CONTENTS

Chapter One

Fluctuations and the Nature of Mutations

1.1 CHANCE FAVORS THE INDEPENDENT MUTATION

Evolution via natural selection denotes any nonrandom change in the genetic makeup of a population due to the differential reproduction and/or survival of individuals. As such, evolution via natural selection requires standing variation to facilitate dynamic change in populations, again and again, over generation and generation. Mutations in the genome of replicating organisms are the grist for this long wheel of evolutionary change. Yet, in the early part of the twentieth century, scientists had not yet identified the molecular basis for heredity. Big questions remained in the field. Big questions that, for us, have become matters to be read and memorized in textbooks. But to start in the process of integrating quantitative methods into our study of living systems requires that we try, however difficult, to displace ourselves from the present time and put ourselves in the mindset of others.

Early molecular biologists faced a profound challenge: what was the basis for the generation of individual variation? The existence of diversity was never in question, but how such diversity came into being was. The two major theories differed radically with respect to the nature of the link between the introduction of variation and its differential selection. Are mutations dependent on selection or independent of selection? The idea that mutations depend on selection seems heretical to modern practitioners of quantitative biology. Yet it was not certainly not always the case. Charles Darwin's theory of evolution via natural selection presumed that variation was introduced in some kind of heritable material. The differential success in survival and reproduction became the mechanism to "select" for a subset of variants. Then those more fit variants would produce new offspring, different again from them, and so on. In essence, mutations are independent of selection.

The contrasting idea is often attributed to Jean-Baptiste Lamarck, a French biologist active in the late eighteenth and early nineteenth centuries. To understand Lamarckian evolution, it is worth sharing a few examples. First, consider a parent who decides to join a gym. She (or he) gets strong. Will the child be more likely to have bigger muscles than if the parent had skipped the gym and stuck to a steady diet of barbecue and ice cream? It seems unlikely, but according to Lamarckian evolution the answer would, in fact, be yes. Another example. The classic one. Consider a female giraffe grazing in the Serengeti. Food is sparse, so the female giraffe must stretch and stretch to reach its preferred acacia leaves. One day

the giraffe has a calf. Would the calf have a shorter neck had the mother not had to stretch as far? This is the essence of Lamarckian evolution: it posits that experiences that change the phenotypic state of a parent will be passed on heritably to its offspring. In other words, "mutations depend on selection." The examples of the gym aficionado and the long-necked giraffe seem improbable. But as anyone who follows the field of epigenetics knows, present experiences can shape the phenotypes of offspring, often in profound ways.

But Luria and Delbrück did not work with humans or giraffes. Instead, working with microbes and their viruses afforded them a quantitative framework to directly address these two hypotheses. It was already known from the work of Frederick Twort (1915) and Felix d'Herelle (1917) that viruses could infect, lyse, and propagate on bacteria (for a historical perspective, see Summers 1999). These bacteriophage were relatively specific in their activity. That is, some phage could spread on certain bacteria but not others. The difference between a phage-resistant and phage-susceptible strain could be identified through a simple colony assay where the number of resistant bacteria were measured on agar plates. Hence, in the case of microbes and viruses, the two hypotheses can be summarized as follows:

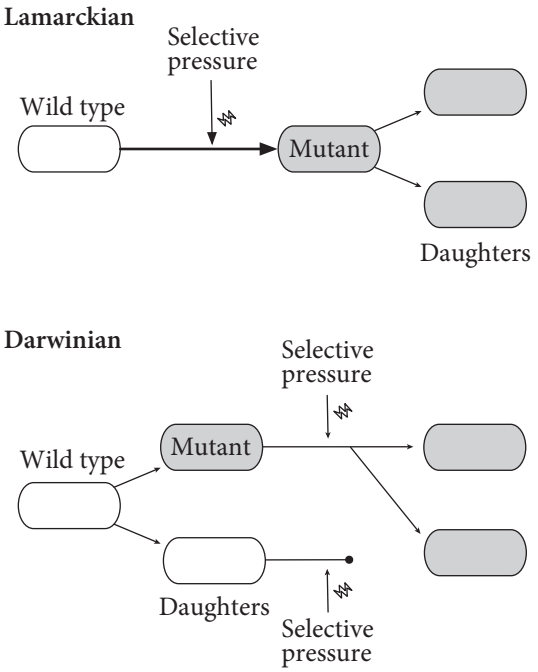


Figure 1.1: Schematic of Lamarckian and Darwinian views of selection and mutation. (Top) Lamarckian selection changes the cell, and daughter cells heritably retain the characteristic of the mother. (Bottom) Darwinian selection changes the functional traits/phenotypes of the cell, but differences in daughter cells (when they arise via mutation; see gray versus white) are independent of the selective experience of the mother.

Spontaneous mutation The change from virus sensitivity to virus resistance happens spontaneously to cells, irrespective of their interaction with viruses. This spontaneous change is rare.

Acquired heritable immunity A small fraction of infected cells survive and acquire an immune state, which can be passed on heritably to daughter cells. This acquired change is rare.

The hypotheses of spontaneous mutation and acquired heritable immunity map roughly to Darwinian and Lamarckian evolution, respectively (Figure 1.1). Yet these two hypotheses should have profoundly different consequences on the variability expected in colony counts of cells that can no longer be killed by viruses, even if their recent ancestor could.

The Luria and Delbrück paper is a seminal event in the history of biological sciences. It proved transformative to understanding the nature of evolution—showing that heritable changes in cellular state were independent of, rather than dependent on, selective forces. The finding is particularly striking given that the work was completed 10 years before the discovery of the double-helix structure of DNA was made possible by Francis Crick, Rosalind Franklin, James Watson, and others (Judson 1979). For Luria

and Delbrück, the selective force was the killing power of bacteriophage. Bacteriophage ("phage") are viruses that exclusively infect and kill bacteria. Yet the first image of a virus was only seen under a microscope 4 years before! Nonetheless, phage and bacteria were already becoming the workhorses driving discoveries into cellular function.

It was in this context that Salvador Luria—a biologist from the University of Indiana—and Max Delbrück—a physicist at the University of Vanderbilt—initiated what was to become a long-term and deservedly famous collaboration. (The rest of this chapter refers to their collaborative work with the initials LD.) Yet the success of LD's ideas came slowly. Their 1943 paper, "Mutations of Bacteria from Virus Sensitivity to Virus Resistance," is difficult to read (Luria and Delbrück 1943). The difficulty is not ours alone, separated as we are in time by 70 years and missing context. Perhaps the authors simply wrote in different ways and the hints at their underlying methods in the text would have been well understood by their peers. This seems unlikely.

Recall that the work of Alfred Lotka and Vito Volterra on predator-prey dynamics was not yet 20 years in the past. The integration of mathematics and biology was hardly commonplace. Moreover, unlike the bulk of models of biological systems, the work of Luria and Delbrück combines elements of both continuous and discrete mathematics. Perhaps it was only Delbrück who truly understood the mathematical nature of his arguments. Indeed, 10 years later, Esther Lederberg and Joshua Lederberg leveraged their ingenious idea of replica plating to show the clonal nature of virus resistance in bacteria (Lederberg and Lederberg 1952). It was then that LD's ideas began to gain acceptance not only because of authority but through the adage "seeing is believing."

The following sections lay out the core arguments to decide whether mutations are dependent on or independent of selection. In doing so, it is critical to review the nature of the heritable state as well as the experimental details and mechanistic hypotheses at stake. This chapter reviews multiple lines of evidence in support of the competing hypotheses, including the quantitative predictions for both the mean and variance of mutant colonies. As we will see, the history of the LD experiments and their outsized influence on the foundations of molecular biology lie in an "irreproducibility opportunity" (Figure 1.2). This schematic provides a visual recapitulation of the kind of data that LD observed—in which some of their experimental replicas included zero (or very few) resistant colonies and others included hundreds. As it turned out, the large-scale disagreement among replicate experiments was precisely the evidence needed to distinguish between the Darwinian and Lamarckian hypotheses. And, by the end of this chapter, you will have a sense of how important this variability was (and is) to understanding something fundamental about how life works.

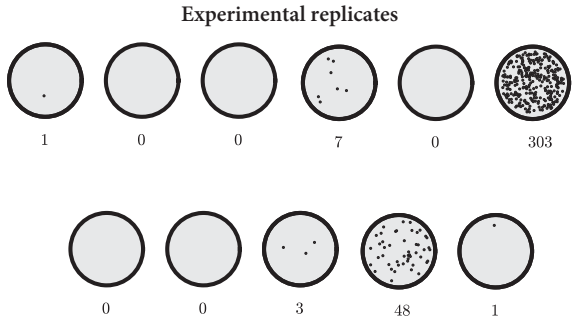


Figure 1.2: Schematic of colony assays illustrating the apparent lack of reproducibility in the Luria-Delbrück experiment. The number of colonies is listed below each plate; these numbers in each experimental replica correspond to experiment 17 (Luria and Delbrück 1943). What would you have done with such large variation between experiments? Is this failure? Or, instead, something more profound? How and why this lack of reproducibility explains the very nature of mutations forms the centerpiece of this chapter.

1.2 CELLULAR PHENOTYPES

Bacteriophage (or phage) are viruses that exclusively infect and lyse bacteria. Infection is initiated via encounter between the virus particle and the surface of the bacterial cell. After encounter and successful adsorption, the genetic material of the phage is injected into the cytoplasm of the bacteria where phage genes redirect the bacterial machinery, including transcriptional enzymes and ribosomes, to copy the viral DNA and produce viral proteins. These viral proteins self-assemble into capsids, which are then packed with viral DNA; through a timed process, viral encoded enzymes—including holins and lysins—make small holes in the inner membrane and cell wall of the bacterial cell. As a consequence, the cell explodes and dozens, if not hundreds or more, virus particles are released. The infection and lysis of bacteria, like *E. coli B*, by phage, like phage α, is depicted in Figure 1.3. This process can be scaled up, millions and billions of times over. Indeed, that is precisely what Luria did.

He did so, as the story goes, on a Sunday. The Saturday night before, in January 1943, Luria had been at a faculty dance and social (those were different times (Luria 1984)). The social included slot machines, which generally yield nothing but occasionally pay off in large jackpots. Luria had observed similarly large, rare events in his experiments to probe the change from virus sensitivity to resistance among bacteria (see an example in Figure 1.2). He reasoned: what if such events were not a mistake in his experimental design but rather a feature of the resistance process itself? These slot machines and their jackpots were the catalyst Luria needed to revisit his own thinking on the nature of mutations (Judson 1979). To test the idea, Luria returned to his laboratory and conducted the prototype of what became the experimental observations at the core of the 1943 LD paper. It is worth explaining precisely what those experiments entail.

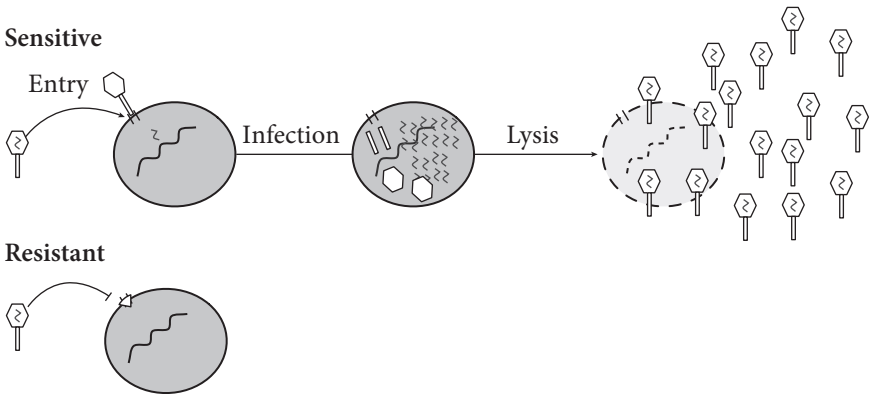


Figure 1.3: Infection and lysis of sensitive bacterial cells by viruses. (Top) In the case of sensitive cells, viruses inject their genetic material into hosts, the virus genome replicates, virus capsids self-assemble and are packed with virus genomes, and then virus particles are released back into the environment following the lysis of cells. (Bottom) In the case of resistant cells, viruses are unable to adsorb, infect, and lyse the cell. Note that, more generally, resistance to infection can be due to extracellular and/or intracellular mechanisms (Labrie et al. 2010).

Luria's experimental design culminates in the interaction of phage T1 with bacteria (now known as *E. coli B*) on agar plates. Despite having many more viruses than bacteria, Luria had observed that bacteria can and do survive the interaction. To reach this point requires the following steps (Figure 1.4). First, a culture of bacteria is grown up overnight. Such cultures typically include bacteria at densities on the order of 10^8 per ml. If grown in a 100 ml flask, this represents over 10 billion bacteria. In parallel, viruses are added to a culture of bacteria. The replication of viruses inside sensitive bacteria leads to the release of large numbers of viruses, which reinfect new cells and release more viruses, such that total virus densities can rapidly exceed 10^9 per ml. Ensuring the culture exclusively contains viruses requires another step. Chloroform is often added to eliminate any remaining bacteria, the culture spun down, and the supernatant removed to extract a *viral lysate*, i.e., a culture of virus particles. This is how the experiment starts. Next, the viral lysate is poured atop agar plates and bacteria are added. The vast majority of bacteria should be infected by viruses and lyse. Yet a few, sometimes hundreds, are not killed. These bacteria replicate, beginning with just a single cell until they form a clustered group of thousands to tens of thousands of bacterial cells on the plate. These dense assemblages of bacteria that arise from a single bacterium are termed a *colony*. How many colonies appear, how often no colonies appear, and the variation in colony counts between replicate plates forms the heart of the Luria and Delbrück experiment.

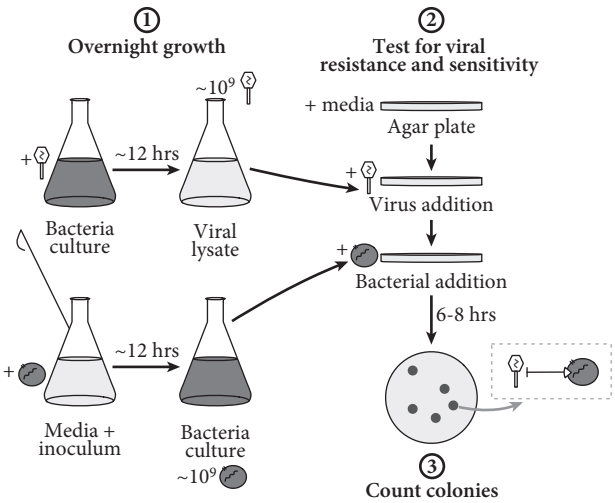


Figure 1.4: The Luria-Delbrück experiment, including overnight growth of bacteria and phage, mixing on agar plates, and colony counting.

The results from different experiments are shown in Table 1.1, where the columns denote distinct experiments and the rows denote distinct counts of the number of colonies in a series of biological replicates as measured in distinct agar plates. There are many striking features of these results. First, there are many replicas with zero resistant colonies. Yet there are also many replicas with dozens if not hundreds of resistant colonies. Imagine yourself staring at this very data, not knowing what Luria and Delbrück discovered. What would you have done? If you are a PhD student, ask yourself: would you show these results to your adviser? Or, instead, would you have thought: there's a mistake in the experiment. It's not repeatable. Yet that lack of repeatability, i.e., the zeros and the jackpots together, is the critical clue to understanding the nature of mutations.

1.3 MUTATIONS THAT DEPEND ON SELECTION

What if Lamarck was right and mutations depend on selection? For a moment, disregard the potential mechanism by which bacteria acquire resistance and/or immunity. Instead, consider what would happen in the event that N bacteria on the agar plate were each exposed

Table 1.1: Number of resistant colonies observed in Luria and Delbrück's 1943 experiment.

	Replica																			
Experiment	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	10	18	125	10	14	27	3	17	17											
10	29	41	17	20	31	30	7	17												
11	30	10	40	45	183	12	173	23	57	51										
15	6	5	10	8	24	13	165	15	6	10										
16	1	0	3	0	0	5	0	5	0	6	107	0	0	0	1	0	0	64	0	35
17	1	0	0	7	0	303	0	0	3	48	1	4								
21a	0	0	0	0	8	1	0	1	0	15	0	0	19	0	0	17	11	0	0	
21b	38	28	35	107	13															

Note: The table here includes a subset of the original data published in (Luria and Delbrück 1943). Each row is a different experiment and each column is a different replicate within that experiment. The number of replicates for each experiment was not fixed. Each replicate denotes a biological replicate referring to distinct numbers of cultures examined in each experiment.

to one or more viruses, and that such interactions can trigger a heritable immunity mechanism. To do so, assume that the probability of an acquired mutation is μ_a. These mutant bacteria are not killed by viruses and pass on this resistance trait to daughter cells. What then is the probability of observing m resistant bacteria, i.e., the "mutants"? This probability is equivalent to flipping a biased coin N times, with the successful outcome occurring quite rarely; e.g., if $\mu_a = 10^{-8}$, then the mutation occurs one in every one hundred million trials. Formally, the probability can be written as follows:

$$p(m|N,\mu_a) = \overbrace{\frac{N!}{(N-m)!m!}}^{\text{permutations}} \overbrace{\mu_a^m}^{\text{mutation}} \overbrace{(1-\mu_a)^{(N-m)}}^{\text{sensitive}}. \tag{1.1}$$

The first term denotes the number of ways to choose exactly m of N individuals. For example, when $m = 1$, then this combinatorial prefactor is N, i.e., the mutation that does occur could have occurred to any one of the N bacteria in the population. Similarly, when $m = 2$, then this combinatorial prefactor is $N(N-1)/2$, i.e., the number of ways to choose two unique members of a population of size N, and so on. The remaining factors correspond to the probability that a mutation with probability μ_a occurs precisely m times and, by virtue of the size of the population, that a mutation does not occur—with probability $1 - \mu_a$—precisely $N - m$ times (explaining why these counts appear as exponents).

This formula denotes the binomial distribution, but saying that does not seem particularly helpful. If you were to calculate this formula on the computer, you might find that calculating massive factorials is not altogether helpful. Instead, we should consider the fact that the experiment was done in a particular regime, that is, when N is a very, very large number, on the order of 10^8 or greater. Similarly, the mutation rate, although unknown, was almost certainly a very small number, on the order of 10^{-8}. In other words, the probability for observing m mutants can be readily calculated in certain limits, e.g., when $N \gg 1$

and $0 < \mu_a \ll 1$. In that regime, the binomial distribution reduces to

$$p(m|N, \mu_a) = \frac{(N\mu_a)^m e^{-N\mu_a}}{m!}. \tag{1.2}$$

This formula denotes the Poisson distribution and is the limit of a binomial distribution given many trials and small probability of success (see the technical appendix for a detailed derivation).

The Poisson distribution has a number of interesting properties. For one, the mean is equal to the value of the argument in both the exponential and the polynomial term: $\bar{m} = N\mu_a$. Moreover, the variance of a Poisson distribution is equal to the mean. Hence, the standard deviation—the square root of the variance—should be $\sigma = \bar{m}^{1/2}$. This scaling implies that replica experiments should yield small fluctuations, such that the standard deviation σ increases more slowly than the mean $\bar{m}$. One way to measure the smallness of fluctuations is to consider the ratio of the standard deviation to the mean, termed the *relative error*. For example, if there are 10 mutants on average, then the standard deviation should be 3 and the relative error, or $\sigma/\bar{m}$, should scale like $\bar{m}^{-1/2}$, or 1/3. Similarly, if there are 100 mutants on average, then the standard deviation should be 10 and the standard error should be 1/10. In essence, there should be a relatively consistent number of mutants between trials. As a result, the acquired immunity hypothesis predicts that repeated experiments should tend to have similar levels of colonies despite the randomness associated with the mutational process.

With this model in hand, let us now adopt the perspective of an experimentalist and try to infer the most likely mutation rate given measurements. In essence, rather than asking how many mutants we expect to see if we know the true mutation rate, we would like to ask: what is the most likely mutation rate compatible with the observations we make? To find the answer, we must turn to the data.

One set of data is reproduced in Table 1.1. The table includes the numbers of resistant colonies in a series of replicate experiments. The number of resistant colonies differ. Some are small, in some cases there are no resistant colonies whatsoever, and some are large, quite large compared to others, e.g., hundreds versus a handful. There are at least two ways to use this data. First is to note that if the process of mutation depends on selection, then we should expect sometimes not to see any mutants at all. This probability is

$$p(0|N, \mu_a) = \frac{(N\mu_a)^0 e^{-N\mu_a}}{0!} = e^{-N\mu_a}. \tag{1.3}$$

Hence, given an observation, f_0, of the fraction of replicates with zero colonies, then the best estimate of the acquired mutation rate should be

$$\hat{\mu}_a = -\frac{\log f_0}{N}. \tag{1.4}$$

This method has advantages but also drawbacks. First, in the event that $f_0 = 0$, then the mutation rate is undefined. In the event that occurs, it is still possible to use a bound, e.g., the frequency of zero events should be $f_0 < 1/s$ where s is the number of replicates. There are other approaches. Note that the average number of mutant colonies, $\bar{m}_{obs}$, is another

feature of the Poisson distribution. It is predicted to be $N\mu_a$. Hence, it may be reasonable to assume that the best estimate of the acquired mutation rate should be $\hat{\mu}_a = \frac{\bar{m}_{obs}}{N}$. This is in fact sensible. Using the mean as the basis for estimating the mutation rate is equivalent to the *maximum likelihood* estimate of the unknown rate.

Formally, we would like to estimate the value of the "acquired" mutation rate that is the most likely value given observations. The choice of the adverb "most" implies there is a range of potential values to choose from. To begin, denote the joint probability of mutation rate and observed number of mutants as $P(\mu_a, m)$. This joint probability can be written as

$$P(\mu_a|m)p(m) = L(m|\mu_a)q(\mu_a). \tag{1.5}$$

This expression leverages the law of total probability such that P denotes the posterior probability of the parameter given the data, p denotes the probability of the data, L denotes the likelihood of the data given a parameter, and q denotes the prior probability of the parameter. Here P is the posterior probability and and L is the likelihood function of observing a certain number of mutant colonies given a known mutation rate. This equation can be rewritten as follows:

$$P(\mu_a|m) = \frac{L(m|\mu_a)q(\mu_a)}{p(m)} \tag{1.6}$$

where $p(m)$ and $q(\mu_a)$ are probability distributions of the data and of the prior of the parameter to be estimated, respectively. Now consider two values of the mutation rate—μ_a and μ_a'—and ask: which is more compatible with observations? To answer this question requires comparing the ratio of the posterior probabilities,

$$\frac{P(\mu_a|m)}{P(\mu_a'|m)} = \frac{L(m|\mu_a)q(\mu_a)p(m)}{L(m|\mu_a')q(\mu_a')p(m)}. \tag{1.7}$$

In the event there is no a priori reason to favor one mutation rate over another, then $q(\mu_a) = q(\mu_a')$. This is what statisticians mean by *uninformed priors*. Using such uninformed priors yields

$$\frac{P(\mu_a|m)}{P(\mu_a'|m)} = \frac{L(m|\mu_a)}{L(m|\mu_a')}. \tag{1.8}$$

In other words, to find the μ_a that is most likely, in a posterior sense, one should find the value of μ_a that maximizes the likelihood. As shown in Figure 1.5, the likelihood function has a zero derivative in μ_a at its maximum (more generally, this is true of both local minima and maxima for functions of one variable). Hence, rather than simulating the likelihood for every possible observation, it is possible to identify a general formula for the maximum likelihood estimate, $\hat{\mu}_a$. The technical appendix explains how to take a first derivative of this likelihood, yielding the maximum likelihood estimate

$$\hat{\mu}_a = m/N. \tag{1.9}$$

In this case, the value as inferred by the mean is the right choice. Caution: This equivalence between the mean and maximum likelihood solution need not always be the case.

But herein lies the problem. The variance of the Poisson distribution is equal to the mean. So we should expect that estimated variances are similar to estimated means. Moreover, we should expect that the standard deviation, which is the square root of the variance, should be smaller than the mean. For example, if there are 10 colonies on average per plate, then the acquired heritable immunity hypothesis predicts that plates will have nearby values, e.g., 5, 12, 7, 9, and so on. Moreover, large deviations should be very rare. This is not the case in the experiments of LD (as seen in Table 1.1). Hence, although it may be possible to estimate an acquired mutation rate through the zero-colony or average-colony methods, the data already suggests that these rates correspond to a quantitative feature of the incorrect mechanism. To consider another mechanism requires that we evaluate the number of mutations that would arise in different replicates if mutations were independent of selection.

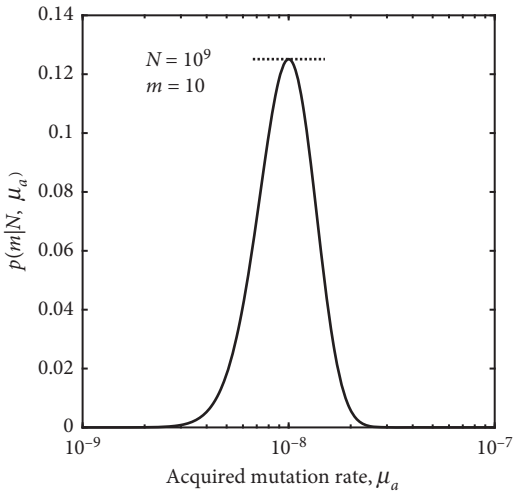


Figure 1.5: Likelihood, $L(m|\mu_a)$ from Eq. (1.2), given variation in μ_a from 10^{-9} to 10^{-7} and the observation that $m = 10$ when $N = 10^9$. As expected, the maximum likelihood corresponds to $\mu_a = m/N$ or $\hat{\mu}_a = 10^{-8}$. Note that the dashed lines provide a visualization of the zero first derivative corresponding to the value of $\hat{\mu}_a$.

1.4 INDEPENDENT MUTATIONS: A CONTINUOUS MODEL

1.4.1 Spontaneous mutations—dynamics

LD proposed a different approach to the origin of resistant mutants in their experiment. Perhaps resistance mutations in the bacteria did not arise due to interactions with the virus. Instead, what happens if the mutant bacteria were already there, waiting, as it were, to be revealed through the process of interacting with viruses that would otherwise kill them? This is the core idea of mutations being independent of selection—and of the Darwinian concept of evolution via natural selection. But how many mutants should there be? This is where theory becomes essential. If mutants arise independent of selection, then in principle they could have arisen very early in the experiment or perhaps near the end, in the very last generation of bacteria to divide before viruses were added to the agar plate. If they arose early, then a single resistant cell could divide many times before being plated on a lawn covered in viruses. Such an experiment would yield a *very large* number of resistant colonies. This possibility is worth exploring in detail.

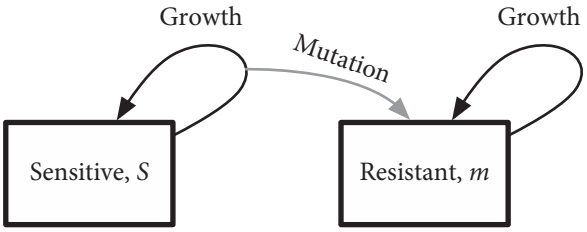


Figure 1.6: Population model of the growth of susceptible bacteria, S, and resistant bacterial mutants, m. Here the S population divides, sometimes yielding mutant bacteria that are resistant to viral infection. The m population also divides and back-mutations leading to virus sensitivity are ignored.

To address this possibility, LD proposed a continuous model of bacterial population dynamics including two populations: sensitive cells and resistant mutants (Figure 1.6). (In practice, it was Delbrück who proposed the mathematical model.) In this continuous model, susceptible bacteria grow at a rate r, but a small fraction of the offspring, μ, mutate

to become resistant. These mutant bacteria also grow, and in the absence of other evidence that there is a link between resistance and growth rates, then LD assume that mutants also grow at a rate r. This model defines a linear dynamical system involving two population types S and m, the number of susceptible and mutant individuals in the population:

$$\frac{dS}{dt} = \overbrace{rS(1-\mu)}^{\text{growth of sensitives}} \tag{1.10}$$

$$\frac{dm}{dt} = \overbrace{\mu rS}^{\text{new mutations}} + \overbrace{rm}^{\text{growth of mutants}} \tag{1.11}$$

This model seems simple in many ways. However, it contains subtleties both in terms of the biology and in terms of the dynamical system itself. First, the model assumes that mutations occur during reproduction. Similar results could hold if mutations occur at any moment. Second, for biologists, writing equations in this way is not necessarily intuitive. In my experience, the instinct of most biologists when asked to translate a mechanism into a model is to think in terms of update rules, i.e., the value of the population at the next time, $x(t+1) = \ldots$, rather than in the changes in the population at the current time, dx/dt. Hence, if you share that instinct, consider taking a diversion to the technical appendix for how to move between the update perspective and the dynamical systems perspective. Finally, the text sometimes uses the notation dx/dt to denote the derivative of a population/variable with time, and sometimes uses the notation $\dot{x}$ to denote the same thing.

The next challenge is to solve this dynamical system to quantify both the sensitive and mutant cells as a function of time given the mutation rate μ. Try to solve by stepping away from the text with a blank piece of paper and keeping in mind only these two rules: (i) Susceptibles divide and sometimes generate mutants. (ii) Mutants also divide.

Now, if you tried and got stuck, keep in mind that there is a helpful trick: add the two derivatives together to find that $\dot{S} + \dot{m} = r(S+m)$. In other words, the entire population $N = S + m$ is growing at a rate r, while the balance of individuals shifts between S and m. The solution to this exponential growth problem for total cells is $N(t) = N_0 e^{rt}$. There is another observation: the mutants—despite being far more rare—are actually growing faster in a per capita sense than the residents. This is true even before selection was applied! Returning to the equations and solving the $\dot{S}$ equation yields the following (where $m(t) = N(t) - S(t)$):

$$S(t) = N_0 e^{r(1-\mu)t}, \tag{1.12}$$

$$m(t) = N_0 e^{rt}\left(1 - e^{-r\mu t}\right). \tag{1.13}$$

The length of the experiment is on the order of 10–20 generations, i.e., the product rt is a dimensionless number of that magnitude. Hence, we can approximate $e^{-r\mu t} \approx 1 - r\mu t$ given that $\mu \ll 1$, such that the number of mutants is predicted to grow faster than exponentially:

$$m(t) = N_0 e^{rt} \mu r t. \tag{1.14}$$

Note at this point that it is important to reconcile this finding with the objective of LD: to determine whether large fluctuations are consistent with mutations being independent of or

dependent on selection. The notion of consistency implies a particular experimental design in which LD performed a series of replicates—many, but certainly not infinite, a point that we will revisit later in the chapter. Hence, to begin to compare theory and experiment, it is worth generalizing this model to apply not only to *E. coli* and phage but to a larger class of problems. Thus far we have retained the growth rate r. A generalization is enabled by noting that the rate r and time over which the experiment is conducted t also appear together—suggesting that the growth rate is not a particularly important feature of the phenomena. The key is that r, as measured in inverse time, and t, as measured in time, must share the same units—hours, minutes, etc. If they do, then their product will remain the same even if we change units. That gives another clue. The value of r doesn't matter that much; it is the product of r and t that matters. If r is the effective inverse of the division period, then rt is simply a measure of the effective number of divisions in this growing population. Hence, it would seem that we would be better off developing a general theory, rather than one tuned for a particular value of r. The fact that LD could manipulate bacteria over dozens of generations reinforces their prudence in working with bacteria and not giraffes for this particular class of problems.

To formally work in this direction, denote a rescaled time, $\tau = rt$, such that $\mathrm{d}\tau = r\mathrm{d}t$. Hence, $\tau = 1$ is effectively one division, $\tau = 10$ is effectively 10 divisions, and so on. The dynamical equations can be written initially as

$$\frac{\mathrm{d}S}{\mathrm{d}t} = rS(1-\mu) \tag{1.15}$$

$$\frac{\mathrm{d}m}{\mathrm{d}t} = \mu rS + rm \tag{1.16}$$

while dividing both sides by r yields

$$\frac{\mathrm{d}S}{r\mathrm{d}t} = S(1-\mu) \tag{1.17}$$

$$\frac{\mathrm{d}m}{r\mathrm{d}t} = \mu S + m. \tag{1.18}$$

Next, replace $\mathrm{d}\tau = r\mathrm{d}t$, yielding

$$\frac{\mathrm{d}S}{\mathrm{d}\tau} = S(1-\mu) \tag{1.19}$$

$$\frac{\mathrm{d}m}{\mathrm{d}\tau} = \mu S + m \tag{1.20}$$

This last set of equations implies that, irrespective of the growth rate, the sensitive population will grow more slowly than that of the mutation population of resistant bacteria. Using the same logic as before, albeit forgoing the explicit inclusion of the growth rate, yields a prediction for the number of resistant mutants expected after a dimensionless time τ:

$$m(\tau) = N_0 e^{\tau} \mu\tau. \tag{1.21}$$

Therefore, at the final time, the number of mutants is expected to be

$$m(\tau_f) = N_f \mu\tau_f \tag{1.22}$$

where $N_f = N_0 e^{\tau_f}$ is the total number of bacteria exposed to viruses. This is, in modern terms, equivalent to Eq. (6) of LD's paper. This equation can be put into practice. Given an observation of the average number of mutants in replicate experiments, then it is possible to estimate the mutation rate:

$$\hat{\mu} = \frac{m_{obs}}{N_f \tau_f}. \tag{1.23}$$

There are two key caveats here. The first caveat is that this estimate of a mutation rate simply becomes an alternative estimate to that obtained assuming the acquired immunity hypothesis. It may be right, but the fact that we can make such an estimate does not provide the necessary evidence in favor of the hypothesis. The second caveat is that the approach to solving this problem is somewhat nonintuitive, i.e., involving mathematical tricks that tend to obscure the key biological drivers of the variation. Let's try another way, hopefully one that helps build intuition.

1.4.2 Spontaneous mutations—a cohort perspective

According to the spontaneous mutation hypothesis, mutants emerge in the growing bacterial culture before viruses are added. Hence, if a single resistant mutant appeared five generations before bacteria were mixed with viruses, then that single mutant would have given rise to $2^5 = 32$ new mutants, each of which corresponds to an observed, resistant colony on the agar plate. Likewise, if a single resistant mutant appeared seven generations before bacteria were mixed with viruses, then that single mutant would have given rise to $2^7 = 128$ new mutants. Hence, the older a mutant is, the more daughter cells appear in that lineage. Yet there is also a counterbalancing force. Given that the population is growing, it is far more likely that mutants will appear near the end of the experiment, even if those mutants have less time to reproduce. It is possible to formalize this by moving from nonoverlapping generations to continuous dynamics and by estimating the number of mutants in terms of cohorts, grouped by their age of first appearance.

To do so, it is essential to recognize that the rate of appearance of mutants is $\mu S(\tau)$. Hence, in a small interval of time $\mathrm{d}\tau$, a total of $\mu S(\tau)\mathrm{d}\tau$ mutants will emerge (at least on average). Each of these cohorts of new mutants will grow exponentially, reaching a final size $e^{\tau_f - \tau}$ greater by the end of the experiment. Hence, given that mutants can appear at any time, we can write

$$\begin{aligned}
m &= \int_0^{\tau_f} \mathrm{d}\tau \quad \overbrace{(\mu S(\tau))}^{\text{new mutant cohort}} \cdot \overbrace{e^{\tau_f - \tau}}^{\text{growth of mutant cohort}} \\
&= \int_0^{\tau_f} \mathrm{d}\tau \mu N_0 e^{(1-\mu)\tau} e^{\tau_f - \tau} \\
&= \int_0^{\tau_f} \mathrm{d}\tau \mu N_f e^{-\tau_f} e^{(1-\mu)\tau} e^{\tau_f - \tau} \\
&= \int_0^{\tau_f} \mathrm{d}\tau \mu N_f e^{-\mu\tau} \\
&= \left[-N_f e^{-\mu\tau}\right]_0^{\tau_f} \\
&= \left[-N_f \left(e^{-\mu\tau_f} - 1\right)\right]
\end{aligned} \tag{1.24}$$

and using the approximation $e^x \approx 1 + x$ for $|x| \ll 1$ yields

$$m = \mu \tau_f N_f, \tag{1.25}$$

precisely what was derived in the dynamical systems approach in the prior section! This is the same answer, but with less mathematical trickery and more intuition.

Now, of course, there is a problem. Working with continuous dynamics also risks unwittingly creating a "continuous fallacy." The continuous fallacy assumes implicitly that fractions of organisms can grow. But fractional organisms do not grow; they don't even exist! Yet the continuous model described above suggests they do (Figure 1.7). For example, what does it mean if at some point $m = 0.0001$? There is not one-ten-thousandth of a mutant proliferating in the flask before viruses are added. Hence, a deterministic model that assumes the growth of fractional organisms may pose problems when trying to compare results to experiments in which rare events matter. One way to address this issue would be to transform the model from a continuous framework to an entirely stochastic framework (in fact, later work did that (Lea and Coulson 1949)). That approach is one that is amenable to the use of a fully stochastic treatment as well as to computation—as realized through the homework problems recommended for this chapter. However, such analysis is mathematically far more difficult (Lea and Coulson 1949; Kessler and Levine 2013). Instead, another way to address this issue is to use a continuous model, albeit to incorporate the stochastic nature of the appearance of mutants by applying a deterministic growth model only to periods in which at least one mutant is likely to be present. That is the tack taken in the next section.

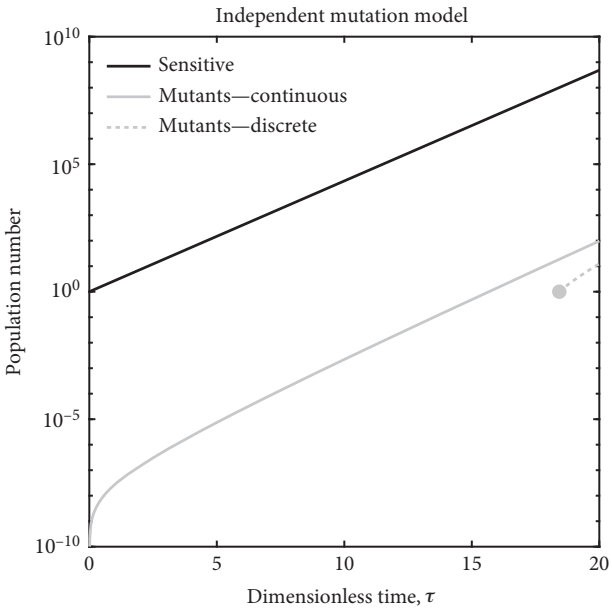


Figure 1.7: Contrasting dynamics of mutants given continuous dynamics and "discrete" approximations. The continuous model assumes that mutants are continuously generated, even at fractional levels. The discrete model assumes that mutants grow continuously only after a single first mutant appears. The variation in timing of the appearance of the first mutant underlies the "jackpot" effect described earlier. Dynamics are simulated assuming $N_0 = 1$ and $\mu = 10^{-8}$.

1.5 MODELING THE GROWTH OF (DISCRETE) MUTANTS

Understanding the implications of the independent mutation hypothesis requires building upon the continuous model while recognizing that mutants are discrete. For example, consider beginning an experiment with approximately 1000 bacteria founded by a single susceptible cell. If the mutation rate is of the order 10^{-8}, it would seem highly unlikely that one of those cells is resistant, indeed with odds on the order of 1/100,000. Yet, in the continuous model, the mutant population is immediately generated, albeit

fractionally, and allowed to grow. This may lead to an overestimate of the expected size of the mutant population, and, by extension, biases in estimating the actual mutation rate (see Figure 1.7).

Instead, to overcome the continuous fallacy, it is imperative to estimate the time τ_0 where the first mutant is likely to appear. In this model, rather than assuming that fractional mutants grow, we expect there should not be mutants, i.e., $m(\tau) = 0$, for $\tau < \tau_0$ and otherwise:

$$S(\tau) = N_{\tau_0} e^{(1-\mu)(\tau-\tau_0)} \tag{1.26}$$

$$m(\tau) = N_{\tau_0} e^{(\tau-\tau_0)} \mu\,(\tau - \tau_0) \tag{1.27}$$

Connecting theory and experiments requires estimating the realized number of mutants at the end of the experiment, i.e., when $\tau = \tau_f$ and given a final number of cells $N_f = N_{\tau_0} e^{(\tau_f-\tau_0)}$. Hence, we can write

$$m(\tau_f) = N_f \mu\,(\tau_f - \tau_0) \tag{1.28}$$

Now recall also that $\tau_f - \tau_0 = \log(N_f/N_{\tau_0})$, which is a feature of the exponential growth of cells. It would seem that we are nearly there in terms of incorporating the discrete nature of mutations in the estimation procedure for μ. Altogether, the experiment yields an observed number of mutants m, a total number of bacteria N_f, as well as the duration of growth τ_f. If we only knew the approximate time at which resistant mutants appear, we would be able to also estimate μ. That time is related to N_{τ_0}. This is where the final puzzle is solved.

If one in a million offspring yielded a resistant bacteria, then one would expect to wait until there were on the order of one million bacteria before finding a mutant. In other words, the time of the first mutant appearance should satisfy $N_{\tau_0}\mu \approx 1$ or alternatively that $N_{\tau_0} \approx 1/\mu$—this is the circular gray point noted in the demonstration example in Figure 1.7. The time (on the x axis), τ, of this point corresponds to the moment when it is likely that a mutant first appears. The number of mutants (on the y axis) is set to 1. The smaller μ is, the larger the population must get before the first mutant appears, and therefore there is less time for this clonal population of mutants to grow exponentially. Substituting this time yields a new estimate of the expected number of mutants at the end of the experiment:

$$m(\tau_f) = N_f \mu \log(N_f \mu). \tag{1.29}$$

Note that if there are C multiple replicates, then the first time a mutant would appear in one of the replicates would be of the order $1/(C\mu)$ such that

$$m(\tau_f) = N_f \mu \log(C N_f \mu). \tag{1.30}$$

Eq. (1.30) can be put into practice. Given an observed average number of mutants m as well as the number of replicates C and population size N_f, it can be used to identify a unique value of μ. This equation is implicit. Nonetheless, it can be "inverted" so as to solve the problem numerically. But even if we have an estimate, this doesn't answer the deeper question: is there sufficient evidence to accept the independent mutation hypothesis and reject the hypothesis that mutations are dependent on selection?

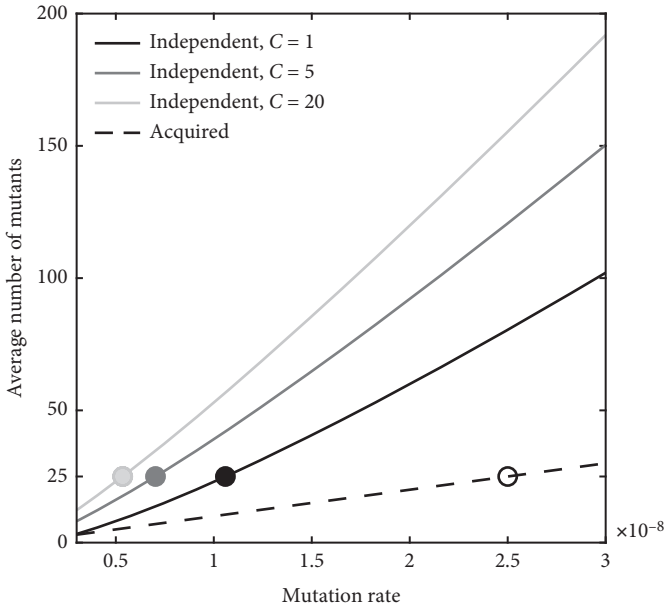


Figure 1.8: The number of expected mutants as a function of unknown mutation size. Here, Eq. (1.30) and Eq. (1.31) are used for the independent and acquired hypotheses, respectively. For the independent hypothesis, the average number of mutants increases logarithmically with C, the number of replicate cultures. Hence, given an observation, it is possible to "invert" the curves and find an estimate of μ_f or μ_a given an observation of m, and the values N_f and C. Here $N_f = 10^9$. The circles denote estimated mutation rates given an observation of 25 for the average number of mutants, solved using a nonlinear zero finding method for the independent mutation case.

Favoring one hypothesis over another requires not just alternative estimates but evidence of the incompatibility of one hypothesis to explain the observed data. Thus far, the theory presented here only utilizes the mean number of colonies to provide two alternative estimates of the mutation rate. Eq. (1.30) links data to an estimate of the mutation rate when mutations are independent of selection. Yet we already derived an alternative equation for the estimated mutation rate when mutations are dependent on selection:

$$m(\tau_f) = N_f \mu_a. \tag{1.31}$$

Figure 1.8 shows the expected number of mutants as a function of μ for three cases of C for the independent mutation hypothesis in contrast to the expected number of mutants in the acquired mutation hypothesis. The same figure also shows how the estimate of μ varies with C and the underlying mutational mechanism given the same observation. That is, if 25 resistant colonies were observed on average, then using the mean information alone would simply lead to distinct estimates of the mutation rate but would not be sufficient to distinguish between the two classes of hypotheses. One measurement, multiple estimates. Distinguishing them requires going beyond means, all the way to the variation.

1.6 VARIANCE OF MUTANTS WHEN MUTATIONS ARE INDEPENDENT OF SELECTION

How much variation is expected among resistant colonies if mutations arise spontaneously, independent of selection? We have already shown that mutants arise in different cohorts, e.g,. from time τ_0 to τ_f. Earlier cohorts may be less likely to arise, but when they do, they lead to a larger number of mutants. Later cohorts are more likely to arise, and when they do, they lead to a smaller number of mutants. Altogether, these cohorts contribute to the expected variation in outcomes across replicate experiments. For example, if there were only mutants at τ_0 and at some other point τ_1, then the total variance in the number of mutants would be

$$\mathrm{Var}(m) = \mathrm{Var}(m|\tau_0) + \mathrm{Var}(m|\tau_1), \tag{1.32}$$

which is to say that expected *variances add!* There are two contributions to the variance of a cohort. First, the number of new mutants generated in a given generation is itself a Poisson random number whose expected value is $N(\tau)\mu$. However, this Poisson random number is multiplied by an exponential factor, corresponding to the proliferation of the cohort. If $x \sim \mathrm{Poisson}(N, \mu)$, then the variance of that random variable multiplied by a constant factor is $\mathrm{Var}(\alpha x) = \alpha^2 \mathrm{Var}(x)$ where α is a constant. In other words, if the cohort grows by a factor of 16, its mean increases that much, but the variance (involving squared values) goes up by a factor of 256! This hints at the possibility that variation in the emergence of early mutants in a growing population before exposure to viruses could underlie the large variation in observed outcomes.

It is possible to assess the variance expected in outcomes by focusing on the case where there are only two potential times when mutants arise:

$$\mathrm{Var}(m) = \left(e^{\tau_f - \tau_0}\right)^2 \mu N(\tau_0) + \left(e^{\tau_f - \tau_1}\right)^2 \mu N(\tau_1). \tag{1.33}$$

However, recall that the numbers of cells are themselves growing exponentially, such that $N(\tau_i) = N_f e^{-(\tau_f - \tau_0)}$, and so

$$\mathrm{Var}(m) = \mu N_f \left[e^{\tau_f - \tau_0} + e^{\tau_f - \tau_1}\right]. \tag{1.34}$$

We can generalize this idea to any value of τ_i between τ_0 and τ_f, i.e., moving to the continuum limit, such that

$$\begin{aligned} \mathrm{Var}(m) &= \mu N_f \int_{\tau_0}^{\tau_f} \mathrm{d}\tau e^{\tau_f - \tau} \\ &= \mu N_f \left[e^{\tau_f - \tau_0} - 1\right] \end{aligned} \tag{1.35}$$

for which we should recall that $\tau_f - \tau_0 \sim \log C\mu N_f$. Finally, we can write

$$\mathrm{Var}(m) = \mu N_f \left(C\mu N_f - 1\right) \approx C\left(\mu N_f\right)^2. \tag{1.36}$$

This equation implies that the variance grows faster than the mean, unlike in the case of mutations that are dependent on selection.

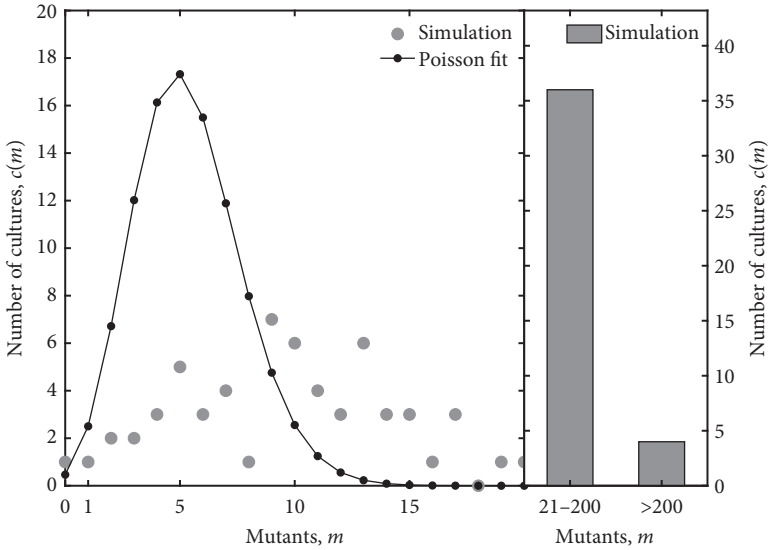


Figure 1.9: Comparison of the expected distribution of mutants assuming the acquired mutation hypothesis to the realized number of mutations given an independent mutation hypothesis. The black solid line denotes the Poisson fit assuming mutation depends on selection. The gray circles denote the results of an LD simulation, with final population size of $\sim 5.4 \times 10^8$ given $\mu = 10^{-8}$. The right panel denotes the large number of jackpots, including four cases where there are far more than 200 mutants in a single experiment out of 100 experiments. As is apparent, such jackpots are wholly unexpected given the Poisson assumption that would arise if mutations were dependent on selection.

Table 1.2: Hallmark features of the acquired mutation hypothesis and spontaneous resistance hypotheses

	Acquired	**Spontaneous continuous**	**Spontaneous discrete**
Mean	$\mu_a N_f$	$\mu N_f \log N_f$	$\mu N_f \log C\mu N_f$
Variance	$\mu_a N_f$	μN_f^2	$C\mu^2 N_f^2$
$\frac{\text{Variance}}{\text{Mean}}$	1	$\frac{N_f}{\log N_f}$	$\frac{C\mu N_f}{\log C\mu N_f}$

Note: The terms *continuous* and *discrete* refer to whether mutant cohorts are assumed to begin at $\tau = 0$ or $\tau = \tau_0$; see text for details. As is apparent, the independent case leads to variance:mean ratios far above 1.

Summarizing these findings requires a focus on the qualitative differences implied by the link between the expected variance and the mean number of resistant colonies. Table 1.2 compares and contrasts the mean, variance, and ratio of variance to mean for both hypotheses. The variance as estimated for the case of mutations independent of selection is $C\left(\mu N_f\right)^2$, whereas the mean is $\mu N_f \log C\mu N_f$ such that the ratio is

$$\frac{\text{Var}}{\text{Mean}} = \frac{C\mu N_f}{\log C\mu N_f}. \tag{1.37}$$

This ratio includes a relatively large number over its log, which should yield a ratio much larger than 1. In the case of mutations that are independent of selection, the bulk of the

variation stems from the very earliest of mutant cohorts, because when they do occur, they grow exponentially, leading to jackpots and large variation. These large jackpots are incompatible with the acquired immunity hypothesis (Figure 1.9). The finding and interpretation of large variances in repeated experiments of phage lysis of bacteria has remained a salient example of the integration of quantitative reasoning of a living system given uncertainty—and such work continues to inspire. As but one example, the interested reader may want to explore recent work showing how delays between the onset of a mutation and change in phenotype impact the population-level mutant distribution (Sun et al. 2018).

It is now up to you to work computationally to help build your intuition as to whether the fluctuations observed are large enough to reject the mechanism that mutations are dependent on selection in favor of the mechanism that mutations are independent of selection. In doing so, a full stochastic framework is used in the homework problems to test the limits of both the simple models and the theoretical predictions—which come with a caveat. The scaling presented in the last column of Table 1.2 differs from the scaling found in a fully stochastic treatment (Lea and Coulson 1949; Zheng 1999). The reason is that correcting for the appearance of the first mutant applies to the sample statistics with a finite number of replicates and not necessarily to the expected mean and variance in the limit of infinite replicates. This claim is equivalent to considering the limit that $CN_0\mu \ll 1$ or equivalently that $CN_0 \ll \frac{1}{\mu}$, i.e., a mutant is unlikely to have already been present at the start of the experiment. As C gets very large then it is more and more likely that a mutant will be present at the very start of one of the C replicates—and the continuous fallacy stops being a fallacy. Problem 6 in this chapter explores how the mean and variance increase with mutation rate given sufficiently small C. Although the quantitative details differ, the fluctuations remain large.

1.7 ON (IN)DIRECT INFERENCE

This chapter has explored a key concept in modern biology. The work of LD is particularly notable for its integration of mathematical theory, physical intuition, and model-data integration as a means to understand the nature of mutation. The conceptual notion of spontaneous mutation versus acquired hereditary immunity can be seen in the schematic in Figure 1.10. As is apparent, the possibility for jackpots is enhanced when lineages (i.e., a bacterium and its descendants) all have the property of resistance. This possibility of jackpots is to be expected when mutations arise spontaneously during the growth process and are unrelated to the selection pressure. Hence, irreproducibility is a hallmark of a particular biological mechanism. The work of LD showed that mutations arose independent of selection and were not acquired as a result of interaction with a selective pressure. Their 1943 paper and its findings were cited when Luria and Delbrück received their Nobel Prize in Physiology or Medicine along with Alfred Hershey in 1969.

We now accept this paper as having established the independence of mutation from selection. It informs not just foundational work but interpretation of the emergence of the frequency and variation in cancer cells (Fidler and Kripke 1977). Yet it took a decade for the "biometric" approach of Luria and Delbrück (sensu Esther Lederberg and Joshua Lederberg) to be accepted. The acceptance was not just because of a gradual increase in

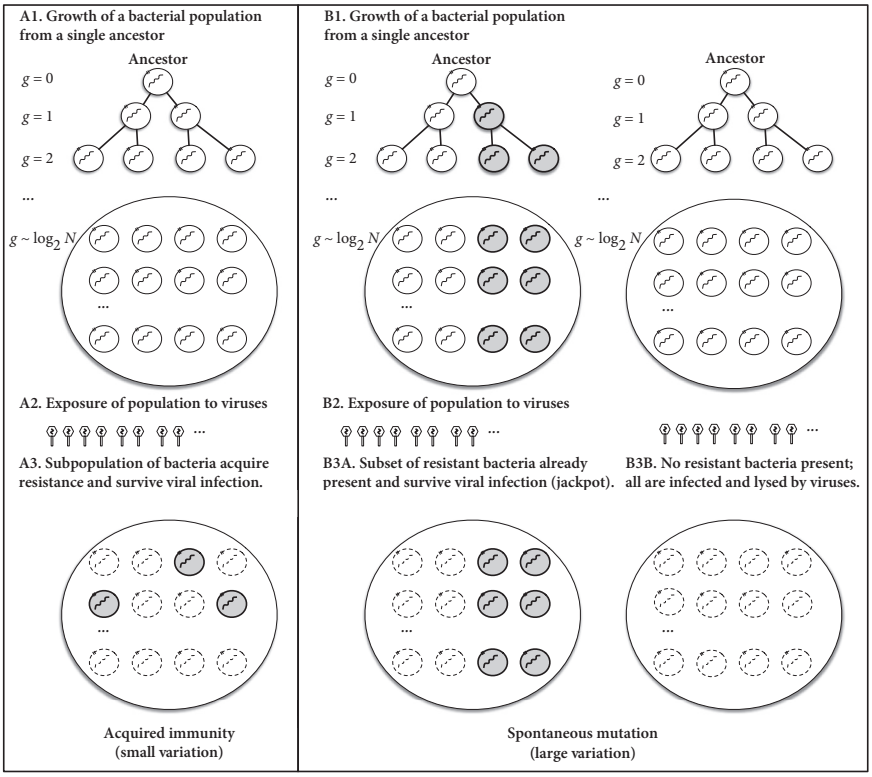


Figure 1.10: Schematic illustration of the acquired heritable immunity mechanism (left) and the spontaneous mutation mechanism (right), including differences in the number of resistant colonies—adapted from J. S. Weitz, *Quantitative Viral Ecology* (with permission) (Weitz 2015).

quantitative rigor in cellular and molecular biology. Indeed, the subject of understanding the basis for the "Luria-Delbrück" distribution continues even now (Kessler and Levine 2013). Instead, fellow researchers were eventually convinced by the dissemination of the elegant replica plating method of Esther Lederberg and Joshua Lederberg (1952), likely even more so than by beautiful mathematics (Figure 1.11; note that Esther Lederberg has been underappreciated for the scope of her contributions (see Schaechter (2014)). The idea of the replica plating method is that a bacterial lawn, likely with preexisting resistant mutants, can be transferred to multiple plates. The transfer is meant to preserve the existing spatial structure of bacteria, including bacterial mutants. These multiple replicate plates are then exposed to a phage lysate. Hence, if the position of the resistant mutants in each replicate plate were similar, that would show—visually—that the property of phage resistance was already present before the interaction with the virus. The numbered colonies in Figure 1.11 demonstrate this very point—many appear in exactly the same position in at least two plates—and are an example of how much a beautiful experiment design can offer.

Despite this chapter's singular focus on evidence building toward a conclusion that mutations are independent of selection, there is a caveat to this seminal discovery.

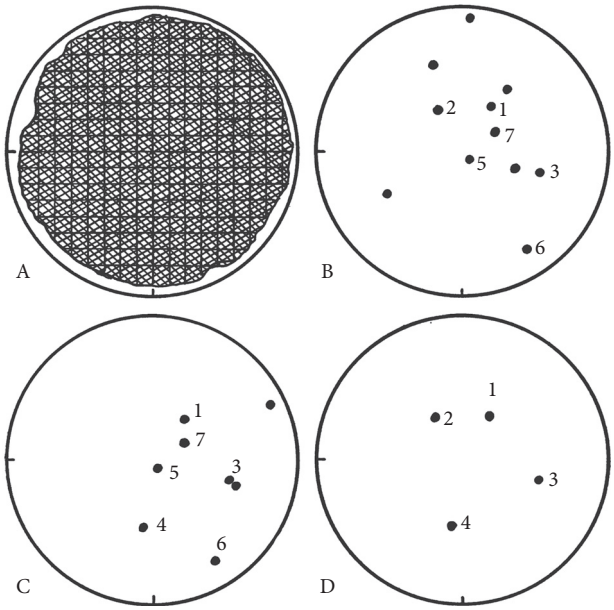


Figure 1.11: Replica plating method to demonstrate viral-resistant mutants are independent of selection. The original plate is shown in A. Colonies' cells resistant to phage T1 are numbered in replica plates B, C, and D. As noted in the original caption, the colocation of resistant colonies in the same location implies they are "derived from small clones of resistant mutants already present at corresponding sites on the plain agar plate, A." Reproduced from Figure 2 of Lederberg and Lederberg (1952).

This caveat is shaped by new research into the origins of genetic variation in microbes. To understand the caveat, it is worth considering the Gedankën experiment: what would have happened to the history of molecular biology if Luria and Delbrück had used *Streptococcus thermophilus* and its phage rather than *E. coli B* and phage T1? The *S. thermophilus* strain utilizes an acquired immune defense system known as CRISPR (clustered regularly interspaced palindromic repeats) or CRISPR-Cas (Barrangou et al. 2007; Makarova et al. 2011). Although CRISPR-Cas is known as the basis for a revolution in genome engineering and biotechnology, at its heart the CRISPR system is a de facto adaptive immune system in bacteria and archaea that enables microbes that survive an infection to become heritably resistant. These microorganisms seem, in some sense, akin to giraffes reaching for acacia leaves and passing on their longer necks to their offspring (Koonin and Wolf 2009). A strange world, but it is the one we live in.

1.8 TAKE-HOME MESSAGES

- Mutations are the generative driver of variation in the evolutionary process.
- Prior to the work of Luria and Delbrück, there was a major unanswered question: are mutations independent of or dependent on selection? Afterward, the consensus shifted: mutations are independent of selection.
- Experimental evidence using phage and bacteria showed that the number of mutational events varied significantly between experiments.
- This large variation, i.e., a lack of reproducibility, was a key hallmark of the independent mutation hypothesis, and counter to predictions of the acquired mutation hypothesis.
- Solving for the Luria and Delbrück distribution is non-trivial; nonetheless, the central concepts of proliferation and mutations among clones is readily analyzed, simulated, and compared to data.
- Although exceptions abound (including CRISPR-Cas immunity), the concept that mutations are independent of selection remains the paradigm in biology.

1.9 HOMEWORK PROBLEMS

A central goal of this book is to help readers develop practical skills to quantitatively reason about living systems given uncertainty. However, each chapter is only part of this process (just like listening to lectures, paper discussions, and in-class work help solidify

understanding). Moreover, for many readers, the mathematical and biological insights provide a partial guide. If seeing is believing, then coding and simulation are a central path to build intuition and insight on the themes developed in this and subsequent chapters. The following homework operates in that spirit and is best approached after working through the exercises in the accompanying computational lab guide. The laboratory guides—in MATLAB, Python, and R—provide insights into how to:

- Sample from random distributions
- Utilize the properties of uniform random distributions to generate random distributions that are nonuniform, e.g., exponential distribution
- Compare and contrast the Poisson with the binomial distribution
- Develop stochastic simulations of growing populations

The homework helps leverage this "toolkit" in order to build intuition on the core ideas of Luria and Delbrück's seminal paper.

The overall objective of these problems is to reproduce the "irreproducibility" of the number of resistant mutants, as observed by LD, and to begin to reach tentative conclusions regarding the confidence on estimated mutation rates and mechanisms in inferring the basis of mutation from resistant colony data. The problems utilize a common set of assumptions initially. That is, in these problems, consider an experiment with C cultures, each of which has N sensitive cells. Every time a cell divides there is a probability μ that one, and only one, of the daughter cells mutates to a resistant form. We assume that the offspring of resistant cells are also resistant, i.e., there are no "back-mutations." Good luck. And remember, you can do it!

PROBLEM 1. Simulating the Luria-Delbrück Experiment over One Generation

Write a program to simulate just one generation of the LD experiment—stochastically. Simulate $C = 500$ cultures, each of which has $N = 1000$ cells and $\mu = 10^{-3}$, i.e., a very high mutation rate. What is the distribution of resistant mutants that you observe across all the cultures? Are they similar or dissimilar to each other? Specify your measurement of $c(m)$, i.e., the number of cultures with m resistant mutants. Is this distribution well fit by a Poisson distribution? If so, what is the best fit shape parameter of the Poisson density function and how does that relate to the microscopic value of the mutation you used to generate the output? Finally, to what extent are the fluctuations "large" or "small"?

PROBLEM 2. Simulating the Luria-Delbrück Experiment Forward One Generation at a Time

Extend the program in Problem 1 by setting $C = 1000$, $N = 400$, and $\mu = 10^{-7}$, while having the population grow over $g = 15$ generations. What choice did you make with respect to modeling the population? If you decided to model each individual cell in each individual culture, explain your rationale. Next, develop a model that represents

the emergence of new resistant cells in each generation in each culture en masse (that is, all at once). (Hint: Think about how prudent use of the Poisson random generation function could help.) The objective here is to develop a working simulation that is both accurate and efficient—in doing so, compare the speed when you use Poisson versus binomial random number generating functions.

PROBLEM 3. Characterizing the Mutant Distribution

Using the simulation in Problem 2, report and describe the shape of $c(m)$—the number of cultures with m resistant mutants. Describe and characterize the shape of this distribution and contrast it to that expected under the acquired hypothesis. What is the mean and variance? Are fluctuations large or small? Finally, to what extent are there jackpot cultures? Can you define a principled way to identify those jackpots? To what extent do fluctuations become small when jackpots are excluded?

PROBLEM 4. Why Are There Jackpots?

Characterize the "age" of each mutant using your LD simulation. That is, if you have not already done so, keep track of the first appearance of each mutant and characterize the relative importance of mutants of different "ages" to the total number at the end. Do early/late mutants contribute disproportionately to the total number? Do early/late mutants more strongly influence variation?

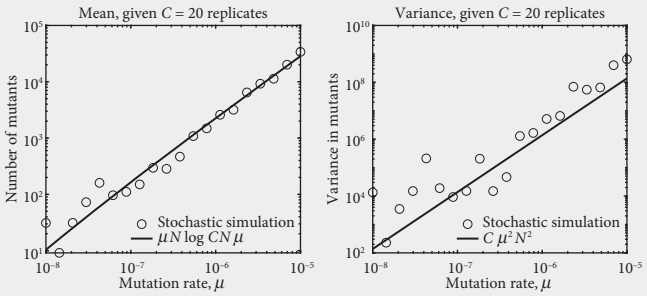
PROBLEM 5. Moving Backward like Luria and Delbrück from Observations to Estimates

Take the results from the first 100 of your experiments in Problem 2, that is, the number of resistant mutants, $m_1, m_2, \ldots, m_{100}$. Treat these 100 numbers as data. Now write a program that leverages results of new simulations to infer the most likely value of μ, the mutation rate, and (if possible) a 95% confidence interval for it—in doing so, pretend that you do not know what μ is in advance. Consider using two pieces of evidence: (i) the number of cultures with no resistant mutants; (ii) the mean number of resistant mutants. Finally, ask: is your data also consistent with the acquired resistance hypothesis? Why or why not?

PROBLEM 6. Scaling of the Mean and Variance of Mutants with μ

Using a stochastic simulation of the LD problem (ideally, using a Poisson approximation to the mutant generation at each step), simulate the expansion of a bacterial population with zero mutants at $t = 0$ and $N_0 = 1000$ over 18 generations so that the final

population has more than two billion cells. Upon division of a wild-type cell, there is a μ probability that one of the daughter cells will be a mutant—both mutant and wild-type cells divide each generation. Modulate the mutation rate from $\mu = 10^{-8}$ to $\mu = 10^{-5}$ and use a value of $C = 20$. If your model works, it should look like the following:



Using your model, explore lower and higher values of C, e.g., $C = 10$ and $C = 40$. Compare and contrast your findings with the scaling in Table 1.2, accounting for the emergence of mutants after the start of the experiment. Using these same settings, compare the LD estimates, assuming mutants grow deterministically from the start. In discussing your results, provide a rationale for gaps between theory and simulation.

1.10 TECHNICAL APPENDIX

One of the key features of a compelling theory and model is that it is often easy to understand in retrospect but hard to complete in the absence of the solution. Such theories are like puzzles and any good solver has a repertoire of techniques. This appendix reviews basic techniques and provides additional information to help fill in the gaps in the main text. These reviews may be of particular value to readers with strong biological training who may not have seen these mathematical methods used in concert with the analysis of biological problems. Indeed, why remember how to do a Taylor expansion or calculate the binomial coefficient if it is never called upon in your research?

Factorials The term $N!$ denotes a factorial, or $N \times (N-1) \times (N-2) \times \cdots \times 2 \times 1$.

Binomial coefficients The binomial coefficient counts the number of ways to permute the location of k events of N trials. Formally, the coefficient can be written as $\binom{N}{k}$ such that

$$\binom{N}{k} \equiv \frac{N!}{(N-k)!k!}. \tag{1.38}$$

For example, consider the case where $N = 5$ and $k = 2$, for which there are $5!/(3! \times 2!) = 10$ distinct combinations of arranging two positive events out of 5 trials. The binomial coefficient can be understood as follows. First, consider the way in which k events can be selected out of N trials in a particular sequence. The first event has N options, the second $N-1$, the third $N-2$, and so on, so that the total number of events

is $N(N-1)(N-2)\cdots(N-(k-1))$. However, the resulting placement overcounts the number of unique configurations; e.g., for $N=5$ and $k=2$, it is possible to first select trial 1 and then trial 2 or to first select trial 2 and then trial 1. The degree of overcounting can be calculated by counting the combinations of the identity of the positive events; i.e., the first event has k options, the second has $k-1$, and so on. That is to say, there are $k!$ ways to permute the identity of the positive events. Hence, the total number of ways to permute the location of k positive events in N trials is

$$\frac{N(N-1)(N-2)\ldots(N-(k-1))}{k!} = \frac{N!}{(N-k)!k!}, \tag{1.39}$$

precisely the binomial coefficient listed above.

From the binomial to Poisson Consider the probability that k events take place, each with probability μ, out of N trials, i.e,

$$p(k|N,\mu) = \frac{N!}{(N-k)!k!}\mu_a^k(1-\mu_a)^{(N-k)}. \tag{1.40}$$

In the Luria experiments, there are many bacteria ($N \gg 1$) and the probability of mutation is rare, $\mu \ll 1$. Denote the average number of events as $\hat{m} = N\mu$. In the limit of large bacterial populations and rare acquisition of virus resistance, then

$$\begin{aligned} p(k|N,\mu) &= \frac{N(N-1)(N-2)\cdots(N-(k-1))}{k!}\mu^k(1-\mu)^{(N-k)} \\ &= \frac{N^k\mu^k(1-1/N)(1-2/N)\cdots(1-(k-1)/N)}{k!}(1-\mu)^{(N-k)} \\ &\approx \frac{\hat{m}^k}{k!}\left(1-\left(\sum_{i=0}^{k-1} i\right)/N\right)e^{-\hat{m}}(1+\mu k) \\ &\approx \frac{\hat{m}^k e^{-\hat{m}}}{k!} \end{aligned} \tag{1.41}$$

Note that this approximation utilizes a Taylor expansion of exponentials, i.e., $e^{-x} \approx 1-x$ and $e^x \approx 1+x$ for small values of x, and also ignores small terms, e.g., those in which a number is divided by N or multiplied by a factor of μ. The resulting Eq. (1.41) corresponds to the Poisson distribution with an expected value of $\hat{m}$ for the number of successful events given N trials.

Taylor expansion A Taylor expansion represents a formal mechanism to approximate the values of arbitrary functions near a reference point using a sequence of polynomials. It is easiest to illustrate the concept of Taylor expansions for functions of one variable, like $f(x)$, but the concepts can be extended to multiple dimensions, e.g., $g(x,y)$. In the one-dimensional case, the value of $f(x)$ near the reference point x_0 can be approximated as

$$f(x) \approx f(x_0) + \frac{\mathrm{d}f}{\mathrm{d}x}\Big|_{x=x_0}(x-x_0) + \frac{\mathrm{d}^2\mathrm{f}}{\mathrm{d}x^2}\Big|_{x=x_0}(x-x_0)^2 + \cdots. \tag{1.42}$$

In essence, the Taylor expansion assumes that function values near a reference point can be approximated by starting at the function value of the reference and then fitting a tangent to it. The slope of the tangent defines the "rise" of the function, i.e., $\mathrm{d}f/\mathrm{d}x$ evaluated at $x = x_0$, and the "run" is the difference between the point of interest and the reference, i.e., $(x - x_0)$. The linear approximation can be extended using the curvature—the second derivative—multiplied by the difference squared, i.e., $(x - x_0)^2$. As a result, the Taylor expansion approximates arbitrary functions as a combination of polynomials. In practice, we will rarely utilize Taylor expansions other than to first (i.e., linear) or second (i.e., quadratic) order. For example, e^x for values of x near 0 can be represented as $1 + x + x^2/2 + \cdots$. Similarly, $\log(1 + x)$ for values of x near 0 can be represented as $x - x^2/2 + \cdots$. Note that retaining higher-order terms can sometimes prove essential (as will be shown in later chapters).

Solving the exponential growth equation Consider the dynamical system

$$\frac{\mathrm{d}x}{\mathrm{d}t} = rx. \tag{1.43}$$

This exponential growth equation can be solved, first by dividing both sides by x and multiplying both sides by $\mathrm{d}t$, and then by integrating:

$$\int \frac{\mathrm{d}x}{x} = \int r\mathrm{d}t$$

such that

$$\log x = rt + C \tag{1.44}$$

where C is a constant of integration. Exponentiating both sides leads to the relationship $x(t) = \tilde{C}e^{rt}$ where $\tilde{C} = e^C$. Given the initial conditions, $x(t = 0) = x_0$, then one can write the complete solutions $x(t) = x_0 e^{rt}$.

Maximum likelihood of the Poisson distribution Consider the Poisson distribution, with expected mean $\lambda = \mu N$, such that

$$p(m|N, \lambda) = \frac{\lambda^m e^{-\lambda}}{m!}. \tag{1.45}$$

The most likely value of λ given an observation of m corresponds to a *maximum* in p. To find such a maximum, first take a derivative with respect to λ, set the derivative to zero, and then double-check that the second derivative at this point is negative (which it will be). The derivative of Eq. (1.45) yields

$$\frac{m\lambda^{m-1}e^{-\lambda}}{m!} = \frac{\lambda^m e^{-\lambda}}{m!}. \tag{1.46}$$

Canceling factors on both sides yields

$$\frac{m}{\lambda} = 1. \tag{1.47}$$

Recall that $\lambda = \mu N$, so that if $m = m_{obs}$, then

$$\hat{\mu} = \frac{m_{obs}}{N} \tag{1.48}$$

where $\hat{\mu}$ is the most likely estimate of the mutation rate. This maximum likelihood estimate of the mutation rate is the ratio of the observed number of mutants to the total number of cells. Technically, one should double-check that this value is the maximum likelihood—it is.

Law of total probability Consider the joint probability of two variables, $p(x, y)$. This joint probability can be written in two ways: $p(x|y)p(y)$ or $p(y|x)p(x)$. The equivalence of these two forms is the basis for Bayes' rule.

From updates to derivatives Consider a population x, which experiences changes due to interactions with other populations and environmental factors. Denote the rate of change as $f(x)$ such that the rate of change depends on parameters but also on the population level x. How can we build a model that describes the trajectory of the population over time? To do so, consider changes that occur over some interval Δt. In that case,

$$\overbrace{x(t+\Delta t)}^{\text{next value}} = \overbrace{x(t)}^{\text{current value}} + \overbrace{f(x)\times\Delta t}^{\text{increment}}. \tag{1.49}$$

Yet the value $x(t+\Delta t)$ can be thought of as a function, the value of x changing over time. A Taylor expansion can be used to approximate the future value of this function given the current value, i.e.,

$$x(t+\Delta t) \approx x(t) + \left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)\Delta t. \tag{1.50}$$

Substituting in this expansion to the update rule yields

$$x(t) + \left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)\Delta t = x(t) + f(x)\Delta t, \tag{1.51}$$

which after canceling terms yields

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x). \tag{1.52}$$

Hence, the continuous rate of change of a population can be expressed in terms of a series of small increments. The usual way to obtain the dynamical systems representation is to note that the derivative is the limit of the finite differences of x, i.e.,

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \lim_{\Delta t\to 0}\frac{x(t+\Delta t) - x(t)}{\Delta t}. \tag{1.53}$$

Nonetheless, it has been my experience in teaching that biologists prefer to begin with the concept of an update rule and reduce to the continuous limit from there, rather than formally beginning the other way around.

INDEX