Some mathematical tools for the Lenski experiment

Bernard Ycart, Agnès Hamon, Joël Gaffé, Dominique Schneider

Some mathematical tools for the Lenski experiment

Bernard Ycart* Agnès Hamon† J. Gaffé‡ D. Schneider§

Abstract

The Lenski experiment is a long term daily reproduction of *Escherichia coli*, that has evidenced phenotypic and genetic evolutions along the years. Some mathematical models, that could be usefull in understanding the results of that experiment, are reviewed here: stochastic and deterministic growth, mutation appearance and fixation, competition of species.
Keywords: cell kinetics; branching processes; Luria-Delbrück distribution
MSC: 92D25

1 Introduction

This is primarily intended as a review of a very small part of the vast litterature on mathematical models applied to population biology. The application that we have in mind is the *Lenski experiment* (referred to as LE hereafter, see [4, 43, 64, 71, 72, 73]). Here is a much simplified description (see Philippe *et al.* [64] for a more detailed account). The ancestral strain was *Escherichia Coli B*. Twelve populations have been propagated by daily serial transfer, using a 1:100 dilution, in the same defined environment, for more than 40,000 generations. The growth medium supports a maximum of 5×10^7 cells per mL; therefore the number of cells in each one of the twelve 10 mL vessels varies daily from 5×10^6 to 5×10^8. Our main focus will be on modelling the appearance and fixation of mutations along successive generations. The total number of mutational events that have happened during the 40,000 generations in each population is estimated at 3×10^9 ([64], p. 849). Our main question here is: how did some of the mutant strains survive and eventually invade the population? We argue that only beneficial mutations (*i.e.* increasing the fitness) can have survived the experimental process. Indeed when a nonbeneficial mutation occurs a given day, the proportion of mutants in the population remains extremely small, and therefore the chances that mutants survive the 1:100 sampling to the next day are very small. On the contrary, a strain carrying a beneficial mutation survives if mutant cells can be found in sizeable amounts at the moment of dilution. Provided the new strain survives the first day dilution, chances are that its proportion in the global population will rapidly increase. When that proportion is close to one, the initial strain has a high probability to be wiped out by daily dilution. Our objective here is to provide a quantified basis to the above assertions.

Although our emphasis will mainly be on probability, we shall use some elementary population dynamics, on which our basic references are Murray [58] and Kot [45]. The stochastic tools used here, essentially discrete sampling and birth-and-death processes, are presented in many excellent textbooks. For probability theory, Feller's treatise (in particular the second volume [22]) remains a good reference. Ross [68] and Tuckwell

*LJK, CNRS UMR 5224, Univ. Grenoble-Alpes, France Bernard.Ycart@imag.fr
†LJK, CNRS UMR 5224, Univ. Grenoble-Alpes, France Agnes.Hamon@imag.fr
‡LAPM, CNRS UMR 5163, Univ. Grenoble-Alpes, France Joel.Gaffe@ujf-grenoble.fr
§LAPM, CNRS UMR 5163, Univ. Grenoble-Alpes, France Dominique.Schneider@ujf-grenoble.fr

[77] are oriented towards applications in biology. The basic theory of continuous time Markov processes is developed in Bharucha-Reid [11] and Çinlar [18]. Ethier & Kurtz [20] give a more advanced treatment, in particular of asymptotics and convergence to diffusion processes. Some textbooks, such as Gardiner [25] are particularly oriented to applications. Being one of the founders of the theory, Bartlett [8, 9] emphasizes population modelling. Moran [57] develops genetic applications. Some more recent textbooks have been devoted to stochastic modelling in biology, such as Ewens [21] and Wilkinson [83]; Allen [2] and Durrett [17] are of particular interest to us. The course given on "Stochastic Population Systems" by D. Dawson to the Summer School in probability at PIMS-UBC in 2009 covers all the material that we shall use and much more. The lecture notes are available on line[1].

The paper is organized as follows. The daily sampling process will be examined first, in section 2. Next, we shall review in section 3 some deterministic models for the evolution of two competing populations (normal and mutant cells) in a constrained environment. On a very simple model, it will shown that the final proportion of mutants (*i.e.* before the next dilution) can be deduced from the initial proportion (*i.e.* right after the previous dilution) and the two division rates of normal and mutant cells. Section 4 will be devoted to the stochastic counterpart of the deterministic models examined in section 3. The emphasis will be on the convergence of stochastic to deterministic models. Stochastic mutation models will be reviewed in section 5. There we shall focus on the first appearance of mutations and the proportion of mutants at the end of the first day. All models considered here are based on several parameters, among which at least individual division rates and mutation probabilities. The question of estimating those parameters is therefore crucial, and it will be reviewed in section 6.

2 Daily sampling

The problem of beneficial mutations being lost in populations with periodic bottlenecks has been studied by Wahl and Gerrish [82]. Our aim here is to give a simple quantitative evaluation of the survival of mutations from one day to the next. Recall that the maximal capacity of the medium is $n_f = 5 \times 10^8$. At the end of each day, n_f cells are present on the population, out of which $n_0 = n_f/100$ are sampled. Assume that at the end of one day, N normal cells and $M = n_f - N$ mutant cells are present. How many mutant cells will remain after the 1:100 dilution? The simplest model assumes equiprobability: assuming that the medium is homogeneous, each n_0-sized sample has the same probability to be chosen for the next day. Therefore the number of mutants surviving dilution follows the hypergeometric distribution with parameters n_f, M and n_0. In other words, the chances to find exactly n normal cells and $m = n_0 - n$ mutant cells after dilution are:

$$\frac{\binom{M}{m}\binom{N}{n}}{\binom{n_f}{n_0}} .$$

Since n_f is large, and n_0 is small compared to n_f, the hypergeometric distribution can be correctly approximated by the binomial with parameters n_0 and $p = M/n_f$ (proportion of mutants). As a consequence, the probability to find exactly m mutants is

$$\frac{\binom{M}{m}\binom{N}{n}}{\binom{n_f}{n_0}} \simeq \binom{n_0}{m} p^m (1-p)^{n_0-n} .$$

If p is large enough (say $p > 10^{-4}$), the binomial distribution above can be approximated by a Gaussian, with expectation $n_0 p$ and variance $n_0 p(1-p)$. In other words, the

[1] http://www.math.ubc.ca/ db5d/SummerSchool09/LectureNotes.html

proportion $p' = m/n_0$ will be close to p, a 95% fluctuation interval being:

$$p \pm 1.96 \frac{\sqrt{p(1-p)}}{\sqrt{n_0}} \ .$$

Since $n_0 = 5\,10^6$, one can express it differently by saying that with 98.73% probability, p' is at distance at most 5×10^{-4} of p.

However, if p is too small, the approximation above is no longer valid. A better approximation is given by the Poisson distribution with parameter $n_0 p$:

$$\frac{\binom{M}{m}\binom{N}{n}}{\binom{n_f}{n_0}} \simeq \mathrm{e}^{-n_0 p} \frac{(n_0 p)^m}{m!} \ .$$

We should like to emphasize here that when M is small (a few units), chances that all mutant cells disappear after dilution are high. Indeed, the probability that no mutant cell remains is:

$$\frac{\binom{N}{n_0}}{\binom{n_f}{n_0}} \simeq \mathrm{e}^{-n_0 M/n_f} \simeq 1 - \frac{M}{100} \ .$$

For instance, if $M = 10$ mutant cells are present, there is a 90% probability that they will all disappear after dilution. From the 100-fold increase, one can evaluate the daily number of generations from a given cell to be of order 6 ($2^6 = 64$) to 7 ($2^7 = 128$). The table below gives the values of the probability of disappearance for each value of $M = 2^d$, d (the number of generations) ranging from 0 to 8. Numerical calculations (from the exact hypergeometric distribution) were obtained through R [67].

divisions	0	1	2	3	4	5	6	7	8
cells	1	2	4	8	16	32	64	128	256
proba of disappearance	0.99	0.98	0.96	0.92	0.85	0.72	0.53	0.28	0.08

3 Deterministic growth models

The simplest deterministic model, that of exponential growth, is named after Malthus [53]:

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = \nu N(t) \ ,$$

where $N(t)$ denotes the number of cells at time t, and ν is the individual division rate (IDR) of each cell. That model had been considered long before Malthus, in particular by Euler. What Malthus actually discussed in 1798 was precisely the growth limitation due to lack of resources.

> It may safely be pronounced, therefore, that population, when unchecked, goes on doubling itself every twenty-five years, or increases in a geometrical ratio.
>
> The rate according to which the productions of the earth may be supposed to increase, it will not be easy to determine. Of this, however, we may be perfectly certain, that the ratio of their increase must be totally of a different nature from the ratio of the increase of population.

Quetelet was the first one to propose to account for growth limitations by substracting a quadratic correction term to the differential equation above, in 1836 (see [66] p. 288).

> La population tend à croître selon une progression géométrique.
>
> La résistance, ou la somme des obstacles à son développement, est, toutes choses égales d'ailleurs, comme le carré de la vitesse avec laquelle la population tend à croître.

Two years later, Verhulst [79] discussed various other limitating terms and compared the models so obtained to experimental data. The model with a quadratic limitation is now known as Verhulst model, or else *logistic model*.

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = \nu N(t)\left(1 - \frac{N(t)}{n_f}\right) ,$$

where n_f is the maximal population sustained by the environment. The model was generalized to several species competing in the same environment by Lotka [50] in 1925, and Volterra [80] in 1926. Volterra discussed the different models in a course given in 1928 to the newly founded Institut Henri Poincaré. The lecture notes "Lessons on the mathematical theory of the struggle for life" [81], to which we shall refer, appeared in 1931. Soon after, Lotka and Volterra predictions were confronted to experimental data coming from all sorts of different contexts, in particular by Gause [26].

Here are the first lines of Volterra [81], from section 1.1 entitled "Deux espèces se disputant la même nourriture".

> Supposons que, avec une nourriture en quantité suffisante pour satisfaire complètement la voracité de ces êtres, il y ait des coefficients d'accroissement positifs et constants ε_1, ε_2. Si nous nous plaçons maintenant dans le cas réel d'espèces vivant dans un milieu délimité, la nourriture diminuera quand les nombres N_1, N_2 des individus des deux espèces augmenteront et cela fera baisser la valeur des coefficients d'accroissement. Si l'on représente la nourriture dévorée par unité de temps par $F(N_1, N_2)$ fonction nulle avec N_1, N_2 ensemble, tendant vers l'infini avec chacune des variables et fonction croissante de chacune d'elles, il sera assez naturel de prendre comme coefficients d'accroissement
>
> $$\varepsilon_1 - \gamma_1 F(N_1, N_2) , \qquad \varepsilon_2 - \gamma_2 F(N_1, N_2) ,$$
>
> γ_1, γ_2 étant des constantes positives correspondant aux deux espèces et à leurs besoins respectifs de nourriture.
>
> D'où le système différentiel traduisant le développement des espèces.
>
> $$\frac{\mathrm{d}N_1}{\mathrm{d}t} = [\varepsilon_1 - \gamma_1 F(N_1, N_2)]N_1 , \qquad \frac{\mathrm{d}N_2}{\mathrm{d}t} = [\varepsilon_2 - \gamma_2 F(N_1, N_2)]N_1 . \qquad (1)$$
>
> Maintenant se pose le problème mathématique d'étudier les intégrales N_1, N_2 de ce système, avec des valeurs initiales N_1^0, N_2^0 positives pour $t = t_0$.
>
> On peut démontrer que pour tout intervalle fini (t_0, T) il y a une solution unique, de deux fonctions continues, restant comprises entre deux nombres positifs, le plus grand ne dépendant pas de l'extrémité T de l'intervalle (c'est-à-dire que N_1, N_2 restent bornés).
>
> Étudions ce qui arrive quand le temps s'écoule indéfiniment. En transcrivant (1) sous la forme
>
> $$\frac{\mathrm{d}\log N_1}{\mathrm{d}t} = \varepsilon_1 - \gamma_1 F(N_1, N_2) , \qquad \frac{\mathrm{d}\log N_2}{\mathrm{d}t} = \varepsilon_2 - \gamma_2 F(N_1, N_2) , \qquad (1')$$
>
> il vient par combinaison
>
> $$\gamma_2 \frac{\mathrm{d}\log N_1}{\mathrm{d}t} - \gamma_1 \frac{\mathrm{d}\log N_2}{\mathrm{d}t} = \varepsilon_1\gamma_2 - \varepsilon_2\gamma_1 ,$$
>
> puis
>
> $$\frac{N_1^{\gamma_2}}{N_2^{\gamma_1}} = \frac{(N_1^0)^{\gamma_2}}{(N_2^0)^{\gamma_1}} \, \mathrm{e}^{(\varepsilon_1\gamma_2 - \varepsilon_2\gamma_1)(t-t_0)} . \qquad (2)$$

Négligeons le cas infiniment peu probable où

$$\varepsilon_1\gamma_2 - \varepsilon_2\gamma_1 = 0$$

et supposons, en permutant au besoin les espèces, que

$$\varepsilon_1\gamma_2 - \varepsilon_2\gamma_1 > 0 \quad \text{ou} \quad \frac{\varepsilon_1}{\gamma_1} > \frac{\varepsilon_2}{\gamma_2} \; ;$$

alors, d'après (2),

$$\lim_{t=\infty} \frac{N_1^{\gamma_2}}{N_2^{\gamma_1}} = +\infty \; . \tag{3}$$

N_1 restant borné, N_2 tend donc vers 0.

Nous concluerons donc que *la seconde espèce, celle de $\frac{\varepsilon}{\gamma}$ le plus petit, s'épuise* et disparaît, *tandis que la première subsiste.*

More as an example than a realistic model, we shall consider a particular case of Volterra's system. Let us denote by $N(t)$ the number of normal cells at time t, and by $M(t)$ the number of mutant cells. Let ν and μ be their respective IDR's, and n_f denote the maximal capacity (total number of cells sustained by the medium). Our model, that we shall call *Volterra Model* (VM) even though it is only a very particular case of (1) deemed "infiniment peu probable" by Volterra, will be:

$$\left\{ \begin{array}{rcl} \dfrac{\mathrm{d}N(t)}{\mathrm{d}t} & = & \nu N(t)\left(1 - \dfrac{N(t)+M(t)}{n_f}\right) \\ \dfrac{\mathrm{d}M(t)}{\mathrm{d}t} & = & \mu M(t)\left(1 - \dfrac{N(t)+M(t)}{n_f}\right) \end{array} \right. \tag{VM}$$

In other words, the growth of both normal and mutant cells is equally restricted, proportionally to their sum. It stops when that sum reaches the maximal capacity n_f. Of course, the IDR's ν and μ should be estimated, and we shall review estimation methods in section 6. For the moment, we shall assume that the time scale has been set so that $\nu = 1$. Therefore $\mu = \mu/\nu$ can be viewed as the *fitness* of mutants, and we shall mainly consider here the case $\mu \geqslant 1$.

No explicit solution can be given to the system of differential equations (VM). However, it is quite easy to compute a numerical solution. Figure 1 plots typical trajectories for $N(t)$ and $M(t)$, all starting from the same initial values $N(0) = 0.9 \times (5 \times 10^6)$ and $M(0) = 0.1 \times (5 \times 10^6)$. The value of μ (the relative fitness of mutants) ranges from 1.2 to 2. All curves have the same logistic-type shape. They increase exponentially at first (when the nutrient limitation is not yet significant), then converge asymptotically to a limit value as t tends to $+\infty$. The calculations show that the asymptotic value is reached in practice for relatively small values of t. All numerical calculations and plots were made through Scilab [14].

Let us denote by $N(\infty)$ and $M(\infty)$, the limits as t tends to infinity of $N(t)$ and $M(t)$. They can be computed using Volterra's method:

$$\left\{ \begin{array}{rcl} \dfrac{(N(\infty))^\mu}{(M(\infty))^\nu} & = & \dfrac{(N(0))^\mu}{(M(0))^\nu} \\ N(\infty) + M(\infty) & = & n_f \end{array} \right.$$

Our main interest is in the evolution of the relative *proportion* of mutants at time t, denoted by $P(t)$:

$$P(t) = \frac{M(t)}{N(t) + M(t)} \; .$$

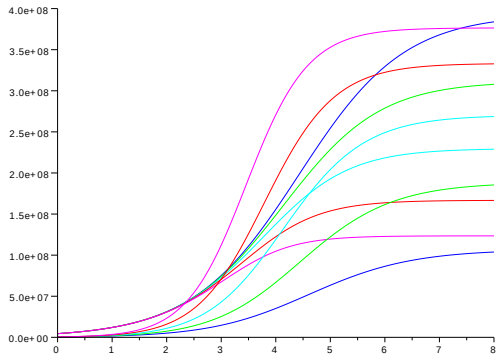

Figure 1: Evolution in time of the numbers of normal and mutant cells in time in the Volterra model. The initial proportion of mutants is 0.1. The IDR of normal cells is 1. That of mutant cells is 1.2 (blue), 1.4 (green), 1.6 (light blue), 1.8 (red), 2 (purple).

Figure 2 shows the evolution in time of $P(t)$, with the same values of the parameters as before. Unlike the general case studied by Volterra, the two populations reach an equilibrium where the proportion of mutants is not 1. Nevertheless the IDR of mutants being larger than that of normal cells, $P(t)$ increases in time. It converges to an asymptotic value, denoted by $P(\infty)$. That "limit proportion of mutants" can be viewed as a function both of the initial proportion $P(0)$ and the fitness μ/ν. Figure 3 plots that function, for small values of $P(0)$ and fitnesses ranging from 1 to 5.

Consider a given strain of mutants with fitness $\mu > 1$, and denote by f_μ the (increasing) function that maps $P(0)$ onto $P(\infty)$. Say that the mutation has taken place on day 0, and denote by P_0 the (small but positive) proportion of mutant cells at the end of day 0. After dilution, the proportion of mutant cells at the beginning of day 1 will be a random variable with expectation P_0 (see the discussion in section 2). Provided it is non null, at the end of day 1 the new proportion will be the image by f_μ of the initial

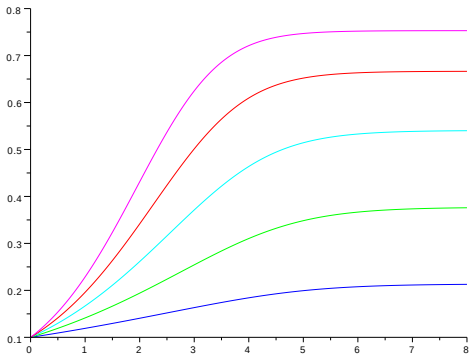

Figure 2: Evolution in time of the proportion of mutant cells in time in the Volterra model. The initial proportion of mutants is 0.1. The IDR of normal cells is 1. That of mutant cells is 1.2 (blue), 1.4 (green), 1.6 (light blue), 1.8 (red), 2 (purple).

one. Dilution transforms it in another random variable P_1 with expectation $f_\mu(P_0)$, and so on. Denote by P_n the proportion of mutants at the beginning of day n. At the end of day n, it will be $f_\mu(P_n)$, then at the beginning of day $n+1$, P_{n+1} will be a new random variable with expectation $f_\mu(P_n)$. The sequence of random variables so defined is a Markov chain, converging to 1 (mutants will eventually invade the whole population). Using renewal theory techniques, the distribution of the (random) number of days before complete invasion can be precisely estimated. The rapid growth of f_μ (figure 3) makes it quite likely that even starting at day 0 with a small proportion of mutants, the number of days to complete invasion will be relatively small, conditionally of course to the fact that mutants do not disappear upon dilution in the first days. An estimate for the number of days until complete invasion can be obtained by iteratively applying f_u, starting with $p_0 = 1/n_0$ (a single mutant on day 0). The results are plotted on figure 4: with a relative fitness of 1.2, it takes about 25 days for mutant cells to invade the population, whereas

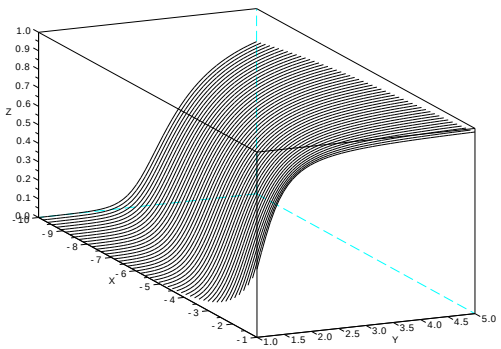

Figure 3: Limit proportion of mutant cells in the Volterra model (Z-axis, normal scale). The initial proportion $P(0)$, ranges from $e^{-1} = 0.368$ to $e^{-10} = 4.5 \times 10^{-5}$ (X-axis, log-scale). The 50 fitnesses are linearly spaced between 1 and 5 (Y-axis, normal scale).

with a relative fitness of 2, it takes 5 days.

Symmetrically, the same techniques will show that for a nonbeneficial mutation ($\mu \leqslant 1$), there are very little chances that the mutation will survive day 2, even if it survives dilution on day 1.

A more sophisticated model, adapted to the context of bacterial growth experiments, was introduced by Herbert, Ellsworth & Telling, and independenty by Powell in 1956 (see [65, 33], and references therein). A description is given in chapter 12 of Kot [45]. The model takes into account the nutrient, and follows the evolution of bacterial cells as a Michaelis-Menten enzymatic reaction (see [36] for a historical account of Michaelis & Menten's original paper). To the number of normal and mutant cells $N(t)$ and $M(t)$, one adds the substrate (nutrient) concentration $S(t)$. The model, that we shall call (somewhat arbitrarily) *Powell Model* or PM is given by the following system of differential equations

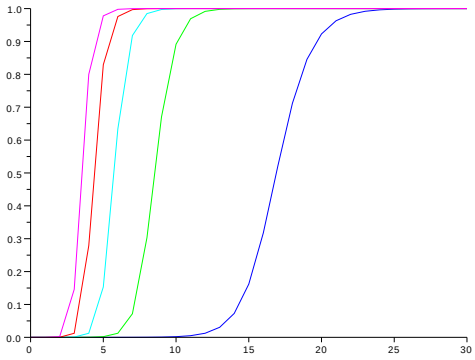

Figure 4: Daily evolution for the proportion of mutant cells. The initial proportion is $1/(5\,10^6)$ at the beginning of day 0. The IDR of normal cells is 1. That of mutant cells is 1.2 (blue), 1.4 (green), 1.6 (light blue), 1.8 (red), 2 (purple).

(see Kot [45], system (12.16) p. 205).

$$\left\{\begin{array}{rcl} \dfrac{\mathrm{d}S(t)}{\mathrm{d}t} & = & D(S_i - S(t)) - \dfrac{1}{y_n}\dfrac{\nu S(t)N(t)}{k_n + S(t)} - \dfrac{1}{y_m}\dfrac{\mu S(t)N(t)}{k_m + S(t)} \\ \dfrac{\mathrm{d}N(t)}{\mathrm{d}t} & = & \dfrac{\nu S(t)N(t)}{k_n + S(t)} - DN(t) \\ \dfrac{\mathrm{d}M(t)}{\mathrm{d}t} & = & \dfrac{\mu S(t)M(t)}{k_m + S(t)} - DN(t) \end{array}\right. \tag{PM}$$

The parameters are the following.

- ν, μ: the IDR's of normal and mutant cells,
- y_n, y_m: the yield coefficients of normal and mutant cells (consumption of substrate by bacteria per unit of time),

- k_n, k_m: the half-saturation constants of normal and mutant cells,
- D: the dilution,
- S_i: the inflowing substrate concentration.

In our case, one can consider that $S_i = 0$ for each single day. Besides being closer to the experimental reality, the PM has another interesting feature. In the VM we considered that the evolutionary advantage of the mutants translated only into a faster division rate. It might also yield a higher consumption of nutrients, which would even accelerate the invasion. A mathematical study of the PM is no more difficult than for the VM; but it would require estimates for 7 parameters instead of 3. Nevertheless, we believe that the qualitative conclusions that were drawn from the VM would still remain true for the PM; they are summarized below.

1. nonbeneficial mutations disappear over a few days, through the effect of daily dilution
2. beneficial mutations, if the strain survives dilution on the first day, have a high probability to invade the full population, and eventually to eliminate normal cells, after at most a few tens of days.
3. On a single day, two competing strains of bacteria increase at different rates, according to their relative fitnesses; their populations eventually stabilize to asymptotic values, the relative proportion of the fittest being strictly larger at the end of the day than at the beginning.

We also wish to point out that the models above can be easily extended to more than 2 competing species. As a common feature, the generalizations will obey the *principle of competitive exclusion* (see section 3.5 of Murray [58]): in the long term, the fittest species eliminates the others.

The fixation along successive generations of beneficial mutations has been discussed in many references and textbooks. Papp *et al.* [63] recently wrote an interesting review on systems-biology approaches to genomic evolution. Lang *et al.* [48] propose a thorough experimental study in different strains of yeast. Campos & Wahl [12, 13] discuss the effect of population bottlenecks. Hermisson & Pfaffelhuber [32] study genetic hitchhiking under recurrent mutations.

4 Stochastic growth models

The probabilistic modelling of population growth started in 1925 with Yule [84]. Such an early date is somewhat misleading. The burst of the theory took place in a few years around 1950 with such important contributors as Bartlett [5, 6, 7], Harris [31, 10] and Kendall [39, 40, 41]. More references and a perspective on the early historical development can be found in the discussion following Armitage's presentation to the Royal Statistical Society [3], the chairman being Bartlett. Kendall [41] gives another useful early review. Of course the same material has since appeared in many textbooks, starting with Bartlett's [8, 9] (see also Bharucha-Reid [11]). Since the 1950's, stochastic population models have remained a lively subject: see *e.g.* Pakes [62] for a more recent review. Aldous [1] gives an interesting perpective relating Yule's founding paper to present researches.

The basic hypothesis of all models is that all bacteria behave independently in the same manner; of course this should be understood in the stochastic sense: the times at which all bacteria divide are independent and identically distributed random variables. The common distribution function of these random times will be denoted by F. The two simplest particular cases are:

- *deterministic:* $F(t) = \mathbb{I}_{[t_0,+\infty)}(t)$ (the division time is fixed and equal to t_0).

- *Markovian:* $F(t) = (1 - \mathrm{e}^{-\nu t})\mathbb{I}_{[0,+\infty)}(t)$ (the division time is exponentially distributed with parameter ν).

These are the cases where an explicit mathematical treatment is easily feasible, and we shall focus on the second one. More general families of distributions have been considered, including Gamma distributions (see Kendall [39, 41] for a discussion and comparison with experimental data).

When a division occurs, the cell that divides is turned into a random number of new bacteria, each having a fresh division date. Let φ be the generating function of the random number of "children" produced at division time:

$$\varphi(z) = \mathbb{E}[z^K] = \sum_{k=0}^{+\infty} z^k \, \mathbb{P}[K = k] \, ,$$

where K is the random number of children at any division. Even though we shall be concerned exclusively by the deterministic case $K \equiv 2$ (or $\varphi = z^2$), let us mention that the general case is the basis of the famous Galton-Watson model of branching processes. Notice also that K can be null, which corresponds to a death.

Generating functions of integer valued random variables will be used extensively in what follows. Let us recall the following basic properties:

1. If two random variables N and M are independent, then the generating function of $N + M$ is the product of the generating functions of N and M.
2. The generating function of N has a k-th left derivative at $z = 1$ if and only if $\mathbb{E}[N^k]$ exists. In that case,

$$\lim_{z \to 1^-} \frac{\mathrm{d}\mathbb{E}[z^N]}{\mathrm{d}z} = \mathbb{E}[N(N-1)\cdots(N-k+1)] \, .$$

The basic object of interest is the number of bacteria alive at time t, when one single cell is present at time 0; its generating function will be denoted by $G(z, t)$.

$$G(z,t) = \sum_{n=0}^{+\infty} z^n \, \mathbb{P}[N_t = n \,|\, N_0 = 1] \, ,$$

where N_t denotes the (random) number of bacteria living at time t, under the hypothesis that $N_0 \equiv 1$. The function G is related to F and φ through the *Bellman-Harris integral equation* (see [10]).

$$G(z,t) = \left(\int_0^t \varphi(G(z, t-s)) \, \mathrm{d}F(s) \right) + z(1 - F(t)) \, . \tag{BH}$$

Even though we shall not provide a rigorous proof, giving an intuitive interpretation of that equation can still be useful for future developments. Let us argue as follows. The distribution function $F(t)$ is the probability that the division of the initial cell has taken place no later than t; $1 - F(t)$ is the probability it has occurred after t, $\mathrm{d}F(s)$ is the probability that it occurs between s and $s + \mathrm{d}s$. If at time t the division has not occurred (probability $(1 - F(t))$) then the number of bacteria is still equal to 1, and its generating function is z: this accounts for the second part of the right hand side, $z(1 - F(t))$. Suppose now that a division has occurred at some time s between 0 and t (in $[s, s + \mathrm{d}s]$ with probability $\mathrm{d}F(s)$). Assume it leads to k new cells, each starting a new lineage. Since all lineages develop with identical rules, the population at time t, stemming from one lineage which started at time s, will have generating function $G(z, t - s)$. The total number will be the sum of all k lineages, which are supposed to be independent: its

generating function will be the k-th power $G^k(z, t-s)$. Now since k cells are produced with probability $\mathbb{P}[K=k]$, the generating function at time t becomes

$$\sum_{k=0}^{+\infty} G^k(z,t-s)\,\mathbb{P}[K=k] = \varphi(G(z,t-s))\ .$$

This accounts for the first term in the (BH) equation.

As a particular case, when $F(t) = (1-\mathrm{e}^{\nu t})\mathbb{I}_{[0,+\infty)}(t)$ (exponential distribution, Markovian case), and $K \equiv 2$ ($\varphi(z) = z^2$), the Bellman-Harris equation (BH) becomes:

$$G(z,t) = \left(\int_0^t G^2(z,t-s)\nu\mathrm{e}^{-\nu s}\,\mathrm{d}s\right) + z\mathrm{e}^{-\nu t}\ .$$

Inside the integral, change $t-s$ into u:

$$G(z,t) = \left(\mathrm{e}^{-\nu t}\int_0^t G^2(z,u)\nu\mathrm{e}^{\nu u}\,\mathrm{d}u\right) + z\mathrm{e}^{-\nu t}\ .$$

or else:

$$G(z,t)\mathrm{e}^{\nu t} = \left(\int_0^t G^2(z,u)\nu\mathrm{e}^{\nu u}\,\mathrm{d}u\right) + z\ .$$

The derivative in time is:

$$\frac{\partial G(z,t)}{\partial t}\mathrm{e}^{\nu t} + G(z,t)(\nu\mathrm{e}^{\nu t}) = G^2(z,t)(\nu\mathrm{e}^{\nu t})\ .$$

Simplifying by $\mathrm{e}^{\nu t}$ and rearranging terms, leads to the following differential equation.

$$\frac{\partial G(z,t)}{\partial t} = \nu(G^2(z,t) - G(z,t))\ . \tag{DE}$$

The solution to (DE) for $G(z,0) = z$ is easily computed:

$$G(z,t) = \frac{\mathrm{e}^{-\nu t}z}{1-(1-\mathrm{e}^{-\nu t})z}\ .$$

This is the generating function of the geometric distribution with parameter $p = \mathrm{e}^{-\nu t}$. In other terms, for all $n \geqslant 1$, the probability that n bacteria are alive at time t is:

$$\mathbb{P}[N_t = n] = \mathrm{e}^{-\nu t}(1-\mathrm{e}^{-\nu t})^{n-1}\ .$$

The first two moments of N_t are:

$$\mathbb{E}[N_t] = \frac{1}{p} = \mathrm{e}^{\nu t} \qquad \text{and} \qquad \mathrm{Var}[N_t] = \frac{1-p}{p^2} = \mathrm{e}^{2\nu t} - \mathrm{e}^{\nu t}\ .$$

The Markov process $\{N_t\,,\ t \geqslant 0\}$ is one of the simplest examples of a birth-and-death process (actually a "pure birth" process). It is usually named after Yule (from [84]), who derived the geometric distribution of N_t. Furry [24] introduced the same process in another context. Novozhilov *et al.* [59] give a very simple presentation of birth-and-death processes used in biology.

In the LE, the number of bacteria varies from $n_0 = 5\times10^6$ to $n_f = 5\times10^8$. So it is legitimate to consider large number approximations. If n_0 bacteria are initially present, their lineages evolve independently, according to the distribution that has just been described: at each time t, the total population is distributed as the sum of n_0 independent random variables, each following the geometric distribution with parameter

$\mathrm{e}^{-\nu t}$, that is a Pascal distribution with parameters n_0 and $\mathrm{e}^{-\nu t}$. If n_0 is large, it can be approximated by a Gaussian (normal) distribution with same mean and variance, *i.e.*:

$$\mathbb{E}[N_t] = \frac{n_0}{p} = n_0 \mathrm{e}^{\nu t} \qquad \text{and} \qquad \mathrm{Var}[N_t] = n_0 \frac{1-p}{p^2} = n_0 \mathrm{e}^{2\nu t} - n_0 \mathrm{e}^{\nu t} \; .$$

Observe that for n_0 large, the standard-deviation is small compared to the expectation:

$$\mathbb{E}[N_t] = \frac{n_0}{p} = n_0 \mathrm{e}^{\nu t} \qquad \text{and} \qquad \sqrt{\mathrm{Var}[N_t]} \simeq \sqrt{n_0 \mathrm{e}^{2\nu t}} = \frac{\mathbb{E}[N_t]}{\sqrt{n_0}} \; .$$

Therefore, as a first approximation, the population is expected to grow deterministically, as $n_0 \mathrm{e}^{\nu t}$.

The Yule process is the stochastic counterpart of the deterministic exponential growth model. Let us denote by $N(t)$ and $n(t)$ the respective numbers of cells in the stochastic model and in the deterministic one. The deterministic model is specified by an ordinary differential equation.

$$\frac{\mathrm{d}n(t)}{\mathrm{d}t} = \nu \, n(t) \; ,$$

or equivalently by the integral equation

$$n(t) = n_0 + \int_0^t \nu \, n(s) \, \mathrm{d}s \; .$$

The basis of the stochastic model is a random time scale, specified by a Poisson process. Let $\{Y(t) \, , \; t \geqslant 0\}$ be a unit Poisson process: Y has jumps of size 1 at successive instants, separated by exponentially distributed durations with expectation 1. The stochastic process $N(t)$ can be defined by the following integral equation, which is the stochastic counterpart of the deterministic one.

$$N(t) = n_0 + Y\left(\int_0^t \nu N(s) \, \mathrm{d}s \right) \; .$$

The distribution at time t of $N(t)$ is the solution of a system of ordinary differential equations, the *Chapmann-Kolmogorov* equations (also called the *Chemical Master Equation* in the context of kinetics). Denoting by $p_n(t)$ the probability that $N(t) = n$, the equation for $n > 0$ is:

$$\frac{\mathrm{d}p_n(t)}{\mathrm{d}t} = -\nu n p_n(t) + \nu(n-1) p_{n-1}(t) \; .$$

As we have seen, that system has an explicit solution. However, it is a very particular case and no explicit solution can be hoped for in more general situations. Also, the fact that the expectation of $N(t)$ is exactly equal to the deterministic solution $n(t)$ is quite rare. What is general however, is the large number approximation that was explained above. We shall rewrite it in a way that can be easily generalized. Its foundation is the long term approximation for the Poisson process Y. At (large) time $n_0 u$, $Y(n_0 u)$ equals $n_0 u$ on average, with fluctuations of order $\sqrt{n_0}$, described by a Brownian motion.

$$\lim_{n \to +\infty} \frac{Y(n_0 u) - n_0 u}{\sqrt{n_0}} = W(u) \; ,$$

where Y is a unit Poisson process, W is the standard Brownian motion and the limit is understood in distribution. Consider now the integral equation defining $N(t)$ and divide both members by n_0.

$$\frac{N(t)}{n_0} = 1 + \frac{1}{n_0} Y\left(\int_0^t \nu N(s) \, \mathrm{d}s \right) \; .$$

Denote by $U(t)$ the ratio $\frac{N(t)}{n_0}$.

$$U(t) = 1 + \frac{1}{n_0} Y\left(\int_0^t \nu n_0 U(s)\,\mathrm{d}s\right) .$$

Approximating $Y(n_0 u)$ by $n_0 u + \sqrt{n_0} W(u)$:

$$U(t) \simeq 1 + \int_0^t \nu U(s)\,\mathrm{d}s + \frac{1}{\sqrt{n_0}} W\left(\int_0^s \nu U(s)\,\mathrm{d}s\right) .$$

The diffusion process U is the solution of a *Stochastic Differential Equation* (called the *Chemical Langevin Equation* in kinetics)

$$\mathrm{d}U(t) = \nu U(t)\,\mathrm{d}t + \frac{1}{\sqrt{n_0}} \sqrt{\nu U(t)}\,\mathrm{d}W(t) .$$

Of course for very large n_0, the diffusion term vanishes and the equation becomes the deterministic differential equation (the *Reaction Rate Equation* in kinetics).

What we have just seen on the example of the Yule process is an illustration of a very general modelling principle. Three different *scales* can be considered.

1. *Microcopic scale*: stochastic jump process with discrete state space. The random fluctuations are governed by Poisson processes;
2. *Mesoscopic scale*: stochastic diffusion process with continous state space. The random fluctuations are governed by Brownian motions;
3. *Macroscopic scale*: deterministic function of time, solution of a differential system of equations.

The three scales are coherent in the sense that each scale is a large number approximation of the previous one.

In order to illustrate the three scales principle, we shall introduce here a microscopic model of competition between normal and mutant cells, matching the Volterra model studied in section 3. Let $N(t)$ and $M(t)$ still denote the numbers of normal and mutant cells. At each division, one of the two counts increases by one unit. When $(N(t), M(t)) = (n, m)$, the respective rates of increase of normal and mutant cells are:

$$\rho(n,m) = \nu n \left(1 - \frac{n+m}{n_f}\right) \quad \text{and} \quad \sigma(n,m) = \mu m \left(1 - \frac{n+m}{n_f}\right) .$$

We shall call this model *competitive process* (CP). A loose description of the CP can be given as follows.

1. When n normal and m mutant cells are present, the next division will occur after a random time, following the exponential distribution with parameter $\rho(n,m) + \sigma(n,m)$.
2. Upon next division,
   (a) with probability $\frac{\rho(n,m)}{\rho(n,m)+\sigma(n,m)}$, $N(t)$ will increase by 1, $M(t)$ remaining unchanged.
   (b) with probability $\frac{\sigma(n,m)}{\rho(n,m)+\sigma(n,m)}$, $M(t)$ will increase by 1, $N(t)$ remaining unchanged.

The formal description uses two independent unit Poisson processes, Y_1 and Y_2.

$$\left\{ \begin{array}{rcl} N(t) & = & N(0) + Y_1\left(\displaystyle\int_0^t \rho(N(s), M(s))\,\mathrm{d}s\right) \\ M(t) & = & M(0) + Y_2\left(\displaystyle\int_0^t \sigma(N(s), M(s))\,\mathrm{d}s\right) \end{array} \right. \qquad \text{(CP)}$$

Let p_0 be the initial proportion of mutant molecules, so that $N(0) = n_0(1-p_0)$ and $M(0) = n_0 p_0$. Assuming that n_0 is large, we reproduce the diffusion approximation scheme that was exposed above for the Yule process. Dividing both equations by n_0, then setting

$$U(t) = \frac{N(t)}{n_0} \quad \text{and} \quad V(t) = \frac{M(t)}{n_0} \ ,$$

one gets:

$$\left\{ \begin{array}{rcl} U(t) & = & 1 - p_0 + \dfrac{1}{n_0} Y_1 \left(\displaystyle\int_0^t \nu n_0 U(s) \left(1 - \frac{n_0}{n_f}(U(s)+V(s)) \right) \mathrm{d}s \right) \\ V(t) & = & p_0 + \dfrac{1}{n_0} Y_2 \left(\displaystyle\int_0^t \mu n_0 V(s) \left(1 - \frac{n_0}{n_f}(U(s)+V(s) \right) \mathrm{d}s \right) \end{array} \right.$$

Replace now $Y_1(n_0 u)$ and $Y_2(n_0 v)$ by:

$$n_0 u + \sqrt{n_0} W_1(u) \quad \text{and} \quad n_0 v + \sqrt{n_0} W_2(v) \ ,$$

where W_1 and W_2 are two independent standard Brownian motions. The CM becomes:

$$\left\{ \begin{array}{rcl} U(t) & \simeq & 1 - p_0 + \displaystyle\int_0^t \nu U(s) \left(1 - \frac{n_0}{n_f}(U(s)+V(s) \right) \mathrm{d}s \\ & & + \dfrac{1}{\sqrt{n_0}} W_1 \left(\displaystyle\int_0^t U(s) \left(1 - \frac{n_0}{n_f}(U(s)+V(s) \right) \mathrm{d}s \right) \\ V(t) & \simeq & p_0 + \displaystyle\int_0^t \mu V(s) \left(1 - \frac{n_0}{n_f}(U(s)+V(s) \right) \mathrm{d}s \\ & & + \dfrac{1}{\sqrt{n_0}} W_2 \left(\displaystyle\int_0^t V(s) \left(1 - \frac{n_0}{n_f}(U(s)+V(s) \right) \mathrm{d}s \right) \end{array} \right.$$

Thus the stochastic process (U, V) is a bidimensional diffusion, solution to the following stochastic differential equations.

$$\left\{ \begin{array}{rcl} \mathrm{d}U(t) & = & \nu U(t) \left(1 - \dfrac{n_0}{n_f}(U(t)+V(t) \right) \mathrm{d}t \\ & & + \dfrac{1}{\sqrt{n_0}} \sqrt{U(t) \left(1 - \dfrac{n_0}{n_f}(U(t)+V(t)) \right)} \mathrm{d}W_1(t) \\ \mathrm{d}V(t) & = & \mu V(t) \left(1 - \dfrac{n_0}{n_f}(U(t)+V(t) \right) \mathrm{d}t \\ & & + \dfrac{1}{\sqrt{n_0}} \sqrt{\mu V(t) \left(1 - \dfrac{n_0}{n_f}(U(t)+V(t)) \right)} \mathrm{d}W_2(t) \end{array} \right.$$

This is the mesoscopic scale model. Of course, by neglecting the diffusion term in $\frac{1}{\sqrt{n_0}}$, one gets the deterministic (or macroscopic) Volterra model of section 3.

How should one choose among the three modelling scales? Figure 5 shows a simulation of 10 trajectories of the CM, for $n_0 = 5 \times 10^3$ and $n_f = 5 \times 10^4$, the relative fitness of mutants being 1.6. On the same graphics, the trajectory of the deterministic Volterra model has been superposed. As can be observed, the dispersion of random trajectories around the deterministic ones is small. For the more realistic values of $n_0 = 5 \times 10^6$ and $n_f = 5 \times 10^8$, the random trajectories would be almost undistinguishable from the deterministic ones. However, this is only true for a relatively large value of p_0. If only a few mutant cells were initially present, stochastic models would certainly give more reliable results than the deterministic one.

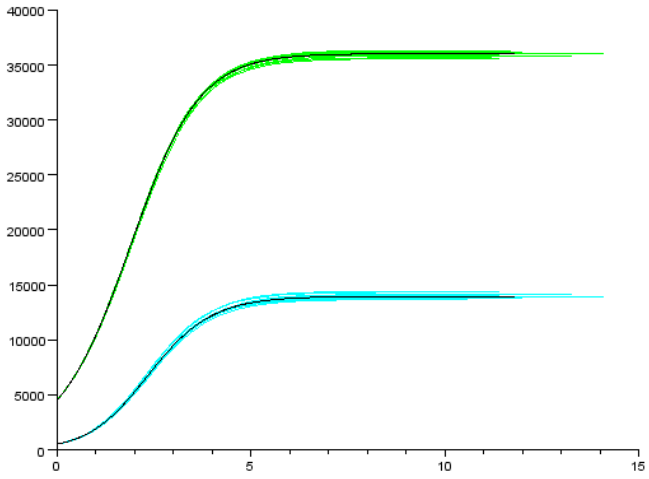

Figure 5: Ten trajectories of the stochastic competition model for the numbers of normal (green) and mutant cells (blue). The trajectory of the corresponding deterministic model is plotted in black. The initial and final numbers of cells are $n_0 = 5 \times 10^3$ and $n_f = 5 \times 10^4$. The initial proportion of mutant cells is 0.1. The IDR's of normal and mutant cells are 1 and 1.6.

Once again, our intention in proposing that very simple model, was to illustrate the possible treatments rather than imposing it as the most realistic. More sophisticated logistic-type stochastic models were proposed long ago: see Novozhilov [59] and section 6.8 p. 242 of Allen [2]. Logistic growth processes have been the object of several mathematical studies, in particular by Tan & Piantadosi [76], or more recently by Lambert [47]. Refinements include for instance spatial random dispersal [78] or local regulation [19].

5 Mutation models

The history of mutation models is as long as that of stochastic growth processes, since estimates of mutation probabilities were already present in Yule's founding paper [84].

However, it really started with Luria and Delbrück experiments [51]. Interestingly enough, they used the same *Escherichia Coli B* as the LE. Their idea was to introduce a virus that killed most bacteria, except those having acquired resistance by mutation. This allowed them to count mutant bacteria. Repeating the experiment, they estimated the distribution of the (random) number of mutants. The data seemed to indicate a distribution with a much heavier tail than expected: most counts had very few mutant cells, but in a sizeable proportion of the counts, they found a relative high number of mutants. In view of our discussion in section 2, that feature is crucial for the survival of mutations through daily sampling in the LE. Indeed we have shown that mutations have good chances to survive daily dilution, only when they are carried by sufficiently many cells. Luria and Delbrück used a very simple deterministic model, assuming that cell counts doubled at each multiple of a given fixed period. Nevertheless, that model allowed them to derive an asymptotic distribution of the number of mutants, that showed indeed a heavy tail behavior. Other models were later consider, in particular by Lea & Coulson [49], Harris [31], Bartlett [5, 6], Kendall [40, 41], and Armitage [3]. Mandelbrot [54] proposed a general proof for the convergence to the Luria-Delbrück distribution. More recently, the importance of that distribution for the treatment of mutation experimental data has stemmed further researches such as [69, 52, 38]. The Luria-Delbrück distribution is known only through its generating function and no close form exists for its probabilities. However an efficient recursive formula permits to calculate them explicitly. Pakes [61] gives an easy derivation of that formula. An example of generalization is given by Dewanji *et al.* [16]. Presently, the most active author on the subject is Zheng; he wrote a useful mathematical review in [85]. Historical reviews include Sarkar [69] and Zheng [88].

The objective of this section is to present Bartlett's derivation of the Luria-Delbrück distribution under the following modelling assumptions (see [6], [3], p. 37, [9], section 4.31 p. 124, and more recently Zheng [87]).

1. At time 0 a homogeneous population of n_0 "normal" cells is given (n_0 is large);
2. normal cells divide at a constant rate ν;
3. when a division occurs:
   (a) with probability $1-p$ two normal cells are produced,
   (b) with probability p one normal and one mutant cells are produced,

   (the mutation probability p is small);
4. mutant cells divide at a constant rate μ;
5. any other mutation than normal to mutant is excluded;
6. all random events (divisions and mutations) are mutually independent.

Let G denote the bivariate generating function for the numbers of normal and mutant cells, starting with a single normal cell at time 0.

$$G(z,y,t) = \sum_{n=0}^{+\infty}\sum_{m=0}^{+\infty} z^n y^m \, \mathbb{P}[N(t)=n\,,\ M(t)=m \,|\, N(0)=1\,,\ M(0)=0]\,.$$

If the population starts with a single mutant cell, only mutant cells can be produced later, and there is no need to consider a bivariate generating function. We shall denote by H the generating fonction of the number of mutant cells, starting with a single mutant cell at $t=0$.

$$H(y,t) = \sum_{m=0}^{+\infty} y^m \, \mathbb{P}[M(t)=m \,|\, M(0)=1]\,.$$

Its calculation from the Bellman-Harris equation was already exposed in section 4.

$$H(y,t) = \frac{y\mathrm{e}^{-\mu t}}{1-y+y\mathrm{e}^{-\mu t}}\,.$$

The generating functions G and H are related through another Bellman-Harris equation.

$$\begin{aligned} G(z,y,t) &= \Big(\int_0^t \Big((1-p)G^2(z,y,t-s) \\ &\quad +p\,G(z,y,t-s)H(y,t-s)\Big)\nu \mathrm{e}^{-\nu s}\,\mathrm{d}s\Big) + z\mathrm{e}^{-\nu t}\ . \end{aligned} \tag{BH2}$$

The justification of (BH2) is quite similar to that of (BH), given in section 4. The initial normal cell divides at a time which is exponentially distributed with parameter ν. At time t, it may not have divided yet (with probability $\mathrm{e}^{-\nu t}$), and the generating function is still z. If it divides at some time s between 0 and t (in $[s, s+\mathrm{d}s]$ with probability $\nu \mathrm{e}^{-\nu s}\mathrm{d}s$), it turns either into 2 normal cells with probability $1-p$, or into a normal and a mutant cell with probability p. The two new cells start independent lineages of their own, accounted for by $G^2(z,y,t-s)$ if two normal cells are produced, and by $G(z,y,t-s)H(y,t-s)$ if one normal and one mutant are produced. As in section 4, (BH2) can be transformed into an ordinary differential equation.

$$\frac{\partial G(z,y,t)}{\partial t} = \nu((1-p)G^2(z,y,t) + p\,G(z,y,t)H(y,t) - G(z,y,t))\ . \tag{DE2}$$

This is a linear first order equation in the inverse $G^{-1}(z,y,t)$:

$$\frac{\partial G^{-1}(z,y,t)}{\partial t} = \nu(p-1) + \nu G^{-1}(z,y,t)(1-p\,H(y,t))\ .$$

Replacing $H(y,t)$ by its explicit expression, the general solution to the homogeneous equation is found to be proportional to:

$$\mathrm{e}^{\nu t}(1-y+y\mathrm{e}^{-\mu t})^{p\frac{\nu}{\mu}}\ .$$

Using the initial value $G(z,y,0)=z$,

$$G^{-1}(z,y,t) = \mathrm{e}^{\nu t}(1-y+y\mathrm{e}^{-\mu t})^{p\frac{\nu}{\mu}}\left(\frac{1}{z} + \nu(p-1)\int_0^t \mathrm{e}^{-\nu s}(1-y+y\mathrm{e}^{-\mu s})^{-p\frac{\nu}{\mu}}\,\mathrm{d}s\right)\ .$$

When $\mu=\nu$, the primitive can be explicitly calculated. This is the only case considered by Bartlett.

$$\int_0^t \mathrm{e}^{-\nu s}(1-y+y\mathrm{e}^{-\nu s})^{-p}\,\mathrm{d}s = \frac{1}{(1-p)\nu y}\left(1-(1-y+y\mathrm{e}^{-\nu t})^{1-p}\right)\ .$$

Hence Bartlett's explicit expression:

$$G^{-1}(z,y,t) = \frac{(1-y+y\mathrm{e}^{-\nu t})^p}{\mathrm{e}^{-\nu t}z} + \frac{(1-y+y\mathrm{e}^{-\nu t})}{y\mathrm{e}^{-\nu t}} - \frac{(1-y+y\mathrm{e}^{-\nu t})^p}{y\mathrm{e}^{-\nu t}}\ .$$

This gives the generating function of both counts, starting with one normal cell. The generating function of mutant cells alone is obtained by setting $z=1$ in the expression above. The generating function of mutant cells, starting with n_0 normal cells is the $n_0{}^{\text{th}}$ power:

$$G^{n_0}(1,y,t) = \left(\frac{(1-y+y\mathrm{e}^{-\nu t})^p}{\mathrm{e}^{-\nu t}} + \frac{(1-y+y\mathrm{e}^{-\nu t})}{y\mathrm{e}^{-\nu t}} - \frac{(1-y+y\mathrm{e}^{-\nu t})^p}{y\mathrm{e}^{-\nu t}}\right)^{-n_0}\ .$$

We want an asymptotic value for this expression, as p tends to 0, n_0 and t to $+\infty$. Let us replace the two terms in $(1-y+y\mathrm{e}^{-\nu t})^p$ by their order 1 expansion:

$$(1-y+y\mathrm{e}^{-\nu t})^p = 1 + p\log(1-y+y\mathrm{e}^{-\nu t}) + o(p)\ .$$

Remember also that t is large, so that:

$$\log(1-y+y\mathrm{e}^{-\nu t}) = \log(1-y) + O(\mathrm{e}^{-\nu t})$$

One gets:

$$G^{n_0}(1,y,t) = \left(1 + p\frac{y-1}{y\mathrm{e}^{-\nu t}}\log(1-y) + o(p) + pO(\mathrm{e}^{-\nu t})\right)^{-n_0}$$

The non-trivial limit is obtained as as p tends to 0, n_0 and t to $+\infty$, in such a way that

$$\lim n_0\, p\, \mathrm{e}^{\nu t} = \alpha\ ,$$

where α is a positive real number. Then:

$$\lim G^{n_0}(1,y,t) = \exp\left(\alpha\frac{1-y}{y}\log(1-y)\right) = (1-y)^{\alpha\frac{1-y}{y}}\ .$$

The *Luria-Delbrück* distribution with parameter α is defined as the probability distribution on integers with generating function:

$$g_\alpha(y) = (1-y)^{\alpha\frac{1-y}{y}}\ .$$

The function g_α has no left derivative at $y=1$: the Luria-Delbrück distribution has no moment of any order. Denote by p_n the corresponding probabilities:

$$g_\alpha(y) = \sum_{m=0}^{+\infty} p_m\, y^m\ .$$

The first value is obtained for $y=0$: $p_0 = \mathrm{e}^{-\alpha}$. There is no explicit expression for p_m as a function of m. An equivalent as m tends to infinity is $p_m \sim \frac{\alpha}{m^2}$. The exact values can be numerically computed through the recursive formula:

$$p_m = \frac{\alpha}{m}\sum_{i=0}^{m-1}\frac{p_i}{n-i+1}\ .$$

See Ma *et al.* [52], Pakes [61], and Kemp [38] for simple derivations of the main results on the p_n's. As an example, the table below gives some values for the probability than more than 50 mutant cells remain, for different values of α.

α	1	2	3	4	5	6	7	8	9	10
$\mathbb{P}[X>50]$	0.021	0.045	0.072	0.102	0.136	0.172	0.213	0.256	0.302	0.349

As expected from a heavy tail distribution, there are quite reasonable chances to get sizeable amounts of mutant cells.

How about the Lenski experiment? As we have seen, the initial number of normal cells in any of the twelve 10 mL vessels is 5×10^6. Let us take $n_0 = 6\times10^7$. For the mutation probability, Philippe *et al.* [64] give $p = 5\times10^{-10}$ per base pair. This is sensibly lower than the value given by Kendall [41], who simply says that the mutation probability is lesser than 10^{-7}. As we have seen in section 4, $\mathrm{e}^{\nu t}$ is the expected number of cells stemming from one initial cell. In the LE, the daily increase is 100-fold. So we shall retain $\mathrm{e}^{\nu t} = 10^2$. The parameter for the Luria-Delbrück distribution is:

$$\alpha = n_0\, p\, \mathrm{e}^{\nu t} = 3\ .$$

For the Luria-Delbrück distribution with parameter 3, the probability to get no mutant cells is $p_0 = 0.05$. The probability to get more than 100 mutant cells is 0.034.

The asymptotics of the number of mutants in the general case $\mu \neq \nu$ is discussed by Jaeger and Sarkar [34]. To the best of our knowledge, no derivation from Bartlett's model with $\mu \neq \nu$ has been made. However, we do not think it would change by much the conclusions. If $\mu > \nu$ (beneficial mutation), mutant cells will multiply faster on average, (thus be more numerous) than predicted by the Luria-Delbrück distribution. In other words, the Luria-Delbrück distribution function is an upper bound for the distribution function of the number of mutant cells in the general case. The heavy tail property can only be reinforced.

The Bartlett model that was exposed here is not unique. Luria-Delbrück distributions can be obtained through other models, including the earlier deterministic growth model of Lea and Coulson [49] (see Zheng's review [85]). It has been extended to other types of non-Markovian growth models by Dewandji *et al.* [16].

6 Parameter estimation

Gause [26], in the introduction to his section "On the mechanism of competition in yeast cells", cites early 30's publications while making quite clear statements on the relation between experimental data and mathematical models. We could not say it better.

> No mathematical theories can be accepted by biologists without a most careful experimental verification. We can but agree with the following remarks made in Nature (H. T. H. P. '31) concerning the mathematical theory of the struggle for existence developed by Vito Volterra: "This work is connected with Prof. Volterra's researches on integro-differential equations and their applications to mechanics. In view of the simplifying hypothesis adopted, the results are not likely to be accepted by biologists until they have been confirmed experimentally, but this work has as yet scarcely begun." First of all, very reasonable doubts may arise whether the equations of the struggle for existence given in the preceding chapter express the essence of the processes of competition, or whether they are merely empirical expressions. everybody remembers the attempt to study from a purely formalistic viewpoint the phenomena of heredity by calculating the likeness between ancestors and descendants. This method did not give the means of penetrating into the mechanism of the corresponding processes and was consequently entirely abandoned. In order to dissipate these doubts and to show that the above-given equations actually express the mechanism of competition, we shall now turn to an experimental analysis of a comparatively simple case. It has been possible to measure directly the factors regulating the struggle for existence in this case, and thus to verify some of the mathematical theories.
>
> Generally speaking, biologists usually have to deal with empirical equations. The essence of such equations is admirably expressed in the following words of Raymond Pearl ('30): "The worker in practically any branch of science is more or less frequently confronted with this sort of problem: he has a series of observations in which there is clear evidence of a certain orderliness, on the one hand, and evident fluctuations from that order, on the other hand. What he obviously wishes to do... is to emphasize the orderliness and minimize the fluctuations about it... He would like an expression, exact if possible, or, failing that, approximate, of the law if there be one. This means a mathematical expression of the functional relation between the variables...
>
> "It should be made clear at the start that there is, unfortunately, no methods known to mathematics which will tell anyone in advance of the trial what is either the correct or even the best mathematical function with which to graduate a particular set of data. The choice of the proper mathematical

> function is essentially, at its very best, only a combination of good judgment and good luck. In this realm, as in every other, good judgment depends in the main only upon extensive experience. What we call good luck in this sort of connection has also about the same basis. The experienced person in this branch of applied mathematics knows at a glance what general class of mathematical expression will take a course, when plotted, on the whole like that followed by the observations. He furthermore knows that by putting as many constants into his equation as there are observations in the data he can make his curve hit all the observed points exactly, but in so doing will have defeated the very purpose with which he started, which was to emphasize the law (if any) and minimize the fluctuations, because actually if he does what has been described he emphasizes the fluctuations and probably loses completely any chance of discovering a law.
>
> "Of mathematical functions involving a small number of constants there are but relatively few... In short, we live in a world which appears to be organized in accordance with relatively few and relatively simple mathematical functions. Which of these one will choose in starting off to fit empirically a group of observations depends fundamentally, as has been said, only on good judgment and experience. There is no higher guide" (pp. 407-408).

The mathematical models that have been presented so far have no predictive value, until they have been confronted to experimental data. The parameters should be estimated, the data have to be adjusted, and the goodness-of-fit must be tested. The adjustment of population models to bacteria growth experiments has been the object of countless publications, of which we have retained a few among the most recent.

Miao *et al.* [56] give an interesting review of parameter estimation methods for deterministic models. The general methods for adjusting models are presented by Jaqaman and Danuser in [35]. Gutenkunst *et al.* [29] argue on sloppy parameter sensitivity in systems biology models. All models have at least in common growth rates and mutation probabilities, as parameters to be estimated from the data. Although most references focus on mutation rates, the estimation of growth rates was recently studied by Maruvka *et al.* [55]. Mutation rates can be estimated by adjusting observed number of mutants in bacterial growth experiments. The method is usually referred to as *fluctuation analysis*. It has been presented in many references, including Koch [44], Sarkar *et al.* [70], Stewart *et al.* [75], Jones *et al.* [37]. Dedicated softwares were made available by Zheng [86] and Hall *et al.* [30]. Several refinements in the use of the Luria-Delbrück probabilities have been proposed, in particular by Kepler & Oprea [42, 60].

The estimation of mutation rates was studied as early as 1974 by Crump & Hoel [15]. The quality of mutation rates estimates is discussed by Stewart [74]. Different estimation methods are presented by Foster [23]. Improvements have been proposed by Koziol [46], and by Gerrish [27]. Grant *et al.* [28] made a case study on *Mycobacterium tubercolosis*.

7 Conclusion

The few classics of mathematical modelling that have been reviewed here are hoped to be of some potential use in the description of different aspects of the Lenski experiment. As already said, none of them has any scientific validity as long as they have not been confronted to real data. However, the models share some common qualitative features that are summarized here for comparison with actual observations. Instead of categorizing the models according to mathematical coherence as we did before, we will summarize their predictions on four different questions about the LE.

1. *Day 0: Which number of cells can carry a given mutation that was not present the*

previous day?
All existing models use asymptotics as the initial number of cells is large and the probability of mutations is low. The random number of new mutants follows a heavy-tailed distribution of which the Luria-Delbrück distribution is a prototype. This means that with a reasonable probability, sizeable amounts of mutants can be present at the end of any given day.

2. *Night 0: What are the chances that a mutation having appeared on day 0 disappears through dilution?*
Since the daily sampling eliminates each day 99% of the cells, mutants present in small quantities have high chances to disappear. In particular, non-beneficial mutations do not survive. Beneficial mutations carried by enough cells (at least a few tens), have good chances to be represented the next days.

3. *Day 1: If two different strains of cells are present at the beginning of a given day, what will their proportion be at the end of the day?*
If a beneficial mutation is carried by a certain proportion of the population, even a weak one, that proportion will gradually increase along the day with the multiplication of cells. So with high probability, the initial proportion will be larger on the next day.

4. *Day N: How many days will it take before a beneficial mutation is carried by all cells in the population?*
The initial proportion of day $N+1$ is an increasing function of that of the previous day. The better the fitness of mutants, the steeper that function. When the initial proportion of mutants is already strong, there is a high probability that the less fit cells will be wiped out by the next daily sampling. The invasion of the full population will take from a few days to a month.

References

[1] D.J. Aldous. Stochastic models add descriptive statistics for phylogenetic trees, from Yule to today. *Statist. Sci.*, 16(1):23–34, 2001.

[2] L.J.S. Allen. *Stochastic processes with applications to biology.* Pearson – Prentice Hall, New-Jersey, 2003.

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

[4] J.E. Barrick, D.S. Yu, S.H. Yoon, H. Jeong, T.K. Oh, D. Schneider, R.E. Lenski, and J.F. Kim. Genome evolution and adaptation in a long-term experiment with *Escherichia Coli*. *Evolution*, 461(29):1243–1249, 2009.

[5] M.S. Bartlett. Some evolutionary stochastic processes. *J. R. Statist. Soc. B*, 11(2):211–229, 1949.

[6] M.S. Bartlett. The dual recurrence relation for multiplicative processes. *Math. Proc. Camb. Phil. Soc.*, 47(4):821–825, 1951.

[7] M.S. Bartlett. Processus stochastiques ponctuels. *Ann. IHP*, 14(1):35–60, 1954.

[8] M.S. Bartlett. *Stochastic population models in ecology and epidemiology.* Methuen, London, 1960.

[9] M.S. Bartlett. *An introduction to stochastic processes, with special reference to methods and applications.* Cambridge University Press, 1966.

[10] R. Bellman and T. Harris. On age-dependent binary branching processes. *Ann. Math.*, 55(2):280–295, 1952.

[11] A.T. Bharucha-Reid. *Elements of the theory of Markov processes and their Applications*. McGraw-Hill, London, 1960.

[12] P.R.A. Campos and L.M. Wahl. The effects of population bottlenecks on clonal interference, and the adaptation effective population size. *Evolution*, 63(4):950–958, 2009.

[13] P.R.A. Campos and L.M. Wahl. The adaptation rate of asexuals: deleterious mutations, clonal interference and population bottlenecks. *Evolution*, 64(7):1773–1783, 2010.

[14] Consortium Scilab. *Scilab: Le logiciel libre de calcul numérique*. Consortium Scilab, Digiteo, Paris, France, 2011.

[15] K.S. Crump and D.G. Hoel. Mathematical models for estimating mutation rates in cell populations. *Biometrika*, 61(2):237–252, 1974.

[16] A. Dewanji, E.G. Luebeck, and S.H. Moolgavkar. A generalized Luria-Delbrück model. *Math. Biosci.*, 197:140–152, 2005.

[17] R. Durrett. *Probability models for DNA sequence evolution*. Springer, New-York, 2008.

[18] E. Çinlar. *Introduction to stochastic processes*. Prentice Hall, New York, 1975.

[19] A.M. Etheridge. Survival and extinction in a locally regulated population. *Ann. Appl. Probab.*, 14(1):188–214, 2004.

[20] S.N. Ethier and T.G. Kurtz. *Markov processes: characterization and convergence*. Wiley series in Probability and Statistics. Wiley, New-York, 2005.

[21] W.J. Ewens. *Mathematical population genetics: theoretical introduction*. Springer, New York, 2004.

[22] W. Feller. *An introduction to probability theory and its applications*, volume II. Wiley, London, 1971.

[23] P.L. Foster. Methods for determining spontaneous mutation rates. *Methods Enzymol.*, 409:195–213, 2006.

[24] W.H. Furry. On fluctuation phenomena in the passage of high energy electrons through lead. *Phys. Rev.*, 52:569–581, 1937.

[25] C.W. Gardiner. *Handbook of stochastic methods for physics, chemistry and the natural sciences*. Springer, Berlin, 2004.

[26] G.F. Gause. *The struggle for existence*. Williams & Wilkins, Baltimore, 1934.

[27] P.J. Gerrish. A simple formula for obtaining markedly improved mutation rates estimates. *Genetics*, 180:1773–1778, 2008.

[28] A. Grant, C. Arnold, N. Thorne, S. Gharbia, and A. Underwood. Mathematical modelling of *Mycobacterium tuberculosis* VNTR loci estimates a very slow mutation rate for the repeats. *J. Mol. Evol.*, 66:565–574, 2008.

[29] R.N. Gutenkunst, J.J. Waterfall, F.P. Casey, K.S. Brown, C.R. Myers, and J.P. Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.*, 3:1871–1878, 2007.

[30] B.M. Hall, C. Ma, P. Liang, and K.K. Singh. Fluctuation Analysis CalculatOR (FALCOR): a web tool for the determination of mutation rate using Luria-Delbrück fluctuation analysis. *Bioinformatics*, 25(12):1564–1565, 2009.

[31] T.E. Harris. Some mathematical models for branching processes. *2d Berk. Symp.*, pages 305–328, 1951.

[32] J. Hermisson and P. Pfaffelhuber. The pattern of genetic hitchhiking under recurrent mutation. *Elec. J. Probab.*, 13(68):2069–2106, 2008.

[33] S.B. Hsu, S. Hubbell, and P. Waltman. A mathematical theory for single-nutrient competition in continuous cultures of micro-organisms. *SIAM J. Applied. Math.*, 32(2):366–383, 1977.

[34] G. Jaeger and S. Sarkar. On the distribution of bacterial mutants: the effects of differential fitness of mutants and non-mutants. *Genetica*, 96:217–223, 1995.

[35] K. Jaqaman and G. Danuser. Linking data to models: data regression. *Nat. Rev. Mol. Cell Biol.*, 7:1–10, Nov 2006.

[36] K.A. Johnson and R.S. Goody. The original Michaelis constant: translation of the 1913 Michaelis-Menten paper. *Biochemistry*, 50(4):8264–8269, 2011.

[37] M.E. Jones, S.M. Thomas, and A. Rogers. Luria-Delbrück fluctuation experiments: design and analysis. *Genetics*, 136:1209–1216, 1994.

[38] A.W. Kemp. Comments on the Luria-Delbrück distribution. *J. Appl. Probab.*, 31(3):822–828, 1994.

[39] D.G. Kendall. On the role of variable generation time in the development of a stochastic birth process. *Biometrika*, 35(3-4):316–330, 1948.

[40] D.G. Kendall. Stochastic processes and population growth. *J. R. Statist. Soc. B*, 11(2):230–282, 1949.

[41] D.G. Kendall. Les processus stochastiques de croissance en biologie. *Ann. IHP*, 13(1):43–108, 1952.

[42] T.B. Kepler and M. Oprea. Improved inference of mutation rates: I: an integral representation for the Luria-Delbrück distribution. *Theor. Pop. Biol.*, 59(1):41–48, 2001.

[43] A.I. Khan, D.M. Dinh, D. Schneider, R.E. Lenski, and T.F. Cooper. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science*, 332(6034):1193–1196, 2011.

[44] A.L. Koch. Mutation and growth rates from Luria-Delbrück fluctuation tests. *Mutat. Res.*, 95(2-3):129–143, 1982.

[45] M. Kot. *Elements of mathematical ecology.* Cambridge University Press, 2001.

[46] J.A. Koziol. A note on efficient estimation of mutation rates using Luria-Delbrück fluctuation analysis. *Mutat. Res.*, 249:275–280, 1991.

[47] A. Lambert. The branching process with logistic growth. *Ann. Appl. Probab.*, 15(2):1506–1535, 2005.

[48] G.I. Lang, D. Botstein, and M.M. Desai. Genetic variation and the fate of beneficial mutations in asexual populations. *Genetics*, 188(3):647–661, 2011.

[49] D.E. Lea and C.A. Coulson. The distribution of the number of mutants in bacterial populations. *J. Genetics*, 49:264–285, 1949.

[50] A.J. Lotka. *Elements of physical biology.* Williams & Wilkins, Baltimore, 1925.

[51] D.E. Luria and M. Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28:491–511, 1943.

[52] W.T. Ma, G.v.H. Sandri, and S. Sarkar. Analysis of the Luria-Delbrück distribution using discrete convolution powers. *J. Appl. Probab.*, 29(2):255–267, 1992.

[53] T.R. Malthus. *An essay on the principle of population.* Johnson, London, 1798.

[54] B. Mandelbrot. A population birth-and-mutation process, I: explicit distributions for the number of mutants in an old culture of bacteria. *J. Appl. Probab.*, 11(3):437–444, 1974.

[55] Y.E. Maruvka, N.M. Shnerb, Y. Bar-Yam, and J. Wakeley. Recovering population parameters from a single gene genealogy: an unbiased estimator of the growth rate. *Mol. Biol. Evol.*, 28(5):1617–1631, 2011.

[56] H. Miao, X. Xia, A.S. Perelson, and H. Wu. On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM Rev.*, 53(1):3–39, 2011.

[57] P.A.P. Moran. *The statistical processes of evolutionary theory.* Clarendon Press, Oxford, 1962.

[58] J.D. Murray. *Mathematical biology.* Springer, New-York, 1989.

[59] A.S. Novozhilov, G.P. Karev, and E.V. Koonin. Biological applications of the theory of birth-and-death processes. *Briefings Bioinfo.*, 7(1):70–85, 2005.

[60] M. Oprea and T.B. Kepler. Improved inference of mutation rates II: generalization of the Luria-Delbrück distribution for realistic cell-cycle time distributions. *Theor. Pop. Biol.*, 59(1):49–59, 2001.

[61] A.G. Pakes. Remarks on the Luria-Delbrück distribution. *J. Appl. Probab.*, 30(4):991–994, 1993.

[62] A.G. Pakes. Biological applications of branching processes. In D.N. Shanbhag and C.R. Rao, editors, *Stochastic Processes: Modelling and Simulation*, volume 21 of *Handbook of Statistics*, pages 693 – 773. Elsevier, 2003.

[63] B. Papp, A.A. Notebaart, and C. Pál. Systems biology approaches for predicting genomic evolution. *Nature Rev. Genetics*, 12:591–602, 2011.

[64] N. Philippe, E. Crozat, R.E. Lenski, and D. Schneider. Evolution of global regulatory network during a long-term experiment with *Escherichia Coli*. *BioEssays*, 29(9):846–860, 2007.

[65] E.O. Powell. Criteria for the growth of contaminants and mutants in continuous culture. *J. Gen. Microbiol.*, 18:259–268, 1958.

[66] A. Quetelet. *Essai de physique sociale, tome premier.* Hauman, Bruxelles, 1836.

[67] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[68] S. Ross. *Introduction to probability models.* Academic Press, 10^{th} edition, 2010.

[69] S. Sarkar. Haldane's solution of the Luria-Delbrück distribution. *Genetics*, 127:257–261, 1991.

[70] S. Sarkar, W.T. Ma, and G.v.H. Sandri. On fluctuation analysis: a new, simple and efficient method for computing the expected number of mutants. *Genetica*, 85:173–179, 1992.

[71] D. Schneider, E. Duperchy, E. Coursange, R.E. Lenski, and M. Blot. Long-term experimental evolution in *Escherichia Coli* IX. characterization of insertion sequence-mediated mutations and rearrangements. *Genetics*, 156(2):477–488, 2000.

[72] D. Schneider, E. Duperchy, J. Dupeyrot, E. Coursange, R.E. Lenski, and M. Blot. Genomic comparisons among *Escherichia Coli* strains b, k12, and O157:H7 using IS elements as molecular markers. *BMC Microbiol.*, 2:18, 2002.

[73] M.T. Stanek, T.F. Cooper, and R.E. Lenski. Identification and dynamics of a beneficial mutation in a long term evolution experiment with *Escherichia Coli*. *BMC Evol. Biol.*, 9:302, 2009.

[74] F.M. Stewart. Fluctuation tests: how reliable are the estimates of mutation rates? *Genetics*, 137:1139–1146, 1994.

[75] F.M. Stewart, D.M. Gordon, and B.R. Levin. Fluctuation analysis: the probability distribution of the number of mutants under different conditions. *Genetics*, 124:175–185, 1990.

[76] W.Y. Tan and S. Piantadosi. On stochastic growth processes with application to stochastic logistic growth. *Statist. Sinica*, 1:527–540, 1991.

[77] H.C. Tuckwell. *Elementary applications of probability theory.* Chapman & Hall/CRC, Boca Raton, FL, 1995.

[78] H.C. Tuckwell and J.A. Koziol. Logistic population growth under random dispersal. *Bull. Math. Biology*, 49(4):495–506, 1987.

[79] P.F. Verhulst. Notice sur la loi que la population suit dans son accroissement. In J.G. Garnier and A. Quetelet, editors, *Correspondance mathématique et physique*, volume 10, pages 113–121. Société Belge de Librairie, Bruxelles, 1838.

[80] V. Volterra. Fluctuations in the abundance of a species considered mathematically. *Nature*, 118:558–560, 1926.

[81] V. Volterra. *Leçons sur la théorie mathématique de la lutte pour la vie.* Gauthier-Villars, Paris, 1931.

[82] L.M. Wahl and P.J. Gerrish. The probability that beneficial mutations are lost in populations with periodic bottlenecks. *Evolution*, 55:2606–2610, 2001.

[83] D.J. Wilkinson. *Stochastic modelling for systems biology.* Chapman & Hall/CRC, Boca Raton, FL, 2006.

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

[85] Q. Zheng. Progress of a half century in the study of the Luria-Delbrück distribution. *Math. Biosc.*, 162:1–32, 1999.

[86] Q. Zheng. Statistical and algorithmic methods for fluctuation analysis with SALVADOR as an implementation. *Math. Biosci.*, 176:237–252, 2002.

[87] Q. Zheng. On Bartlett's formulation of the Luria-Delbrück mutation model. *Math. Biosci.*, 215:48–54, 2008.

[88] Q. Zheng. The Luria-Delbrück distribution: early statistical thinking about evolution. *Chance*, 23(2):15–18, 2010.