1
Tools of the trade

> It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.
>
> —*Pierre-Simon Laplace*[1]

1.1 Probability: The calculus of uncertainty

All measurements and observations, forecasts and inferences, are subject to uncertainty. These uncertainties reflect a lack of precise knowledge arising from the limitations of one's time, which restricts the amount of data that can be collected, or instrumentation, which determines the resolution with which signals or information can be acquired, or the fundamental laws of nature, which give rise to intrinsically random processes whose exact outcomes cannot be predicted irrespective of the apparatus and observation time. Although a well-ordered world governed by deterministic laws with no uncertainties may seem desirable at times, such a world will never be – and, in any event, would make for a rather dull place indeed.

To deal with the vagaries of nature one ordinarily must turn to the principles of mathematics bearing on probability and statistics. I will make no attempt to define probability. For one thing, innocuous as the subject may sound, it has spawned two schools of thought whose members have gone after one another (in a manner of speaking) like Crips and Bloods. So, from a practical standpoint, I would rather not begin a book with remarks likely to inflame any group of readers. Second, and more to the point, probability is a sufficiently basic concept that, in trying to capture its meaning in a few words, one ends up using tautological expressions like "chance" or "odds" or "likelihood" that do not really explain anything. The latter term, in fact, is not even a synonym, but is quite distinct from probability as will become apparent later when we encounter Bayes' theorem or make use of the method of maximum likelihood.

[1] Quoted by Mark Kac, "Probability" in *The Mathematical Sciences* (MIT Press, Cambridge, 1969) 239.

Let it suffice, therefore, to say that, if you are reading this book, you are already familiar with the basic idea of probability in at least two contexts.

(a) The first is as the relative frequency of occurrence of an event. Suppose the sample space – i.e. list of all possible outcomes – of some process comprises events A, B, C whose frequencies of occurrence in $N = 100$ observations are respectively $N_A = 20$, $N_B = 50$, and $N_C = 30$. (The total number must sum to N.) Then, assuming a random process generated these events, one can estimate the probability of event A by the ratio $P(A) = N_A/N = 1/5$, with corresponding expressions for the other events. We read this as one chance in five or a probability of 20%.

(b) The second is as a statement of the plausibility of occurrence of an event. Thus, given meteorological data such as the current temperature, humidity, cloud cover, wind speed and direction, etc., a meteorologist might pronounce a 40% chance of rain for tomorrow. Tomorrow's weather occurs but once; one cannot replay it one hundred times and construct a table of outcomes and frequencies. The probability estimate relies in part on prior knowledge of the occurrences of similar past weather patterns.

The two senses of probability reflect the two schools of thought, referred to usually as "frequentist" and "Bayesian". There are subtle issues connected with both understandings of probability. In the frequentist case (a), for example, a more complete and accurate definition of probability would have N approach infinity, which is no problem for a mathematician, but would pose a crushing burden on an experimental physicist. The Bayesian case (b) avoids resorting to multiple hypothetical replications of an experiment in order to deduce the desired probabilities for a particular experiment, but the method seems to entail a hunch or guess dependent on the analyst's prior knowledge. Since different analysts may have different states of knowledge, the subjectivity of a Bayesian-derived estimate of probability appears to clash with a general expectation that probability should be a well-defined mathematical quantity. (One would hesitate to use calculus if he thought the value of an integral depended on who calculated it.)

At this point I will simply state that both approaches to the calculation of probability are employed in the sciences (and elsewhere); both are mathematically justifiable; both often lead to the same or comparable results in "straightforward" cases. For all the philosophical differences between the two approaches, it may be argued that the frequentist deduction of probability is actually a special case of the Bayesian method. Thus, when the two methods lead to significantly divergent outcomes, the underlying cause (if all calculations were executed correctly) arises from different underlying assumptions regarding the process or system under scrutiny. With that conclusion for the moment, let us move on.

1.2 Rules of engagement

Although philosophical differences may persist regarding the estimation or inference of probabilities, there is no disagreement over the mathematical rules for combining probabilities once they are known. Suppose A and B are two independent events with respective probabilities $P(A)$ and $P(B)$. Then

(a) the probability that A and B both occur is

$$P(AB) = P(A)P(B);$$

(b) the probability that A or B occurs is

$$P(A + B) = P(A) + P(B).$$

Note: the simultaneous occurrence of events is expressed symbolically by multiplication (AB); the exclusive occurrence of events is expressed symbolically by addition ($A + B$).

If A and B are not necessarily independent, one might want to know what is the probability of A occurring, given that B has occurred. This is the conditional probability of A given B, written as $P(A|B)$ and defined by the relation

$$P(A|B) \equiv P(AB)/P(B). \tag{1.2.1}$$

From a frequentist point of view, the foregoing expression may be interpreted as the ratio (theoretically, in the limit of an infinitely large number of trials; practically, for a "reasonably" large number of trials) of the number of events in which A and B occur together to the number of events in which B occurred irrespective of the occurrence of A.

It is common symbolism to represent the *non*-occurrence of an event by an overbar; thus $\overline{A}$ represents all outcomes that do not include event A. From the foregoing considerations, therefore, we can succinctly express two fundamental rules of conditional probability:

$$\textit{inclusivity} \qquad P(A|B) + P\left(\overline{A}|B\right) = 1, \tag{1.2.2}$$

$$\textit{Bayes' theorem} \qquad P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}. \tag{1.2.3}$$

The first rule (1.2.2) signifies that, after B occurs, A either occurs or it does not; those are the two mutually exclusive outcomes that exhaust all possibilities. Note that it is *not* generally true that $P(A|B) + P\left(A|\overline{B}\right) = 1$. Rather, given $P(A|B)$ and Bayes' theorem, it is demonstrable that

$$P(A|B) + P\left(A|\overline{B}\right) = \frac{P(A) + P(A|B) - 2P(AB)}{1 - P(B)}, \tag{1.2.4}$$

as shown in an appendix.

The second rule (1.2.3), although called Bayes' theorem, is a logical consequence of the laws of probability accepted by frequentists and Bayesians alike. It is regularly used in the sciences to relate $P(H|D)$, the probability of a particular hypothesis or model, given known data, to $P(D|H)$, the more readily calculable probability that a process of interest produces the known data, given the adoption of a particular hypothesis. In this way, Bayes' theorem is the basis for scientific inference, used to test or compare different explanations of some phenomenon.

The parts of Eq. (1.2.3), relabeled as

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}, \tag{1.2.5}$$

are traditionally identified as follows. $P(H)$ is the "prior" probability; it is what one believes about hypothesis H before doing an experiment or making observations to acquire more information. $P(D|H)$ is the "likelihood" function of the hypothesis H. $P(H|D)$ is the "posterior" probability. The flow of terms from right to left is a mathematical representation of how science progresses. Thus, by doing another experiment to acquire more data – let us refer to the outcomes of the two experiments as D_1 and D_2 – one obtains the chain of inferences

$$P(H|D_2D_1) = \frac{P(D_2|D_1H)P(D_1|H)P(H)}{P(D_2D_1)} \tag{1.2.6}$$

with the new posterior on the left and the sequential acquisition of information shown on the right.

As an example, consider the problem of inferring whether a coin is two-headed (i.e. biased) or fair without being able to examine it – i.e. to decide only by means of the outcomes of tosses. Before any experiment is done, it is reasonable to assign a probability of ½ to both hypotheses: (a) H_0, the coin is fair; (b) H_1, the coin is biased. Thus

$$\text{ratio of priors:} \qquad \frac{P(H_0)}{P(H_1)} = 1.$$

Suppose the outcome of the first toss is a head h. Then the posterior relative probability becomes

$$\text{first toss:} \qquad \frac{P(H_0|h)}{P(H_1|h)} = \frac{P(h|H_0)P(H_0)}{P(h|H_1)P(H_1)} = \frac{(\frac{1}{2})(\frac{1}{2})}{(1)(\frac{1}{2})} = \frac{1}{2}.$$

Let the outcome of the second toss also be h. Assuming the tosses to be independent of one another, we then have

$$\text{second toss:} \qquad \frac{P(H_0|h_2,h_1)}{P(H_1|h_2,h_1)} = \frac{P(h_2|h_1,H_0)P(h_1|H_0)P(H_0)}{P(h_2|h_1,H_1)P(h_1|H_1)P(H_1)} = \frac{(\frac{1}{2})(\frac{1}{2})(\frac{1}{2})}{(1)(1)(\frac{1}{2})} = \frac{1}{4}.$$

It is evident, then, that the ratio of posteriors following n consecutive tosses resulting in h would be

$$n\text{th toss:} \qquad \frac{P(H_0|h_n \ldots h_1)}{P(H_1|h_n \ldots h_1)} = \frac{1}{2^n}.$$

Thus, although without direct examination one could not say with 100% certainty that the coin was biased, it would be a good bet (odds of H_0 over H_1: 1:4096) if 12 tosses led to straight heads.

It is important to note, however, that unlikely events can and do occur. No law of physics prevents a random process from leading to 12 straight heads. Indeed, the larger the number of trials, the more probable it will be that a succession of heads of any specified length will eventually turn up. In the nuclear decay experiments we consider later in the book, the equivalent of 20 h in a row occurred.

The probability of an outcome can be highly counter-intuitive if thought about in the wrong way. Consider a different application of Bayes' theorem. Suppose the probability of being infected with a particular disease is 5 in 1000 and your diagnostic test comes back positive. This test is not 100% reliable, however, but let us say that it registers accurately in 95% of the trials. By that I mean that it registers positive (+) if a person is sick (s) and negative (–) if a person is not sick ($\bar{s}$). What is the probability that you are sick?

From the given information and the rules of probability, we have the following numerical assignments.

Probability of infection $P(s) = 0.005$
Probability of no infection $P(\bar{s}) = 0.995$
Probability of correct positive: $P(+|s) = 0.95$
Probability of false negative $P(-|s) = 1 - P(+|s) = 0.05$
Probability of correct negative $P(-|\bar{s}) = 0.95$
Probability of false positive $P(+|\bar{s}) = 1 - P(-|\bar{s}) = 0.05$.

Then from Bayes' theorem it follows that the probability of being sick, given a positive test, is

$$P(s|+) = \frac{P(+|s)P(s)}{P(+|s)P(s) + P(+|\bar{s})P(\bar{s})} = \frac{(0.95)(0.005)}{(0.95)(0.005) + (0.95)(0.995)} = 0.087$$

or 8.7%, which is considerably less worrisome than one might have anticipated on the basis of the high reliability of the test. Bayes' theorem, however, takes account as well of the low incidence of infection.

1.3 Probability density function and moments

In the investigation of stochastic[2] (i.e. random) processes, the physical quantity being measured or counted is often represented mathematically by a random variable.

[2] The world "stochastic" derives from a Greek root for "to aim at", referring to a guess or conjecture.

A random variable is a quantity whose value at each observation is determined by a probability distribution. For example, the number of radioactive nuclei decaying within some specified time interval is a discrete random variable; the length of time between two successive decays is a continuous random variable. Once the probability distribution is known – or at least approximated – the probability for any outcome (or combination of outcomes) can be calculated, as well as any statistical moments (provided they exist).

If we let X stand for a discrete random variable whose set of realizable values $\{x_i\ i = 1,2,\ldots N\}$ are the possible outcomes to an experiment with corresponding probability distribution $\{p_i\}$, then the probability that the experiment leads to *some* outcome in the set is the normalization or completeness requirement $P = \sum_{i=1}^{N} p_i = 1$. The average – i.e. mean value – of some function of the outcomes, $f(X)$, is expressed symbolically by angular brackets

$$\langle f(X)\rangle = \sum_{i=1}^{N} f(x_i)p_i. \tag{1.3.1}$$

Thus the nth moment of the distribution of X is defined to be

$$\mu_n \equiv \langle X^n\rangle = \sum_{i=1}^{N} x_i^n p_i. \tag{1.3.2}$$

Several particularly significant moments or combinations of moments include:

$$\text{mean:} \qquad \mu_X \equiv \mu_1 = \langle X\rangle = \sum_{i=1}^{N} x_i p_i, \tag{1.3.3}$$

$$\text{variance:} \quad \text{var}(X) \equiv \sigma_X^2 = \left\langle (X-\mu_X)^2 \right\rangle = \mu_2 - \mu_1^2, \tag{1.3.4}$$

from which the standard deviation σ_X is calculated. We also have

$$\text{skewness: } Sk_X \equiv \left\langle \left(\frac{X-\mu_X}{\sigma_X}\right)^3 \right\rangle = \frac{\mu_3 - 3\mu_2\mu_1 + 2\mu_1^3}{\sigma_X^3}, \tag{1.3.5}$$

which is a measure of the asymmetry of a probability distribution about its center, and

$$\text{kurtosis: } K_X \equiv \left\langle \left(\frac{X-\mu_X}{\sigma_X}\right)^4 \right\rangle = \frac{\mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^4}{\sigma_X^4}, \tag{1.3.6}$$

which is a measure of the degree of flatness of a distribution near its center. It is ordinarily not necessary to go beyond the fourth moment in applying statistics to experimental distributions.

With regard to notation, the subscript X designating the random variable of interest may be omitted from the symbols for statistical functions where no confusion results.

To a continuous random variable X is associated a probability density function (pdf) $p(x)$, such that the probability that X lies within the range $(x, x + dx)$ is $p(x)dx$. The normalization requirement and moments of X are now given by integrals rather than sums:

$$\int_{-\infty}^{\infty} p(x)dx = 1 \qquad m_n = \int_{-\infty}^{\infty} x^n p(x)dx. \tag{1.3.7}$$

The range of integration can always be taken to span the full real axis by requiring, if necessary, the pdf to vanish for specific segments. Thus, if X is a non-negative-valued random variable, then one defines $p(x) = 0$ for $x < 0$.

The cumulative distribution function (cdf) $F(x)$ – sometimes referred to simply as the distribution – is the probability $\Pr(X \leq x)$, which, geometrically, is the area under the plot of the pdf up to the point x:

$$\Pr(X \leq x) \equiv F(x) = \int_{-\infty}^{x} p(x')dx'. \tag{1.3.8}$$

It therefore follows by use of Leibnitz's equation from elementary calculus

$$\frac{d}{dx}\int_{a(x)}^{b(x)} F(x,y)dy = \frac{db}{dx}F(x,b) - \frac{da}{dx}F(x,a) + \int_{a(x)}^{b(x)} \frac{\partial F(x,y)}{\partial x}dy \tag{1.3.9}$$

that differentiation of the cdf yields the pdf: $p(x) = dF/dx$. This is a practical way to obtain the pdf, as we shall see later, under circumstances where it is easier to determine the cdf directly.

1.4 The binomial distribution: "bits" [*Bin*(1, *p*)] and "pieces" [*Bin*(*n*, *p*)]

The binomial distribution, designated $Bin(n, p)$, is perhaps the most widely encountered discrete distribution in physics, and it plays an important role in the research described in this book. Consider a binomial random variable X with two outcomes per trial:

$$X = \begin{cases} \text{success} \equiv 1 \text{ with probability } p \\ \text{failure} \equiv 0 \text{ with probability } q = 1 - p. \end{cases} \tag{1.4.1}$$

The number of distinct ways of getting k successes in n independent trials, which is represented by the random variable $Y = X_1 + X_2 + \cdots + X_n$, where each subscript

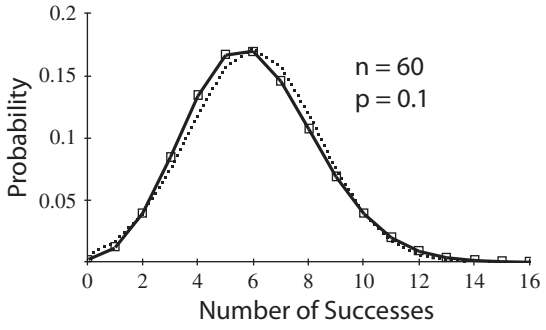

Fig. 1.1 Probability of x successes out of n trials for binomial distribution (solid) $Bin(n, p) = Bin(60, 0.1)$ and corresponding approximate normal distribution (dotted) $N(\mu, \sigma^2) = N(6,5.4)$.

labels a trial, is the coefficient of $p^k q^{n-k}$ in the binomial expansion $(p+q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}$ with combinatorial coefficient $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Thus, the binomial probability function can be written in the form

$$P(x|n,p) = \binom{n}{p} p^x q^{n-x} \qquad (n \geq x \geq 0), \tag{1.4.2}$$

which shows explicitly the two parameters of the distribution. It is then straightforward, albeit somewhat tedious, to calculate from (1.3.2) the statistical quantities

$$\mu = np \qquad \text{var} = npq \qquad Sk = \frac{(q-p)}{\sqrt{npq}} \qquad K = \frac{3(n-2)pq + 1}{npq} \tag{1.4.3}$$

and others as needed. If the probability of obtaining either outcome is the same ($p = q = \frac{1}{2}$), the distribution is symmetric and the skewness vanishes. For $p < q$ the skewness is positive, which means the distribution skews to the right as shown in Figure 1.1. In the limit of infinitely large n, the kurtosis approaches 3, which is the value for the standard normal distribution (to be considered shortly). A distribution with high kurtosis is more sharply peaked than one with low kurtosis; the tails are "fatter" (in statistical parlance), signifying a higher probability of occurrence of outlying events.

In calculating statistical moments with the binomial probability function, the trick to performing the ensuing summations is to transform them into operations on the binomial expression $(p + q)^n$ whose numerical value is 1. For illustration, consider the steps in calculation of the mean

$$\langle X \rangle = \sum_{x=0}^{n} \binom{n}{x} x p^x q^{n-x} = p \frac{d}{dp} \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = p \frac{d}{dp} (p+q)^n = np(p+q)^{n-1} \xrightarrow{q=1-p} np$$

where only in the final step does one actually substitute the value of the sum: $p + q = 1$. For higher moments, one applies $p\frac{d}{dp}$ the requisite number of times. There is a

more convenient way to achieve the same goal (with additional advantages as well) by means of a generating function, which will be introduced shortly.

1.5 The Poisson distribution: counting the improbable

The Poisson distribution, symbolized by $Poi(\mu)$, is perhaps second on the list of most widely encountered discrete distributions in physics. It is the distribution that one virtually always thinks of in connection with counting particles from disintegrating nuclei or photons from radiating atoms. More generally, it characterizes the statistics of phenomena whereby the probability of an occurrence is very low, but the number of trials is very large. Seen in that light, the Poisson distribution is a special case of the binomial distribution, and one can derive the probability function of a Poisson random variable X

$$P(x|\mu) = e^{-\mu}\frac{\mu^x}{x!} \qquad (x = 0, 1, 2\ldots) \tag{1.5.1}$$

directly from $P(x|n, p)$ by appropriately taking limits $p \to 0$ and $n \to \infty$ such that the mean $\mu = np$ remains constant. This is a tedious calculation, and a more efficient way is again afforded by use of a generating function.

The moments of the Poisson distribution are calculable from relation (1.3.2) with substitution of probability function (1.5.1). The sums are completed by the same device employed in the previous section, except that now one operates with $\mu\frac{d}{d\mu}$ on the expression $\sum_{x=0}^{\infty}\frac{\mu^x}{x!} = e^{\mu}$. For example, consider the first and second moments

$$\langle X\rangle = e^{-\mu}\sum_{x=0}^{\infty}x\frac{\mu^x}{x!} = e^{-\mu}\left(\mu\frac{d}{d\mu}\right)\sum_{x=0}^{\infty}\frac{\mu^x}{x!} = e^{-\mu}\mu e^{\mu} = \mu$$

$$\langle X^2\rangle = e^{-\mu}\sum_{x=0}^{\infty}x^2\frac{\mu^x}{x!} = e^{-\mu}\left(\mu\frac{d}{d\mu}\right)\left(\mu\frac{d}{d\mu}\right)\sum_{x=0}^{\infty}\frac{\mu^x}{x!} = e^{-\mu}\left(\mu\frac{d}{d\mu}\right)^2 e^{\mu} = \mu + \mu^2$$

from which follows the equality

$$\langle X\rangle = \mathrm{var}(X) = \mu, \tag{1.5.2}$$

which is a characteristic feature of the Poisson distribution. By analogous manipulations one obtains the skewness and kurtosis

$$Sk = \mu^{-1/2} \qquad K = 3 + \frac{1}{\mu}. \tag{1.5.3}$$

Since μ is never negative in a Poisson distribution (physically, it is a distribution of *counted* objects), Sk is also seen to be a non-negative function and therefore the Poisson distribution always skews to the right. Also, since $K > 3$, the

distribution is more sharply peaked and has fatter tails than a standard normal distribution. The above two expressions suggest, however, that as the mean gets larger, the Poisson distribution approaches the shape of the normal distribution. That this is indeed the case will be shown more rigorously by means of generating functions.

1.6 The multinomial distribution: histograms

The multinomial distribution is a generalization of the binomial distribution. It is the theoretical basis for a histogram: the graphical representation of counted or measured data sorted into categories (called classes) of specified value. Consider a random variable X representing the result of an experiment (i.e. single trial) with a multiplicity r of possible outcomes $\{x_i\, i = 1 \ldots r\}$ with corresponding probabilities $\{p_i\}$. Then the probability that in n trials the outcome x_i will occur n_i times is obtained from expansion of the nth power of a multinomial form $(p_1 + p_2 + \cdots + p_r)^n$, which leads to the expression

$$P(n_1, n_2, \ldots n_r\,|n; p_1, p_2, \ldots p_r) \equiv P\big(\{n_i\}|n; \{p_i\}\big) = \binom{n}{n_1 \ldots n_r} p_1^{n_1} p_2^{n_2} \ldots p_r^{n_r} = n! \prod_{i=1}^{r} \frac{p_i^{n_i}}{n_i!}. \tag{1.6.1}$$

The two-tiered symbol

$$\binom{n}{n_1 \ldots n_r} \equiv \frac{n!}{\prod_{i=1}^{r} n_i!} \qquad \text{with} \quad \sum_{i=1}^{r} n_i = n \tag{1.6.2}$$

defined above is the multinomial combinatorial coefficient.

The form of $P(\{n_i\}|n;\{p_i\})$ may be understood in the following way, which is a generalization of the way one would deduce the binomial probability distribution.

- The probability that n_i independent events of type x_i occur is $p_i^{n_i}$.
- Thus, the probability that a *particular* sequence of n_1 x_1s, n_2 x_2s, $\ldots$ n_r x_rs occurs is $p_1^{n_1} p_2^{n_2} \ldots p_r^{n_r}$ since all trials are independent of one another.
- However, this sequence could occur in $\binom{n}{n_1 \ldots n_r}$ different ways.

It is useful to demonstrate this combinatorial statement since the multinomial distribution enters significantly (in the form of a histogram) in all the experimental investigations to be discussed in the book.

The number of ways one can partition a set of size n into r ordered subsets such that the first has size n_1, the second has size n_2, etc., and where $n_1 + n_2 + \cdots + n_r = n$ is the product