Estimation of Mutation Rates from Fluctuation Experiments via Probability Generating Functions[*]

Stephen Montgomery-Smith[†], Anh Le[‡], George Smith[§]
Sidney Billstein[‡], Hesam Oveys[†], Dylan Pisechko[‡], Austin Yates[‡]

Abstract

This paper calculates probability distributions modeling the Luria-Delbrück experiment. We show that by thinking purely in terms of generating functions, and using a 'backwards in time' paradigm, that formulas describing various situations can be easily obtained. This includes a generating function for Haldane's probability distribution due to Ycart. We apply our formulas to both simulated and real data created by looking at yeast cells acquiring an immunization to the antibiotic canavanine.

This paper is somewhat incomplete, having been last significantly modified in March 29, 2014. However the first author feels that this paper has some worthwhile ideas, and so is going to make this paper publicly available.

1 Introduction

The famous experiment of Luria and Delbrück [5], [12] determined whether mutations in bacteria arise via Darwinian evolution or some kind of Lamarckian process. The experiment consisted of taking small samples of bacteria, and then allowing them to replicate up to a known large number of cells on many plates, and then looking at the distribution of number of cells on each plate that had acquired an immunity to an antibiotic. The mutation was carefully chosen to be a forward mutation, that is, the probability of the mutation taking place was very much larger than the probability of the mutation reversing itself.

If the immunity was acquired by some kind of Lamarckian process, then one would naturally expect the distribution of surviving cells in the plates to follow a Poisson distribution. In particular, the variance would have a value very close to the mean, and the probability of a plate having a so called 'jackpot,' that is, a much larger number of surviving cells than the mean, would be vanishingly small.

[*] Research supported by NSF DMS Grant 0928053, PRISM: Mathematics in Life Sciences

[†] Department of Mathematics, University of Missouri, Columbia MO 65211

[‡] Student in Mathematics of Life Sciences Program, University of Missouri, Columbia MO 65211

[§] Department of Biology, University of Missouri, Columbia MO 65211

However, Luria and Delbrück argued that if the immunity was acquired by Darwinian evolution, then the variance of the number of surviving cells would be much larger than the mean, and furthermore, the probability of any plate having a 'jackpot' would be large enough that it would be observed reasonably often.

Their experiment conclusively validated the latter assumption. Soon after this, scientists sought after a formula for the distribution of surviving cells in each plate. According to [11], the biologist J.B.S. Haldane produced a formula for the distribution in 1946. But his work was never widely published at that time, and in 1949 Lea and Coulson [4] produced a different distribution, which seems to be the basis of most research these days.

A great advantage of their distribution is that it can be given by a simple generating function, whereas the calculations described by Haldane seem to be very time consuming, and requires enumerating combinatorial structures. For example, the paper [6] gives an algorithm whereby the Lea-Coulson distribution can be calculated quickly. The survey paper [9] gives many ways in which one can calculate the *fluctuation rate*, that is, the probability of a single offspring cell acquiring the mutation. And the paper [1] gives a Bayesian approach to estimating the fluctuation rate.

The primary goal of this paper is to promote an approach which uses generating functions from the very beginning. The other change we propose to the approaches taken in other papers is to think 'backwards in time' rather than 'forwards in time,' that is, instead of considering how mutants might have developed from initial conditions, look at the final situation and reason out where the mutations must have come from. We will illustrate this approach under several conditions. In particular we will be able to obtain a generating function for the Haldane distribution which allows for rapid calculations.

2 The experimental setup to be modeled

We assume that in each plate we start from a small number of cells, none of which have the mutation. Then the cells replicate repeatedly. When each cell replicates, one of the offspring can acquire the mutation with a probability μ, which is extremely small, effectively infinitesimal. We continue the replication process until we have obtained a population of n cells, where n is effectively infinite. We assume that $m = \mu n$ is of order 1.

We shall use the following terminology (which is common in the literature). We shall call a cell that acquires the mutation, but not from its parent, a *mutation*. A cell that has the mutant gene will be called a *mutant*. We will assume that the probability of a mutant have offspring that are not mutants is zero. This can either come from supposing that the genetic change is much more likely to happen than to be reversed (that is, it is a forward mutation), or by realizing that the population of mutants is so much smaller than the number of non-mutants, that even if the probability of a mutant having non-mutant offspring were of the order μ, we wouldn't observe this anyway.) Thus all mutations are mutants, but not vice versa.

The quantity m is the fundamental parameter that describes the probability distribution of the random variable R which is the number of cells that have acquired the mutation. All

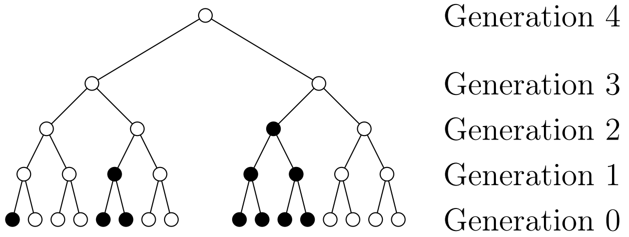

Figure 1: Five generations producing 7 mutants from 3 mutations.

the formulas described in this paper will be of the form

$$\Pr(R = r) = p_r(m) = q_r(m)e^{-\alpha m} \tag{2.1}$$

where $m = \mu n$ is the fundamental parameter that we try to estimate, α is a real number, and $q_r(m)$ is a polynomial of degree r.

We assume that all cells take approximately the same amount of time between cell divisions. We use the letter g to denote the number of generations backwards in time, setting $g = 0$ to denote when the experiment finishes. Under this assumption, the cell divisions that took place recently (that is, with g small), are assumed to take place synchronously. However the bulk population may or may not be assumed to be dividing asynchronously. We give each mutant a *generation number* $g \geq 1$, which says this mutant was created g generations ago. In future papers we will drop the assumption of the same amount of time between cell divisions.

Another assumption we make is that the population of non-mutant cells is effectively infinite. Thus the number of non-mutant cells is effectively the same as the number of all cells g generations ago, and does not depend upon the number of mutations created.

Finally, we will allow for the possibility that cells may die or become non-functional. If cell death does not occur, then the total number of cells g generations ago is $2^{-g}n$. Figure 1 illustrates the process looking only four generations back, but in our formulas we assume that the total number of generations is effectively infinite.

Haldane produced a distribution for the Luria-Delbrück experiment based on the assumption that all the cells divide with perfect synchronicity, and no cells die. The formula can be described thus: given an integer $r \geq 0$, let P_r denote the set of sequences $(a_0, a_1, \dots)$ of non-negative integers, with only finitely many non-zero terms, such that $\sum_{s=0}^\infty a_s 2^s = r$. Then

$$\Pr(R = r) = e^{-m} \sum_{(a_s) \in P_r} (m/2)^{\sum_{s=0}^\infty a_s} \Big/ \prod_{s=0}^\infty a_s! \, 2^{s a_s} \tag{2.2}$$

(Note that [11] has a typographical error in stating this formula. Note also that his g is our $m/2$.)

3 Generating functions for probability distributions

If R is a random variable that takes values in the non-negative integers, then the *probability generating function* of R is defined by the formula

$$G_R(x) = \sum_{r=0}^{\infty} \Pr(R = r) x^r \tag{3.1}$$

Note that G_R is an analytic function, with radius of convergence at least as big as 1. We will make much use of the following well known results.

Proposition 3.1. *Let X_k ($1 \leq k < \infty$) be a sequence of non-negative integer valued independent random variables.*

1. *If the sum $Z = \sum_{k=1}^{\infty} X_k$ converges almost surely, then*

$$G_Z(x) = \prod_{k=1}^{\infty} G_{X_k}(x) \tag{3.2}$$

2. *If X_k are identically distributed, and N is a non-negative integer valued random variable that is independent of the X_k, and if $Z = \sum_{n=1}^{N} X_k$, then*

$$G_Z(x) = G_N(G_{X_k}(x)) \tag{3.3}$$

3. *If N is a Poisson random variable with mean λ, then*

$$G_N(x) = e^{\lambda(x-1)} \tag{3.4}$$

All the distributions we shall consider will have a generating function of the form

$$G_R(x) = e^{-\alpha m} e^{m H_R(x)} \tag{3.5}$$

where $\alpha > 0$, and $H_R(x)$ is an analytic function satisfying $H_R(0) = 0$ with Taylor series

$$H_R(x) = \sum_{k=1}^{\infty} h_k x^k \tag{3.6}$$

with $h_k \geq 0$ for all $k \geq 1$.

For example if the mutation acquisition were Lamarckian, then R would have the Poisson distribution with mean m, that is

$$p_r(m) = e^{-m} m^r / r!, \quad G_R = e^{-m} e^{mx} \tag{3.7}$$

and it is well known that the generating function is

Another example is the Lea-Coulson distribution [4]. This assumes that the mutation acquisition is Darwinian, but it also approximates the discrete replication process by a continuous process. They obtain a distribution that satisfies (3.5) with $\alpha = 1$ and

$$H_R(x) = \sum_{k=1}^{\infty} \frac{x^k}{k(k+1)} = \frac{x + (1-x)\log(1-x)}{x} \tag{3.8}$$

There are two ways to compute the polynomials q_r from the power series for $H_R(x)$. One way is to place $H_R(x)$ into the Taylor's series for e^x. This involves multiplication of polynomials. We discuss this method in Section 10.

The other algorithm is described in Ma, Sandri, and Sarkar [6]:

$$q_0(m) = 1 \tag{3.9}$$

$$q_r(m) = m \sum_{s=1}^{r} \frac{s}{r}\, h_s q_{r-s}(m) \tag{3.10}$$

In practice, this algorithm worked very well. Note that to compute $q_r(m)$, one only needs to know h_s for $s \le r$.

4 The generating function approach

The approach adopted throughout this paper is to compute the generating function for R, the number of mutations. Let us first illustrate this method to derive the Haldane distribution. We assume the cell division is completely synchronous, and that no cells die or malfunction.

While this paper was being prepared, we discovered that the following result is a special case of a result that appeared in 2013 [15].

Theorem 4.1. *If R has the Haldane distribution, then its generating function is given by equation* (3.5) *with $\alpha = 1$ and*

$$H_R(x) = \sum_{g=0}^{\infty} \frac{x^{2^g}}{2^{g+1}} \tag{4.1}$$

Hence from the Ma-Sandri-Sarkar algorithm equation (3.10) *we obtain the rapid formula*

$$q_r(m) = \frac{m}{2r} \sum_{g=0}^{\lfloor \log_2 r \rfloor} q_{r-2^g}(m) \tag{4.2}$$

From equation (4.2) we obtain the polynomials given in Table 1.

Proof. We know that g generations ago, the population count is $n/2^g$. All of these were created from cells dividing $g+1$ generations ago, and therefore half of these have copied genotypes, the rest having the original genotype. Thus the number of mutants created at

r	$\Pr(X = r)$
0	e^{-m}
1	$\frac{m}{2}\, e^{-m}$
2	$\left(\frac{m^2}{8} + \frac{m}{4}\right) e^{-m}$
3	$\left(\frac{m^3}{48} + \frac{m^2}{8}\right) e^{-m}$
4	$\left(\frac{m^4}{384} + \frac{m^3}{32} + \frac{m^2}{32} + \frac{m}{8}\right) e^{-m}$
5	$\left(\frac{m^5}{3840} + \frac{m^4}{192} + \frac{m^3}{64} + \frac{m^2}{16}\right) e^{-m}$
6	$\left(\frac{m^6}{46080} + \frac{m^5}{1536} + \frac{m^4}{256} + \frac{7m^3}{384} + \frac{m^2}{32}\right) e^{-m}$
7	$\left(\frac{m^7}{645120} + \frac{m^6}{15360} + \frac{m^5}{1536} + \frac{m^4}{256} + \frac{m^3}{64}\right) e^{-m}$
8	$\left(\frac{m^8}{10321920} + \frac{m^7}{184320} + \frac{m^6}{12288} + \frac{m^5}{1536} + \frac{25m^4}{6144} + \frac{m^3}{256} + \frac{m^2}{128} + \frac{m}{16}\right) e^{-m}$
9	$\left(\frac{m^9}{185794560} + \frac{m^8}{2580480} + \frac{m^7}{122880} + \frac{m^6}{11520} + \frac{3m^5}{4096} + \frac{m^4}{512} + \frac{m^3}{256} + \frac{m^2}{32}\right) e^{-m}$
10	$\left(\frac{m^{10}}{3715891200} + \frac{m^9}{41287680} + \frac{m^8}{1474560} + \frac{7m^7}{737280} + \frac{5m^6}{49152} + \frac{61m^5}{122880} + \frac{m^4}{768} + \frac{5m^3}{512} + \frac{m^2}{64}\right) e^{-m}$

Table 1: Probabilities according to Haldane's distribution.

that time is N_g, a Poisson random variable with parameter $\frac{1}{2}\mu n/2^g = 2^{-g-1}m$. The number of mutations that arise from the gth generation mutants is $2^g N_g$. Hence

$$R = \sum_{g=0}^{\infty} 2^g N_g \tag{4.3}$$

Furthermore, because we suppose the total population of cells to be very large, we can assume that the random variables N_g are independent. Thus we obtain the generating function

$$\begin{aligned} G_R(x) &= \prod_{g=0}^{\infty} G_{N_g}(x^{2^g}) = \prod_{g=0}^{\infty} e^{2^{-g-1}m(x^{2^g}-1)} \\ &= \exp\left(m \sum_{g=0}^{\infty} 2^{-g-1}(x^{2^g} - 1)\right) = e^{-m} \exp\left(m \sum_{g=0}^{\infty} 2^{-g-1} x^{2^g}\right) \end{aligned} \tag{4.4}$$

□

This approach can be generalized as follows. First, we define times t_k, $(k = 0, 1, 2, \dots)$, with $0 = t_0 < t_1 < t_2 < \dots$. The units of time will be generations, and as with generation number, a positive number denotes the number of generations before the experiment is completed. Let us suppose that a time t_k, $(k \geq 1)$, and only at those times, some or all of the cells divide. For each k, we let n_k denote the total population of cells at time t_k, and let $\pi_k = n_k/n$. Thus $1 = \pi_0 \geq \pi_1 \geq \pi_2 \geq \cdots > 0$. Because we have assumed that the total population of cells is effectively infinite, we can assume $t_k \to \infty$ as $k \to \infty$.

We consider two random variables:

1. for each possible $k \geq 0$, the number of mutants N_k created at negative time t_k, which will always be a Poisson random variable with parameter $\pi_k - \pi_{k+1}$;
2. for each possible $k \geq 0$, the number of mutations M_k that arise from any one mutant which was created at time t_k.

Then the total number of mutations will always be given be the formula

$$R = \sum_{k=0}^{\infty} \sum_{l=1}^{N_k} M_k^{(l)} \tag{4.5}$$

where $M_g^{(l)}$ denotes independent copies of M_l. Hence we obtain the generating function

$$\begin{aligned} G_R(x) &= \prod_{k=0}^{\infty} G_{N_k}(G_{M_k}(x)) = \prod_{k=1}^{\infty} e^{m(\pi_k - \pi_{k+1})(G_{M_k}(x)-1)} \\ &= \exp\left(m \sum_{k=0}^{\infty} (\pi_k - \pi_{k+1})(G_{M_k}(x) - 1) \right) \end{aligned} \tag{4.6}$$

Let's explain first how the Haldane distribution fits into this paradigm. Since cell division is synchronous, we set $t_g = g$ where the time is measured in multiples the time between cell divisions. Since no cells die, we have that $\pi_g = 2^{-g}$, and $M_g = 2^g$, so that $G_{M_g}(x) = x^{2^g}$. Substituting into equation (4.6), we obtain equation (4.1).

We can also obtain the Lea-Coulson distribution using this paradigm. Note the argument we present is in their paper as a “second proof" (and indeed the whole “generating function approach" is inspired by their paper). We still suppose that no cells die, but we assume that the formulas $\pi_k = 2^{-t_k}$ and $M_k = 2^{t_k}$ still hold when t_k is not an integer. To get the most even spacing of the generating of cells, we set $t_k = \log_2(k+1)$. Then $\pi_k - \pi_{k+1} = 1/(k+1) - 1/(k+2) = 1/(k+1)(k+2)$, and $M_k = k+1$. Thus equation (3.8) follows after substituting the summation variable k by $k-1$.

5 Haldane's distribution versus the Lea-Coulson distribution

The Lea-Coulson assumes that the generations are somehow being created continuously. From this point of view, it would seem that their distribution is unrealistic.

However there also are problems with the Haldane distribution. It is obvious that the probability of obtaining 128 mutants is far higher than the probability of obtaining 127 mutants. If one had a single mutation, say, seven generations ago, this will result in 128 mutants. But it is possible that a few of these mutants will either die (and perhaps any of their mutant ancestors died), or fail to be transferred to the plate properly. Thus while Haldane's distribution puts a very low probability of getting 127 mutants compared to 128

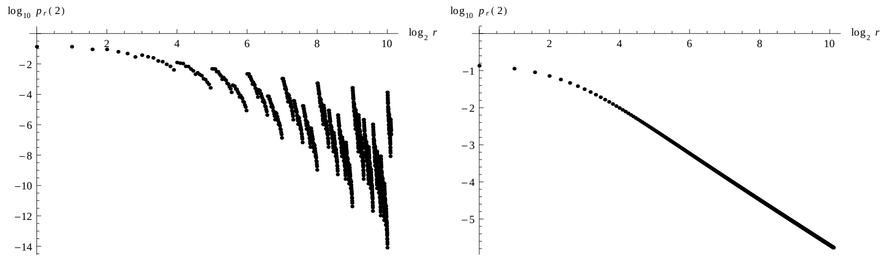

Figure 2: Log/Log graphs of $p_r(2)$ against r, for Haldane's distribution, and for the Lea-Coulson distribution.

mutants, nevertheless from an experimental point of view, it is very likely that not all of the 128 mutants will be observed, and the chances of seeing 127 or less mutants is comparable with the probability of seeing 128 mutants.

To show this effect, in Figure 2 we give log/log plots of $p_r(m)$ to r, for the value $m = 2$, for both distributions. On the x-axis we show $\log_2(r)$, and on the y-axis we show $\log_{10}(p_r(2))$. Notice that for the Lea-Coulson distribution that this graph approaches a straight line as $r \to \infty$. In fact, in [6], it is shown that $p_r(m) \approx c_m r^{-2}$ as $r \to \infty$, where c_m depends only on m.

We will look at several ways to modify the Haldane distribution.

1. Allow that some cells may die.

2. Examine the situation when only a fixed proportion of cells are plated with the antibiotic.

3. Consider asynchronous cell division.

4. Allow that the mutants replicate at a different rate than the non-mutants [3].

We will explore the third case only superficially, as we have not yet obtained formulas that fully handle this issue.

6 Accounting for cell death

Let us now assume that a cell dies with probability θ before it replicates. Thus from one cell, the expected number of cells that come from this cell one generation later is $2(1-\theta)$. Thus

$$\pi_g = (2(1-\theta))^{-g} \tag{6.1}$$

The computation of G_{M_g} is given by a recursion relationship. From a mutant created g generations ago, with probability θ this mutant will die, and with probability $1-\theta$ there will by $X+Y$ cells, where X and Y are independent and have the same distribution as M_{g-1}. Thus we obtain

$$G_{M_0}(x) = x \tag{6.2}$$
$$G_{M_{g+1}}(x) = \theta + (1-\theta)[G_{M_g}(x)]^2 \tag{6.3}$$

These formulas are then substituted into equation (4.6) to obtain equation equation (2.1) with

$$\alpha = \sum_{g=0}^{\infty} (1-2\theta)(2(1-\theta))^{-g-1}(G_{M_g}(0)-1) \tag{6.4}$$
$$H_R(x) = \sum_{g=0}^{\infty} (1-2\theta)(2(1-\theta))^{-g-1}(G_{M_g}(x)-G_{M_g}(0)) \tag{6.5}$$

It is unlikely we will find an explicit formula for solving the recurrence relation (6.3), because this is in essence the same recurrence relation that is used to define the famous Mandelbrot set [7].

7 Plating only a fixed proportion of the cells

This is discussed for the Lea-Coulson distribution in [13, 14].

Let us now assume that we perform an experiment such that the probability of obtaining $R = r$ mutants is given by p_r, with corresponding generating function $F_R(x) = e^{-\alpha m} \exp(mH_R(x))$. Now suppose that we only plate a proportion $\theta \in (0,1]$ of the cells. How many mutants will we observe? Let us call the number of mutants observed R_θ.

Theorem 7.1. *The random variable R_θ has generating function*

$$G_\theta(x) = G_R(1-\theta+\theta x) = e^{-\alpha_\theta m} \exp(mH_\theta(x)) \tag{7.1}$$

where

$$\alpha_\theta = \alpha - H_R(1-\theta) \tag{7.2}$$
$$H_\theta(x) = H_R(1-\theta+\theta x) - H_R(1-\theta) \tag{7.3}$$

Proof. We have that R_θ has the same distribution as $\sum_{r=1}^{R} I_r$, where I_r is a sequence of independent random variables taking the value 1 with probability θ, and the value 0 with probability $1-\theta$. The result follows from applying Proposition 3.1.

Another way to obtain this formula is to use the binomial distribution

$$\Pr(R_\theta = r) = \sum_{n=r}^{\infty} \Pr(R=n) \Pr(R_\theta = r \big| R = n) = \sum_{n=r}^{\infty} \binom{n}{r} \theta^r (1-\theta)^{n-r} p_n \tag{7.4}$$

and then rearrange the resulting double sum. □

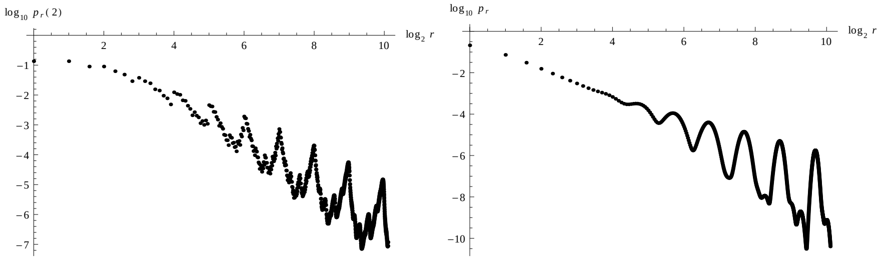

Figure 3: Log/Log graph of $p_r = \Pr(R_\theta = r)$ against r, when cells die with probabilities $\theta = 0.01$ (left hand side) or $\theta = 0.1$ (right hand side), and R has the Haldane distribution with $m = 2$.

Usually, the most timely and accurate numerical method for computing α_θ, and coefficients of the power series for H_θ, is by direct evaluation from the power series of the derivatives of $H(x)$ at $x = 1 - \theta$. For the Lea-Coulson probability measure, it makes sense to compute α directly by substituting equation (3.8). One might also be tempted to compute the coefficients of H_θ by symbolically differentiating H_R given in equation (3.8). However the symbolic derivatives become very unwieldy.

8 Asynchronous cell division

There are two notions of asynchronous cell division that we shall consider.

1. *slight asynchronous cell division* We suppose that cell division is asynchronous enough so that at time g generations before the experiment ends, the cell population is given by $2^{-g}n$, even if g is not an integer, but not so asynchronous so that over a few generations the cell division is well approximated by synchronous cell division.

2. *full asynchronous cell division* We suppose that cell division is sufficiently asynchronous that we should base our model on a probability distribution that describes the time for a cell to divide.

Accounting for full asynchronous cell division seems to be very difficult, and we have not yet solved this problem[1]. In this chapter we shall consider slight asynchronous cell division, and show that it makes no difference to the Haldane distribution.

We assume that cell splitting takes place every g/n time units, where at the end we let $n \to \infty$. Therefore in equation (4.6), we set $t_k = k/n$, $\pi_k = 2^{-k/n}$, and $G_{M_k} = 2^{[k/n]}$, where

[1] Note added in August 13, 2016: this problem has been solved in the thesis of Hesam Oveys [8].

$[x]$ denotes the integer part of x. Then writing $k/n = g + j/n$ where $g = [k/n]$, we obtain

$$\begin{aligned} G_R(x) &= \exp\left(m \sum_{g=0}^{\infty} \sum_{j=0}^{n-1} (2^{1/n} - 1) 2^{-g-(j+1)/n} (x^{2^g} - 1) \right) \\ &= \exp\left(m \sum_{g=0}^{\infty} 2^{-g-1} (x^{2^g} - 1) \right) \end{aligned} \tag{8.1}$$

where the last step comes from summing the geometric series. Thus the formula is independent of n.

9 Mutants replicating at a different rate than non-mutants

The analogous formula to the Lea-Coulson was calculated by [3]. We will do the analog for the Haldane distribution. While this paper was being prepared, we found out that the this result appeared in 2013 [15].

Let g denote the number of generations back in time as counted in the time for a mutant to replicate. Suppose that the time for a non-mutant to replicate is τ times the time for a mutant to replicate. Then we can apply equation (4.6) with $M_g = 2^g$, and $\pi_g = 2^{\tau g}$, to obtain equation equation (2.1) with

$$\alpha = -1, \quad H_R(x) = \sum_{g=0}^{\infty} (1 - 2^{-\tau}) 2^{-\tau g} x^{2^g} \tag{9.1}$$

10 Comparison of the Ma-Sandri-Sarkar algorithm with using the Fast Fourier Transform

Another way to compute the probability generating function given by equation (3.5) is to expand it using Taylor's series for e^x, and obtain

$$G_R(x) = e^{-\alpha m} \sum_{n=0}^{\infty} \frac{1}{n!} m^n [H_R(x)]^n \tag{10.1}$$

In order to calculate $p_r(m)$, it is only necessary to sum the series to the rth term, and to calculated $[H_R(x)]^n$ up to x^r. It is well known that $[H_R(x)]^n$ can be rapidly calculated using the Fast Fourier Transform. If one performs a algorithm complexity analysis, one finds that the time to compute $p_r(m)$ for $m \le n$ is $O(n^3)$ using the Ma-Sandri-Sarkar algorithm, but is $O(n^2 \log n)$ when using the Fast Fourier Transform. (The exception is when using the Ma-Sandri-Sarkar for Haldane's distribution, when the time taken is also $O(n^2 \log n)$.)

A difficulty with the Fast Fourier Transform is getting sufficient accuracy in the coefficients of $G_R(x)$. If one calculates $[H_R(x)]^n$ using long multiplication (and this is, in effect, what the Ma-Sandri-Sarkar algorithm is doing), then because all the coefficients of $H_R(x)$ are positive, there will be no canceling large numbers to produce small numbers in computing the coefficients. However, the Fast Fourier Transform combines the various coefficients using complex numbers, and so the accuracy of the final answer is limited by the size of the largest coefficient.

However there is also the fast polynomial package FLINT [2], which comes standard with the Sage software package [10]. While all our computations were performed with Sage, we found that in our situations that the Ma-Sandri-Sarkar was much faster than using the built in polynomial multiplication.

11 Analyzing data using likelihood functions/Bayesian statistics

In this section, we will show how to compute the likelihood function to estimate the value of m from experimentally obtained data. Suppose we have performed n separate experiments, using identical but independent protocols, and obtain counts of plates $r_1, r_2, \dots, r_n$. Then the likelihood function is

$$L(m) = \prod_{k=1}^{n} \Pr(R = r_k) \tag{11.1}$$

where R has the distribution that we believe most correctly represents our situation.

The Bayesian approach is to suppose that $L(m)$ represents a probability distribution for a random variable M, where $\Pr(M \in [m_1, m_2])$ represents the 'belief' that we have that the parameter m lies between m_1 and m_2. This is related to $L(m)$ via the formula

$$\Pr(M \in [m, m + dm]) = C f(m) L(m) dm \tag{11.2}$$

where here we think of dm is being infinitesimal. Here $f(m)dm$ is usually called the prior distribution, and the constant C is chosen so that the integral of the right hand side is zero. The paper [1] describes estimates of M using Bayesian statistics, and they use prior distributions like 1m m^{-1} or m^{-2}.

Thus the area under likelihood function represents the probability distribution of M if $f(m) = 1$. However, since m is some kind of scaling factor, it makes more sense to draw the graph of $L(m)$ on a graph where the x-axis represents $log(m)$. Then the area under likelihood function represents the probability distribution of M if $f(m) = 1/m$.

If we treat M as a random variable, then it makes sense to also compute the expected value and variance of M. The computing of C, $E(M)$ and $\text{var}(M)$ requires computing certain integrals of the form of linear combinations of $\int_0^\infty m^a e^{-bm} \, dm$, and so can be easily be computed using a change of variables and the Γ function.

However, since we are interested in graphing these using a log scale on the x axis, to get an idea of which range of m we should use for plotting, we want to compute $E(\log(M))$ and $\text{var}(\log(M))$. These are not so easily computed, but numerical integration does an outstanding job.

In any case, we would hope that the amount of data collected should be enough so that the formulas $E(M)$ or $\exp(E(\log(M)))$ or the maximum of $L(m)$, should give similar answers, even if different prior distributions are used. And we would hope to see a bell-shape distribution, so we should get similar answers for the estimate of the middle 68.3%-ile of belief, that is, $E(M) \pm \operatorname{var}(M)$ or $\exp(E(\log(M)) \pm \operatorname{var}(\log(M)))$.

Finally, in performing the computations, the numbers involved tend to be much smaller than much computer software can represent. For example, the commonly used double precision IEEE floating point representation cannot represent numbers much smaller than 10^{-308}, and values far smaller than this will often arise in calculating the likelihood function. However we used the software package sage [10], and this was well capable of handling very small numbers.

12 Simulations

13 Analysis of real data

The following experiments were performed by students who were part of the 'Mathematics in Life Science' program at the University of Missouri.

The mutations in question confer resistance to an antibiotic (canavanine) in yeast cells. There were two types of resistant mutants, one giving red colonies, the other giving white colonies. Our experiment focused on the red mutants, because we were able to sequence their DNA to find the exact mutation responsible for resistance.

In Experiment A, the students plated 70 60µl cultures, each arising from a 25ml culture. In Experiment B was to plate 50 60µl cultures from a single 25ml culture. Each culture started at a cell density of 10^5 cells/ml and ended at a cell density of 10^8 cells/ml. See Figure 4.

In Experiment A, two of the 70 cultures were accidentally destroyed. The counts for the other 68 cultures were as follows: 2, 4, 0, 0, 0, 0, 244, 55, 0, 141, 0, 0, 0, 1, 0, 511, 0, 0, 0, 0, 4, 0, 0, 0, 0, 5, 95, 2, 0, 0, 11, 0, 0, 0, 5, 0, 0, 49, 0, 1, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 3, 0, 0, 0, 11, 1, 16, 0, 0, 0, 1, 0, 1, 4, 1, 2, 0, 0. These were analyzed using both the Haldane and Lea-Coulson distributions, with $\theta = 0.0024$.

mean = 0.976689, stand dev = 0.126925.

The counts for Experiment B were as follows: 155, 188, 189, 191, 173, 161, 164, 221, 191, 221, 148, 173, 186, 152, 195, 86, 90, 154, 133, 149, 165, 182, 162, 144, 129, 65, 165, 159, 151, 183, 170, 130, 140, 118, 162, 132, 154, 142, 134, 151, 150, 89, 147, 143, 142, 191, 93, 119, 163, 133. Since these were all taken from the same culture, we only took the total value, which was 7628, and analyzed this using both the Haldane and Lea-Coulson distributions with $\theta = 0.12$.

The counts for the 50 60µl cultures suggest we hit a jackpot. However in looking at the distribution created by the Bayesian method, the data seems to rather confidently give a large figure for m, which we strongly suspect is too large. One way to interpret this data is

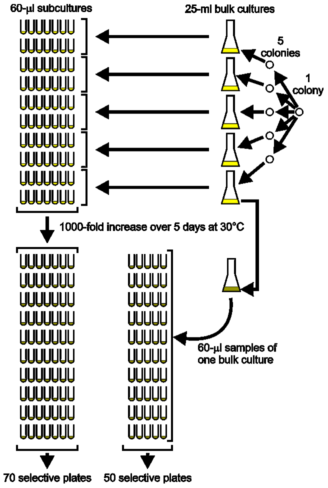

Figure 4: A diagram illustrating the experiment.

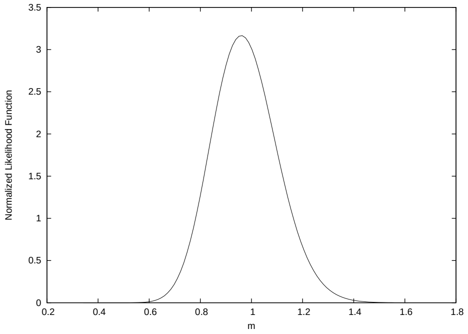

Figure 5: The normalized likelihood function for real data.

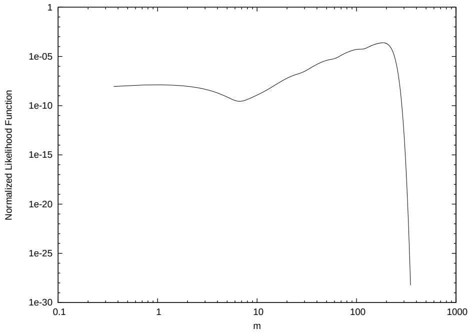

Figure 6: The normalized likelihood function for the bulk culture.

to use a log-log plot for the likelihood function. In this case it can be seen ...... issues with how to interpret probabilities obtained using the Bayesian method.

Acknowledgments

The first named author would like to thank Miriam Golomb for many useful conversations.

References

[1] Asteris, G. and Sarkar, S. (1996) Bayesian Procedures for the Estimation of Mutation Rates from Fluctuation Experiments, Genetics, 142, 313-326.

[2] FLINT: Fast Library for Number Theory, `http://www.flintlib.org`.

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

[4] Lea, D.E. and Coulson, C.A. (1949) The distribution of the number of mutants in bacterial populations. Journal of Genetics 49, 264-285.

[5] Luria, S. E. and Delbrück, M. (1943) Mutations of bacteria from virus sensitivity to virus resistance. Genetics 28, 491-511.

[6] Ma, W.T., Sandri, G.Vh. and Sarkar S., (1992) Analysis of the Luria-Delbrück Distribution Using Discrete Convolution Powers. Journal of Applied Probability, 29, 255-267.

[7] Mandelbrot, B. (1980) Fractal aspects of the iteration of $z \mapsto \lambda z(1-z)$ for complex λ, z, Annals NY Acad. Sci. 357, 249-259.

[8] Oveys, Hesam. (2015) Age-dependent Branching Processes and Applications to the Luria-Delbrück Experiment, Ph.D. thesis, University of Missouri.

[9] Rosche, W.A. and Foster, P.L. (2000) Determining Mutation Rates in Bacterial Populations, Methods 20, 4-17.

[10] Sage: An open-source mathematics software system, `http://www.sagemath.org`.

[11] Sarkar S., (1991) Haldane's solution of the Luria-Delbrück distribution. Genetics 127, 257-261.

[12] Smith, George P., Golomb, Miriam, Billstein, Sidney K., and Montgomery-Smith, Stephen, (2015) An Enduring Legacy: The Luria-Delbrück Fluctuation Test as a Classroom Investigation in Darwinian Evolution, American Biology Teacher 77(8), 614-619.

[13] Stewart FM, Gordon DM, Levin BR. Fluctuation analysis: the probability distribution of the number of mutants under different conditions. Genetics. 1990 Jan;124(1):175-185.

[14] Stewart FM. Fluctuation analysis: the effect of plating efficiency. Genetica. 1991;84(1):51-5.

[15] Ycart B., (2013) Fluctuation Analysis: Can Estimates Be Trusted? PLoS ONE 8(12): e80958. doi: 10.1371/journal.pone.0080958.