Matthew Schwartz
Statistical Mechanics, Spring 2025

# 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, we 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. In the continuous case, we call $P(x)$ the probability density, so that $\int_{x_1}^{x_2} P(x)\,dx$ is the probility of finding $x$ values between $x_1$ and $x_2$. Probability densities only become probabilities when integrated.

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

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

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

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

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

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

Probabilities distributions 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

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

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 mean of the square and the square of the mean:

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

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.

For example, a Gaussian has two parameters, $x_0$ and $\sigma_0$. 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 write a Gaussian in this way ($e^{-x^2/2\sigma}$ rather than, say, $e^{-\lambda x^2}$).

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. That's not true for all distributions. For example, with a constant probability distribution, there is a 58% chance of finding $x$ within $1\sigma$ 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 by 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

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

This is the definition of the mathematical operation of **convolution** between two functions. 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

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

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 $\sigma$ 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 on one side than the other. We might find some crazy function $P(x) = \frac{0.74}{L}\ln(1+e^{2x/L})$ for these probabilities. Then, by numerical integration we find

$$\langle x\rangle = 0.59\, L, \quad \langle x^2\rangle = 0.42\, L^2, \quad \sigma = 0.28\, L \tag{21}$$

Also, $\int_{\langle x\rangle-\sigma}^{\langle x\rangle+\sigma} P(x)dx = 0.6$ so 60% within $\langle x\rangle \pm \sigma$. This is just a contrived example. You should be able to compute $\langle x\rangle$ and $\sigma$ with any function $P(x)$, at least numerically, and you will generally find that not exactly 68% are within $\langle x\rangle \pm \sigma$, but often you get something close.

## 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 l**aw of large numbers** states that

| The average of the results from a set of independent trials varies less and less the more trials are performed |
|---|

More mathematically, we can state it this way

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

Thus as $N\to\infty$, the standard deviation of the average $\frac{\sigma}{\sqrt{N}} \to 0$.

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 $\sigma_N$ 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

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

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 $\sigma$, but the relation between $\sigma_N$ with $N$ molecules and $\sigma$ 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!

More precisely the **central limit theorem** states that

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

There are a lot of ways to prove it. I find the "moment" approach the most accessible, as discussed next. Another proof using convolutions an Fourier transforms in is Appendix C.

## 3.1 CLT proof using moments

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 intuitively, 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 $\sigma$ 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 $\sigma$ so that the width is always 1. Then the curve will not get any smaller as $N \to 0$ because its 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 $\sigma$, are dimensionless numbers given by the formula

$$M_n = \begin{cases} 0 & , n \text{ odd} \\ 2^{-\frac{n}{2}} \dfrac{n!}{(\frac{n}{2})!} & , n \text{ even} \end{cases} \tag{38}$$

These $M_n$ completely determine 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

$$\langle x^3 \rangle_N = \int_{-\frac{L}{2}}^{\frac{L}{2}} dx_1 \cdots dx_N P(x_1) \cdots P(x_N) \Big( \frac{x_1 + \cdots + x_N}{N} \Big)^3 \tag{39}$$

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

$$\langle x^3 \rangle_N = \frac{1}{N^2} \langle x^3 \rangle \tag{40}$$

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

$$S_N = \frac{\langle (x-\bar{x})^3 \rangle_N}{\sigma_N^3} = \frac{\langle (x-\bar{x})^3 \rangle / N^2}{(\sigma/\sqrt{N})^3} = \frac{S_1}{\sqrt{N}} \tag{41}$$

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

$$\langle x^4\rangle_N = \int_{-\frac{L}{2}}^{\frac{L}{2}} dx_1\cdots dx_N P(x_1)\cdots P(x_N)\Big(\frac{x_1+\cdots+x_N}{N}\Big)^4 \tag{42}$$

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,

$$\langle x^4\rangle_N = \frac{1}{N^3}\langle x^4\rangle + \frac{3(N-1)}{N^3}\langle x^2\rangle\langle x^2\rangle \tag{43}$$

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

$$K_N = \frac{\langle (x-\bar{x})^4\rangle_N}{\sigma_N^4} = \frac{1}{\sigma^4/N^2}\Big[\frac{1}{N^3}\langle x^4\rangle + \frac{3(N-1)}{N^3}\langle x^2\rangle\langle x^2\rangle\Big] = \frac{K_1}{N} + 3\Big(1-\frac{1}{N}\Big) \tag{44}$$

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 $N\to\infty$

$$S_N\to 0,\quad K_N\to 3,\quad (M_5)_N\to 0,\quad (M_6)_N\to 15,\quad (M_7)_N\to 0,\quad (M_8)_N\to 105,\quad \cdots \tag{45}$$

What we are seeing is that at large $N$ all of the higher moments go to those of a Gaussian! If you work out the details, the general formula is

$$(M_r)_N\to\begin{cases} 0 & ,n\,\text{odd}\\ 2^{-\frac{r}{2}}\dfrac{r!}{(\frac{r}{2})!} & ,n\,\text{even}\end{cases} \tag{46}$$

In exact agreement with the moments of a Gaussian. Thus we always get a Gaussian and the central limit is proven. Another proof using convolutions is in Appendix C.

## 3.2 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 when $x$ is close to the boundaries of the box than if $x$ is central. One way to see this is graphically

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 $3N(N-1)$.

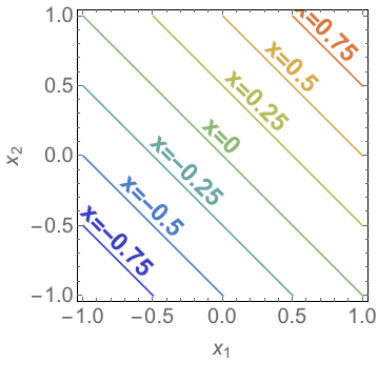



**Figure 1.** The regions in the $x_1/x_2$ plane with mean value $x$ are diagonal lines for $L=2$. The length of the line is the probability $P_2(x)$. For $x=0$, the line is longest and probability greatest. For $x=1$, the line reduces to a point and the probability to zero.

To be quantitative, the easiest way to calculate the probability is with the Dirac $\delta$ function $\delta(x)$ (see Appendix A for a refresher on $\delta(x)$). Using the $\delta$-function, we can write the probability for getting a mean value $x=\frac{x_1+x_2}{2}$ as

$$P_2(x)=\int_{-\frac{L}{2}}^{\frac{L}{2}}dx_1P(x_1)\int_{-\frac{L}{2}}^{\frac{L}{2}}dx_2\,P(x_2)\delta\Big(\frac{x_1+x_2}{2}-x\Big) \tag{47}$$

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

$$\begin{aligned}\int_{-\frac{L}{2}}^{\frac{L}{2}}dx\,P_2(x)&=\int_{-\frac{L}{2}}^{\frac{L}{2}}dx\int_{-\frac{L}{2}}^{\frac{L}{2}}dx_1P(x_1)\int_{-\frac{L}{2}}^{\frac{L}{2}}dx_2\,P(x_2)\delta\Big(\frac{x_1+x_2}{2}-x\Big)\\&=\int_{-\frac{L}{2}}^{\frac{L}{2}}dx_1P(x_1)\int_{-\frac{L}{2}}^{\frac{L}{2}}dx_2\,P(x_2)=1\end{aligned} \tag{48}$$

where we have used the $\delta$-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 $\delta$-function using Eq. (82), giving

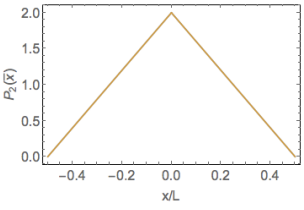
$$P_2(x)=2\int_{-\frac{L}{2}}^{\frac{L}{2}}dx_1P(x_1)\int_{-\frac{L}{2}}^{\frac{L}{2}}dx_2\,P(x_2)\delta(x_1+x_2-2x) \tag{49}$$

Now, the $\delta$-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 $2x-\left(-\frac{L}{2}\right)=\frac{L}{2}+2x$. In other words, we have

$$P_2(x<0)=2\int_{-\frac{L}{2}}^{\frac{L}{2}+2x}dx_1P(x_1)P(2x-x_1) \tag{50}$$

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

$$L^2P_2(x)=\begin{cases}2L+4x, & x<0\\ 2L-4x, & x>0\end{cases}= \tag{51}$$




You can also check this by evaluating Eq. (47) with Mathematica:

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

Plot[P, {x, -1, 1}]

For $N=3$ we compute

$$P_3(x)=\int_{-\frac{L}{2}}^{\frac{L}{2}}dx_1P(x_1)\int_{-\frac{L}{2}}^{\frac{L}{2}}dx_2\,P(x_2)\int_{-\frac{L}{2}}^{\frac{L}{2}}dx_3\,P(x_3)\delta\Big(\frac{x_1+x_2+x_3}{3}-x\Big) \tag{52}$$

and so on. These successive approximations look like

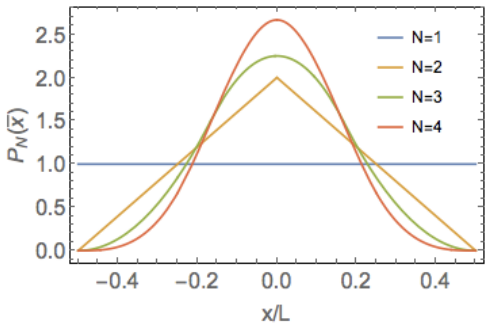



**Figure 2.** The average position of $N=1,2,3,4$ particles, each of which separately has a flat probability distribution.

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,

$$P_N(x)\to\sqrt{\frac{N}{2\pi\sigma^2}}\exp\Big(-N\frac{(x-\bar{x})^2}{2\sigma^2}\Big) \tag{53}$$

Sometimes we sum 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 $\sigma$ convolved with itself $N$ times approaches a Gaussian with mean $N\bar{x}$ and standard deviation $\sqrt{N}\sigma$ as $N\to\infty$.

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

$$P_N^{\text{sum}}(x)=\underbrace{P*P*\cdots*P}_{N}\to\frac{1}{\sqrt{2\pi\sigma^2N}}\exp\Big(-\frac{(x-N\bar{x})^2}{2\sigma^2N}\Big) \tag{54}$$

A proof of the CLT using convolutions is in Appendix B.

We put the "sum" superscript to remind ourselves that we sum the values from each draw from $P(x)$ rather than average their values. The relation is simply

$$P_N^{\text{sum}}(x)=\frac{1}{N}P_N\Big(\frac{x}{N}\Big) \tag{55}$$

The $\frac{1}{N}$ comes from the fact that the probability distributions are differential, so we should technically write $P_N^{\text{sum}}(x)dx=P_N\big(\frac{x}{N}\big)\,d\frac{x}{N}$. Note when we average $\bar{x}\to\bar{x}$ and $\sigma\to\frac{\sigma}{\sqrt{N}}$ and when we sum $\bar{x}\to N\bar{x}$ and $\sigma\to\sqrt{N}\sigma$, so either way

$$\frac{\sigma}{\bar{x}}\to\frac{1}{\sqrt{N}}\frac{\sigma}{\bar{x}} \tag{56}$$

Thus a foolproof way to think of the scaling is that the dimensionless ratio $\frac{\sigma}{\bar{x}}$ should decrease as $\frac{1}{\sqrt{N}}$.

### 3.3 Why we take logarithms in statistical mechanics

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 Gaussian is an unusual function. It is flat near the peak, but then quickly drops off and has a long tail. Since the function is smooth near the peak, it's hard to know what's going on at the tail from expanding near the peak. In particular, you have to work very hard to get information about points with $x \gtrsim \sigma$ from information at the peak. Taking the logarithm puts the peak and the tail on the same footing. Of course, we can't get something for nothing: taking logarithms alone won't solve any problems. But taking logarithms often makes it easier to solve problems. We will see many examples of this as the course progresses.

## 4 Poisson distribution

In many physical situations, there is a large number $N$ of possible events each occurring with very small probability $\lambda$ for a given time interval. 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

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

$\lambda$ 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 (compared to $\frac{1}{\lambda}$), $t=\Delta t$. If the rate to decay is $dP=\lambda dt$ then the probability of not decaying in time $t=\Delta t$ is

$$P_{\text{no decay}}(\Delta t)=1-\lambda\Delta t \tag{61}$$

For the system to survive to a time $2\Delta t$ with no decays, it would have to not decay in $\Delta t$ and then not decay again in the next $\Delta t$. 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

$$P_{\text{no decay}}(2\Delta t)=(1-\lambda\Delta t)^2 \tag{62}$$

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

$$P_{\text{no decay}}(t)=\lim_{N\to\infty}\Big(1-\lambda\frac{t}{N}\Big)^N=e^{-\lambda t} \tag{63}$$

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,

$$P_{1\text{ decay}}(t)=\lim_{N\to\infty} N\underbrace{\Big(1-\lambda\frac{t}{N}\Big)^{N-1}}_{N-1\text{ no decays}}\underbrace{\Big(\lambda\frac{t}{N}\Big)}_{\text{one decay}}=\lim_{N\to\infty}-t\partial_t\Big(1-\lambda\frac{t}{N}\Big)^N \tag{65}$$

In the third term, we have simply rewritten the expression in a smart way with a derivative so we can reduce it to a previously solved problem – a powerful physicist trick. Now we switch the order of the limit and the $\partial_t$ and use Eq. (63) to get

$$P_{1\text{ decay}}(t)=-t\partial_t P_{\text{nodecay}}(t)=\lambda t e^{-\lambda t} \tag{66}$$

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

$$P_{2\text{decays}}(t)=\lim_{N\to\infty}\underbrace{\frac{N(N-1)}{2}}_{\text{pick 2 to decay}}\underbrace{\Big(1-\lambda\frac{t}{N}\Big)^{N-2}}_{N-2\quad\text{no decays}}\underbrace{\Big(\lambda\frac{t}{N}\Big)^2}_{\text{two decays}}=\frac{1}{2}t^2\,\partial_t^2 P_{\text{no decay}}(t)=\frac{(\lambda t)^2}{2}e^{-\lambda t} \tag{67}$$

For general $m$ the result is

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

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 $\lambda$ and the events are uncorrelated.

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

$$\sum_m P_m(t)=\sum\frac{(\lambda t)^m}{m!}e^{-\lambda t}=1 \tag{69}$$

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 $t$. Moreover, while it is a normalized probability in $m$, it is a simply a function (not a probability distribution) of $t$. There is not a sense in which $\int dt\,P_m(t)=1$; this doesn't even have the right units.

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 $m$

$$\langle m \rangle = \sum_m m P_m(z) = \sum_m m \frac{(\lambda t)^m}{m!} e^{-\lambda t} = \lambda t \tag{70}$$

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 $\lambda t$. 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

$$\sigma = \sqrt{\langle m^2 \rangle - \langle m \rangle^2} = \sqrt{\lambda t} \tag{71}$$

Again, you can check this yourself as an exercise.

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

$$\frac{\sigma}{\langle m \rangle} = \frac{1}{\sqrt{\lambda t}} \tag{72}$$

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 $1 \pm 1$ particle to decay. If we wait 2 lifetimes, we expect $2 \pm \sqrt{2}$ to decay ($t = \frac{2}{\lambda}$, $\langle m \rangle = 2$ and $\sigma = \frac{2}{\sqrt{2}} = \sqrt{2}$). If we wait 100 lifetimes, we expect $100 \pm 10$ 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:

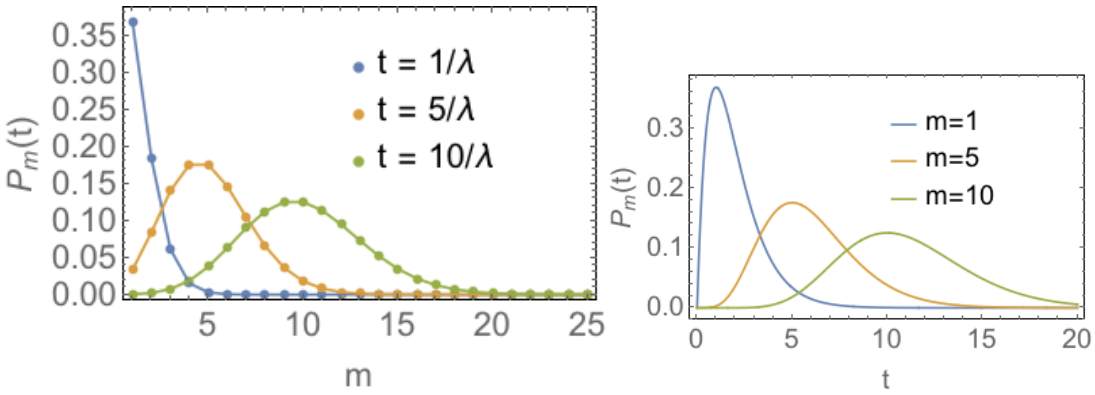



**Figure 3.** The Poisson distribution as a function of the discrete index $m$ for various times (left) and time, for various values of $m$ (right)

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 $m!$ and want to expand at large $m$, you should immediately think **Stirling's approximation**:

$$x! \approx e^{-x} x^x \times (\cdots) \tag{73}$$

or equivalently

$$\ln x! \approx x \ln x - x + \cdots \tag{74}$$

For a simple derivation, see Appendix B. We will use this expansion a lot.

The log of the Poisson distribution is

$$\ln P_m(t) = \ln\left[\frac{(\lambda t)^m}{m!}e^{-\lambda t}\right] = m\ln(\lambda t) - \lambda t - \ln m! \tag{75}$$

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

$$\ln P_m(t) \xrightarrow[m\gg 1]{} m\ln(\lambda t) - \lambda t - m\ln m + m + \cdots = m\ln\frac{\lambda t}{m} + (m-\lambda t) + \cdots \tag{76}$$

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. (76) vanish. The next term is

$$\frac{\partial}{\partial m}\ln P_m(t)\bigg|_{m=\lambda t} = \lim_{m\to\lambda t}[\ln(\lambda t) - \ln m] = 0 \tag{77}$$

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

$$\frac{\partial^2}{\partial m^2}\ln P_m(t)\bigg|_{m=\lambda t} = \lim_{m\to\lambda t}\left[-\frac{1}{m}\right] = -\frac{1}{\lambda t} \tag{78}$$

Thus,

$$\ln P_m(t) = -\frac{1}{2\lambda t}(m-\lambda t)^2 + \cdots \tag{79}$$

and therefore

$$P_m(t) \xrightarrow[m\gg 1]{} \frac{1}{\sqrt{2\pi\lambda t}}e^{-\frac{(m-\lambda t)^2}{2\lambda t}} \tag{80}$$

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 Summary

In this lecture, we introduced the basic concepts from probability that will be useful for statistical mechanics. The key concepts are

- Normalized **probability distributions** $P(x)$ with $\int dx\, P(x) = 1$
- **Mean**: $\bar{x} = \langle x\rangle = \int dx\, x P(x)$
- **Variance** $\text{var} = \int dx (x-\bar{x})^2 P(x)$,
- **Standard deviation** or **width** $\sigma = \sqrt{\text{var}}$
- **Gaussian** distribution $P(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\bar{x})^2}{2\sigma^2}\right)$ has mean $\bar{x}$ and width $\sigma$.
- If you draw $x$ from Gaussian it is 68% likely to between $\bar{x}-\sigma$ and $\bar{x}+\sigma$.
- The **convolution** of two distributions is defined as $(P_A * P_B)(z) = \int_{-\infty}^{\infty} dx\, P_A(z-x)P_B(x)$. It describes the probabilty of getting $z$ as the sum of a number draw from $P_A$ and another number drawn from $P_B$.
- Given a probabilty distribition $P(x)$ with mean $\bar{x}$ and width $\sigma$, you can construct a new probabilty distribution $P_N(x)$ by averaging over $N$ draws from $P(x)$. The **central limit thoerem (CLT)** says that as $N\to\infty$ this new distribution will approach a Gaussian with the same mean as $P(x)$ ($\bar{x}_N = \bar{x}$) and a smaller standard deviation $\sigma_N \approx \frac{\sigma}{\sqrt{N}}$. All other properties of $P(x)$ are lost after this averaging at large $N$.
- The CLT also implies that if we *sum* (rather than average) the values from draws, the mean grows like $\bar{x}_N \approx N\bar{x}$ and the standard deviation like $\sigma_N \approx \sqrt{N}\sigma$. If we

- Because of the CLT, Gaussians are very common. Their exponential decay encourages us to study logarithms of distributions, which turns fast-varying exponentials into slow-varying polynomials: $\ln e^{-\frac{x^2}{2\sigma^2}} = -\frac{x^2}{2\sigma^2}$.
- When we have a rate $dP = \lambda dt$ for an event happening that is independent of time, then the probability of having $m$ events after a time $t$ is described by the **Poisson distribution** $P_m(t) = \frac{(\lambda t)^m}{m!} e^{-\lambda t}$.
- **Stirling's approximation** is that $N! \approx \sqrt{2\pi N} N^N e^{-N}$ at large $N$. This works very well, even at $N = 1$.

# Appendix A Dirac $\delta$-function

The Dirac $\delta$-function is very useful in physics, from quantum mechanics to statistical mechanics. The $\delta$-function is not really a function but rather a distribution. $\delta(x)$ is zero everywhere except at $x = 0$. When you integrate a function against $\delta(x)$ you pick up the value of that function at 0. That is

$$\int dx \delta(x) f(x) = f(0) \tag{81}$$

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

$$\delta(ax) = \frac{1}{a}\delta(x) \tag{82}$$

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

$$\int dx \delta(ax) f(x) = \int d\frac{x}{a} \delta(x) f\Big(\frac{x}{a}\Big) = \frac{1}{a} f(0) = \int dx \Big[\frac{1}{a}\delta(x)\Big] f(x) \tag{83}$$

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

$$\delta(x) = \lim_{\sigma \to 0} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \tag{84}$$

As a check, note that the integral over the Gaussian is 1 regardless of $\sigma$, so the $\delta$ 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 $\delta$ function.

# Appendix B Stirling's approximation

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

$$\ln N! = \ln N + \ln(N-1) + \ln(N-2) + \cdots + \ln 1 = \sum_{j=1}^{N} \ln j \tag{85}$$

For large $N$ we then write the sum as an integral

$$\ln N! = \sum_{j=1}^{N} \ln j \approx \int_1^N dj \ln j = N \ln N - N - 1 \approx N \ln N - N \tag{86}$$

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

$$N! \approx \sqrt{2\pi N} N^N e^{-N} \tag{87}$$

An alternative derivation is to given an integral representation of the factorial as a $\Gamma$ function: $n! = \Gamma(n+1) = \int_0^\infty x^n e^{-x} dx$. For example, Mathematica can simply series expand this around $n = \infty$ to reproduce Eq. (87). Try it!

The next order correction to this is down by $\frac{1}{12N}$, which gets small fast. You can check that Stirling's approximation is off by less than 8% already at $N = 1$ and by less than 2% by $N = 3$. For Avogadro's number $N = 6 \times 10^{23}$ it is off by one part in $10^{25}$.

## Appendix C Central limit theorem from convolutions

Here's another slick proof of the central limit theorem. We start with the definition

$$P_N(x) = \int dx_1 ... dx_n P(x_1) ... P(x_n) \delta\Big(\frac{x_1 + \cdots + x_n}{N} - x\Big) \tag{88}$$

Now we write the $\delta$ function in Fourier space as

$$\delta(x_1 + \cdots + x_n - x) = \int \frac{dk}{2\pi} e^{ik(x_1 + \cdots + x_n - x)} \tag{89}$$

So that

$$P_N(x) = \int dk \int dx_1 ... dx_n P(x_1) ... P(x_n) e^{ik\left(\frac{x_1 + \cdots + x_n}{N} - x\right)} \tag{90}$$

Defining the Fourier transform of $P$ as

$$\tilde{P}(k) = \int dx e^{ikx} P(x) \tag{91}$$

we then have

$$P_N(x) = \int \frac{dk}{2\pi} \Big[\tilde{P}\Big(\frac{k}{N}\Big)\Big]^N e^{ikx} \tag{92}$$

Eq. (92) is just the statement that Fourier transforms turns convolutions into products.Then,

$$\tilde{P}\Big(\frac{k}{N}\Big) = \int dx e^{i\frac{kx}{N}} P(x) = \int dx \Big(1 + \frac{ikx}{N} - \frac{1}{2}\Big(\frac{kx}{N}\Big)^2 + \frac{1}{3!}\Big(\frac{ikx}{N}\Big)^3 + \cdots\Big) P(x) \tag{93}$$

$$= 1 + i\frac{k}{N}\langle x\rangle - \frac{k^2}{2}\frac{\langle x^2\rangle}{N^2} - i\frac{k^3\langle x^3\rangle}{6N^3} + \cdots \tag{94}$$

Now if we didn't do anything else, then as $N \to \infty$ we see immediately that $\tilde{P}\Big(\frac{k}{N}\Big) \to 1$ and so $P_N(x) \to \delta(x)$. This is because the whole distribution is shrinking down to be around $x = 0$. That result is not wrong, but it's happening because the width of the Guassian is going to zero. We want to work to first subleading order in $\frac{1}{N}$, i.e. keep the factor of $N$ in the width.

To proceed, let's first set $\langle x\rangle = 0$ by shifting the offset of $x$. Then $\langle x^2\rangle = \sigma^2$ and we have

$$\Big[\tilde{P}\Big(\frac{k}{N}\Big)\Big]^N = \Big[1 - \frac{k^2}{2}\frac{\sigma^2}{N^2} - i\frac{k^3\langle x^3\rangle}{6N^3} + \cdots\Big]^N \tag{95}$$

Now, the biggest terms in this product will come from taking as many factors of 1 as we can. If you take one of the other terms you pay at least $\frac{1}{N^2}$. So taking either all $1's$ or only one non-1 term, we get

$$\Big[\tilde{P}\Big(\frac{k}{N}\Big)\Big]^N = 1 - N\Big(\frac{k^2}{2}\frac{\sigma^2}{N^2} - i\frac{k^3\langle x^3\rangle}{6N^3} + \cdots\Big) + \cdots \tag{96}$$

From here we see that the first term subleading in $\frac{1}{N}$ is the $\sigma^2$ term and the rest is order $\frac{1}{N^2}$ or lower. Thus to get the answer write to order $\frac{1}{N}$ we write

$$\left[\tilde{P}\left(\frac{k}{N}\right)\right]^N = \left[1-\frac{1}{N}\left(\frac{k^2}{2}\frac{\sigma^2}{N}\right)\right]^N + \mathcal{O}\left(\frac{1}{N^2}\right) \xrightarrow[N\gg 1]{} e^{-\frac{k^2}{2}\left(\frac{\sigma^2}{N}\right)} + \mathcal{O}\left(\frac{1}{N^2}\right) \tag{97}$$

where $e^x = \lim_{N\to\infty}\left(1+\frac{x}{N}\right)^N$ was used. We can then compute the inverse Fourier transform to get

$$P_N(x) = \int_{-\infty}^{\infty}\frac{dk}{2\pi}e^{-\frac{k^2}{2}\left(\frac{\sigma^2}{N}\right)}e^{ikx} = \sqrt{\frac{2N}{\sigma^2}}\int_{-\infty}^{\infty}\frac{dk}{2\pi}e^{-\frac{k^2}{2}}e^{ik\left(\frac{\sqrt{2N}x}{\sigma}\right)} = \sqrt{\frac{N}{2\pi\sigma^2}}e^{-\frac{Nx^2}{2\sigma^2}} \tag{98}$$

which is the desired result, the central limit theorem.