Counting, Probability, and Statistics

Table of Contents

• basic ideas of probability related to coin tossing and similar random events
• combinations and permutations
• binomial coefficients and Pascal's triangle
• application to Maxwell-Boltzmann, Fermi-Dirac, and Bose-Einstein statistics
• mean, variance, and standard deviation for discrete and continuous distributions
• binomial, Poisson, and Gaussian distributions
• uncertainty analysis: propagation of random errors

Tools of the Trade

• Stirling's approximation
• error function

Possible Future Additions

• least-squares fitting
• correlation coefficient, reduced chi square, and other statistical tests
• shot noise, kinetic theory, information theory, ergodic hypothesis, random walk
• mean thermal occupancy of states, Planck spectrum
• induction proofs and examples

Basic Probability Theory

We define a *sample space* as a list of all possible distinct outcomes of an experiment. Suppose we label the result of the first outcome as x_1, the second as x_2, and so on up to x_N if there is an exhaustive total of N mutually exclusive outcomes. (For the moment, we are supposing that the number of possible outcomes is finite and definite; later we will generalize to continuous variables that have an infinite number of different outcomes.) By "exhaustive" and "mutually exclusive" we mean that one and only one listed outcome occurs each time the experiment is performed such that all possible outcomes are included in the sample space and two outcomes cannot occur simultaneously. We label these results in general as x_i where i runs from 1 to N. For example, x_i might represent the number of heads you get when you toss three pennies simultaneously (in which case $\{x_1 = 0\,,\ x_2 = 1\,,\ x_3 = 2\,,\ x_4 = 3\}$ would be an organized way to list the sample space). These outcomes are in general not equally likely. We wish to work out the *probability* p_i that the result of the experiment is x_i. What we mean by this is that if we repeat the same experiment a very large number of times, the fraction of times that we get result x_i will approach p_i ; it would become exactly p_i in the limit that the experiment were repeated an

infinite number of times. From this, we deduce the following *operational* definition of probability. Suppose we conduct an experiment some large number of times, say n times. If the number of times we obtained result i was n_i, then $p_i = n_i / n$. (Strictly speaking, p_i is only *approximately* equal to n_i / n. The equality is exact only in the limit as $n \to \infty$ which we cannot of course achieve in practice.) Notice that

$$\sum_{i=1}^{N} n_i = n \quad \Rightarrow \quad \sum_{i=1}^{N} p_i = 1 . \tag{1}$$

We call the set of values $\{p_1, p_2, \ldots, p_N\}$ the (finite) *probability distribution* and Eq. (1) is the *normalization* condition.

For example, if we draw one playing card from an ordinary shuffled deck, the probability that it will be the two of diamonds is 1/52. More generally, if all N outcomes of some experiment are equally likely, then every outcome has the same probability $p_1 = p_2 = \cdots = p_N = 1/N$. Clearly Eq. (1) is satisfied in this case.

Next suppose we toss a coin. If the coin is a fair toss (the coin is not "loaded" nor thrown in some fashion that predisposes one face to preferentially land up, and rare events such as landing on edge are excluded) then there is a probability of 1/2 of getting heads (h) and a probability of 1/2 of getting tails (t). If we now toss the coin a second time and list out the sample space for the set of two tosses, we get {hh, ht, th, tt} where the first letter in any paired entry indicates the result of the first toss and the second letter that of the second toss. Since each of these 4 outcomes is equally likely, each has a probability of 1/4. We notice that $1/4 = (1/2) \times (1/2)$ which we state in words as: The probability of any outcome for two coin tosses is the square of the probability of any outcome for one coin toss. This equality is a special case of a general principle. Suppose p_A is the probability that some event A will occur, and $p_{B \text{ if } A}$ is the probability that some event B will occur if event A has already occurred. Then the probability that A occurs followed by B is just the product of the two individual probabilities, $p_A p_{B \text{ if } A}$. For example, suppose there are 5 black balls and 10 white balls in a bag. We draw one out at random and then, without replacing it, draw out a second ball. What is the probability that the first ball is white and the second ball is black? The probability of drawing a white ball the first time is 10/15. The probability of subsequently drawing a black ball is 5/14, because there are 14 balls left and 5 of them are black if we drew a white ball out first. Thus the overall probability is $(10/15) \times (5/14) = 5/21$. On the other hand, if we return the first ball to the bag before drawing another ball, then the probability of getting a black ball on the second try is *independent* of what color ball we got on the first try and we write $p_{B \text{ if } A} = p_B$. Thus the probability of getting first a white ball and then a black ball is now $(10/15) \times (5/15) = 2/9$, which is a little smaller than 5/21.

As an aside, the probabilities in this example assumed that we drew the balls out of the bag *at random*. Needless to say, it is very difficult to decide when some event *really* is random. For example, consider a purported "random number generator" that is supposed to randomly output digits between 0 and 9. It outputs the following sequence: 1620569660240580381501935 11. Is this sequence of digits random? In fact, it's not; I simply copied it from the beginning of the 18th line of the first 100,000 digits of π from the website http://www.geom.uiuc.edu/~huberty/math5337/groupe/digits.html. Examples of this sort will probably convince you that it's extremely difficult to tell if some events are truly "random" or not; in fact, the preceding sequence of digits is probably as good as true "randomness" for most practical purposes. This is an interesting philosophical issue, but we do not have time to pursue it further here. We will just assume that randomness is possible and that in practice events such as flipping coins and drawing cards from a shuffled deck are close enough to random that we can declare their outcomes to be equally likely. (The standard Copenhagen interpretation of quantum mechanics presumes the concept of randomness when one considers Schrödinger's cat, two-slit interference, the Einstein-Podolsky-Rosen paradox, and similar ideas.)

Returning now to our example of coin flipping, each toss is independent of the others, and hence the probability of getting any particular *sequence* for two subsequent tosses (such as ht) is $p_A p_B = (1/2)\times(1/2) = 1/4$. By a "sequence," we mean that we distinguish the order of the two coin tosses, so that ht and th are distinct results. Another way to get such a sequence would be to toss two coins simultaneously (rather than flipping one coin twice in a row) *provided* we can keep track of which coin is which. (For example, one coin might be a penny and the other a dime. Or we might flip one coin with our left hand and the other with our right hand. One way or another, we can label one coin as "A" and the other as "B" and always keep track of which is which.) Clearly there is a one-to-one correspondence between sequences resulting from flipping one coin n times and those resulting from flipping n distinguishable coins simultaneously.

On the other hand, if we cannot distinguish the n coins tossed simultaneously (or equivalently, if we do not keep track of the order of the results when we flip one coin n times), then the best we can do is count the number of heads and tails. For instance, suppose I have four quarters, one of which is painted green, one blue, one red, and one yellow. I throw them up in the air and catch them. An example of a *sequence* is: the green coin is heads, the blue is tails, the red is tails, and the yellow is heads. But the *count* is 2 heads, regardless of whether I can distinguish color or not.

Now suppose we are interested in determining the probability of getting a certain number of heads when we toss a coin n times. For example, there are only three possible counts for two coin tosses: $x_1 = 0$ (zero heads which can only occur for sequence tt), $x_2 = 1$ (one head which occurs for either sequence th or ht), and $x_3 = 2$ (two heads which only arises for sequence hh).

The probability of any outcome is equal to the sum of the probabilities of the sequences contributing to that particular outcome. Thus the probability of tossing a coin twice and getting zero heads is $p_1 = 1/4$, that of getting one head is $p_2 = 1/4 + 1/4 = 1/2$, and that of getting two heads is $p_3 = 1/4$. Notice that the sum of the probabilities of all possible counts must equal 1, because that is the sum of the probabilities of all possible sequences by our normalization condition in Eq. (1).

In general then to determine the probability of getting a certain count, we need to determine the number of possible sequences that correspond to any given count. This is called the number of *combinations*. For example, if I flip *n* coins, in how many different ways can I get *r* heads? (We assume that $0 \le r \le n$ if this question is to make sense.) We denote this number of combinations by $C(n,r)$. Other common notations for this quantity are ${}_nC_r$ and $\binom{n}{r}$. (Note that fixing the value of *r* corresponds to getting exactly *r* heads and $n-r$ tails.) In order to calculate this number, we first calculate a related value called the number of *permutations* of *n* objects taken *r* at a a time. This is denoted as $P(n,r)$ or ${}_nP_r$. For example, suppose the objects are the first six letters of the alphabet {A,B,C,D,E,F} and that I want to form all distinct three-letter words (whether grammatical or not!) with no repeated characters, so that $n = 6$ and $r = 3$. The first letter can be any of the *n* choices, so there are *n* ways I can select it. Now for the selection of the second character, only $n-1$ letters remain, so I have $n-1$ choices for it. Proceeding similarly, I continue until I come to the selection of the final character, i.e., the *r*th letter in the word, for which there are $n-r+1$ choices.

Exercise: By considering the preceding alphabetical example of $n = 6$ and $r = 3$, enumerate explicitly the number of ways you have of choosing each character in a word. In particular, convince yourself that the number of ways of choosing the last character is equal to $n-r+1$ and not to $n-r$.

We multiply together the number of ways of getting each character to get the total number of permutations (for exactly the same reason that we multiplied p_A and $p_{B \text{ if } A}$ above to find the probability of getting event A followed by event B), so that

$$\begin{aligned} P(n,r) &= n(n-1)(n-2)\cdots(n-r+1) \\ &= \frac{n(n-1)(n-2)\cdots(1)}{(n-r)(n-r-1)\cdots(1)} = \frac{n!}{(n-r)!} \end{aligned} \tag{2}$$

where in the second step I multiplied and divided by the product of the numbers from $n-r$ down to 1 in order, and in the last step I introduced the *factorial* notation.

Exercise: Suppose that you are restricted to the first 4 letters of the alphabet. List out all possible 2-letter words that do not repeat a letter. (Develop an organized system for listing them so that

you do not miss any possibilities!) Check that the number of different permutations you came up with agrees with the prediction of Eq. (2). In contrast, if you are allowed to repeat letters, how many different possible 2-letter words are there and why?

Note in particular that the number of permutations of n objects, all n of which are taken, is $P(n,n) = n!$ as you can see from Eq. (2) because we *define* 0! to be equal to 1.

Now suppose that we are only interested in forming groups of letters and do not care about the order of the letters within each group. For instance, in our previous example of 3-letter selections out of the first 6 letters of the alphabet, suppose that CAB and FAD count separately, but BAD and DAB do not. Such groups are the *combinations* of 6 letters taken 3 at a time. We can compute the number of combinations given the number of permutations by dividing the latter by the number of permutations of the members of a group among themselves, so that

$$C(n,r) = \frac{P(n,r)}{P(r,r)} = \frac{n!}{(n-r)!r!}. \tag{3}$$

Exercise: Suppose I have four quarters, one of which is painted green, one blue, one red, and one yellow. I throw them up in the air and catch them. An example of a sequence is: the green coin is heads, the blue is tails, the red is tails, and the yellow is heads. The count in this case is 2 heads. The total number of sequences of 4 coins that have a count of 2 heads is $C(4,2) = 6$. Can you spell out the other five possible sequences of our 4 coins that give 2 heads? Now suppose that I specify that the first head I caught was the yellow one, the second head I caught was the green one, and that the blue and red coins were tails. This is an example of a permutation of 2 heads out of 4 coins. How many other possible permutations of 2 heads out of 4 coins are there?

An important application of Eq. (3) is to finding the *binomial expansion* of $(x+y)^n$ where n is a positive integer. For example, suppose we wish to find the *coefficient* (i.e., the numerical prefactor) of xy^2 in the expansion of $(x+y)^3$. We proceed as follows. Noting that

$$(x+y)^3 = (x+y)(x+y)(x+y), \tag{4}$$

we can get xy^2 in 3 possible ways: the x in the first parenthetical term can multiply a y from each of the second and third parenthetical terms, or a y from the first parenthetical term can multiply an x from the second and a y from the third parenthetical term, or finally a y from the first and second parenthetical terms can multiply an x from the third parenthetical term. Therefore the *binomial coefficient* of xy^2 in the expansion of $(x+y)^3$ is 3.

Exercise: Fully expand by hand the product in Eq. (4). Simplify your result and show that one term in it is $3xy^2$.

Another way to express what we just said is that we need to select one x from each of three parenthetical terms, for a total of $C(3,1) = 3$ combinations according to Eq. (3). Generalizing this idea, we see that

$$(x+y)^n = \sum_{m=0}^{n} C(n,m)\, x^{n-m} y^m \,. \tag{5}$$

Exercise: Let $n = 3$ and fully expand the sum in Eq. (5), using Eq. (3) to calculate each of the binomial coefficients $C(n,m)$. Show that your answer is identical with the fully expanded product from the preceding exercise.

Equation (5) is such a useful result that it is helpful to have another way to remember it. We can construct what is called *Pascal's triangle* which is the following pyramidal structure of integers,

$$\begin{array}{ccccccccc} & & & & 1 & & & & \\ & & & 1 & & 1 & & & \\ & & 1 & & 2 & & 1 & & \\ & 1 & & 3 & & 3 & & 1 & \\ 1 & & 4 & & 6 & & 4 & & 1 \end{array} \tag{6}$$

where the rows are to be indexed in order from top to bottom by n equaling 0, 1, 2, 3, and 4. We can similarly extend the triangle indefinitely to rows 5, 6, 7, …. The entries in a given row n then correspond to the binomial coefficients for values of m increasing in order from 0 up to n. For example, row 3 (which is 1 3 3 1) tells us that

$$(x+y)^3 = 1x^3y^0 + 3x^2y^1 + 3x^1y^2 + 1x^0y^3 = x^3 + 3x^2y + 3xy^2 + y^3 \,. \tag{7}$$

The triangle is constructed by requiring that any entry equal the sum of the two entries immediately above it (with zeroes implicitly understood to exist to the left of the first 1 and to the right of the last 1 in each row). This allows you to easily extend the triangle up to any row n you need.

Exercise: Extend Pascal's triangle up to $n = 7$. (It is worth noting that the second entry in any row is the row number n. That is, you need *eight* rows in all to reach $n = 7$, because the top row corresponds to $n = 0$.) Use your result to write down $(x+y)^7$ without having to actually multiply $(x+y)$ by itself seven times (which would be a rather tedious process)!

Application to Statistical Mechanics

We will now consider three different ways of putting N balls into n boxes. In statistical mechanics, the "balls" can represent molecules, electrons, photons, or other particles and the

"boxes" correspond to particular states or regions of phase (position–momentum) space. For example, we may be distributing electrons among the allowed energy levels of a (multi-electron) atom. Or we may be determining the distribution of positions and velocities of gas molecules within a chamber.

Let us first consider the case of classical, distinguishable particles. This would apply to an ideal monatomic gas in which we picture atoms as miniature billard balls that obey the rules of *classical* mechanics. That is, if we are given the initial positions and velocities of all N atoms, in principle we can determine the trajectory of every particle and keep track of where it is at any subsequent instant. In this sense, the particles are *distinguishable* and may as well be painted different colors. We say that such particles obey *Maxwell-Boltzmann (MB) statistics*. Suppose we chop phase space up into n discrete bins and wish to calculate the probability p_{MB} of finding a certain distribution of gas atoms among the bins, say N_1 "balls" in the first "box," N_2 in the second, and so on, subject to the restriction that

$$\sum_{i=0}^{n} N_i = N \ . \tag{8}$$

(Some boxes may be empty, while others may contain many balls.) We can determine this probability by dividing the number of different arrangements of the balls that correspond to this particular distribution by the total number of different possible arrangements of the balls. To see how to calculate the first number, consider a specific example. Suppose we have 15 balls and 6 boxes and want to distribute them as {3,1,4,2,3,2} where this notation means we put 3 balls in the first box, 1 in the second box, and so forth. The number of ways we can group the balls in this way is the product

$$\begin{aligned} &C(15,3)C(12,1)C(11,4)C(7,2)C(5,3)C(2,2) \\ &\qquad = \frac{15!}{3!12!}\frac{12!}{1!11!}\frac{11!}{4!7!}\frac{7!}{2!5!}\frac{5!}{3!2!}\frac{2!}{2!0!} = \frac{15!}{3!1!4!2!3!2!} \end{aligned} \tag{9}$$

because we choose 3 out of the original 15 balls to put in the first box, then choose 1 out of the remaining 12 balls to put in the second box, and so on. Equation (3) was used to obtain the second step, which was then simplified to get the third step. Continuing, to calculate the total number of different possible arrangements of the balls, we can put the first ball into any of the 6 boxes, and likewise for the second, third, …, and fifteenth balls. Thus the total number of possible arrangements is

$$6 \times 6 \times \cdots \times 6 = 6^{15} \ . \tag{10}$$

Generalizing the results in Eqs. (9) and (10) and ratioing them, we conclude that

$$p_{\mathrm{MB}} = \frac{N!\,n^{-N}}{\prod_{i=1}^{n} N_i!} \tag{11}$$

where the notation in the denominator refers to the product of factorials $N_1!N_2!\cdots N_n!$ analogous to what appears in the denominator of Eq. (9).

Next we consider the statistics of *indistinguishable* particles behaving according to the rules of *quantum mechanics*. We must assume that the particles can move around and get mixed up; if they are nailed down like the atoms in a solid crystal, then we can distinguish them by their lattice site. Also the particles must be small enough that they behave quantum mechanically. (More precisely, we can combine these two conditions and state that the *de Broglie wavelengths* of the particles must *overlap* significantly during the period of observation. The degree of overlap depends on temperature, which is why many quantum effects such as superconductivity are only seen at low temperatures.) For example, we might be referring to the properties of the electrons in a metal or of the phonons (lattice vibrational normal modes) in an insulator. It turns out that the statistics of such particles depend on whether their *spin* is an integer or a half-integer (specifically half of an *odd* integer—obviously half of an *even* integer is simply another integer). Particles with integer spin (e.g., ^{4}He nuclei which consist of 2 neutrons and 2 protons) are called *bosons* and obey *Bose-Einstein (BE) statistics*; particles with half-integer spin (e.g., individual electrons, protons, or neutrons which have spin $S = \frac{1}{2}$) are called *fermions* and obey *Fermi-Dirac (FD) statistics*. The key difference between fermions and bosons is that only the former obey the *Pauli Exclusion Principle*. That is, two identical fermions cannot occupy the same quantum state simultaneously. (For example, two electrons in an atom cannot have the same four quantum numbers n, l, m_l, and m_s.) In terms of our model of balls and boxes, this means that we can put at most one ball in each box. The balls and boxes are like people and chairs—each seat can either be vacant (*unoccupied*) or *occupied* by one person. With this in mind, let us calculate the probability given say 6 chairs and 4 people that the first 2 chairs will be unoccupied. Clearly we are choosing to occupy 4 chairs out of 6 available, for a total of $C(6,4)$ possible combinations. Hence the probability of any one combination (in our case, that the first two chairs are vacant) is $1/C(6,4) = 1/15$. More generally, if we have N balls and n boxes, the probability of any particular distribution is

$$p_{\mathrm{FD}} = \frac{N!\,(n-N)!}{n!} \tag{12}$$

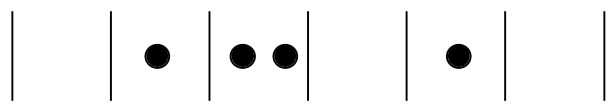
where we are assuming $n \geq N$ if there is to be any allowable distribution at all.

Notice that Eq. (12) presupposes that every distinct combination (or what we might call a *distinguishable arrangement* of the balls) is *equally* likely. Suppose that this remains true for

bosons, for which we are allowed to put any number of balls we like into each box. (In particular, at low enough temperatures all of the particles will "condense" into the ground state, a phenomenon known as *Bose-Einstein condensation* which is responsible for such remarkable properties as superfluidity.) Notice that this equiprobability of arrangements is *not* true for the Maxwell-Boltzmann distribution. For example, suppose that $N = 4$ and $n = 10$. Then according to Eq. (11), the probability of finding 2 particles in each of the first two states is 6 times more likely than the probability of finding all 4 particles in the first state.

Exercise: Suppose that $N = 4$ and $n = 10$. If $N_1 = 4$ and $N_2 = \cdots = N_{10} = 0$, show that $p_{\mathrm{MB}} = 0.01\%$. In contrast, recalculate the probability if $N_1 = N_2 = 2$ and $N_3 = \cdots = N_{10} = 0$.

We see that "bunched" Maxwell-Boltzmann distributions in which a few states are heavily occupied are less likely than "uniform" distributions in which the particles are more evenly spread across the available states. In contrast, for the Bose-Einstein distribution, we wish to ensure that all (distinguishable) distributions are equally probable. (This means the balls and boxes are like friends and dorm rooms, respectively, in which the people prefer to group together than to be found alone.) To figure out how to do this, consider the following picture of 4 balls arranged in 6 boxes.



The edges of the boxes are indicated by vertical lines, with adjacent boxes sharing a common edge, and the balls by dots. Thus this particular distribution corresponds to $N_1 = 0$, $N_2 = 1$, $N_3 = 2$, $N_4 = 0$, $N_5 = 1$, and $N_6 = 0$. We can obtain any other distinguishable arrangement by pushing the common edges leftward or rightward across the balls, leaving the first and last (outermost) edges fixed in place. Any such distribution corresponds to an arrangement of the 5 interior edges and the 4 balls, that is of 9 objects in all. Out of 9 possible positions, we need to choose 5 positions for the edges and 4 positions for the dots. Thus there are $C(9,5)$ or equivalently $C(9,4)$ equally likely arrangements.

Exercise: Use Eq. (3) to prove that $C(i,j) = C(i,i-j)$ for any positive integers *i* and *j* such that $j < i$. What does this tell you about the symmetry of Pascal's triangle?

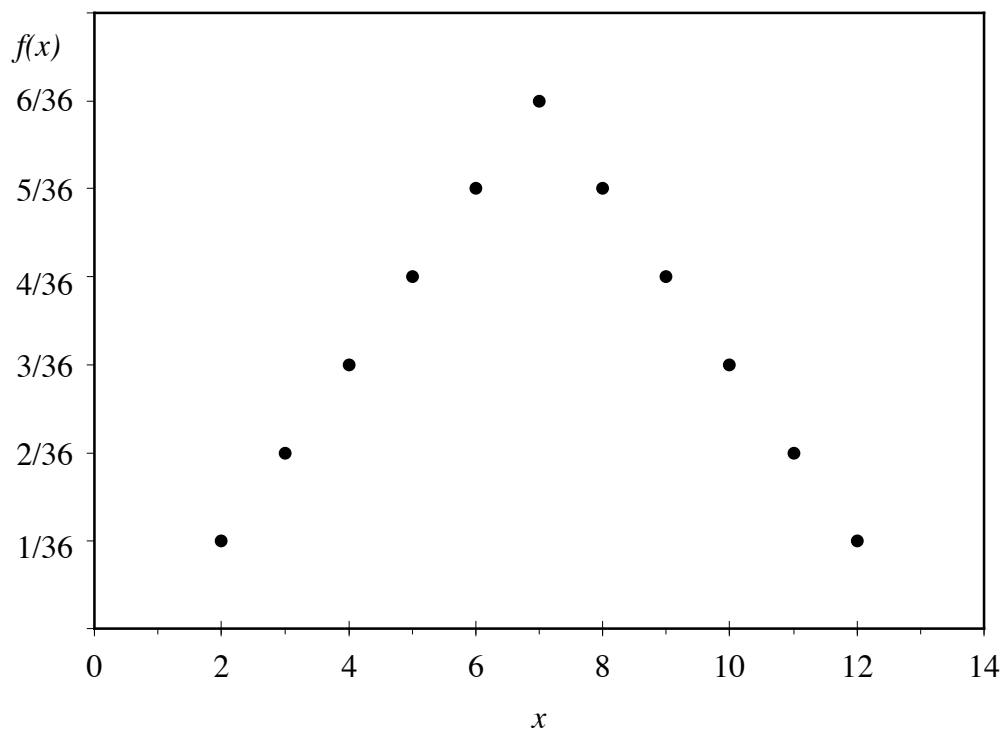
So in general, if we have *N* balls and *n* boxes, the probability of any particular distribution is

$$p_{\mathrm{BE}} = \frac{1}{C(N+n-1,N)} = \frac{N!\,(n-1)!}{(n+N-1)!} \tag{13}$$

which, unlike the Fermi-Dirac case in Eq. (12), is well-defined regardless of whether *n* is bigger or smaller than *N*.

Mean and Standard Deviation

The sample space for a finite probability distribution is the set of values x_i where i is an index running from 1 up to some finite value N. The probability of any particular value x_i is p_i. This probability is clearly some function of x_i, which we call the probability function $f(x_i)$. For example, here is a graph of the probability function $f(x)$ where x is equal to the sum of the numbers on two dice rolled randomly.



For example, $f(9) = 4/36$ because there are 4 distinct sequences of values of two rolls of a die that add up to 9—namely {6,3}, {5,4}, {4,5}, and {3,6}—out of a total of $36 = 6 \times 6$ different possible sequences of two rolls of a die in all.

Now suppose that x represents experimental measurements of the period of a pendulum and that we have made some total number n of distinct measurements of it. If we obtained the value x_i in n_i of the trials, then the *average* or *mean* value of x is defined to be equal to

$$\bar{x} \equiv \frac{1}{n}\sum_i n_i x_i \ . \tag{14}$$

For example, suppose we used a stopwatch with one-tenth of a second accuracy and measured the period to be 1.9 s once, 2.0 s in five trials, 2.1 s six times, and 2.2 s twice. Then the mean value of the period is $(1 \times 1.9\,\text{s} + 5 \times 2.0\,\text{s} + 6 \times 2.1\,\text{s} + 2 \times 2.2\,\text{s}) / (1 + 5 + 6 + 2) = 2.06\,\text{s}$.

Recall from our previous discussion that the ratio n_i / n is an approximation to the value of the probability p_i if we make a large number of trials. Consequently Eq. (14) can be rewritten as

$$\bar{x} \cong \sum_i x_i p_i = \sum_i x_i f(x_i) \,, \tag{15}$$

and this result becomes exact in the limit that we make an infinite number of random trials. Clearly we would report this as our "best" value of the measured quantity, the period of the pendulum in our example. In quantum mechanics, this quantity is often called the *expectation value* and written as $\langle x \rangle$.

We also wish to report some measure of the spread in our measured values. One's first thought might be to calculate the discrepancy between any particular value x_i and the mean value $\bar{x}$, and then average these discrepancies together. On second thought, however, it is not hard to see that this "average discrepancy" is zero.

Exercise: Prove this assertion. That is, show that $\sum n_i (x_i - \bar{x}) = 0$. After you have finished the mathematical proof, explain the result intuitively. How can it be that we take a bunch of nonzero discrepancies and average them together to get zero?

To get a nonzero result, we can first square the discrepancies to ensure that they are all positive, then average them together to get a nonzero result called the *variance*, and finally take its square root to obtain a measure of the spread called the *standard deviation*,

$$\Delta x \equiv \sqrt{\frac{1}{n} \sum_i n_i (x_i - \bar{x})^2} \cong \sqrt{\sum_i (x_i - \bar{x})^2 f(x_i)} \,, \tag{16}$$

where again the second approximation becomes exact in the limit that we make infinite independent measurements. We can see from this definition that the standard deviation is equal to the *root-of-the-mean-of-the-squares* (rms) of the discrepancies.

Exercise: Why couldn't we just define the standard deviation to be equal to the variance? Answer this by returning to the pendulum example above, computing the standard deviation of the measured period T, and reporting your experimental result as $\bar{T} \pm \Delta T$ in seconds. In contrast, what units does the variance have?

Another common notation for the mean and standard deviation is μ and σ, respectively. Their ratio σ / μ is then a measure of the fractional or *relative* error, in contrast to σ which can be called the *absolute* error.

So far our distributions have only taken on a *discrete* (finite) set of possible different values. Let us now consider what happens if a random variable can adopt any value within some continuous range. Note that the range can either be finite (e.g., the position of a cart along a 1-m

track) or infinite (e.g., the location of a free electron in three-dimensional space) in size. In one dimension, we can always choose a finite range to extend from 0 to L by appropriate choice of the origin and direction of the x-axis, whereas an infinite range extends from $-\infty$ to $+\infty$. Keep in mind however that a finite range can be considered to be a subset of an infinite range with $f(x) = 0$ for all $x < 0$ or $x > L$.

To introduce the notion of the probability function $f(x)$ for a continuous distribution, let's start with the simple example of a cart moving along a horizontal frictionless track of length L and elastically rebounding off springs at the two ends of the track. Based on the reasoning developed in the first complete paragraph after Eq. (1), the probability of finding the cart at any exact position x on the track is zero, because there are an infinite number of possible positions between 0 and L, and $1/\infty = 0$. Clearly it makes more sense to instead ask for the probability of finding the cart within some small interval dx of positions centered on x. (For example, suppose $L = 1\text{ m}$ and $x = 50\text{ cm}$. If we want to know or can measure the cart's position to within ±1%, that corresponds to a total interval of $dx = 2\text{ cm}$. What is the probability that the cart is located between the 49 and 51 cm marks along the track?) Since the cart spends fraction dx/L of its time in the interval dx, the probability of finding it in that interval is dx/L.

Now let's replot at the top of the next page the preceding graph of $f(x)$ for the probability of achieving a particular sum when two dice are rolled, but this time include horizontal line segments of length 1 centered on each dot. Observe that the area under a horizontal line segment at any particular value of x is $f(x)\cdot 1 = f(x)$. If we round x off to the nearest integer x_i, we could use this area as the probability of getting that particular sum of the dice. For example, the probability of getting a value for x between 5.5 and 6.5 (i.e., of getting a sum of 6) is 5/36. We can write 5/36 as the area under the histogram from 5.5 to 6.5,

$$p_6 = \int_{5.5}^{6.5} f(x)dx\,. \tag{17}$$

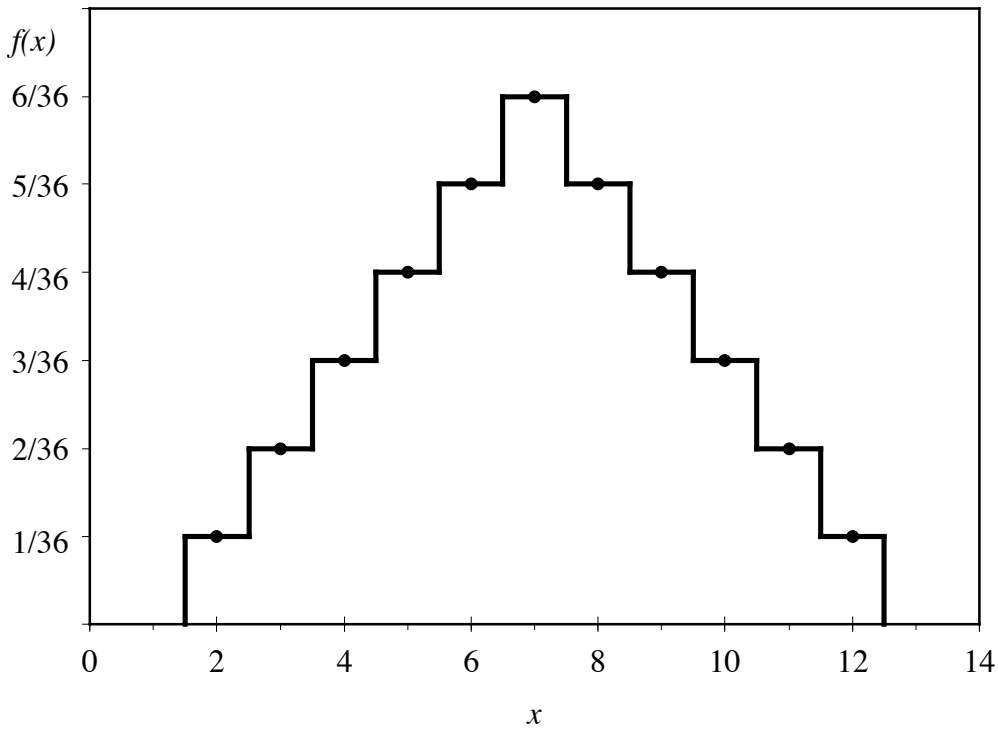
We now *define* the probability function $f(x)$ for a continuous distribution such that the area $f(x)dx$ under a differential interval dx located at x is the probability of finding the result to be between $x - \frac{1}{2}dx$ and $x + \frac{1}{2}dx$ (or more simply between x and $x + dx$). In the example of the cart above, we have $f(x)dx = dx/L$ and therefore $f(x) = 1/L$ (provided that x is between 0 and L). Note that the generalization of Eq. (1) for a continuous distribution becomes

$$\int_0^L f(x)dx = 1 \tag{18a}$$

for a finite range, or more generally

$$\int_{-\infty}^{\infty} f(x)\,dx = 1 \tag{18b}$$

for an infinite range.



In the cart example, the probability function $f(x)$ was a constant (namely $1/L$) over the range of x. Now suppose we tilt the track so that the cart just reaches the top end (with zero speed) and then turns around and slides back down. In that case, the probability of finding the cart in an interval dx along the length of the track is proportional to the time spent in that interval, $f(x)dx \propto dt = dx/\upsilon$ where υ is the speed of that cart at that point. Using conservation of mechanical energy, one can show that $\upsilon = \sqrt{2g\sin\theta(L-x)}$ where θ is the angle of the track. Therefore $f(x)dx \propto dx/\sqrt{L-x}$ after dropping all constants such as g and θ. To turn the proportionality into an equality, we need to *normalize* the right-hand side by dividing it by the constant N equal to

$$N \equiv \int_0^L \frac{dx}{\sqrt{L-x}} = 2\sqrt{L}\ . \tag{19}$$

This normalization procedure ensures that Eq. (18a) will be satisifed. We thus obtain the final result $f(x) = 1\big/2\sqrt{L(L-x)}$. Note that although $f(x)$ becomes infinite at $x = L$, nevertheless

the area under $f(x)$ from 0 to L is finite!

Exercise: Derive the expression for υ given above. (Hint: Note that $y = x\sin\theta$ is the height of the cart above the bottom end of the track.) Also, substitute the expression for $f(x)$ given above into the left-hand side of Eq. (18a) and perform the integral to verify that it is equal to 1, thereby demonstrating that it is properly normalized.

We can now generalize the definition of mean in Eq. (15) to a continuous distribution as

$$\bar{x} = \mu = \int_{-\infty}^{\infty} x\, f(x)\, dx\,, \tag{20}$$

where it is understood for a finite range that $f(x) = 0$ for all x outside of the range. Similarly the variance in Eq. (16) becomes

$$(\Delta x)^2 = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx\,. \tag{21}$$

The mean is sometimes called the *centroid* or *first moment* of the normalized distribution function $f(x)$, while the variance is called the *second moment* (about the mean). These two quantities appear in special guises in mechanics and in electricity & magnetism. Consider the linear mass density $\lambda(x)$ of a rod in units of kg/m. Then the center of mass of the rod is defined as

$$x_{\text{cm}} = \int_{-\infty}^{\infty} x\,\lambda(x)\, dx \Big/ \int_{-\infty}^{\infty} \lambda(x)\, dx\,, \tag{22}$$

which we can think of as an average value of position x along the rod *weighted* by the density of the rod at that position. Notice that the denominator in Eq. (22) equals the total mass M (or *zeroth moment*) of the rod and plays the same role as the normalization constant N in Eq. (19). If we define a *normalized* mass density by $f(x) = \lambda(x)/M$, then Eq. (20) is equal to the center of mass. Using the same normalized mass density in Eq. (21) gives the *moment of inertia* of the rod about its center of mass. In similar fashion, if $f(x)$ is the linear charge density along a rod in units of C/m, then we can relate Eqs. (20) and (21) to the electric dipole and quadrupole moments.

Binomial, Gaussian, and Poisson Distributions

Suppose we repeat n independent trials of some process (such as rolling a die) and that there are two possible outcomes of each trial: success with probability p, and failure with probability $q = 1 - p$. Then the probability of getting exactly x successes in the n trials is

$$f(x) = C(n,x)p^x q^{n-x} \tag{23}$$

and is called the *binomial probability function* (or *distribution*). This formula is the product of $C(n,x)$, the number of combinations of x successes out of n trials, and $p^x q^{n-x}$, the probability that any particular combination will consist of x successes and $n-x$ failures. Note that if we sum Eq. (23) over all possible values of x we get

$$\sum_{x=0}^{n} f(x) = \sum_{x=0}^{n} C(n,x)p^x q^{n-x} = (p+q)^n = 1. \tag{24}$$

The second equality follows from the *binomial expansion* in Eq. (5), and the final equality follows from the fact that $q = 1-p$ and is an expression of the *normalization* of $f(x)$ analogous to Eq. (1).

If we make histogram plots of the binomial distribution for various values of n and p, one can make a number of interesting observations about them. First, the area under the curve is always 1 because of Eq. (24). Second, $f(x)$ peaks at the most probable value $x \approx np$ (as proven in problem 9); this result is consistent with the operational definition of probability $p \approx x/n$ (for large n) discussed on the first page of this handout. For example, if $p = 1/2$ (such as if we are flipping a coin n times and counting heads), then the most probable count is that we get $\frac{1}{2}n$ heads if n is even or $\frac{1}{2}(n \pm 1)$ heads if n is odd. For this value of p, the histogram is symmetric about $x = n/2$. On the other hand, if $p > 1/2$ then the plot is asymmetric and peaks for large values of x (and likewise it peaks at small values of x if $p < 1/2$). Third and finally, the mean and standard deviation for the binomial distribution are $\mu = np$ and $\sigma = \sqrt{npq}$, respectively (as proven in problem 10). Therefore the peak becomes wider (and correspondingly shorter in height because the area under the curve is fixed at 1) as n increases for a given value of p. In particular, if $p = 1/2$ then $\mu = n/2$ (which lies within 1 of the most probable value) and $\sigma^2 = \mu/2$.

We can derive an approximation to the binomial distribution for the important special case that the probability p of success is small and the number n of trials is large (i.e., $p << 1$ and therefore $\mu << n$). Start by writing Eq. (23) as the product of five terms,

$$f(x) = \frac{1}{x!}\frac{n!}{(n-x)!}p^x(1-p)^n(1-p)^{-x}. \tag{25}$$

The second term is equal to $n(n-1)(n-2)\cdots(n-x+1)$. Since p is small, we only get a significant value for $f(x)$ when x is small compared to n and in this limit, the second term is approximately $n \times n \times n \cdots \times n = n^x$. Combining that result with the third term we get $(np)^x = \mu^x$. The fifth term is approximately equal to $1 + px \approx 1$ for small p and x. Finally, substituting $p = \mu/n$, the fourth term can be written in the limit of large n as

$$\lim_{n\to\infty}\left[1+\frac{(-\mu)}{n}\right]^n = e^{-\mu}. \tag{26}$$

Exercise: The final result in Eq. (26) uses the fact that

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x.$$

To prove this relation, take the logarithm of the limit and interchange the order (which we can do because the logarithm is a continuous function) to get

$$\lim_{n\to\infty}\ln\left(1+\frac{x}{n}\right)^n = \lim_{n\to\infty}\frac{\ln(1+x/n)}{1/n}.$$

Next use l'Hôpital's rule (make sure you differentiate with respect to n not x) to show that this limit is equal to x. Taking the anti-logarithm now completes the proof.

Collecting together all of our results, we get

$$f(x) = \frac{\mu^x}{x!}e^{-\mu} \tag{27}$$

which is called the *Poisson distribution*. The mean of the distribution is μ, while the standard deviation is $\sqrt{npq} \to \sqrt{\mu}$ since $q = 1 - p \to 1$ for small p. Notice that the Poisson distribution is therefore characterized by a single parameter (μ), unlike the binomial distribution which is characterized by two (n and p).

Exercise: Verify that the Poisson distribution is normalized. That is, check that

$$\sum_{x=0}^{\infty} f(x) = 1 \tag{28}$$

which is the generalization of Eq. (24) to a large number of trials, $n \to \infty$.

The Poisson distribution is particularly useful in physics for describing the probability $p_x(t) = f(x)$ of observing exactly x radioactive decays of a sample during a time interval t short compared to the mean lifetime τ (or to the half-life $t_{1/2} = \tau \ln 2$) where $\mu = t/\tau$ is the average number of decays observed in the time interval t. For example, if we measure 1800 counts from a source in 10 hours, then the average number of counts μ per minute is 3. Consequently the probability of collecting x counts in a minute is $3^x e^{-3}/x!$ and so we would expect to observe 6 counts in about 5% out of 600 of the minute-long intervals for which the counting experiment was performed. In this particular numerical example, μ is small (on the order of 1) and a plot of the Poisson distribution results in an asymmetric peak (near 3) having a tail extending to large

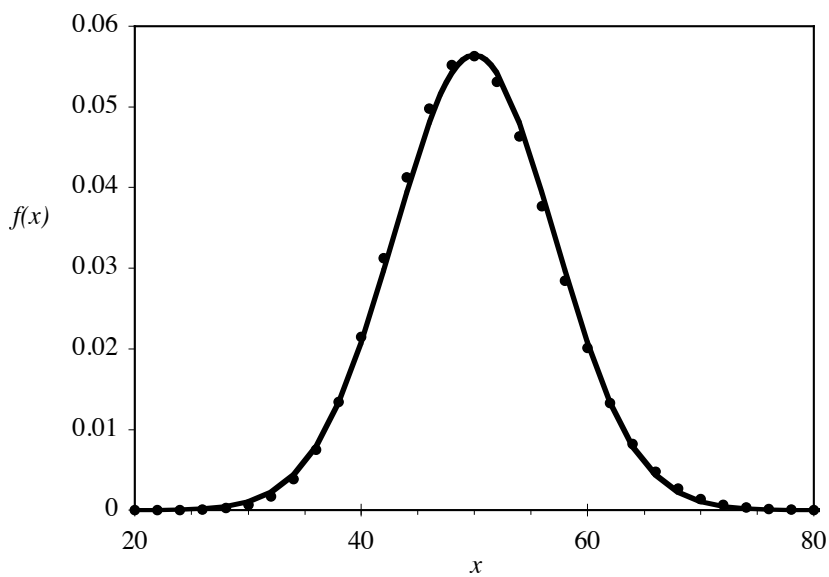
values of x. On the other hand, if μ is large (compared to 1) then the peak has a bell shape. This observation suggests that we attempt to fit it with a continuous *Gaussian distribution*,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{29}$$

also called a *normal* distribution. Here μ is the mean (and thus the location of the peak since it is symmetric), σ is the standard deviation (as proven in problem 11), and the numerical prefactor in front of the exponential has been chosen to normalize the distribution in accordance with Eq. (18b).

Exercise: Verify that the Gaussian distribution is normalized. To do so, substitute Eq. (29) into (18b) and change variables from x to $X \equiv (x-\mu)/\sigma\sqrt{2}$; call the result I. Next change the dummy variable from X to Y to get I written in a different way. Now form the product of these two integrals and pull both integral signs out front. Interpret $dXdY$ as the area element in a plane and change to polar coordinates for which the area element is $rdrd\phi$. Rewrite the integrand in terms of r and ϕ rather than in terms of X and Y. Also change the limits of integration so that they range over the entire plane in polar coordinates. You can now separately perform the radial and angular integrals. The final answer should be 1.

In order to fit Eq. (29) to a Poisson distribution, we require $\sigma = \sqrt{\mu}$. The resulting Gaussian is plotted as the solid curve in the following graph for the case of $\mu = 50$. For comparison, Eq. (27) is plotted as the dots for the same value of μ.



We see from this figure that a Poisson (and therefore also a binomial) distribution are well approximated by a Gaussian if both μ and n are large (compared to 1).

In the above figure, we set $\sigma = \sqrt{50} \approx 7.07$ which is one measure of the width of the Gaussian. There are several other physically useful measures of the width. One is the half-width-at-half-maximum (HWHM) Δx, which is the x distance that you have to move away from μ in order for the ordinate (i.e., the y coordinate) to fall to half of its value. At $x = \mu$, Eq. (29) has the peak value $f(\mu) = 1/\sigma\sqrt{2\pi}$. For example, $f(\mu) \approx 0.056$ in the preceding figure. We seek the value of Δx for which $f(\mu + \Delta x) = \frac{1}{2} f(\mu)$. Using Eq. (29) we find that the HWHM is

$$\Delta x = \sigma\sqrt{2\ln 2}\,. \tag{30}$$

For example, $\Delta x \approx 8.32$ for the above example. That is, $f(x)$ falls from its peak value of 0.056 to half of its peak value (0.028) when x is changed from the mean (50) to the sum of the mean and HWHM (58.32). Note that the HWHM is a little larger than the standard deviation. If you double the HWHM, you get the full-width-at-half-maximum (FWHM). Other measures of the width for specialized applications (such as in optics) are the x distances you have to move away from μ in order for the ordinate to fall to $1/e$ or $1/e^2$ of its peak value. Each of these definitions of width is some numerical multiple of σ (because σ carries units of width) where the numerical factor is usually near 1 in value. For example, in Eq. (30) the numerical multiple is 1.177.

Exercise: If x is in cm, what are the units of μ, of σ^2, and of $f(x)$ in Eq. (29)?

Experimental Error Analysis

In science, the term *error* refers not to a mistake but to a difference (or deviation or discrepancy) between a measured value and the "true" value. A synonym for error is *uncertainty*. We can divide errors into two possible sources: systematic and random. *Systematic* errors arise from miscalibrations in equipment and from the bias of the observers. For example, if a metal ruler has been calibrated at room temperature but is then used outdoors in the winter, the ruler's length will thermally contract and the measured lengths will be slightly larger than they should be. (Why are the measured values systematically larger instead of smaller than the actual lengths?) Other examples are failure to correct for parallax (angle of observation) when using a meter stick, failure to properly zero (or tare) a scale, and failure to account for human reaction time when starting and stopping a stopwatch. It is the responsibililty of a careful experimenter to consider and correct for systematic errors. In the preceding examples, the thermal contraction of a ruler of known composition can be calculated, one's eye can be moved to perpendicular incidence when using a meter stick, a known weight can be used to test a scale's accuracy, and a longer time interval can be measured with a stopwatch to minimize the effects of reaction time. In what follows, it is assumed that all systematic errors have been reduced to negligible levels.

The second category of uncertainties is *random* error which arises from inherent statistical fluctations in the measured object, the measuring instrument, or the nature of the experiment

itself. For example, if we measure the length of a box at different places along its width even with a "perfect" ruler, there will be some variation due to slight manufacturing nonparallelism of the sides and surface roughness. In addition, the ruler itself is imperfect and can only be read to perhaps a quarter of the smallest divisions on it (millimeters say), leading to some intrinsic uncertainty. As another example, if we repeatedly measure the period of a pendulum released from a particular angle, we will find some range of values because of slight differences in how the bob is released, in air drag, in the elasticity of the string, and the like. (Some of these effects might be construed as systematic and could be reduced by enclosing the pendulum in a box to reduce air currents, constructing a more reproducible bob release system, and so on. Nevertheless, there will always remain some fluctuations in any experimental system. Consider for instance the measurement of the half-life of a particular radioactive element.)

For a given experimental setup, the random error cannot be eliminated. Instead our goal is to estimate how big it is likely to be, so that we can report that a measured value is within some fractional or absolute amount of the true value. A subtle but important issue now arises. There is no *single* true value—instead, the infinite set of all possible actual values (such as the length of a box or the period of a pendulum) are distributed in some range according to what is called the *parent distribution* having some particular shape, mean, standard deviation, and so forth. On the other hand, what we measure are a finite number (say n) of experimental values that determine what is called the *sample distribution*. Our best estimate of the parent mean if we measure x_i in trial i is

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{31}$$

in agreement with Eq. (14). Two alternatives quantities that are often quoted in statistical discussions are the *median* (the value of x such that there are an equal number of larger and smaller measurements) and the *mode* (the most frequently measured value). Our best estimate of the parent variance is

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 \ . \tag{32}$$

This formula is similar to Eq. (16) except that we have replaced n by $n-1$ in the denominator. A rigorous mathematical derivation of this replacement is lengthy. It suffices for our purposes to point out that $n-1$ represents the number of degrees of freedom remaining after we have used up one to compute the mean from Eq. (31). In particular, if we made only one observation, there would be no way to estimate the variance, in agreement with the fact that Eq. (32) then gives the indefinite result 0/0 (whereas it would not if there were an n in the denominator). In any case, if

we make a large number of measurements, then the difference between n and $n-1$ is small and not worth a long analysis.

For completeness in case you encounter it elsewhere, one can also derive the standard deviation in our estimate of the mean and prove that it is equal to $\sigma/\sqrt{n}$. This quantity is often called the *standard error*. In other words, the actual parent mean is unlikely to differ from our estimate μ by more than a few multiples of the standard error. As expected, our estimate approaches the actual value in the limit of an increasing number n of measurements. Note however that what we typically report in an experiment is not the standard error but instead our estimate σ of the parent standard deviation, because we want to know how much variation there can be in the x values. To put it another way, σ is the standard deviation of a single measurement.

Next let's take up the matter of *error propagation*. Suppose that we make measurements of three quantities (x,y,z) and wish to combine those into one overall formula $f(x,y,z)$. For example, x might be length, y the width, and z the height of a box of volume $f = xyz$. Specifically let (x,y,z) be our best (i.e., mean) values of each measurement, as separately computed using Eq. (31), and let's denote the standard deviation in each calculated from Eq. (32) as (dx,dy,dz). Here we are using differential notation, assuming that the fractional error in each (e.g., dx/x) is small (i.e., no more than a few percent). If the errors are not small, then we should instead denote them as $(\Delta x,\Delta y,\Delta z)$ and their treatment will be briefly discussed later.

Recall that a first-order Taylor expansion of f about the mean values (x,y,z) is

$$df = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial z}dz\,. \tag{33}$$

In words, this equation states that the change in f is equal to the sum of the changes in f when each independent variable is separately varied (with the other two variables held fixed, thereby explaining the need for *partial* derivatives). Equation (33) enables us to compute the standard deviation of the new quantity f; the only proviso is that we need to take the absolute value of each term in Eq. (33) because errors must add. In principle it is the only error propagation formula that we need. In practice, however, it is faster to perform computations by working out Eq. (33) for various special cases:

Product Rule—If $f = xyz$ then Eq. (33) becomes

$$df = yzdx + xzdy + xydz \quad \Rightarrow \quad \frac{df}{f} = \frac{dx}{x} + \frac{dy}{y} + \frac{dz}{z}\,. \tag{34}$$

The final result says that we can find the fractional error in f by summing up the fractional errors in the variables. This rule also holds for quotients. For example,

if we measure the distance traveled by a cart $x = 31.2 \pm 1.1$ cm in a time of $t = 3.45 \pm 0.12$ s, then our best estimate of the average speed of the cart is $\upsilon = (31.2 \text{ cm}) / (3.45 \text{ s}) = 9.04$ cm/s with a percent error of $100(1.1/31.2 + 0.12/3.45) = 7.0\%$.

Power Rule—If we have a term like x^n in the formula for *f*, then it contributes a fractional error of $n\,dx/x$. This result is consistent with the product rule when *n* is an integer; for instance if $n = 3$ and we identify *y* and *z* with *x* then Eq. (34) becomes $df/f = 3dx/x$. For example, the period of a pendulum is given by the formula $T = 2\pi\sqrt{L/g}$. Suppose we know the length of the string to within 1% and the gravitational field to 0.1% (i.e. to within ±0.01 m/s^2), then the percent error in the period is 0.55%.

Sum Rule— If $f = x + y + z$ then Eq. (33) becomes

$$df = dx + dy + dz\,. \tag{35}$$

This result states that the absolute error in *f* is the sum of the absolute errors in the variables. The rule also holds for differences. For example, if we measure the length and width of a rectangular tabletop to be $1.50 \text{ m} \pm 3.5\%$ and $1.10 \text{ m} \pm 2.5\%$, respectively, then its perimeter is 5.20 ± 0.09 m.

Note that these rules change if we are propagating standard errors rather than parent errors. Specifically, one then needs to *square* each of the terms on the right-hand side of Eq. (33) if the variables are uncorrelated, as is proven in statistics textbooks. (Taking squares also avoids the issue of having to take absolute values of terms.)

Equation (33) becomes increasingly inaccurate as the fractional errors in the variables increase. An alternative method of propagating errors is as follows. First compute *f* using the mean values of the variables and call the result f_{ave}. Next choose the values of the variables within their range of variation (for example between $x - \Delta x$ and $x + \Delta x$) to find the value of *f* (call it f_{dis}) that maximizes the discrepancy between f_{dis} and f_{ave}. We can then compute the error in *f* as $\Delta f = |f_{\max} - f_{\text{dis}}|$.

Exercise: As an example, suppose that $f(A,\theta) = A\cos(\theta)$ where $A = 2.3 \text{ m} \pm 25\%$ and $\theta = 0.86 \text{ rad} \pm 34\%$. Compute Δf and compare it to *df* calculated from Eq. (33) remembering to take the absolute value of each term. (By the way, recall that the derivative of $\cos\theta$ is equal to $-\sin\theta$ only if θ is in radians!) The answers are $\Delta f = 0.92$ m and $df = 0.88$ m, showing that even for fairly sizeable percent errors, Eq. (33) works pretty well.

TOOLS OF THE TRADE

Stirling's Approximation

Many statistical formulae contain factorials in them. Examples in the present Handout are Eqs. (3) and (27). Often we need to evaluate those factorials for large values of their argument. For instance, if Eq. (13) is applied to a gas, we may have a mole of particles, corresponding to Avogadro's number (6.022×10^{23}). Stirling's approximation is a useful approximation to $n!$ when n is a large integer. It can be written as a product of terms,

$$n! = n^n e^{-n} \chi(n) . \tag{36}$$

The third term is the divergent infinite series

$$\chi(n) = \sqrt{2\pi n}\left(1 + \frac{1}{12n} + \frac{1}{288n^2} + \cdots\right). \tag{37}$$

However it is not hard to see that $\chi(n)/n^n e^{-n} \to 0$ as $n \to \infty$.

Exercise: Prove this assertion as follows. Start by noting that $\chi(n) \approx \sqrt{2\pi n}$ in the limit of large n. Next rewrite $n^n e^{-n}$ as $e^{n(\ln n - 1)} \approx e^{n \ln n}$ in the same limit. Now complete the proof using l'Hôpital's rule.

Therefore if we take the logarithm of Eq. (36), we can drop the third term in comparison to the other two terms giving an asymptotic expression for large n,

$$\ln(n!) \approx n \ln n - n . \tag{38}$$

Here is a quick derivation of Eq. (38),

$$\ln(n!) = \sum_{m=1}^{n} \ln m \approx \int_1^n \ln x \, dx \approx n \ln n - n . \tag{39}$$

(You should review this derivation carefully and make sure you understand each step in it.) As a somewhat contrived example of the application of Stirling's formula we can compute the following limit,

$$\lim_{n \to \infty} \frac{(n!)^{1/n}}{n} = \frac{(n^n e^{-n})^{1/n}}{n} = \frac{1}{e} . \tag{40}$$

Note that $\chi^{1/n} \to 1$ as $n \to \infty$. For example try evaluating $(2\pi n)^{1/2n}$ on a calculator for $n = 1000$. Therefore the final answer in this example is exact.

Error Function

There are several slightly different definitions of this function. We will adopt the one that is most useful for our purposes,

$$\mathrm{erf}(z) \equiv \frac{1}{\sqrt{2\pi}} \int_{-z}^{z} e^{-t^2/2} dt \ . \tag{41}$$

The integrand can be recognized as a Gaussian expressed in terms of the dimensionless deviation $t \equiv (x-\mu)/\sigma$ and the prefactor of $1/\sqrt{2\pi}$ normalizes the Gaussian so that $\mathrm{erf}(\infty)$, which is the complete area under the bell curve, is equal to 1. In contrast, $\mathrm{erf}(1)$ represents the area within ±1 standard deviation of the peak, $\mathrm{erf}(2)$ represents the area within ±2 standard deviations of the peak, and so forth. These values are often used when "grading on a curve" for example. Here are a few selected values of Eq. (41), as computed numerically:

z	$\mathrm{erf}(z)$
0	0
1	0.683
2	0.954
3	0.997

For example, the second entry tells us that the probability that a measurement will lie within one standard deviation of the true value (i.e., the parent mean) is 68%. If you wish to increase your confidence interval to near 100%, you can instead use an error range of $\pm 3\sigma$.

Problems

1. Prove that $p_A p_{B \text{ if } A} = p_B p_{A \text{ if } B}$, i.e., the probability that event A occurs followed by event B is the same as the probability that the two events occur in the opposite order. Do <u>not</u> assume the two events are independent. (The proof is trivial if they are!)

2. Prove that the probability that either *or* both of events A and B will occur is $p_A + p_B - p_A p_{B \text{ if } A}$. Now suppose that Alice and Bob are independently working on a problem and that Alice has a 75% chance of solving it, while Bob only has a 50% probability. What is the probability that at least one of them will solve it? (Answer: 87.5%.)

3. You are dealt two cards from a shuffled deck. What is the probability of getting two aces? If you know that one is an ace, what is the probability that the other is an ace? If you know that one is the ace of spades, what is the probability that the other is an ace? Each of these questions has a different answer, clearly demonstrating that the more information you have, the more your odds go up.

4. Show that in any random group of 23 people, the probability is greater than 50% that two of them have the same birthday. (Do this by calculating the exact probability.)

5. If $n >> N$, show that $p_{\text{FD}} \approx N! n^{-N} \approx p_{\text{BE}}$. <u>Hint</u>: Argue in each case that the inverse ratio of the factorials involving n is approximately equal to the product of N numbers each of which is approximately equal to n.

6. A game of chance involves throwing 3 balls into 5 cups numbered 1 through 5. Each cup can only hold one ball. Any balls which do not initially land in an available cup roll around the tops of the cups until they fall into an empty one. It costs \$1 each time you play, but you win 10¢ times the sum of the numbers on the cups that your balls fall into. If you play the game 100 times, how much money would you expect to win or lose in all? (Answer: lose \$10.)

7. Use the first equality in Eq. (16) to prove that $(\Delta x)^2 = \overline{x^2} - \overline{x}^2$. In words, this result says that the variance in x equals the difference between the mean of the squares of x and the square of the mean of x. In practice it is often easier to calculate the standard deviation using this relation instead of Eq. (16). [Note that parentheses are required because the expression Δx^2 is ambiguous: it could mean either $(\Delta x)^2$ or $\Delta(x^2)$. A similar ambiguity in principle is present in the expression dx^2, but by convention it is defined to mean $(dx)(dx)$, because if $d(x^2)$ were intended one would instead write $2x\,dx$.] Be careful: Do not erroneously assume that $(a-b)^2 = a^2 - b^2$ in your proof!

8. According to quantum mechanics, the probability of finding the electron in a volume element dV is proportional to $e^{-2r/a}dV$ for the ground state of a hydrogen atom. Write dV in spherical coordinates and integrate over the full ranges of the two angular variables. Now find the radial electron density $f(r)$ by normalizing this probability function over the radial range. Also prove that the most probable value of r is a (i.e., the value of r which maximizes f) and sketch $f(r)$ by hand over its complete range from 0 to ∞. (Note: a is called the Bohr radius and is equal to 0.53 Å.)

9. We wish to prove that the most probable value of x for the binomial distribution is the integer(s) nearest in value to np. To do so, let's temporarily allow x to take on nonintegral values and find x and $x+1$ which symmetrically bracket the peak of the binomial distribution in the sense that $f(x) = f(x+1)$. Substitute Eq. (23) into both sides of this equality and use the fact that $q = 1 - p$ to solve for x. Then complete the proof appropriately.

10. (a) We want to prove that $\mu = np$ for the binomial distribution. Start by substituting Eq. (23) into Eq. (15) where you can drop the subscripts i and instead sum x from 0 up to n. Now substitute $y = x - 1$ and $m = n - 1$, factor $(m+1)p$ out of the sum, and use the normalization condition from Eq. (24) to complete the proof.
(b) Use similar tricks as in part (a) to prove that $\sigma^2 = npq = np - np^2$. <u>Hint</u>: First use the result of problem 7 to simplify Eq. (16).

11. Substitute Eq. (29) into the integral on the right-hand side of Eq. (21) and perform the integration to verify that it is equal to σ^2, thereby proving that σ in Eq. (29) defines the standard deviation. <u>Hint</u>: Define the integral

$$I \equiv \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-a(x-\mu)^2/2\sigma^2} dx ,$$

change variables from x to $X \equiv \sqrt{a}(x-\mu)/\sigma\sqrt{2}$, and then use normalization to show that $I = a^{-1/2}$. Now use parameter calculus by differentiating I with respect to a, multiplying the result by $-2\sigma^2$, and then setting a equal to 1. (To complete the proof you need to do this twice, once for I written as the above integral, and secondly when I is evaluated as $a^{-1/2}$. If you do not understand this logic, consult the handouts on integration.)

12. Derive Eq. (30).

13. Derive the exponential rule for error propagation. That is if we are computing x in terms of the variable t according to $x(t) = Ae^{t/\tau}$ where A and τ are constants, then $dx/x = dt/\tau$. (This looks like the rule for products except that we have τ not t in the denominator on the right-hand side.)

14. Consider a set of n two-level spins (i.e., they can either be in the spin-up or spin-down state) which are fixed at particular (distinguishable) lattice positions. Suppose that N of them are in the up state (and the remaining $n-N$ of them in the down state). The number of ways that we can choose N out of n is

$$\Omega = \frac{n!}{(n-N)!N!}$$

from Eq. (3) and in the present context is called the *multiplicity*. The Boltzmann entropy is defined in terms of the multiplicity as $S = k\ln\Omega$ where $k = 1.38\times10^{-23}$ J/K is called the Boltzmann constant. Use Eq. (38) to prove that S is a maximum when half of the spins are in each state.

15. Show that $\mathrm{erf}(x-ix) = (1-i)\sqrt{2/\pi}\left[C(x)+iS(x)\right]$ where

$$C(x) \equiv \int_0^x \cos(u^2)\,du \quad \text{and} \quad S(x) \equiv \int_0^x \sin(u^2)\,du$$

are called the *Fresnel integrals* and are relevant to diffraction theory in optics. <u>Hint</u>: Make the substitutions $t = (1-i)u$ and $x = z/(1-i)$ in Eq. (41).

References

P.R. Bevington, *Data Reduction and Error Analysis for the Physical Sciences* (McGraw-Hill, New York NY, 1969).

M.L. Boas, *Mathematical Methods in the Physical Sciences*, 3rd ed. (Wiley, Hoboken NJ, 2006).