Probabilities

Robert B. Griffiths
Version of 12 January 2010

References:

Feller, *An introduction to probability theory and its applications*, Vol. 1, 3d ed (Wiley 1968). See Introduction, Ch. I, Ch. V

DeGroot and Schervish, *Probability and Statistics*, 3d ed (Addison-Wesley, 2002), Chs 1, 2, 3, 4

Very compact introductions to material relevant to quantum mechanics:

QCQI = *Quantum Computation and Quantum Information* by Nielsen and Chuang (Cambridge, 2000), App. 1

CQT = *Consistent Quantum Theory* by Robert Griffiths (Cambridge, 2002), Secs. 5.1, 8.2, 9.1, 9.2

Contents

1 Basics

★ Ordinary (classical) probability theory uses three key concepts: sample space $\mathcal{S}$, event algebra $\mathcal{B}$, probabilities $\mathcal{P}$.

1.1 Sample space

★ Sample space $\mathcal{S}$ of *mutually exclusive possibilities*, thought of as outcomes of an ideal experiment, one and only one of which actually occurs, or is true, in a particular case.

◦ Examples. Coin toss: H or T. Die: $\{s = 1, 2, 3, 4, 5, 6\}$

◦ A sample space can be either discrete or continuous (e.g., integers, real numbers), and in the former case either finite or infinite. For our purposes it suffices to consider finite discrete sample spaces.

1.2 Event algebra

★ Event algebra $\mathcal{B}$ consisting of a collection of subsets of $\mathcal{S}$. In the simplest situation it is the collection of all subsets of $\mathcal{S}$, but is sometimes a smaller collection. The collection $\mathcal{B}$ must be closed under the operations of union, intersection, and complement, and must contain $\mathcal{S}$ as one of its elements. Elements of $\mathcal{B}$ are known as "events."

◦ The symbol $\mathcal{B}$ is used because this is a Boolean algebra under the usual set-theoretic operations.

• Example: For a coin there are $2^2 = 4$ possible subsets of the set $\{H, T\}$, namely $\emptyset$ (the empty set; yes it must be included in the algebra), $\{H\}$, $\{T\}$, $\{H, T\}$.

◦ We won't worry too much about the difference between T and $\{T\}$.

• More interesting example. For a die, $s = 6$ is a *simple* event, corresponding to a single element from the sample space, while $s \geq 3$ is an example of a *compound* event, as it corresponds to the subset $\{3, 4, 5, 6\}$. Another compound event: s is even. The adjectives "simple" and "compound" are not always used, but are sometimes convenient.

□ Exercise. How big is the algebra containing all subsets of $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ in the case of a die?

• A case where $\mathcal{B}$ does not consist of all subsets: We are only interested in whether the number of spots on the die is even or odd. One event in $\mathcal{B}$ is $\{1, 3, 5\}$, a second is $\{2, 4, 6\}$, and $\mathcal{B}$ contains their intersection $\emptyset$ and their union $\mathcal{S}$, so four subsets in all.

□ Exercise. What is the smallest Boolean algebra that contains $\{1, 3\}$ in the case of a die? That contains both $\{1, 3\}$ and $\{1, 4, 5\}$?

1.3 Probabilities

★ Probabilities $\mathcal{P}$: a set of real numbers between 0 and 1 assigned to the events in $\mathcal{B}$ in a manner satisfying certain rules. In the case of a discrete sample space with $\mathcal{B}$ consisting of all its subsets one can proceed as follows. For each $s \in \mathcal{S}$ choose a number $p(s)$ between 0 and 1, the probability of s, in such a way that

$$\sum_s p(s) = 1. \tag{1}$$

Then to each B in $\mathcal{B}$ assign the probability

$$\Pr(B) = \sum_{s \in B} p(s). \tag{2}$$

• Probability theory by itself does *not* tell one what numbers to choose for the different probabilities $p(s)$. These typically represent some idealization, e.g., an ideal coin should turn up H or T with equal probability, or results of measurements, or some sort of guess. Often one chooses the probability distribution so that it depends on a collection of (real) parameters, and then adjusts the parameters in some way, e.g., to obtain agreement with experiments.

◦ Quantum mechanics is unique as a physical theory in that certain probabilities represent a "law of nature," i.e., enter the theory as axioms.

★ The probability for the *conjunction* of two events A and B, corresponding to the subset $A \cap B$ of $\mathcal{S}$, can be written as $\Pr(A \cap B)$ or $\Pr(A \wedge B)$, but the usual custom, followed in these notes, is to use a comma between A and B:

$$\Pr(A, B) = \Pr(A \wedge B) = \Pr(A \cap B) \tag{3}$$

• Similarly $\Pr(A, B, C)$, etc.

• There seems to be no similarly compact notation for disjunction. Use $\Pr(A \vee B)$ or $\Pr(A \cup B)$ for A OR B understood as "A or B or both."

□ Exercise. Use (2) to find a formula expressing $\Pr(A \vee B)$ in terms of $\Pr(A)$, $\Pr(B)$, and $\Pr(A, B) = \Pr(A \wedge B)$.

1.4 Ensembles

★ The notion of an *ensemble* provides a useful way of "visualizing" probabilities. One imagines a large number of "identical" systems—think of a large collection of coins, or of dice—with each element in the ensemble described by the same sample space, but different elements of the ensemble in different states, with the number of systems in the ensemble in state s proportional to $p(s)$. Equivalently, one can imagine carrying out the same experiment over and over again a large number of times, and each time recording the outcome.

• Thus if there are N systems, the number in the state s will be $Np(s)$. Of course this may not be an integer, but you should imagine that N is so large that this really does not matter much. For this reason one sometimes speaks of an "infinite ensemble."

• In such an ensemble, $p(s)$ is the *relative frequency* with which s occurs. More generally, the fraction of elements of the ensemble for which property A is true is given by $\Pr(A)$.

• By contrast, when an experiment (such as tossing a coin) is carried out a small number of times, the actual set of outcomes may be significantly different from $Np(s)$. Analyzing what to expect in such a "finite ensemble" is a nontrivial exercise in probability theory.

1.5 Conditional probabilities

★ Conditional probabilities play an important role in applications of probability theory. Let A and B be two events in $\mathcal{B}$. Then

$$\Pr(A|B) = \Pr(A, B)/\Pr(B), \tag{4}$$

is the *conditional probability of A given B*, assuming that $\Pr(B) > 0$, so that the right side is defined. Multiplying through by $\Pr(B)$ yields

$$\Pr(A, B) = \Pr(A|B)\Pr(B). \tag{5}$$

It is worthwhile memorizing both formulas.

• The intuitive significance of $\Pr(A|B)$ is that if one takes the ensemble corresponding to one's probability distribution and considers only those systems for which property B is true, then $\Pr(A|B)$ is the relative frequency of property A in this new ensemble.

• As a function of its first argument A, $\Pr(A|B)$ satisfies all the rules for a set of probabilities on the event algebra $\mathcal{B}$. For this reason, if B is held fixed during a particular discussion, or is evident from the context, it is sometimes omitted from the notation: $\Pr(A)$ may stand for $\Pr(A|B)$ when it would be pedantic to keep mentioning B.

2 Random Variables

2.1 Random variables, averages, indicator functions

★ A *random variable* is a real (or sometimes complex) valued function on the sample space.

○ Bad terminology, but we're stuck with it.

• For a random variable V on a discrete sample space with points labeled by s we may use $V(s) = V(s)$ for the value of the random variable V at the point s.

★ The *mean* or *average* or *expectation* of a random variable V is

$$E(V) = \langle V \rangle = \sum_s p(s)V(s) \tag{6}$$

★ A useful class of random variables are the *indicator functions* which take only two values, 0 and 1. The indicator P_B for a subset B of $\mathcal{S}$ is defined by

$$P_B(s) = \begin{cases} 1 & \text{if } s \in B \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

• Using the indicator function we can write

$$\Pr(B) = \langle B \rangle. \tag{8}$$

2.2 Probability distributions

★ The *probability distribution* of a random variable V is defined in the following way in the case of a discrete sample space:

$$\Pr(V{=}v) = \Pr(\{s : V(s){=}v\}) = \sum_{s:V(s)=v} p(s) \tag{9}$$

• Note that if $v \neq w$, the subsets of $\mathcal{S}$ where $V(s){=}v$ and $V(s){=}w$ are disjoint. Furthermore, since V must take some value at every $s \in \mathcal{S}$, otherwise it is not what mathematicians call a function, it follows that

$$\sum_v \Pr(V{=}v) = 1. \tag{10}$$

• As long as one is only interested in the random variable V, or V and other random variables that are functions of V, it is possible to define a new sample space $\mathcal{V}$ which simply consists of the different possible values v of V. This is often helpful. Note that (10) is then the counterpart of (1).

2.3 Joint and conditional distributions

★ Let V and W be two random variables defined on the same sample space. The *joint probability distribution* is defined by

$$\Pr(V{=}v, W{=}w) = \Pr(\{s : V(s){=}v, W(s){=}w\}), \tag{11}$$

where remember that the comma is to be read as "AND": $V(s){=}v$ AND $W(s){=}w$. It satisfies

$$\sum_{v,w} \Pr(V{=}v, W{=}w) = 1. \tag{12}$$

• The *marginal* distributions for V and W are given by

$$Pr(V{=}v) = \sum_w \Pr(V{=}v, W{=}w), \quad Pr(W{=}w) = \sum_v \Pr(V{=}v, W{=}w). \tag{13}$$

• All of this generalizes in a pretty obvious way to three or more random variables, e.g.,

$$\Pr(U{=}u, W{=}w) = \sum_v \Pr(U{=}u, V{=}v, W{=}w). \tag{14}$$

★ Conditional probability distributions:

$$\Pr(V{=}v|W{=}w) = \Pr(V{=}v, W{=}w)/\Pr(W{=}w), \quad \Pr(V{=}v, W{=}w) = \Pr(V{=}v|W{=}w)\Pr(W{=}w). \tag{15}$$

The two expressions are equivalent as long as the first is well defined, i.e., $\Pr(W{=}w)$ is not zero, and both should be memorized.

• Compare with (4) and (5). The conditional $\Pr(V{=}v|W{=}w)$ can be regarded as an "ordinary" probability distribution so far as the first argument is concerned and as such follows the usual rules, e.g., $\sum_v \Pr(V{=}v|W{=}w) = 1$.

★ Notation of the kind $\Pr(V{=}v, W{=}w)$ is unambiguos but awkward. One often replaces it with $\Pr(v, w)$ or something similar when the meaning is clear from the context.

◦ Often $P(v, w)$ or $p(v, w)$ are used in place of $\Pr(v, w)$, which is fine as long as one does not have other uses for P or p.

2.4 Independent random variables

★ Two random variables V and W are said to be *(statistically) independent* provided

$$\Pr(V{=}v, W{=}w) = \Pr(V{=}v)\Pr(W{=}w) \tag{16}$$
$$\Pr(V{=}v|W{=}w) = \Pr(V{=}v). \tag{17}$$

The two definitions are equivalent as long as $\Pr(W{=}w)$ is not zero, and both should be memorized. It is the second which motivates the term "independent," as it says that even if the value of W is given, this does not influence the probability distribution of V. And of course the roles of V and W can be interchanged:

$$\Pr(W{=}w|V{=}v) = \Pr(W{=}w). \tag{18}$$

is also equivalent to (16), and hence (17), provided $\Pr(V{=}v)$ is not zero.

• This notion of independence can also be applied to pairs of events A and B in the event algebra $\mathcal{B}$: they are (statistically) independent provided

$$\Pr(A, B) = \Pr(A)\Pr(B). \tag{19}$$

□ Exercise. Show that (19) corresponds to statistical independence (in the sense defined earlier) of the characteristic functions P_A and P_B.

★ The statistical independence of three random variables is given by the (almost) obvious generalization

$$\Pr(U{=}u, V{=}v, W{=}w) = \Pr(U{=}u)\Pr(V{=}v)\Pr(W{=}w) \tag{20}$$

of (16), and similarly for four or more random variables.

• The formula generalizing (19) to three events is

$$\Pr(A, B, C) = \Pr(A)\Pr(B)\Pr(C), \tag{21}$$

with an obvious generalization to four or more.

★ Note that (20) implies *but is not implied by* the conditions of pair independence:

$$\Pr(u, v) = \Pr(u)\Pr(v), \quad \Pr(u, w) = \Pr(u)\Pr(w), \quad \Pr(v, w) = \Pr(v)\Pr(w) \tag{22}$$

in a compact notation.

□ Exercise. Find an explicit example of three events A, B, and C such that

$$\Pr(A, B) = \Pr(A)\Pr(B), \quad \Pr(A, C) = \Pr(A)\Pr(C), \quad \Pr(B, C) = \Pr(B)\Pr(C) \tag{23}$$

but (21) is *not* true. [Hint. Feller Sec. V.3.]

★ Random variables or events that are *not* (statistically) independent are *(statistically) dependent*.

• The term "correlated" (along with the associated "correlations") is used in an informal way to mean "statistically dependent," but risks confusion with the quantity defined in (26) below.

2.5 Variance, covariance, correlation

★ The *variance* of a random variable V is the square of the *standard deviation* σ:

$$\mathrm{Var}(V) = \sigma_V^2 = E[(V - E(V))^2] = E(V^2) - [E(V)]^2 \tag{24}$$

★ If V and W are two random variables their *covariance* is defined to be

$$\mathrm{Cov}(V, W) = E\{[V - E(V)][W - E(W)]\} \tag{25}$$

• One then defines the *correlation* as

$$\text{correlation}\,(V, W) = \frac{\mathrm{Cov}(V, W)}{\sigma_V \sigma_W} \tag{26}$$

where σ_V and σ_W are the standard deviations of V and W.

• The term "correlation" is, however, also used in a more general sense to indicate a relationship (not necessarily precisely defined) between two random variables that are not statistically independent.

3 Stochastic Processes

3.1 Examples; sample spaces

★ Probability theory is often applied to a sequence of events occurring at successive times: dice are rolled repeatedly, rain falls or doesn't fall on successive days, a Brownian particle is first here and next there, etc. Such *stocahstic* or *random processes* are described by the usual machinery of probability theory, but because of the temporal ordering and the sort of questions one is interested in they have acquired some special terminology.

• Stochastic processes are used to describe the time development of systems whose dynamics is not deterministic. Before the advent of quantum mechanics the idea was widespread that the world is fundamentally deterministic in its dynamics, so that stochastic processes represent a "coarse-grained" description of real processes. However, the quantum world possesses an intrinsic randomness, so stochastic processes are now a fundamental part of physics.

★ Consider a coin tossed three times in a row, with outcome heads H or tails T at each time. The eight possible sequences or *histories*, HHH, HHT, HTH, $\ldots TTT$, constitute the *sample space* of mutually exclusive possibilities. Note that temporal order is important: HTT is not the same history as THT.

• The event algebra contains 2^8 possible subsets of histories, including the empty set, though one could also use a smaller Boolean algebra. For example, one could ignore what happens on the second toss.

• Note that the sample space for tossing a coin three times in a row is the same as that for simultaneously tossing three different coins. This second "atemporal" perspective is sometimes useful: even though the formalism is the same, one's physical intuition is a bit different.

★ Random walks are used to model many physical processes. Perhaps the simplest example is that of a particle moving or "hopping" in one dimension; we assume its location x can take only integer values. Let x_t be its position at time t. We shall only consider discrete values of t, and it is convenient to assume they

are integers. The sample space for a walk that begins at $t = 0$ and terminates at time $t = f$ then consists of sequences of the form

$$\mathbf{x} = (x_0, x_1, x_2, \ldots x_f) \tag{27}$$

• One often imposes the restriction that $|x_{t+1} - x_t| \leq 1$: during a single time interval the particle either stays where it is or hops one step to the right or to the left. It is useful to view a walk subject to this restriction in the following way. Define

$$s_t = x_{t+1} - x_t, \tag{28}$$

and let the sample space consist of sequences

$$\mathbf{s} = (s_0, s_2, \ldots s_{f-1}). \tag{29}$$

Given the starting position x_0 of the walker the positions at later times are determined by the s_t values, so interesting statistical properties of the walk can be deduced from the probability distribution $\Pr(\mathbf{s})$.

3.2 Probabilities

★ Just as in other applications of probability theory there are no set rules which specify in advance the probabilities in a stochastic process. They are often chosen so as to provide a simple easy-to-calculate model of a physical situation, perhaps with parameters which can be fitted to experiment. (Quantum physics is exceptional in that at least to some extent the rules for assigning probabilities are part of a fundamental physical theory.)

★ The simplest assumption is that events at successive times are statistically independent, and their probabilities are identical. Thus in the case of a coin one might suppose that the probability of its landing heads is p_H, tails is p_T, and assign to HTT the probability $p_H p_T^2$, and similarly for other elements of the sample space.

• Another example is provided by a random walk in which we assume that probabilities for s_t to be -1, 0, or $+1$ are nonnegative numbers p, q, and r that sum to 1, and the probability of the sequence $\mathbf{s}$ in (29) is the product of the probabilities for the s_j values it contains.

3.3 Markov chains

★ The simplest stochastic processes are those in which the events at successive times are statistically independent. However, one is often interested in constructing models in which there is some correlation between successive events. A Markov chain provides what is probably the simplest example of a correlated process.

• Assume that the sample space consists of histories which are sequences of the form (27), where f can be as large as one wants, but x_t can only take on a finite set of values, say the integers between 1 and N. The event algebra shall consist of all subsets of histories of this form.

★ A Markov chain results when one assigns probabilities $\Pr(\mathbf{x})$ in a way that satisfies the *Markov condition*:

$$\Pr(x_{t+1}|x_t, x_{t-1}, x_{t-2}, \ldots x_1) = \Pr(x_{t+1}|x_t). \tag{30}$$

for $t = 2, 3, \ldots$.

• The intuitive interpretation is that whereas the probability of x_{t+1} depends on x_t at the immediately preceding time, there is no influence from values of x at still earlier times. The process has a very short "memory" that extends backwards only one step in time.

• The random walk discussed earlier has this property when the probabilities of the s_t in (28) are statistically independent. Where the walker will be at time $t + 1$ is influenced by its location at time t, but not by its earlier history.

★ The probability distribution $\Pr(\mathbf{x})$ for a Markov chain is entirely determined by the *initial* distribution $\Pr(x_0)$ and the collection $\Pr(x_{t+1}|x_t)$ of *transition probabilities* for $t = 0, 1, 2, \ldots$. The simplest situation is that of a *stationary* Markov chain in which

$$\Pr(x_{t+1}|x_t) = M(x_{t+1}, x_t), \tag{31}$$

where $M(x', x) = M_{x'x}$ is a fixed matrix of nonnegative numbers that does not depend upon the time t.

◦ A *nonstationary* Markov chain is one in which the Markov matrix M in (31) depends on t. In what follows we only consider stationary Markov processes.

□ Exercise. Show that the initial distribution and the transition probabilities together determine $\Pr(\mathbf{x})$ by first writing down a formula for $\Pr(x_0, x_1)$ and then one for $\Pr(x_0, x_1, x_2)$. At this point it should be obvious how to continue on and get a formula for $\Pr(\mathbf{x}) = \Pr(x_0, \ldots x_f)$.

• Since summing a conditional probability over its first argument (the one on the left) must add to 1, the Markov matrix M must satisfy

$$\sum_{x'} M(x', x) = \sum_{x'} M_{x'x} = 1. \tag{32}$$

for every value of x. That is, each column of the matrix sums to 1.

★ Let us call

$$\rho_t(x) = \Pr(x_t{=}x), \tag{33}$$

which is the marginal probability distribution at time t, a *single time* probability. If one thinks of it as a column vector ρ_t with components labeled by x, then

$$\rho_{t+1} = M \cdot \rho_t, \tag{34}$$

where the right side is understood in the usual sense of a matrix multiplying a column vector producing another column vector.

★ WARNING! The definition of the matrix M given above is "natural" for physicists. Books on probability written by professional mathematicians typically use an alternative definition in which the Markov matrix is the transpose of the one defined here, and (34) must be written in the form: row vector = row vector times matrix. Obviously it doesn't make much difference which scheme one uses, as long as one knows what it is.