Lecture 02: Calculus, Probability, and Combinatorics

In these notes, we review the mathematics of change, chance, and counting.

1 Change, Chance, and Counting

Like all areas of physics, statistical physics makes use of calculus, the mathematics of change. Through calculus, we can define functions and compute the way these functions vary their independent variables. But the "statistical" in the statistical physics' name suggests we will need additional mathematical tools beyond the ones developed in an introductory physics course. These tools can in general be framed our ability to answer certain questions concerning chance, and counting. The mathematics of chance is more formally known as **probability theory**, and the mathematics of counting is more formally termed **combinatorics**). Probability theory provides ways to define and compute the properties (e.g., averages, standard deviations) of random events. Combinatorics provides systematic ways for counting elements in a set. With the tools from calculus, chance, and counting we should be able to answer questions like

1. How do we determine the local maxima of $f(x) = x^3/3 - x$?
2. The probability density for a random variable t is $p(x) = e^{-t/\tau}/\tau$. What is the variance of t?
3. How many ways can we fill 3 different electoral positions when we have 9 people?

The physical relevance of these questions may be unclear right now, but each one is related to techniques we will need when we study statistical physics. So our framing question for these notes is

Framing Question
What are the basic mathematics governing change, chance, and counting?

2 Calculus: Mathematics of change

A function in a single real variable is a quantity that takes in a real number (i.e., a number that has no factors of the imaginary number $i = \sqrt{-1}$) and outputs another number, such that each input has only a single output. Common examples of functions are

$$f(x) = e^x, \qquad f(x) = x^3 + x^2 + x + 1, \qquad \text{and} \qquad f(x) = \sin(x). \tag{1}$$

Less common examples, (both of which we will see in this course) are

$$f(x) = \Gamma(x+1), \qquad \text{and} \qquad f(x) = \tanh(x). \tag{2}$$

2.1 Differentiation

When we have a function, we are often also interested in the rate of change of the function with respect to its independent variable. If we wanted to know the average rate of change in $f(x)$ as x varied from $x = x_0$ to $x = x_1$, we would compute

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0}. \qquad \text{[Average rate of change from } x = x_0 \text{ to } x = x_1\text{.]} \tag{3}$$

However, more often we are not merely interested in an average rate of change over an interval, but in a rate of change at a *specific point*. In such a scenario, we have to compute the **derivative** of the function at the point of interest. The derivative of $f(x)$ at the point $x = x_0$ is defined as

$$\lim_{h\to 0} \frac{f(x_0+h)-f(x)}{h}, \qquad \text{[Instantaneous rate of change at } x = x_0\text{.]} \tag{4}$$

where h is a real number and $\lim_{h\to 0}$ means we take the value of the associated expression as h approaches 0. The quantity in Eq.(4) is variously denoted as $f'(x_0)$, $df/dx|_{x=x_0}$, and $df(x_0)/dx$. Some common examples of derivatives are

$$\frac{d}{dx}e^x = e^x \qquad \frac{d}{dx}\sin(x) = \cos(x)\,, \qquad \frac{d}{dx}x^n = nx^{n-1}. \tag{5}$$

We can schematically represent derivatives of functions as the action of the derivative operator on the original function:

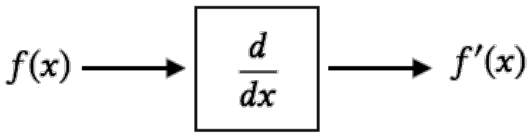

Figure 1: The action of a derivative takes a function and outputs a new function.

Derivatives will be important in our subsequent development of statistical physics, not only because they encode information on instantaneous rates of change, but also because they allow us to determine at what values in its domain a function has local minima or local maxima. Say we had the function

$$f(x) = \frac{x^3}{3} - x. \tag{6}$$

Plotting the function, we obtain the graph shown in Fig. 2. If we wanted to determine, at what values of x the function assumes the points denoted "top of hill" and "bottom of valley" we could use calculus to do so. Using their formal names, the "top of the hill" is a **local maximum** of the function and the "bottom of the valley" is a **local minimum** of the function. The local maximum of a function is the point where the value of the function is larger than any other values in some arbitrarily small vicinity of that point. The local minimum is similarly defined.

From the properties of derivatives, we recognize that at both the top of the hill and the bottom of the valley, the rate of change of the function is zero:

$$f'(x) = 0 \text{ at both local minima and local maxima} \tag{7}$$

However, as we move from the left to the right of the local maximum, the derivative of $f(x)$ changes from positive to negative. Thus while the rate of change of $f(x)$ is zero at the local maximum, *the rate of change of this rate of change* is negative at the local maximum, that is, at the local minimum the rate of change is decreasing. Similarly, the rate of change in moving from the left of the local minimum to the right of the local minimum goes from negative to positive, that is, at the local maximum the rate of change is increasing. Thus, the properties of the first derivative and the second derivative at an optimal value of a function determines what kind of optimal value it is:

- A function has a local minimum at a point x_0, if $f'(x_0) = 0$ and $f''(x_0) > 0$.
- A function has a local maximum at a point x_0, if $f'(x_0) = 0$ and $f''(x_0) < 0$.

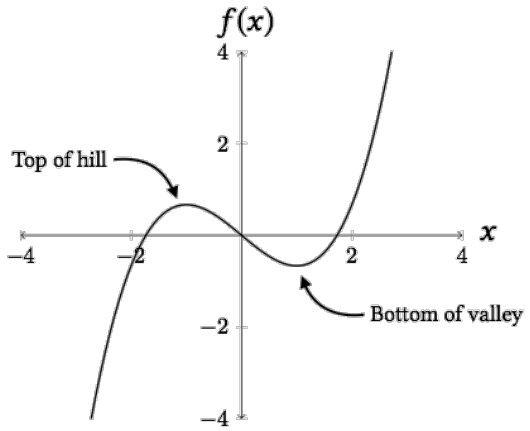

Figure 2: The function $f(x) = x^3/3 - x$ has a local minimum and a local maximum. We can compute the location of the former by finding the value of x where $f'(x) = 0$ and $f''(x) > 0$. We can compute the location of the latter by finding the value of x where $f'(x) = 0$ and $f''(x) < 0$.

2.2 Integration

Another question we may ask about differentiation is whether it has an inverse operation. Is there an operation that, when given the derivative of a function, returns the function itself? In a calculus course, we find that the answer is indeed yes. The operation that is the inverse of the derivative is called the anti-derivative or **indefinite integral** of a function, and it is denoted symbolically as $\int dx$. For example, when acting on the derivative $f'(x)$, this operation yields

$$\int dx\, f'(x) = f(x) + C, \tag{8}$$

where C is a constant which must be added to $f(x)$ to represent the most general form of a function which yields $f'(x)$ when differentiated. Depicted schematically, Eq.(14) is

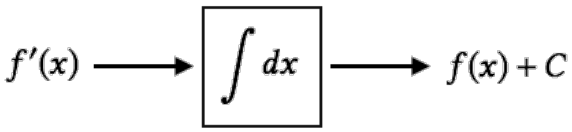

Figure 3: The action of an indefinite integral takes a function and outputs a new function. It is the inverse operation of the derivative in Fig. 1.

The relationship between derivatives and indefinite integrals that is shown in Eq.(14) can also be expressed in terms of **definite integrals**. A definite integral represents the area between a function and the x-axis from an initial point on that axis to a final point. Expressing Eq.(14) as a definite integral leads to the **fundamental theorem of calculus**. The simplest form of the theorem states that for functions $f(x)$ that are continuous and differentiable over their entire domain, we have

$$\int_a^b dx\, f'(x) = f(b) - f(a). \tag{9}$$

Conceptually, the theorem states that the summation of the rate of change of a function from one point to another point is equal to the difference between the values of the function at those two points. For example,

we can evaluate the integral of $\sin^2(x)$ from 0 to 2π using Eq.(15):

$$\begin{aligned}\int_0^{2\pi} dx\, \sin^2 x &= \int_0^{2\pi} dx\, \frac{1}{2}\left(1-\cos 2x\right) \\ &= \left[\frac{x}{2} - \frac{1}{4}\sin(2x)\right]_0^{2\pi} \\ &= \frac{2\pi}{2} - \frac{1}{4}\sin(4\pi) - \frac{0}{2} + \frac{1}{4}\sin(0) = \pi. \end{aligned} \tag{10}$$

Unlike differentiation, which can be performed algorithmically, it is not possible to calculate the integral of every function we encounter. However, there are two standard techniques for computing integrals which will be useful throughout this course:

- **u-substitution:** Because the derivative of a composite function $f(g(x))$ is $f'(g(x))g'(x)$ the anti-derivative of $f'(g(x))g'(x)$ is $f(g(x))$. Thus, by Eq.(15), we have

$$\int_a^b dx\, g'(x) f'\left(g(x)\right) = f\left(g(b)\right) - f\left(g(a)\right). \tag{11}$$

  We can write this equation in a simpler form by defining $u = g(x)$ and substituting this definition into the integral. Noting that $du/dx = g'(x)$ and thus that $dx\, g'(x) = du$, we have

$$\int_{u_a}^{u_b} du\, f'\left(u\right) = f(u_b) - f(u_a), \tag{12}$$

  where $u_b = g(b)$ and $u_a = g(a)$. This is the essence of the u-substitution technique: Finding a factor in the integral which is the derivative of the argument of a composite function. This technique is useful in evaluating the integrals of functions like xe^{-x^2} and $\cos(x)\ln\left(\sin(x)\right)$.

- **Integration by parts:** By the product rule of derivatives, we know that

$$\frac{d}{dx}(fg) = \frac{df}{dx}g + f\frac{dg}{dx}, \tag{13}$$

  for functions $f(x)$ and $g(x)$. We can integrate both sides of Eq.(19) and use Eq.(15), to find

$$fg\Big]_{x=a}^{x=b} = \int_a^b dx\, \frac{df}{dx}g + \int_a^b dx\, f\frac{dg}{dx}, \tag{14}$$

  or

$$\int_a^b dx\, \frac{df}{dx}g = fg\Big]_{x=a}^{x=b} - \int_a^b dx\, f\frac{dg}{dx}. \tag{15}$$

  We thus are able to evaluate the integral of a product of two functions by "reducing" the number of derivatives applied to one function and "increasing" the number of derivatives applied to the other. This technique aids in the evaluation of integrals of functions like x^4e^{-x} and $e^x\sin(x)$.

2.3 Taylor Series

The final topic of calculus that we will review provides a way to express special functions like $\ln(1+x)$ and $\tan x$ as sums of functions like x, x^2, and so on. We begin with the fundamental theorem of calculus:

$$\int_{x_0}^{x} dx_1\, f'(x_1) = f(x) - f(x_0), \tag{16}$$

or, rearranging the equation a bit,

$$f(x) = f(x_0) + \int_{x_0}^{x} dx_1\, f'(x_1), \tag{17}$$

where x_1 is an integration variable whose exact label is not important[1]. Eq.(23) expresses the function $f(x)$ as an integral of its derivative offset by a constant. This equation is valid for any function $f(x)$ that has a well-defined derivative over its entire domain. Let's suppose this is true for our $f(x)$. We can, for example, express $f'(x_1)$ itself as an integral. Doing so, we have

$$f'(x_1) = f'(x_0) + \int_{x_0}^{x_1} dx_2\, f''(x_2), \tag{18}$$

where $f''(x)$ is both the second derivative of $f(x)$ and the first derivative of $f'(x)$, and x_2 is a new integration variable. Inserting Eq.(24) into Eq.(23) yields

$$\begin{aligned} f(x) &= f(x_0) + \int_{x_0}^{x} dx_1 \left[f'(x_0) + \int_{x_0}^{x_1} dx_2\, f''(x_2) \right] \\ &= f(x_0) + \int_{x_0}^{x} dx_1\, f'(x_0) + \int_{x_0}^{x} dx_1 \int_{x_0}^{x_1} dx_2\, f''(x_2) \\ &= f(x_0) + (x - x_0) f'(x_0) + \int_{x_0}^{x} dx_1 \int_{x_0}^{x_1} dx_2\, f''(x_2), \end{aligned} \tag{19}$$

where in the final line we used the fact that $f'(x_0)$ is independent of x to simplify the integration. We can repeat this procedure for $f''(x_2)$. We replace $f(x)$ in Eq.(23) with $f''(x_2)$ to obtain

$$f''(x_2) = f''(x_0) + \int_{x_0}^{x_2} dx_3\, f'''(x_3). \tag{20}$$

Inserting this expression into Eq.(25) and using

$$\int_{x_0}^{x} dx_1 \int_{x_0}^{x_1} dx_2 = \int_{x_0}^{x} dx_1\, (x_1 - x_0) = \frac{1}{2}(x - x_0)^2, \tag{21}$$

we have

$$f(x) = f(x_0) + (x - x_0) f'(x_0) + \frac{1}{2}(x - x_0)^2 f''(x_0) + \int_{x_0}^{x} dx_1 \int_{x_0}^{x_1} dx_2 \int_{x_0}^{x_2} dx_3\, f'''(x_3). \tag{22}$$

You may already recognize the pattern. If we were to continue this procedure indefinitely we would obtain

$$f(x) = f(x_0) + (x - x_0) f'(x_0) + \frac{1}{2}(x - x_0)^2 f''(x_0) + \frac{1}{3!}(x - x_0)^3 f'''(x_0) + \frac{1}{4!}(x - x_0)^4 f^{(4)}(x_0) + \cdots, \tag{23}$$

where $f^{(k)}$ is the kth derivative of f. We can write Eq.(29) more succinctly using sum notation:

$$f(x) = \sum_{k=0}^{\infty} \frac{(x - x_0)^k}{k!} f^{(k)}(x_0) \tag{24}$$

Eq.(30) is called the **Taylor series** of the function $f(x)$ expanded about the point $x = x_0$. The Taylor series allows us to express a function as a summation (with appropriate coefficients) of powers of x. What the above derivation obscures is the fact that this sum does not always result in a finite quantity. For this class, we sidestep this issue, because we would have to use more theoretical tools from calculus to establish the

[1] We could, for example, replace x_1 with u, y, or "smiley face" and we would get the same result.

conditions under which Eq.(30) is finite. For our purposes, we will only apply Eq.(30) to functions whose domain of validity in x is well known.

Often we are interested in Taylor series about the point $x = 0$, in which case the expression Eq.(30) simplifies. Below we list a few such Taylor series for common functions in mathematics. On each line we also include the domain of x in which these Taylor Series are valid.

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^k}{k!} + \cdots \qquad \text{valid for all } x \tag{25}$$

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots + (-1)^k \frac{x^{2k+1}}{(2k+1)!} + \cdots \qquad \text{valid for all } x \tag{26}$$

$$\ln(1+x) = x - \frac{x^2}{2} - \frac{x^3}{3} + \cdots + (-1)^{k-1}\frac{x^k}{k} + \cdots \qquad \text{valid for } |x| < 1 \tag{27}$$

$$(1+x)^n = 1 + nx + \binom{n}{2}x^2 + \binom{n}{3}x^3 + \cdots + \binom{n}{k}x^k + \cdots \quad \begin{cases} n \text{ is an integer} & \text{valid for all } x \\ \text{otherwise} & \text{valid for } |x| < 1 \end{cases} \tag{28}$$

It is worth noting that Eq.(34) is both the Taylor Series and binomial expansion of $(1+x)^n$.

3 Probability

3.1 Discrete random variables

In our later development of statistical physics, we will constantly deal with the mathematics of chance through questions concerning probability and averages. In order to build up the mathematics relevant to these questions, it will prove simplest to begin with an example.

Say we had a fair six-sided die. Rolling the die and allowing it to settle on a side would yield an upturned face with dots numbering either one, two , three, four, five, or six. Because the outcome of the roll is random, we call this outcome a **random variable**. More formally, a random variable is a quantity that results from performing a random experiment. For this system of a single die, we label our general random variable as j where j can be any element of the set $\{1, 2, 3, 4, 5, 6\}$.

Because j takes on integer values (and is, more generally, countable), the system of the rolled die is said to concern a **discrete random variable**. The possible values of our random variable are called **events** such that when we roll the die, we observe one of the possible events. Because we have a fair die, the probability of each of the possible events is the same. Labeling p_j as the probability of observing j dots upon a roll of the die (i.e., the probability of observing event j), we have

$$p_j = 1/6 \qquad \text{for } j \text{ an element of } \{1, 2, 3, 4, 5, 6\}. \tag{29}$$

We could have inferred the result Eq.(35) intuitively, but more rigorously it came from noting that the the probability to observe *any* event must be 1. That is the sum of the probabilities of all the possible values of j must be 1:

$$\sum_{j=1}^{N} p_j = 1, \tag{30}$$

where $N = 6$. Given that p_j is the same for all j, we then obtain $6p_1 = 1$ or $p_1 = 1/6$, which is the probability of obtaining any *particular* value of j.

We can ask another question: What is the average value of a roll for this die? Namely if we were to roll the die (say) 120 times, to then sum all of the obtained values on the upturned side, and to finally divide by 120, what would we expect to get[2]? Because the probability of each event is $1/6$, we would expect to get

[2]This calculation is mostly correct, but when we actually implement multiple trials of a random experiment, the value we actually obtain is rarely the exact theoretical value. Instead, the experimental value has some probability (which is not equal to 1) of being in

$1/6 \times 120 = 20$ repeats of each side of the die. In this case, the average value would be

$$\begin{aligned}\text{Average of } j &= \frac{1\times 20 + 2\times 20 + 3\times 20 + 4\times 20 + 5\times 20 + 6\times 20}{120} \\ &= \frac{1}{6}\times 1 + \frac{1}{6}\times 2 + \frac{1}{6}\times 3 + \frac{1}{6}\times 4 + \frac{1}{6}\times 5 + \frac{1}{6}\times 6 \\ &= \frac{15}{6}, \end{aligned} \tag{31}$$

From the second line of Eq.(37), we see that the average value of j is the sum of the j for all events weighted by the probability of the event. Denoting this average (also called the **mean**) as $\langle j\rangle$, we can write this result as

$$\langle j\rangle \equiv \sum_{j=1}^{N} j\, p_j. \tag{32}$$

The utility of the above expression for the average of j is that it is easy to generalize. If the average of j is the probability weighted sum of j over all possible values of j, then we can infer that the average of any $f(j)$ (i.e., function of j) is the probability weighted sum of $f(j)$ over the possible values of j. Using our standard notation, the average of a quantity $f(j)$ is then

$$\langle f(j)\rangle \equiv \sum_{j=1}^{N} f(j)\, p_j. \tag{33}$$

The "average" operation Eq.(39) has some properties which will be useful later on. For constants c_1 and c_2 and functions $f_1(j)$ and $f_2(j)$, we can use Eq.(39) to show

$$\langle c_1 f_1(j) + c_2 f_2(j)\rangle = c_1\langle f_1(j)\rangle + c_2\langle f_2(j)\rangle, \tag{34}$$

namely the sum of an average is the average of the sum, and constant coefficients can be factored from our averages.

The most important application of Eq.(39) is to the calculation of $\langle j\rangle$, the mean of our random variable. As we discussed, the mean represents the value we would expect to obtain if we ran our probability trials (rolling the die in this case), multiple times, summed up all the values for the trials, and divided by the number of trials. The second most important application of Eq.(39) is to the calculation of how spread out the possible values of j are in our distribution. This quantity is termed the variance of j and is defined as

$$\sigma_j^2 \equiv \left\langle \left(j - \langle j\rangle\right)^2 \right\rangle. \tag{35}$$

The variance of a random variable is the average of the square of the difference between the value of the random variable and the mean of the random variable. We can find a simpler form for this quantity by using our previous definition of $\langle f(j)\rangle$. We have

$$\begin{aligned}\sigma_j^2 &= \sum_j \left(j - \langle j\rangle\right)^2 p_j \\ &= \sum_j \left(j^2 - 2j\langle j\rangle + \langle j\rangle^2\right) p_j \\ &= \sum_j j^2 p_j - \sum_j 2j\langle j\rangle p_j + \sum_j \langle j\rangle^2 p_j \end{aligned}$$

some finite range around the theoretical average.

$$= \langle j^2 \rangle - 2\langle j \rangle \sum_j j p_j + \langle j \rangle^2 \sum_j p_j$$
$$= \langle j^2 \rangle - 2\langle j \rangle^2 + \langle j \rangle^2, \tag{36}$$

or simply

$$\sigma_j^2 = \langle j^2 \rangle - \langle j \rangle^2. \tag{37}$$

For the case of our die, we can show that the variance of j has the value

$$\sigma_j^2 = \frac{107}{12} \simeq 8.9. \tag{38}$$

Because the variance is defined in terms of the square of $j - \langle j \rangle$, it does not precisely represent how spread out the distribution of values is about the mean. Rather, to get a spread with the correct dimensions, we need to take the square root of the variance. Doing so gives us the **standard deviation**. For the die, the standard deviation is

$$\sigma_j \simeq 2.98. \tag{39}$$

Roughly, the standard deviation provides a scale for how wide our probability distribution is. In an experiment with multiple trials, this standard deviation also sets the scale for how much our experimentally observed mean deviates from the theoretically calculated mean a result which has larger relevance in what is known as the **central limit theorem**, but exploring this fact will require some additional work in probability theory.

3.2 Continuous random variables

Now, that we have reviewed the basics of probability theory with discrete random variables, we can consider **continuous random variables**. We will take an example from microbiology.

E. coli is a species of bacteria that lives in the human intestine. *E. coli* reproduces by splitting itself into two identical copies. The time t at which this splitting occurs is random and thus constitutes a random variable for a probabilistic process. Because time is continuous, so too is our random variable t and thus the discrete summations considered above do not apply.

Instead, what we need is a way to define probabilities for continuous random variables. Because a continuous domain has an infinite number of values within any finite interval, we cannot simply define a probability to be at any one value, for any such probability would be zero. For example, if we have $p = 1/N$ for N possible values of the random variable then $p \to 0$ as $N \to \infty$. What we require is a new mathematical quantity called a **probability density function**. For our example of dividing *E. coli*, our probability density is a function of t and can be denoted $p(t)$. It is defined as follows:

> **Probability density $p(t)$ (defined):** If $p(t)$ is the probability density for the division time t of *E. coli*, then, for Δt sufficiently small, the probability that *E. coli* divides at a time between t and $t + \Delta$ is
> $$p(t)\Delta t. \tag{40}$$

The units of $p(t)$ are 1/time and hence it does not represent an *absolute* probability, but rather (as its name suggests) a density. To find an absolute probability, we have to relax the natural desire to find the probability of any single event (because single event probabilities are zero for continuous distributions) and satisfy ourselves with finding the probability to be in a space of values. This requires us to define a quantity called the **cumulative probability density**. Given that division times for *E. coli* must be greater than or equal to zero, we can define the cumulative probability that a division has occurred by time t as

$$P(t' < t) = \int_0^t dt'\, p(t'). \tag{41}$$

By the fundamental theorem of calculus, Eq.(47) tells us that the derivative of the cumulative probability is equal to the probability density:

$$\frac{d}{dt}P(t' < t) = p(t). \tag{42}$$

Moreover, because the *E. coli* cell must divide at some point, we know that as time goes to infinity, the probability of division must be 1. That is $P(t' < \infty) = 1$. With Eq.(47), we then have the normalization requirement

$$1 = \int_0^\infty dt'\, p(t'). \tag{43}$$

Eq.(49) is the continuous analog of Eq.(36). Similarly, analogous to Eq.(39) we can compute the average of any function of t as the probability-weighted integration of said function over all possible values of t:

$$\langle f(t)\rangle = \int_0^\infty dt'\, f(t')p(t'). \tag{44}$$

For example, the mean division time $\langle t\rangle$ can be computed from

$$\langle t\rangle = \int_0^\infty dt'\, t'\, p(t'). \tag{45}$$

Let us practice evaluating these averages using a concrete expression for the probability distribution of *E. coli*'s time of division. The distribution of division times for $E.coli$ is known to have an **exponential distribution**:

$$p(t) = \frac{1}{\tau}e^{-t/\tau}, \tag{46}$$

where $\tau = 30$ min[3].

Computing Eq.(51), using the identity in Eq.(52), we find

$$\langle t\rangle = \int_0^\infty dt\, \frac{t}{\tau}e^{-t/\tau} = \tau\int_0^\infty dx\, xe^{-x} = \tau, \tag{47}$$

where we used a change of variables in the first equality and integration by parts in the final equality. Thus, by this probability model, the mean division time of *E. coli* is $\tau = 30$ min. Let us compute the standard deviation. The formula Eq.(43) still applies in the case of continuous probability distributions. Applying it we have

$$\begin{aligned}\sigma_t^2 &= \langle t^2\rangle - \langle t\rangle^2 \\ &= \int_0^\infty dt\, \frac{t^2}{\tau}\, e^{-t/\tau} - \tau^2 \\ &= 2\tau^2 - \tau^2 = \tau^2.\end{aligned} \tag{48}$$

Therefore the standard deviation is $\tau = 30$ min, the same as the mean. Thus τ sets both the scale for the average division time of the bacterium and the width of the distribution of division times.

4 Combinatorics

In addition to being able to compute probabilities, we will need to know how to count. You might feel you are already quite adept at counting, but the type of counting we will need to do is of a rather more specialized kind. For example, we will be interested in determining the number of ways to arrange two

[3]Bacteria actually have a distribution of division times [1]. We choose a single value for the purposes of this example.

identical particles amongst N possible spaces for those particles, or in the ways to create pairings between multiple elements such that no pairing is from some special set of elements. In general, the problems of counting we will consider will concern a set of elements and determining the number of ways to construct groups of these elements according to some given constraint. The mathematics of how to perform this kind of counting is called **combinatorics**.

For our later work we will only need two definitions from this rather large field: permutations and combinations.

Say we have the three letters A, B, and C. A **permutation** of these letters is a particular ordering of them. (A, B, C) is one permutation and (C, A, B) is another. Given this definition, an obvious question is exactly how many permutations are there for a given collection of letters, or, more generally, elements. We can answer this question by counting the possibilities for each placement of a letter. We have three letters, so there are 3 possible choices for the entry in the first place. After we make this choice, there are 2 possible choices for the second place, and then 1 possible choice for the last place. Therefore, there are $3 \times 2 \times 1 = 6$ possible permutations of A, B, C.

This solution can be easily generalized to answer the question "How many permutations do we have of N distinct elements A_1, A_2, $\ldots$, A_N. By similar reasoning to that above, we find that the number of permutations of this N-distinct-elements set is

$$N \times (N-1) \times \cdots \times 2 \times 1. \tag{49}$$

This quantity is so important that it is given its own symbol in mathematics: We denote it $\boldsymbol{N!}$, and it is termed "N factorial".

Exercise: Permutations with repeated elements
Now, Eq.(55) only counts the number of ways to order N elements of a list when each element is unique. If an element appears multiple times in the list, we would need to divide Eq.(55), by the number of equivalent ways to reorder this collection of repeated elements. For example, if instead of A, B, and C in the above example, we had A, A, and C, then there are still 3 elements and 3! ways to order them, but because A appears twice, we can rearrange the As in 2! ways for each ordering without changing the actual permutation. So to find the *true* number of permutations of AAC we divide 3! by 2! to get 3. Similarly, if we had AAA, then we would find that there are $3!/3! = 1$ permutation. In general, if we have a list of N elements where element 1 occurs n_1 times, element 2 occurs n_2 times, and so on up until element M occurring n_M times, then the number of permutations of this list is

$$\frac{N!}{n_1! \times n_2! \times \cdots \times n_M!}, \qquad \text{where } n_1 + n_2 + \cdots + n_M = N. \tag{50}$$

From here we can ask a related question: what if we had N distinct elements and we wanted to form a list of $r \leq N$ elements. How many such lists could we create? We can answer this question by repeating the element-by-element multiplication procedure used above, but instead of going all the way to the last element, we stop after we have considered r elements Thus the number of ways to create ordered lists of r elements when we have N unique elements is

$$N \times (N-1) \times \cdots \times (N-(r-1)) = \frac{N!}{(N-r)!}. \tag{51}$$

4.1 Combinations

Now, that we have discussed permutations, we are now prepared to tackle combinations. We can think of a **combination** as a permutation where order does not matter. For example, while there are $3! = 6$ permutations of the letters A, B, and C, there is only *one* three-letter "combination" of these letters. At the level of combinations, the orderings (A, B, C), (C, A, B), and so on are equivalent because they are all composed of

the same three letters. Because of a combination's indifference to order, we can modify our previous results on permutations to obtain analogous results to similarly phrased questions on combinations.

Say we wanted to know the number of ways we could create three-letter combinations of the letters A, B, C, D, and E. We know from Eq.(57) that the number of ways to create a three-letter *ordered list* from five letters is $5!/(5-3)!$. But in moving from an ordered list to a combination, order is not important and so the $3!$ ways of ordering these three-letter lists are all equivalent. Thus, to account for the irrelevance of order we need to divide the permutation result Eq.(57) by the number of ways to re-order the chosen elements. For our example, this leaves us with $5!/3!(5-3)!$. More generally, the number of ways to select r elements from a list of N different elements is

$$\frac{N!}{r!(N-r)!}. \tag{52}$$

Eq.(59) is also important enough to have its own notation. We denote it as ${}_NC_r$ or $\binom{N}{r}$ and call it "N choose r".

Permutations and combinations form the foundation for many questions in the mathematics of counting but they also feature heavily in probability theory. As an example, consider the binomial theorem. Expressed in terms of the notation for "N choose r", we have

$$(x+y)^N = \sum_{r=0}^{N} \binom{N}{r} x^r y^{N-r}. \tag{53}$$

We can use Eq.(60) to determine the probability of getting r events of one kind when we perform N trials of an experiment with only two possible values. Let us say we had an unfair coin where there was a probability $3/4$ of getting heads and a probability $1/4$ of getting tails. Here is the question:

Question: If we flipped the coin 10 times, what is the probability of getting 5 heads in *any* order?

Because our random variable (the side the coin lands on) can take only one of two values, if we get 5 heads, we must also get $10-5=5$ tails. The probability of getting 5 heads and 5 tails in one particular ordering of heads and tails is

$$\text{Prob. of } 5 \text{ heads for } \textit{a particular} \text{ order of results} = \left(\frac{3}{4}\right)^5 \left(\frac{1}{4}\right)^5. \tag{54}$$

However, in answering the main question, we do not care about the order in which these results occur. Getting the event consisting of 5 heads in a row and then 5 tails in a row is unlikely, but it should be included in our probability. Similarly, the event of getting heads and tails in alternating fashion should also be included in our probability. Thus, the probability of getting 5 heads and 5 tails in *any* order is Eq.(61) multiplied by the number of ways to choose 5 events for a heads outcome out of a sequence of 10 events. This latter quantity is simply $\binom{10}{5}$, and so we have

$$\text{Prob. of } 5 \text{ heads in } \textit{any} \text{ order} = \binom{10}{5} \left(\frac{3}{4}\right)^5 \left(\frac{1}{4}\right)^5. \tag{55}$$

More generally, Eq.(62) arises from the set of distributions called **binomial distributions**. Taking $x=p$ to be the probability of getting heads and $y=1-p$ to be the probability of getting tails, we can write Eq.(60) as

$$1 = \sum_{k=0}^{N} \binom{N}{k} p^k (1-p)^{N-k} \equiv \sum_{k=0}^{N} P_N(k;p), \tag{56}$$

where

$$P_N(k;p) = \binom{N}{k} p^k (1-p)^{N-k}, \tag{57}$$

is the probability of getting k heads from N coin throws when the probability of a single heads is p. For the particular values of N, k, and p pertinent to our question we have

$$P_{10}(5; 3/4) = \binom{10}{5} \left(\frac{3}{4}\right)^5 \left(\frac{1}{4}\right)^5, \tag{58}$$

which reproduces Eq.(62).

References

[1] S. Cooper, "Bacterial growth and division," *Reviews in Cell Biology and Molecular Medicine*.