(c) A partition (of arbitrary cardinality) of Ω.

(d) The level sets of $\sin x$ ($\Omega = R^1$).

(e) The σ-field in Problem 3.5.

4.9. 2.9 2.10↑ In connection with Example 4.8 and Problem 2.10, prove these facts:

(a) Every set in $\sigma(\mathscr{A})$ is a union of $\mathscr{A}$-equivalence classes.

(b) If $\mathscr{A} = [A_\theta: \theta \in \Theta]$, then the $\mathscr{A}$-equivalence classes have the form $\bigcap_\theta B_\theta$, where for each θ, B_θ is A_θ or A_θ^c.

(c) Every finite σ-field is generated by a finite partition of Ω.

(d) If $\mathscr{F}_0$ is a field, then each singleton, even each finite set, in $\sigma(\mathscr{F}_0)$ is a countable intersection of $\mathscr{F}_0$-sets.

4.10. 3.2↑ There is in the unit interval a set H that is nonmeasurable in the extreme sense that its inner and outer Lebesgue measures are 0 and 1 (see (3.9) and (3.10)): $\lambda_*(H) = 0$ and $\lambda^*(H) = 1$. See Problem 12.4 for the construction.

Let $\Omega = (0, 1]$, let $\mathscr{G}$ consist of the Borel sets in Ω, and let H be the set just described. Show that the class $\mathscr{F}$ of sets of the form $(H \cap G_1) \cup (H^c \cap G_2)$ for G_1 and G_2 in $\mathscr{G}$ is a σ-field and that $P[(H \cap G_1) \cup (H^c \cap G_2)] = \frac{1}{2}\lambda(G_1) + \frac{1}{2}\lambda(G_2)$ consistently defines a probability measure on $\mathscr{F}$. Show that $P(H) = \frac{1}{2}$ and that $P(G) = \lambda(G)$ for $G \in \mathscr{G}$. Show that $\mathscr{G}$ is generated by a countable subclass (see Problem 2.11). Show that $\mathscr{G}$ contains all the singletons and that H and $\mathscr{G}$ are independent.

The construction proves this: *There exist a probability space $(\Omega, \mathscr{F}, P)$, a σ-field $\mathscr{G}$ in $\mathscr{F}$, and a set H in $\mathscr{F}$, such that $P(H) = \frac{1}{2}$, H and $\mathscr{G}$ are independent, and $\mathscr{G}$ is generated by a countable subclass and contains all the singletons.*

Example 4.10 is somewhat similar, but there the σ-field $\mathscr{G}$ is not countably generated and each set in it has probability either 0 or 1. In the present example $\mathscr{G}$ is countably generated and $P(G)$ assumes every value between 0 and 1 as G ranges over $\mathscr{G}$. Example 4.10 is to some extent unnatural because the $\mathscr{G}$ there is not countably generated. The present example, on the other hand, involves the pathological set H. This example is used in Section 33 in connection with conditional probability; see Problem 33.11.

4.11. (a) If $A_1, A_2, \ldots$ are independent events, then $P(\bigcap_{n=1}^\infty A_n) = \prod_{n=1}^\infty P(A_n)$ and $P(\bigcup_{n=1}^\infty A_n) = 1 - \prod_{n=1}^\infty (1 - P(A_n))$. Prove these facts and from them derive the second Borel–Cantelli lemma by the well-known relation between infinite series and products.

(b) Show that $P(\limsup_n A_n) = 1$ if for each k the series $\sum_{n>k} P(A_n | A_k^c \cap \cdots \cap A_{n-1}^c)$ diverges. From this deduce the second Borel–Cantelli lemma once again.

(c) Show by example that $P(\limsup_n A_n) = 1$ does not follow from the divergence of $\sum_n P(A_n | A_1^c \cap \cdots \cap A_{n-1}^c)$ alone.

(d) Show that $P(\limsup_n A_n) = 1$ if and only if $\sum_n P(A \cap A_n)$ diverges for each A of positive probability.

(e) If sets A_n are independent and $P(A_n) < 1$ for all n, then $P[A_n \text{ i.o.}] = 1$ if and only if $P(\bigcup_n A_n) = 1$.

4.12. (a) Show (see Example 4.21) that $\log_2 n + \log_2 \log_2 n + \theta \log_2 \log_2 \log_2 n$ is an outer boundary if $\theta > 1$. Generalize.

(b) Show that $\log_2 n + \log_2 \log_2 \log_2 n$ is an inner boundary.

4.13. Let φ be a positive function of integers, and define B_φ as the set of x in $(0, 1)$ such that $|x - p/2^i| < 1/2^i\varphi(2^i)$ holds for infinitely many pairs p, i. Adapting the proof of Theorem 1.6, show directly (without reference to Example 4.12) that $\sum_i 1/\varphi(2^i) < \infty$ implies $\lambda(B_\varphi) = 0$.

4.14. 2.19↑ Suppose that there are in $(\Omega, \mathscr{F}, P)$ independent events $A_1, A_2, \ldots$ such that, if $\alpha_n = \min\{P(A_n), 1 - P(A_n)\}$, then $\sum \alpha_n = \infty$. Show that P is nonatomic.

4.15. 2.18↑ Let F be the set of square-free integers—those integers not divisible by any perfect square. Let F_l be the set of m such that $p^2 | m$ for no $p \le l$, and show that $D(F_l) = \prod_{p \le l}(1 - p^{-2})$. Show that $P_n(F_l - F) \le \sum_{p > l} p^{-2}$, and conclude that the square-free integers have density $\prod_p (1 - p^{-2}) = 6/\pi^2$.

4.16. 2.18↑ Reconsider Problem 2.18(d). If D were countably additive on $f(\mathscr{M})$, it would extend to $\sigma(\mathscr{M})$. Use the second Borel–Cantelli lemma.

SECTION 5. SIMPLE RANDOM VARIABLES

Definition

Let $(\Omega, \mathscr{F}, P)$ be an arbitrary probability space, and let X be a real-valued function on Ω; X is a *simple random variable* if it has finite range (assumes only finitely many values) and if

$$[\omega: X(\omega) = x] \in \mathscr{F} \tag{5.1}$$

for each real x. (Of course, $[\omega: X(\omega) = x] = \varnothing \in \mathscr{F}$ for x outside the range of X.) Whether or not X satisfies this condition depends only on $\mathscr{F}$, not on P, but the point of the definition is to ensure that the probabilities $P[\omega: X(\omega) = x]$ are defined. Later sections will treat the theory of general random variables, of functions on Ω having arbitrary range; (5.1) will require modification in the general case.

The $d_n(\omega)$ of the preceding section (the digits of the dyadic expansion) are simple random variables on the unit interval: the sets $[\omega: d_n(\omega) = 0]$ and $[\omega: d_n(\omega) = 1]$ are finite unions of subintervals and hence lie in the σ-field $\mathscr{B}$ of Borel sets in $(0, 1]$. The Rademacher functions are also simple random variables. Although the concept itself is thus not entirely new, to proceed further in probability requires a systematic theory of random variables and their expected values.

The run lengths $l_n(\omega)$ satisfy (5.1) but are not simple random variables, because they have infinite range (they come under the general theory). In a discrete space, $\mathscr{F}$ consists of all subsets of Ω, so that (5.1) always holds.

It is customary in probability theory to omit the argument ω. Thus X stands for a general value $X(\omega)$ of the function as well as for the function itself, and $[X = x]$ is short for $[\omega: X(\omega) = x]$

A finite sum

(5.2) $$X=\sum_i x_i I_{A_i}$$

is a random variable if the A_i form a finite partition of Ω into $\mathcal{F}$-sets. Moreover, every simple random variable can be represented in the form (5.2): for the x_i take the range of X, and put $A_i=[X=x_i]$. But X may have other such representations because $x_iI_{A_i}$ can be replaced by $\sum_j x_i I_{A_{ij}}$ if the A_{ij} form a finite decomposition of A_i into $\mathcal{F}$-sets.

If $\mathcal{G}$ is a sub-σ-field of $\mathcal{F}$, a simple random variable X is *measurable* $\mathcal{G}$, or *measurable with respect to* $\mathcal{G}$, if $[X=x]\in\mathcal{G}$ for each x. A simple random variable is by definition always measurable $\mathcal{F}$. Since $[X\in H]=\bigcup[X=x]$, where the union extends over the finitely many x lying both in H and in the range of X, $[X\in H]\in\mathcal{G}$ for every $H\subset R^1$ if X is a simple random variable variable measurable $\mathcal{G}$.

The σ-field $\sigma(X)$ *generated by* X is the smallest σ-field with respect to which X is measurable; that is, $\sigma(X)$ is the intersection of all σ-fields with respect to which X is measurable. For a finite or infinite sequence $X_1,X_2,\ldots$ of simple random variables, $\sigma(X_1,X_2,\ldots)$ is the smallest σ-field with respect to which *each* X_i is measurable. It can be described explicitly in the finite case:

Theorem 5.1. *Let $X_1,\ldots,X_n$ be simple random variables.*

(i) *The σ-field $\sigma(X_1,\ldots,X_n)$ consists of the sets*

(5.3) $$[(X_1,\ldots,X_n)\in H]=[\omega:(X_1(\omega),\ldots,X_n(\omega))\in H]$$

for $H\subset R^n$; H in this representation may be taken finite.

(ii) *A simple random variable Y is measurable $\sigma(X_1,\ldots,X_n)$ if and only if*

(5.4) $$Y=f(X_1,\ldots,X_n)$$

for some $f: R^n\to R^1$.

PROOF. Let $\mathcal{M}$ be the class of sets of the form (5.3). Sets of the form $[(X_1,\ldots,X_n)=(x_1,\ldots,x_n)]=\bigcap_{i=1}^n[X_i=x_i]$ must lie in $\sigma(X_1,\ldots,X_n)$; each set (5.3) is a finite union of sets of this form because $(X_1,\ldots,X_n)$, as a mapping from Ω to R^n, has firite range. Thus $\mathcal{M}\subset\sigma(X_1,\ldots,X_n)$.

On the other hand, $\mathcal{M}$ is a σ-field because $\Omega=[(X_1,\ldots,X_n)\in R^n]$, $[(X_1,\ldots,X_n)\in H]^c=[(X_1,\ldots,X_n)\in H^c]$, and $\bigcup_j[(X_1,\ldots,X_n)\in H_j]=[(X_1,\ldots,X_n)\in\bigcup_jH_j]$. But each X_i is measurable with respect to $\mathcal{M}$, because $[X_i=x]$ can be put in the form (5.3) by taking H to consist of those $(x_1,\ldots,x_n)$ in R^n for which $x_i=x$. It follows that $\sigma(X_1,\ldots,X_n)$ is contained in $\mathcal{M}$ and therefore equals $\mathcal{M}$. As intersecting H with the range (finite) of $(X_1,\ldots,X_n)$ in R^n does not affect (5.3), H may be taken finite. This proves (i).

Assume that Y has the form (5.4)—that is, $Y(\omega)=f(X_1(\omega),\ldots,X_n(\omega))$ for every ω. Since $[Y=y]$ can be put in the form (5.3) by taking H to consist of those $x=(x_1,\ldots,x_n)$ for which $f(x)=y$, it follows that Y is measurable $\sigma(X_1,\ldots,X_n)$.

Now assume that Y is measurable $\sigma(X_1,\ldots,X_n)$. Let $y_1,\ldots,y_r$ be the distinct values Y assumes. By part (i), there exist sets $H_1,\ldots,H_r$ in R^n such that

$$[\omega: Y(\omega)=y_i]=[\omega:(X_1(\omega),\ldots,X_n(\omega))\in H_i].$$

Take $f=\sum_{i=1}^r y_iI_{H_i}$. Although the H_i need not be disjoint, if H_i and H_j share a point of the form $(X_1(\omega),\ldots,X_n(\omega))$, then $Y(\omega)=y_i$ and $Y(\omega)=y_j$, which is impossible if $i\neq j$. Therefore each $(X_1(\omega),\ldots,X_n(\omega))$ lies in exactly one of the H_i, and it follows that $f(X_1(\omega),\ldots,X_n(\omega))=Y(\omega)$. ■

Since (5.4) implies that Y is measurable $\sigma(X_1,\ldots,X_n)$, it follows in particular that functions of simple random variables are again simple random variables. Thus X^2, e^{tX}, and so on are simple random variables along with X. Taking f to be $\sum_{i=1}^n x_i$, $\prod_{i=1}^n x_i$, or $\max_{i\le n} x_i$ shows that sums, products, and maxima of simple random variables are simple random variables.

As explained on p. 57, a sub-σ-field corresponds to partial information about ω. From this point of view, $\sigma(X_1,\ldots,X_n)$ corresponds to a knowledge of the values $X_1(\omega),\ldots,X_n(\omega)$. These values suffice to determine the value $Y(\omega)$ if and only if (5.4) holds. The elements of the $\sigma(X_1,\ldots,X_n)$-partition (see (4.16)) are the sets $[X_1=x_1,\ldots,X_n=x_n]$ for x_i in the range of X_i.

Example 5.1. For the dyadic digits $d_n(\omega)$ on the unit interval, d_3 is not measurable $\sigma(d_1,d_2)$; indeed, there exist ω' and ω'' such that $d_1(\omega')=d_1(\omega'')$ and $d_2(\omega')=d_2(\omega'')$ but $d_3(\omega')\neq d_3(\omega'')$, an impossibility if $d_3(\omega)=f(d_1(\omega),d_2(\omega))$ identically in ω. If such an f existed, one could unerringly predict the outcome $d_3(\omega)$ of the third toss from the outcomes $d_1(\omega)$ and $d_2(\omega)$ of the first two. ■

Example 5.2. Let $s_n(\omega)=\sum_{k=1}^n r_k(\omega)$ be the partial sums of the Rademacher functions—see (1.14). By Theorem 5.1(ii) s_k is measurable $\sigma(r_1,\ldots,r_n)$ for $k\le n$, and $r_k=s_k-s_{k-1}$ is measurable $\sigma(s_1,\ldots,s_n)$ for $k\le n$. Thus $\sigma(r_1,\ldots,r_n)=\sigma(s_1,\ldots,s_n)$. In random-walk terms, the first n positions contain the same information as the first n distances moved. In gambling terms, to know the gambler's first n fortunes (relative to his initial fortune) is the same thing as to know his gains and losses on each of the first n plays. ■

Example 5.3. An indicator I_A is measurable $\mathscr{G}$ if and only if A lies in $\mathscr{G}$. And $A \in \sigma(A_1, \dots, A_n)$ if and only if $I_A = f(I_{A_1}, \dots, I_{A_n})$ for some $f\colon R^n \to R^1$. ■

Convergence of Random Variables

It is a basic problem, for given random variables X and $X_1, X_2, \dots$ on a probability space $(\Omega, \mathscr{F}, P)$, to look for the probability of the event that $\lim_n X_n(\omega) = X(\omega)$. The normal number theorem is an example, one where the probability is 1. It is convenient to characterize the complementary event: $X_n(\omega)$ fails to converge to $X(\omega)$ if and only if there is some ϵ such that for no m does $|X_n(\omega) - X(\omega)|$ remain below ϵ for all n exceeding m—that is to say, if and only if, for some ϵ, $|X_n(\omega) - X(\omega)| \ge \epsilon$ holds for infinitely many values of n. Therefore,

$$\text{(5.5)} \qquad \left[\lim_n X_n = X\right]^c = \bigcup_\epsilon \left[|X_n - X| \ge \epsilon \text{ i.o.}\right],$$

where the union can be restricted to rational (positive) ϵ because the set in the union increases as ϵ decreases (compare (2.2)).

The event $[\lim_n X_n = X]$ therefore always lies in the basic σ-field $\mathscr{F}$, and it has probability 1 if and only if

$$\text{(5.6)} \qquad P\left[|X_n - X| \ge \epsilon \text{ i.o.}\right] = 0$$

for each ϵ (rational or not). The event in (5.6) is the limit superior of the events $[|X_n - X| \ge \epsilon]$, and it follows by Theorem 4.1 that (5.6) implies

$$\text{(5.7)} \qquad \lim_n P\left[|X_n - X| \ge \epsilon\right] = 0.$$

This leads to a definition: If (5.7) holds for each positive ϵ, then X_n is said to *converge to X in probability*, written $X_n \to_P X$.

These arguments prove two facts:

Theorem 5.2. (i) *There is convergence* $\lim_n X_n = X$ *with probability* 1 *if and only if* (5.6) *holds for each* ϵ.

(ii) *Convergence with probability* 1 *implies convergence in probability.*

Theorem 1.2, the normal number theorem, has to do with the convergence with probability 1 of $n^{-1}\sum_{i=1}^n d_i(\omega)$ to $\frac{1}{2}$. Theorem 1.1 has to do instead with the convergence in probability of the same sequence. By Theorem 5.2(ii), then, Theorem 1.1 is a consequence of Theorem 1.2 (see (1.30) and (1.31)). The converse is not true, however—convergence in probability does not imply convergence with probability 1:

Example 5.4. Take $X \equiv 0$ and $X_n = I_{A_n}$. Then $X_n \to_P X$ is equivalent to $P(A_n) \to 0$, and $[\lim_n X_n = X]^c = [A_n \text{ i.o.}]$. Any sequence $\{A_n\}$ such that $P(A_n) \to 0$ but $P[A_n \text{ i.o.}] > 0$ therefore gives a counterexample to the converse to Theorem 5.2(ii).

Consider the event $A_n = [\omega\colon l_n(\omega) \ge \log_2 n]$ in Example 4.15. Here, $P(A_n) \le 1/n \to 0$, while $P[A_n \text{ i.o.}] = 1$ by (4.26), and so this is one counterexample. For an example more extreme and more transparent, define events in the unit interval in the following way. Define the first two by

$$A_1 = (0, \tfrac{1}{2}], \qquad A_2 = (\tfrac{1}{2}, 1].$$

Define the next four by

$$A_3 = (0, \tfrac{1}{4}], \qquad A_4 = (\tfrac{1}{4}, \tfrac{1}{2}], \qquad A_5 = (\tfrac{1}{2}, \tfrac{3}{4}], \qquad A_6 = (\tfrac{3}{4}, 1].$$

Define the next eight, $A_7, \dots, A_{14}$, as the dyadic intervals of rank 3. And so on. Certainly, $P(A_n) \to 0$, and since each point ω is covered by one set in each successive block of length 2^k, the set $[A_n \text{ i.o.}]$ is all of $(0, 1]$. ■

Independence

A sequence $X_1, X_2, \dots$ (finite or infinite) of simple random variables is by definition *independent* if the classes $\sigma(X_1), \sigma(X_2), \dots$ are independent in the sense of the preceding section. By Theorem 5.1(i), $\sigma(X_i)$ consists of the sets $[X_i \in H]$ for $H \subset R^1$. The condition for independence of $X_1, \dots, X_n$ is therefore that

$$\text{(5.8)} \qquad P[X_1 \in H_1, \dots, X_n \in H_n] = P[X_1 \in H_1] \cdots P[X_n \in H_n]$$

for linear sets $H_1, \dots, H_n$. The definition (4.10) also requires that (5.8) hold if one or more of the $[X_i \in H_i]$ is suppressed; but taking H_i to be R^1 eliminates it from each side. For an infinite sequence $X_1, X_2, \dots$, (5.8) must hold for each n. A special case of (5.8) is

$$\text{(5.9)} \qquad P[X_1 = x_1, \dots, X_n = x_n] = P[X_1 = x_1] \cdots P[X_n = x_n].$$

On the other hand, summing (5.9) over $x_1 \in H_1, \dots, x_n \in H_n$ gives (5.8). Thus the X_i are independent if and only if (5.9) holds for all $x_1, \dots, x_n$.

Suppose that

$$\text{(5.10)} \qquad \begin{array}{ccc} X_{11} & X_{12} & \cdots \\ X_{21} & X_{22} & \cdots \\ \vdots & \vdots & \end{array}$$

is an independent array of simple random variables. There may be finitely or infinitely many rows, each row finite or infinite. If $\mathscr{A}_i$ consists of the finite intersections $\bigcap_j [X_{ij} \in H_j]$ with $H_j \subset R^1$, an application of Theorem 4.2 shows that the σ-fields $\sigma(X_{i1}, X_{i2}, \ldots)$, $i = 1, 2, \ldots$ are independent. As a consequence, $Y_1, Y_2, \ldots$ are independent if Y_i is measurable $\sigma(X_{i1}, X_{i2}, \ldots)$ for each i.

Example 5.5. The dyadic digits $d_1(\omega), d_2(\omega), \ldots$ on the unit interval are an independent sequence of random variables for which

$$P[d_n = 0] = P[d_n = 1] = \tfrac{1}{2}. \tag{5.11}$$

It is because of (5.11) and independence that the d_n give a model for tossing a fair coin.

The sequence $(d_1(\omega), d_2(\omega), \ldots)$ and the point ω determine one another. It can be imagined that ω is determined by the outcomes $d_n(\omega)$ of a sequence of tosses. It can also be imagined that ω is the result of drawing a point at random from the unit interval, and that ω determines the $d_n(\omega)$. In the second interpretation the $d_n(\omega)$ are all determined the instant ω is drawn, and so it should further be imagined that they are then revealed to the coin tosser or gambler one by one. For example, $\sigma(d_1, d_2)$ corresponds to knowing the outcomes of the first two tosses—to knowing not ω but only $d_1(\omega)$ and $d_2(\omega)$—and this does not help in predicting the value $d_3(\omega)$, because $\sigma(d_1, d_2)$ and $\sigma(d_3)$ are independent. See Example 5.1. ■

Example 5.6. Every permutation can be written as a product of cycles. For example,

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 5 & 1 & 7 & 4 & 6 & 2 & 3 \end{pmatrix} = (1562)(37)(4).$$

This permutation sends 1 to 5, 2 to 1, 3 to 7, and so on. The cyclic form on the right shows that 1 goes to 5, which goes to 6, which goes to 2, which goes back to 1; and so on. To standardize this cyclic representation, start the first cycle with 1 and each successive cycle with the smallest integer not yet encountered.

Let Ω consist of the $n!$ permutations of $1, 2, \ldots, n$, all equally probable; $\mathscr{F}$ contains all subsets of Ω, and $P(A)$ is the fraction of points in A. Let $X_k(\omega)$ be 1 or 0 according as the element in the kth position in the cyclic representation of the permutation ω completes a cycle or not. Then $S(\omega) = \sum_{k=1}^{n} X_k(\omega)$ is the number of cycles in ω. In the example above, $n = 7$, $X_1 = X_2 = X_3 = X_5 = 0$, $X_4 = X_6 = X_7 = 1$, and $S = 3$. The following argument shows that $X_1, \ldots, X_n$ are independent and $P[X_k = 1] = 1/(n - k + 1)$. This will lead later on to results on $P[S \in H]$.

The idea is this: $X_1(\omega) = 1$ if and only if the random permutation ω sends 1 to itself, the probability of which is $1/n$. If it happens that $X_1(\omega) = 1$—that ω fixes 1—then the image of 2 is one of $2, \ldots, n$, and $X_2(\omega) = 1$ if and only if this image is in fact 2; the conditional probability of this is $1/(n-1)$. If $X_1(\omega) = 0$, on the other hand, then ω sends 1 to some $i \neq 1$, so that the image of i is one of $1, \ldots, i-1, i+1, \ldots, n$, and $X_2(\omega) = 1$ if and only if this image is in fact 1; the conditional probability of this is again $1/(n-1)$. This argument generalizes.

But the details are fussy. Let $Y_1(\omega), \ldots, Y_n(\omega)$ be the integers in the successive positions in the cyclic representation of ω. Fix k, and let A_v be the set where $(X_1, \ldots, X_{k-1}, Y_1, \ldots, Y_k)$ assumes a specific vector of values $v = (x_1, \ldots, x_{k-1}, y_1, \ldots, y_k)$. The A_v form a partition $\mathscr{A}$ of Ω, and if $P[X_k = 1 | A_v] = 1/(n-k+1)$ for each v, then by Example 4.7, $P[X_k = 1] = 1/(n-k+1)$ and X_k is independent of $\sigma(\mathscr{A})$ and hence of the smaller σ-field $\sigma(X_1, \ldots, X_{k-1})$. It will follow by induction that $X_1, \ldots, X_n$ are independent.

Let j be the position of the rightmost 1 among $x_1, \ldots, x_{k-1}$ ($j = 0$ if there are none). Then ω lies in A_v if and only if it permutes $y_1, \ldots, y_j$ among themselves (in a way specified by the values $x_1, \ldots, x_{j-1}$, $x_j = 1$, $y_1, \ldots, y_j$) and sends each of $y_{j+1}, \ldots, y_{k-1}$ to the y just to its right. Thus A_v contains $(n-k+1)!$ sample points. And $X_k(\omega) = 1$ if and only if ω also sends y_k to y_{j+1}. Thus $A_v \cap [X_k = 1]$ contains $(n-k)!$ sample points, and so the conditional probability of $X_k = 1$ is $1/(n-k+1)$. ■

Existence of Independent Sequences

The *distribution* of a simple random variable X is the probability measure μ defined for all subsets A of the line by

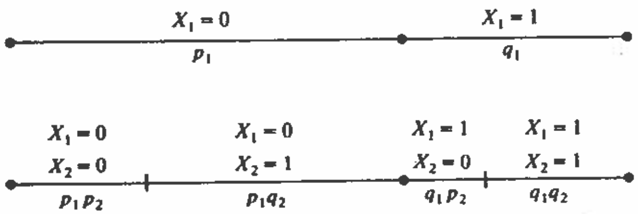
$$\mu(A) = P[X \in A]. \tag{5.12}$$

This does define a probability measure. It is discrete in the sense of Example 2.9: If $x_1, \ldots, x_l$ are the distinct points of the range of X, then μ has mass $p_i = P[X = x_i] = \mu\{x_i\}$ at x_i, and $\mu(A) = \sum p_i$, the sum extending over those i for which $x_i \in A$. As $\mu(A) = 1$ if A is the range of X, not only is μ discrete, it has finite support.

Theorem 5.3. *Let $\{\mu_n\}$ be a sequence of probability measures on the class of all subsets of the line, each having finite support. There exists on some probability space $(\Omega, \mathscr{F}, P)$ an independent sequence $\{X_n\}$ of simple random variables such that X_n has distribution μ_n.*

What matters here is that there are finitely or countably many distributions μ_n. They need not be indexed by the integers; any countable index set will do.

PROOF. The probability space will be the unit interval. To understand the construction, consider first the case in which each μ_n concentrates its mass on the two points 0 and 1. Put $p_n = \mu_n\{0\}$ and $q_n = 1 - p_n = \mu_n\{1\}$. Split $(0,1]$ into two intervals I_0 and I_1 of lengths p_1 and q_1. Define $X_1(\omega) = 0$ for $\omega \in I_0$ and $X_1(\omega) = 1$ for $\omega \in I_1$. If P is Lebesgue measure, then clearly $P[X_1 = 0] = p_1$ and $P[X_1 = 1] = q_1$, so that X_1 has distribution μ_1.



Now split I_0 into two intervals I_{00} and I_{01} of lengths p_1p_2 and p_1q_2, and split I_1 into two intervals I_{10} and I_{11} of lengths q_1p_2 and q_1q_2. Define $X_2(\omega) = 0$ for $\omega \in I_{00} \cup I_{10}$ and $X_2(\omega) = 1$ for $\omega \in I_{01} \cup I_{11}$. As the diagram makes clear, $P[X_1 = 0,\ X_2 = 0] = p_1p_2$, and similarly for the other three possibilities. It follows that X_1 and X_2 are independent and X_2 has distribution μ_2. Now X_3 is constructed by splitting each of $I_{00}, I_{01}, I_{10}, I_{11}$ in the proportions p_3 and q_3. And so on.

If $p_n = q_n = \frac{1}{2}$ for all n, then the successive decompositions here are the decompositions of $(0,1]$ into dyadic intervals, and $X_n(\omega) = d_n(\omega)$.

The argument for the general case is not very different. Let $x_{n1}, \ldots, x_{nl_n}$ be the distinct points on which μ_n concentrates its mass, and put $p_{ni} = \mu_n\{x_{ni}\}$ for $1 \le i \le l_n$.

Decompose[1] $(0,1]$ into l_1 subintervals $I_1^{(1)}, \ldots, I_{l_1}^{(1)}$ of respective lengths $p_{11}, \ldots, p_{1l_1}$. Define X_1 by setting $X_1(\omega) = x_{1i}$ for $\omega \in I_i^{(1)}$, $1 \le i \le l_1$. Then (P is Lebesgue measure) $P[\omega\colon X_1(\omega) = x_{1i}] = P(I_i^{(1)}) = p_{1i}$, $1 \le i \le l_1$. Thus X_1 is a simple random variable with distribution μ_1.

Next decompose each $I_i^{(1)}$ into l_2 subintervals $I_{i1}^{(2)}, \ldots, I_{il_2}^{(2)}$ of respective lengths $p_{1i}p_{21}, \ldots, p_{1i}p_{2l_2}$. Define $X_2(\omega) = x_{2j}$ for $\omega \in \bigcup_{i=1}^{l_1} I_{ij}^{(2)}$, $1 \le j \le l_2$. Then $P[\omega\colon X_1(\omega) = x_{1i},\ X_2(\omega) = x_{2j}] = P(I_{ij}^{(2)}) = p_{1i}p_{2j}$. Adding out i shows that $P[\omega\colon X_2(\omega) = x_{2j}] = p_{2j}$, as required. Hence $P[X_1 = x_{1i},\ X_2 = x_{2j}] = p_{1i}p_{2j} = P[X_1 = x_{1i}]P[X_2 = x_{2j}]$, and X_1 and X_2 are independent.

The construction proceeds inductively. Suppose that $(0,1]$ has been decomposed into $l_1 \cdots l_n$ intervals

$$I_{i_1 \ldots i_n}^{(n)}, \qquad 1 \le i_1 \le l_1, \ldots, 1 \le i_n \le l_n, \tag{5.13}$$

of lengths

$$P\left(I_{i_1 \ldots i_n}^{(n)}\right) = p_{1,i_1} \cdots p_{n,i_n}. \tag{5.14}$$

Decompose $I_{i_1 \ldots i_n}^{(n)}$ into l_{n+1} subintervals $I_{i_1 \ldots i_n 1}^{(n+1)}, \ldots, I_{i_1 \ldots i_n l_{n+1}}^{(n+1)}$ of respective lengths $P(I_{i_1 \ldots i_n}^{(n)})p_{n+1,1}, \ldots, P(I_{i_1 \ldots i_n}^{(n)})p_{n+1,l_{n+1}}$. These are the intervals of the next decomposition. This construction gives a sequence of decompositions (5.13) of $(0,1]$ into subintervals; each decomposition satisfies (5.14), and each refines the preceding one. If μ_n is given for $1 \le n \le N$, the procedure terminates after N steps; for an infinite sequence it does not terminate at all.

For $1 \le i \le l_n$, put $X_n(\omega) = x_{ni}$ if $\omega \in \bigcup_{i_1 \ldots i_{n-1}} I_{i_1 \ldots i_{n-1} i}^{(n)}$. Since each decomposition (5.13) refines the preceding, $X_k(\omega) = x_{ki_k}$ for $\omega \in I_{i_1 \ldots i_k \ldots i_n}^{(n)}$. Therefore, each element of (5.13) is contained in the element with the same label $i_1 \ldots i_n$ in the decomposition

$$A_{i_1 \ldots i_n} = \left[\omega\colon X_1(\omega) = x_{1i_1}, \ldots, X_n(\omega) = x_{ni_n}\right], \quad 1 \le i_1 \le l_1, \ldots, \quad 1 \le i_n \le l_n.$$

The two decompositions thus coincide, and it follows by (5.14) that $P[X_1 = x_{1i_1}, \ldots, X_n = x_{ni_n}] = p_{1,i_1} \cdots p_{n,i_n}$. Adding out the indices $i_1, \ldots, i_{n-1}$ shows that X_n has distribution μ_n and hence that $X_1, \ldots, X_n$ are independent. But n was arbitrary. ■

In the case where the μ_n are all the same, there is an alternative construction based on probabilities in sequence space. Let S be the support (finite) common to the μ_n, and let p_u, $u \in S$, be the probabilities common to the μ_n. In sequence space S^∞, define product measure P on the class $\mathscr{C}_0$ of cylinders by (2.21). By Theorem 2.3, P is countably additive on $\mathscr{C}_0$, and by Theorem 3.1 it extends to $\mathscr{C} = \sigma(\mathscr{C}_0)$. The coordinate functions $z_k(\cdot)$ are random variables on the probability space $(S^\infty, \mathscr{C}, P)$; take these as the X_k. Then (2.22) translates into $P[X_1 = u_1, \ldots, X_n = u_n] = p_{u_1} \cdots p_{u_n}$, which is just what Theorem 5.3 requires in this special case.

Probability theorems such as those in the next sections concern independent sequences $\{X_n\}$ with specified distributions or with distributions having specified properties, and because of Theorem 5.3 these theorems are true not merely in the vacuous sense that their hypotheses are never fulfilled. Similar but more complicated existence theorems will come later. For most purposes the probability space on which the X_n are defined is largely irrelevant. Every independent sequence $\{X_n\}$ satisfying $P[X_n = 1] = p$ and $P[X_n = 0] = 1 - p$ is a model for Bernoulli trials, for example, and for an event like $\bigcup_{n=1}^\infty [\sum_{k=1}^n X_k > \alpha n]$, expressed in terms of the X_n alone, the calculation of its probability proceeds in the same way whatever the underlying space $(\Omega, \mathscr{F}, P)$ may be.

[1] If $b - a = \delta_1 + \cdots + \delta_l$ and $\delta_i \ge 0$, then $I_i = (a + \sum_{j<i}\delta_j, a + \sum_{j\le i}\delta_j]$ decomposes $(a,b]$ into subintervals $I_1, \ldots, I_l$ with lengths of δ_i. Of course, I_i is empty if $\delta_i = 0$.

It is, of course, an advantage that such results apply not just to some canonical sequence $\{X_n\}$ (such as the one constructed in the proof above) but to every sequence with the appropriate distributions. In some applications of probability within mathematics itself, such as the arithmetic applications of run theory in the preceding section, the underlying Ω does play a role.

Expected Value

A simple random variable in the form (5.2) is assigned *expected value* or *mean value*

$$E[X]=E\left[\sum_i x_i I_{A_i}\right]=\sum_i x_i P(A_i). \tag{5.15}$$

There is the alternative form

$$E[X]=\sum_x xP[X=x], \tag{5.16}$$

the sum extending over the range of X; indeed, (5.15) and (5.16) both coincide with $\Sigma_x \Sigma_{i: x_i = x} x_i P(A_i)$. By (5.16) the definition (5.15) is consistent: different representations (5.2) give the same value to (5.15). From (5.16) it also follows that $E[X]$ depends only on the distribution of X; hence $E[X]=E[Y]$ if $P[X=Y]=1$.

If X is a simple random variable on the unit interval and if the A_i in (5.2) happen to be subintervals, then (5.15) coincides with the Riemann integral as given by (1.6). More general notions of integral and expected value will be studied later. Simple random variables are easy to work with because the theory of their expected values is transparent and free of technical complications.

As a special case of (5.15) and (5.16),

$$E[I_A]=P(A). \tag{5.17}$$

As another special case, if a constant α is identified with the random variable $X(\omega)\equiv\alpha$, then

$$E[\alpha]=\alpha. \tag{5.18}$$

From (5.2) follows $f(X)=\Sigma_i f(x_i) I_{A_i}$, and hence

$$E[f(X)]=\sum_i f(x_i)P(A_i)=\sum_x f(x)P[X=x], \tag{5.19}$$

the last sum extending over the range of X. For example, the kth *moment* $E[X^k]$ of X is defined by $E[X^k]=\Sigma_y yP[X^k=y]$, where y varies over the range of X^k, but it is usually simpler to compute it by $E[X^k]=\Sigma_x x^k P[X=x]$, where x varies over the range of X.

If

$$X=\sum_i x_i I_{A_i}, \qquad Y=\sum_j y_j I_{B_j} \tag{5.20}$$

are simple random variables, then $\alpha X+\beta Y=\Sigma_{ij}(\alpha x_i+\beta y_j)I_{A_i\cap B_j}$ has expected value $\Sigma_{ij}(\alpha x_i+\beta y_j)P(A_i\cap B_j)=\alpha\Sigma_i x_i P(A_i)+\beta\Sigma_j y_j P(B_j)$. Expected value is therefore *linear*:

$$E[\alpha X+\beta Y]=\alpha E[X]+\beta E[Y]. \tag{5.21}$$

If $X(\omega)\le Y(\omega)$ for all ω, then $x_i\le y_j$ if $A_i\cap B_j$ is nonempty, and hence $\Sigma_{ij}x_i P(A_i\cap B_j)\le\Sigma_{ij}y_j P(A_i\cap B_j)$. Expected value therefore *preserves order*:

$$E[X]\le E[Y] \qquad \text{if } X\le Y. \tag{5.22}$$

(It is enough that $X\le Y$ on a set of probability 1.) Two applications of (5.22) give $E[-|X|]\le E[X]\le E[|X|]$, so that by linearity,

$$|E[X]|\le E[|X|]. \tag{5.23}$$

And more generally,

$$|E[X-Y]|\le E[|X-Y|]. \tag{5.24}$$

The relations (5.17) through (5.24) will be used repeatedly, and so will the following theorem on expected values and limits. If there is a finite K such that $|X_n(\omega)|\le K$ for all ω and n, the X_n are *uniformly bounded*.

Theorem 5.4. *If $\{X_n\}$ is uniformly bounded, and if $X=\lim_n X_n$ with probability 1, then $E[X]=\lim_n E[X_n]$.*

PROOF. By Theorem 5.2(ii), convergence with probability 1 implies convergence in probability: $X_n\to_P X$. And in fact the latter suffices for the present proof. Increase K so that it bounds $|X|$ (which has finite range) as well as all the $|X_n|$; then $|X-X_n|\le 2K$. If $A=[|X-X_n|\ge\epsilon]$, then

$$|X(\omega)-X_n(\omega)|\le 2KI_A(\omega)+\epsilon I_{A^c}(\omega)\le 2KI_A(\omega)+\epsilon$$

for all ω. By (5.17), (5.18), (5.21), and (5.22),

$$E[|X-X_n|]\le 2KP[|X-X_n|\ge\epsilon]+\epsilon.$$

But since $X_n \to_P X$, the first term on the right goes to 0, and since ϵ is arbitrary, $E[|X - X_n|] \to 0$. Now apply (5.24). ■

Theorems of this kind are of constant use in probability and analysis. For the general version, Lebesgue's dominated convergence theorem, see Section 16.

***Example** 5.7.* On the unit interval, take $X(\omega)$ identically 0, and take $X_n(\omega)$ to be n^2 if $0 < \omega \le n^{-1}$ and 0 if $n^{-1} < \omega \le 1$. Then $X_n(\omega) \to X(\omega)$ for every ω, although $E[X_n] = n$ does not converge to $E[X] = 0$. Thus theorem 5.4 fails without some hypothesis such as that of uniform boundedness. See also Example 7.7. ■

An extension of (5.21) is an immediate consequence of Theorem 5.4:

Corollary. *If $X = \Sigma_n X_n$ on an $\mathscr{F}$-set of probability 1, and if the partial sums of $\Sigma_n X_n$ are uniformly bounded, then $E[X] = \Sigma_n E[X_n]$.*

Expected values for independent random variables satisfy the familiar product law. For X and Y as in (5.20), $XY = \Sigma_{ij} x_i y_j I_{A_i \cap B_j}$. If the x_i are distinct and the y_j are distinct, then $A_i = [X = x_i]$ and $B_j = [Y = y_j]$; for independent X and Y, $P(A_i \cap B_j) = P(A_i)P(B_j)$ by (5.9), and so $E[XY] = \Sigma_{ij} x_i y_j P(A_i)P(B_j) = E[X]E[Y]$. If X, Y, Z are independent, then XY and Z are independent by the argument involving (5.10), so that $E[XYZ] = E[XY]E[Z] = E[X]E[Y]E[Z]$. This obviously extends:

$$E[X_1 \cdots X_n] = E[X_1] \cdots E[X_n] \tag{5.25}$$

if $X_1, \ldots, X_n$ are independent.

Various concepts from discrete probability carry over to simple random variables. If $E[X] = m$, the *variance* of X is

$$\operatorname{Var}[X] = E\left[(X - m)^2\right] = E[X^2] - m^2; \tag{5.26}$$

the left-hand equality is a definition, the right-hand one a consequence of expanding the square. Since $\alpha X + \beta$ has mean $\alpha m + \beta$, its variance is $E[((\alpha X + \beta) - (\alpha m + \beta))^2] = E[\alpha^2 (X - m)^2]$:

$$\operatorname{Var}[\alpha X + \beta] = \alpha^2 \operatorname{Var}[X]. \tag{5.27}$$

If $X_1, \ldots, X_n$ have means $m_1, \ldots, m_n$, then $S = \Sigma_{i=1}^n X_i$ has mean $m = \Sigma_{i=1}^n m_i$, and $E[(S - m)^2] = E[(\Sigma_{i=1}^n (X_i - m_i))^2] = \Sigma_{i=1}^n E[(X_i - m_i)^2] + 2\Sigma_{1 \le i < j \le n} E[(X_i - m_i)(X_j - m_j)]$. If the X_i are independent, then so are the $X_i - m_i$, and by (5.25) the last sum vanishes. This gives the familiar formula for the variance of a sum of independent random variables:

$$\operatorname{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \operatorname{Var}[X_i]. \tag{5.28}$$

Suppose that X is nonnegative; order its range: $0 \le x_1 < x_2 < \cdots < x_k$. Then

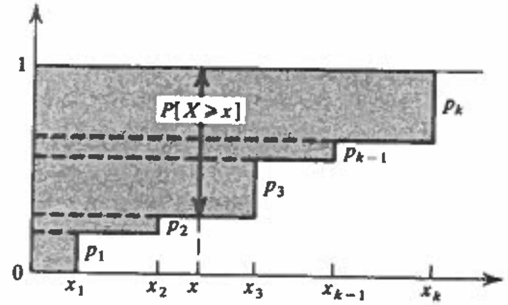
$$\begin{aligned} E[X] &= \sum_{i=1}^k x_i P[X = x_i] \\ &= \sum_{i=1}^{k-1} x_i (P[X \ge x_i] - P[X \ge x_{i+1}]) + x_k P[X \ge x_k] \\ &= x_1 P[X \ge x_1] + \sum_{i=2}^k (x_i - x_{i-1}) P[X \ge x_i]. \end{aligned}$$

Since $P[X \ge x] = P[X \ge x_1]$ for $0 \le x \le x_1$ and $P[X \ge x] = P[X \ge x_i]$ for $x_{i-1} < x \le x_i$, it is possible to write the final sum as the Riemann integral of a step function:

$$E[X] = \int_0^\infty P[X \ge x]\,dx. \tag{5.29}$$

This holds if X is nonnegative. Since $P[X \ge x] = 0$ for $x > x_k$, the range of integration is really finite.

There is for (5.29) a simple geometric argument involving the "area over the curve." If $p_i = P[X = x_i]$, the area of the shaded region in the figure is the sum $p_1 x_1 + \cdots + p_k x_k = E[X]$ of the areas of the horizontal strips; it is also the integral of the height $P[X \ge x]$ of the region.



Inequalities

There are for expected values several standard inequalities that will be needed. If X is nonnegative, then for positive α (sum over the range of X) $E[X]=\sum_x xP[X=x]\ge\sum_{x:\,x\ge\alpha}xP[X=x]\ge\alpha\sum_{x:\,x\ge\alpha}P[X=x]$. Therefore,

$$P[X\ge\alpha]\le\frac{1}{\alpha}E[X] \tag{5.30}$$

if X is nonnegative and α positive. A special case of this is (1.20). Applied to $|X|^k$, (5.30) gives *Markov's inequality*,

$$P[|X|\ge\alpha]\le\frac{1}{\alpha^k}E\left[|X|^k\right], \tag{5.31}$$

valid for positive α. If $k=2$ and $m=E[X]$ is subtracted from X, this becomes the *Chebyshev* (or Chebyshev–Bienaymé) *inequality*:

$$P[|X-m|\ge\alpha]\le\frac{1}{\alpha^2}\operatorname{Var}[X]. \tag{5.32}$$

A function φ on an interval is *convex* [A32] if $\varphi(px+(1-p)y)\le p\varphi(x)+(1-p)\varphi(y)$ for $0\le p\le 1$ and x and y in the interval. A sufficient condition for this is that φ have a nonnegative second derivative. It follows by induction that $\varphi(\sum_{i=1}^{l}p_ix_i)\le\sum_{i=1}^{l}p_i\varphi(x_i)$ if the p_i are nonnegative and add to 1 and the x_i are in the domain of φ. If X assumes the value x_i with probability p_i, this becomes *Jensen's inequality*,

$$\varphi(E[X])\le E[\varphi(X)], \tag{5.33}$$

valid if φ is convex on an interval containing the range of X.

Suppose that

$$\frac{1}{p}+\frac{1}{q}=1,\qquad p>1,\quad q>1. \tag{5.34}$$

Hölder's inequality is

$$E[|XY|]\le E^{1/p}[|X|^p]\cdot E^{1/q}[|Y|^q]. \tag{5.35}$$

If, say, the first factor on the right vanishes, then $X=0$ with probability 1, hence $XY=0$ with probability 1, and hence the left side vanishes also. Assume then that the right side of (5.35) is positive. If a and b are positive, there exist s and t such that $a=e^{p^{-1}s}$ and $b=e^{q^{-1}t}$. Since e^x is convex,

$e^{p^{-1}s+q^{-1}t}\le p^{-1}e^s+q^{-1}e^t$, or

$$ab\le\frac{a^p}{p}+\frac{b^q}{q}.$$

This obviously holds for nonnegative as well as for positive a and b. Let u and v be the two factors on the right in (5.35). For each ω,

$$\left|\frac{X(\omega)Y(\omega)}{uv}\right|\le\frac{1}{p}\left|\frac{X(\omega)}{u}\right|^p+\frac{1}{q}\left|\frac{Y(\omega)}{v}\right|^q.$$

Taking expected values and applying (5.34) leads to (5.35).

If $p=q=2$, Hölder's inequality becomes *Schwarz's inequality*:

$$E[|XY|]\le E^{1/2}[X^2]\cdot E^{1/2}[Y^2]. \tag{5.36}$$

Suppose that $0<\alpha<\beta$. In (5.35) take $p=\beta/\alpha$, $q=\beta/(\beta-\alpha)$, and $Y(\omega)=1$, and replace X by $|X|^\alpha$. The result is *Lyapounov's inequality*,

$$E^{1/\alpha}[|X|^\alpha]\le E^{1/\beta}\left[|X|^\beta\right],\qquad 0<\alpha\le\beta. \tag{5.37}$$

PROBLEMS

5.1. (a) Show that X is measurable with respect to the σ-field $\mathscr{G}$ if and only if $\sigma(X)\subset\mathscr{G}$. Show that X is measurable $\sigma(Y)$ if and only if $\sigma(X)\subset\sigma(Y)$.

(b) Show that, if $\mathscr{G}=\{\varnothing,\Omega\}$, then X is measurable $\mathscr{G}$ if and only if X is constant.

(c) Suppose that $P(A)$ is 0 or 1 for every A in $\mathscr{G}$. This holds, for example, if $\mathscr{G}$ is the tail field of an independent sequence (Theorem 4.5), or if $\mathscr{G}$ consists of the countable and cocountable sets on the unit interval with Lebesgue measure. Show that if X is measurable $\mathscr{G}$, then $P[X=c]=1$ for some constant c.

5.2. 2.19↑ Show that the unit interval can be replaced by any nonatomic probability measure space in the proof of Theorem 5.3.

5.3. Show that $m=E[X]$ minimizes $E[(X-m)^2]$.

5.4. Suppose that X assumes the values $m-\alpha, m, m+\alpha$ with probabilities $p, 1-2p, p$, and show that there is equality in (5.32). Thus Chebyshev's inequality cannot be improved without special assumptions on X.

5.5. Suppose that X has mean m and variance σ^2.

(a) Prove *Cantelli's inequality*

$$P[X-m\ge\alpha]\le\frac{\sigma^2}{\sigma^2+\alpha^2},\qquad \alpha\ge 0.$$

(b) Show that $P[|X-m| \geq \alpha] \leq 2\sigma^2/(\sigma^2+\alpha^2)$. When is this better than Chebyshev's inequality?

(c) By considering a random variable assuming two values, show that Cantelli's inequality is sharp.

5.6. The polynomial $E[(t|X|+|Y|)^2]$ in t has at most one real zero. Deduce Schwarz's inequality once more.

5.7. (a) Write (5.37) in the form $E^{\beta/\alpha}[|X|^\alpha] \leq E[(|X|^\alpha)^{\beta/\alpha}]$ and deduce it directly from Jensen's inequality.

(b) Prove that $E[1/X^p] \geq 1/E^p[X]$ for $p>0$ and X a positive random variable.

5.8. (a) Let f be a convex real function on a convex set C in the plane. Suppose that $(X(\omega), Y(\omega)) \in C$ for all ω and prove a two-dimensional Jensen's inequality:

$$f(E[X], E[Y]) \leq E[f(X,Y)]. \tag{5.38}$$

(b) Show that f is convex if it has continuous second derivatives that satisfy

$$f_{11} \geq 0, \qquad f_{22} \geq 0, \qquad f_{11}f_{22} \geq f_{12}^2. \tag{5.39}$$

5.9. ↑ Hölder's inequality is equivalent to $E[X^{1/p}Y^{1/q}] \leq E^{1/p}[X] \cdot E^{1/q}[Y]$ $(p^{-1}+q^{-1}=1)$, where X and Y are nonnegative random variables. Derive this from (5.38).

5.10. ↑ *Minkowski's inequality* is

$$E^{1/p}[|X+Y|^p] \leq E^{1/p}[|X|^p] + E^{1/p}[|Y|^p], \tag{5.40}$$

valid for $p \geq 1$. It is enough to prove that $E[(X^{1/p}+Y^{1/p})^p] \leq (E^{1/p}[X] + E^{1/p}[Y])^p$ for nonnegative X and Y. Use (5.38).

5.11. For events $A_1, A_2, \ldots$, not necessarily independent, let $N_n = \sum_{k=1}^n I_{A_k}$ be the number to occur among the first n. Let

$$\alpha_n = \frac{1}{n}\sum_{k=1}^n P(A_k), \qquad \beta_n = \frac{2}{n(n-1)} \sum_{1 \leq j < k \leq n} P(A_j \cap A_k). \tag{5.41}$$

Show that

$$E[n^{-1}N_n] = \alpha_n, \qquad \operatorname{Var}[n^{-1}N_n] = \beta_n - \alpha_n^2 + \frac{\alpha_n - \beta_n}{n}. \tag{5.42}$$

Thus $\operatorname{Var}[n^{-1}N_n] \to 0$ if and only if $\beta_n - \alpha_n^2 \to 0$, which holds if the A_n are independent and $P(A_n) = p$ (Bernoulli trials), because then $\alpha_n = p$ and $\beta_n = p^2 = \alpha_n^2$.

5.12. Show that, if X has nonnegative integers as values, then $E[X] = \sum_{n=1}^\infty P[X \geq n]$.

5.13. Let $I_i = I_{A_i}$ be the indicators of n events having union A. Let $S_k = \sum I_{i_1} \cdots I_{i_k}$, where the summation extends over all k-tuples satisfying $1 \leq i_1 < \cdots < i_k \leq n$. Then $s_k = E[S_k]$ are the terms in the inclusion–exclusion formula $P(A) = s_1 - s_2 + \cdots \pm s_n$. Deduce the inclusion–exclusion formula from $I_A = S_1 - S_2 + \cdots \pm S_n$. Prove the latter formula by expanding the product $\prod_{i=1}^n (1-I_i)$.

5.14. Let $f_n(x)$ be n^2x or $2n-n^2x$ or 0 according as $0 \leq x \leq n^{-1}$ or $n^{-1} \leq x \leq 2n^{-1}$ or $2n^{-1} \leq x \leq 1$. This gives a standard example of a sequence of continuous functions that converges to 0 but not uniformly. Note that $\int_0^1 f_n(x)\,dx$ does not converge to 0; relate to Example 5.7.

5.15. By Theorem 5.3, for any prescribed sequence of probabilities p_n, there exists (on some space) an independent sequence of events A_n satisfying $P(A_n) = p_n$. Show that if $p_n \to 0$ but $\sum p_n = \infty$, this gives a counterexample (like Example 5.4) to the converse of Theorem 5.2(ii).

5.16. ↑ Suppose that $0 \leq p_n \leq 1$ and put $\alpha_n = \min\{p_n, 1-p_n\}$. Show that, if $\sum \alpha_n$ converges, then on some discrete probability space there exist independent events A_n satisfying $P(A_n) = p_n$. Compare Problem 1.1(b).

5.17. (a) Suppose that $X_n \to_P X$ and that f is continuous. Show that $f(X_n) \to_P f(X)$.

(b) Show that $E[|X - X_n|] \to 0$ implies $X_n \to_P X$. Show that the converse is false.

5.18. 2.20↑ The proof given for Theorem 5.3 for the special case where the μ_n are all the same can be extended to cover the general case: use Problem 2.20.

5.19. 2.18↑ For integers m and primes p, let $\alpha_p(m)$ be the exact power of p in the prime factorization of m: $m = \prod_p p^{\alpha_p(m)}$. Let $\delta_p(m)$ be 1 or 0 as p divides m or not. Under each P_n (see (2.34)) the α_p and δ_p are random variables. Show that for distinct primes $p_1, \ldots, p_u$,

$$P_n[\alpha_{p_i} \geq k_i, i \leq u] = \frac{1}{n}\left\lfloor \frac{n}{p_1^{k_1}\cdots p_u^{k_u}} \right\rfloor \to \frac{1}{p_1^{k_1}\cdots p_u^{k_u}} \tag{5.43}$$

and

$$P_n[\alpha_{p_i} = k_i, i \leq u] \to \prod_{i=1}^u \left(\frac{1}{p_i^{k_i}} - \frac{1}{p_i^{k_i+1}}\right). \tag{5.44}$$

Similarly,

$$P_n[\delta_{p_i} = 1, i \leq u] = \frac{1}{n}\left\lfloor \frac{n}{p_1 \cdots p_u} \right\rfloor \to \frac{1}{p_1 \cdots p_u}. \tag{5.45}$$

According to (5.44), the α_p are for large n approximately independent under P_n, and according to (5.45), the same is true of the δ_p.

For a function f of positive integers, let

$$E_n[f] = \frac{1}{n}\sum_{m=1}^{n} f(m) \tag{5.46}$$

be its expected value under the probability measure P_n. Show that

$$E_n[\alpha_p] = \sum_{k=1}^{\infty}\frac{1}{n}\left\lfloor\frac{n}{p^k}\right\rfloor \to \frac{1}{p-1}; \tag{5.47}$$

this says roughly that $(p-1)^{-1}$ is the average power of p in the factorization of large integers.

5.20. ↑ (a) From Stirling's formula, deduce

$$E_n[\log] = \log n + O(1). \tag{5.48}$$

From this, the inequality $E_n[\alpha_p] \le 2/p$, and the relation $\log m = \sum_p \alpha_p(m)\log p$, conclude that $\sum_p p^{-1}\log p$ diverges and that there are infinitely many primes.
(b) Let $\log^* m = \sum_p \delta_p(m)\log p$. Show that

$$E_n[\log^*] = \sum_p \frac{1}{n}\left\lfloor\frac{n}{p}\right\rfloor \log p = \log n + O(1). \tag{5.49}$$

(c) Show that $\lfloor 2n/p\rfloor - 2\lfloor n/p\rfloor$ is always nonnegative and equals 1 in the range $n < p \le 2n$. Deduce $E_{2n}[\log^*] - E_n[\log^*] = O(1)$ and conclude that

$$\sum_{p\le x}\log p = O(x). \tag{5.50}$$

Use this to estimate the error removing the integral-part brackets introduces into (5.49), and show that

$$\sum_{p\le x} p^{-1}\log p = \log x + O(1). \tag{5.51}$$

(d) Restrict the range of summation in (5.51) to $\theta x < p \le x$ for an appropriate θ, and conclude that

$$\sum_{p\le x}\log p \asymp x, \tag{5.52}$$

in the sense that the ratio of the two sides is bounded away from 0 and ∞.
(e) Use (5.52) and truncation arguments to prove for the number $\pi(x)$ of primes not exceeding x that

$$\pi(x) \asymp \frac{x}{\log x}. \tag{5.53}$$

(By the prime number theorem the ratio of the two sides in fact goes to 1.) Conclude that the rth prime p_r satisfies $p_r \asymp r\log r$ and that

$$\sum_p \frac{1}{p} = \infty. \tag{5.54}$$

SECTION 6. THE LAW OF LARGE NUMBERS

The Strong Law

Let $X_1, X_2, \ldots$ be a sequence of simple random variables on some probability space $(\Omega, \mathscr{F}, P)$. They are *identically distributed* if their distributions (in the sense of (5.12)) are all the same. Define $S_n = X_1 + \cdots + X_n$. The *strong law of large numbers*:

Theorem 6.1. *If the X_n are independent and identically distributed and $E[X_n] = m$, then*

$$P\Big[\lim_n n^{-1}S_n = m\Big] = 1. \tag{6.1}$$

PROOF. The conclusion is that $n^{-1}S_n - m = n^{-1}\sum_{i=1}^{n}(X_i - m) \to 0$ with probability 1. Replacing X_i by $X_i - m$ shows that there is no loss of generality in assuming that $m = 0$. The set in question does lie in $\mathscr{F}$ (see (5.5)), and by Theorem 5.2(i), it is enough to show that $P[|n^{-1}S_n| \ge \epsilon \text{ i.o.}] = 0$ for each ϵ.

Let $E[X_i^2] = \sigma^2$ and $E[X_i^4] = \xi^4$. The proof is like that for Theorem 1.2. First (see (1.26)), $E[S_n^4] = \sum E[X_\alpha X_\beta X_\gamma X_\delta]$, the four indices ranging independently from 1 to n. Since $E[X_i] = 0$, it follows by the product rule (5.25) for independent random variables that the summand vanishes if there is one index different from the three others. This leaves terms of the form $E[X_i^4] = \xi^4$, of which there are n, and terms of the form $E[X_i^2 X_j^2] = E[X_i^2]E[X_j^2] = \sigma^4$ for $i \ne j$, of which there are $3n(n-1)$. Hence

$$E\big[S_n^4\big] = n\xi^4 + 3n(n-1)\sigma^4 \le Kn^2, \tag{6.2}$$

where K does not depend on n.

By Markov's inequality (5.31) for $k = 4$, $P[|S_n| \ge n\epsilon] \le Kn^{-2}\epsilon^{-4}$, and so by the first Borel–Cantelli lemma, $P[|n^{-1}S_n| \ge \epsilon \text{ i.o.}] = 0$, as required. ∎

Example 6.1. The classical example is the strong law of large numbers for Bernoulli trials. Here $P[X_n = 1] = p$, $P[X_n = 0] = 1 - p$, $m = p$; S_n represents the number of successes in n trials, and $n^{-1}S_n \to p$ with probability 1. The idea of probability as frequency depends on the long-range stability of the success ratio S_n/n. ∎