COMMON SENSE RULES IN PROBABILITY THEORY

The probability of event A is

$$P_{\mathcal{A}} = \frac{\text{How many times A happened}}{\text{Number of events}} = \frac{N_{\mathcal{A}}}{N} \ .$$

In fact, to avoid "luck"/"misfortune" fluctuations, one has to take the limit of $N_{\mathcal{A}}, N \to \infty$:

$$P_{\mathcal{A}} = \lim_{N\to\infty} \frac{N_{\mathcal{A}}}{N} \ .$$

In practice $N = \infty$ is not possible, but we will see later that $N \sim 10^{23}$ is often the case, and **FAPP** (For All Practical Purposes) this is a sufficiently large number.

Probability that something will happen is 1, i.e.,

$$\sum_{\mathcal{A}} P_{\mathcal{A}} \ = 1 \ . \quad - \textbf{Normalization condition} \ .$$

Just like in computing, parallel measurements are faster; instead of repeating the same measurement many times we may try to do the same measurement on many identical systems simultaneously and then count outcomes. These systems are called an *ensemble*.

We may also invert the definition of the probability to say that if $P_{\mathcal{A}}$ is the probability of A to happen, then in N events A will happen approximately $N_{\mathcal{A}} \ \approx P_{\mathcal{A}} \ N$ times, or

$$< N_{\mathcal{A}} \ >= P_{\mathcal{A}} \ N \ , \quad -\textbf{The expectation value}$$

of A to happen in N events.

If A and B are completely independent events (mathematically we may say so, but in physics one may always complain (rightly) that everything depends on everything; however in most cases this dependence can be neglected). Given $P_{\mathcal{A}}$ and $P_{\mathcal{B}}$, the probability of the event to be either A **or** B is the **sum** :

$$P_{\mathcal{A} \ or \mathcal{B}} \ = P_{\mathcal{A}} \ + P_{\mathcal{B}}$$

since $N_{\mathcal{A} \ or \mathcal{B}} \ = N_{\mathcal{A}} \ + N_{\mathcal{B}}$. Also, the probability of having A in the first measurement **and** B in the second independent measurement is

$$P_{\mathcal{A} \ \& \mathcal{B}} \ = P_{\mathcal{A}} \ * P_{\mathcal{B}} \ \ ,$$

i.e., probabilities **multiply** ("guessing" all the time is harder)

These are simple rules, but starting from "bricks", a complicated building may be constructed. Other rules are also simple (although we will not need them), like given $P_{\mathcal{A}}$ and $P_{\mathcal{B}}$ and **joint probabilities** for *dependent* events

$$P_{\mathcal{A}}\ (\mathcal{B}\ )\ -\ \text{The prob. of A to happen if B happened ;}$$

$$P_{\mathcal{B}}\ (\mathcal{A}\ )\ -\ \text{The prob. of B to happen if A happened ;}$$

the probability of A to happen if B did not happen is just

$$P_{\mathcal{A}}\ (Not\ \mathcal{B}\ ) = P_{\mathcal{A}}\ - P_{\mathcal{A}}\ (\mathcal{B}\ )\ ,$$

(it reads: subtract from $N_{\mathcal{A}}$ those cases when A happened after B).

Let us apply these rules to calculate what is the probability that A happens *exactly* $N_{\mathcal{A}}$ times in N events. We need this to understand how good is our expectation that $N_{\mathcal{A}}\ =< N_{\mathcal{A}}\ >$ (as an example imagine tossing coins and asking for the probability of having $N_{\mathcal{A}}$ heads in N attempts). If we specify some sequence of outcomes, like

$$Not\mathcal{A}\ ,\ \mathcal{A}\ ,\ \mathcal{A}\ ,\ Not\mathcal{A}\ ,\ Not\mathcal{A}\ ,\ Not\mathcal{A}\ \ \mathcal{A}\ \ldots$$

then assuming that all events happen independently from each other, the probability of the sequence will be

$$P^{sequence}(N_{\mathcal{A}}\ ) = P_{\mathcal{A}}\ ^{N_{\mathcal{A}}}\ (1 - P_{\mathcal{A}}\ )^{N-N_{\mathcal{A}}}\ \ -\ \ \text{Product}$$

But there are many sequences giving $N_{\mathcal{A}}$ - one has to count the number of ways of placing $N_{\mathcal{A}}$ characters A in N boxes

$$\underbrace{|\ \ |A|\ \ |A|A|\ \ |\ \ |\ \ |A|}_{Nevents}$$

To place the first one we have N places, then only $N-1$ left for the second one, etc. , i.e., $N!/(N-N_{\mathcal{A}}\ )!$ combinations. But $N_{\mathcal{A}}\ !$ of them are identical since all A are the same, thus we have $({N \atop N_{\mathcal{A}}})$ different combinations, and (to simplify notations lets use $p = P_{\mathcal{A}}$ and $k = N_{\mathcal{A}}$)

$$P(k) = p^k(1-p)^{N-k}({N \atop k}) \equiv p^k(1-p)^{N-k}\frac{N!}{(N-k)!k!}\ .$$

This is **binomial distribution**.

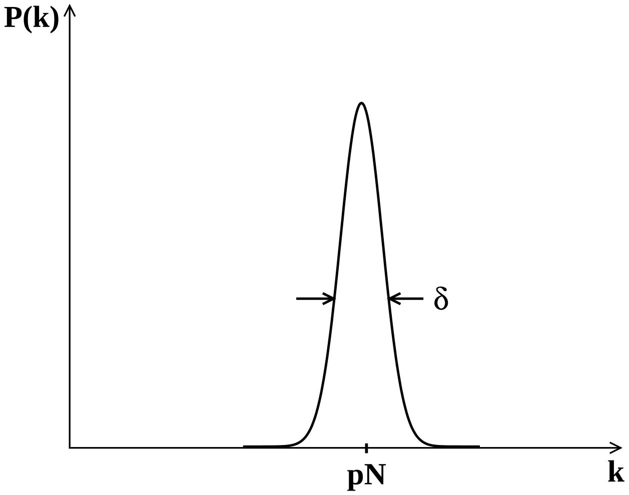
We are ready to see the difference between our expectations $< k >= pN$ and the actual statistics of possible outcomes. For large N and k one may use the Stirling's expression for the factorial function $N! \approx \sqrt{2\pi N} N^N e^{-N}$, to get

$$({}^N_k) \approx \frac{1}{\sqrt{2\pi N}} \left(\frac{N}{k}\right)^{k+1/2} \left(\frac{N}{N-k}\right)^{N-k+1/2} .$$

Substituting this into $P(k)$, we find

$$P(k) \approx \frac{1}{\sqrt{2\pi N p(1-p)}} \exp\left\{(k+1/2)\ln[pN/k]-(N-k+1/2)\ln[(N-k)/(N-pN)]\right\}$$

this function is strongly peaked when N and k are large and the maximum is at the expectation value pN.



We can actually find the form of $P(k)$ by assuming that $|k - pN| \ll pN$. We simply substitute $k = pN + \delta$ into the exponent and keep only terms of order δ (there will be none) and δ^2/N, neglecting δ/N and $(\delta/N)^2$ (we will see shortly that typical fluctuations are of order $N^{1/2}$, and thus $\delta^2/N \sim 1$, but $\delta/N \sim N^{-1/2} \ll 1$. This calculation takes time but is straightforward:

$$P(k) \approx \frac{1}{\sqrt{2\pi}\sigma} e^{-(k-pN)^2/2\sigma^2} ,$$

where $\sigma = \sqrt{Np(1-p)}$. This is **Gaussian distribution**. The parameter σ which measures the typical spread, or variance, of the k values is called the mean-square-root-deviation (m.s.r.d.). Since $\sigma \sim \sqrt{k(1-p)}$ the peak is very narrow for large k.

Narrow? Well, relatively narrow, because the absolute value of σ increases with k as $k^{1/2}$, but the ratio σ/k tends to zero as $k^{-1/2}$.

> All this applies to large N and k only. In "finance",
> out of 1000 brokers one will almost certainly (by pure luck,
> not knowledge) guess correctly stocks variations up/down 10
> times in a row. But the whole population of Earth is not
> enough to find a lucky one who will guess correctly
> 100 times in a row.

So far we have been discussing discrete variables, and defined $P_{\mathcal{A}} = \lim_{N\to\infty} N_{\mathcal{A}} \,/N$. If events A form a continuum of possibilities, then we have to use probability densities:

$$\text{Probability of A } = d\mathcal{A}\ \rho(\mathcal{A}\ ) = \lim_{dN\to\infty} N_{\mathcal{A}}\ /N\ ,$$

and $dN_{\mathcal{A}}$ will be proportional to $d\mathcal{A}$. One may also use an identical definition of the probability density as:

$$\text{Prob.of } \mathcal{A}_{\ 1} < \mathcal{A}\ < \mathcal{A}_{\ 2} = \int_{\mathcal{A}_{\ 1}}^{\mathcal{A}_{\ 2}} d\mathcal{A}\ \rho(\mathcal{A}\ ) = \lim_{N\to\infty} \frac{\int_{\mathcal{A}_{\ 1}}^{\mathcal{A}_{\ 2}} dN_{\mathcal{A}}}{N}\ .$$

The continuous version of the Gaussian probability density is very similar

$$\rho(x) = \frac{1}{\sqrt{2\pi}\sigma}\ e^{-(x-<x>)^2/2\sigma^2}\ ,$$

and we will see shortly how it derives for the large ensemble of independent variables. $<x>$ is the expectation value, or **average**

$$<x>= \int dx x\rho(x)$$

and σ=r.m.s.d.

$$\sigma =< (x- <x>)^2 >= \int dx(x- <x>)^2\rho(x) =< x^2 > - < x >^2\ .$$

In statistical analysis of errors (e.g., in Monte Carlo
simulations) when the final result is not deterministic,
but rather probabalistic with Gaussian density distribution,
one is usually giving error bars in terms of σ to
specify how confident he is about the answer

$$\text{Error bar} = \ \sigma \qquad \Longrightarrow \left\{ \begin{array}{l} \text{The answer is within the} \\ \text{limits specified with probability} \\ \int_{<x>-\sigma}^{<x>+\sigma} dx\rho(x) = 0.683 \end{array} \right.$$

$$\text{Error bar} = 3\sigma \qquad \Longrightarrow \left\{ \begin{array}{l} \text{The answer is within the} \\ \text{limits specified with probability} \\ \int_{<x>-3\sigma}^{<x>+3\sigma} dx\rho(x) = 0.998 \end{array} \right.$$

$< x >$ and σ are the first *moments* of the distribution. In general the n-th moment is defined as

$$< x^n >= \int dx x^n \rho(x)$$

For Gaussian distribution

$$< (x- < x >)^{2n} >= \frac{2^n \sigma^{2n}}{\sqrt{\pi}} \Gamma(n + \frac{1}{2}) \ ; \qquad [\Gamma(x) = \text{Gamma function}]$$

Higher moments are sometimes used to distinguish between the distributions; e.g., the third moment in $< (x- < x >)^3 > \equiv < x^3 > -3 < x^2 >< x > +2 < x >^3$ tells us how asymmetric is the distribution function.

Another important distribution arises when we study *random processes*. In radioactive decay we have the probability of decay in time dt to be

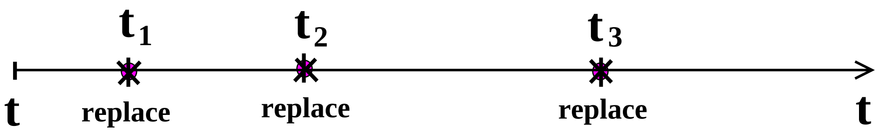
$$dt/\tau \qquad \text{OR} \qquad \gamma dt$$

where τ is the decay time, and $\gamma = 1/\tau$ the decay rate. Then

$$(1 - dt/\tau) \qquad \text{is the prob. not to decay in time } dt$$

$$(1 - dt/\tau)^{t/dt} \equiv e^{-\gamma t} \ \text{ the prob. not to decay in time } t$$

Suppose now that we "repair" the decaying system (e.g., if it is a bulb burning out)



What is the probability of having done m repairs in time t? We have to sum/integrate over all possibilities for the m events to happen:

$$P(m) = \int_0^t \gamma dt_m \int_0^{t_m} \gamma dt_{m-1} \dots \int_0^{t_2} \gamma dt_1 \; e^{-\gamma t_1} e^{-\gamma(t_2-t_1)} \dots e^{-\gamma(t-t_m)}$$

This expression is nothing but the sum/integral of product probabilities for the particular sequence of events. All exponents collect into $e^{-\gamma t}$ independent of the set of $\{t_i\}$, and the integral then is just $(\gamma t)^m/m!$. Thus we find

$$P(m) = \frac{(\gamma t)^m}{m!} e^{-\gamma t} \quad \textbf{The Poisson distribution}$$

Also, if we have an ensemble of $N \gg 1$ decaying systems and select small Δt such that $(\gamma \Delta t)N = a \sim 1$, where a is the average number of events/decays happening in N systems in time Δt, then the probability of m decays to happen in time Δt is

$$\left({}^{N}_{m} \right)(\gamma \Delta t)^m (1-\gamma \Delta t)^{N-m} \approx \frac{N^m (\gamma \Delta t)^m e^{-(\gamma \Delta t)N}}{m!} = \frac{a^m e^{-a}}{m!}$$

also a Poisson distribution. We have seen already that if a is large, then the Poisson distribution goes to Gaussian.

Characteristic functions and sums of random variables

There is a very useful trick to deal with probability densities of independent variables. Define:

$$\text{FOURIER TRANSFORM}: \quad \phi(k) = \int dx e^{ikx} \rho(x) \equiv \langle e^{ikx} \rangle$$

$$\text{INVERSE TRANSFORM}: \quad \rho(x) = \frac{1}{2\pi} \int dk e^{-ikx} \phi(k)$$

Moments are very easy with $\phi(k)$ - they are simply related to the Taylor expansion of $\phi(k)$

$$\phi(k) = \sum_{n=0}^{\infty} \frac{k^n R_n}{n!}$$

but also

$$\langle e^{ikx} \rangle = \sum_{n=0}^{\infty} \frac{(ik)^n \langle x^n \rangle}{n!}$$

i.e., $R_n = i^n \langle x^n \rangle$. Finally

$$\langle x^n \rangle = \frac{\phi^{(n)}(k)}{i^n} \qquad \text{at } k = 0$$

For example,

$$\langle x \rangle = -id\phi/dk|_{k=0} \ ; \quad \langle x^2 \rangle = -d^2\phi/dk^2|_{k=0} \ ; \ \text{ etc.}$$

The second expression may be also used to compute the dispersion as

$$\langle x^2 \rangle - \langle x \rangle^2 = -d^2\phi/dk^2 + (d\phi/dk)^2 = -\frac{d}{dk}\left[\frac{d\phi/dk}{\phi}\right]_{k=0}$$

since $\phi(0) = 1$ (normalization condition for $\rho(x)$). Thus

$$\sigma = -\frac{d^2}{dk^2}\left[\ln \phi\right]_{k=0}$$

Taylor expansion coefficients of $\ln \phi(k)$ are known as commulants:

$$\ln \phi(k) = \sum_{n=0}^{\infty} \frac{(ik)^n C_n}{n!}$$

$$C_3 = \langle (x - \langle x \rangle)^3 \rangle \quad - \text{ The distribution asymmetry}$$

The real advantage of $\phi(k)$ is found when one has to deal with independent events. Let's consider two independent variables described by $\rho_x(x)$ and $\rho_y(y)$. No assumptions about their form is made here except that they have Fourier transforms. What is the probability density of the sum $z = x + y$? As before we count all possibilities

$$\rho_z(z) = \underbrace{\int dx \rho(x)}_{\text{Let } x} \underbrace{\int dy \rho(y)}_{\text{Let } y} \qquad \underbrace{\delta(x+y-z)}_{\text{Count only when } x+y=z}$$

or

$$\rho_z(z) = \int dx \rho(x) \rho_y(z-x) \quad - \quad \text{Convolution}$$

Fourier transform:

$$\begin{aligned}
\phi(k) &= \int dz e^{ikz} \int dx \rho_x(x) \rho_y(z-x) \\
&= \underbrace{\int d(z-x) e^{ik(z-x)} \rho_y(z-x)}_{\text{Shift}} \int dx e^{ikx} \rho_x(x) \\
&= \phi_x(k) \phi_y(k) \quad - \quad \textbf{Fourier transforms multiply}
\end{aligned}$$

More generally, if $x_1, x_2, \ldots, x_N$ are independent, then

$$\phi_{z=\sum x_i} = \prod_{i=1}^{N} \phi_i(k)$$

and

$$\ln \phi_z = \sum_{i=1}^{N} \ln \phi_i \quad - \quad \textbf{Logarithms add}$$

Since logarithms of the Fourier transforms add, their Taylor expansions also add:

$$C_n^{(z)} = \sum_{i=1}^{N} C_n^{(i)} \quad - \quad \textbf{Commulants add}$$

Our previous result can be easily reproduced now - if all variables are identical (i.e., they are just outcomes of an ensemble measurement) then $\langle z \rangle = N \langle x \rangle$ and $\sigma_z = \sqrt{N} \sigma_x$. This rule applies for any distribution ρ_x provided N is large. It is not just that average values and mean square deviations add, but also the functional form of $\rho(z)$ becomes universal! We can prove that it is Gaussian for large N - this result is called

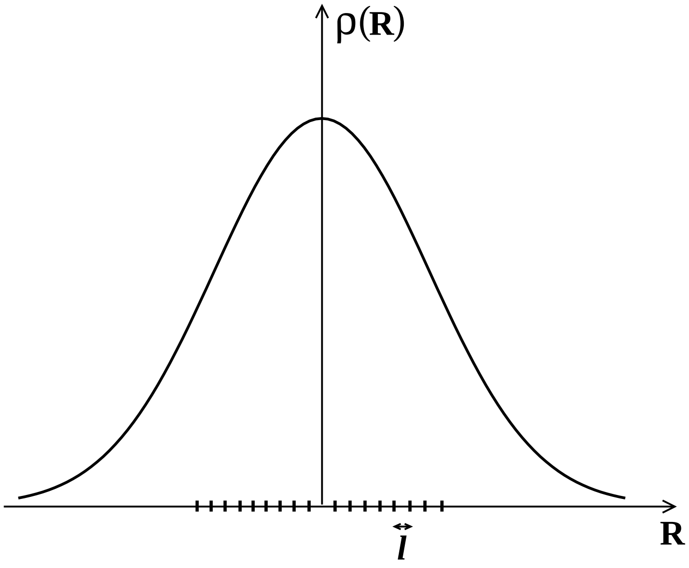
The Central Limit Theorem, or CLT.
Proof:

$$\begin{aligned}
\text{Let's} \quad z &= \left\{\sum_{j=1}^{N}[x_j - \langle x_j\rangle]\right\}/\sqrt{N} \\
\text{Then:} \quad \phi_z(k) &= \langle e^{ik[\sum_{j=1}^{N}(x_j - \langle x_j\rangle)]/\sqrt{N}}\rangle \\
&= e^{-ik\sum_{j=1}^{N}\langle x_j\rangle/\sqrt{N}}\langle e^{ik\sum_{j=1}^{N}x_j/\sqrt{N}}\rangle \\
&= e^{-ik\sum_{j=1}^{N}\langle x_j\rangle/\sqrt{N}}\prod_{j=1}^{N}\phi_j(k/\sqrt{N}) \\
\text{Take ln:} \quad \ln[\phi_z(k)] &= -ik\sum_{j=1}^{N}\langle x_j\rangle/\sqrt{N} + \sum_{j=1}^{N}\ln\phi_j(k/\sqrt{N}) \\
\text{Expand in k:} \quad &= -ik\sum_{j=1}^{N}\langle x_j\rangle/\sqrt{N} + \sum_{j=1}^{N}(k/\sqrt{N})\phi_j'(0) - \frac{k^2}{2N}\sum_{j=1}^{N}\sigma_j^2 \\
&\quad +(\text{terms} \sim \tfrac{1}{\sqrt{N}} \to 0) \\
\text{Thus:} \quad \phi_z(k) &= e^{-k^2\sigma^2/2} \\
\text{Inverse transform } \rho_z(z) &= \frac{1}{\sqrt{2\pi\sigma}}\, e^{-z^2/2\sigma^2} \quad - \quad \text{Gaussian}
\end{aligned}$$

Random Walks Let's apply CLT to the problem of random walks (or lattice diffusion if you like). Random walk means that each step we make is decided with equal probability (1/2) to be to the left or to the right. The displacement is then $\pm l$, i.e., each time $\rho(\Delta x) = 1/2(\delta(\Delta x - l) + \delta(\Delta x + l))$.

Question: How far one may go in N steps, and what is the distribution of distances $R = \sum_{j=1}^{N}\Delta x_j$? CLT says :

$$\begin{cases}
\langle R\rangle = N\langle x\rangle = 0 \\
\langle R^2\rangle = N\langle x^2\rangle = Nl^2 \\
\text{and } \rho(R) \text{ is Gaussian: } \rho(R) \longrightarrow \exp\{-R^2/2\langle R^2\rangle\}/(\sqrt{2\pi\langle R^2\rangle})
\end{cases}$$



Discretization is not too important because $\sigma_R \gg l$

I.e., we are still around the origin (in one dimension, in fact, we keep coming back!) and diffusively spread out as a square-root in the number of steps, or time, if 1 step is a unit of time, $R \sim \sqrt{t}$. Apart from lattice diffusion, where particles hop between lattice sites, random walks also apply to the Brownian motion of molecules in solutions.