and Statistical Analysis*. New York : Springer, 2013.

[72] J.S. Murphy, F.R. Landsberger, T. Kikuchi et I. Tamm. « Occurrence of cell division is not exponentially distributed : differences in the generation times of sister cells can be derived from the theory of survival of populations ». *Proc. Natl. Acad. Sci. USA* 81 (1984), p. 2379–2384.

[73] I. Nasell. « Extinction and quasi-stationnary in the Verhlust logistic model ». *J. Theoret. Biol.* 211 (2001), p. 11–27.

[74] H.T. Nguyen. *An Introduction to Random Sets*. Boca Raton : Chapman & Hall/CRC, 2006.

[75] A.G. Pakes. « Remarks on the Luria-Delbrück distribution ». *J. Appl. Probab.* 30.4 (1993), p. 991–994.

[76] R. Pemantle. « Tree-indexed processes ». *Statist. Sci.* 10.2 (1995), p. 200–213.

[77] R Development Core Team. *R : A Language and Environment for Statistical Computing*. Vienna : R Foundation for Statistical Computing, 2008.

[78] O. Rahn. « A chemical explanation of the variability of the growth rate ». *J. Gen. Physiol.* 15 (1932), p. 257–277.

[79] W.A. Rosche et P.L. Foster. « Determining mutation rates in bacterial populations ». *Methods* 20.1 (2000), p. 1–17. doi : 10.1006/meth.1999.0901.

[80] B. Rémillard et R. Theodorescu. « Inference based on the empirical probability generating function for mixtures of Poisson distributions ». *Statist. Decisions* 18 (2000), p. 349–366.

[81] S. Sarkar. « Haldane's solution of the Luria-Delbrück distribution ». *Genetics* 127 (1991), p. 257–261.

[82] S. Sarkar, W.T. Ma et G.v.H. Sandri. « On fluctuation analysis : a new, simple and efficient method for computing the expected number of mutants ». *Genetica* 85 (1992), p. 173–179.

[83] E.J. Stewart, R. Madden, G. Paul et F. Taddei. « Aging and death in an organism that reproduces by morphologically symmetric division ». *PLoS Biology* 3.2 (2005), p. 295–300.

[84] F.M. Stewart. « Fluctuation analysis : the effect of plating efficiency ». *Genetica* 84.1 (1991), p. 51–55.

[85] F.M. Stewart. « Fluctuation Tests : How Reliable Are the Estimates of Mutation Rates ? » *Genetics* 137.4 (1994), p. 1139–1146.

[86] F.M. Stewart, D.M. Gordon et B.R. Levin. « Fluctuation analysis : the probability distribution of the number of mutants under different conditions ». *Genetics* 124.1 (1990), p. 175–185.

[87] W.Y. Tan. « On distribution theories for the number of mutants in cell populations ». *SIAM J. Appl. Math.* 42.4 (1982), p. 719–730.

[88] W.Y. Tan. « A stochastic Gompertz birth-death process ». *Statist. Probab. Lett.* 4.4 (1986), p. 25–28.

[89] W.Y. Tan et S. Piantadosi. « On stochastic growth processes with application to stochastic logistic growth ». *Statist. Sinica* 1 (1991), p. 527–540.

[90] P.F. Verhulst. « Notice sur la loi que la population suit dans son accroissement ». *Correspondance mathématique et physique*. Sous la dir. de J.G. Garnier et A. Quetelet. T. 10. Société Belge de Librairie, Bruxelles, 1838, p. 113–121.

[91] P.F. Verhulst. « Recherches mathématiques sur la loi d'accroissement de la population ». *Nouveaux mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles* 18 (1845), p. 14–54.

[92] P. Wang, L. Robert, J. Pelletier, W.L. Dang, F. Taddei, A. Wright et S. Jun. « Robust growth of *Escherichia Coli.* » *Curr. Biol.* 20 (2010), p. 1099–1103.

[93] L. Wasserman. *All of Statistics : a concise course in statistical inference*. New York : Springer, 2004.

[94] J. Werngren et S.E. Hoffner. « Drug susceptible *Mycobacterium tuberculosis Beijing* genotype does not develop motation-conferred resistance to Rifampin at an elevated rate ». *J. Clin. Microbiol.* 41.4 (2003), p. 1520–1524.

[95] R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing*. 3e. Amsterdam : Elsevier, 2012.

[96] X. Wu, E.D. Strome, Q. Meng, P.J. Hastings, S.E. Plon et M. Kimmel. « A robust estimator of mutation rates ». *Mut. Res.* 661.1-2 (2009), p. 101–109.

[97] B. Ycart. « Fluctuation analysis : can estimates be trusted ? » *PLoS One* 8.12 (2013), p. 1–12. URL : http://dx.doi.org/10.1371%2Fjournal.pone.0080958.

[98] B. Ycart. « Fluctuation analysis with cell deaths ». *J. Appl. Probab. Statist* 9.1 (2014), p. 13–29.

[99] B. Ycart et N. Veziris. « Unbiased estimates of mutation rates under fluctuating final counts ». *PLoS One* 9.7 (2014), p. 1–10. URL : http://dx.doi.org/10.1371/journal.pone.0101434.

[100] G.U. Yule. « A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S. » *Phil. Trans. Roy. Soc. London Ser. B* 213 (1925), p. 21–87.

[101] Q. Zheng. « Progress of a half century in the study of the Luria-Delbrück distribution ». *Math. Biosci.* 162 (1999), p. 1–32.

[102] Q. Zheng. « Statistical and algorithmic methods for fluctuation analysis with SALVADOR as an implementation ». *Math. Biosci.* 176.2 (2002), p. 237–252.

[103] Q. Zheng. « New algorithms for Luria-Delbrück fluctuation analysis ». *Math. Biosci.* 196.2 (2005), p. 198–214.

[104] Q. Zheng. « A note on plating efficiency in fluctuation experiments ». *Math. Biosci.* 216 (2007), p. 150–153.

[105] Q. Zheng. « On Haldane's formulation of Luria-Delbrück's mutation model ». *Math. Biosci.* 209 (2007), p. 500–513.