The Luria-Delbruck experiment

Imagine doing an experiment a large number of times, and finding a completely different outcome on every run. How would you respond to such scenario? Most likely you will discard the data, and attempt to improve the experimental setup – perhaps attempt to get a more precise measurement setup, or try to ensure that the initial conditions are identical in all runs. In this chapter, we will describe a classical experiment in biology, where this happened. Specifically, the standard deviation of the measurement far exceeded the mean. Instead of being a nuisance or an unwanted result, this observation by itself led to the most important results of the experiments. Delbruck and Luria were examining how many bacteria in a culture survive a viral attack. The combined work of Luria – who did the experiment – and the modeling work of Delbruck, led the pair to an understanding that it is mutations (rather than adaptation) that provide bacteria with resistance to bacteriophages, a finding for which they were awarded the Nobel prize. You are encouraged to read their original paper on this from 1943 [1], which is a beautiful case of modeling and simple maths leading to profound results.

1 The experimental setup and results

The experiment done by Luria is simple to define: we start with a few *E. coli* bacteria (50-500 in the original experiment), and let them grow. They reproduce by asexual reproduction, with each one growing and dividing into two daughter cells within about $\tau_d = 20$ minutes, i.e., at time t the number of cells will be:

$$N(t) = N_0 \cdot 2^{t/\tau_d}. \tag{1}$$

After getting about 10^9 cells, Luria exposed the bacteria to a virus (a bacteriophage), that killed nearly all of the cells. Nevertheless, several bacteria survived in most of the experiments. By plating the cells and counting the number of individual colonies that emerged, he could know approximately how many cells managed to survive the viral attack. This number was the main result of his experiment. Annoyingly, this number varied enormously from experiment to experiment, no matter how hard Luria tried to control things! See Fig. 1, taken from their original paper.

We also note that once the cells developed resistance to the virus, all of their offspring would also be resistant: to prove this, Luria showed that exposing the progeny of cells from the surviving colonies to the virus does not kill them.

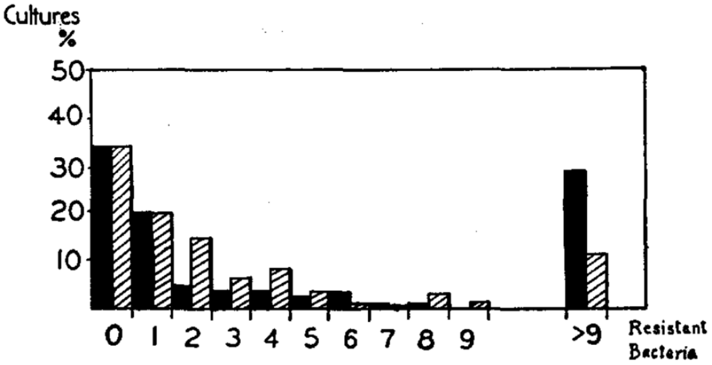

Figure 1: Histogram of the number of survivors in the experiment (black bars) showing the incredibly high variability. The gray dashed boxes are the results of the theoretical model that elucidates this behavior – based on the assumption of randomly occurring mutations.

2 What's going on?

As of 1943, there were two possible hypotheses as to the mechanism through which bacteria develop resistance to viruses: in the first, "Lamarckian" approach, the cells attempt to adapt to the virus, after exposure. The majority do not succeed in adapting, and die, but a small fraction adapts and survives. Note that in this scenario, the cells only start to adapt in the *final* generation, once they are exposed to the virus.

In the second hypothesis, resistance comes from a beneficial mutation that provides the cell with immunity against the virus. In this case, if the mutation happened early on in the experiment, all of the progeny of that cell will also be resistant.

In the next section we will show that if the Lamarckian adaptation hypothesis is correct, the probability distribution of the number of surviving cells approximately follows a Poisson distribution – since every cell has some finite probability p to adapt, and the adaptation of every cell is an independent process. We will also show that the variance to mean ratio of the number of survivors over many trials would be 1, and hence this model cannot account for the high variance:mean ratio observed experimentally. Because of this, the adaptation hypothesis can be ruled out!

In the mutation hypothesis, the crucial thing to note is that the final survivor population is exponentially sensitive to the generation in which the mutation occurred - in the unlikely event that it happened early on, we will have a huge number of resistent cells at the end; while if it happened very late in the experiment only a handful of cells will survive. Therefore the distribution of the number of survivors is very broad, with an enormous variance:mean ratio, which we will quantify in the next section.

Together, the modeling of Delbruck and Luria allowed them to rule out the adaptation hypothesis and accept the mutation hypothesis. Remarkably, it is the stochasticity and fluctuations in the results of these experiments which proved to be the “smoking gun” of the model. Only via their quantitative analysis could they support their biologically relevant conclusion.

Note: also today the question of how bacteria develop resistance (to antibiotics rather than bacteriophages) is highly relevant to society, with exciting ongoing research unravelling some surprising strategies used by the cells.

3 Quantitative analysis

Adaptation Consider first the adaptation model. A single cell present when the virus is applied either adapts and survives with some small probability p (hence contributing $x_i = 1$ to the number of surviving cells) or dies (and contributes $x_i = 0$ to the total number of surviving cells). This is also known as a Bernoulli process, or coin-flipping. Therefore, to find the probability of having X survivors out of the entire population of N cells, we have to consider N “coin-flips” with a probability of success p for every flip. Hence:

$$P(X) = \binom{N}{X} p^X (1-p)^{N-X}, \tag{2}$$

i.e., it is a binomial distribution. When p is very small, which is the case in the Delbruck-Luria experiments, we can approximate the distribution for $X \ll N$ as:

$$P(X) \approx \frac{1}{X!} N^X p^X (1-p)^N, \tag{3}$$

where we used $\binom{N}{X} = N(N-1)...(N-X+1)/X! \approx N^X/X!$. Note that using the Taylor expansion of $\log(1+\epsilon)$, we have:

$$N \log(1 - X/N) \approx -X \rightarrow (1 - X/N)^N \approx e^{-X}. \tag{4}$$

Therefore we find:

$$(1-p)^N = (1 - \frac{Np}{N})^N \approx e^{-Np}, \tag{5}$$

Defining $\lambda \equiv Np$ we find that:

$$P(X) \approx \frac{\lambda^X}{X!} e^{-\lambda}. \tag{6}$$

This is the Poisson distribution. It is easy to check that the distribution is indeed normalized, using the following Taylor expansion formula to sum over the X's:

$$e^{\lambda} = 1 + \lambda + \lambda^2/2!.... \tag{7}$$

Consider a single cell present when the virus is applied, corresponding to a single "coin-flip". The expectation value of its contribution to the number of living cells at the end of the experiment, X, is clearly $\mathbf{E}(x_i) = p$, and the variance (for the single cell contribution) is:

$$Var[x_i] = p(1-p)^2 + (1-p)(0-p)^2 = (1-p)p((1-p)+p) = p(1-p) \approx p = \mathbf{E}[x_i]. \tag{8}$$

Therefore, for a *single* coin-flip, the variance and mean are approximately the same for small p!

Furthermore, the total number of survivors X is simply the sum of N such independent variables, each having approximately the same variance and mean. Thus we find:

$$\mathbf{E}[X] = Np \approx Var[x], \tag{9}$$

as we have asserted previously. In fact, since the Poisson distribution arises from the binomial one in the limit where $p \to 0$, this relation then becomes *exact* for the Poisson distribution. This can also be verified directly by calculating its two first moments.

Mutation

Throughout this section, we would like to make the approximation that the probability of having a mutation occur on top of an already existing mutation (in a single experiment) is small (not to be confused with the probability of having more than one survivor!). This assumption hinges on the mutation probability p being sufficiently small (*how* small we shall shortly quantify). Before proceeding, let's imagine that we have just completed the experiment and would like to estimate the mutation probability p (assuming that this is the correct hypothesis). An important point to note is that while the number of *survivors* is highly variable (since an early mutation gives rise to a huge progeny of survivors), the number of *mutations* M that occur somewhere along the family tree must follow Poisson statistics, since mutation events are independent:

$$P(M) \approx \frac{\hat{\lambda}^M}{M!} e^{-\hat{\lambda}}. \tag{10}$$

Here, $\hat{\lambda} = pN_{events}$, where N_{events} is the total number of edges in our lineage tree (since mutations can occur on each of them), hence $N_{events} = N_0 \sum_{n=1}^{g} 2^n \approx 2N$.

Now, since there are zero survivors if and only if there were zero mutations, we conclude that the fraction of experiments in which there were no survivors should be approximately equal to:

$$P(M=0) \approx e^{-2Np}. \tag{11}$$

Since that number is high in the experiment, we conclude that Np must be a relatively small number (about 1/2 for the original experiment). As we shall shortly estimate, this implies that the probability for a mutation to occur in the lineage of a pre-existing mutation is negligible– a fact that will make the calculation of the mean and variance straightforward (since the surviving offspring of one mutation can be added independently to those of another mutation).

To see this, note that the probability of a mutation to occur in the lth replication event is:

$$P_l = pN_0 2^l. \tag{12}$$

The number of replication events that follow this one is about $f \cdot 2^{g-l}$, with $1 \leq f < 2$ (can you see why?). Hence the probability of *another* mutation occurring in the lineage of the first mutation is:

$$P_{second|first} = pf2^{g-l}. \tag{13}$$

Therefore the probability of getting *two* mutations in the same lineage is approximately:

$$\sum_l P_l P_{second|first} < 2gp^2 N_0 2^g = 2gp^2 N. \tag{14}$$

This is a ridiculously small number for the experimental parameters, hence we can safely neglect this event.

Consider now the cells of the nth generation. Their number is $N_0 \cdot 2^n$, so the expected number of mutations occurring in the reproduction events leading to this generation is:

$$\langle M_n \rangle = N_0 \cdot 2^n \cdot p, \tag{15}$$

where the $\langle \rangle$ denotes averaging over many trials.

The expected number of survivors originating from a mutation in the nth generation is:

$$\langle S_n \rangle = N_0 \cdot 2^n \cdot p \cdot 2^{g-n} = N_0 \cdot 2^g \cdot p, \tag{16}$$

where g is the total number of generations. The fact that this number is independent of n is crucial for our understanding of the DL experiment. It means that the survivors are equally likely to come from early or late generation, since although the odds are small for the mutation to occur earlier (due to the much smaller number of cells), this rare event will result in a huge number of survivors.

To see this more formally, let us calculate the mean and variance of the number of survivors. We have:

$$S_{total} = \sum_{j=1}^{g} S_j. \tag{17}$$

(writing the random variable S_{total} in this way *hinges* on our previous result that we can neglect the probability for a mutation to occur in the lineage of another).

From Eq. (16) we can write:

$$\langle S_{total} \rangle = \sum_{n=1}^{g} N_0 \cdot 2^n p \left(2^{g-n}\right) = N \cdot g \cdot p, \tag{18}$$

where $N = N_0 \cdot 2^g$, the growth of the initial population over g generations. Following a similar logic, let us consider the second moment of S_{total}. Eq. (17) allows us to write the random variable S_{total} as a sum of contributions from mutations that occurred in a specific generation j (where we rely on the excellent approximation discussed above that a mutation is highly unlikely to occur in the “wake” of another). Therefore we have:

$$\langle S^2_{total}\rangle \approx \sum_{j=1}^{g}\sum_{i=1}^{g}\langle S_i S_j\rangle . \tag{19}$$

The diagonal elements in the double-sum lead to:

$$S_{diagonal} = \sum_{n=1}^{g} N_0 \cdot 2^n p \left(2^{g-n}\right)^2 , \tag{20}$$

where $N_0 \cdot 2^n p$ is the expected number of mutations in the nth generation and $\left(2^{g-n}\right)^2$ is their contribution to the second moment.

The sum of Eq. (20) can readily be evaluated:

$$S_{diagonal} = \langle S^2_{total}\rangle \approx pN \sum_{n=1}^{g} 2^{g-n} \approx pN2^g . \tag{21}$$

The off-diagonal, in a similar fashion, give:

$$\sum_{i\neq j} (N_0 \cdot 2^i p)(N_0 \cdot 2^j p) \left(2^{g-i}\right) \left(2^{g-j}\right) \approx (Npg)^2 . \tag{22}$$

Note that for the experiment, where $Np = O(1)$, this contribution is negligible compared to the diagonal one, and equal to the square of the mean.

We therefore find that the variance is given by:

$$Var[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 \approx S_{diagonal} \approx Np2^g . \tag{23}$$

We therefore conclude that the variance:mean ratio is ridiculously high, of the order of 2^g!

In fact, the mean and variance calculation reflect the expected result if we repeat the experiments an enormous – and impractical – number of times. In order to find typical variance:mean ratio for the experiments, we have to consider the finite number Q of experiments performed. In your problem set you will simulate the experiment, which can be used to provide a good estimate for the expected variance:mean ratio for a given Q. We shall now repeat the argumentation from their original paper in order to estimate this ratio analytically.

Estimating the typical variance:mean ratio

When the number of experiments C is relatively small (e.g., a few hundreds in the original set of experiments), it is highly unlikely that in the *theoretical* expectation values of Eqs. (18, 21) we will actually sample mutations occurring in the early generations. What is the first generation that we can *typically* sample in C experiments? We can demand that a mutation in that generation occurs in *one* of the C runs of the experiments. This leads to the following equation:

$$N_0 2^{l_{min}} pC = 1 \rightarrow l_{min} = -\log_2(N_0 pc). \tag{24}$$

Taking for example $N_0 = 100$, $p = 5 \cdot 10^-9$ and $C = 200$, we find: $l_{min} \approx 9$. We therefore have to revisit the sums of Eqs. (18, 21), and start the sum not from the first generation, but from this minimal generation number. This leads to a profoundly different estimate for the variance:mean ratio – much lower than the previous estimate, but nevertheless much larger than 1. This estimate turned out to be in good agreement with the experimental results, lending further support to the random mutation hypothesis.

Heavy-tailed distributions

All of the approximately $2N$ replication events may independently lead to a mutation. Consider now a random mutation occurring somewhere on the lineage tree, without specifying in which generation it occurred (note that dividing the mutations according to this criterion was very helpful in making the previous quantitative estimates though!). The probability of a mutation occurring in the lth generation is given by Eq. (12). This leads to $X = 2^{g-l}$ survivors. If we now treat the number of survivors X as a continuous variable, we will find that:

$$P(X) = P_l[l = g - \log_2(X)]/|\frac{dx}{dl}|. \tag{25}$$

We have $|\frac{dx}{dl}| = e^{\log(2)(g-l)} \log(2)$, hence:

$$P(X) = pN_0 2^{g-\log_2(X)}/|\frac{dx}{dl}| \propto \frac{1}{X^2}. \tag{26}$$

Therefore the number of survivors can be thought of as a sum of independent random variables, each drawn from a power-law tailed distribution. The so-called "heavy-tail" of this distribution is precisely what underlies the rare "jackpot" events of the distribution of the number of survivors, leading e.g. to the anomalously large variance:mean ratio. In the limit of a large number of generations g (not corresponding to the realistic situation), a remarkable generalization of the central limit theorem can be utilized to show that this sum of independent random variables converges to a universal distribution – albeit not a Gaussian as in the case of distributions with a *finite* variance [2]. This family is known as Lévy stable distributions, and for the particular case of the distribution of Eq. (26), the distribution pops up in physics as well and is known as the Landau distribution. This connection between the Delbruck-Luria experiment and the generalized central limit theorem was first noted by Mandelbrot in Ref. [3], and dealt with more carefully mathematically in Ref. [4].

Numerical simulation In fact, it is incredibly easy to simulate the model discussed above. The following code generates the distribution of outcomes of many runs of the experiments, for realistic parameters.

```
p=5*10^-9; N0=100; generation_num=20; experiment_number=10000;
Nf=N0*2^generation_num;
for exp=1:experiment_number
    mutant=0;
    non_mutant=N0;
    for j=1:generation_num
        tmp= poissrnd(2*p*non_mutant);
        mutant=2*mutant + tmp;
        non_mutant=non_mutant*2 - tmp;
    end;
    res(exp)=mutant;
end;
bins=[0:10];
q=hist(res,bins);
bar(bins,q/(sum(q)*(bins(2)-bins(1))));
```

The result is shown in Fig. 2. Note that in this code the divisions are *synchronous* (i.e., all occuring at the same time) leading (unrealistically) to peaks at powers of 2 (can you see why?) One may similarly write code for *asynchronous* divisions to remove this artifact, though both cases exhibit the large variance:mean ratio, the main point of this experiment and analysis.

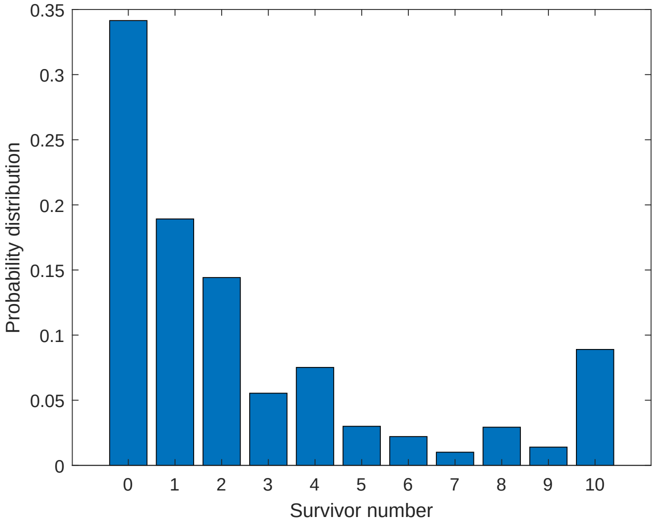

Figure 2: Histogram of the number of survivors in a numerical simulation of the experiment.

References

[1] Luria, Salvador E., and Max Delbrück. ”Mutations of bacteria from virus sensitivity to virus resistance.” Genetics 28.6 (1943): 491.

[2] Amir, A., 2020. An elementary renormalization-group approach to the generalized central limit theorem and extreme value distributions. Journal of Statistical Mechanics: Theory and Experiment, 2020(1), p.013214.

[3] Mandelbrot, B., 1974. A population birth-and-mutation process, I: explicit distributions for the number of mutants in an old culture of bacteria. Journal of Applied Probability, 11(3), pp.437-444.

[4] Kessler, D.A. and Levine, H., 2013. Large population solution of the stochastic Luria–Delbrück evolution model. Proceedings of the National Academy of Sciences, 110(29), pp.11682-11687.