Fluctuation analysis on mutation models with birth-date dependence

Adrien Mazoyer

Fluctuation analysis on mutation models with birth-date dependence

Adrien Mazoyer

November 17, 2017

Abstract

The classic Luria-Delbrück model can be interpreted as a Poisson compound (number of mutations) of exponential mixtures (developing time of mutant clones) of geometric distributions (size of a clone in a given time). This "three-ingredients" approach is generalized in this paper to the case where the split instant distributions of cells are not i.i.d. : the lifetime of each cell is assumed to depend on its birth date. This model takes also into account cell deaths and non-exponentially distributed lifetimes. Previous results on the convergence of the distribution of mutant counts are recovered. The particular case where the instantaneous division rates of normal and mutant cells are proportional is studied. The classic Luria-Delbrück and Haldane models are recovered. Probability computations and simulation algorithms are provided. Robust estimation methods developed for the classic mutation models are extended to the new model: their properties of consistency and asymptotic normality remain true; their asymptotic variance are computed. Finally, the estimation biases induced by considering classic mutation models instead of inhomogeneous model are studied with simulation experiments.

1 Introduction

Mutation models are probabilistic descriptions of the growth of a population of cells, in which scarce mutations randomly occur. Data are samples of integers, interpreted as final numbers of mutant cells. The frequent appearance in the data of very large mutant counts, usually called "jackpots", evidences heavy-tailed probability distributions. Mutation models have two objectives: study the distribution of the number of mutant cells at the end of the growth process; perform fluctuation analysis on data to estimate the probability for a mutant to appear at any division.

Any classic mutation model can be interpreted as the result of the three following ingredients [10]:

1. a random number of mutations occurring 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 expectation of that distribution is the product of the mutation probability by the total number of divisions;

2. from each mutation, a clone of mutant cells growing during a random time. Due to exponential growth, most mutations occur close to the end of the process, and the developing time of a random clone has exponential distribution. The rate of that distribution is the *relative fitness*, i.e., the ratio of the growth rate of normal cells to that of mutants;

3. the number of mutant cells that any clone developing for a given time will produce. The distribution of this number depends on the modeling assumptions, in particular the lifetimes of mutants.

This approach leads to a family of probability distributions which depend on the expected number of mutations and the relative fitness. One of the most used mutation models is the well known Luria-Delbrück model [20]. Mathematical descriptions were introduced by Lea and Coulson [18], followed by Armitage [3] and Bartlett [5]. In that model, division times of mutant cells were supposed to be exponentially distributed. Thus a clone develops according to a Yule process [41, p. 35]; [4, p. 109], and its size at any given time follows a geometric distribution. The distribution of final mutant counts is also explicit when lifetimes of mutant cells are supposed to be constant. This latter model is called Haldane model by Sarkar [29]; an explicit form of the asymptotic distribution is given by Ycart [38]. General lifetimes have also been studied in [38], but no explicit distribution is available apart from the exponential and constant lifetimes. Other extensions of the Luria-Delbrück model take into account the case where cells have a certain probability to die rather than divide [2, sec. 3.1]; [6, 14, 39], where final number of cells are random [2, 14, 40].

As mentioned above, the main statistical objective of mutation models is the estimation of the probability for a mutant cell to appear upon any given cell division. It is computed dividing estimate of the mean number of mutations by the mean final number of cells. Computing robust estimates is of crucial importance in medical applications, like cancer tumor relapse or multidrug resistance of *Mycobacterium Tuberculosis* for instance. Estimates are realizations of an estimator which is a random variable depending on the considered sample. A robust estimator satisfies two properties: consistency, and explicit asymptotic distribution. Thus confidence intervals and p-values can be computed.

Luria and Delbrück [20] have proposed two estimators. The first estimator, called p_0 estimator, is based on the relation between the probability to get a null count in the sample and the mean number of mutations: taking the negative logarithm of the relative frequency of zeros among the sample gives a robust estimate of the expected number of mutations [20, eq. (5)]. Remark that if the sample does not contain null counts, the method cannot be applied. The second estimator proposed in the same article is based on

the relation between the mean number of mutants, the sample size, and the final mutant of cells [20, eq. (8)]. Since this estimator does not have expectation, it is not consistent and should not be used. A wide panel of estimation methods has been proposed since then [27, 9]. Most of these methods are based on empirical median of the mutant count to reduce the heavy tail effects [18, eq. (25)], [13, eq. (6)]. Even if some median methods give good results in practice, the consistency property is not satisfied or cannot be checked: indeed the empirical median is not a robust estimator of the median for discrete distributions. Thus other methods which satisfy the properties of interest should be considered. Since the distribution of final numbers has an explicit form, the Maximum Likelihood (ML) seems to be an obvious optimal choice [21, 31, 42]. However, the computation of the likelihood and its derivatives can be numerically unstable because of the jackpots. One of the possible ways to reduce such tail effects is "Winsorization" of the sample [37, sec. 2.2], which consists in replacing any value of the sample that pass a certain bound by the bound itself. The last method exposed here uses the probability generating function [28, 10], and is called Generating Function (GF) method. This method is comparable in terms of efficiency to the ML method. Moreover, this method has a good numerical stability and a negligible computing time. The p_0, ML, and GF methods provide asymptotically normal estimators of the mean number of mutation. The estimation of the mutation probability can then be deduced dividing estimation of mean number of mutations by the mean final number of cells. The fluctuations of the final number of cells can also be taken into account [40], in order to get a more accurate estimation. Sometimes, data are samples of couples of integers, interpreted as final numbers of mutants and final numbers of cells. In that case, the Maximum Likelihood can be used to estimate directly the mutation probability [40]. Moreover, the relative fitness can also be estimated.

The lifetimes of the cells are supposed to be i.i.d. in the classic mutation models. Thus, the population grows exponentially or dies out [4, 12]. However, the growth is in practice logistic [16]: it is exponential until an inflexion instant when the growth begins to slow and eventually tends to zero. Indeed, during an experiment, a colony of cells grows in an environment which contains a finite amount of resources. Then a cell born at a instant s_1 will complete its lifetime faster than a cell born at a instant $s_2 \ll s_1$. The Verhulst model [35] is one of the most known deterministic growth model which takes into account this limitation. Logistic-type stochastic models are described by Allen [1, sec. 9.4.2], and mathematically studied by several authors among which [33, 34, 17]. Stewart et al. [32] proposed an approach to take into account the decreasing rate of division as the cells run out of resources. Houchmandzadeh [11] described a discrete approach with a general growth model for mutant clones. However, none of these studies provide results for the non-i.i.d. lifetimes case, in particular on the distribution of final mutant count.

In a previous work [23], an extension of the classic mutation models to the case where the split instant of a cell depends on its birth date has been proposed. The results on the asymptotic distribution of the mutant count were very similar to that of classic Luria-Delbrück model. Therefore the methods of estimation described above should be directly

adapted to the model with birth-date dependence. As for the homogeneous case, the three methods provide consistent and asymptotically normal estimator for the parameters of interest. However, fast simulation cannot be deduced from the approach exposed in [23]. Such algorithms are necessary to perform large scale simulation studies. Another approach of the model is proposed here: as for the homogeneous mutation models, the distribution of the final mutant count can still be interpreted as the result of three ingredients. As a direct consequence, a fast simulation algorithm can be deduced. The asymptotic results on the distribution of the mutant count of [23] are recovered and extended to the case where the death of normal cells are taken into account.

General modeling assumptions are described in Section 2. The three-ingredients approach is exposed and used to prove the convergence in distribution of the final mutant count in the Section 3. Probability computation and simulation algorithms are exposed in Section 4. The case where the hazard functions associated to the split instant distribution of normal and mutant cells are proportional is studied. In particular, the Luria-Delbrück distribution with cell deaths [39] is recovered. The Haldane model is also recovered, and extended to the case where mutant cell deaths are taken into account. The p_0, ML, and GF methods are generalized to the inhomogeneous model in these specific cases in Section 5. Estimation biases induced by considering classic mutation models instead of model with birth-date dependence are illustrated with simulation experiments in Section 6.

2 Hypotheses and models

Notations and hypotheses are described in this section. A rigorous definition of the probabilistic model as a tree-indexed process has already been given in [23, sec. 2]. Thus, the dynamics are shortly described, and the modeling assumptions will be summarized at the end of this section.

Consider a normal cell born at a given instant s. At a random instant (called here a *final instant*) with cumulative distribution function (cdf) $F_\nu(s,\cdot)$, the cell produces one normal and one mutant cell with probability π (this event is called a *mutation*), two normal cells with probability $1-\pi-\gamma$, or dies with probability γ. Consider now a mutant cell, born at a given time s. At a random instant with cdf $F_\mu(s,\cdot)$, the mutant produces two mutant cells with probability $1-\delta$ or dies with probability δ. Starting from a single cell, whatever its nature, the set of all descendants constitutes a *clone*. Thus the *clone size* at a given time t denotes the number of cells alive at time t in the clone. Consider a given cell, the mutation or death events are independent from its final instant. Two cells are independent conditionally on their common ancestor. Therefore, the clones stemming from these cells are also independent conditionally on this ancestor. Remark that those dependence assumptions hold whatever the nature of the considered cells. At the beginning of the process, the population contains a given number n of normal cells and no mutants.

Some details about F_ν and F_μ are now given. The final instant of a cell born at a given time $s \geqslant 0$ cannot be smaller than s. Thus, both cdf satisfies $F_\nu(s,t) = 0$ and $F_\mu(s,t) = 0$ for $t \leqslant s$. Moreover, the total number of cells is in practice bounded by the carrying capacity. It corresponds to the maximum sustainable population: the closer to this bound the number of cells, the slower the growth of the population. In other words, some cells do not produce descendants before the end of the growth process. Thus, the distribution of the final instant of any cell may have a positive mass at infinity. The cdfs $F_\nu(s,\cdot)$ and $F_\mu(s,\cdot)$ are cdfs of subprobability measures on $\mathbb{R}_+$, i.e. the limit of $F_\nu(s,t)$ and $F_\mu(s,t)$ as t tends to infinity may be strictly smaller than 1 for any s in $\mathbb{R}_+$. For more details about subprobability measures, see for example [25, p. 170]. Thus, $F_\nu(s,\cdot)$ and $F_\mu(s,\cdot)$ are assumed to be cdfs on the extended real line $\overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{+\infty\}$ for any $s \in \mathbb{R}_+$. Construction of such cdfs is described in [23]. Additive assumptions on the cdf $F_\nu(s,\cdot)$ are now precised. For any $s \geqslant 0$, let $F(s,\cdot)$ be a cdf on $\overline{\mathbb{R}}_+$ such that $F(s,t) = 0$ if $t \leqslant s$. The cdf F will satisfy $(\mathcal{H})$ if there exists a cdf of subprobability on $\mathbb{R}_+$, denoted by $\tilde{F}(s,\cdot)$, such that the following holds:

$(\mathcal{H}_1)$ the cdf $\tilde{F}$ is differentiable with respect to s and t, and decreasing in s;

$(\mathcal{H}_2)$ $\lim_{t\to+\infty} \tilde{F}(s,t) \leqslant 1$ for all $s \in \mathbb{R}_+$ and $\tilde{F}(s,t) = 0$ if $t \leqslant s$;

$(\mathcal{H}_3)$ for any $s \geqslant 0$, $F(s,\cdot)$ is deduced from $\tilde{F}(s,\cdot)$ by

$$F(s,t) = \tilde{F}(s,t)\mathbb{1}_{t\in[0\,;+\infty)} + \mathbb{1}_{t=+\infty}\,;$$

$(\mathcal{H}_4)$ the function h defined for all $(s,t) \in \mathbb{R}_+^2$ by

$$h(s,t) = -\log\left(1 - \tilde{F}(s,t)\right),$$

satisfies for any $t \geqslant s$:

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

By definition, h is positive, differentiable with respect to s and t, increasing in t, decreasing in s, and for any $(s,t) \in \mathbb{R}_+^2$:

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

There exists a positive, continuous, $\mathbb{R}_+$-valued function λ such that:

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

The function h can be interpreted as the cumulative hazard rate on an interval $[s\,;t]$ associated to F. The function λ can be interpreted as the instantaneous hazard rate

associated to F. The cdf $F(s,\cdot)$ is then defined on $\overline{\mathbb{R}}_+$ for any $s \in \mathbb{R}_+$ by

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

Moreover, consider a positive, continuous, and $\mathbb{R}_+$-valued function λ. Then, the cdf deduced from λ by (1) satisfies $(\mathcal{H})$. In particular, if $\lambda(t)$ tends to 0 as t tends to infinity, then the limit of $h(0,t)$ is finite. In that case, the limit of $\tilde{F}(s,t)$ is smaller than 1 for any $s \geqslant 0$.

Notice also that if $T(s)$ is a random variable with cdf $F(s,\cdot)$, such that F satisfies $(\mathcal{H})$, then:

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

This is quite similar to the *memorylessness* property of exponential distributions.

Condition $(\mathcal{H}_4)$ is quite restrictive. However, the mean growth of the population can be adjusted to any positive, continuous, increasing and $\mathbb{R}_+$-valued function [23]. Remark that condition $(\mathcal{H}_4)$ is equivalent to satisfy:

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

Therefore, for any $\mathbb{R}_+$-valued cdf G, the cdf $\tilde{F}$ defined by

$$\tilde{F}(s,t) = \frac{G(t) - G(s)}{1 - G(s)}\,,$$

satisfies $(\mathcal{H}_4)$.

From now on, F_ν is assumed to satisfy $(\mathcal{H})$. Its related function defined in $(\mathcal{H}_4)$ will be denoted by h_ν. The limit of $h_\nu(0,t)$ as t tends to infinity will be denoted by $h_{\nu,\infty}$. The instantaneous division rate associate to F_ν will be denoted by λ_ν. No additive assumptions on F_μ are required yet. The modeling assumptions can then be summarized as follows:

- at time 0, n normal cells are present;
- the final instant of any cell depends on its nature and its birth date;
- the final instant of a normal cell born at time s is a random variable with cdf $F_\nu(s,\cdot)$ which satisfies $(\mathcal{H})$;
- upon completion of the lifetime of a normal cell:

  - with probability π one normal and one mutant cells are produced;
  - with probability γ the cell dies out;
  - with probability $1-\pi-\gamma$ two normal cells are produced;
- the final instant of a mutant cell born at time s is a random variable with cdf $F_\mu(s,\cdot)$ defined on $\overline{\mathbb{R}}_+$;
- upon completion of the lifetime of a mutant cell:
  - with probability δ the cell dies out;
  - with probability $1-\delta$ two mutant cells are produced;
- for any cell, the events of death or mutation do not depend on its final instant;
- two cells, whatever their nature, are independent conditionally on their common ancestor;
- two clones are independent conditionally on the common ancestor of the two cells which started those clones.

3 "Three-ingredients" approach

The generalization of the "three-ingredients" interpretation exposed in the introduction is described in this section. Let $(\tau_n)_{n\in\mathbb{N}}$ be a sequence of observation instants, tending to infinity as n tends to infinity. Let $(\pi_n)_{n\in\mathbb{N}}$ be a sequence of mutation probabilities, tending to 0 as n tends to infinity. Moreover, assume

$$\lim_{n\to+\infty} \pi_n n \mathrm{e}^{h_\nu(0,\tau_n)} = \alpha\,,$$

where α is some fixed positive real number. Remark that the constant α corresponds in the classic case to the mean number of mutations. Denote by $Z_n(t)$ the number of mutations in an interval $[0\,;t]$, starting with n normal cells at time 0. The increasing sequence of mutation instants will be denoted by $\left(T_i^{(n)}\right)_{i\in\mathbb{N}}$.

Similarly to [10], a three-ingredients approach of model mutations can be given for the non-homogeneous birth date case.

Proposition 3.1. *Assume $\gamma = 0$. Let $\pi = (\pi_n)_{n\in\mathbb{N}}$ and $\tau = (\tau_n)_{n\in\mathbb{N}}$ two sequences, and α a positive real such that:*

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

Then:

($\mathcal{A}_1^{(0)}$) *As n tends to infinity, the distribution of the total number of mutations $Z_n(\tau_n)$ tends to the Poisson distribution with parameter*

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

where

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

($\mathcal{A}_2^{(0)}$) *As n tends to infinity, the joint distribution of the vector $\left(T_1^{(n)}, \ldots, T_k^{(n)}\right)$ of k fixed number of mutation instants in an interval $[0\,;t]$ converges to the order statistics of a k sample of the distribution:*

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

i.e. the distribution $\lambda_\nu(u)\mathrm{e}^{-h_\nu(u,t)}$ truncated on $[0\,;t]$.

In particular, from assertions ($\mathcal{A}_2^{(0)}$), the probability generating function (pgf) of the size at a given time t of any mutant clone is given by

$$\mathcal{I}(z,t) = \int_0^t \psi(z,u,t) \frac{\lambda_\nu(u)\mathrm{e}^{-h_\nu(u,t)}}{1 - \mathrm{e}^{-h_\nu(0,t)}} \mathrm{d}u , \tag{2}$$

where $\psi(z,s,t)$ is the pgf of the size at time t of a clone stemming from a mutant born at time s.

Proposition 3.1.

Assertion ($\mathcal{A}_1^{(0)}$) Consider the binary branching process with a single initial cell and without mutations. Denote by $N_1(t)$ the total number of cells at instant t. For any $t \geqslant 0$, consider the sequence $(N_n(t))_{n \in \mathbb{N}}$ defined for any $n > 0$ by

$$N_n(t) = \sum_{i=1}^{n} N_1^{(i)}(t) , \tag{3}$$

where the $N_1^{(i)}(t)$ are i.i.d. copies of $N_1(t)$. For any $n > 0$, $N_n(t)$ denotes the total number of cells living at time t in n independent copies of $N_1(t)$. According to Proposition 3.2. of [23]:

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

Let $\varepsilon > 0$. According to the law of large numbers, there exists for any $t > 0$ an integer $n_0(t)$ such that for any $n > n_0(t)$:

$$\mathbb{P}\left[\left|\frac{N_n(t)}{n\mathrm{e}^{h_\nu(0,t)}} - 1\right| < \left(1 - \mathrm{e}^{-h_\nu(0,t)}\right)\varepsilon\right] = 1 . \tag{4}$$

Since:

$$\frac{N_n(t)-n}{n\left(\mathrm{e}^{h_\nu(0,t)}-1\right)}-1=\frac{1}{1-\mathrm{e}^{-h_\nu(0,t)}}\left(\frac{N_n(t)}{n\mathrm{e}^{h_\nu(0,t)}}-\mathrm{e}^{-h_\nu(0,t)}\right)-1$$

$$=\frac{1}{1-\mathrm{e}^{-h_\nu(0,t)}}\left(\frac{N_n(t)}{n\mathrm{e}^{h_\nu(0,t)}}-1\right),$$

for any $n > n_0(t)$:

$$\mathbb{P}\left[\left|\frac{N_n(t)-n}{n\left(\mathrm{e}^{h_\nu(0,t)}-1\right)}-1\right|<\varepsilon\right]=1\,. \tag{5}$$

Thus, the number of cell divisions during the time interval $[0\,;t]$ is almost surely equivalent to $n\left(\mathrm{e}^{h_\nu(0,t)}-1\right)$. Let $\vartheta > 0$. Since τ_n tend to infinity, there exists $n_1 \in \mathbb{N}$ such that for any $n > n_1$:

$$\tau_n > \vartheta\,. \tag{6}$$

Therefore, according to (5) and (6), for any $n > \max(n_0(\vartheta), n_1)$:

$$\mathbb{P}\left[\left|\frac{N_n(\tau_n)-n}{n\left(\mathrm{e}^{h_\nu(0,\tau_n)}-1\right)}-1\right|<\varepsilon\right]=1\,. \tag{7}$$

In other words, the total number of divisions $N_n(\tau_n) - n$ is almost surely equivalent to $n\left(\mathrm{e}^{h_\nu(0,\tau_n)}-1\right)$. Since the mutant clones develop according to a different dynamic, the number of divisions of normal cells in a mutation model does not have the same distribution. However, since mutations are rare, the difference remains negligible: start with the n independent copies, and mark independently the division as potentially mutant with probability π_n. Denote by $(X_n(t))_{n\in\mathbb{N}}$ the sequence of the number of marked divisions at time t in the n copies. Since $\pi_n n\mathrm{e}^{h_\nu(0,\tau_n)}$ tends to α as n tends to infinity, the distribution of $X_n(\tau_n)$ converges to the Poisson distribution with parameter $m = \alpha(1-\mathrm{e}^{-h_{\nu,\infty}})$. Thus the number of marked divisions remains bounded in probability. Consider now that the clones stemming from the marked divisions are mutant clones. If a division is marked, all the marked divisions occurring in the clone stemming from it will be ignored. In other words, the number of mutations $Z_n(\tau_n)$ is smaller than the number of marked divisions $X_n(\tau_n)$ with probability 1. Moreover, since $X_n(\tau_n)$ is bounded in probability, the difference between $Z_n(\tau_n)$ and $X_n(\tau_n)$ is also bounded. Since this difference is bounded, the number of mutations occasions is equivalent in probability to $n\left(\mathrm{e}^{h_\nu(0,\tau_n)}-1\right)$. In other words, for any $\varepsilon > 0$, there exists $n_2 \in \mathbb{N}$ such that for any $n > n_2$:

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

Then the distribution of $Z_n(\tau_n)$ tends to the Poisson distribution with parameter m. Hence ($\mathcal{A}_1^{(0)}$).

Assertion ($\mathcal{A}_2^{(0)}$) From (5), the number of mutation occasions in an interval $[0\,;t]$ is equivalent in probability to $\lfloor n\left(\mathrm{e}^{h_\nu(0,t)}-1\right)\rfloor$. Then for any $k\in\mathbb{N}$ and any $t\geqslant 0$:

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

with

$$\iota_n(k,t)=\left(\begin{array}{c}\lfloor n\left(\mathrm{e}^{h_\nu(0,t)}-1\right)\rfloor\\ k\end{array}\right)\pi_n^k(1-\pi_n)^{\lfloor n\left(\mathrm{e}^{h_\nu(0,t)}-1\right)\rfloor-k}\,,$$

where $\lfloor x\rfloor$ is the only relative integer which satisfies for any $x\in\mathbb{R}$:

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

i.e. the integer part of x. Hence:

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

Moreover, by construction of $\{X_n(t)\}_{t\geqslant0}$, the process $\{Z_n(t)\}_{t\geqslant0}$ satisfies the following properties:

1. $\{Z_n(t)\}_{t\geqslant0}$ is simple, i.e. for any $t\in\mathbb{R}_+$:
$$\lim_{\Delta t\to0}\frac{1}{\Delta t}\mathbb{P}\left[Z_n(t+\Delta t)-Z_n(t)>1\right]=0\,;$$
2. $\{Z_n(t)\}_{t\geqslant0}$ has independent increments;
3. $Z_n(0)=0$ with probability 1.

Denote by ξ_n the intensity of the process $\{Z_n(t)\}_{t\geqslant0}$ defined for any $t\in\mathbb{R}_+$ by

$$\xi_n(t)=\lim_{\Delta t\to0}\frac{1}{\Delta t}\mathbb{P}\left[Z_n(t+\Delta t)-Z_n(t)=1\right]\,.$$

For any $\boldsymbol{t}=(t_1,\ldots,t_k)\in\mathbb{R}_+^k$, the event $\left(T_1^{(n)}=t_1,\ldots,T_k^{(n)}=t_k\right)$ will be denoted by $\boldsymbol{T^{(n)}}=\boldsymbol{t}$ For any $k\in\mathbb{N}$ and $\boldsymbol{t}\in\mathbb{R}_+^k$ and conditionally to $\boldsymbol{T^{(n)}}=\boldsymbol{t}$, the distribution of the $(k+1)$-th mutation instant $T_{k+1}^{(n)}$ is given by Proposition A.2

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

The joint distribution of the first k mutation instants $\boldsymbol{T^{(n)}}$ is also given by Proposition A.2:

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

Consider now the inhomogeneous Poisson process $\{Y_n(t)\}_{t\geqslant 0}$ with expectation:

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

For any $k \in \mathbb{N}$, denote by $\boldsymbol{S^{(n)}} = \left(S_1^{(n)}, \ldots, S_k^{(n)}\right)$ the k first occurrence instants of the process $\{Y_n(t)\}_{t\geqslant 0}$. Then, for any $k \in \mathbb{N}$, $t \in \mathbb{R}_+$ and $\boldsymbol{t} \in \mathbb{R}_+^k$, the three following assertions hold:

$$\lim_{n\to+\infty} \frac{\mathbb{P}\left[Z_n(t) = k\right]}{\mathbb{P}\left[Y_n(t) = k\right]} = 1\,, \tag{8}$$

and:

$$\lim_{n\to+\infty} \frac{f_{\boldsymbol{T^{(n)}}}(\boldsymbol{t})}{f_{\boldsymbol{S^{(n)}}}(\boldsymbol{t})} = 1\,, \tag{9}$$

and:

$$\lim_{n\to+\infty} \frac{f_{\left(T_{k+1}^{(n)} \,|\, \boldsymbol{T^{(n)}}=\boldsymbol{t}\right)}(t)}{f_{\left(S_{k+1}^{(n)} \,|\, \boldsymbol{S^{(n)}}=\boldsymbol{t}\right)}(t)} = 1\,. \tag{10}$$

Thus, for any $k \in \mathbb{N}$, $t \in \mathbb{R}_+$, and conditionally to $Z_n(t) = k$ the distribution of the vector $\boldsymbol{T^{(n)}}$ is given for any $\boldsymbol{t} \in \mathbb{R}_+^k$ by

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

Thus, according to (8), (9) and (10):

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

for any $k \in \mathbb{N}$. Conditionally to $Y_n(t) = k$, the vector $\boldsymbol{S^{(n)}}$ is distributed as the order statistics of k sample of the distribution (Proposition A.1):

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

Hence $(\mathcal{A}_2^{(0)})$. □

Remark that ($\mathcal{A}_2^{(0)}$) considers mutation instants, while similar results for homogeneous cases consider developing times of mutant clones. For example, according to Theorems 2.1 and 3.1 of [15], the joint distribution of developing times of k fixed mutant clones converges to the product of k independent exponential distributions.

The asymptotic pgf of the total mutant counts can be explicited, as long as the pgf ψ of a clone size is known.

Theorem 3.1. *Assume $\gamma = 0$. Let $\pi = (\pi_n)_{n\in\mathbb{N}}$ and $\tau = (\tau_n)_{n\in\mathbb{N}}$ two sequences, and α a positive real such that:*

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

As n tends to infinity, the pgf of the number of mutants at time τ_n starting with n normal cells tends to the pgf:

$$\phi(z) = \exp\left\{-m\left(1-\mathcal{I}(z)\right)\right\}\,, \tag{11}$$

where

$$\begin{aligned}\mathcal{I}(z) &= \lim_{t\to+\infty} \mathcal{I}(z,t) \\ &= \frac{1}{1-\mathrm{e}^{-h_{\nu,\infty}}} \lim_{t\to+\infty} \int_0^t \psi(z,u,t)\lambda_\nu(u)\mathrm{e}^{-h_\nu(u,t)}\mathrm{d}u\,.\end{aligned} \tag{12}$$

A first probabilistic interpretation of Theorem 3.1 is the generalization of the "three-ingredients" description exposed in the introduction as follows:

1. a random number of mutations occurring with small probability among a large number of cell divisions. According to ($\mathcal{A}_1^{(0)}$), the number of mutations approximately follows a Poisson distribution with expectation $m = \alpha\left(1-\mathrm{e}^{-h_{\nu,\infty}}\right)$;

2. each mutation appears at a random instant. According to ($\mathcal{A}_2^{(0)}$), any mutation instant asymptotically follows the distribution $\lambda_\nu(u)\mathrm{e}^{-h_\nu(u,t)}$ truncated on the interval $[0\,;t]$;

3. a mutant clone started at instant s develops according to a given process. Its size at time t follows the distribution with pgf $\psi(z,s,t)$.

Remark that the second ingredient concerns the instants at which a given number of mutant clones are started, instead of their developing times. Observe that Theorem 3.1 holds whether if F_μ satisfies ($\mathcal{H}$) or not. The Haldane model can be considered as an example: the final instants of the normal cell are exponentially distributed with rate λ and the lifetimes of the mutant cells are equal to a constant a. Then the cdf $F_\mu(s,\cdot)$ is defined for any $t \geqslant s$ by

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

and does not satisfy $(\mathcal{H})$. Computation of pgfs and probabilites for Haldane model have already been done in [23]. The pgf $\mathcal{I}$ is given by

$$\mathcal{I}(z) = \sum_{i \geqslant 0} b_i(z) \mathrm{e}^{-\lambda i a} \left(1 - \mathrm{e}^{-\lambda a}\right) . \tag{13}$$

For any $i \geqslant 0$, the pgf b_i represents the size of a clone stemming from a mutant born at time s in the interval $[s + ia\,; s + (i+1)a]$. In other words, $b_0(z) = z$, and for all $i > 0$:

$$b_i(z) = \delta + (1 - \delta) \left(b_{i-1}(z)\right)^2 . \tag{14}$$

The probabilities of the final mutant count will be recalled in next section. In particular, the results obtained by Ycart [38] are recovered when $\delta = 0$ and $a = \log(2)$.

Proposition 3.1 can be extended to the case $\gamma \geqslant 0$.

Proposition 3.2. *Let $\pi = (\pi_n)_{n\in\mathbb{N}}$ et $\tau = (\tau_n)_{n\in\mathbb{N}}$ two sequences, and α a positive real such that:*

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

where

$$h^*_\nu(s,t) = (1 - 2\gamma) h_\nu(s,t) ,$$

and

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

the probability that a clone stemming from a normal cell born at time 0 is not died out at time t. Then:

$(\mathcal{A}_1^{(\gamma)})$ *As n tends to infinity, the distribution of the total number of mutations $Z_n(\tau_n)$ tends to the Poisson distribution parameter:*

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

where

$$h^*_{\nu,\infty} = \lim_{t\to+\infty} h^*_\nu(0,t) , \quad \text{and} \quad \omega_\infty = \lim_{t\to+\infty} \omega(t) ;$$

$(\mathcal{A}_2^{(\gamma)})$ *As n tends to infinity, the joint distribution of the vector $\left(T_1^{(n)}, \ldots, T_k^{(n)}\right)$ of k mutation instants in an interval $[0\,; t]$ converges to that of the order statistics of a k sample of the distribution*

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

Proposition 3.2.

Assertion ($\mathcal{A}_1^{(\gamma)}$) Consider again the binary branching process with a single initial cell and without mutations: each cell can die out with probability γ instead of dividing into two cells. Denote by $N_1(t)$ the total number of cells at instant t. For any $t \geqslant 0$, consider the sequence $(N_n(t))_{n\in\mathbb{N}}$ defined by (3). For any $t > 0$, each copy of $N_1(t)$ may survive until t with probability $\omega(t)$. By the same reasoning as for assertion ($\mathcal{A}_1^{(0)}$), the number of divisions in the surviving clones occurring in $[0\,;t]$ is almost surely equivalent to $n\left(\omega(t)\mathrm{e}^{h_\nu^*(0,t)}-1\right)$. Moreover, the number of divisions in the proportion $1-\omega(t)$ of dying clones is bounded and can be neglected. Thus, the total number of divisions in the n copies is almost surely equivalent to $n\left(\omega(\tau_n)\mathrm{e}^{h_\nu^*(0,\tau_n)}-1\right)$. Let us mark independently and with probability π_n the cells divisions. Consider $(X_n(t))_{n\in\mathbb{N}}$ the sequence of marked divisions at time t in the n copies. By the same reasoning as for assertion ($\mathcal{A}_1^{(0)}$), the distribution of the number of mutations $Z_n(\tau_n)$ tends to the Poisson distribution with parameter $m=\alpha\left(1-\left(\omega_\infty \mathrm{e}^{h_{\nu,\infty}^*}\right)^{-1}\right)$.

Assertion ($\mathcal{A}_2^{(\gamma)}$) The number of mutation occasions in an interval $[0\,;t]$ is equivalent in probability to $n\left(\omega(t)\mathrm{e}^{h_\nu^*(0,t)}-1\right)$. Therefore:

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

where

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

Moreover:

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

For any $t \geqslant 0$ and as n tends to infinity, the distribution of $Z_n(t)$ is equivalent to the Poisson distribution with parameter:

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

Assertion ($\mathcal{A}_2^{(\gamma)}$) is then deduced by the same reasoning as for assertion ($\mathcal{A}_2^{(0)}$). □

Therefore, Theorem 3.1 can also be extended to the case where $\gamma \geqslant 0$.

Theorem 3.2. *Let $\pi = (\pi_n)_{n\in\mathbb{N}}$ and $\tau = (\tau_n)_{n\in\mathbb{N}}$ two sequences, and α a positive real such that:*

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

As n tends to infinity, the pgf of the number of mutants at time τ_n starting with n normal cells tends to the pgf (11) with:

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

and

$$\mathcal{I}(z) = \frac{1}{\omega_\infty - \mathrm{e}^{-h^*_{\nu,\infty}}} \lim_{t\to+\infty} \int_0^t \psi(z,u,t)\left(\omega'(u) + \lambda_\nu(u)(1-2\gamma)\omega(u)\right)\mathrm{e}^{-h^*_\nu(u,t)}\mathrm{d}u\,. \tag{15}$$

From now on, assume that F_μ satisfies $(\mathcal{H})$ and denote by h_μ its related function defined in $(\mathcal{H}_4)$. There exists a positive, continuous, $\mathbb{R}_+$-valued function λ_μ such that:

$$h_\mu(s,t) = \int_s^t \lambda_\mu(u)\mathrm{d}u\,.$$

From Proposition 3.2 of [23], the pgf of the size at time t of a clone stemming from a mutant cell born at time s is given by

$$\psi(z,s,t) = \frac{\delta(1-z) + \mathrm{e}^{-h^*_\mu(s,t)}((1-\delta)z-\delta)}{(1-\delta)(1-z) + \mathrm{e}^{-h^*_\mu(s,t)}((1-\delta)z-\delta)}\,, \tag{16}$$

where, for any $(s,t)\in\mathbb{R}^2_+$:

$$h^*_\mu(s,t) = (1-2\delta)h_\mu(s,t)\,.$$

Assume also there exists $\rho > 0$ such that for any $t \geqslant 0$:

$$(1-2\gamma)\lambda_\nu(t) = \rho(1-2\delta)\lambda_\mu(t)\,. \tag{17}$$

The constant ρ can be interpreted as the instantaneous ratio of hazard functions λ_ν and λ_μ. The assumption of proportional hazard functions is not new: in survival analysis, it is known as the Cox proportional-hazard regression model, which is widely used. Observe that, in the case where λ_μ is a positive constant, the constant ρ corresponds to the relative fitness in Luria-Delbrück model. This designation will be kept from now on. Under the asymptotic context of Thereom 3.2, the distribution of final mutant count depends on the mean number of mutation m, the fitness ρ, the death parameters γ and δ, and the limit of $h_\mu(0,t)$.

Theorem 3.3. *Let $\pi = (\pi_n)_{n\in\mathbb{N}}$ and $\tau = (\tau_n)_{n\in\mathbb{N}}$ two sequences, and α a positive real such that:*

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

Under (17) and as n tends to infinity, the pgf of the number of mutants at time τ_n starting with n normal cells tends to the pgf (11) with:

$$m = \alpha\left(1 - \left(\omega_\infty \mathrm{e}^{\rho h^*_{\mu,\infty}}\right)^{-1}\right)\,,$$

and

$$\mathcal{I}(z) = \frac{(1-2\gamma)\rho}{\omega_\infty - \mathrm{e}^{-\rho h^*_{\mu,\infty}}} \int_{\mathrm{e}^{-h^*_{\mu,\infty}}}^1 \frac{\delta(1-z) + v((1-\delta)z-\delta)}{(1-\delta)(1-z)+v((1-\delta)z-\delta)} Q(v) v^{\rho-1}\mathrm{d}v\,, \tag{18}$$

where

$$h_{\mu,\infty} = \lim_{t\to+\infty} h_\mu(0,t)\,, \quad h^*_{\mu,\infty} = (1-2\delta)h_{\mu,\infty}\,,$$

and

$$Q(v) = \frac{1-\gamma-2\gamma v\mathrm{e}^{-\rho h^*_{\mu,\infty}}}{\left(1-\gamma-\gamma v\mathrm{e}^{-\rho h^*_{\mu,\infty}}\right)^2}\,. \tag{19}$$

In particular, when $\gamma = 0$, (18) becomes:

$$\mathcal{I}(z) = \frac{1}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} \int_{\mathrm{e}^{-h^*_{\mu,\infty}}}^1 \frac{\delta(1-z) + v((1-\delta)z-\delta)}{(1-\delta)(1-z)+v((1-\delta)z-\delta)} v^{\rho-1}\mathrm{d}v\,, \tag{20}$$

which corresponds to Theorem 5.1 of [23].

Theorem 3.3. Plugging (16) in (15) and changing $\mathrm{e}^{-h_\mu(s,t)}$ into v leads to

$$\mathcal{I}(z) = \frac{(1-2\gamma)\rho}{\omega_\infty - \mathrm{e}^{-\rho h^*_{\mu,\infty}}} \lim_{t\to+\infty} \int_{\mathrm{e}^{-h^*_\mu(0,t)}}^1 \frac{\delta(1-z) + v(1-\delta)z-\delta)}{(1-\delta)(1-z)+v(1-\delta)z-\delta)} \frac{\left(1-\gamma-2\gamma v\mathrm{e}^{-\rho h^*_\mu(0,t)}\right)}{\left(1-\gamma-\gamma v\mathrm{e}^{-\rho h^*_\mu(0,t)}\right)^2} v^{\rho-1}\mathrm{d}v\,.$$

Since for any $v \in [0\,;1]$ and $t \geqslant 0$:

$$\frac{\left(1-\gamma-2\gamma v\mathrm{e}^{-\rho h^*_\mu(0,t)}\right)}{\left(1-\gamma-\gamma v\mathrm{e}^{-\rho h^*_\mu(0,t)}\right)^2} \leqslant \frac{1}{1-\gamma}\,,$$

applying the dominated convergence theorem leads to the result. □

A first example consists in considering a non-negative and increasing function f on $\mathbb{R}_+$, with finite limit f_∞ as t tends to infinity. Let h_μ be defined for (s,t) in $\mathbb{R}_+^2$ by

$$h_\mu(s,t) = \log\left(\frac{f(t)}{f(s)}\right) . \tag{21}$$

Then the expected size of a mutant clone started at time s is $\left(\frac{f(t)}{f(s)}\right)^{1-2\delta}$. In other words, it is possible to fit the average trajectory of the development of the clones to any appropriate function of time defining h_μ as (21). Moreover, only the ratio of f_∞ over $f(0)$ has an influence on $\mathcal{I}(z)$.

As another particular case, if $h_{\mu,\infty}$ is infinite, the Luria-Delbrück with cells deaths distribution [39] is recovered.

Corollary 3.1. *Let $\pi = (\pi_n)_{n\in\mathbb{N}}$ and $\tau = (\tau_n)_{n\in\mathbb{N}}$ two sequences, and α a positive real such that:*

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

Assume $h_{\mu,\infty} = +\infty$. Under (17) and as n tends to infinity, the pgf of the number of mutants at time τ_n starting with n normal cells tends to the pgf (11) with $m = \alpha$ and:

$$\mathcal{I}(z) = \rho \int_0^1 \frac{\delta(1-z) + v((1-\delta)z - \delta)}{(1-\delta)(1-z) + v((1-\delta)z-\delta)} v^{\rho-1}\mathrm{d}v . \tag{22}$$

In other words, the Luria-Delbrück with cells deaths distribution can be extended to the case where $F_\nu(s,\cdot)$ and $F_\mu(s,\cdot)$ are non-exponential distributions, as long as they satisfy $(\mathcal{H})$, are cdfs of true measures on $\mathbb{R}_+$, and such that the associated functions λ_ν and λ_μ are proportional.

From now on, the different mutation models will be referred to from the notations of Table 1.

4 Probability calculations

Computation and simulation algorithms for the distribution of the final mutant counts are described here. Consider first the case of IMM models. The pgf (15) can be written as

$$\mathcal{I}(z,t) = \sum_{k\geqslant 0} r_k(t) z^k ,$$

where r_k is defined for any $k \in \mathbb{N}$ and $t \in \mathbb{R}_+$ by

$$r_k(t) = \frac{1}{\omega(t) - \mathrm{e}^{-h_\nu^*(0,t)}} \int_0^t p_k(u,t)\left(\omega'(u) + \lambda_\nu(u)(1-2\gamma)\omega(u)\right)\mathrm{e}^{-h_\nu^*(u,t)}\mathrm{d}u .$$

$IMM(m, \gamma, \delta, F_\nu, F_\mu)$	Inhomogeneous mutation models where the final instant of a normal (resp. mutant) cell born at time s has cdf $F_\nu(s, \cdot)$ (resp. $F_\mu(s, \cdot)$) (Theorem 3.2)
$ILD(m, \rho, \gamma, \delta, h_{\mu,\infty})$	IMM models where F_μ satisfies $(\mathcal{H})$ under (17) (Theorem 3.3)
$LD(m, \rho, \delta)$	ILD models where $h_{\mu,\infty} = +\infty$ (Corollary 3.1)
$H(m, \rho, \delta)$	Haldane model

Table 1: Mutation models designations. The parameters m, ρ, γ, δ, and $h_{\mu,\infty}$ respectively, denote the mean number of mutation, the fitness parameter, the probability of dying for a normal cell, the probability of dying for a mutant cell, the asymptotic cumulative division rate of mutants. F_ν and F_μ, respectively, denote the cdf of the final instant for normal cells, the cdf of final instant for mutant.

and the $p_k(s,t)$'s are probabilities of the size at time t of a mutant clone started at time s. These probabilities mainly depend on the model assumptions. For example, if the cdf F_μ satisfies $(\mathcal{H})$, the $p_k(s,t)$'s are given by Proposition 3.3 of [23]

$$p_0(s,t) = \frac{\delta(1 - \mathrm{e}^{-h^*_\mu(s,t)})}{1 - \delta - \delta \mathrm{e}^{-h^*_\mu(s,t)}},$$

and for $k > 0$:

$$p_k(s,t) = (1 - p_0(s,t))P(s,t)(1 - P(s,t))^{k-1},$$

where:

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

Thus:

$$r_0(t) = \frac{1}{\omega(t) - \mathrm{e}^{-h^*_\nu(0,t)}} \int_0^t \frac{\delta\left(1 - \mathrm{e}^{-h^*_\mu(u,t)}\right)}{1 - \delta - \delta \mathrm{e}^{-h^*_\mu(u,t)}} \left(\omega'(u) + \lambda_\nu(u)(1-2\gamma)\omega(u)\right) \mathrm{e}^{-h^*_\nu(u,t)} \mathrm{d}u,$$

and for all $k > 0$:

$$r_k(t) = \frac{1}{\omega(t) - \mathrm{e}^{-h^*_\nu(0,t)}} \int_0^t \left(1 - \frac{\delta\left(1 - \mathrm{e}^{-h^*_\mu(u,t)}\right)}{1 - \delta - \delta \mathrm{e}^{-h^*_\mu(u,t)}}\right) P(u,t)(1 - P(u,t))^{k-1} \left(\omega'(u) + \lambda_\nu(u)(1-2\gamma)\omega(u)\right) \mathrm{e}^{-h^*_\nu(u,t)} \mathrm{d}u,$$

Then, (15) can be given by

$$\mathcal{I}(z) = \sum_{k \geqslant 0} r_k z^k,$$

where for any $k \in \mathbb{N}$, r_k is the limit of $r_k(t)$ as t tends to infinity. In particular, the r_k's for ILD models are given by

$$r_0 = \frac{(1-2\gamma)\rho}{\omega_\infty - \mathrm{e}^{-\rho h^*_{\mu,\infty}}} \int_{\mathrm{e}^{-h^*_{\mu,\infty}}}^1 \frac{\delta - \delta v}{1-\delta-\delta v} v^{\rho-1} Q(v)\mathrm{d}v\,,$$

and for all $k > 0$:

$$r_k = \frac{(1-2\gamma)\rho(1-2\delta)^2(1-\delta)^{k-1}}{\omega_\infty - \mathrm{e}^{-\rho h^*_{\mu,\infty}}} \int_{\mathrm{e}^{-h^*_{\mu,\infty}}}^1 \frac{(1-v)^{k-1}}{(1-\delta-\delta v)^{k+1}} Q(v) v^\rho \mathrm{d}v\,, \tag{23}$$

where Q is given by (19). Remark that for $\gamma = \delta = 0$, (23) leads to:

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

where $B_x(\theta,\beta)$ is the incomplete Beta function defined for any $x \in (0\,;1)$ by

$$B_x(\theta,\beta) = \int_0^x w^{\theta-1}(1-w)^{\beta-1}\mathrm{d}w\,,$$

and $B(\theta,\beta)$ is the complete Beta function, i.e. $B(\theta,\beta) = B_1(\theta,\beta)$.

As an other example, consider now H models. The pgf $\mathcal{I}$ is then given by (13). For any $i \in \mathbb{N}$, let us denote by $\left(p_k^{(i)}\right)_{k\in\mathbb{N}}$ the probabilities associated to the pgf (14). Then pgf $\mathcal{I}$ can be rewritting as

$$\begin{aligned} \mathcal{I}(z) &= \sum_{i\geqslant 0} \mathrm{e}^{-\lambda i a}\left(1-\mathrm{e}^{-\lambda a}\right) \sum_{k\geqslant 0} p_k^{(i)} z^k \\ &= \sum_{k\geqslant 0} z^k \sum_{i\geqslant 0} \mathrm{e}^{-\lambda i a}\left(1-\mathrm{e}^{-\lambda a}\right) p_k^{(i)}\,. \end{aligned}$$

The probabilities $(r_k)_{k\in\mathbb{N}}$ are then given by

$$r_k = \sum_{i\geqslant 0} \mathrm{e}^{-\lambda i a}\left(1-\mathrm{e}^{-\lambda a}\right) p_k^{(i)}\,.$$

Then the probabilities $(r_k)_{k\in\mathbb{N}}$ can be computed if the probabilities $\left(p_k^{(i)}\right)_{k\in\mathbb{N}}$ are known for all $i \in \mathbb{N}$. In practice, $\left(r_k^{(i)}\right)_{k\in\mathbb{N}}$ can be identified using Fast Fourier Transform.

The probabilities $(q_k)_{k\in\mathbb{N}}$ of the final mutant count can finally be explicited using the following algorithm [7] for Poisson compounds:

$$q_0 = \mathrm{e}^{-m(1-r_0)}\,, \tag{24}$$

and for any $k > 0$:

$$q_k = \frac{m}{k} \sum_{i=1}^{k} i r_i q_{k-i} \,. \tag{25}$$

The simulation algorithms are now introduced. Consider a inhomogeneous Poisson process $\{Y(t)\}_{t\geqslant 0}$ with a given expectation $\Lambda(t)$. Denote by $(S_i)_{i\in\mathbb{N}}$ the occurrence instants of the process $\{Y(t)\}_{t\geqslant 0}$. Conditionally to $Y(t) = k$, $\left(\frac{\Lambda(S_1)}{\Lambda(t)}, \ldots, \frac{\Lambda(S_k)}{\Lambda(t)}\right)$ has same distribution as the order statistics of a k sample of the uniform distribution on $(0\,;1)$ (Proposition A.2). Consider first IM models with $\gamma = 0$. As a direct consequence of assertion $(\mathcal{A}_2^{(0)})$, the following corollary holds:

Corollary 4.1. *Assume $\gamma = 0$. Conditionally on $Z_n(t) = k$ and as n tends to infinity, the vector $\left(\frac{\mathrm{e}^{h_\nu(0,T_1)}-1}{\mathrm{e}^{h_\nu(0,t)}-1}, \ldots, \frac{\mathrm{e}^{h_\nu(0,T_k)}-1}{\mathrm{e}^{h_\nu(0,t)}-1}\right)$ converges in distribution to the order statistics of a k sample of the uniform distribution on $(0\,;1)$..*

Then, a k sample of mutation instants $T_1, \ldots, T_k$ in an interval $[0\,;t]$ is drawn by:

1. sample k uniform variables $U_1, \ldots, U_k$;
2. deduce the order statistics $U_{(1)}, \ldots, U_{(k)}$;
3. apply for any $i = 1 \ldots k$:

$$T_i = h_{\nu,0}^{-1}\left[\log\left(U_{(i)}\left(\mathrm{e}^{h_\nu(0,t)} - 1\right) + 1\right)\right] \,,$$

where for any s, $h_{\nu,s}^{-1}$ is the function which satisfies:

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

From $(\mathcal{H}_4)$, this function is well defined. For example, if h_ν is defined as (21), then:

$$h_{\nu,s}^{-1}(u) = f^{-1}\left[\mathrm{e}^u f(s)\right] \,.$$

Remark that IMM models with $\gamma \geqslant 0$ could also be considered, as a consequence of assertion $(\mathcal{A}_2^{(\gamma)})$;

Corollary 4.2. *Conditionally on $Z_n(t) = k$, the vector $\left(\frac{\omega(T_1)\mathrm{e}^{h_\nu^*(0,T_1)}-1}{\omega(t)\mathrm{e}^{h_\nu^*(0,t)}-1}, \ldots, \frac{\omega(T_k)\mathrm{e}^{h_\nu^*(0,T_k)}-1}{\omega(t)\mathrm{e}^{h_\nu^*(0,t)}-1}\right)$ converges as n tends to infinity to the order statistics of a k sample of the uniform distribution on $(0\,;1)$.*

Note that the computation of the inverse function of

$$\begin{aligned} g(t) &= \omega(t)\mathrm{e}^{h_\nu^*(0,t)} - 1 \\ &= \frac{1-2\gamma}{(1-\gamma)\mathrm{e}^{-h_\nu^*(0,t)} - \gamma\mathrm{e}^{-2h_\nu^*(0,t)}} - 1 \,, \end{aligned}$$

has not been studied here. However, Newton-Pahson algorithm can be used to approximate it. Thus, as for the case where $\gamma = 0$, the simulation of a k sample of mutation instants is possible.

Consider now ILD models with $\gamma = 0$. According to Corollary 4.1, a k sample of mutation instants $T_1, \ldots, T_k$ in an interval $[0\,;t]$ is drawn by:

1. sample k uniform variables $U_1, \ldots, U_k$;
2. deduce the order statistics $U_{(1)}, \ldots, U_{(k)}$;
3. apply for any $i = 1 \ldots k$:

$$T_i = h_{\mu,0}^{-1}\left[\frac{1}{\rho(1-2\delta)}\log\left(U_{(i)}\left(\mathrm{e}^{\rho h_\mu^*(0,t)}-1\right)+1\right)\right], \tag{26}$$

where for any s, $h_{\mu,s}^{-1}$ is the function which satisfies:

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

Hence the following algorithm is used to compute sample of ILD distribution with $\gamma = 0$:

1. simulate a random number k of mutations, according to the Poisson distribution with parameter m;
2. compute the clone size stemming from each mutation:
   (a) simulate the mutation instant s as above (with $t = +\infty$);
   (b) compute $p_0(s, +\infty)$ and $P(s, +\infty)$;
   (c) make the random choice:
   - with probability $p_0(s, +\infty)$, output 0 (extinction of the clone);
   - with probability $1 - p_0(s, +\infty)$, output a geometric random number with parameter $h_\mu(s, +\infty)$.
3. sum the k clone sizes.

Remark that this algorithm can be used only if $h_{\mu,\infty}$ is finite. Otherwise, the mutations instants cannot been computed applying (26) on uniform variables. However, from Corollary 3.1, the LD distribution is recovered when $h_{\mu,\infty} = +\infty$. Thus the algorithm exposed by Ycart [39] can be used.

5 Parameter estimation

This section is dedicated to the extension of the p_0, ML, and GF methods mentioned in the introduction. Their construction will be quickly recalled. The ILD and H models with $\gamma = 0$ are considered here. The methods exposed in this section perform estimation only for m and ρ. Indeed, the fluctuations of the distribution of final mutant counts with respect to δ are very small [39]. In practice, only the magnitude of δ can be measured [30, 8]. The same is also true for the parameter $h_{\mu,\infty}$: it is related to the death parameter δ as follows:

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

where $M(s,t)$ is the size at time t of a mutant clone started at time s. In other words, estimating $h_{\mu,\infty}$ is possible only if both estimates of the death parameter and the expected final size of a mutant clone started at time 0 are available. Thus, the identifiability of the model is hard in practice. In that sense, δ and $h_{\mu,\infty}$ are assumed to be known, which is not realistic. In this section, $(X_1, \ldots, X_n)$ will denote a sample of n i.i.d. random variables following ILD or H models.

The p_0 method was the first method introduced by Luria and Delbrück [20] to estimate m for classic mutation models when $\delta = 0$. It uses the fact that the probability of null counts is e^{-m}. This relation remains true for ILD models, whether $h_{\mu,\infty}$ is finite or not. Hence m can be estimated by

$$\hat{m}_0 = -\log\left(\hat{q}_0\right) ,$$

where $\hat{q}_0$ is the relative frequency of zero among mutant counts. Thus, $\hat{m}_0$ is a consistent and asymptotically estimator of m. From the Δ-method (see for example [36, p. 79]), its asymptotic variance is given by

$$v_{\hat{m}_0} = \frac{1-q_0}{nq_0} .$$

An extension of the p_0 estimator to the case where $\delta > 0$ has been described by Ycart [39] for LD models (called Fixed Point estimator). It uses the fact that the probability of extinction of a supercriticial process is a fixed point of the pgf of the number of cells at time t, in a clone starting from one cell at time 0 [4, Lemma 1, p.141]. This probability is given by Theorem 1 of [4, Chap. I]:

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

Under ILD models, δ_* is a fixed point of the pgf (16) and (20). Therefore, a consistent and asymptotically normal estimator of m is then given by

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

where $\hat{\phi}_n$ denotes the empirical pgf of the sample:

$$\hat{\phi}_n(z) = \frac{1}{n}\sum_{i=1}^{n} z^{X_i}\,. \tag{27}$$

For any $z \in (0,1)$, $\hat{\phi}_n(z)$ is a consistent and asymptotically normal estimator of $\phi(z)$. From the Δ-method, the asymptotic variance of $\hat{m}_0$ is given by

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

Remark that the p_0 estimator of m depends only on δ. In other words, the p_0 method does not depend on the growth model, and can be used for any mutation model. However, this method does not diretly yield an estimator of ρ. If an estimate is desired, the Maximum Likelihood can be used for ρ only, setting $m = \hat{m}_0$.

As the probabilities q_k's and their derivatives with respect to m and ρ are explicit for ILD and H models, the Maximum Likelihood method seems to be an obvious choice to estimate m and ρ. Moreover, an equivalent of r_k for large values of k can be computed rewriting (23) as:

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

where the constant $\overline{\delta}$ is given by

$$\overline{\delta} = \frac{1-2\delta}{1-\delta}\,.$$

The following equivalent is then obtained:

$$r_k \underset{k\to+\infty}{\sim} \frac{\rho\overline{\delta}^2}{1-\mathrm{e}^{-h_{\nu,\infty}}}\left(\frac{\Gamma(\rho+1)}{k^{\rho+1}}\left(\frac{1-\delta}{1-2\delta}\right)^{\rho+1} - \int_0^{\mathrm{e}^{-h^*_{\mu,\infty}}} \frac{(1-v)^{k-1}}{(1-\delta_* v)^{k+1}} v^\rho \mathrm{d}v\right), \tag{28}$$

where Γ is the Gamma function. Remark that for $h_{\mu,\infty} = +\infty$, Formula (3.4) of [39] is recovered. Consider first ILD models. The derivatives of the r_k's with respect to ρ are now given:

$$\frac{\partial r_0}{\partial \rho} = \frac{\rho}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}}\int_{\mathrm{e}^{-h^*_{\mu,\infty}}}^1 \frac{\delta-\delta v}{1-\delta-\delta v} v^{\rho-1}\log(v)\mathrm{d}v + \left(\frac{1}{\rho} - \frac{h^*_{\mu,\infty}\mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}}\right) r_0\,,$$

and for any $k > 0$:

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

From (28), an asymptotic equivalent for large k is given by

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

where $\digamma$ is the Digamma function. Remark also that for $\delta = 0$:

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

The gradient of the q_k's can then be deduced from the r_k's and their derivatives:

$$\frac{\partial q_0}{\partial m} = -(1 - r_0) q_0 \quad \text{and} \quad \begin{aligned}\frac{\partial q_0}{\partial \rho} =& - \left[\frac{\partial m}{\partial \rho} (1 - r_0) - m \frac{\partial r_0}{\partial \rho} \right] q_0 \\ =& - m \left[\frac{h^*_{\mu,\infty} \mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1 - \mathrm{e}^{-\rho h^*_{\mu,\infty}}} (1 - r_0) - \frac{\partial r_0}{\partial \rho} \right] q_0\end{aligned}$$

The derivative of ϕ with respect to m is given by

$$\begin{aligned}\frac{\partial \phi}{\partial m}(z) =& - (1 - \mathcal{I}(z)) \phi(z) \\ =& - \left[\sum_{j \geqslant 0} q_j z^j - \left(\sum_{i \geqslant 0} r_i z^i \right) \left(\sum_{j \geqslant 0} q_j z^j \right) \right] \\ =& - \left[\sum_{j \geqslant 0} q_j z^j - \sum_{i,j \geqslant 0} r_i q_j z^{i+j} \right]\end{aligned}$$

On the other hand:

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

Hence for any $k > 0$:

$$\frac{\partial q_k}{\partial m} = -\left(q_k - \sum_{\substack{i,j\geqslant 0 \\ i+j=k}} r_i q_j \right) = \sum_{i=1}^{k} r_i q_{k-i} - q_k \,. \tag{29}$$

Similarly, the derivative of ϕ with respect to ρ is given by

$$\begin{aligned}
\frac{\partial \phi}{\partial \rho}(z) &= -\left[\frac{\partial m}{\partial \rho}(1-\mathcal{I}(z)) - m\frac{\partial h}{\partial \rho}(z) \right] \phi(z) \\
&= -\, m \left[\frac{h^*_{\mu,\infty} \mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}}(1-\mathcal{I}(z)) - \frac{\partial h}{\partial \rho}(z) \right] \phi(z) \\
&= -\, m \left[\frac{h^*_{\mu,\infty} \mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} \left(1 - \sum_{i\geqslant 0} z^i r_i \right) - \left(\sum_{i\geqslant 0} \frac{\partial r_i}{\partial \rho} z^i \right) \right] \left(\sum_{j\geqslant 0} q_j z^j \right) \\
&= -\, m \left[\frac{h^*_{\mu,\infty} \mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} - \sum_{i\geqslant 0} z^i \left(\frac{h^*_{\mu,\infty} \mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} r_i + \frac{\partial r_i}{\partial \rho} \right) \right] \left(\sum_{j\geqslant 0} q_j z^j \right)
\end{aligned}$$

Hence for any $k > 0$:

$$\frac{\partial q_k}{\partial \rho} = -m \left[\frac{h^*_{\mu,\infty} \mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} q_k - \sum_{i=1}^{k} q_{k-i} \left(\frac{h^*_{\mu,\infty} \mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} r_i + \frac{\partial r_i}{\partial \rho} \right) \right] . \tag{30}$$

Define the log-likelihood by

$$\begin{aligned}
\ell(X_1, \ldots, X_n) &= \sum_{i=1}^{n} \log(q_{X_i}) \\
&= \sum_{j=1}^{M} \left[\log(q_j) \sum_{i=1}^{n} \mathbb{1}_{X_i = j} \right] , \qquad (31)
\end{aligned}$$

where $M = \max_j X_j$ is the sample maximum. The couple $(\hat{m}_{ML}, \hat{\rho}_{ML})$ maximizing (31) is consistent and asymptotically normal [19, Theo. 5.1, Chap.6]. The asymptotic variances of $\hat{m}_{ML}$ and $\hat{\rho}_{ML}$ are given by

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

where $I = (I_{i,j})_{i,j\in\{1,2\}}$ is the following information matrix:

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

Using the algorithms (25), (29) and (30), the log-likelihood and its derivatives can be calculated iteratively. However, the formulas must be applied to vectors as large as M. Therefore, as for the homogeneous case [10], the procedure can be very long and numerically unstable. In practice, this instability problem can be avoided using Winsorization [37, sec. 2.2.]: any value of the sample that pass a certain bound is replaced by the bound itself. All information above the bound is lost. In extreme cases where the sample minimum is greater than the bound, i.e. for large m and/or small ρ, irrelevant results will be returned.

As mentioned previously, the optimization of the likelihood with respect to δ and $h_{\mu,\infty}$ is hard in practice. However, the derivatives of the r_k's and the q_k's with respect to δ and $h_{\mu,\infty}$ can also be iteratively computed. Consider first the derivatives with respect to δ:

$$\begin{aligned}\frac{\partial r_0}{\partial \delta} =& \frac{2\rho h_{\mu,\infty} \mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} \left(r_0 - \frac{\delta\left(1-\mathrm{e}^{h^*_{\mu,\infty}}\right)\rho \mathrm{e}^{(\rho-1)h^*_{\mu,\infty}}}{1-\delta-\delta \mathrm{e}^{h^*_{\mu,\infty}}} \right) \\ &+ \frac{1}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} \int_{\mathrm{e}^{-h^*_{\mu,\infty}}}^{1} \frac{(1-v)\rho v^{\rho-1}}{(1-\delta-\delta v)^2} \mathrm{d}v \,,\end{aligned}$$

and

$$\begin{aligned}\frac{\partial q_0}{\partial \delta} &= -\left[\frac{\partial m}{\partial \delta}(1-r_0) - m\frac{\partial r_0}{\partial \delta}\right] q_0 \\ &= -m\left[\frac{2\rho h_{\mu,\infty}\mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}}(1-r_0) - \frac{\partial r_0}{\partial \delta}\right] q_0 \,.\end{aligned}$$

For all $k > 0$:

$$\begin{aligned}\frac{\partial r_k}{\partial \delta} =& \frac{1}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} \Bigg\{ r_k \left(\frac{-2\delta(1-\delta) - (1-2\delta)(k-1)}{(1-2\delta)(1-\delta)} + 2\rho h^*_{\mu,\infty} \mathrm{e}^{-\rho h^*_{\mu,\infty}} \right) \\ &+ (1-2\delta)(1-\delta)^{k-1} \Bigg[- \frac{\left(1-\mathrm{e}^{-h^*_{\mu,\infty}}\right)^{k-1}}{1-\delta-\delta \mathrm{e}^{-h^*_{\mu,\infty}}} 2h^*_{\mu,\infty}\rho \mathrm{e}^{-\rho h^*_{\mu,\infty}} \\ &+ \int_{\mathrm{e}^{-h^*_{\mu,\infty}}}^{1} \frac{(1-v)^{k-1}(1+v)(k+1)}{(1-\delta-\delta v)^{k+1}} \rho v^{\rho} \mathrm{d}v \Bigg] \Bigg\} \,.\end{aligned}$$

Since the derivative of ϕ with respect to δ is given by

$$\begin{aligned}\frac{\partial \phi}{\partial \delta}(z) =& -\left[\frac{\partial m}{\partial \delta}(1-\mathcal{I}(z)) - m\frac{\partial h}{\partial \delta}(z)\right]\phi(z)\\ =& -m\left[\frac{2\rho h_{\mu,\infty}\mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}}(1-\mathcal{I}(z)) - \frac{\partial h}{\partial \delta}(z)\right]\phi(z)\\ =& -m\left[\frac{2\rho h_{\mu,\infty}\mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} - \sum_{i\geqslant 0} z^i\left(\frac{\partial r_i}{\partial \delta} + \frac{2\rho h_{\mu,\infty}\mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{\rho h^*_{\mu,\infty}}} r_i\right)\right]\left(\sum_{j\geqslant 0} q_j z^j\right).\end{aligned}$$

Hence, for any $k>0$:

$$\frac{\partial q_k}{\partial \delta} = -m\left[\frac{2\rho h_{\mu,\infty}\mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} q_k - \sum_{i=1}^{k} q_{k-i}\left(\frac{\partial r_i}{\partial \delta} - \frac{2h_{\mu,\infty}\mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} r_i\right)\right].$$

The derivatives of the r_k's and the q_k's with respect to $h^*_{\mu,\infty}$ can be computed as follows:

$$\frac{\partial r_0}{\partial h^*_{\mu,\infty}} = \frac{\rho \mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}}\left(-r_0 + \frac{\delta\left(1-\mathrm{e}^{-h^*_{\mu,\infty}}\right)}{1-\delta-\delta\mathrm{e}^{-h^*_{\mu,\infty}}}\right),$$

thus:

$$\begin{aligned}\frac{\partial q_0}{\partial h^*_{\mu,\infty}} =& -\left[\frac{\partial m}{\partial h^*_{\mu,\infty}}(1-r_0) - m\frac{\partial r_0}{\partial h^*_{\mu,\infty}}\right]q_0\\ =& m\left[\frac{\rho \mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}}(1-r_0) + \frac{\partial r_0}{\partial h^*_{\mu,\infty}}\right]q_0.\end{aligned}$$

And for any $k>0$:

$$\frac{\partial r_k}{\partial h^*_{\mu,\infty}} = \frac{\rho \mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}}\left(-r_k + \frac{\mathrm{e}^{-h^*_{\mu,\infty}}\left(1-\mathrm{e}^{-h^*_{\mu,\infty}}\right)^{k-1}}{\left(1-\delta-\delta\mathrm{e}^{-h^*_{\mu,\infty}}\right)^{k+1}}\right)$$

Since the derivative of ϕ with respect to $h^*_{\mu,\infty}$ is given by

$$\begin{aligned}\frac{\partial \phi}{\partial h^*_{\mu,\infty}}(z) =& -\left[\frac{\partial m}{\partial h^*_{\mu,\infty}}(1-\mathcal{I}(z)) - m\frac{\partial \mathcal{I}(z)}{\partial h^*_{\mu,\infty}}\right]\phi(z)\\ =& m\left[\frac{\rho \mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}}(1-\mathcal{I}(z)) + \frac{\partial \mathcal{I}(z)}{\partial h^*_{\mu,\infty}}\right]\phi(z)\\ =& m\left[\frac{\rho \mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} + \sum_{i\geqslant 0} z^i\left(\frac{\partial r_i}{\partial h^*_{\mu,\infty}} - \frac{\rho \mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} r_i\right)\right]\left(\sum_{j\geqslant 0} q_j z^j\right).\end{aligned}$$

Hence, for any $k > 0$:

$$\frac{\partial q_k}{\partial h^*_{\mu,\infty}} = m\left[\frac{\rho \mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} q_k + \sum_{i=1}^{k} q_{k-i}\left(\frac{\partial r_i}{\partial h^*_{\mu,\infty}} - \frac{\rho \mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} r_i\right)\right] .$$

Consider now h models. In that case, the derivatives of the r_k's with respect to ρ and δ are given by

$$\begin{aligned}\frac{\partial r_k}{\partial \rho} &= \sum_{i\geqslant 0} r_k^{(i)}\left(-ia\mathrm{e}^{-na\rho} + (i+1)a\mathrm{e}^{-(i+1)a\rho}\right) \\ &= \sum_{i\geqslant 0} r_k^{(n)} a\mathrm{e}^{-ia\rho}\left((i+1)\mathrm{e}^{-a\rho} - n\right) ,\end{aligned}$$

and

$$\begin{aligned}\frac{\partial r_k}{\partial \delta} &= \sum_{i\geqslant 0} \frac{\partial p_k^{(i)}}{\partial \delta}\left(\mathrm{e}^{-ia\rho} - \mathrm{e}^{-(i+1)a\rho}\right) + p_k^{(i)}\left(\frac{i\rho}{(1-\delta)}\mathrm{e}^{-ia\rho} - \frac{(i+1)\rho}{(1-\delta)}\mathrm{e}^{-(i+1)a\rho}\right) \\ &= \left(1-\mathrm{e}^{-a\rho}\right)\sum_{i\geqslant 0}\frac{\partial p_k^{(i)}}{\partial \delta}\mathrm{e}^{-ia\rho} + 2\rho\mathrm{e}^{-a}\sum_{i\geqslant 0} p_k^{(i)}\mathrm{e}^{-ia\rho}\left(i-(i+1)\mathrm{e}^{-a\rho}\right) .\end{aligned}$$

The derivatives $\left(\frac{\partial p_k^{(i)}}{\partial \delta}\right)_{k\in\mathbb{N}}$ can be computed in a similar way as for the probabilities $\left(p_k^{(i)}\right)_{k\in\mathbb{N}}$. For any $z \in (0\,;1)$:

$$\begin{aligned}\frac{\partial b_i}{\partial \delta}(z) &= \sum_{k\geqslant 0}\frac{\partial p_k^{(i)}}{\partial \delta} z^k \\ &= 1 - \left(b_{i-1}(z)\right)^2 + 2(1-\delta)\frac{\partial b_{i-1}}{\partial \delta}(z) b_{i-1}(z) ,\end{aligned}$$

and

$$\frac{\partial b_0}{\partial \delta}(z) = 0 .$$

The sequence $\left(\frac{\partial p_n^{(i)}}{\partial \delta}\right)_{k\geqslant 0}$ can then be deduced for any $i \in \mathbb{N}$ from the sequenceof polynoms $\left(\frac{\partial b_i}{\partial \delta}\right)_{i\geqslant 0}$ using Fast Fourier Transform. However, each polynom b_i has 2^i coefficients. The computation of these coefficients is very long. As mentioned earlier, the ML method has already several numerical issues. Therefore, using the ML method is not recommended under Haldane model with $\delta > 0$.

An alternative to the ML method relies on the generating function ϕ of final mutant count. Indeed, the parameter of a Poisson compound can be easily estimated [28, 22]. The Gf method proposed in [10, Section 4] uses this approach to estimate m and ρ under LD models when $\delta = 0$. This method can also be used for H models [38] or LD models with $\delta > 0$ [39]. It is extended here to the ILD models. The pgf $\mathcal{I}$ and its derivative with respect to ρ are required in this method. Thus they are implemented as the following numerically stable expressions :

$$\mathcal{I}(z) = \delta_* + \frac{z_*(1-\delta_*)}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} \int_{\mathrm{e}^{-h^*_{\mu,\infty}}}^1 \frac{\rho v^\rho}{1+z_* v} \mathrm{d}v \,,$$

and

$$\frac{\partial \mathcal{I}(z)}{\partial \rho} = \frac{z_*(1-\delta_*)}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} \left\{ \left[1 - \frac{\rho h^*_{\mu,\infty} \mathrm{e}^{-\rho h^*_{\mu,\infty}}}{1-\mathrm{e}^{-\rho h^*_{\mu,\infty}}} \right] \int_{\mathrm{e}^{-h^*_{\mu,\infty}}}^1 \frac{v^\rho}{1+z_* v} \mathrm{d}v \right.$$
$$\left. + \int_{\mathrm{e}^{-h^*_{\mu,\infty}}}^1 \frac{\rho v^\rho}{1+z_* v} \log(v) \mathrm{d}v \right\} \,,$$

where the constant δ_* and z_* are given by

$$\delta_* = \frac{\delta}{1-\delta} \,, \quad \text{and} \quad z_* = \frac{z-\delta_*}{1-z} \,.$$

Consider z_1, z_2, z_3 in $(0\,;1)$. The GF estimators of m and ρ are the following:

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

where $\mathcal{I}_x$ is the pgf (20) for ILD models or (13) for H models setting $\rho = x$, $\hat{\phi}_n$ is the empricial pgf (27) and:

$$g(x) = \frac{\mathcal{I}_x(z_1) - 1}{\mathcal{I}_x(z_2) - 1} \quad \text{and} \quad \hat{y}_n = \frac{\log\left(\hat{\phi}_n(z_1)\right)}{\log\left(\hat{\phi}_n(z_2)\right)} \,.$$

From Theorem 3.4 of [28] and Δ-method, it can be proved that the couple of estimators $(\hat{m}_{GF}, \hat{\rho}_{GF})$ is strongly consistent and asymptotically normal, with explicit asymptotic variance.

Proposition 5.1. *Let z_1, z_2, z_3 in $(0\,;1)$, two by two distinct. Consider the random vector*

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

and its asymptotic covariance matrix $C = (c(z_i, z_j))_{i,j=1,2,3}$ *given by*

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

Then, the couple of random variables:

$$\sqrt{n}\,((\hat{m}_{GF}, \hat{\rho}_{GF}) - (m, \rho)) \tag{32}$$

converges in distribution to the bivariate centered normal distribution with covariance matrix $A^t C A$*, where* $A = (a_{i,j})_{\substack{i=1,2,3\\ j=1,2}}$ *is a* 3×2 *matrix with:*

$$a_{1,1} = \frac{m a_{1,2}}{\mathcal{I}(z_3) - 1} \frac{\partial \mathcal{I}(z_3)}{\partial \rho} \quad ; \quad a_{1,2} = \frac{\mathcal{I}(z_2) - 1}{m\phi(z_1)\left(\frac{\partial \mathcal{I}(z_1)}{\partial \rho}(\mathcal{I}(z_2) - 1) - \frac{\partial \mathcal{I}(z_2)}{\partial \rho}(\mathcal{I}(z_1) - 1)\right)}$$

$$a_{2,1} = \frac{m a_{2,2}}{\mathcal{I}(z_3) - 1} \frac{\partial \mathcal{I}(z_3)}{\partial \rho} \quad ; \quad a_{2,2} = \frac{\mathcal{I}(z_1) - 1}{m\phi(z_2)\left(\frac{\partial \mathcal{I}(z_2)}{\partial \rho}(\mathcal{I}(z_1) - 1) - \frac{\partial \mathcal{I}(z_1)}{\partial \rho}(\mathcal{I}(z_2) - 1)\right)}$$

$$a_{3,1} = \frac{1}{\phi(z_3)(\mathcal{I}(z_3) - 1)} \quad ; \quad a_{3,2} = 0$$

The proof of 5.1 has already been exposed by Hamon and Ycart [10]. By definition, the GF estimators depend on the arbitrary values of z_1, z_2 and z_3. They can be seen as tuning parameters and should be appropriately chosen according to the sample to minimize the asymptotic variances of Proposition 5.1. Since the unknown values of m and ρ influence also the variance, optimal values for z_1, z_2 and z_3 cannot be easily identified. However, the fluctuations of the variances with respect to these three parameters are quite small. Their values have been set in [10, p. 1262] according to simulation experiments.

The fact is that the GF estimators are in practice comparable in precision to ML estimators, with a much broader range of calculability, a better numerical stability, and a negligible computing time. For that reason, ML optimization can be initialized by GF estimates, to improve both numerical stability and computing time.

Remark that the estimation of ρ requires to identify the zero of the monotone function $g(\rho) - \hat{y}_n$. In practice, the research domain is bounded. This can be a problem if the sample does not contain jackpot, i.e. for large theoretical values of ρ. However in that case, it should be considered that a mutation model is not adapted to the data.

6 Simulation study

If the model used for the estimation does not correspond to the theoretical model, the estimates can be biased. Simulation experiments have been performed to observe the bias induced by estimating under LD or H models when data are realizations of ILD model. The death parameters γ and δ are assumed to be zero. These simulation studies have

been implemented in R [26], using the R package `flan` [24], which is available on CRAN (https://cran.r-project.org/package=flan).

As exposed in [24], the GF method is at least equivalent to the p_0 and ML methods in terms of mean squared error. Since it is also the most performant in terms of computational time, the simulation experiments have been performed using this estimation method.

Assume that h_μ is defined as (21), where f is solution of the logistic equation:

$$f'(t) = f(t)\left(1 - \frac{f(t)}{f_\infty}\right),$$

where f_∞ denotes the finite limit of f as t tends to infinity, i.e. the carrying capacity. Thus:

$$f(t) = \frac{f_\infty}{1 + \left(\frac{f_\infty}{f(0)} - 1\right)\mathrm{e}^{-t}}.$$

The associated cdf F_μ is then given for any $(s,t) \in \mathbb{R}_+ \times \overline{\mathbb{R}}_+$ by

$$F_\mu(s,t) = \left(1 - \frac{f(s)}{f(t)}\right)\mathbb{1}_{t\in[s\,;+\infty)} + \mathbb{1}_{t=+\infty}, \tag{33}$$

and the asymptotic cumulative division rate $h_{\mu,\infty}$ by

$$h_{\mu,\infty} = \log\left(\frac{f_\infty}{f(0)}\right).$$

Recall that the choice of the function f is a matter only for the simulation: mutant count distribution depends only on the ratio $h_{\mu,\infty}$ (Theorem 3.3). In that sense, $f(0)$ is set here to 1.

Simulation experiments have been made along the following lines for each of the 24 sets of parameters $m = (0.5, 2, 4, 8)$, $\rho = (0.8, 1, 1.2)$ and $h_{\mu,\infty} = (\log(100), \log(10^4))$:

1. draw 10000 samples of size 100, under inhomogeneous model with cdf (33);
2. for each sample, compute GF estimates of m and ρ under ILD, LD and H models;
3. for each model, observe the empirical distribution of relative bias $\hat{\theta}/\theta$, where $\hat{\theta}$ is the estimator of the true value θ.

Boxplots of Figures 1 and 2 show the empirical distributions of the estimators $\hat{m}_{GF}$ and $\hat{\rho}_{GF}$ obtained under LD model (left boxplots), H model (center boxplots), and ILD model(right boxplots). Red lines mark theoretical values, blue lines mark relative biases of 0.9 and 1.1.

According to visual observations, LD and H models correctly estimate small values of m. Positive biases which increase as ρ decreases are observed for $m = 4$ and $m = 8$.

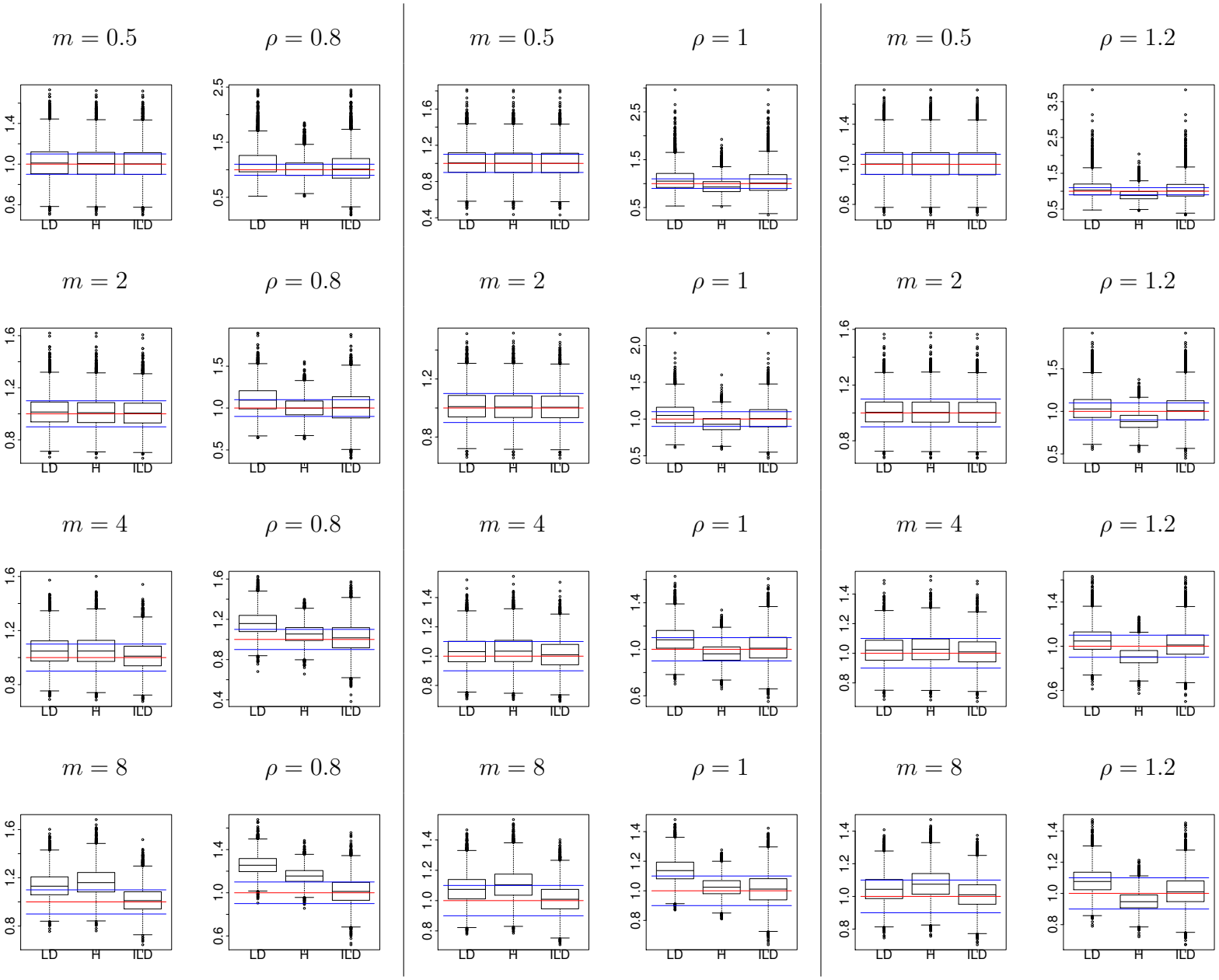

Figure 1: **GF estimates under *LD*, *H*, and *ILD* models on data drawn with *ILD* model ($h_{\mu,\infty} = \log(100)$).** For each of the 12 sets of parameters $m = (0.5, 2, 4, 8)$ (rows) and $\rho = (0.8, 1, 1.2)$ (columns), 10^4 samples of size 100 of the $ILD(m, \rho, 0, 0, h_{\mu,\infty})$ distribution were simulated with cdf (33) and $h_{\mu,\infty} = \log(100)$. For each column, the first three boxplots represent the distribution of the 10^4 ratio $\hat{m}_{GF}/m$ obtained under LD model (left), H model (center), and ILD model (right) ; the last three boxplots represent the distribution of the 10^4 ratio $\hat{\rho}_{GF}/\rho$ obtained under LD model (left), H model (center), and ILD model (right). Red horizontal lines mark theoretical value. Blue horizontal lines mark relative biases of 0.9 and 1.1.

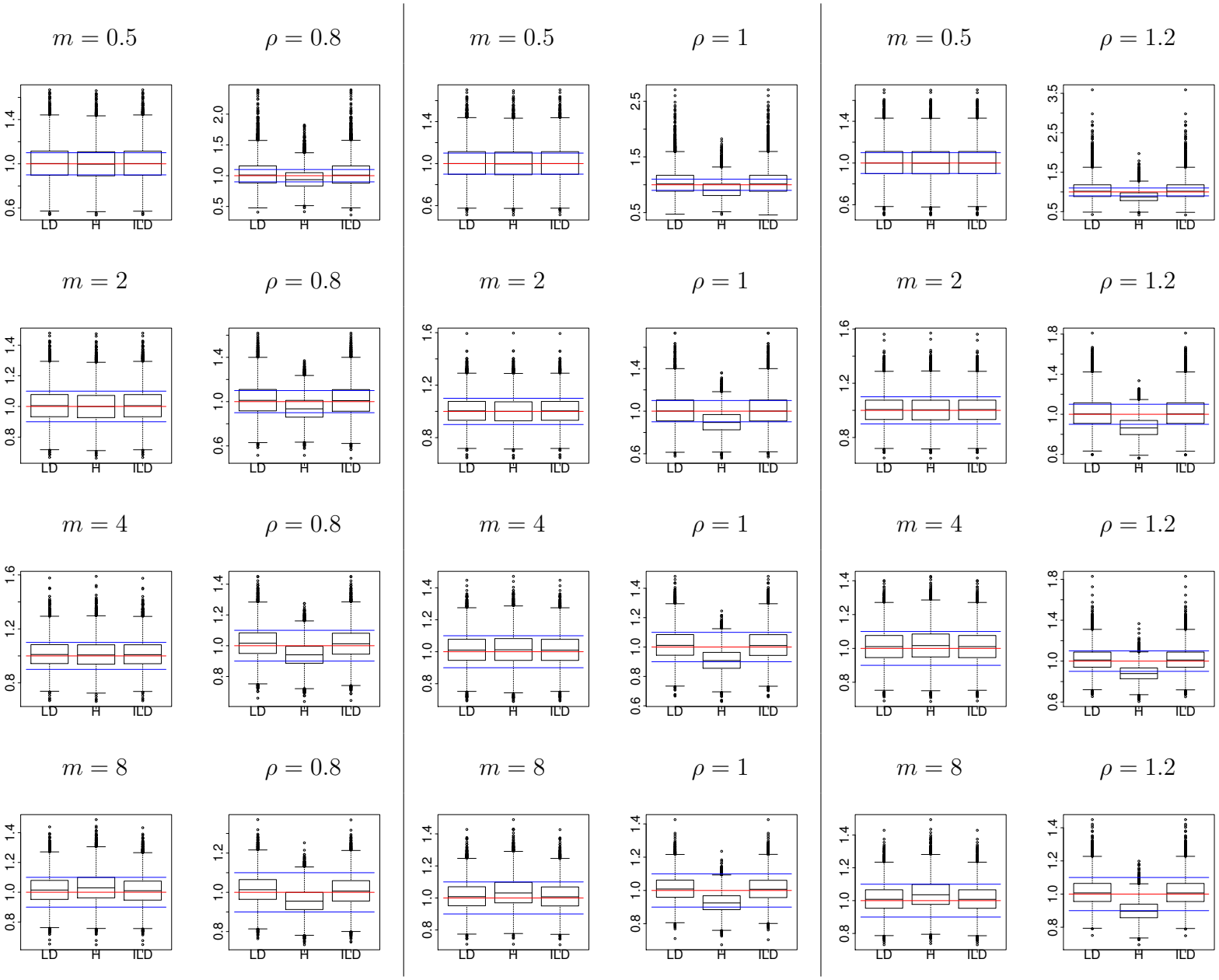

Figure 2: **GF estimates under *LD*, *H*, and *ILD* models on data drawn with *ILD* model ($h_{\mu,\infty} = \log(10^4)$).** For each of the 12 sets of parameters $m = (0.5, 2, 4, 8)$ (rows) and $\rho = (0.8, 1, 1.2)$ (columns), 10^4 samples of size 100 of the $ILD(m, \rho, 0, 0, h_{\mu,\infty})$ distribution were simulated with cdf (33) and $h_{\mu,\infty} = \log(10^4)$. For each column, the first three boxplots represent the distribution of the 10^4 ratio $\hat{m}_{GF}/m$ obtained under LD model (left), H model (center), and ILD model (right) ; the last three boxplots represent the distribution of the 10^4 ratio $\hat{\rho}_{GF}/\rho$ obtained under LD model (left), H model (center), and ILD model (right). Red horizontal lines mark theoretical values. Blue horizontal lines mark relative biases of 0.9 and 1.1.

Moreover, LD model overestimates the fitness ρ. This bias seems to increase as m increases and/or ρ decreases. For the extreme case where $m = 8$ and $\rho = 0.8$, a wide proportion of the estimates have a relative bias larger than 20%. The behavior of ρ estimates under H model, is harder to interpret. For larger value of $h_{\mu,\infty}$, there is no distinction between LD and ILD models. In particular, the H model seems to underestimate ρ.

Two facts have to be noticed. First, even if the whole population of cells has a logistic growth, it is possible in practice to stop the experiment before the inflexion instant. Thus, the assumption of an exponential growth could be considered and the LD distribution could be used to perform estimations. Moreover, the initial number of cells is in practice of order 10^3–10^4, the final number of cells of order 10^8–10^9. According to Figure 2, the value of $h_{\mu,\infty}$ is then such that the bias induced by considering an exponential growth instead of logistical growth should be negligible.

7 Conclusion

An extension for the classic mutation models to the case where the final instant of a cell depends on its birth date has been proposed. Results are based on the decomposition as three ingredients of any mutation model. This approach led to a family of distributions for the asymptotic mutant count. These distributions depend on the expected number of mutations m, the death probabilities γ and δ for normal and mutant cells, and the final instant cdf $F_\nu(s,\cdot)$ and $F_\mu(s,\cdot)$ for normal and mutant cells born at a given time s. The previous results obtained with an analytic approach are recovered and generalized to the case where $\gamma > 0$. Computation of probabilities and simulation algorithms have been described. The Luria-Delbrück distribution with cell deaths and the Haldane model are recovered. The latter has also been extended to the case where $\delta > 0$. The consequence for statistical inference has been treated: robust estimation methods have been extended to the ILD models; biases induced by considering classic mutation models instead of birth-date dependence model have been studied with simulation experiments. Considering the order of $h_{\mu,\infty}$ in practice, the biases on m and ρ induced in practice by considering the LD distribution instead of the ILD distribution seems negligible. The R package `flan` which has been used for simulation study is available on CRAN (https://cran.r-project.org/package=flan). Note that `flan` does not take into account the death of normal cells (i.e. $\gamma > 0$) for now. The extension to the case $\gamma > 0$ is being developed and should appear on CRAN. In the same time, a web-tool based on `flan` and R package `shiny` is also being developed. It allows to use some features of `flan` (in particular hypothesis testing and simulation) without any installation or knowledge in R. A first version is already available at https://toltex-shiny.u-ga.fr/RodaShiny/ShinyFlan/.

A Point processes

The main properties of point processes used in this paper are shortly exposed in this appendix. Consider a sequence $(T_i)_{i\in\mathbb{N}}$. Each T_i represents the occurring time of a given event, such as a mutation. Let us denote by $\{N(t)\}_{t\geqslant 0}$ the associated point process, defined for any $t \in \mathbb{R}_+$ by

$$N(t) = \max\{i \in \mathbb{N}\,;\, T_i \leqslant t\}\ ,$$

which represents the cumulated number of events occurring in $[0\,;t]$. Assume that the following holds:

1. $T_0 = 0$ and $N(0) = 0$ with probability 1;
2. The paths $\{N(t)\}_{t\geqslant 0}$ are *cadlag* and increasing ;
3. the counting process $\{N(t)\}_{t\geqslant 0}$ is simple, i.e. for any $t \in \mathbb{R}_+$:

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

Consider now the two following functions:

1. The intensity of the process $\{N(t)\}_{t\geqslant 0}$ defined by

$$\begin{aligned}\xi(t, N(t)) &\equiv \xi(t, N(t), T_1, \ldots, T_{N(t)})\\ &= \lim_{\Delta t\to 0} \frac{1}{\Delta t}\mathbb{P}\left[N(t+\Delta t) - N(t) = 1 \,|\, N(t), T_1, \ldots, T_{N(t)}\right]\ .\end{aligned} \tag{34}$$

   This intensity represents the probability of occurring after a time t, conditionally to the number of events before t and their respective appearance time. Thus ξ is a random variable depending on $N(t), T_1, \ldots, T_{N(t)}$. To simplify the redaction, it will be sometimes denoted here by $\xi(t, N(t))$.

2. The conditional intensity defined by

$$\tilde{\xi}(t, n) = \mathbb{E}[\xi(t, N(t)) \,|\, N(t) = n] \tag{35}$$

   The function $\tilde{\xi}(t, n)$ is deterministic. It represents the probability of occurring after a time t, knowing that n events have occurred before t.

The distribution of the event number occurring before a given time can then be explicited.

Proposition A.1. *For any $t \in \mathbb{R}_+$, the distribution of the event number occurring before a given time t is given by*

$$\mathbb{P}\left[N(t) = 0\right] = \exp\left(-\int_0^t \tilde{\xi}(u, 0)\mathrm{d}u\right)\ , \tag{36}$$

and for any $n > 0$,

$$\mathbb{P}\left[N(t) = n\right] = \int_{0<t_1<\cdots<t_n} \left[\prod_{i=1}^{n} \tilde{\xi}(t_i, i-1)\right] \exp\left(-\sum_{i=0}^{n}\left(\int_{t_i}^{t_{i+1}} \tilde{\xi}(u, i)\mathrm{d}u\right)\right) \mathrm{d}t_1 \ldots \mathrm{d}t_n\,, \tag{37}$$

with convention $t_0 = 0$ et $t_{k+1} = t$.

Consider now the distribution of the occurring instants. For any vector $\boldsymbol{t^{(n)}} = (t_1, \ldots, t_n) \in \mathbb{R}_+^n$, the event $(T_1 = t_1, \ldots, T_n = t_n)$ will be denoted by $\boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}$.

Proposition A.2. *The probability distribution function (pdf) of the instant T_1 of the first event is given by*

$$f_{T_1}(t) = \xi(t, 0) \exp\left(- \int_0^t \xi(u, 0) \mathrm{d}u \right) . \tag{38}$$

Conditionally to $\boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}$, the pdf of the instant T_{n+1} of the $n+1$-th event is given by

$$f_{\left(T_{n+1} | \boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}\right)}(t) = \xi(t, n, \boldsymbol{t^{(n)}}) \exp\left(- \int_{t_n}^t \xi(u, n, \boldsymbol{t^{(n)}}) \mathrm{d}u \right) \mathbb{1}_{0 < t_1 < \cdots < t_n} . \tag{39}$$

The joint pdf of the vector $\boldsymbol{T^{(n)}}$ is given by

$$f_{\boldsymbol{T^{(n)}}}(\boldsymbol{t^{(n)}}) = \left[\prod_{i=1}^n \xi(t_i, i-1, \boldsymbol{t^{(i-1)}}) \right] \exp\left(- \sum_{i=1}^n \int_{t_{i-1}}^{t_i} \xi(u, i-1, \boldsymbol{t^{(i-1)}}) \mathrm{d}u \right) \mathbb{1}_{0 < t_1 < \cdots < t_n} . \tag{40}$$

Assume now that the process $\{N(t)\}_{t \geqslant 0}$ is a time inhomogeneous Poisson process. In that case, the intensity (34) is a deterministic time function:

$$\xi(t, N(t), T_1, \ldots, T_{N(t)}) = \xi(t) .$$

Note that the increments are independent. The conditional intensity (35) is given by

$$\tilde{\xi}(t, N(t)) = \mathbb{E}\left[\xi(t) \,|\, N(t) \right] = \xi(t) .$$

According to Proposition A.1, the distribution of the event number occurring before a given time can be explicited. The direct consequence of Proposition A.2 provides the distribution of the occurring instant for a given number of events.

Proposition A.3. *For any $t \in \mathbb{R}_+$, $N(t)$ follows the Poisson distribution with expectation*

$$m(t) = \int_0^t \xi(u) \mathrm{d}u .$$

The pdf of the instant T_1 of the first event is given by

$$f_{T_1}(t) = \xi(t) \mathrm{e}^{-m(t)} . \tag{41}$$

Conditionally to $\boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}$, the pgf of the $(n+1)$-th instant T_{n+1} is given by

$$f_{\left(T_{n+1} | \boldsymbol{T^{(n)}} = \boldsymbol{t^{(n)}}\right)}(t) = \xi(t) \mathrm{e}^{m(t_n) - m(t)} \mathbb{1}_{0 < t_1 < \cdots < t_n} . \tag{42}$$

The joint pdf of the vector $\boldsymbol{T^{(n)}}$ is given by

$$f_{\boldsymbol{T^{(n)}}}(\boldsymbol{t^{(n)}}) = \left[\prod_{i=1}^n \xi(t_i) \right] \mathrm{e}^{-m(t_n)} \mathbb{1}_{t_1 < \cdots < t_n} . \tag{43}$$

The inhomogeneous Poisson processes satisfy a wide panel of usefull properties. One of them concerns the joint distribution of $\boldsymbol{T^{(n)}}$, conditionally to $N(t) = n$.

Corollary A.1. *Conditionally to $N(t) = n$, the joint pdf of the vector $\boldsymbol{T^{(n)}}$ is the same as the order statistics of a n sample of the distribution $\dfrac{\xi(u)}{m(t)}\mathbb{1}_{u\in[0\,;\,t]}$, i.e.*

$$f_{\left(\boldsymbol{T^{(n)}}\,|\,N(t)=n\right)}(\boldsymbol{t^{(n)}}) = n!\left[\prod_{i=1}^{n}\frac{\xi(t_i)}{m(t)}\mathbb{1}_{t_i\in[0\,;\,t]}\right]\mathbb{1}_{0<t_1<\cdots<t_n}\,.$$

Corollary A.1 can also be written as Corollary A.2.

Corollary A.2. *Conditionally to $N(t) = n$, the joint pdf of the vector $\left(\dfrac{m(T_1)}{m(t)},\ldots,\dfrac{m(T_n)}{m(t)}\right)$ is the same as the order statistics of a n sample of the uniform distribution on $(0\,;1)$.*

References

[1] L.J.S. Allen. *Stochastic processes with applications to Biology*. 2nd. Chapman and Hall/CRC, 2010.

[2] W.P. Angerer. "An explicit representation of the Luria-Delbrück distribution". In: *J. Math. Biol.* 42.2 (2001), pp. 145–174.

[3] P. Armitage. "The statistical theory of bacterial populations subject to mutation". In: *J. R. Statist. Soc. B* 14 (1952), pp. 1–40.

[4] K.B. Athreya and P.E. Ney. *Branching Processes*. Berlin Heidelberg: Springer, 1972.

[5] M.S. Bartlett. *An Introduction to Stochastic Processes, with Special Reference to Methods and applications*. 3rd. Cambridge University Press, 1978.

[6] A. Dewanji, E.G. Luebeck, and S.H. Moolgavkar. "A generalized Luria-Delbrück model". In: *Math. Biosci.* 197.2 (2005), pp. 140–152.

[7] P. Embrechts and J. Hawkes. "A limit theorem for tails of discrete infinitely divisible laws with applications to fluctuation theory". In: *J. Austral. Math. Soc. Series A* 32 (1982), pp. 412–422.

[8] F. Fontaine, E.J. Stewart, A.B. Lindner, and F. Taddei. "Mutations in two global regulators lower individual mortality in *Escherichia Coli*". In: *Mol. Microbio.* 67.1 (2008), pp. 2–14.

[9] P.L. Foster. "Methods for Determining Spontaneous Mutation Rates". In: *Method. Enzymol.* 409 (2006), pp. 195–213.

[10] A. Hamon and B. Ycart. "Statistics for the Luria-Delbrück distribution". In: *Elect. J. Statist.* 6 (2012), pp. 1251–1272.

[11] B. Houchmandzadeh. "General formulation of Luria-Delbrück distribution of the number of mutants". In: *Physical Review E : Statistical, Nonlinear, and Soft Matter Physics* 92 (2015), p. 012719.

[12] P. Jagers. "Stabilities and instabilities in population dynamics". In: *J. Appl. Probab.* 29 (1992), pp. 770–780.

[13] M.E. Jones, S.M. Thomas, and A. Rogers. "Luria-Delbrück Fluctuation Experiments: Design and Analysis". In: *Genetics* 136 (1994), pp. 1209–1216.

[14] N.L. Komarova, L. Wu, and P. Baldi. "The fixed-size Luria-Delbrück model with a nonzero death rate". In: *Math. Biosci.* 210.1 (2007), pp. 253–290.

[15] T. Kuzcek. "Almost sure limit results for the supercritical Bellman-Harris process". In: *J. Appl. Probab.* 19.3 (1982), pp. 668–674.

[16] A.K. Laird. "Dynamics of tumor growth". In: *Brit. J. Cancer* 18 (1964), pp. 490–502.

[17] A. Lambert. "The branching process with logistic growth". In: . *Appl. Probab.* 15.2 (2005), pp. 1506–1535.

[18] D.E. Lea and C.A. Coulson. "The distribution of the number of mutants in bacterial populations". In: *J. Genet.* 49.3 (1949), pp. 264–285.

[19] E.L. Lehmann and G. Casella. *Theory of Point Estimation*. 2nd. Springer Texts in Statistics. New York: Springer, 2003.

[20] S.E. Luria and M. Delbrück. "Mutations of bacteria from virus sensitivity to virus resistance". In: *Genetics* 28.6 (1943), pp. 491–511.

[21] W.T. Ma, G.v.H. Sandri, and S. Sarkar. "Analysis of the Luria-Delbrück distribution using discrete convolution powers". In: *J. Appl. Probab.* 29.2 (1992), pp. 255–267.

[22] M. Marcheselli, A. Baccini, and L. Barabesi. "Parameter estimation for the discrete stable family". In: *Commun. Statist. Theory Methods* 37.6-7 (2008), pp. 815–830.

[23] A. Mazoyer. "Time inhomogeneous mutation models with birth-date dependence". In: *Bull. Math. Biol.* (2017). DOI: 10.1007/s11538-017-0357-3.

[24] A. Mazoyer, R. Drouilhet, S. Despréaux, and B. Ycart. "flan: An R package for inference on mutation models". In: *R J.* 9.1 (2017).

[25] H.T. Nguyen. *An Introduction to Random Sets*. Boca Raton: Chapman & Hall/CRC, 2006.

[26] R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing, 2008.

[27] W.A. Rosche and P.L. Foster. "Determining mutation rates in bacterial populations". In: *Methods* 20.1 (2000), pp. 1–17.

[28] B. Rémillard and R. Theodorescu. "Inference based on the empirical probability generating function for mixtures of Poisson distributions". In: *Statist. Decisions* 18 (2000), pp. 349–366.

[29] S. Sarkar. "Haldane's solution of the Luria-Delbrück distribution". In: *Genetics* 127 (1991), pp. 257–261.

[30] E.J. Stewart, R. Madden, G. Paul, and F. Taddei. "Aging and death in an organism that reproduces by morphologically symmetric division". In: *PLoS Biology* 3.2 (2005), pp. 295–300.

[31] F.M. Stewart. "Fluctuation Tests: How Reliable Are the Estimates of Mutation Rates?" In: *Genetics* 137.4 (1994), pp. 1139–1146.

[32] F.M. Stewart, D.M. Gordon, and B.R. Levin. "Fluctuation analysis: the probability distribution of the number of mutants under different conditions". In: *Genetics* 124.1 (1990), pp. 175–185.

[33] W.Y. Tan. "A stochastic Gompertz birth-death process". In: *Statist. Probab. Lett.* 4.4 (1986), pp. 25–28.

[34] W.Y. Tan and S. Piantadosi. "On stochastic growth processes with application to stochastic logistic growth". In: *Statist. Sinica* 1 (1991), pp. 527–540.

[35] P.F. Verhulst. "Notice sur la loi que la population suit dans son accroissement". In: *Correspondance mathématique et physique*. Ed. by J.G. Garnier and A. Quetelet. Vol. 10. Société Belge de Librairie, Bruxelles, 1838, pp. 113–121.

[36] L. Wasserman. *All of Statistics: a concise course in statistical inference*. New York: Springer, 2004.

[37] R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing*. 3rd. Amsterdam: Elsevier, 2012.

[38] B. Ycart. "Fluctuation analysis: can estimates be trusted?" In: *PLoS One* 8.12 (2013), pp. 1–12.

[39] B. Ycart. "Fluctuation analysis with cell deaths". In: *J. Appl. Probab. Statist* 9.1 (2014), pp. 13–29.

[40] B. Ycart and N. Veziris. "Unbiased estimates of mutation rates under fluctuating final counts". In: *PLoS One* 9.7 (2014), pp. 1–10.

[41] G.U. Yule. "A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S." In: *Phil. Trans. Roy. Soc. London Ser. B* 213 (1925), pp. 21–87.

[42] Q. Zheng. "New algorithms for Luria-Delbrück fluctuation analysis". In: *Math. Biosci.* 196.2 (2005), pp. 198–214.