Evolutionary Dynamics: Homework 10

Guido Putignano, Lorenzo Tarricone, Gavriel Hannuna,
Athanasia Sapountzi

February 22, 2024

Problem 2: Probability generating function

Let Z be a random variable such that $Z \in \mathbb{Z}^+$ (where Z is a non-negative integer), and p_i its distribution, i.e.

$$\text{Prob}[Z = i] = p_i.$$

The probability generating function (pgf) of Z is a function of a symbolic argument s defined as the expected value $E[s^Z] = \sum_{i=0}^{\infty} p_i s^i$ and denoted by $f_Z(s)$. We assume that all probability generating functions are absolutely convergent on the interval $[0, 1]$. Note: This is a technical requirement to assure that summand-wise operations are permitted.

Prove the following statements:

(a) The expectation of Z is given by $E[Z] = f'_Z(1)$; (1/2 point)
Solution
First of all, let us calculate the derivative of f(s), and evaluate it at s=1.

$$\frac{df}{ds} = \Sigma_{i=1}^{\infty} i p_i s^{i-1} \Rightarrow \frac{df}{ds}(1) = \Sigma_{i=1}^{\infty} i p_i$$

In this case, i represents the number of individuals in a generation, and p_i is the corresponding probability of obtaining i individuals. This would mean that the equation found before can be rewritten in the form of the expected value of the individual in a generation:

$$E[Z] = \Sigma_{i=1}^{\infty} i p_i = f^{'}(1)$$

(b) The variance of Z is $\text{Var}[Z] = f'_Z(1) + f''_Z(1) - [f'_Z(1)]^2$; (1/2 point)
Solution
We know that the variance of a distribution is defined as:

$$Var[Z] = E[Z^2] - E[Z]^2$$

Given what we showed in section (a), we can say that $Var[Z] = E[Z^2] - [f^{'}(1)]^2$. Now $E[Z^2]$ can be defined in the following way:

$$E[Z^2] = \Sigma_{i=0}^{\infty} i^2 p_i$$

We saw that the degree of the polynomial multiplied by p_1 is indicative of the order of the derivation performed on f(s), hence we might expect this sum to be simplified with the second derivative of f(s).

$$f^{''}(1) = \Sigma_{i=2}^{\infty} i(i-1)p_i = \Sigma_{i=0}^{\infty} i^2 p_i - i p_i$$

The sum $\Sigma_{i=0}^{\infty} i p_i$ is actually defined as $f^{'}(1)$, so we can rewrite the previous equation as:

$$f^{''}(1) = E[Z^2] - f^{'}(1) \Rightarrow E[Z^2] = f^{''}(1) + f^{'}(1)$$

So finally $Var[Z] = f^{''}(1) + f^{'}(1) - [f^{''}(1)]^2$

(c) $\frac{d^k f_Z}{ds^k}\Big|_{s=0} = k! p_k$; (1 point)

Solution

This comes directly from the laws of derivation of an exponential function x^a (Poisson distribution). We can see from the previous definition of f(s) that it is a geometric sequence, where p_k is multiplied by an exponential of s^k. Let us derivate this sum k times

$$\frac{df}{ds} = \Sigma_{k=0}^{\infty} k p_k s^{k-1}$$

$$\frac{d^2 f}{ds^2} = \Sigma_{k=0}^{\infty} k(k-1) p_k s^{k-2}$$

$$\dots$$

$$\frac{d^k f}{ds^k} = \Sigma_{k=0}^{\infty} k \cdot (k-1) \dots 2 \cdot 1 \cdot p_k s^0 = \Sigma_{k=0}^{\infty} k! p_k$$

Any k between 0 and k-1 will cause the sum to be 0 since one term of the k! will be 0, so we can directly write this sum as $\frac{d^k f}{ds^k} = k! p_k$

(d) If Z and Y are two independent random variables in $\mathbb{Z}^+$, then $f_{Z+Y}(s) = f_Z(s) f_Y(s)$; (1 point)

Solution

This conclusion can be derived directly from the properties of expectations E. Previously we defined $f_Z(s) = E[s^Z]$, so now we can expand this definition with 2 independent random variables Z and Y $f_{Z+Y}(s) = E[s^{Z+Y}]$. Let us define $Prob(Y = k) = p_k$ and a previously stated, $Prob(Z = i) = p_i$, then:

$$E[s^{Z+Y}] = \Sigma_{k=0}^{\infty} \Sigma_{i=0}^{\infty} p_k p_i s^{i+k} = \Sigma_{k=0}^{\infty} \Sigma_{i=0}^{\infty} p_k p_i s^i s^k = E[s^Z] E[s^Y]$$

The last expression is equivalent by definition to $f_Z(s) f_Y(s)$ so

$$f_{Z+Y}(s) = f_Z(s) f_Y(s)$$

(e) If Y is a $\mathbb{Z}^+$-valued random variable and $\{Z^{(i)}, i \geq 1\}$ a sequence of independent identically distributed random variables in $\mathbb{Z}^+$ independent of Y, then $V = \sum_{i=1}^{Y} Z^{(i)}$ has the pgf $f_V(s) = f_Y[f_{Z^{(1)}}(s)]$.

Solution
Firs of all, let us open the right-hand expression $f_Y[f_{Z^{(1)}}(s)]$

$$f_Y[f_{Z^{(1)}}(s)] = f_Y[E[s^{Z^{(1)}}]] = E[E[s^{Z^{(1)}}]^Y] =$$

Knowing two properties of expectations $E[E[x]] = E[x]$ and that $E[XY] = E[X]E[Y]$ (since $Z^{(i)}$s are independent) we can simplify the previous expression as follows

$$= E[s^{Z^{(1)}}] \cdot E[s^{Z^{(1)}}].... \cdot E[s^{Z^{(1)}}]$$

Multiplied Y times.
On the other hand, we have $f_V(s)$ which from what we found in (d) is equal to:

$$f_V(s) = f_{Z^{(1)}} \cdot f_{Z^{(2)}}.... \cdot f_{Z^{(Y)}}$$

Let us remember that the $Z^{(i)}$ are taken from the same distribution, meaning they have the same p_i. Hence all of the pgfs are equal

$$f_V(s) = f_{Z^{(1)}} \cdot f_{Z^{(1)}}.... \cdot f_{Z^{(1)}} = E[s^{Z^{(1)}}] \cdot E[s^{Z^{(1)}}].... \cdot E[s^{Z^{(1)}}]$$

This is indeed what we previously found when expanding the right-hand side of the initial equation.

Problem 3: The Luria-Delbrück experiment

The Luria–Delbrück experiment (Salvador E Luria and Max Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. Genetics, 28(6):491, 1943) tests two hypotheses for how bacteria acquire resistance to the virus. The first hypothesis (adaptive immunity) states that the mutations leading to resistance to the virus were caused by an induced activation (exposure to the virus). The second hypothesis (random mutation) states that the mutations to resistance may occur any time prior to the addition of the virus. The experiment demonstrated that in bacteria, genetic mutations arise in the absence of selective pressure rather than being a response to it.

Experiment setup (see the figure below): Several bacterial cultures are grown from a single cell into separate culture tubes. After a period of growth, the cultures are exposed to the virus. If the resistance to the virus was caused by adaptive immunity, then each plate should contain roughly the same number of resistant colonies. Otherwise (the random mutation case) the number of resistant colonies on each plate should vary (the variance greater than the mean).

Cells grow at a rate β, such that $N(t) = N(0)e^{\beta t}$, and they give stochastically rise to a mutant (resistant) offspring with rate α. Hence the number of cells that directly arise through mutation are a non-homogeneous Poisson process with time-dependent rate $\lambda(t) = \alpha N(t)$. Thus, the distribution of the number of mutations that occur in $[0, t]$ is Poissonian with parameter $\Lambda(t) = \int_0^t \lambda(\tau)d\tau$. In the absence of the virus, mutant cells grow at the same rate as normal bacteria.

(a) Compute the probability $P_0(t)$ that no mutations have occurred at time t. Show that the mutation rate α can be estimated as $\alpha = \frac{\beta \ln \rho}{1-e^{\beta t}}$, where ρ is the ratio of experiments in which resistance was not found (estimator for $P_0(t)$). Assume $N(0) = 1$.

Hint: Use

$$\begin{aligned} P_0(t) &= P(0 \text{ mutants in } [0, \Delta t]) P(0 \text{ mutants in } [\Delta t, 2\Delta t]) \dots P(0 \text{ mutants in } [t - \Delta t, t]) \\ &\approx (1 - \alpha N(0)\Delta t) \dots (1 - \alpha N(t - \Delta t)\Delta t) \\ &\approx e^{-\alpha N(0)\Delta t} \dots e^{-\alpha N(t-\Delta t)\Delta t}, \end{aligned}$$

and let $\Delta t \to 0$. Explain the assumptions made in this calculation. (2 points)

Solution

We partition the time interval $[0, t]$ into n sub-intervals, each of length Δt. The probability of having no mutants at time t is equivalent to the probability of observing no mutants across all these intervals. Since the individual intervals are independent of each other we have:

$$P_0(t) = P(0 \text{ mutants in } [0, \Delta t]) \cdot P(0 \text{ mutants in } [\Delta t, 2\Delta t]) \cdot \dots \cdot P(0 \text{ mutants in } [t - \Delta t, t])$$

We assume that $\Delta t \to 0$, so the size of the population can be considered constant in each time interval.

The probability that a cell mutated in time interval $[t, t + \Delta t]$ is calculated as: $\alpha N(t)\Delta t$.

The probability that there was no mutation: $1 - \alpha N(t)\Delta t$

Thus, we have:

$$\begin{aligned} P_0(t) &= P(0 \text{ mutants in } [0, \Delta t]) \cdot P(0 \text{ mutants in } [\Delta t, 2\Delta t]) \cdot \dots \cdot P(0 \text{ mutants in } [t - \Delta t, t] \\ &= (1 - \alpha N(0)\Delta t)(1 - \alpha N(\Delta t)\Delta t) \cdots (1 - \alpha N(t - \Delta t)\Delta t) \\ &\approx e^{-\alpha N(0)\Delta t} e^{-\alpha N(\Delta t)\Delta t} \cdots e^{-\alpha N(t-\Delta t)\Delta t} \\ &= e^{-\alpha(N(0)\Delta t + N(\Delta t)\Delta t + \dots + N(t-\Delta t)\Delta t)} \end{aligned}$$

As we mentioned above, $\Delta t \to 0$ so we have:

$-\alpha(N(0)\Delta t + N(\Delta t)\Delta t + \dots + N(t - \Delta t)\Delta t) = -\alpha \int_0^t N(\tau) d\tau$

Cells grow at rate β :

$$\begin{aligned} &= -\alpha \int_0^t N(0) e^{\beta \tau} d\tau \\ &= -\alpha N(0) \frac{1}{\beta} \left[e^{\beta \tau} \right]_0^t \\ &= -\frac{\alpha}{\beta} N(0) \left(e^{\beta t} - 1 \right) \end{aligned}$$

So we calculate:

$$P_0(t) \approx \rho = e^{-\frac{\alpha}{\beta}N(0)\left(e^{\beta t}-1\right)}$$

Given ρ be the estimator for $P_0(t)$. We can define that the estimator for α is calculated as

$$\rho = e^{-\frac{\alpha}{\beta}\left(e^{\beta t}-1\right)} \Leftrightarrow \alpha = \frac{\beta \ln \rho}{1 - e^{\beta t}}$$

(b) Derive α from the expression for P_0 derived in the lecture for the Galton-Watson process and explain the differences. (1 point)

Solution

$P_0(t)$ For the Galton-Watson process $P_0(t)$ is computed as:

$P_0(t) = (1-\alpha)^{2^{t+1}-2}$

Let ρ be the estimator for $P_0(t)$, we have

$\rho = (1-\alpha)^{2^{t+1}-2} \Leftrightarrow \ln \rho = (2^{t+1}-2)\ln(1-\alpha)$

For small α,

$$\ln \rho \approx \left(2^{t+1}-2\right)(-\alpha) \Leftrightarrow \alpha \approx \frac{\ln \rho}{2\left(1-2^t\right)}$$

We observe a significant resemblance between this expression and the one found in the Luria-Delbrück experiment, where $\alpha = \frac{\beta \ln \rho}{1-e^{\beta t}}$. Both approximations for α exhibit proportionality to $\ln \rho$. These two expressions are equal when $\frac{\beta}{1-e^{\beta t}} = \frac{1}{2(1-2^t)}$.

In the case of the Galton-Watson process described in the lecture, each cell in each generation gives rise to exactly 2 offspring. Consequently, the number of cells after t generations is 2^t. The expected number of cells at time t is $e^{\beta t}$

(c) Compute the expected number of mutant cells at time t, $m(t)$, and their variance $\sigma^2(t)$.

Hint: The expected number of new mutant cells that arise in the interval $\tau + d\tau$ is:

$$\nu(\tau + d\tau) = \lambda(\tau)d\tau.$$

Compute to which size these newly generated subclones have grown to at time t and express $m(t)$ as their superposition. Consider a similar strategy for the variance. Use that the new mutant cells arising in $\tau + d\tau$ are Poissonian variables and remember that $\mathrm{Var}[aX] = a^2\mathrm{Var}[X]$. (2 points)

Solution

The expected number of new mutant cells that arise in the interval $\tau + d\tau$ is:

$\nu(\tau + d\tau) = \lambda(\tau)d\tau$

Since the mutant cells have a growth rate β, the growth of mutants generated in time interval $\tau + d\tau$ up until time t can be represented by an

analogous expression to $N(t) = N(0)e^{\beta t}$, where $\nu(\tau + d\tau)$ is analogous to N(0) and the $t - \tau$ is analogous to t.

$$\begin{aligned}\nu\left(\tau + d\tau\right)e^{\beta(t-\tau)} &= \lambda(\tau)d\tau e^{\beta(t-\tau)} \\ &= \alpha N(\tau)e^{\beta(t-\tau)}d\tau \\ &= \alpha N(0)e^{\beta\tau}e^{\beta(t-\tau)}d\tau \\ &= \alpha N(0)e^{\beta t}d\tau\end{aligned}$$

If we integrate over the interval from 0 to t, we get the expected number of mutant cells at time t :

$$\begin{aligned}m(t) &= \int_0^t \alpha N(0)e^{\beta t}d\tau \\ &= \alpha N(0)e^{\beta t}\int_0^t d\tau \\ &= \alpha N(0)e^{\beta t}(t-0) \\ &= \alpha N(0)e^{\beta t}t \\ &= \alpha N(t)t\end{aligned}$$

Since we have a Poisson process:

$$\sigma^2(\tau + d\tau) = \nu(\tau + d\tau) = \lambda(\tau)d\tau$$

The number of mutants after time t:

$$\nu(\tau + d\tau)e^{\beta(t-\tau)}$$

Thus, the variance of the number of mutants after time t is:

$$\operatorname{Var}\left[\nu(\tau + d\tau)e^{\beta(t-\tau)}\right] = e^{2\beta(t-\tau)}\operatorname{Var}[\nu(\tau + d\tau)] = e^{2\beta(t-\tau)}\lambda(\tau)d\tau$$

We now take the integral over the time interval from 0 to t to get:

$$\begin{aligned}\sigma^2(t) &= \int_0^t e^{2\beta(t-\tau)}\lambda(\tau)d\tau \\ &= \alpha N(0)\int_0^t e^{\beta\tau}e^{2\beta(t-\tau)}d\tau \\ &= \alpha N(0)e^{2\beta t}\int_0^t e^{-\beta\tau}d\tau \\ &= \frac{\alpha}{\beta}N(0)e^{2\beta t}\left(1 - e^{-\beta t}\right) \\ &= \frac{\alpha}{\beta}N(t)\left(e^{\beta t} - 1\right) \\ \sigma^2(t) &= \frac{\alpha}{\beta}N(t)\left(e^{\beta t} - 1\right)\end{aligned}$$

(d) Luria and Delbrück used the mean and variance to distinguish the proposed mechanism of mutations stochastically accumulating prior to viral infection from an active adaptation scenario. Suppose that in the adaptation case, bacteria have no resistance, but stochastically acquire resistance upon infection with high rate δ. In this short period of time, the population size can be considered constant. What would be the resulting relation between the expected number of resistant cells and their variance? Compare this with your results from the accumulation scenario, part (c). (1 point)

Solution

Consider the adaptation scenario where bacteria initially lack resistance but stochastically develop resistance upon infection at a high rate δ. Over a brief time span $[0, t]$, we can treat the population size as constant.

$$m(t) = \sigma^2(t) = \int_0^t \delta N(0) d\tau = \delta t.$$

In the adaptation case:

$$\frac{\sigma^2(t)}{m(t)} = 1$$

whereas in the accumulation scenario, we have

$$\frac{\sigma^2(t)}{m(t)} = \frac{\frac{\alpha}{\beta} N(t)(N(t)-1)}{\alpha t N(t)} = \frac{e^{\beta t} - 1}{\beta t} \to 1 \text{ as } t \to 0$$

Hence, it is reasonable to anticipate that over a brief timeframe, both situations will exhibit comparable variance-mean ratios. Over a longer duration, the accumulation scenario is expected to exhibit significantly greater variability in comparison to the adaptation scenario. Hence, analyzing the magnitude of this ratio serves as a means to differentiate between mechanisms that more accurately explain how bacteria develop resistance to a virus. Based on the studies done by Luria and Delbrück, the accumulation scenario is deemed more plausible.