32.11. 32.2↑ Absolute continuity of a set function φ with respect to a measure μ is defined just as if φ were itself a measure: $\mu(A) = 0$ must imply that $\varphi(A) = 0$. Show that, if this holds and μ is σ-finite, then $\varphi(A) = \int_A f d\mu$ for some integrable f. Show that $A^+ = [\omega: f(\omega) \geq 0]$ and $A^- = [\omega: f(\omega) < 0]$ give a Hahn decomposition for φ. Show that the three variations satisfy $\varphi^+(A) = \int_A f^+ \, d\mu$, $\varphi^-(A) = \int_A f^- \, d\mu$, and $|\varphi|(A) = \int_A |f| \, d\mu$. *Hint:* To construct f, start with (32.2).

32.12. ↑ A *signed measure* φ is a set function that satisfies (32.1) if $A_1, A_2, \ldots$ are disjoint and may assume one of the values $+\infty$ and $-\infty$ but not both. Extend the Hahn and Jordan decompositions to signed measures

32.13. 31.22↑ Suppose that μ and ν are a probability measure and a σ-finite measure on the line and that $\nu \ll \mu$. Show that the Radon–Nikodym derivative f satisfies

$$\lim_{h \to 0} \frac{\nu(x-h, x+h]}{\mu(x-h, x+h]} = f(x)$$

on a set of μ-measure 1.

32.14. Find on the unit interval uncountably many probability measures μ_p, $0 < p < 1$, with supports S_p such that $\mu_p\{x\} = 0$ for each x and p and the S_p are disjoint in pairs.

32.15. Let $\mathcal{F}_0$ be the field consisting of the finite and the cofinite sets in an uncountable Ω. Define φ on $\mathcal{F}_0$ by taking $\varphi(A)$ to be the number of points in A if A is finite, and the negative of the number of points in A^c if A is cofinite. Show that (32.1) holds (this is not true if Ω is countable). Show that there are no negative sets for φ (except the empty set), that there is no Hahn decomposition, and that φ does not have bounded range.

SECTION 33. CONDITIONAL PROBABILITY

The concepts of conditional probability and expected value with respect to a σ-field underlie much of modern probability theory. The difficulty in understanding these ideas has to do not with mathematical detail so much as with probabilistic meaning, and the way to get at this meaning is through calculations and examples, of which there are many in this section and the next.

The Discrete Case

Consider first the conditional probability of a set A with respect to another set B. It is defined of course by $P(A|B) = P(A \cap B)/P(B)$, unless $P(B)$ vanishes, in which case it is not defined at all.

It is helpful to consider conditional probability in terms of an observer in possession of partial information.† A probability space $(\Omega, \mathcal{F}, P)$ describes

†As always, *observer*, *information*, *know*, and so on are informal, nonmathematical terms; see the related discussion in Section 4 (p. 57).

the working of a mechanism, governed by chance, which produces a result ω distributed according to P; $P(A)$ is for the observer the probability that the point ω produced lies in A. Suppose now that ω lies in B and that the observer learns this fact and no more. From the point of view of the observer, now in possession of this partial information about ω, the probability that ω also lies in A is $P(A|B)$ rather than $P(A)$. This is the idea lying back of the definition.

If, on the other hand, ω happens to lie in B^c and the observer learns of this, his probability instead becomes $P(A|B^c)$. These two conditional probabilities can be linked together by the simple function

$$f(\omega) = \begin{cases} P(A|B) & \text{if } \omega \in B, \\ P(A|B^c) & \text{if } \omega \in B^c. \end{cases} \tag{33.1}$$

The observer learns whether ω lies in B or in B^c; his new probability for the event $\omega \in A$ is then just $f(\omega)$. Although the observer does not in general know the argument ω of f, he can calculate the value $f(\omega)$ because he knows which of B and B^c contains ω. (Note conversely that from the value $f(\omega)$ it is possible to determine whether ω lies in B or in B^c, unless $P(A|B) = P(A|B^c)$—that is, unless A and B are independent, in which case the conditional probability coincides with the unconditional one anyway.)

The sets B and B^c partition Ω, and these ideas carry over to the general partition. Let $B_1, B_2, \ldots$ be a finite or countable partition of Ω into $\mathscr{F}$-sets, and let $\mathscr{G}$ consist of all the unions of the B_i. Then $\mathscr{G}$ is the σ-field generated by the B_i. For A in $\mathscr{F}$, consider the function with values

$$f(\omega) = P(A|B_i) = \frac{P(A \cap B_i)}{P(B_i)} \quad \text{if } \omega \in B_i, \qquad i = 1, 2, \ldots. \tag{33.2}$$

If the observer learns which element B_i of the partition it is that contains ω, then his new probability for the event $\omega \in A$ is $f(\omega)$. The partition $\{B_i\}$, or equivalently the σ-field $\mathscr{G}$, can be regarded as an experiment, and to learn which B_i it is that contains ω is to learn the outcome of the experiment. For this reason the function or random variable f defined by (33.2) is called the *conditional probability of A given $\mathscr{G}$* and is denoted $P[A\|\mathscr{G}]$. This is written $P[A\|\mathscr{G}]_\omega$ whenever the argument ω needs to be explicitly shown.

Thus $P[A\|\mathscr{G}]$ is the function whose value on B_i is the ordinary conditional probability $P(A|B_i)$. This definition needs to be completed, because $P(A|B_i)$ is not defined if $P(B_i) = 0$. In this case $P[A\|\mathscr{G}]$ will be taken to have any constant value on B_i; the value is arbitrary but must be the same over all of the set B_i. If there are nonempty sets B_i for which $P(B_i) = 0$, $P[A\|\mathscr{G}]$ therefore stands for any one of a family of functions on Ω. A specific such function is for emphasis often called a *version* of the conditional

probability. Note that any two versions are equal except on a set of probability 0.

Example 33.1. Consider the Poisson process. Suppose that $0 \le s \le t$, and let $A = [N_s = 0]$ and $B_i = [N_t = i]$, $i = 0, 1, \ldots$. Since the increments are independent (Section 23), $P(A|B_i) = P[N_s = 0]P[N_t - N_s = i]/P[N_t = i]$, and since they have Poisson distributions (see (23.9)), a simple calculation reduces this to

$$P\left[N_s = 0 \| \mathscr{G}\right]_\omega = \left(1 - \frac{s}{t}\right)^i \quad \text{if } \omega \in B_i, \qquad i = 0, 1, 2, \ldots. \tag{33.3}$$

Since $i = N_t(\omega)$ on B_i, this can be written

$$P\left[N_s = 0 \| \mathscr{G}\right]_\omega = \left(1 - \frac{s}{t}\right)^{N_t(\omega)}. \tag{33.4}$$

Here the experiment or observation corresponding to $\{B_i\}$ or $\mathscr{G}$ determines the number of events—telephone calls, say—occurring in the time interval $[0, t]$. For an observer who knows this number but not the locations of the calls within $[0, t]$, (33.4) gives his probability for the event that none of them occurred before time s. Although this observer does not known ω, he knows $N_t(\omega)$, which is all he needs to calculate the right side of (33.4). ■

Example 33.2. Suppose that $X_0, X_1, \ldots$ is a Markov chain with state space S as in Section 8. The events

$$[X_0 = i_0, \ldots, X_n = i_n] \tag{33.5}$$

form a finite or countable partition of Ω as $i_0, \ldots, i_n$ range over S. If $\mathscr{G}_n$ is the σ-field generated by this partition, then by the defining condition (8.2) for Markov chains, $P[X_{n+1} = j \| \mathscr{G}_n]_\omega = p_{i_n j}$ holds for ω in (33.5). The sets

$$[X_n = i] \tag{33.6}$$

for $i \in S$ also partition Ω, and they generate a σ-field $\mathscr{G}_n^0$ smaller than $\mathscr{G}_n$. Now (8.2) also stipulates $P[X_{n+1} = j \| \mathscr{G}_n^0]_\omega = p_{ij}$ for ω in (33.6), and the essence of the Markov property is that

$$P\left[X_{n+1} = j \| \mathscr{G}_n\right] = P\left[X_{n+1} = j \| \mathscr{G}_n^0\right]. \tag{33.7}$$

■

The General Case

If $\mathscr{G}$ is the σ-field generated by a partition $B_1, B_2, \ldots$, then the general element of $\mathscr{G}$ is a disjoint union $B_{i_1} \cup B_{i_2} \cup \cdots$, finite or countable, of certain of the B_i. To know which set B_i it is that contains ω is the same thing

as to know which sets in $\mathscr{G}$ contain ω and which do not. This second way of looking at the matter carries over to the general σ-field $\mathscr{G}$ contained in $\mathscr{F}$. (As always, the probability space is $(\Omega, \mathscr{F}, P)$.) The σ-field $\mathscr{G}$ will not in general come from a partition as above.

One can imagine an observer who knows for each G in $\mathscr{G}$ whether $\omega \in G$ or $\omega \in G^c$. Thus the σ-field $\mathscr{G}$ can in principle be identified with an experiment or observation. This is the point of view adopted in Section 4; see p. 57. It is natural to try and define conditional probabilities $P[A\|\mathscr{G}]$ with respect to the experiment $\mathscr{G}$. To do this, fix an A in $\mathscr{F}$ and define a finite measure ν on $\mathscr{G}$ by

$$\nu(G) = P(A \cap G), \qquad G \in \mathscr{G}.$$

Then $P(G) = 0$ implies that $\nu(G) = 0$. The Radon–Nikodym theorem can be applied to the measures ν and P on the measurable space $(\Omega, \mathscr{G})$ because the first one is absolutely continuous with respect to the second.† It follows that there exists a function or random variable f, measurable $\mathscr{G}$ and integrable with respect to P, such that† $P(A \cap G) = \nu(G) = \int_G f\,dP$ for all G in $\mathscr{G}$.

Denote this function f by $P[A\|\mathscr{G}]$. It is a random variable with two properties:

(i) $P[A\|\mathscr{G}]$ *is measurable* $\mathscr{G}$ *and integrable.*
(ii) $P[A\|\mathscr{G}]$ *satisfies the functional equation*

$$\int_G P[A\|\mathscr{G}]\,dP = P(A \cap G), \qquad G \in \mathscr{G}. \tag{33.8}$$

There will in general be many such random variables $P[A\|\mathscr{G}]$, but any two of them are equal with probability 1. A specific such random variable is called a *version* of the conditional probability.

If $\mathscr{G}$ is generated by a partition $B_1, B_2, \ldots$ the function f defined by (33.2) is measurable $\mathscr{G}$ because $[\omega\colon f(\omega) \in H]$ is the union of those B_i over which the constant value of f lies in H. Any G in $\mathscr{G}$ is a disjoint union $G = \bigcup_k B_{i_k}$, and $P(A \cap G) = \sum_k P(A|B_{i_k})P(B_{i_k})$, so that (33.2) satisfies (33.8) as well. Thus the general definition is an extension of the one for the discrete case.

Condition (i) in the definition above in effect requires that the values of $P[A\|\mathscr{G}]$ depend only on the sets in $\mathscr{G}$. An observer who knows the outcome of $\mathscr{G}$ viewed as an experiment knows for each G in $\mathscr{G}$ whether it contains ω or not; for each x he knows this in particular for the set $[\omega'\colon P[A\|\mathscr{G}]_{\omega'} = x]$,

†Let P_0 be the restriction of P to $\mathscr{G}$ (Example 10.4), and find on $(\Omega, \mathscr{G})$ a density f for ν with respect to P_0. Then, for $G \in \mathscr{G}$, $\nu(G) = \int_G f\,dP_0 = \int_G f\,dP$ (Example 16.4). If g is another such density, then $P[f \neq g] = P_0[f \neq g] = 0$.

and hence he knows in principle the functional value $P[A\|\mathscr{G}]_\omega$ even if he does not know ω itself. In Example 33.1 a knowledge of $N_t(\omega)$ suffices to determine the value of (33.4)—ω itself is not needed.

Condition (ii) in the definition has a gambling interpretation. Suppose that the observer, after he has learned the outcome of $\mathscr{G}$, is offered the opportunity to bet on the event A (unless A lies in $\mathscr{G}$, he does not yet know whether or not it occurred). He is required to pay an entry fee of $P[A\|\mathscr{G}]$ units and will win 1 unit if A occurs and nothing otherwise. If the observer decides to bet and pays his fee, he gains $1 - P[A\|\mathscr{G}]$ if A occurs and $-P[A\|\mathscr{G}]$ otherwise, so that his gain is

$$(1 - P[A\|\mathscr{G}])I_A + (-P[A\|\mathscr{G}])I_{A^c} = I_A - P[A\|\mathscr{G}].$$

If he declines to bet, his gain is of course 0. Suppose that he adopts the strategy of betting if G occurs but not otherwise, where G is some set in $\mathscr{G}$. He can actually carry out this strategy, since after learning the outcome of the experiment $\mathscr{G}$ he knows whether or not G occurred. His expected gain with this strategy is his gain integrated over G:

$$\int_G (I_A - P[A\|\mathscr{G}])\,dP.$$

But (33.8) is exactly the requirement that this vanish for each G in $\mathscr{G}$. Condition (ii) requires then that each strategy be fair in the sense that the observer stands neither to win nor to lose on the average. Thus $P[A\|\mathscr{G}]$ is the just entry fee, as intuition requires.

Example 33.3. Suppose that $A \in \mathscr{G}$, which will always hold if $\mathscr{G}$ coincides with the whole σ-field $\mathscr{F}$. Then I_A satisfies conditions (i) and (ii), so that $P[A\|\mathscr{G}] = I_A$ with probability 1. If $A \in \mathscr{G}$, then to know the outcome of $\mathscr{G}$ viewed as an experiment is in particular to know whether or not A has occurred. ■

Example 33.4. If $\mathscr{G}$ is $\{0, \Omega\}$, the smallest possible σ-field, every function measurable $\mathscr{G}$ must be constant. Therefore, $P[A\|\mathscr{G}]_\omega = P(A)$ for all ω in this case. The observer learns nothing from the experiment $\mathscr{G}$. ■

According to these two examples, $P[A\|\{0, \Omega\}]$ is identically $P(A)$, whereas I_A is a version of $P[A\|\mathscr{F}]$. For any $\mathscr{G}$, the function identically equal to $P(A)$ satisfies condition (i) in the definition of conditional probability, whereas I_A satisfies condition (ii). Condition (i) becomes more stringent as $\mathscr{G}$ decreases, and condition (ii) becomes more stringent as $\mathscr{G}$ increases. The two conditions work in opposite directions and between them delimit the class of versions of $P[A\|\mathscr{G}]$.

Example 33.5. Let Ω be the plane R^2 and let $\mathscr{F}$ be the class $\mathscr{R}^2$ of planar Borel sets. A point of Ω is a pair (x, y) of reals. Let $\mathscr{G}$ be the σ-field consisting of the vertical strips, the product sets $E \times R^1 = [(x, y): x \in E]$, where E is a linear Borel set. If the observer knows for each strip $E \times R^1$ whether or not it contains (x, y), then, as he knows this for each one-point set E, he knows the value of x. Thus the experiment $\mathscr{G}$ consists in the determination of the first coordinate of the sample point. Suppose now that P is a probability measure on $\mathscr{R}^2$ having a density $f(x, y)$ with respect to planar Lebesgue measure: $P(A) = \iint_A f(x, y)\,dx\,dy$. Let A be a horizontal strip $R^1 \times F = [(x, y): y \in F]$, F being a linear Borel set. The conditional probability $P[A\|\mathscr{G}]$ can be calculated explicitly.

Put

$$\varphi(x, y) = \frac{\int_F f(x, t)\,dt}{\int_{R^1} f(x, t)\,dt}. \tag{33.9}$$

Set $\varphi(x, y) = 0$, say, at points where the denominator here vanishes; these points form a set of P-measure 0. Since $\varphi(x, y)$ is a function of x alone, it is measurable $\mathscr{G}$. The general element of $\mathscr{G}$ being $E \times R^1$, it will follow that φ is a version of $P[A\|\mathscr{G}]$ if it is shown that

$$\int_{E \times R^1} \varphi(x, y)\,dP(x, y) = P(A \cap (E \times R^1)). \tag{33.10}$$

Since $A = R^1 \times F$, the right side here is $P(E \times F)$. Since P has density f, Theorem 16.11 and Fubini's theorem reduce the left side to

$$\int_E \left\{ \int_{R^1} \varphi(x, y) f(x, y)\,dy \right\} dx = \int_E \left\{ \int_F f(x, t)\,dt \right\} dx$$

$$= \iint_{E \times F} f(x, y)\,dx\,dy = P(E \times F).$$

Thus (33.9) does give a version of $P[R^1 \times F\|\mathscr{G}]$. ■

The right side of (33.9) is the classical formula for the conditional probability of the event $R^1 \times F$ (the event that $y \in F$) given the event $\{x\} \times R^1$ (given the value of x). Since the event $\{x\} \times R^1$ has probability 0, the formula $P(A|B) = P(A \cap B)/P(B)$ does not work here. The whole point of this section is the systematic development of a notion of conditional probability that covers conditioning with respect to events of probability 0. This is accomplished by conditioning with respect to *collections* of events—that is, with respect to σ-fields $\mathscr{G}$.

Example 33.6. The set A is by definition independent of the σ-field $\mathscr{G}$ if it is independent of each G in $\mathscr{G}$: $P(A \cap G) = P(A)P(G)$. This being the same thing as $P(A \cap G) = \int_G P(A)\,dP$, A is independent of $\mathscr{G}$ if and only if $P[A\|\mathscr{G}] = P(A)$ with probability 1. ■

The σ-field $\sigma(X)$ generated by a random variable X consists of the sets $[\omega\colon X(\omega) \in H]$ for $H \in \mathscr{R}^1$; see Theorem 20.1. The conditional probability of A given X is defined as $P[A\|\sigma(X)]$ and is denoted $P[A\|X]$. Thus $P[A\|X] = P[A\|\sigma(X)]$ by definition. From the experiment corresponding to the σ-field $\sigma(X)$, one learns which of the sets $[\omega'\colon X(\omega') = x]$ contains ω and hence learns the value $X(\omega)$. Example 33.5 is a case of this: take $X(x, y) = x$ for (x, y) in the sample space $\Omega = R^2$ there.

This definition applies without change to random vector, or, equivalently, to a finite set of random variables. It can be adapted to arbitrary sets of random variables as well. For any such set $[X_t,\ t \in T]$, the σ-field $\sigma[X_t,\ t \in T]$ it generates is the smallest σ-field with respect to which each X_t is measurable. It is generated by the collection of sets of the form $[\omega\colon X_t(\omega) \in H]$ for t in T and H in $\mathscr{R}^1$. The *conditional probability* $P[A\|X_t,\ t \in T]$ *of* A *with respect to this set of random variables* is by definition the conditional probability $P[A\|\sigma[X_t,\ t \in T]]$ of A with respect to the σ-field $\sigma[X_t,\ t \in T]$.

In this notation the property (33.7) of Markov chains becomes

$$P[X_{n+1} = j\|X_0, \ldots, X_n] = P[X_{n+1} = j\|X_n]. \tag{33.11}$$

The conditional probability of $[X_{n+1} = j]$ is the same for someone who knows the present state X_n as for someone who knows the present state X_n and the past states $X_0, \ldots, X_{n-1}$ as well.

Example 33.7. Let X and Y be random vectors of dimensions j and k, let μ be the distribution of X over R^j, and suppose that X and Y are *independent*. According to (20.30),

$$P[X \in H, (X, Y) \in J] = \int_H P[(x, Y) \in J]\,\mu(dx)$$

for $H \in \mathscr{R}^j$ and $J \in \mathscr{R}^{j+k}$. This is a consequence of Fubini's theorem; it has a conditional-probability interpretation. For each x in R^j put

$$f(x) = P[(x, Y) \in J] = P[\omega'\colon (x, Y(\omega')) \in J]. \tag{33.12}$$

By Theorem 20.1(ii), $f(X(\omega))$ is measurable $\sigma(X)$, and since μ is the distribution of X, a change of variable gives

$$\int_{[X \in H]} f(X(\omega))P(d\omega) = \int_H f(x)\mu(dx) = P([(X, Y) \in J] \cap [X \in H]).$$

Since $[X \in H]$ is the general element of $\sigma(X)$, this proves that

$$f(X(\omega)) = P[(X,Y) \in J \| X]_\omega \tag{33.13}$$

with probability 1. ■

The fact just proved can be written

$$\begin{aligned} P[(X,Y) \in J \| X]_\omega &= P[(X(\omega), Y) \in J] \\ &= P[\omega' : (X(\omega), Y(\omega')) \in J]. \end{aligned}$$

Replacing ω' by ω on the right here causes a notational collision like the one replacing y by x causes in $\int_a^b f(x,y)\,dy$.

Suppose that X and Y are independent random variables and that Y has distribution function F. For $J = [(u,v)\colon \max\{u,v\} \le m]$, (33.12) is 0 for $m < x$ and $F(m)$ for $m \ge x$; if $M = \max\{X,Y\}$, then (33.13) gives

$$P[M \le m \| X]_\omega = I_{[X \le m]}(\omega) F(m) \tag{33.14}$$

with probability 1. All equations involving conditional probabilities must be qualified in this way by the phrase *with probability 1*, because the conditional probability is unique only to within a set of probability 0.

The following theorem is useful for checking conditional probabilities.

Theorem 33.1. *Let $\mathscr{P}$ be a π-system generating the σ-field $\mathscr{G}$, and suppose that Ω is a finite or countable union of sets in $\mathscr{P}$. An integrable function f is a version of $P[A \| \mathscr{G}]$ if it is measurable $\mathscr{G}$ and if*

$$\int_G f\,dP = P(A \cap G) \tag{33.15}$$

holds for all G in $\mathscr{P}$.

PROOF. Apply Theorem 10.4. ■

The condition that Ω is a finite or countable union of $\mathscr{P}$-sets cannot be suppressed; see Example 10.5.

Example 33.8. Suppose that X and Y are independent random variables with a common distribution function F that is positive and continuous. What is the conditional probability of $[X \le x]$ given the random variable $M = \max\{X,Y\}$? As it should clearly be 1 if $M \le x$, suppose that $M > x$. Since $X \le x$ requires $M = Y$, the chance of which is $\frac{1}{2}$ by symmetry, the conditional probability of $[X \le x]$ should by independence be $\frac{1}{2}F(x)/F(m) = \frac{1}{2}P[X \le x | X \le m]$ with the random variable M substituted

for m. Intuition thus gives

$$P[X \le x \| M]_\omega = I_{[M \le x]}(\omega) + \frac{1}{2} I_{[M > x]}(\omega) \frac{F(x)}{F(M(\omega))}. \tag{33.16}$$

It suffices to check (33.15) for sets $G = [M \le m]$, because these form a π-system generating $\sigma(M)$. The functional equation reduces to

$$P[M \le \min\{x, m\}] + \frac{1}{2} \int_{x < M \le m} \frac{F(x)}{F(M)}\, dP = P[M \le m,\ X \le x]. \tag{33.17}$$

Since the other case is easy, suppose that $x < m$. Since the distribution of (X, Y) is product measure, it follows by Fubini's theorem and the assumed continuity of F that

$$\int_{x < M \le m} \frac{1}{F(M)}\, dP = \iint_{\substack{u \le \iota \\ x < \iota \le m}} \frac{1}{F(v)}\, dF(u)\, dF(v) + \iint_{\substack{\iota \le u \\ x < u \le m}} \frac{1}{F(u)}\, dF(u)\, dF(v) = 2(F(m) - F(x)),$$

which gives (33.17). ■

Example 33.9. A collection $[X_t\colon t \ge 0]$ of random variables is a *Markov process in continuous time* if for $k \ge 1$, $0 \le t_1 \le \cdots \le t_k \le u$, and $H \in \mathscr{R}^1$,

$$P\left[X_u \in H \| X_{t_1}, \ldots, X_{t_k}\right] = P\left[X_u \in H \| X_{t_k}\right] \tag{33.18}$$

holds with probability 1. The analogue for discrete time is (33.11). (The X_n there have countable range as well, and the transition probabilities are constant in time, conditions that are not imposed here.)

Suppose that $t \le u$. Looking on the right side of (33.18) as a version of the conditional probability on the left shows that

$$\int_G P\left[X_u \in H \| X_t\right] dP = P([X_u \in H] \cap G) \tag{33.19}$$

if $0 \le t_1 \le \cdots \le t_k = t \le u$ and $G \in \sigma(X_{t_1}, \ldots, X_{t_k})$. Fix t, u, and H, and let k and $t_1, \ldots, t_k$ vary. Consider the class $\mathscr{P} = \bigcup \sigma(X_{t_1}, \ldots, X_{t_k})$, the union extending over all $k \ge 1$ and all k-tuples satisfying $0 \le t_1 \le \cdots \le t_k = t$. If $A \in \sigma(X_{t_1}, \ldots, X_{t_k})$ and $B \in \sigma(X_{s_1}, \ldots, X_{s_j})$, then $A \cap B \in \sigma(X_{r_1}, \ldots, X_{r_j})$, where the r_α are the s_β and the t_γ merged together. Thus $\mathscr{P}$ is a π-system. Since $\mathscr{P}$ generates $\sigma[X_s\colon s \le t]$ and $P[X_u \in H \| X_t]$ is measurable with respect to this σ-field, it follows by (33.19) and Theorem 33.1 that $P[X_u \in H \| X_t]$ is a version of $P[X_u \in H \| X_s,\ s \le t]$:

$$P\left[X_u \in H \| X_s,\ s \le t\right] = P\left[X_u \in H \| X_t\right], \qquad t \le u, \tag{33.20}$$

with probability 1.

This says that for calculating conditional probabilities about the future, the present $\sigma(X_t)$ is equivalent to the present and the *entire* past $\sigma[X_s: s \le t]$. This follows from the apparently weaker condition (33.18). ■

Example 33.10. The Poisson process $[N_t: t \ge 0]$ has independent increments (Section 23). Suppose that $0 \le t_1 \le \cdots \le t_k \le u$. The random vector $(N_{t_1}, N_{t_2} - N_{t_1}, \ldots, N_{t_k} - N_{t_{k-1}})$ is independent of $N_u - N_{t_k}$, and so (Theorem 20.2) $(N_{t_1}, N_{t_2}, \ldots, N_{t_k})$ is independent of $N_u - N_{t_k}$. If J is the set of points $(x_1, \ldots, x_k, y)$ in R^{k+1} such that $x_k + y \in H$, where $H \in \mathscr{R}^1$, and if ν is the distribution of $N_u - N_{t_k}$, then (33.12) is $P[(x_1, \ldots, x_k, N_u - N_{t_k}) \in J] = P[x_k + N_u - N_{t_k} \in H] = \nu(H - x_k)$. Therefore, (33.13) gives $P[N_u \in H \| N_{t_1}, \ldots, N_{t_k}] = \nu(H - N_{t_k})$. This holds also if $k = 1$, and hence $P[N_u \in H \| N_{t_1}, \ldots, N_{t_k}] = P[N_u \in H \| N_{t_k}]$. The Poisson process thus has the Markov property (33.18); this is a consequence solely of the independence of the increments. The extended Markov property (33.20) follows. ■

Properties of Conditional Probability

Theorem 33.2. *With probability* 1, $P[\varnothing \| \mathscr{G}] = 0$, $P[\Omega \| \mathscr{G}] = 1$; *and*

$$0 \le P[A \| \mathscr{G}] \le 1 \tag{33.21}$$

for each A. *If* $A_1, A_2, \ldots$ *is a finite or countable sequence of disjoint sets, then*

$$P\Big[\bigcup_n A_n \| \mathscr{G}\Big] = \sum_n P[A_n \| \mathscr{G}] \tag{33.22}$$

with probability 1.

PROOF. For each version of the conditional probability, $\int_G P[A\|\mathscr{G}]\,dP = P(A \cap G) \ge 0$ for each G in $\mathscr{G}$; since $P[A\|\mathscr{G}]$ is measurable $\mathscr{G}$, it must be nonnegative except on a set of P-measure 0. The other inequality in (33.21) is proved the same way.

If the A_n are disjoint and if G lies in $\mathscr{G}$, it follows (Theorem 16.6) that

$$\int_G \Big(\sum_n P[A_n\|\mathscr{G}]\Big)\,dP = \sum_n \int_G P[A_n\|\mathscr{G}]\,dP = \sum_n P(A_n \cap G)$$
$$= P\Big(\Big(\bigcup_n A_n\Big) \cap G\Big).$$

Thus $\sum_n P[A_n\|\mathscr{G}]$, which is certainly measurable $\mathscr{G}$, satisfies the functional equation for $P[\bigcup_n A_n\|\mathscr{G}]$, and so must coincide with it except perhaps on a set of P-measure 0. Hence (33.22). ■

Additional useful facts can be established by similar arguments. If $A \subset B$, then

$$P[B - A\|\mathscr{G}] = P[B\|\mathscr{G}] - P[A\|\mathscr{G}], \qquad P[A\|\mathscr{G}] \leq P[B\|\mathscr{G}]. \tag{33.23}$$

The inclusion–exclusion formula

$$P\left[\bigcup_{i=1}^{n} A_i\|\mathscr{G}\right] = \sum_i P[A_i\|\mathscr{G}] - \sum_{i<j} P[A_i \cap A_j\|\mathscr{G}] + \cdots \tag{33.24}$$

holds. If $A_n \uparrow A$, then

$$P[A_n\|\mathscr{G}] \uparrow P[A\|\mathscr{G}], \tag{33.25}$$

and if $A_n \downarrow A$, then

$$P[A_n\|\mathscr{G}] \downarrow P[A\|\mathscr{G}]. \tag{33.26}$$

Further, $P(A) = 1$ implies that

$$P[A\|\mathscr{G}] = 1, \tag{33.27}$$

and $P(A) = 0$ implies that

$$P[A\|\mathscr{G}] = 0. \tag{33.28}$$

Of course (33.23) through (33.28) hold with probability 1 only.

Difficulties and Curiosities

This section has been devoted almost entirely to examples connecting the abstract definition (33.8) with the probabilistic idea lying back of it. There are pathological examples showing that the interpretation of conditional probability in terms of an observer with partial information breaks down in certain cases.

Example 33.11. Let $(\Omega, \mathscr{F}, P)$ be the unit interval Ω with Lebesgue measure P on the σ-field $\mathscr{F}$ of Borel subsets of Ω. Take $\mathscr{G}$ to be the σ-field of sets that are either countable or cocountable. Then the function identically equal to $P(A)$ is a version of $P[A\|\mathscr{G}]$: (33.8) holds because $P(G)$ is either 0 or 1 for every G in $\mathscr{G}$. Therefore,

$$P[A\|\mathscr{G}]_\omega = P(A) \tag{33.29}$$

with probability 1. But since $\mathscr{G}$ contains all one-point sets, to know which

elements of $\mathscr{G}$ contain ω is to know ω itself. Thus $\mathscr{G}$ viewed as an experiment should be completely informative—the observer given the information in $\mathscr{G}$ should know ω exactly—and so one might expect that

$$P[A\|\mathscr{G}]_\omega = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases} \tag{33.30}$$

This is Example 4.10 in a new form. ■

The mathematical definition gives (33.29); the heuristic considerations lead to (33.30). Of course, (33.29) is right and (33.30) is wrong. The heuristic view breaks down in certain cases but is nonetheless illuminating and cannot, since it does not intervene in proofs, lead to any difficulties.

The point of view in this section has been "global." To each fixed A in $\mathscr{F}$ has been attached a function (usually a family of functions) $P[A\|\mathscr{G}]_\omega$ defined over all of Ω. What happens if the point of view is reversed—if ω is fixed and A varies over $\mathscr{F}$? Will this result in a probability measure on $\mathscr{F}$? Intuition says it should, and if it does, then (33.21) through (33.28) all reduce to standard facts about measures.

Suppose that $B_1, \ldots, B_r$ is a partition of Ω into $\mathscr{F}$-sets, and let $\mathscr{G} = \sigma(B_1, \ldots, B_r)$. If $P(B_1) = 0$ and $P(B_i) > 0$ for the other i, then one version of $P[A\|\mathscr{G}]$ is

$$P[A\|\mathscr{G}]_\omega = \begin{cases} 17 & \text{if } \omega \in B_1, \\ \dfrac{P(A \cap B_i)}{P(B_i)} & \text{if } \omega \in B_i,\ i = 2, \ldots, r. \end{cases}$$

With this choice of version for each A, $P[A\|\mathscr{G}]_\omega$ is, as a function of A, a probability measure on $\mathscr{F}$ if $\omega \in B_2 \cup \cdots \cup B_r$, but not if $\omega \in B_1$. The "wrong" versions have been chosen. If, for example,

$$P[A\|\mathscr{G}]_\omega = \begin{cases} P(A) & \text{if } \omega \in B_1, \\ \dfrac{P(A \cap B_i)}{P(B_i)} & \text{if } \omega \in B_i,\ i = 2, \ldots, r, \end{cases}$$

then $P[A\|\mathscr{G}]_\omega$ is a probability measure in A for each ω. Clearly, versions such as this one exist if $\mathscr{G}$ is finite.

It might be thought that for an arbitrary σ-field $\mathscr{G}$ in $\mathscr{F}$ versions of the various $P[A\|\mathscr{G}]$ can be so chosen that $P[A\|\mathscr{G}]_\omega$ is for each fixed ω a probability measure as A varies over $\mathscr{F}$. It is possible to construct a

counterexample showing that this is not so.† The example is possible because the exceptional ω-set of probability 0 where (33.22) fails depends on the sequence $A_1, A_2, \ldots$; if there are uncountably many such sequences, it can happen that the union of these exceptional sets has positive probability whatever versions $P[A\|\mathscr{G}]$ are chosen.

The existence of such pathological examples turns out not to matter. Example 33.9 illustrates the reason why. From the assumption (33.18) the notably stronger conclusion (33.20) was reached. Since the set $[X_u \in H]$ is fixed throughout the argument, it does not matter that conditional probabilities may not, in fact, be measures. What does matter for the theory is Theorem 33.2 and its extensions.

Consider a point ω_0 with the property that $P(G) > 0$ for every G in $\mathscr{G}$ that contains ω_0. This will be true if the one-point set $\{\omega_0\}$ lies in $\mathscr{F}$ and has positive probability. Fix any versions of the $P[A\|\mathscr{G}]$. For each A the set $[\omega: P[A\|\mathscr{G}]_\omega < 0]$ lies in $\mathscr{G}$ and has probability 0; it therefore cannot contain ω_0. Thus $P[A\|\mathscr{G}]_{\omega_0} \geq 0$. Similarly, $P[\Omega\|\mathscr{G}]_{\omega_0} = 1$, and, if the A_n are disjoint, $P[\bigcup_n A_n\|\mathscr{G}]_{\omega_0} = \sum_n P[A\|\mathscr{G}]_{\omega_0}$. Therefore, $P[A\|\mathscr{G}]_{\omega_0}$ is a probability measure as A ranges over $\mathscr{F}$.

Thus conditional probabilities behave like probabilities at points of positive probability. That they may not do so at points of probability 0 causes no problem because individual such points have no effect on the probabilities of sets. Of course, *sets* of points individually having probability 0 do have an effect, but here the global point of view reenters.

Conditional Probability Distributions

Let X be a random variable on $(\Omega, \mathscr{F}, P)$, and let $\mathscr{G}$ be a σ-field in $\mathscr{F}$.

Theorem 33.3. *There exists a function $\mu(H, \omega)$, defined for H in $\mathscr{R}^1$ and ω in Ω, with these two properties*:

(i) *For each ω in Ω, $\mu(\cdot, \omega)$ is a probability measure on $\mathscr{R}^1$.*
(ii) *For each H in $\mathscr{R}^1$, $\mu(H, \cdot)$ is a version of $P[X \in H\|\mathscr{G}]$.*

The probability measure $\mu(\cdot, \omega)$ is a *conditional distribution* of X given $\mathscr{G}$. If $\mathscr{G} = \sigma(Z)$, it is a conditional distribution of X given Z.

PROOF. For each rational r, let $F(r, \omega)$ be a version of $P[X \leq r\|\mathscr{G}]_\omega$. If $r \leq s$, then by (33.23),

$$F(r, \omega) \leq F(s, \omega) \tag{33.31}$$

†The argument is outlined in Problem 33.11. It depends on the construction of certain nonmeasurable sets.

for ω outside a $\mathscr{G}$-set A_{rs} of probability 0. By (33.26),

$$F(r,\omega) = \lim_n F(r+n^{-1},\omega) \tag{33.32}$$

for ω outside a $\mathscr{G}$-set B_r of probability 0. Finally, by (33.25) and (33.26),

$$\lim_{r\to-\infty} F(r,\omega) = 0, \qquad \lim_{r\to\infty} F(r,\omega) = 1 \tag{33.33}$$

outside a $\mathscr{G}$-set C of probability 0. As there are only countably many of these exceptional sets, their union E lies in $\mathscr{G}$ and has probability 0.

For $\omega \notin E$ extend $F(\cdot,\omega)$ to all of R^1 by setting $F(x,\omega) = \inf[F(r,\omega): x < r]$. For $\omega \in E$ take $F(x,\omega) = F(x)$, where F is some arbitrary but fixed distribution function. Suppose that $\omega \notin E$. By (33.31) and (33.32), $F(x,\omega)$ agrees with the first definition on the rationals and is nondecreasing; it is right-continuous; and by (33.33) it is a probability distribution function. Therefore, there exists a probability measure $\mu(\cdot,\omega)$ on $(R^1,\mathscr{R}^1)$ with distribution function $F(\cdot,\omega)$. For $\omega \in E$, let $\mu(\cdot,\omega)$ be the probability measure corresponding to $F(x)$. Then condition (i) is satisfied.

The class of H for which $\mu(H,\cdot)$ is measurable $\mathscr{G}$ is a λ-system containing the sets $H = (-\infty, r]$ for rational r; therefore $\mu(H,\cdot)$ is measurable $\mathscr{G}$ for H in $\mathscr{R}^1$.

By construction, $\mu((-\infty, r],\omega) = P[X \le r \| \mathscr{G}]_\omega$ with probability 1 for rational r; that is, for $H = (-\infty, r]$ as well as for $H = R^1$,

$$\int_G \mu(H,\omega) P(d\omega) = P([X \in H] \cap G)$$

for all G in $\mathscr{G}$. Fix G. Each side of this equation is a measure as a function of H, and so the equation must hold for all H in $\mathscr{R}^1$. ■

Example 33.12. Let X and Y be random variables whose joint distribution ν in R^2 has density $f(x,y)$ with respect to Lebesgue measure: $P[(X,Y) \in A] = \nu(A) = \iint_A f(x,y)\,dx\,dy$. Let $g(x,y) = f(x,y)/\int_{R^1} f(x,t)\,dt$, and let $\mu(H,x) = \int_H g(x,y)\,dy$ have probability density $g(x,\cdot)$; if $\int_{R^1} f(x,t)\,dt = 0$, let $\mu(\cdot,x)$ be an arbitrary probability measure on the line. Then $\mu(H, X(\omega))$ will serve as the conditional distribution of Y given X. Indeed, (33.10) is the same thing as $\int_{E\times R^1} \mu(F,x)\,d\nu(x,y) = \nu(E\times F)$, and a change of variable gives $\int_{[X\in E]} \mu(F, X(\omega)) P(d\omega) = P[X \in E, Y \in F]$. Thus $\mu(F, X(\omega))$ is a version of $P[Y \in F \| X]_\omega$. This is a new version of Example 33.5. ■