Matthew Schwartz
Statistical Mechanics, Spring 2019

Lecture 1: Probability

1 Basic probability

We are going to be dealing with systems with enormous degrees of freedom, typically governed by Avogadro's number $N_A = 6.02 \times 10^{23}$. This is the number of hydrogen atoms in a gram, or more intuitively, the number of molecules of water in a tablespoon. Even a tiny cell, with a diameter of only 100 microns ($10^{-4}\,m$), contains a trillion molecules. In most areas of physics, we work with small numbers (the fine structure constant $\alpha = \frac{1}{137}$ for example), and calculate things as a Taylor series in the coupling $f(\alpha) = \sum c_n \alpha^n$, often keeping only the leading term $f(\alpha) \approx c_1 \alpha$. In statistical mechanics, work with a large number N and calculate things as a Taylor expansion in $\frac{1}{N}$, often keeping only the leading term ($N = \infty$). The key to doing this is not to ask what each particle is doing, which would be both impossible and impractical, but rather to ask what the *probability* is that a particle is doing something. It is imperative therefore to begin statistical mechanics with statistics.

In general, we will be interested in probabilities of states of a system which we write as P_a or $P(a)$. The parameter a represents the microstate – e.g. the positions $\{\vec{q}_i\}$ and momenta $\{\vec{p}_i\}$ of all the particles in a gas, or the square of the wavefunction $|\psi(\vec{q})|^2$ in quantum mechanics. We will sometimes think of a as a discrete index (e.g. if we flip a coin, it can land heads up with $P_H = \frac{1}{2}$ or tails up with $P_T = \frac{1}{2}$) and sometimes continuous, writing $P(x)\,dx$ for the differential probability.

We will get to know a number of different probability distributions:

$$\text{Gaussian:} \quad P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-x_0)^2}{2\sigma^2}\right) \tag{1}$$

$$\text{Poisson:} \quad P_m(t) = \frac{t^m}{m!}\, e^{-t} \tag{2}$$

$$\text{Binomial:} \quad B_N(m) = a^m b^{N-m} \frac{N!}{m!(N-m)!} \tag{3}$$

$$\text{Lorentzian:} \quad P(x) = \frac{\Gamma}{2\pi} \frac{1}{(x-x_0)^2 + \left(\frac{\Gamma}{2}\right)^2} \tag{4}$$

$$\text{Flat:} \quad P(x) = \text{constant} \tag{5}$$

Probabilities are always normalized so that they integrate/sum to 1:

$$\int dx\, P(x) = 1, \qquad \sum_a P_a = 1 \tag{6}$$

Given a probability distribution, we can calculate the expected value of any observable by integrating/summing against the probability. For example, the expected value of x (the **mean**) is

$$\bar{x} \equiv \langle x \rangle = \int dx\, x\, P(x) \tag{7}$$

or the mean-square is

$$\langle x^2 \rangle = \int dx\, x^2\, P(x) \tag{8}$$

The **variance** of a distribution is the difference between the square of the mean and mean of the square

$$\text{Var} \equiv \langle x^2 \rangle - \langle x \rangle^2 \tag{9}$$

The square root of the variance is called the **standard deviation**.

$$\sigma \equiv \sqrt{\langle x^2 \rangle - \langle x \rangle^2} \tag{10}$$

While the mean has the intuitive interpretation as the expected outcome, variance is more subtle. Indeed, developing intuition for variance is a key to mastering statistics. The key point is that the expected value is worthless if you don't know how likely that value is.

For example, a Gaussian has two parameters, x_0 and σ_0. The first parameter is the mean:

$$\langle x\rangle = \int_{-\infty}^{\infty} dx\, x \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{(x-x_0)^2}{2\sigma_0^2}\right) = x_0 \tag{11}$$

The mean of x^2 is

$$\langle x^2\rangle = \sigma^2 + x_0^2 \tag{12}$$

So that the standard deviation is $\sigma = \sqrt{\langle x^2\rangle - \langle x\rangle^2} = \sigma_0$. This is why we usually just write σ instead of σ_0 for the parameter of the Gaussian.

The standard deviation has an interpretation as the width of a distribution – how far you can go from the mean before the probability has decreased substantially. For example, in a Gaussian, the probability of finding x between $x_0 - \sigma$ and $x_0 + \sigma$ is

$$(\Delta x)_{1\sigma} = \int_{x_0-\sigma}^{x_0+\sigma} dx\, P(x) = 0.68 \tag{13}$$

So, *for a Gaussian*, there is a 68% that the values of x fall within 1 standard deviation of the mean.

We will often be interested in situations where the mean is zero. Then the standard deviation is equivalent to the **root-mean-square**

$$x_{\text{RMS}} = \sqrt{\langle x^2\rangle} \tag{14}$$

For example, in a gas the velocities point in random directions, so $\langle \vec{v}\rangle = 0$. Thus the characteristic speed of a gas is characterized not be the mean but by the RMS velocity $v_{\text{RMS}} = \sqrt{\langle \vec{v}^2\rangle}$.

Another important concept is how probability distributions behave when they are combined. For example, say $P_A(x)$ and $P_B(y)$ are the probabilities of winning x dollars when betting on horse A and y dollars when betting on horse B. The probability of getting z total dollars is then

$$P_{AB}(z) = \int_{-\infty}^{\infty} dx\, P_A(z-x) P_B(x) \tag{15}$$

We say P_{AB} is the **convolution** of P_A and P_B and write it as

$$P_{AB} = P_A * P_B \tag{16}$$

Convolutions are extremely important in statistical mechanics, since we often measure only the sum of a great many independent processes. For example, the pressure on the wall of a container is due to the sum of the forces of all the little molecules hitting it, each with its own probability.

1.1 Examples

Consider the system of a gas molecule bouncing around in a 1D box of size L centered on $x = 0$. If there are no external forces and no position-dependent interactions, the molecule is equally likely to be anywhere in the box. So

$$P(x) = \frac{1}{L} \tag{17}$$

The mean value of the position of the molecule is

$$\langle x\rangle = \frac{1}{L}\int_{-\frac{L}{2}}^{\frac{L}{2}} dx\, x = 0 \tag{18}$$

Similarly, the mean value of x^2 is

$$\langle x^2\rangle = \frac{1}{L}\int_{-\frac{L}{2}}^{\frac{L}{2}} dx\, x^2 = \frac{L^2}{12} \tag{19}$$

So that the standard deviation is

$$\sigma = \sqrt{\langle x^2\rangle - \langle x\rangle^2} = \frac{1}{\sqrt{12}} L = 0.29\, L \tag{20}$$

Note that the probability of finding x within $\langle x\rangle \pm \sigma$ is $\frac{2\sigma}{L} = 58\%$. It is not 68% because the probability distribution is not Gaussian. This illustrates that the interpretation of σ as a 68% confidence interval is not always accurate.

Suppose instead that there is some electric field so that the particles in the box are more likely to be one side than the other. We might find some crazy function $P(x) = \frac{1.36}{L}\ln(1+e^{2x/L})$ for these probabilities. Then, by numerical integration we find

$$\langle x\rangle = 0.113\,L, \quad \langle x^2\rangle = 0.087\,L^2, \quad \sigma = 0.272\,L \tag{21}$$

with 61% within $\langle x\rangle \pm \sigma$. This is just a contrived example. It is easy to compute $\langle x\rangle$ and σ with any function $P(x)$, at least numerically.

2 Law of large numbers

An extremely important result from probability is that even if $P(x)$ is very complicated, when you average over many measurements, the result dramatically simplifies. More precisely, the law of large numbers states that

- If $P(x)$ has standard deviation σ, then the probability $P_N(x)$ of finding an average value x from N draws from $P(x)$ will have standard deviation $\frac{\sigma}{\sqrt{N}}$.

To derive the law of large numbers, lets consider the probability distribution for the center of mass of molecules in a box. Say there are N molecules in the box and the probability function of finding each is $P(x)$. Some examples for $P(x)$ are Section 1.1. We assume that the probabilities for each molecule are independent – having one at x does not tell us anything about where the others might be. In this case, what is the mean value of the center of mass of the system? We'll write $\langle x\rangle_N$, $\langle x^2\rangle_N$ and σ_N for quantities involving the N-body system and drop the subscript for the $N=1$ case: $\langle x\rangle_1 = \langle x\rangle$ and $\sigma_1 = \sigma$.

For $N=2$, the center of mass is $x = \frac{x_1+x_2}{2}$, so the mean value of the center of mass is

$$\langle x\rangle_2 = \int_{-\frac{L}{2}}^{\frac{L}{2}} dx_1 \int_{-\frac{L}{2}}^{\frac{L}{2}} dx_2 P(x_1)P(x_2)\frac{x_1+x_2}{2} = \int_{-\frac{L}{2}}^{\frac{L}{2}} dx_1 P(x_1)\frac{x_1}{2} + \int_{-\frac{L}{2}}^{\frac{L}{2}} dx_2 P(x_2)\frac{x_2}{2} = \langle x\rangle \tag{22}$$

So the mean value for 2 molecules is the same as for 1 molecule. The expectation of x^2 with 2 molecules is

$$\langle x^2\rangle_2 = \int_{-\frac{L}{2}}^{\frac{L}{2}} dx_1 \int_{-\frac{L}{2}}^{\frac{L}{2}} dx_2 P(x_1)P(x_2)\Big(\frac{x_1+x_2}{2}\Big)^2 \tag{23}$$

$$= \frac{1}{4}\int_{-\frac{L}{2}}^{\frac{L}{2}} dx_1 P(x_1)x_1^2 + \frac{1}{2}\int_{-\frac{L}{2}}^{\frac{L}{2}} dx_1 P(x_1)x_1 \int_{-\frac{L}{2}}^{\frac{L}{2}} dx_2 P(x_2)x_2 + \frac{1}{4}\int_{-\frac{L}{2}}^{\frac{L}{2}} dx_2 P(x_2)x_2^2 \tag{24}$$

$$= \frac{1}{2}\langle x^2\rangle + \frac{1}{2}\langle x\rangle^2 \tag{25}$$

So the standard deviation of the center-of-mass for 2 particles is:

$$\sigma_2 = \sqrt{\langle x^2\rangle_2 - (\langle x\rangle_2)^2} = \sqrt{\frac{1}{2}\langle x^2\rangle + \frac{1}{2}\langle x\rangle^2 - \langle x\rangle^2} = \frac{1}{\sqrt{2}}\sqrt{\langle x^2\rangle - \langle x\rangle^2} = \frac{\sigma}{\sqrt{2}} \tag{26}$$

That is, the standard deviation has shrunk by a factor of $\sqrt{2}$ from the one particle case *for any* $P(x)$.

Now say there are N particles. The mean value of the center of mass is

$$\langle x\rangle_N = \int_{-\frac{L}{2}}^{\frac{L}{2}} dx_1 \cdots dx_N P(x_1)\cdots P(x_N)\Big(\frac{x_1+\cdots+x_N}{N}\Big) = \frac{1}{N}\Bigg[N\int_{-\frac{L}{2}}^{\frac{L}{2}} dx_1 P(x_1)\Bigg] = \langle x\rangle \tag{27}$$

independent of N. The expectation value of x^2 is

$$\langle x^2\rangle_N = \int_{-\frac{L}{2}}^{\frac{L}{2}} dx_1 \cdots dx_N P(x_1)\cdots P(x_N)\Big(\frac{x_1+\cdots+x_N}{N}\Big)^2 \tag{28}$$

When we expand $(x_1+\cdots+x_N)^2$ there are N terms that give $\langle x^2\rangle$ and the remaining (N^2-N) terms are the same as $\langle x_1 x_2\rangle = \langle x\rangle^2$. So,

$$\langle x^2\rangle_N = \int_{-\frac{L}{2}}^{\frac{L}{2}} dx_1 \cdots dx_N P(x_1)\cdots P(x_N)\frac{1}{N^2}[Nx_1^2 + (N^2-N)x_1x_2] \tag{29}$$

$$= \frac{1}{N}\langle x^2\rangle + \left(1-\frac{1}{N}\right)\langle x\rangle^2 \tag{30}$$

Therefore

$$\sigma_N = \sqrt{\langle x^2\rangle_N - \langle x\rangle^2} = \frac{1}{\sqrt{N}}\sqrt{\langle x^2\rangle - \langle x\rangle^2} = \frac{\sigma}{\sqrt{N}} \tag{31}$$

The appearance of $\sqrt{N}$ is called **the law of large numbers**. Note that Eq. (31), describing how the standard deviation scales as we average over many molecules, holds for any function $P(x)$. Different $P(x)$ will give different values of σ, but the relation between σ_N with N molecules and σ with one molecule is universal.

For the gas in the box with a flat $P(x) = \frac{1}{L}$, as in Section 1.1, the expected value of the center of mass is $\langle x\rangle_N = 0$, just like for any individual gas molecules, and the standard deviation is $\sigma_N = \frac{\sigma}{\sqrt{N}} \approx 10^{-11}\frac{L}{\sqrt{12}}$. Thus, even though we don't know very well where any of the molecules are, we know the center of mass to extraordinary precision.

The law of large numbers is the reason that statistical mechanics is possible: we can compute macroscopic properties of systems (like the center of mass, or pressure, or all kinds of other things) with great confidence even if we don't know exactly what is going on at the microscopic level.

3 Central Limit Theorem

We saw how for when we average over a large number N of draws from a probability distribution $P(x)$ the mean stays fixed and the standard deviation shrinks by $\sigma \to \frac{\sigma}{\sqrt{N}}$. What can we say about the shape of the probability distribution $P_N(x)$? It turns out we can say a lot. In fact, in the limit $N \to \infty$ we know $P_N(x)$ exactly: it is a Gaussian! This surprising result is called the central limit theorem.

One way to prove the central limit theorem is by computing moments. If you specify the complete set of moments of a function, you know its shape completely. These moments are

$$\text{mean:} \quad \bar{x} = \langle x\rangle \tag{32}$$

$$\text{variance:} \quad \sigma^2 = \langle x-\bar{x}\rangle^2 = \langle x^2\rangle - \bar{x}^2 \tag{33}$$

$$\text{skewness:} \quad S = \frac{\langle (x-\bar{x})^3\rangle}{\sigma^3} = \frac{1}{\sigma^3}[\langle x^3\rangle - 3\bar{x}\langle x^2\rangle + 2\bar{x}^3] \tag{34}$$

$$\text{kurtosis:} \quad K = \frac{\langle (x-\bar{x})^4\rangle}{\sigma^4} \tag{35}$$

$$n^{\text{th}}\,\text{moment:} \quad M_n = \frac{\langle (x-\bar{x})^n\rangle}{\sigma^n} \tag{36}$$

Skewness measures how asymmetric a distribution is around its mean. Kurtosis measures the 4th derivative, which is a measure of curvature. More intutively, higher kurtosis means a probability distribution has a longer tail, i.e. more outliers from the mean. The higher moments do not have simple interpretations.

Notice that all the higher-order moments are normalized by dividing by powers of σ so that they are dimensionless. To understand this normalization imagine plotting $P_N(x)$, but shift it to center around $x=0$ and rescale the x axis by σ so that the width is always 1. Then the curve will not get any smaller as $N\to 0$ because it's width is fixed to be 1, but its shape may change. The shape is determined by the numbers M_n with $n>2$. See Fig. 2 below for an example.

For the Gaussian probability distribution in Eq. (1) the moments are easy to calculate in Mathematica:

$$\bar{x}=0,\quad \sigma=\sigma,\quad S=0,\quad K=3,\quad M_5=0,\quad M_6=15,\quad M_7=0,\quad M_8=105,\cdots(\textbf{Gaussian}) \tag{37}$$

Note that skewness is zero for a Gaussian because it is symmetric. For a Gaussian, in fact all the odd moments (M_n with n odd) vanish. The even moments, normalized to powers of σ, are dimensionless parameters completely determining the shape of a Gaussian. If a function has all of these moments, it is a Gaussian.

Now let's compute the moments of the center of mass of our N molecules-in-a-box with probability $P(x)$. We'll do this for a general $P(x)$, but shift the domain so that $\langle x\rangle=\bar{x}=0$ in order to simplify the formulas in Eqs. (33)-(36). For example, the 3rd moment of $P_N(x)$ is

$$\langle x^3\rangle_N=\int_{-\frac{L}{2}}^{\frac{L}{2}}dx_1\cdots dx_N P(x_1)\cdots P(x_N)\Big(\frac{x_1+\cdots+x_N}{N}\Big)^3 \tag{38}$$

Since $\langle x\rangle=0$ the only terms in this expression which don't vanish are the ones of the form x_j^3. So

$$\langle x^3\rangle_N=\frac{1}{N^2}\langle x^3\rangle \tag{39}$$

We conclude that the skewness S_N with N molecules is related to the skewness S_1 for 1 molecule by

$$S_N=\frac{\langle(x-\bar{x})^3\rangle_N}{\sigma_N^3}=\frac{S_1/N^2}{(\sigma/\sqrt{N})^3}=\frac{S_1}{\sqrt{N}} \tag{40}$$

In particular, the skewness goes to zero as $N\to\infty$. That is, the distribution becomes more and more symmetric abound the mean as $N\to\infty$.

Now let's look at the 4th moment, kurtosis. Following the same method we need

$$\langle x^4\rangle_N=\int_{-\frac{L}{2}}^{\frac{L}{2}}dx_1\cdots dx_N P(x_1)\cdots P(x_N)\Big(\frac{x_1+\cdots+x_N}{N}\Big)^4 \tag{41}$$

In this case, since $\langle x\rangle=0$, the terms that don't vanish are x_j^4 or $x_j^2x_i^2$ with $i\neq j$. Thinking about the combinatorics a little you can convince yourself that there N terms of the form x_i^4 and $3N(N-1)$ terms of the form $x_i^2x_j^2$.[1] So,

$$\langle x^4\rangle_N=\frac{1}{N^3}\langle x^4\rangle+\frac{3(N-1)}{N^3}\langle x^2\rangle\langle x^2\rangle \tag{42}$$

Then, calling $K_1=\frac{1}{\sigma^4}\langle x^4\rangle$ the kurtosis for $N=1$ we have

$$K_N=\frac{\langle(x-\bar{x})^4\rangle_N}{\sigma_N^4}=\frac{1}{\sigma^4/N^2}\Big[\frac{1}{N^3}\langle x^4\rangle+\frac{3(N-1)}{N^3}\langle x^2\rangle\langle x^2\rangle\Big]=\frac{K_1}{N}+3\Big(1-\frac{1}{N}\Big) \tag{43}$$

This is interesting – it says that as $N\to\infty$ the kurtosis $K_N\to 3$ *independent* of the kurtosis of the one particle probability distribution! So the skewness goes to zero and the kurtosis goes to 3.

For the 6th moment the term which dominates at large N is the non-vanishing one with the largest combinatoric factor: $\langle x^2\rangle^3$. There are ${}_NC_3\times{}_6C_2\times{}_4C_2=\frac{1}{6}N(N-1)(N-2)\times 15\times 2\to 15$ of these. So $(M_6)_N\to 15$. Similarly, $(M_8)_N\to 105$. In other words, for any $P(x)$ we find that as $N\to\infty$

$$S_N\to 0,\quad K_N\to 3,\quad (M_5)_N\to 0,\quad (M_6)_N\to 15,\quad (M_7)_N\to 0,\quad (M_8)_N\to 105,\quad \cdots \tag{44}$$

What we are seeing is that at large N all of the higher moments go to those of a Gaussian!

1. There are $\binom{N}{1}=N$ of the x_j^4 terms. There are $\binom{N}{2}=\frac{N!}{2!(N-2)!}=\frac{N(N-1)}{2}$ possible pairs $i\neq j$ and there are $\binom{4}{2}=6$ ways of picking which two of the 4 terms in the expansion are i. So the total number of these terms is $3N(N-1)$.

This is the **central limit theorem**:

> When *any* probability distribution is sampled N times
> the average of the samples approaches a Gaussian distribution as $N \to \infty$
> with width scaling like $\sigma \sim \frac{1}{\sqrt{N}}$

There are lot of other ways to prove it, but I find the moment approach the most accessible.

3.1 Combining flat distributions

Because the central limit theorem is so important, let's try to understand why it is true more physically. Again, say we have some probability distribution $P(x)$ for molecules in a box, with $-\frac{L}{2} < x < \frac{L}{2}$. We want to pick N molecules and compute their mean position (center of mass position) $x = \frac{1}{N}\sum_j x_j$. What is the probability distribution $P_N(x)$ that the mean value is x?

To be concrete, let's take the flat distribution $P(x) = \frac{1}{L}$. For $N = 1$, we pick only molecule with position x_1. Then $x = x_1$ and so $P(x) = \frac{1}{L}$: any value for the center-of-mass position is equally likely.

Now say $N = 2$, so we pick two molecules with positions x_1 and x_2. What is the probability that they will have mean x? For a given x we need $\frac{x_1 + x_2}{2} = x$. For example if $x = 0$, then for any x_1 there is an x_2 that works, namely $x_2 = -x_1$. However, if the mean is all the way on the edge, $x = \frac{L}{2}$, then not all x_1 work; in fact, we need both x_1 and x_2 to be exactly $\frac{L}{2}$. Thus there are fewer possibilities with x is close to the boundaries of the box than if x is central. One way to see this is graphically

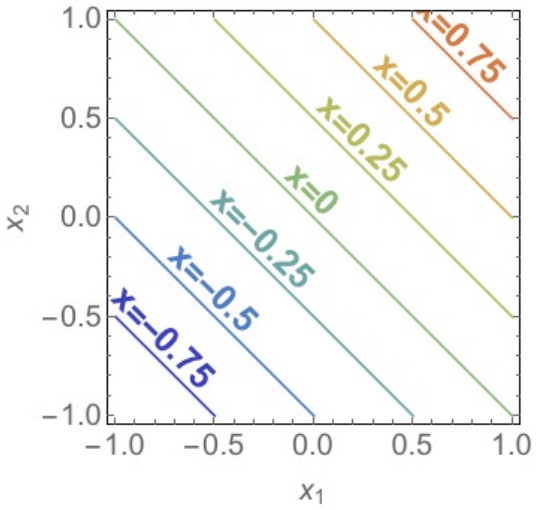


Figure 1. The regions in the x_1/x_2 plane with mean value x are diagonal lines for $L = 2$. The length of the line is the probability $P_2(x)$. For $x = 0$, the line is longest and probability greatest. For $x = 1$, the line reduces to a point and the probability to zero.

3.2 δ-function aside

To be quantitative, the easiest way to calculate the probability is with the Dirac δ function $\delta(x)$. Recall that the δ-function is not really a function but rather a distribution. $\delta(x)$ is zero everyone except at $x = 0$. When you integrate a function against $\delta(x)$ you pick up the value of that function at 0. That is

$$\int dx \delta(x) f(x) = f(0) \tag{45}$$

This is the defining property of $\delta(x)$. The integration region has to include $x = 0$ but is otherwise arbitrary since $\delta(x) = 0$ if $x \neq 0$.

Another useful property of δ-functions is that if we rescale the argument of $\delta(x)$ by a number a then the δ-function rescales by $\frac{1}{a}$. That is,

$$\delta(ax) = \frac{1}{a}\delta(x) \tag{46}$$

To check this, we can change variables from $x \to \frac{x}{a}$ in the integral

$$\int dx\delta(ax)f(x) = \int d\frac{x}{a}\delta(x)f\Big(\frac{x}{a}\Big) = \frac{1}{a}f(0) = \int dx\Big[\frac{1}{a}\delta(x)\Big]f(x) \tag{47}$$

It's sometimes helpful to think of the δ function as the limit of a regular function. There are lots functions whose limits are δ functions. For example, Gaussians:

$$\delta(x) \to \lim_{\sigma\to 0}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}} \tag{48}$$

As a check, note that the integral over the Gaussian is 1 regardless of σ, so the δ function also integrates to 1. As $\sigma \to 0$, the width of the Gaussian goes to zero, so it has zero support away from mean, that is it vanishes except at $x = 0$, just like the δ function.

3.3 Back to flat probability

Using the δ-function, we can write the probability for getting a mean value $x = \frac{x_1+x_2}{2}$ as

$$P_2(x) = \int_{-\frac{L}{2}}^{\frac{L}{2}} dx_1 P(x_1)\int_{-\frac{L}{2}}^{\frac{L}{2}} dx_2\, P(x_2)\delta\Big(\frac{x_1+x_2}{2} - x\Big) \tag{49}$$

This is another way of writing a convolution, as in Eq. (15): $P_2 = P * P$.

As a check, we can verify that this probability distribution is normalized correctly

$$\begin{aligned}\int_{-\frac{L}{2}}^{\frac{L}{2}} dx\, P_2(x) &= \int_{-\frac{L}{2}}^{\frac{L}{2}} dx \int_{-\frac{L}{2}}^{\frac{L}{2}} dx_1 P(x_1)\int_{-\frac{L}{2}}^{\frac{L}{2}} dx_2\, P(x_2)\delta\Big(\frac{x_1+x_2}{2} - x\Big)\\ &= \int_{-\frac{L}{2}}^{\frac{L}{2}} dx_1 P(x_1)\int_{-\frac{L}{2}}^{\frac{L}{2}} dx_2\, P(x_2) = 1\end{aligned} \tag{50}$$

where we have used the δ-function to integrate over x to get to the second line.

To evaluate $P_2(x)$ we first pull a factor of 2 out of the δ-function using Eq. (46), giving

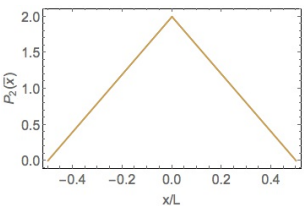
$$P_2(x) = 2\int_{-\frac{L}{2}}^{\frac{L}{2}} dx_1 P(x_1)\int_{-\frac{L}{2}}^{\frac{L}{2}} dx_2\, P(x_2)\delta(x_1+x_2-2x) \tag{51}$$

Now, the δ-function can only fire if its argument hits zero in the integration region. Since $\frac{x_1+x_2}{2} = x$ we can solve for $x_1 = 2x - x_2$. If $x < 0$ then the most x_1 can be is $2x - \Big(-\frac{L}{2}\Big) = \frac{L}{2} + 2x$. In other words, we have

$$P_2(x<0) = 2\int_{-\frac{L}{2}}^{\frac{L}{2}+2x} dx_1 P(x_1)P(2x-x_1) \tag{52}$$

Taking the flat distribution $P(x) = \frac{1}{L}$ this evaluates to $P_2(x<0) = 2L + 4x$. Similarly, for $x > 0$ the limit is $x_1 > 2x - \frac{L}{2}$ and for a flat distribution $P_2(x>0) = 2L - 4x$. Thus we have

$$L^2 P_2(\bar{x}) = \begin{cases} 2L+4x, & x<0 \\ 2L-4x, & x>0 \end{cases} = \tag{53}$$



You can also check this by evaluating Eq. (49) with Mathematica:

```
P=Integrate[DiracDelta[x1+x2-2x],{x1,-1,1},{x2,-1,1}];
Plot[P, {x, -1, 1}]
```

For $N=3$ we compute

$$P_3(x)=\int_{-\frac{L}{2}}^{\frac{L}{2}}dx_1P(x_1)\int_{-\frac{L}{2}}^{\frac{L}{2}}dx_2\,P(x_2)\int_{-\frac{L}{2}}^{\frac{L}{2}}dx_3\,P(x_3)\delta\Big(\frac{x_1+x_2+x_3}{3}-x\Big) \tag{54}$$

and so on. These successive approximations look like

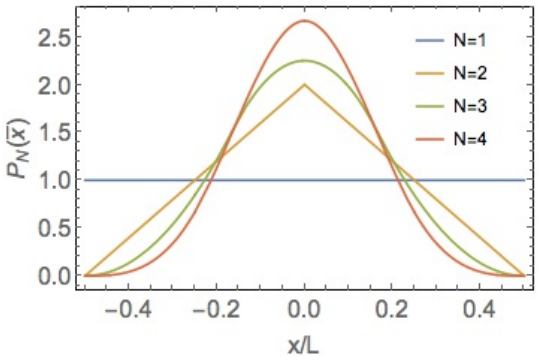


Figure 2. The average position of $N=1,2,3,4$ particles, each of which separately has a flat probability distribution.

We see that already at $N=4$ the flat probability distribution is becoming a Gaussian. Note also that the widths of the distributions are getting narrower.

The **central limit theorem** says that the distribution of the mean of N draws from a probability distribution approaches a Gaussian of width $\frac{\sigma}{\sqrt{N}}$ as $N\to\infty$ *independent* of the original probability distribution. That is,

$$P_N(x)\to\sqrt{\frac{N}{2\pi\sigma^2}}\exp\Big(-N\frac{(x-\bar{x})^2}{2\sigma^2}\Big) \tag{55}$$

Sometimes we add the values of the draws from a distribution instead of averaging them. In this case, the mean grows as $\bar{x}\to N\bar{x}$ and the standard deviation grows like $\sigma\to\sqrt{N}\sigma$. Thus an equivalent phrasing of the central limit theorem is

- **Central Limit Theorem**: A function with mean $\bar{x}$ and standard deviation σ convolved with itself N times approaches a Gaussian with mean $N\bar{x}$ and standard deviation $\sqrt{N}\sigma$ as $N\to\infty$.

Adding the values is what happens when you convolve a function with itself. So for adding the values, the central limit theorem has the form

$$P_N^{\text{sum}}(x)=\underbrace{P\star P\star\cdots\star P}_{N}\to\frac{1}{\sqrt{2\pi\sigma^2N}}\exp\Big(-\frac{(x-N\bar{x})^2}{2\sigma^2N}\Big) \tag{56}$$

3.4 Central limit theorem and Taylor expansions

In statistical mechanics, we will make great use out of the central limit theorem. Generally we have systems composed of enormously large numbers of particles $N\sim$ Avogadro's number $\sim 10^{24}$. The things we measure are macroscopic: the pressure a gas puts on a wall is the *average* pressure. Microscopically, the gas has a bunch of little molecules hitting and bouncing off the wall and the force these molecules impart is constantly varying. We don't care about these tiny fluctuations, just the average. So any time we try to measure something, like the pressure in a gas, or the concentration of a chemical, we will necessarily be averaging over an enormous number of fluctuations. Because of the central limit theorem, the distribution of any macroscopic quantity will be close to a Gaussian around its mean. This central limit theorem itself doesn't tell us what the mean is, or how various macroscopic quantities are related – we need physics for that. But it tells us that we don't need to worry about the precise details of the microscopic description.

Normally when a function $f(x)$ is rapidly falling away from $x \approx \bar{x}$ we Taylor expand $x = \bar{x}$ and keep the first few terms. We can do this for $P_N(x)$ too. However, the Taylor expansion of a Gaussian has an infinite number of terms

$$e^{-\frac{x^2}{2\sigma^2}} = \sum_{m=0}^{\infty} \frac{1}{m!}\left(-\frac{x^2}{2\sigma^2}\right)^m = 1 - \frac{x^2}{2\sigma^2} + \frac{1}{2}\left(\frac{x^2}{2\sigma^2}\right)^2 - \frac{1}{6}\left(\frac{x^2}{2\sigma^2}\right)^3 + \cdots \tag{57}$$

You need all the terms to reconstruct the original Gaussian. However, if we take the logarithm first, then Taylor expand, we find

$$\ln e^{-\frac{x^2}{2\sigma^2}} = -\frac{x^2}{2\sigma^2} \tag{58}$$

with only one term. So it will be extremely convenient to start taking the logarithms of our probabilities. By the central limit theorem, when we average the values,

$$\ln P_N(x) \to -N\frac{(x-\bar{x})^2}{2\sigma^2} + \ln\sqrt{\frac{N}{2\pi\sigma^2}} \tag{59}$$

As $N \to \infty$ there are no higher order terms.

In other words, a very sharp Gaussian is not very smooth – it has a huge sharp peak, like a δ function, so it is hard to Taylor expand. Taking the logarithm makes the function much smoother.

4 Poisson distribution

In many physical situations, there is a large number N of possible events each occurring with very small probability λ. For example, if you put a glass out in the rain, there are lots of possible drops of water that could fall into the glass, but each has a small probability. Or you have lots of friends on Instagram, each one has a small probability of posting something interesting. Or we have a gas of molecules and each one has a small chance of being in some tiny volume. Probabilities in situations like this, where each event is uncorrelated with the previous event, are described by the Poisson distribution.

Let's take a concrete example, radioactive decay. A block of ^{235}U has $N \sim 10^{24}$ atoms each of which can decay with a tiny probability

$$dP = \lambda dt \tag{60}$$

λ is called the **decay rate**. It has units of $\frac{1}{\text{time}}$. For a single atom of ^{235}U, this decay rate is $\lambda = 3 \times 10^{-17}\, s^{-1}$. In a mole of Uranium ($10^{24}$ atoms), 10^7 Uranium atoms decay, on average, each second. What is the chance of seeing m decays in a time t?

Let's start with $m = 0$ and the time t very small, $t = \Delta t$. If the rate to decay is $dP = \lambda dt$ then the probability of not decaying in time $t = \Delta t$ is

$$P_{\text{nodecay}}(\Delta t) = 1 - \lambda \Delta t \tag{61}$$

For the system to survive to a time $2\Delta t$ with no decays, it would have to not decay in Δt and then not decay again in the next Δt. Since the probability of two uncorrelated occurrences (or not-occurrences in this case) is the product of the probabilities, $P(a\&b) = P(a)P(b)$ we then have

$$P_{\text{nodecay}}(2\Delta t) = (1 - \lambda \Delta t)^2 \tag{62}$$

Now we can get all the way to time t by sewing together small times $\Delta t = \frac{t}{N}$ and taking $N \to \infty$. We thus have

$$P_{\text{nodecay}}(t) = \lim_{N\to\infty}\left(1 - \lambda\frac{t}{N}\right)^N = e^{-\lambda t} \tag{63}$$

So that's the $m = 0$ case: no particles decay.

Using this formula, how long will it take for the probability of some decay to be $\frac{1}{2}$? That's the same as the probability of no decay being $1 - \frac{1}{2} = \frac{1}{2}$. So we just solve

$$\frac{1}{2} = e^{-\lambda t_{1/2}} \qquad \Rightarrow \qquad t_{1/2} = \frac{1}{\lambda}\ln 2 = \frac{0.7}{\lambda} \tag{64}$$

We often say $\frac{1}{\lambda}$ is the **lifetime** and $t_{1/2}$ is the **halflife**. The two numbers are related by a factor of ln2: $t_{1/2} = \frac{1}{\lambda}\ln 2$.

Now try $m = 1$. We need the probability that there is exactly one decay in exactly one of the time intervals. There are N intervals we can pick. So

$$P_{1\,\text{decay}}(t) = \lim_{N\to\infty} N\underbrace{\Big(1-\lambda\frac{t}{N}\Big)^{N-1}}_{N-1\,\text{nodecays}}\underbrace{\Big(\lambda\frac{t}{N}\Big)}_{\text{onedecay}} = -t\,\partial_t P_{\text{nodecay}}(t) = \lambda t e^{-\lambda t} \tag{65}$$

For two decays there are $\binom{N}{2} = \frac{N!}{(N-2)!2!} = \frac{1}{2}N(N-1)$ ways and we have

$$P_{2\text{decays}}(t) = \lim_{N\to\infty} \underbrace{\frac{N(N-1)}{2}}_{\text{pick2todecay}}\underbrace{\Big(1-\lambda\frac{t}{N}\Big)^{N-2}}_{N-2\,\text{nodecays}}\underbrace{\Big(\lambda\frac{t}{N}\Big)^{2}}_{\text{twodecays}} = \frac{1}{2}t^2\,\partial_t^2 P_{\text{nodecay}}(t) = \frac{(\lambda t)^2}{2}e^{-\lambda t} \tag{66}$$

For general m the result is

$$\boxed{P_m(t) = \frac{(\lambda t)^m}{m!}e^{-\lambda t}} \tag{67}$$

This is called the **Poisson distribution**. It gives the probability for exactly m events in time t when each event has a probability per unit time of λ and the events are uncorrelated.

In any time t there must have been some number of decays between 0 and ∞. Indeed,

$$\sum_m P_m(t) = \sum \frac{(\lambda t)^m}{m!}e^{-\lambda t} = 1 \tag{68}$$

So that's consistent (as is the t-independence of this sum).

The way we derived the Poisson distribution was for a fixed m, as a function of t. But it can be more useful to think of it as a function of m at a fixed value of t: $P(m,t) = P_m(t)$. Keep in mind though that for fixed t, $P(m,t)$ as a function of m is a discrete probability distribution (meaning m is an integer). In contrast for fixed m, $P(m,t)$ is a continuous function of t.

For a given fixed t, how many particles do we expect to have decayed? In other words, what is the expected value $\langle m\rangle$ in a time t? We compute the mean value for m, by summing the value of m times the probability of getting m

$$\langle m\rangle = \sum_m m P_m(z) = \sum_m m\frac{(\lambda t)^m}{m!}e^{-\lambda t} = \lambda t \tag{69}$$

The last step is a little tricky – see if you can figure out how to do the sum yourself. (You can always run Mathematica if you get stuck on steps like this.) The result implies that the expected number of decays in a time t is λt. It makes sense that if you double the time, twice as many particles decay. How long will it take for half the particles to decay?

The standard deviation of the Poisson distribution is

$$\sigma = \sqrt{\langle m^2\rangle - \langle m\rangle^2} = \sqrt{\lambda t} \tag{70}$$

Again, you can check this yourself as an exercise.

So the Poisson distribution as a function of m at fixed t has mean λt and width $\sqrt{\lambda t}$. Thus the width compared to the mean is

$$\frac{\sigma}{\langle m\rangle} = \frac{1}{\sqrt{\lambda t}} \tag{71}$$

This goes to 0 as $t\to\infty$. In other words, the Poisson distribution is narrower and narrower as t gets larger. What does this mean physically? It means if we wait one lifetime ($t = \frac{1}{\lambda}$) we should expect 1 ± 1 particle to decay. If we wait 2 lifetimes, we expect $2\pm\sqrt{2}$ to decay. If we wait 100 lifetimes, we expect 100 ± 10 to decay. So the longer we wait, not only are there more decays, but we know more precisely how many decays there will be. This is, of course, a consequence of the central limit theorem.

So what do you expect the distribution to look like as $t\to\infty$ or $m\to\infty$? Let's first look numerically. We can plot $P_m(t)$ as a function m, which is a discrete index, or as a function of t, which is continuous:

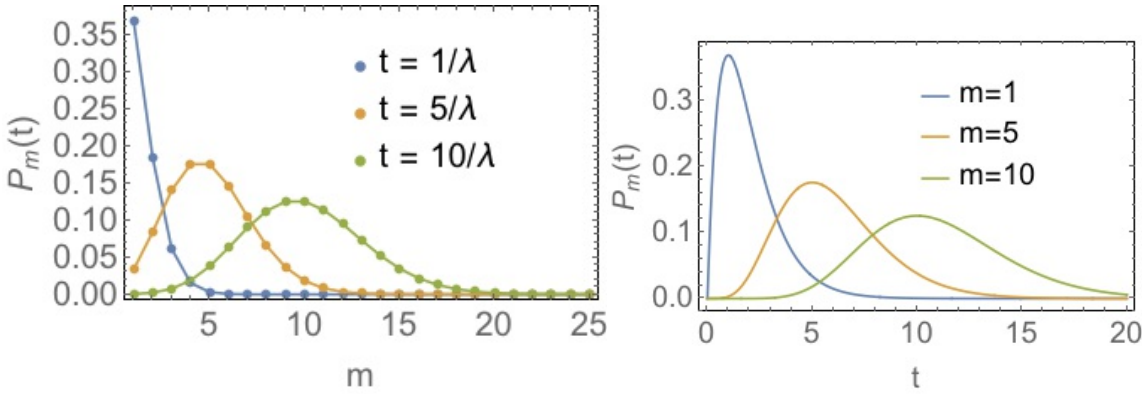


Figure 3. The Poisson distribution as a function of the discrete index m for various times (left) and time, for various values of m (right)

We clearly see the Gaussian shape emerging at large t (left) and at large m (right).

Now let's try to see how the Gaussian form arises analytically. First of all, we want the high statistics limit, which means large t in units of $\frac{1}{\lambda}$ which also means large m. When you see a factor of $m!$ as in the Poisson distribution want to expand at large m, you should immediately think **Stirling's approximation**:

$$x! \approx e^{-x} x^x \times (\cdots) \tag{72}$$

or equivalently

$$\ln x! \approx x \ln x - x + \cdots \tag{73}$$

For a simple derivation, see Section 5. We will use this expansion a lot.

The log of the Poisson distribution is

$$\ln P_m(t) = \ln\left[\frac{(\lambda t)^m}{m!} e^{-\lambda t}\right] = m \ln(\lambda t) - \lambda t - \ln m! \tag{74}$$

Then we use Stirling's approximation for $m!$

$$\ln P_m(t) \xrightarrow[m \gg 1]{} m \ln(\lambda t) - \lambda t - m \ln m + m + \cdots \tag{75}$$

This is still a mess. But we expect $P_m(t)$ to be peaked around its mean $\langle m \rangle = \lambda t$. So let's Taylor expand $\ln P_m(t)$ around $m = \lambda t$. The leading term, from setting $m = \lambda t$ makes Eq. (75) vanish. The next term is

$$\frac{\partial}{\partial m} \ln P_m(t)\Big|_{m=\lambda t} = \lim_{m \to \lambda t} [\ln(\lambda t) - \ln m] = 0 \tag{76}$$

which also vanishes. We have to go one more order in the Taylor expansion to get a nonzero answer:

$$\frac{\partial^2}{\partial m^2} \ln P_m(t)\Big|_{m=\lambda t} = \lim_{m \to \lambda t} \left[-\frac{1}{m}\right] = -\frac{1}{\lambda t} \tag{77}$$

Thus,

$$\ln P_m(t) = -\frac{1}{2\lambda t}(m - \lambda t)^2 + \cdots \tag{78}$$

and therefore

$$P_m(t) \xrightarrow[m \gg 1]{} \frac{1}{\sqrt{2\pi\lambda t}} e^{-\frac{(m-\lambda t)^2}{2\lambda t}} \tag{79}$$

This is a Gaussian with mean $\langle m \rangle = \lambda t$ and width $\sigma = \sqrt{\lambda t}$ exactly as expected by the central limit theorem.

You might not be terribly impressed with this derivation as a check of the central limit theorem. After all, we expanded $\ln P_m$ to second order around $m = \langle m \rangle$. Doing that, for any function P_m is guaranteed to give a Gaussian. But that's really the whole point of the central limit theorem – any function *does* give a Gaussian. So in the end you should be impressed after all.

5 Appendix: Stirling's approximation

There are many ways to derive Stirling's approximation. Here's a relatively easy one. We start by taking the logarithm

$$\ln N! = \ln N + \ln(N-1) + \ln(N-2) + \cdots + \ln 1 = \sum_{j=1}^{N} \ln j \tag{80}$$

For large N we then write the sum as an integral

$$\ln N! = \sum_{j=1}^{N} \ln j \approx \int_1^N dj \ln j = N \ln N - N - 1 \approx N \ln N - N \tag{81}$$

That's the answer.

One can include more terms in the expansion by using the Euler-McLauren formula for the difference between a sum and an integral. For example, the next term is $N! = \sqrt{2\pi N} N^N e^{-N}$. Already at $N = 10$, Stirling's approximation is off by only 1% ($N! = 3.60 \times 10^6$ while $\sqrt{2\pi N} N^N e^{-N} = 3.63 \times 10^6$), and we will be taking $N = 10^{24}$.