

Chapter 1

Elementary Concepts in Statistics and Probability

1.0. Random Variables and Their Distributions

Like the winning numbers of lottery tickets, physical variables often have a random component, with quantum mechanics adding an extra layer of uncertainty to the results. Some variables, such as the spatial position of an object $\mathbf{r} = (x, y, z)$, are continuous ($x \rightarrow x + dx$), allowing differential calculus to be brought to bear. Others are discrete (e.g., the color of a sock in a drawer that contains a number of brown, grey and black socks, the identity of a card in a playing deck of 52, the spin of an electron, etc.) and require a distinct approach.

In this chapter we turn our attention to the most basic of discrete cases, the binary example of a coin toss (e.g., heads or tails.) Although a trained prestidigitator or a specially devised machine might be able to produce coin tosses that are *always* heads, under normal circumstances experience and common sense tell us that a coin toss results in 50% heads and 50% tails. That is, an infinitesimal initial variability in the toss results in maximum variability of the results. We study this phenomenon in detail with the aid of a remarkably simple tool, the *binomial distribution*. A bias toward one or the other outcome is then introduced. At the end of the chapter we generalize to an arbitrary number of discrete possibilities, using the *multinomial distribution*. Taken together, these intuitive results suggests a rôle for, and definition of, *temperature*, as the control parameter for the generation of random events.

1.1. The Binomial Distribution

We can obtain all the binomial coefficients from a simple *generating function* G_N :

$$G_N(p_1, p_2) \equiv (p_1 + p_2)^N = \sum_{n_1=0}^N \binom{N}{n_1} p_1^{n_1} p_2^{n_2}, \quad (1.1)$$

where the $\binom{N}{n_1}$ symbol^a (equally written $\binom{N}{n_2}$) stands for the ratio of factorials $N!/n_1!n_2!$. Recalling the definition of the factorial of a positive integer $n! = 1 \times 2 \times \cdots \times n$, it is also conventional to define $0! = 1$ by extension. This definition is required in order to satisfy a logical identity, that the number of ways to choose N objects out of a set of N is 1, i.e., $\binom{N}{N} = \binom{N}{0} = 1$. Both here and subsequently, $n_2 = N - n_1$.

If the p 's are positive, each term in the sum is positive. If restricted to $p_1 + p_2 = 1$ they add to $G_N(p_1, 1 - p_1) \equiv 1$. Thus, each term in the expansion on the right-hand side of (1.1) can be viewed as a *probability* of sorts.

Generally there are only three requirements for a function to be a probability: it must be non-negative, sum to 1, and it has to express the relative frequency of some stochastic (i.e. *random*) phenomenon in a meaningful way. The binary distribution which ensues from the generating function above can serve to label a coin toss (let 1 be “heads” and 2 “tails”), or to label spins “up” in a magnetic spin system by 1 and spins “down” by 2, or to identify copper atoms by 1 and gold atoms by 2 in a copper-gold alloy, etc. Indeed all non-quantum mechanical binary processes with a statistical component are similar and can be studied in the same way.

It follows (by inspection of Eq. (1.1)) that we can define the probability of n_1 heads and $n_2 = N - n_1$ tails, in N tries, as

$$W_N(n_1) = \binom{N}{n_1} p_1^{n_1} p_2^{n_2}, \quad (1.2)$$

subject to $p_2 = 1 - p_1$. This chapter concerns in part the manner in which one chooses p_1 and p_2 in physical processes. These are the parameters that pre-determine the relative *a priori* probabilities of the two events. Of course, by just measuring the relative frequency of the two events one could determine their respective values *a posteriori* after a sufficiently large number of tries N , and on the way measure all other properties of the binary distribution including as its width (second moment), etc.

^aSpoken: “ N choose n_1 ”.

But this is not required nor is it even desirable. One might attribute $p_1 = p_2 = 1/2$ by *symmetry* to a perfectly milled coin, *without* performing the experiment. Tossing it N times should either confirm the hypothesis or show up a hidden flaw. Similarly one can predict the width of the binary distribution from theory alone, without performing the experiment.

Thus it becomes quite compelling to understand the consequences of a probability distribution at arbitrary values of the parameters. Experiment can then be used not just to determine the numerical values of the parameters but also to detect systematic deviations from the supposed randomness.

These are just some of the good reasons not to insist on $p_1 + p_2 = 1$ at first. By allowing the generating function to depend on *two* independent parameters p_1 and p_2 it becomes possible to derive all manners of useful (or at least, entertaining), identities. In the first of these one sets $p_1 = p_2 = 1$ in (1.1) and immediately obtains the well-known sum rule for binomial coefficients:

$$\sum_{n_1=0}^N \binom{N}{n_1} = 2^N. \quad (1.3a)$$

Setting $p_1 = -p_2$ yields a second, albeit less familiar, sum rule:

$$\sum_{n_1=0}^N \binom{N}{n_1} (-1)^{n_1} = 0. \quad (1.3b)$$

Next, expand $(p_1 + p_2)^{n+k}$ in powers of p_1 and p_2 as in (1.1) while similarly expanding each factor $(p_1 + p_2)^n$ and $(p_1 + p_2)^k$ in powers of p_1 and p_2 . Upon equating the coefficients term by term one derives the “addition theorem” for the binomial coefficients:

$$\binom{n+k}{r} \equiv \sum_{t=0}^r \binom{n}{t} \binom{k}{r-t}. \quad (1.3c)$$

Two special cases of this formula may prove useful. In the first, set $k = 1$, so that t is restricted to $r - 1$ and r . Recalling that $0! \equiv 1$ and $1! = 1$ we deduce $\binom{n+1}{r} = \binom{n}{r} + \binom{n}{r-1}$ from Eq. (1.3c).

In the second example set $k = n = r$. Then (1.3c) yields:

$$\binom{2n}{n} = \sum_{t=0}^n \binom{n}{t} \binom{n}{n-t} = \sum_{t=0}^n \binom{n}{t}^2.$$

Moreover, by retaining p_1 and p_2 as independent variables in G_N it becomes possible to obtain *all* moments of the distribution simply by differen-

tiating the generating function multiple times, following which $p_1 + p_2 = 1$ is imposed. Let us start by evaluating the lowest moment, i.e. the average of n_1 , denoted $\langle n_1 \rangle$ (sometimes also written \bar{n}_1).

$$\begin{aligned} \langle n_1 \rangle &\equiv \sum_{n_1=0}^N n_1 W_N(n_1) = \left\{ p_1 \frac{\partial}{\partial p_1} G_N(p_1, p_2) \right\} \Big|_{p_2=1-p_1} \\ &= \left\{ p_1 \frac{\partial}{\partial p_1} (p_1 + p_2)^N \right\} \Big|_{p_2=1-p_1} = N p_1. \end{aligned} \quad (1.4a)$$

Similarly,

$$\langle n_1^2 \rangle = \left\{ \left(p_1 \frac{\partial}{\partial p_1} \right)^2 (p_1 + p_2)^N \right\} \Big|_{p_2=1-p_1} = (N p_1)^2 + N p_1 (1 - p_1). \quad (1.4b)$$

For higher powers also, $n_1^m \leftrightarrow (p_1 \partial / \partial p_1)^m$ is always the correct substitution. A measure of the width or “second moment” of the distribution may be derived from the *variance* σ , here defined as $\sigma^2 \equiv \langle (n_1 - \langle n_1 \rangle)^2 / N \rangle = (\langle n_1^2 \rangle - \langle n_1 \rangle^2) / N$. In the present example, inserting the result (1.4a) in (1.4b) we find $\sigma = \sqrt{p_1 p_2} = \sqrt{p_1 (1 - p_1)}$. This result can be put to immediate and practical use.

1.2. Length of a Winning Streak

In casino gambling, unlike some other real-life situations, persistence is not a virtue. Take as an example the most favorable situation of a “winning streak.” Under the assumption that a coin toss resulting in heads wins 1 unit whereas tails loses 1 unit, $d = n_1 - n_2 = N - 2n_2$ measures the net winnings (or, if negative, losses). Here $N = n_1 + n_2$ is the number of plays, presumably proportional to the time t spent playing. With an honest coin $p_1 = p_2 = 1/2$ and therefore according to Eq. (1.4a) the average $\langle d \rangle = 0$ and $\langle d^2 \rangle = N^2 - 4N(N/2) + 4\{(N/2)^2 + N(1/2) \times (1/2)\} = N$, according to (1.4b). Then, defining $d_{rms} \equiv (\langle d^2 \rangle - \langle d \rangle^2)^{1/2}$, we see

$$d_{rms} = N^{1/2} \quad (1.5)$$

If the coin toss is unbiased the most probable outcomes after N tries lies randomly between $+N^{1/2}$ and $-N^{1/2}$ and averages to zero. But this is unrealistic: one should also correct for any systematic “tax” or “bias” on the game, which might take the form $-\alpha N$, with $\alpha = |p_2 - p_1| > 0$ assumed to be a small number. Figure 1.1 locates the most probable outcomes, those

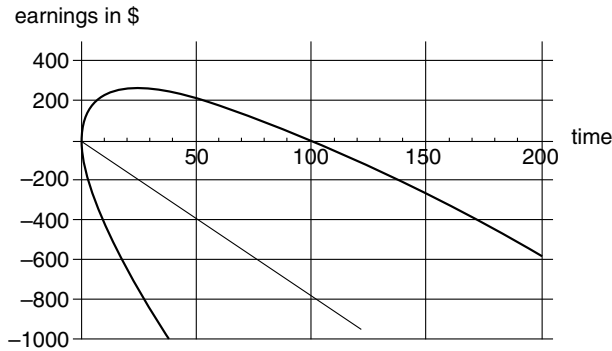


Fig. 1.1. Examples of most probable gambling streaks.

This plot of “Earnings versus Time spent playing”, shows the *most* probable gambling trajectory (according to elementary *biased* RW theory), as a thin straight line. The top curve is the “winning” streak at one standard deviation from the most probable; after an initial winning spurt “in the black” it goes “into the red” after $t > 100$, for the arbitrary parameters used in this illustration. Lower curve (the losing streak), also one standard deviation from the most probable, is negative from the start. Asymptotically *any* reasonably probable trajectory lying between these two curves must end up deeply “in the red”.

lying between the two curves: $-\alpha N \pm N^{1/2}$. It shows the most probable outcomes are all negative once N exceeds $1/\alpha^2$, regardless how small is α . This model inspires the one-dimensional *biased random walk* explored further in Problem 1.1.

1.3. Brownian Motion and the Random Walk

The Scottish botanist Robert Brown observed in 1827 that grains of pollen, coal dust, or other specks of materials in liquid suspension and visible under a microscope, appeared to jump randomly in position and direction. The physical explanation provided by Einstein in 1906 invoked multitudes of invisible molecules striking the visible particles, imparting large numbers of random impulses to them. Let us consider one simplified version of this kinetic theory; a second will follow.

Consider a completely random walker (RW) (either the above mentioned speck or the proverbial “drunken sailor”) whose position from the origin after n steps is \mathbf{r}_n . Each step is $\mathbf{s}_{n+1} = \mathbf{r}_{n+1} - \mathbf{r}_n$. Assume a given step length s_n and a perfectly random direction $\hat{\mathbf{s}}_n$, each step being uncorrelated with those preceding it. By symmetry $\langle \mathbf{s}_n \rangle = 0$. We define $\lambda^2 = \langle \mathbf{s}_n^2 \rangle$ as the

average square step-length. Then if \mathbf{r}_N is the destination,

$$\langle \mathbf{r}_N^2 \rangle = \left\langle \left(\sum_{n=1}^N \mathbf{s}_n \right)^2 \right\rangle = N\lambda^2 + 2 \sum_{n=1}^{N-1} \sum_{m=n+1}^N \langle \mathbf{s}_n \cdot \mathbf{s}_m \rangle. \quad (1.6)$$

Due to the lack of correlations all the terms in the double sum can be factored, i.e. $\langle \mathbf{s}_n \cdot \mathbf{s}_m \rangle = \langle \mathbf{s}_n \rangle \cdot \langle \mathbf{s}_m \rangle = 0$, and vanish. Thus the rms distance achieved by RW is $R = \lambda\sqrt{N}$ and lies with equal probability at any point on the surface of a sphere of radius R in three dimensions (3D), on a circle of radius R in 2D and at the two points $\pm R$ in 1D. As the number of steps can be assumed to be proportional to the elapsed time ($N \propto t$) the most probable distance from the origin increases as $R \propto \sqrt{t}$. This power law is recognized as the signature of classical *diffusive* motion, just as $R \propto t$ is the signature of *ballistic* motion.

Problem 1.1. A given biased RW is defined by $\langle \mathbf{s}_n \rangle = \mathbf{a}$ and $\langle s_n^2 \rangle = \lambda^2$, with \mathbf{a} and λ constants. Determine the two rms loci of this biased random walker after N steps in 1D (as in Fig. 1.1) and generalize to 2D and 3D, as function of a_x/λ , a_y/λ and in 3D, a_z/λ .

1.4. Poisson versus Normal (Gaussian) Distributions

Both of these well-known statistical distributions can be derived as different limiting cases of the binomial distribution. In this regard the Gamma Function $\Gamma(z)$ and *Stirling's approximation* to $N!$ and $\Gamma(N+1)$ prove useful. First, define

$$\Gamma(z) \equiv \int_0^\infty dt t^{z-1} e^{-t} \quad (1.7)$$

as a function that exists everywhere in the complex z -plane except on the negative real axis. After partial integration on t^{z-1} one finds,

$$\Gamma(z) = z^{-1}\Gamma(z+1), \quad \text{i.e. } \Gamma(z+1) = z\Gamma(z). \quad (1.8)$$

$\Gamma(1) = 1$ from Eq. (1.7) (by inspection). Hence $\Gamma(2) = 1$, $\Gamma(3) = 2 \cdot 1$ and by induction on any positive integer N , $\Gamma(N) = (N-1)!$ Hence, $\Gamma(1) = 0! = 1$, and $(-N)! = \infty$ by extension. This allows the limits on the sum in Eq. (1.1) to be extended, from $0 \leq n \leq N$, to $-\infty < n < +\infty$, if desired.

Half-integer arguments are of equal importance. To obtain them, first calculate $\Gamma(\frac{1}{2})$:

$$\begin{aligned} \left\{ \Gamma\left(\frac{1}{2}\right) \right\}^2 &\equiv \int_0^\infty dt t^{-1/2} e^{-t} \int_0^\infty ds s^{-1/2} e^{-s} \\ &= \int_{-\infty}^\infty dx e^{-x^2} \int_{-\infty}^\infty dy e^{-y^2} \\ &= \int_0^{2\pi} d\phi \int_0^\infty dr r e^{-r^2} = \pi \end{aligned} \quad (1.9)$$

(as obtained by substituting $t = x^2$, $s = y^2$, then switching to radial coordinates). Thus: $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, $\Gamma(\frac{3}{2}) = \frac{1}{2}\sqrt{\pi}$, $\Gamma(\frac{5}{2}) = \frac{3}{2} \cdot \frac{1}{2}\sqrt{\pi}$, etc.

Problem 1.2. Find a *formula* for area of unit sphere $A(d)$ in d dimensions.

[Hint : using $\mathbf{r}^2 = r_1^2 + r_2^2 + \dots + r_d^2$, compute :

$$\begin{aligned} \int d^d r e^{-\mathbf{r}^2} &= A(d) \int_0^\infty dr r^{d-1} e^{-r^2} = \frac{A(d)}{2} \int_0^\infty dt t^{(\frac{d}{2}-1)} e^{-t} \\ &= \left[\int_{-\infty}^\infty dx e^{-x^2} \right]^d . \end{aligned}$$

Evaluate $A(d)$ explicitly for $d = 1, 2, 3, 4, 5$.

Stirling's approximation to $\Gamma(N + 1)$ is obtained by setting $t^N e^{-t} \equiv \exp[g(t)]$ and evaluating it by steepest descents. At the maximum of $g(t)$, defined as $t = t_0$, $g'(t) = \partial/\partial t\{-t + N \log t\}|_{t_0} = 0 \Rightarrow t_0 = N$. Approximating $g(t)$ by the first few terms in its Taylor expansion we find $g(t) = g(t_0) + [(t - t_0)^2/2!]g''(t_0) + [(t - t_0)^3/3!]g'''(t_0) + \dots$, with $g''(t_0) = -1/N$. Third and higher derivatives become negligible in the large N limit. Thus, $g(t) = N \log N - N - (1/2N)(t - N)^2$ is to be inserted into the integral for $\Gamma(N + 1)$:

$$\begin{aligned} \Gamma(N + 1) &= N! = \exp(N \log N - N) \int_{-N}^\infty dt \exp\left(-\frac{t^2}{2N}\right) \\ &= \sqrt{2N\pi} \exp(N \log N - N). \end{aligned} \quad (1.10)$$

The logarithm of this result yields the more familiar expression, $\log(N!) = N \log N - N + 1/2 \log(2N\pi)$. We note that in most applications the first two terms suffice and the last term is omitted.

Problem 1.3. Estimate the fractional errors in Stirling's approximation arising from the two sources: the neglect of the next term $(1/3!g'''(t - t_0)^3)$ in the expansion of $g(t)$ and the approximation of $-N$ by $-\infty$ in the limits of integration. Compare Stirling's result with the exact values of $\Gamma(z)$ for $z = 3, 3.5, 10$ and 10.5 and obtain the fractional errors numerically. How well do they agree with your estimate?

The *Poisson* distribution is named after the renowned 19th Century mathematician and physicist who, in the Napoleonic wars, was required to analyze the tragic, albeit uncommon, problem of soldiers kicked to death by mules. Was it greater than random? The distribution that bears his name applies to trick coins, radioactive decay of metastable nuclei, and other instances in which some remarkable event being monitored is highly improbable; it is obtained as a limiting case of the binomial distribution in the $\lim p_1 \rightarrow 0$. Define $\lambda = Np_1$ as a new, finite, parameter of $O(N^0)$, in the thermodynamic $\lim N \rightarrow \infty$. Thus $n_1 \ll N$ and $n_2 \approx N$. It then becomes permissible to approximate $N!/n_2! \equiv N(N-1)\cdots(N-n_1+1)$ by N^{n_1} and $(1-p_1)^{n_2}$ by $(1-\lambda/N)^N \rightarrow e^{-\lambda}$ (recall the definition of e). With these substitutions $W_N(n_1) \rightarrow P(n_1)$, we obtain the Poisson distribution:

$$P(n) = \frac{e^{-\lambda}}{n!} \lambda^n \quad [\text{Poisson}]. \quad (1.11)$$

Remarkably, despite the several approximations, normalization is preserved — that is, the sum rule $\sum P(n) = 1$ continues to be satisfied exactly.

The more familiar, i.e. *normal*, distribution is that due to Gauss. The *Gaussian* distribution can also be derived from the binomial whenever the p 's are both nonzero and the number $N \rightarrow \infty$. Its applications are ubiquitous: the distribution of grades in large classes, of energies in a classical gas, etc.

Because both n_1 and n_2 scale with N the ratios n_1/N and n_2/N can, in some sense, be treated as continuous variables (this is not possible if n_1 is $O(1)$ as in the Poisson distribution). Clearly, $W_N(n_1)$ has its maximum at some \tilde{n}_1 which is then defined as the “most probable” value of n_1 . As $W_N(n_1)$ must be “flat” at the maximum we look for solution of $W_N(n_1 \pm 1) = W_N(n_1)$.

Using Eq. (1.2), we obtain to leading order in $1/N$:

$$1 \approx \left(\frac{N}{\tilde{n}_1} - 1 \right) p_1 (1 - p_1)^{-1}, \quad \text{i.e. } \tilde{n}_1 = p_1 N = \langle n_1 \rangle. \quad (1.12)$$

In other words, the *most probable* values of n_1 and n_2 turn out to be identical to their respective *average* values, $p_1 N$ and $p_2 N$, i.e. $\tilde{n} = \langle n \rangle$.

Next let us use Stirling's approximation, Eq. (1.10), to expand $\log W_N$ about its optimum value. With $\sigma^2 = p_1 p_2$ (recall §1.1), the result is:

$$\log(W_N(n_1)) = \log(W_N(\tilde{n}_1)) - \frac{1}{2} \frac{(n_1 - \tilde{n}_1)^2}{N\sigma^2} + O\left(\frac{(n_1 - \tilde{n}_1)^3}{N^2}\right). \quad (1.13)$$

The entire Taylor series expansion of the exponent is replaced by its leading terms, as discussed further in Problem 1.4. The *form* of the leading term in (1.13) strongly suggests defining a new independent variable $x = n_1/N^{1/2}$ which becomes continuous in the lim. $N \rightarrow \infty$. Clearly its range is from $-N^{1/2}$ to $+N^{1/2}$ and spans the entire real line in the limit. Also, $\tilde{x} = \tilde{n}_1/N^{1/2}$.

In changing the independent variable from n_1 to x we should require that the probability function in the new variable $P(x)$ be given by $dx P(x) = dn_1 W_N(n_1)$, to ensure that $P(x)$ — like $W_N(n_1)$ — remains properly normalized. Then, $P(x) = \frac{dn_1}{dx} W_N(n_1) = N^{1/2} W_N(xN^{1/2})$,

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\tilde{x})^2}{2\sigma^2}} \quad [\text{Gaussian, aka "normal" distribution}]. \quad (1.14)$$

The remainder of the exponent summarized by $O\left(\frac{(n_1 - \tilde{n}_1)^3}{N^2}\right) = O\left(\frac{(x - \tilde{x})^3}{\sqrt{N}}\right) \xrightarrow{N \rightarrow \infty} 0$ vanishes for all finite x in the lim. $N \rightarrow \infty$. This choice of variables is what makes the parameter N disappear from the normal distribution (1.14) and reveals the *limiting function* of the binomial distribution in the large number limit, after the variable n_1 is appropriately scaled. It is not the first, nor will it be the last time that one finds such simplification and universality in the “large number limit.” Actually this limit is called “the *thermodynamic* limit,” for reasons that will become abundantly clear.

Problem 1.4. (A) Treating n_1 as a continuous variable, derive (1.13) and show that the leading correction generically denoted by $O\left(\frac{(n_1 - \tilde{n}_1)^3}{N^2}\right)$ in the text is, precisely, $(p_2^2 - p_1^2)\frac{(n_1 - \tilde{n}_1)^3}{3!N^2\sigma^4}$. Due to the vanishing of the Gaussian distribution $P(x)$ outside the range $\tilde{n}_1 \pm O(\sqrt{N})$, show that this correction and *all* further correction terms are negligible throughout the range where $P(x)$ is finite.

(B) If this is true, the distribution is *should* remain normalized, given that the original, discrete, distribution adds up to 1 ($G_N(p, 1-p) \equiv 1$ in Eq. (1.1).) Using $P(x)$ in (1.14) show that indeed, the corrections all cancel and $\int_{-\infty}^{+\infty} dx P(x) = 1$.

(C) Then, calculate $\langle x^2 \rangle = \int dx P(x)x^2$ and compare your answer to Eqs. (1.4), analyzing the various terms in some detail.

Problem 1.5. Using the normal distribution (1.14) and the result of the preceding Problem, show that:

$$\langle (n_1 - \tilde{n}_1)^k \rangle = \begin{cases} (k-1) \cdots 5 \cdot 3 \cdot 1 \cdot (N\sigma^2)^{k/2} & \text{if } k \text{ is even,} \\ 0 & \text{if } k \text{ is odd.} \end{cases}$$

1.5. Central Limit Theorem (CLT)

This useful theorem formalizes the preceding results and helps derive the normal distribution without reference to the binomial expansion. Definitions and an example follow.

Generally speaking, the CLT applies to *any* process that results from a large number of small contributions — as in the example of Brownian motion, where the motion of a dust particle is the results of its unobservable collisions with a myriad of smaller, invisible, host molecules. To analyze such situations in more detail we must learn to formulate probabilities for two or more variables.

Let us start with 2 and generalize to N by induction. Let $P(s, t)$ be the probability distribution of 2 independent variables s and t over a finite range. The requirements are that P must be non-negative over that range and its integral normalized, $\int ds \int dt P(s, t) = 1$. If one integrates over all t , the remainder is a probability for s . Denote it: $P_1(s) = \int dt P(s, t)$. Unless $P(s, t)$ is a symmetric function, P_1 differs from the analogously defined $P_2(t) = \int ds P(s, t)$.

Averages over arbitrary functions of the two variables are evaluated in the usual way,

$$\langle f(s, t) \rangle = \int ds \int dt f(s, t) P(s, t).$$

Consider $f = g(s)h(t)$ in the special case where s and t are *statistically independent*. Then $\langle f \rangle = \langle g \rangle_1 \langle h \rangle_2$ (the subscripts indicate averages over P_1 or P_2 respectively). This holds for arbitrary g and h iff^b $P(s, t) \equiv P_1(s)P_2(t)$. The generalization is $P(s, t, u) = P_1(s)P_2(t)P_3(u)$, etc. In the following we examine the x -dependence of a one-dimensional RW in which the end-point $x = s_1 + s_2 + \dots + s_N$ is the sum over N individual steps, each assumed statistically independent of the others but all governed by the same probability $p(s)$,^c i.e. $P(s_1, \dots, s_N) = \prod p(s_j)$.

At this point it is helpful to introduce the *Dirac delta function* $\delta(z)$, a function that is zero everywhere except at the origin where it is infinite — such that its integral is 1. With the aid of this singular function we can write,

$$\begin{aligned} P(x) &= \int ds_1 \int ds_2 \cdots \int ds_N \delta\left(x - \sum_{n=1}^N s_n\right) P(s_1, s_2, \dots, s_N) \\ &= \int ds_1 p(s_1) \int ds_2 p(s_2) \cdots \int ds_N p(s_N) \delta\left(x - \sum_{n=1}^N s_n\right). \end{aligned} \quad (1.15a)$$

The Dirac delta function has numerous representations, such as the limit of an infinitely high and narrow rectangle of area 1, etc. The one most particularly helpful here is: $\delta(\Delta x) = (1/2\pi) \int_{-\infty}^{+\infty} dk e^{ik\Delta x}$. As we saw in Sec. 1.2, the most probable $\Delta x \equiv (x - N\langle s_n \rangle)$ is $O(\sqrt{N})$. The dominant contributions to the integration in (1.15a) are from regions where the product $k\Delta x$ is $O(1)$. Therefore the important values of k are at most $O(1/\sqrt{N}) \rightarrow 0$.

$$P(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} dk e^{ikx} \{Q(k)\}^N$$

where

$$\begin{aligned} Q(k) &= \int ds e^{-iks} p(s) \\ &= 1 - \frac{1}{1!} ik \langle s \rangle + \frac{1}{2!} (-ik)^2 \langle s^2 \rangle + \frac{1}{3!} (-ik)^3 \langle s^3 \rangle + \dots \end{aligned} \quad (1.15b)$$

^bIf and only if.

^cThe reader might wish to consider a case in which the p 's differ, e.g. suppose the individual $p_n(s_n)$ to depend explicitly on n .

Collecting powers, we evaluate $\log Q$ to leading orders:

$$Q(k) = e^{-ik\langle s \rangle} e^{-(k^2/2)(\langle s^2 \rangle - \langle s \rangle^2)} e^{O(k^3)}. \quad (1.16)$$

The coefficients in the exponent are the so-called *moments* of the distribution (also denoted *cumulants* or *semi-invariants*) at each individual step. We identify the second moment as the variance $\sigma^2 = \langle s^2 \rangle - \langle s \rangle^2$ of an individual step in the above, to obtain:

$$P(x) = \frac{1}{2\pi} \int dk e^{ikx} e^{-ikN\langle s \rangle} e^{-(Nk^2\sigma^2/2)} e^{N \cdot O(k^3)}. \quad (1.17a)$$

The integration is rendered more transparent by a change of variables. Let $x \equiv N\langle s \rangle + \xi\sqrt{N}$ where ξ is a measure of the fluctuations of x about its systematic (i.e. *biased*) value $N\langle s \rangle$.

In the present application we find that *all moments beyond the second are irrelevant*. For example, $N \cdot O(k^3)$ in the above exponent is $N \cdot N^{-3/2} \rightarrow 0$ throughout the range that contributes most to the integration. Higher moments are smaller yet. Replacing k by the rescaled dummy variable of integration $q = k\sqrt{N}$ we obtain^d $P(\xi)$:

$$P(\xi) = \frac{1}{2\pi} \int dq e^{iq\xi} e^{-q^2\sigma^2/2} e^{N \cdot O(q^3N^{-3/2})} \xrightarrow{\lim \cdot N \rightarrow \infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\xi^2/2\sigma^2}. \quad (1.17b)$$

This is the prototype *normal* distribution for the fluctuations. It is independent of $\langle s_n \rangle$. The derivation has preserved the norm: the integral of $P(\xi)$ over ξ in the range $-\infty$ to $+\infty$ remains precisely 1.

The generalization to the 2D or 3D RW *appears* simple, but that may be misleading. With $\mathbf{r} = N\langle \mathbf{s} \rangle + \boldsymbol{\xi}\sqrt{N}$ one readily derives:

$$P(\boldsymbol{\xi}) = \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\boldsymbol{\xi}^2/2\sigma^2}, \quad \text{in } d \text{ dimensions.} \quad (1.18)$$

However, it remains to define σ^2 appropriately for $d \geq 2$. This is not necessarily straightforward. Consider the two distinct scenarios.

- (1) In this model, the Cartesian components at each step, s_x, s_y, \dots , are uncorrelated (statistically independent) but subject to $\langle s^2 \rangle = a^2$.
- (2) Here the individual step lengths are *constrained* by $s^2 = a^2$, hence the Cartesian components are *not* independent.

The proof of (1.18) and the definition of σ in each of these two instances is left as an exercise for the reader, who is re-directed to Problem 1.1 and

^dNote: to preserve probability $P(\xi) = P(x(\xi))(dx/d\xi) = P(x(\xi))\sqrt{N}$.

its method of solution. (Hint: formulate the constraints by an appropriate use of the Dirac delta function.) We conclude this section with some related observations.

- The step distribution needs to be neither differentiable nor smooth, for the CLT to be valid. For example let each step have equal probability $1/2$ to be either $+a$ or $-a$, independent of the preceding steps. Then $Q(k) \equiv \cos ka$ and even though $p(s_n)$ is a singular function, Eq. (1.17b) continues to be valid with $\sigma^2 = \langle s^2 \rangle = a^2$.
- Finally, if instead of distance traveled we investigate the momentum \mathbf{p} transferred (or impulses imparted) to an object of mass M by numerous collisions with lighter, invisible molecules of mass $m \ll M$, as in Brown's experiments, we now know (essentially, by inspection) that the result *has to be*

$$P(\mathbf{p}) = \frac{1}{(2\pi\sigma^2)^{3/2}} e^{-\mathbf{p}^2/2\sigma^2}. \quad (1.19)$$

We can turn the variance into a measure of temperature T by defining $\sigma^2 = Mk_B T$, identifying with thermal fluctuations and where k_B is the Boltzmann constant. Then the CLT implies the familiar *Maxwell–Boltzmann* distribution in the momentum space of an “ideal” gas of mass M particles. The exponent expresses the energy of the particles in units of $k_B T$, the thermal unit of energy. Note this derivation does not ensure that any similar expression exists for the momentum distribution of the *unseen* molecules of lighter mass m .^e

1.6. Multinomial Distributions, Statistical Thermodynamics

After making some apparently *ad hoc* identifications and generalizations we intend to use the previous results to derive some sort of thermodynamics from “first principles”. This section is designed to motivate the following, more rigorous, chapters.

Assume particles carry r distinguishable labels. For example, $r = 2$ for the coin toss, 6 for dice and $2S + 1$ for spins S ($S = 1/2, 1, 3/2, \dots$). The total number of particles is then $N = \sum n_j$, where n_j is the number of particles with label j running from 1 to r . Independent of any *a priori* probabilities p_j the statistical factors are “ N choose n_1, n_2, \dots ”. The multinomial probability

^eJust as the Gaussian $P(\xi)$ does not mirror arbitrary $p(s)$ in Eqs. (1.15)–(1.17).

distribution is,

$$V_N(n_1, n_2, \dots, n_r) = N! \prod_{j=1}^r \frac{p_j^{n_j}}{n_j!}. \quad (1.20)$$

These are all positive quantities that can be obtained from the expansion of a generalized *generating function* $G = (\sum p_j)^N$. Therefore if the p_j add up to 1, $G = 1$ and the V 's are also normalized.

The V 's are sharply peaked about a maximum. Our experience with the binomial distribution has shown that even though V_N is highly singular its *natural logarithm* can be expanded in a Taylor series about the maximum. Let us first examine this maximum.

$$\log V_N = \left[\sum_{j=1}^r n_j \log(p_j) \right] + \left[\log(N!) - \sum_{j=1}^r \log(n_j!) \right]. \quad (1.21a)$$

We then *arbitrarily* identify the first bracket on the *rhs* of the equation (which involves the parameters p_i), with the negative of the energy of the system. Similarly we identify the second, purely statistical, bracket with the product of its entropy \mathcal{S} and the temperature T .^f Then, the above takes the form,

$$[\log V_N] = \beta[-E] + \beta[T\mathcal{S}]. \quad (1.21b)$$

The overall-factor β has units [energy]⁻¹, introduced to make the expression dimensionless. Each bracketed quantity in (1.21a) is *extensive* ($\propto N$), hence E and \mathcal{S} are also extensive. This requires β and T to be *intensive*, i.e. independent of N .

The signs have been chosen such that the maximum V behaves properly in two known limiting cases. It must correspond to an *energy minimum* (i.e. the “virtual forces” all vanish) when \mathcal{S} and T are held constant, and with *maximal entropy* at constant E and T .

Similarly let us identify the LHS of the equation with $-\beta F$, where F is defined as a “free energy”. Upon dividing by β we recapture the well-known thermodynamic relation: $F = E - T\mathcal{S}$. (Still lacking is identification of β as $1/k_B T$.) Thus, maximizing V is the same as *minimizing the free energy* F .

To minimize the free energy we make use of Stirling's approximation in (1.21a) and optimize *w.r.* to each n_j (treated as a continuous variable.)

$$\frac{\partial}{\partial n_j} \left\{ n_j \log p_j - n_j \left[\log \left(\frac{n_j}{N} \right) - 1 \right] \right\} = 0, \quad \text{for } j = 1, \dots, r. \quad (1.22)$$

^fThese quantities will be given a rigorous definition in the next chapter.

Note that we can add any arbitrary multiple of $(N - \sum n_j) = 0$ to the expression in curly brackets.

The solution of (1.22) yields the most probable values of the n_j 's, denoted \tilde{n}_j as before. We observe once again that the most probable value is also the average value: $\tilde{n}_j = p_j N = \langle n_j \rangle$.

In the mantra of kinetic theory (and of statistical mechanics) the so-called *ergodic hypothesis* occupies a place of honor. Crudely put, it states that any system permitted to evolve will ultimately, or asymptotically, tend to a state of maximal probability — beyond which it must cease to evolve. One makes the connection with thermodynamics by identifying the *most probable configurations* with those found in *thermodynamic equilibrium*. In the axiomatic statistical mechanics outlined in Chapter 4, this common-sense notion is elevated to a high principle.

Here, to make a more explicit connection with elementary thermodynamics, we identify the *a priori* probabilities with Boltzmann factors, $p_j = \exp[-\beta(\varepsilon_j - \mu)]$. The quantity ε_j is the energy of a particle in the j th state and the “chemical potential” μ is chosen to allow p_j to satisfy $\sum p_j = 1$. Some recognizable results can be found upon using this p_j in Eq. (1.21a). For example, the statistically averaged (also, the most probable) internal energy is

$$N \sum_{j=1}^r e^{-\beta(\varepsilon_j - \mu)} (\varepsilon_j - \mu) \equiv E - \mu N,$$

in which E is defined as the total averaged “physical” energy. We shall find similar expressions in connection with the theory of ideal gases.

1.7. The Barometer Equation

To show that the identification of the probabilities p_j with $\exp -\beta\varepsilon_j$, the energy exponential, is not entirely capricious nor limited to discrete systems, let us here derive the empirical *barometer equation*. To start, identify the probability p_j with the relative density $\rho(z)$ (*number of particles per unit volume*) of the atmosphere at a level z above the ocean floor. With $Mg =$ force of gravity on masses M and $d\varepsilon = Mg dz$ (neglecting the slight changes in g with altitude,) this identification is equivalent to:

$$d \log \rho(z) = -\beta d\varepsilon = -\beta Mg dz \quad (1.23)$$

which integrates to $\rho(z) = \rho(0) \exp -\beta Mgz$. The parameter β having units $[\text{energy}]^{-1}$ is once again introduced to ensure that the exponent is dimensionless, we identify it as $\beta = 1/k_{\text{B}}T$, where k_{B} is the Boltzmann constant. According to the ideal gas law, the local pressure is linearly dependent on the local density ($p \propto \rho T$) and thus,

$$p(z) = p(0)e^{-Mgz/kT}, \quad (1.24)$$

the well-known barometer equation. Taken together with (1.19), this formula strongly suggests that in general, *both* kinetic *and* potential energies have equal rôles to play in the exponential “Boltzmann factor.” While this may seem intuitive and unremarkable — considering that either form of energy is readily transformable into the other — a true *proof* requires a more systematic formulation of statistical mechanics, such as that developed in the following three chapters.

1.8. Other Distributions

There are instances in which the *lognormal* distribution plays a role, in which it is the logarithm of the random variable that is normally distributed. Among other, more exotic distributions, we count the *stretched exponential*, the *Lorentzian*, and numerous others for which the moments do or do not exist, as covered in numerous texts on mathematical statistics.^g We deal with some of them elsewhere in this book. But with the sole exception of the “Lorentzian”, $P(x) = \frac{\gamma/\pi}{x^2 + \gamma^2}$, what makes them exotic is their rarity in physical applications, in comparison with either the Gauss or Poisson distributions. On the other hand, the far more esoteric *stochastic matrices* have found their niche in physics ever since their introduction by Wigner,^h who approximated the Hamiltonian of the quantum nuclear many-body problem by a random matrix.

It is helpful to visualize a random matrix as some kind of a quantum-mechanical RW in which the n th step connects not just the n th position of the random walker to the $n + 1$ st, but also to all other positions that he ever has or ever will occupy. One may denote this process a “multiply-connected”

^gSee, e.g., M. Evans, N. Hastings and B. Peacock, *Statistical Distributions*, Wiley, New York, 1993. This 2nd edition describes 39 “major” distributions. We note that their terminology “distribution function” for the *integral over* $P(x)$ differs from common physics usage such as found in the present text.

^hSee M. L. Mehta, *Random Matrices*, Academic, New York, 1967.

RW and represent it by an $N \times N$ real, symmetric, matrix:

$$\tilde{M} = \begin{pmatrix} m_{11} & m_{12} & \cdots & \cdots \\ m_{12} & m_{22} & \cdots & \cdots \\ \vdots & \vdots & \ddots & \\ \vdots & \vdots & & m_{NN} \end{pmatrix}. \quad (1.25)$$

In one well-known example, the individual matrix elements m_{ij} are real, random and normalized as follows: $m_{ij} = \varepsilon_{ij}/\sqrt{N}$, with $\langle \varepsilon_{ij} \rangle = 0$ and $\langle \varepsilon_{ij}^2 \rangle = \sigma^2$. All $N(N+1)/2$ matrix elements on or above the main diagonal are statistically independent. One can prove that such a matrix has N real eigenvalues λ_n which, in the *thermodynamic limit* $N \rightarrow \infty$, are smoothly distributed according to a distribution function $P(\lambda)$ in the interval $-2\sigma < \lambda < +2\sigma$. As in the CLT, the global result does not depend on the distribution of the $p(\varepsilon)$ of the individual random variables. So one can assume for the latter whichever is more convenient: the binomial distribution (each independent $\varepsilon_{ij} = \pm\sigma$ with equal probability) or the continuous Gaussian ensemble, $P(\varepsilon_{ij})$ with $\sigma = 1$.

Problem 1.6. (a) Using a random number generator for the individual ε 's, use your *PC* to numerically diagonalize the matrix M in Eq. (1.25), specifying $N = 101$. List the eigenvalues λ_n in ascending order. Plot the level spacings $\lambda_{n+1} - \lambda_n$ as a function of λ_n .

(b) Estimate how large N must be in order that $P_N(\lambda)$ approach its asymptotic limit function $P(\lambda)$ to within $\pm 1\%$ everywhere in the interval $-2 < \lambda < +2$. (Note: the analytical form of $P(\lambda)$ is derived in a later chapter of this book.)

Problem 1.7. A gambler G is involved in a game that pays off $(100 \pm x)\%$ of his bet, with x a fixed number in the range $100 > x > 0$. The sign \pm is perfectly random. He plays it N times against a much richer opponent (the bank) B , starting with an initial stake $\$D$. At each step he tenders his *entire stake* (including any winnings or losses.) Be warned, this is an example where the “most probable” outcome differs considerably from the *average*.

(A) Show that for any x and despite any fluctuations, with *greatest probability* G 's stake *must* ultimately go to zero exponentially when $N > 10^4$ is large. Is this paradoxical?

(B) Show that if, on the other hand, all winning (and losing) streaks of arbitrary lengths are allowed. G ends up with *precisely* his initial stake $\$D$ *on average*.

(C) By how small an amount must the odds change in order that, *on average*, G always wins (or loses) *consistently* at large N ?
