

Chapter 2

Random Variables, Distribution Functions, and Expectation

A. Random Variables and Distribution Functions

Given a space Ω and a σ -algebra \mathcal{F} of subsets of Ω , recall that a function $f : \Omega \rightarrow \mathbb{R}$ is called *Borel measurable* if $f^{-1}(U) = \{\omega ; f(\omega) \in U\}$ is an element of \mathcal{F} for every Borel set U in \mathbb{R} .

Definition A.1. A *random variable* is a real-valued, Borel measurable function on a probability space.

Definition A.2. If X_1, \dots, X_n are random variables on a common probability space, $(\Omega, \mathcal{F}, \mathbb{P})$, the function $Z = (X_1, \dots, X_n)$ from Ω to \mathbb{R}^n is called a random vector. More generally, if I is an index set, a collection of random variables $\{X_\alpha\}_{\alpha \in I}$ on a common probability space is called a *stochastic process* indexed by I . (In these notes, I will usually be the integers, or the positive integers.)

A random variable is then just a numerical quantity $X(\omega)$ depending on the random outcome ω of a trial. We impose the measurability assumption on X so that we can measure the probability of events of the form $\{\omega ; X(\omega) \in U\}$ for Borel sets U . The following simple example already shows the great usefulness of random variables for discussion of probabilistic problems.

Example A.1 Consider the coin tossing probability space $(\Omega, \mathcal{B}, P_p^\infty)$ where P_p is the probability measure for one toss of the coin with $P(\{1\}) = p$ and $P(\{0\}) = 1 - p$. This is the probability space constructed in Example C.3 of Chapter 1 ($p = 1/2$) or the exercise stated after Example E.2. Suppose we play a gambling game in which the payoff from heads is \$1 and the payoff (in this case, the loss) from tails is -\$1. Let Z_i denote the payoff from toss i ; then, for $\omega \in \{0, 1\}^\infty$,

$$Z_i(\omega) = 2\omega_i - 1.$$

It is easy to check that each Z_i is a measurable function and hence is a random variable. The total gain accumulated from tosses 1 through n is then the random variable $S_n(\omega) = \sum_{i=1}^n Z_i(\omega)$. It is interesting to study $\{S_n\}_{n \geq 1}$ as a stochastic process. It is called the simple random walk; when $p = 1/2$, we say that S_n is a simple, symmetric random walk. The idea is that the Z_i , $1 \leq i < \infty$, represent successive, random and independent steps and S_n represents the position, relative to the origin, of the walker after n steps. We are particularly interested in the long time, i.e. large n , behavior of S_n . For example, we shall see that the strong law of large numbers implies that

$$P_p^\infty \left(\lim_{n \rightarrow \infty} \frac{1}{n} S_n = 2p - 1 \right) = 1.$$

This implies the intuitively obvious fact that if $p > 1/2$, our winnings will tend to increase without bound, while if $p < 1/2$, our losses will accumulate without bound. The critical case, $p = 1/2$ is more interesting. Then the strong law, as stated in Theorem C.1 of Chapter 1, says that $(1/n)S_n$ converges to 0 with probability 1, which does not tell us whether $|S_n|$ can get arbitrarily large or not. Questions about the asymptotic behavior of S_n , its rate of growth, whether it crosses 0 infinitely often, lead directly to central issues of probability theory, including the central limit theorem, the law of the iterated logarithm, and recurrence of Markov chains. \diamond

Remark: Sometimes it is convenient to use *extended* random variables, that is functions on the probability space taking values in the extended real line $\{-\infty\} \cup \mathbb{R} \cup \{\infty\}$ and measurable in the sense that $X^{-1}(\infty)$, $X^{-1}(-\infty)$, and $X^{-1}(U)$, for Borel set U , are all in \mathcal{F} . Extended random variables are convenient when taking limits, suprema, or infima, of sequences of random variables.

Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. We let $\sigma(X)$ denote the smallest sub- σ -algebra of \mathcal{F} with respect to which X is measurable. Thus, if $\mathcal{B}(\mathbb{R})$ is the Borel σ -algebra of \mathbb{R} ,

$$\sigma(X) = X^{-1}(\mathcal{B}(\mathbb{R})) = \{X^{-1}(U); U \in \mathcal{B}(\mathbb{R})\}.$$

This σ -algebra is, in words, just the set of all events concerning the value of X . If $\{X_\alpha; \alpha \in I\}$ is some set of random variables indexed by I , $\sigma(\{X_\alpha; \alpha \in I\})$ denotes the σ -algebra generated by $\cup_{\alpha \in I} \sigma(X_\alpha)$. It will be very convenient in future work to use these definitions. In example A.1, the σ -algebra generated by Z_i is simply the collection of all subsets of the form $\{\omega; \omega_i \in A\}$ where A is a subset of $\{0, 1\}$. The σ -algebra generated by S_n consist of all disjoint unions of sets of the form $\{\omega; \sum_1^n \omega_i = k\}$, where k ranges through the even integers between $-n$ and n if n is even, and through the odd integers between $-n$ and n if n is odd.

Several conventions of notation are common in the treatment of random variables. One usually denotes random variables, as we have done, using capital letters from the end of the alphabet. Although random variables are functions, it is also usual to suppress mention of the independent variable. Thus, $\{\omega; X(\omega) \in U\}$ is written simply as $\{X \in U\}$.

Finally, as a reminder, we recall some basic facts about measurability.

- (1) A function X on (Ω, \mathcal{F}) is Borel measurable if and only if $X^{-1}((-\infty, b]) \in \mathcal{F}$ for every $b \in \mathbb{R}$.
- (2) If $Z = (X_1, \dots, X_n)$ is a random vector then Z is Borel measurable as a map from Ω to \mathbb{R}^n .
- (3) If Z is an \mathbb{R}^n -valued random vector and if $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is Borel measurable, then $f(Z)$ is a random variable. In particular, any algebraic function of random variables, is again a random variable.
- (4) If $X_n, n \geq 1$, is a sequence of (extended) random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, then $\sup X_n, \inf X_n, \limsup X_n, \liminf X_n$, etc. are all (possibly extended) random variables. If $\lim X_n$ exists everywhere, it is likewise a random variable.

The reader may have noticed an ambiguity in our definition of random variable, because, if a random variable is just a measurable function, what role does \mathbb{P} play, if any? In Example A.1, the functions Z_i and S_n are simply measurable functions that have a

clear meaning independent of the probability measure placed on $\{0,1\}^\infty$. Nevertheless, when speaking of a random variable we shall always have in mind a particular probability measure, say \mathbb{P} , on the underlying outcome space. The same function on $(\Omega, \mathcal{F}, \mathbb{M})$, where $\mathbb{M} \neq \mathbb{P}$ shall be considered a different random variable. We do this because, as we shall see, it is not the particular way X is defined as a function that interests us, but the probabilities of the events in $\sigma(X)$.

There is a more refined definition of random variable that does use the underlying probability measure explicitly. Let us say that two measurable functions X and Y on $(\Omega, \mathcal{F}, \mathbb{P})$ are \mathbb{P} -equivalent if $\mathbb{P}(X \neq Y) = 0$. Since \mathbb{P} -equivalent functions are identical from the point of view of the measure \mathbb{P} , one could define random variable to be a \mathbb{P} -equivalence class of measurable functions. In Example A.1, let $X = \lim_{n \rightarrow \infty} S_n/n$ if the limit exists and $X = 2$ otherwise (the value 2 was chosen arbitrarily just to mark the event that the limit does not exist). Then the strong law of large numbers implies that X is P_p^∞ -equivalent to $2p - 1$. Thus the same measurable function is effectively a different constant for different p . The equivalence class definition thus brings out the important role of the probability measure. Despite this conceptual benefit of the equivalence class definition, we shall stick with the definition of a random variable as an ordinary measurable function on a given probability space. In this way, we avoid the awkwardness of having to frame definitions and theorems in the language of equivalence classes rather than actual functions.

B. Laws of Random Variables and Distribution Functions

In Chapter 1, we proposed modelling random phenomena by construction of a probability space. We can also model the outcome of a random trial as the value taken on by a random variable. In fact, this is the preferred method, because it gives greater flexibility and allows the application of the theory of measurable functions and their integrals. In the random variable approach, the nature of the probability space is not so important, and need not correspond in any direct way to the phenomenon being modelled. We begin by illustrating with an example.

Example B.1 Let $\mathcal{B}[0,1)$ denote the Borel subsets of $[0,1)$, and let \mathbb{P} be Lebesgue measure. Then $([0,1), \mathcal{B}[0,1), \mathbb{P})$ is a probability space. Each number x in $[0,1)$ admits a binary expansion

$$x = \sum_1^\infty x_i 2^{-i},$$

where each $x_i \in \{0,1\}$, and this expansion is unique if sequences ending in a string of 1's are excluded. Let $Y_i(x) = x_i$, $1 \leq i < \infty$ assign to x the i^{th} coefficient in its expansion. For example, $Y_1(x) = \mathbf{1}_{[1/2,1)}(x)$, $Y_2(x) = \mathbf{1}_{[1/4,1/2)}(x) + \mathbf{1}_{[3/4,1)}(x)$, and so on, where $\mathbf{1}_A$ denotes the indicator function of A . Now let N be an arbitrary integer and let $(\omega_1, \dots, \omega_N)$ be an arbitrary N -vector of 0's and 1's. We leave it as an exercise that

$$\mathbb{P}(Y_1 = \omega_1, \dots, Y_N = \omega_N) = 2^{-N}.$$

Comparing this result to the definition of $\mathbb{P}^{(N)}$ on $\{0,1\}^N$ in Example C.1 of Chapter 1, we see that for any N , the random variables Y_1, \dots, Y_N on $([0,1], \mathcal{B}[0,1], \mathbb{P})$ is a model of N independent tosses of a fair coin. That is, selecting a point x at random from $[0,1]$ according to Lebesgue measure and forming the sequence $Y_1(x), \dots, Y_N(x)$ is statistically equivalent to flipping a fair coin N times with independent tosses and recording heads as 1 and tails as 0. By extension, the process Y_1, Y_2, \dots on $([0,1], \mathcal{B}[0,1], \mathbb{P})$ is a model of an infinite sequence of independent tosses of a fair coin. \diamond

In the previous example we see that what really matters to us about random variables is not the particular probability space on which they are defined, nor how they are defined, but probabilities, such as $\mathbb{P}(X \in U)$ or $\mathbb{P}(X \in U, Y \in V)$, of the events on the values of the random variables. The next definition is therefore very basic.

Definition B.1 Given a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, the *distribution measure* or *law* of X is the probability measure

$$\mathbb{F}_X(U) := \mathbb{P} \circ X^{-1}(U) := \mathbb{P}(X \in U).$$

defined on the Borel sets U of \mathbb{R} . Similarly, the law of a random vector $Z = (X_1, \dots, X_n)$, often called the *joint law* of (X_1, \dots, X_n) , is the measure $\mathbb{F}_Z := \mathbb{P} \circ Z^{-1}$ on the Borel sets of \mathbb{R}^n .

Two \mathbb{R}^n -valued random vectors Z and Z' , defined on possibly different probability spaces, are said to be *equal in law* if $\mathbb{F}_Z = \mathbb{F}_{Z'}$.

To illustrate, each random variable Y_i in Example B.1 has the same law, $\mathbb{F} = (1/2)(\delta_0 + \delta_1)$, where δ_x is the Dirac delta measure putting a mass of 1 at point x . (Laws like this that put probability mass at only two points are called Bernoulli distributions.) We showed in Example B.1 that the law \mathbb{F}_n of the random vector (Y_1, \dots, Y_N) is essentially the measure $\mathbb{P}^{(N)}$ on $\{0,1\}^N$ modelling N -independent tosses of a fair coin. More exactly,

$$(1) \quad \mathbb{F}_n(V) = \frac{|\{\omega \in \{0,1\}^N; \omega \in V\}|}{2^n} = \mathbb{P}^{(N)}(V).$$

Finally, if we define the random variable $X(\omega) = \omega$ for $\omega \in [0,1]$ on the probability space $([0,1], \mathcal{B}[0,1], \mathbb{P})$ of Example B.1, we see that the law of X is just Lebesgue measure on $[0,1]$. (In this case we say that X has the uniform distribution on $[0,1]$.)

Remark: Notice in regard to the remarks at the end of section A that the law of a random variable is governed by the fixed probability measure \mathbb{P} on the underlying probability space.

In practice, a probability model of a random phenomena is often specified by a law on \mathbb{R}^n , describing the statistical behavior of the outcome, and a random vector with that law. To repeat, little attention is paid to the particular nature of Ω or to the way in which the random variables are defined as functions on Ω . For example, when we discuss an infinite sequence of independent coin tosses in the future, we merely invoke an infinite sequence of random variables with joint distributions as in (1) on some abstract probability space. In

this approach, it is important to answer the question: given a probability measure \mathbb{P} on \mathbb{R}^n , is there in fact some probability space with a random vector Z , such that $\mathbb{P}_Z = \mathbb{P}$? The answer is yes and the construction of Z is simple.

Lemma B.1. Let \mathbb{P} be any probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Then there is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random vector Z on it, such that $\mathbb{P}_Z = \mathbb{P}$.

Proof. Let $(\Omega, \mathcal{F}, \mathbb{P}) := (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathbb{P})$ and let $Z(\omega) = \omega$ for $\omega \in \Omega = \mathbb{R}^n$. \diamond

The probability space constructed in the preceding proof is called the *canonical space* for \mathbb{P} .

A similar question to the one answered by Lemma A.1 can be posed for stochastic processes. The answer is a little more complicated, and we defer discussion to a later section.

We now introduce another important concept.

Definition B.2 The *cumulative distribution function* (also called the *probability distribution function*) of X is the function

$$F_X(b) := \mathbb{P}(X \leq b) (= \mathbb{P}_X((-\infty, b])).$$

Likewise, the *joint cumulative distribution function* of the random vector $Z = (X_1, \dots, X_n)$ is the function

$$F_Z(b_1, \dots, b_n) := \mathbb{P}(X_1 \leq b_1, \dots, X_n \leq b_n) = \mathbb{P}_Z((-\infty, b_1] \times \dots \times (-\infty, b_n]).$$

Cumulative distribution functions are important because they give simple and unique characterizations of probability laws of random vectors. Consider, for example, the random variable case. A simple calculation states that for for $a < b$,

$$(2) \quad \mathbb{P}((a, b]) = \mathbb{P}(a < X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F_X(b) - F_X(a),$$

so the cumulative distribution function determines the probabilities of X for all intervals of the form $(a, b]$, and by extension, for all Borel sets U . Details of the argument are sketched in the next Proposition, which characterizes all cdf's of random variables.

Proposition B.2 A function F is the cumulative distribution function of a random variable if and only if

- (i) F is right continuous and non-decreasing,
 - (ii) $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$.
- Given an F satisfying (i) and (ii), the probability law \mathbb{P} corresponding to F in the sense of equation (2) is unique.

Proof: If F is the cdf (cumulative distribution function) of a random variable X , and $\{\epsilon_n\}$ is a sequence of positive numbers decreasing to 0, then, using the continuity from above of \mathbb{P}

$$\begin{aligned} \lim_{n \rightarrow \infty} F(b + \epsilon_n) &= \lim_{n \rightarrow \infty} \mathbb{P}(X \leq b + \epsilon_n) = \mathbb{P}(\cap_{n \geq 1} \{X \leq b + \epsilon_n\}) \\ &= \mathbb{P}(X \leq b) = F(b), \end{aligned}$$

which proves right continuity at an arbitrary b . The non-decreasing property of F is an immediate consequence of the monotonicity of probability measures.

Let $b_n \downarrow -\infty$. Since $\emptyset = \cap_n \{X \leq b_n\}$, the fact that $\lim_{x \rightarrow -\infty} F(x) = 0$ follows again from continuity from above of \mathbb{P} . Similarly, the property, $\lim_{x \rightarrow \infty} F(x) = 1$, is a due to continuity from below and the observation that $\Omega = \cup_n \{X \leq b_n\}$, if $b_n \uparrow \infty$.

To argue the converse, let F satisfy (i) and (ii). Let \mathcal{R} be the collection of all finite, disjoint unions of intervals of the form $(a, b]$. In the Exercise of Chapter 1, Section A, we proved that \mathcal{R} is an algebra. Let $\cup_1^n (a_i, b_i]$ be an arbitrary element of \mathcal{R} with disjoint intervals $(a_i, b_i]$ and define

$$\mathbb{F}_0(\cup_1^n (a_i, b_i]) := \sum_1^n (F(b_i) - F(a_i)).$$

This definition is motivated by equation (2) above. Now we can check that \mathbb{F}_0 is a finitely additive probability measure on \mathcal{R} and that \mathbb{F}_0 is continuous from above at \emptyset . Carathéodory's extension theorem then implies the existence of a unique probability measure on the Borel sets of \mathbb{R} extending \mathbb{F}_0 . From Lemma B.1, there is a random variable Z whose law coincides with \mathbb{F} . Clearly then $F_Z = F$. \diamond

Because of Proposition B.2, any function satisfying properties (i) and (ii) is called a cumulative distribution function. Also, Proposition B.2 says that we can characterize a random variable in law by specifying its cdf, and this is the approach we most often take.

Note that a cdf need not be left-continuous. Indeed, if we denote the left limit of F_X at a by $F_X(a-)$, continuity from above of probability measures, implies that

$$(4) \quad F_X(b) - F_X(b-) = \mathbb{P}(\cap_{a < b} \{a < X \leq b\}) = \mathbb{P}(X = b).$$

This, discontinuities of F correspond precisely to atoms of the law of X .

In practice, random variables come in two flavors, the discrete and the continuous. A discrete random variable is one which takes on only a countable number of possible values and whose law is therefore an atomic measure. From equation (4), we see that the cdf of a discrete random variable will be piecewise constant with jumps at the atoms of X . In the discrete case, it is more convenient to characterize the law by its probability mass function, specifying the probability mass of each atom. If X is discrete with values in the set $\{a_1, a_2, \dots\}$, its *probability mass function* is

$$p_i = \mathbb{P}(X = a_i), \quad 1 \leq i < \infty.$$

We shall also define continuous random variables.

Definition B.3 A random variable $Z = (X_1, \dots, X_n)$ is continuous if its law is absolutely continuous with respect to Lebesgue measure, that is, if there exists a non-negative function f_Z on \mathbb{R}^n such that

$$\mathbb{F}_Z(U) = \int_U f_Z(x_1, \dots, x_n) dx_1 \cdots dx_n,$$

for all Borel sets $U \subset \mathbb{R}^n$. The function f_Z is called the *probability density* of Z or the *joint density* of (X_1, \dots, X_n) .

Note that a continuous random variable is not necessarily continuous as a function. Neither is a random variable whose law lacks atoms necessarily continuous. Rather, a random variable X is continuous if and only if its cdf, F_X is absolutely continuous, so that there is a density f_X for which

$$(5) \quad F_X(b) = \int_{-\infty}^b f_X(x) dx, \quad \forall b.$$

Proposition B.2 extends to the multi-dimensional case. We shall state the extension and sketch its proof. We begin with the extension of formula (2) to higher dimensions. If $a < b$ and G is a function of n variables, define a new function $\Delta_a^b G$ of $n - 1$ variables by

$$\Delta_a^b G(x_2, \dots, x_n) := G(b, x_2, \dots, x_n) - G(a, x_2, \dots, x_n).$$

Now suppose that $a = (a_1, \dots, a_n)$ and $b = (b_1, \dots, b_n)$ are two vectors such that $a_i < b_i$ for each i . Then an induction argument will show that

$$(6) \quad \mathbb{F}_Z((a_1, b_1] \times \dots \times (a_n, b_n]) = \Delta_{a_n}^{b_n} \dots \Delta_{a_1}^{b_1} F_Z.$$

We leave the proof of (6) as an exercise. By the same arguments as in the one-dimensional case, one can identify all possible distribution functions and show that F_Z uniquely characterizes \mathbb{F}_Z .

Proposition B.3. A function $F : \mathbb{R}^n \rightarrow [0, 1]$ is the distribution function of a random vector if and only if F has the following properties:

- (i) F is non-decreasing and right continuous in each variable;
- (ii) F satisfies

$$\lim_{(x_1, \dots, x_n) \rightarrow \infty} F(x_1, \dots, x_n) = 1 \quad \text{and} \quad \lim_{(x_1, \dots, x_n) \rightarrow -\infty} F(x_1, \dots, x_n) = 0;$$

- (iii) For any $a = (a_1, \dots, a_n)$ and $b = (b_1, \dots, b_n)$ such that $a_i < b_i$ for each $1 \leq i \leq n$,

$$\Delta_{a_n}^{b_n} \dots \Delta_{a_1}^{b_1} F_Z \geq 0.$$

The probability law \mathbb{F}^F corresponding to a function F satisfying (i)-(iii) is unique.

Proof: Suppose that $F = F_Z$ where Z is a random vector. Then (i) and (ii) are consequences of the definition of F_Z in terms of \mathbb{F}_Z and the continuity from above and below of the probability measure \mathbb{F}_Z , as in the proof of Proposition B.2. Property (iii) is a direct consequence of equation (6) and the positivity of \mathbb{F}_Z .

We shall only sketch the proof of the converse, which is mostly a classical exercise in the construction of Borel measures on Euclidean space. Let \mathcal{R} denote the algebra of finite

disjoint unions of rectangles of the form $\prod_1^n (a_i, b_i]$, and define the measure $\mathbb{I}F_0$ on \mathcal{R} using equation (6); that is,

$$\mathbb{I}F_0\left(\prod_1^n (a_i, b_i]\right) := \Delta_{a_n}^{b_n} \cdots \Delta_{a_1}^{b_1} F_Z \geq 0$$

on rectangles. One can show that $\mathbb{I}F_0$ is continuous from above at \emptyset on \mathcal{R} . Carathéodory's theorem then says that $\mathbb{I}F_0$ admits a unique extension, $\mathbb{I}F$, to a countably additive probability measure on the Borel sets of \mathbb{R}^n . Finally lemma B.1 implies that there is a random variable such that $\mathbb{I}F_Z = \mathbb{I}F$, and hence $F = F_Z$. \diamond

C. Expected Values

Definitions. Let X be a random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{I}P)$. Then the *expected value* or *mean* of X is the number

$$E[X] = \int_{\Omega} X(\omega) \mathbb{I}P(d\omega).$$

if X is integrable with respect to $\mathbb{I}P$. We often denote $E[X]$ by μ_X .

If A is an event, we shall let $\mathbf{1}_A$ represent the indicator function of A . Note that,

$$E[\mathbf{1}_A] = \mathbb{I}P(A),$$

an identity we shall use repeatedly without further comment.

To motivate the definition of expected value, consider a discrete random variable X taking on only $\{c_1, \dots, c_n\}$ as its possible values. Then X is a simple function of the form

$$X(\omega) = \sum_1^n c_i \mathbf{1}_{A_i}(\omega),$$

where, without loss of generality, we may assume that A_1, \dots, A_n are disjoint. Now consider repeated, independent trials of X , whose outcome are labelled X_1, X_2, \dots . The mean or *empirical average* of the first n trials is $\bar{X}_n := (X_1 + \dots + X_n)/n$. When n is large, we expect to see the value c_i about $n\mathbb{I}P(X = c_i)$ times in the n trials and so we expect that

$$\begin{aligned} \bar{X}_n &\approx c_1 \mathbb{I}P(X = c_1) + \cdots + c_n \mathbb{I}P(X = c_n) \\ &= c_1 \mathbb{I}P(A_1) + \cdots + c_n \mathbb{I}P(A_n) \\ &= \int_{\Omega} X(\omega) \mathbb{I}P(d\omega). \end{aligned}$$

The last expression is just what we have defined as $E[X]$. The first major limit theorems we prove will be large number laws verifying this interpretation of expected values. We shall state these after we have discussed independence of random variables.

It is immediate from the definition of expected value that it acts linearly on random variables; that is $E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n]$. We shall use this linearity repeatedly.

Examples C.1 (a) Let X_1, \dots, X_n denote the results of n coin tosses, where each toss has probability p of coming up heads. Assume $X = 1$ records a head, while $X = 0$ records a tail. Then $E[X_1] = p$ and $E[X_1 + \cdots + X_n] = np$.

(b) N people wearing hats go to a party. They have such a good time that when they leave they pay not attention to whose hat is whose and just take one at random from the pile. Let Y be the number of persons who leave with their own hat. What is $E[Y]$? We write $Y = X_1 + \dots + X_N$ where $X_i = 1$ if person i gets back his or her own hat, and $X_i = 0$ otherwise. A simple combinatorial argument shows that for each person $E[X_i] = 1/N$. Hence $E[Y] = 1$.

In these notes we shall assume without review the standard theorems from integration theory—monotone convergence, Fatou’s lemma, dominated convergence, Hölder’s inequality, etc.

Recall that in specifying a random variable one normally gives only its cdf and not an explicit construction of it as a function on a probability space. Therefore, we need a formula that computes expected value from the cdf.

Theorem C.1 Let \mathbb{F}_Z be the joint distribution measure of X_1, \dots, X_n and let g be a Borel measurable function of n variables. Then

$$(1) \quad E[g(X_1, \dots, X_n)] = \int_{\mathbb{R}^n} g(x_1, \dots, x_n) \mathbb{F}_Z(d(x_1, \dots, x_n)),$$

in the sense that one side of the equality is well-defined if the other side is. In particular, if F_X is the probability distribution function of X ,

$$(2) \quad E[X] = \int_{-\infty}^{\infty} x dF_X(x).$$

Proof: (Sketch) Verify that equation (1) holds for simple functions g and then pass to the general case by taking limits of simple functions. \diamond

Example C.2 Let X be uniformly distributed on $[0, 1]$. This means that \mathbb{F}_X is Lebesgue measure on $[0, 1]$. Then $E[X] = \int_0^1 x dx = 1/2$.

Equation (2) says that the mean of a random variable is the center of mass of its probability law. To measure the concentration of the law of X about its mean we introduce the *variance*, $\text{Var}(X)$. It is defined by

$$\text{Var}(X) := E[(X - \mu_X)^2].$$

Note by (1) that

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 dF_X(x).$$

In general, for $r \geq 1$, the quantity $E[|X|^r]$, is called the r^{th} moment of X , while $E[|X - \mu_X|^r]$ is called the r^{th} centered moment.

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we shall let $L^p(\mathbb{P})$ denote the collection of all random variables with finite p^{th} moment. It is useful to note that since probability measures are finite measures, existence of an r^{th} moment implies the existence of lower order moments.

Proposition C.3 For a random variable X and $1 \leq s < r$, $E[|X|^s] \leq (E[|X|^r])^{s/r}$.

Proof: This is simply an application of Hölder's inequality,

$$|E[XY]| \leq (E[|X|^p])^{1/p} (E[|Y|^q])^{1/q},$$

where $1/p + 1/q = 1$, with $p = r/s$ and $Y \equiv 1$. \diamond

It is useful to have special names for second order moments of pairs of random variables. Thus, if X and Y are random variables with finite variances, define the *covariance* between X and Y as

$$\text{Cov}(X, Y) := E[(X - \mu_X)(Y - \mu_Y)].$$

By the Cauchy-Schwarz inequality,

$$(3) \quad |\text{Cov}(X, Y)|^2 \leq \text{Var}(X)\text{Var}(Y),$$

so finiteness of the variances of X and of Y suffices for the existence of the covariance. The *correlation* between X and Y is

$$\text{Cor}(X, Y) := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

By (3), the correlation is a number between -1 and 1 . The correlation of two random variables will be close to 1 if Y strongly tends to be above or below its mean when X is respectively above or below its mean, and it will be close to -1 if Y tends to be above its mean when X is below its mean and vice-versa. A small correlation means that the signs of $X - \mu_x$ and $Y - \mu_Y$ tend not to affect each other strongly.

Definition C.2 Random variables X and Y are *uncorrelated* if $\text{Cov}(X, Y) = 0$.

We shall see a significant application of the notion of uncorrelated random variables in the study of large number laws. For now, as an exercise in applying the definitions, we state the following simple, but basic formulas for the computation of variances of linear combinations of random variables. The proof is by direct computation.

Proposition C.4 Let X_1, \dots, X_n be random variables with finite variances. Then

$$(4) \quad \text{Var}\left(\sum_1^n a_i X_i\right) = \sum_1^n a_i^2 \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} a_i a_j \text{Cov}(X_i, X_j).$$

In particular, if X_1, \dots, X_n are uncorrelated, then

$$(5) \quad \text{Var}\left(\sum_1^n a_i X_i\right) = \sum_1^n a_i^2 \text{Var}(X_i)$$

Finally we use expectations to define two transforms of probability laws.

Definition C.3 If X is a random variable the function,

$$\phi_X(\lambda) = E[e^{i\lambda X}], \quad \lambda \in \mathbb{R},$$

is called the *characteristic function* of X . The function,

$$m_X(t) = E[e^{tX}], \quad t \in \mathbb{R},$$

is called the *moment generating function* of X .

Again applying (1),

$$\phi_X(\lambda) = \int_{\mathbb{R}} e^{i\lambda x} dF_X(x), \quad \text{and} \quad m_X(t) = \int_{\mathbb{R}} e^{tx} dF_X(x).$$

In particular, $\phi_X(\lambda)$ is the Fourier transform of the law F_X of X . We shall discuss the characteristic function in depth later on. For now, we will be more interested in the moment generating function. The following properties of the moment generating function are important. The proof is left as an exercise.

Proposition C.5 Assume that there is an $\epsilon > 0$ such that $m_X(t) < \infty$ for all t satisfying $-\epsilon < t < \epsilon$. Then $E[|X|^n]$ is finite for all n and

$$(6) \quad E[|X|^n] = \left. \frac{d^n m_X(t)}{dt^n} \right|_{t=0}.$$

Often, in stating a theorem, or in formulating a random variable model, we want to assume no more of a random variable than that it has finite moments up to a certain order. The mere existence of moments has simple consequences for the probability distribution that can be exploited to great effect. The basic result, *Markov's inequality*, looks even crude, but is amazingly useful.

Proposition C.6 Let Z be a nonnegative random variable ($\mathbb{P}(Z \geq 0) = 1$). Then for all $a > 0$,

$$(7) \quad \mathbb{P}(Z \geq a) \leq \frac{E[Z]}{a}$$

Moreover, if $E[Z] < \infty$, $\lim_{a \rightarrow \infty} a\mathbb{P}(Z \geq a) = 0$.

PROOF: Since $1 \leq Z/a$ on the set $\{Z \geq a\}$,

$$\mathbb{P}(Z \geq a) = E[\mathbf{1}_{\{Z \geq a\}}] \leq \frac{1}{a} E[Z \mathbf{1}_{\{Z \geq a\}}] \leq \frac{1}{a} E[Z],$$

giving (7). From the same calculation, we obtain from the dominated convergence theorem that

$$\lim_{a \rightarrow \infty} a \mathbb{P}(Z \geq a) \leq \lim_{a \rightarrow \infty} E[Z \mathbf{1}_{\{Z \geq a\}}] = 0. \quad \diamond$$

By applying Markov's inequality to higher order moments, one can derive faster decay: if $E[|X|^r] < \infty$,

$$(8) \quad \mathbb{P}(|X| \geq a) \leq \frac{E[|X|^r]}{a^r}.$$

Perhaps the most important application is Chebyshev's inequality:

$$(9) \quad \mathbb{P}(|X - \mu_X| \geq a) \leq \frac{\text{Var}(X)}{a^2},$$

which comes from applying (8) with $X - \mu_X$ in place of X and $r = 2$.

The moment generating function, together with Markov's inequality, provides another bound on tail probabilities. Indeed, for $t > 0$, $\mathbb{P}(X \geq a) = \mathbb{P}(e^{tX} \geq e^{ta})$. Applying Markov's inequality to the last probability, gives

$$(10) \quad \mathbb{P}(X \geq a) \leq e^{-at} m_X(t)$$

But this inequality holds for all $t > 0$, and hence, to get the best possible bound, we should optimize over t . We state the result in the following proposition.

Proposition C.7 Let $\Lambda_X(t) := m_X(t)$ be the logarithmic moment generating function of X . Then

$$(11) \quad \begin{aligned} \mathbb{P}(X \geq a) &\leq \exp\left\{-\sup_{t \geq 0} (at - \Lambda_X(t))\right\}, & \text{and} \\ \mathbb{P}(X \leq a) &\leq \exp\left\{-\sup_{t \leq 0} (at - \Lambda_X(t))\right\}. \end{aligned}$$

PROOF: For the first inequality, just apply (10) and the definition of $\Lambda_X(t)$ and optimize over t . Notice that we have taken the supremum over $t \geq 0$ rather than $t > 0$, but this is harmless, because setting $t = 0$ in $at - \Lambda_X(t)$ yields 0. The second inequality is proved in a similar way. \diamond

A first example of the power of the inequalities in (11) is given for binomial random variables in the next section. Later on, (11) provides the essential step in the derivation of large deviation upper bounds for empirical means of a sequence of independent and identically distributed random variables.

D. Independence of random variables.

Roughly speaking, two random variables X and Y are independent of one another if any event concerning the outcome of X alone is independent of any event concerning the outcome of Y alone. This is made precise in the following definition.

Definition D.1 Random variables X and Y are independent if $\sigma(X)$ and $\sigma(Y)$ are independent σ -algebras. The random vector Z_1 is independent of the random vector Z_2 if $\sigma(Z_1)$ is independent of $\sigma(Z_2)$. The random variables of the family $\{X_\alpha; \alpha \in \mathcal{I}\}$ are (mutually) independent if the σ -algebras of the corresponding family $\{\sigma(X_\alpha); \alpha \in \mathcal{I}\}$ are independent.

Thus, X and Y are independent if and only if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ for any $A \in \sigma(X)$ and any $B \in \sigma(Y)$. Since the typical member of $\sigma(X)$ has the form $\{X \in U\}$ for a Borel set U , independence can be restated as requiring that

$$(1) \quad \mathbb{P}(X \in U, Y \in V) = \mathbb{P}(X \in U)\mathbb{P}(Y \in V) \quad \text{for any Borel sets } U \text{ and } V.$$

Example D.1 Return to Example B.1, in which $Y_1(\omega), Y_2(\omega), \dots$ were the successive digits in the binary expansion of a point ω drawn at random from $[0, 1)$ according to the uniform (Lebesgue) measure. By equation (1) of section B, the law of (Y_1, \dots, Y_N) corresponds to that of N independent coin tosses of a fair coin, for any positive integer N . It follows that Y_1, Y_2, \dots are mutually independent random variables.

The next result expresses the connection between independence and product measures in the framework of random variables.

Proposition D.1 Let $Z = (X_1, \dots, X_n)$ be a random vector. Then the following are equivalent.

- (a) The random variables X_1, \dots, X_n are independent.
- (b) The distribution measure \mathbb{F}_Z is the product measure:

$$\mathbb{F}_Z = \mathbb{F}_{X_1} \times \cdots \times \mathbb{F}_{X_n}.$$

- (c) For each integer n ,

$$F_{Z_n}(x_1, \dots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n).$$

Proof: The random variables X_1, \dots, X_n are independent if and only if for any Borel sets A_1, \dots, A_n

$$\mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n).$$

Written in terms of distribution measures, this equation says

$$\mathbb{F}_Z(A_1 \times \cdots \times A_n \times \mathbb{R}^\infty) = \mathbb{F}_{X_1}(A_1) \cdots \mathbb{F}_{X_n}(A_n),$$

and requiring this to be true for all n and choices of Borel sets A_i means, by definition, that \mathbb{F}_Z is the product of the \mathbb{F}_{X_i} . Thus we see that (a) and (b) are equivalent.

To see that (b) implies (c), note that from (b)

$$\begin{aligned}\mathbb{F}_{Z_n}(x_1, \dots, x_n) &= \mathbb{F}_Z((-\infty, x_1] \times (-\infty, x_n] \times \mathbb{R}^\infty) \\ &= \mathbb{F}_{X_1}((-\infty, x_1]) \cdots \mathbb{F}_{X_n}((-\infty, x_n]) = F_{X_1}(x_1) \cdots F_{X_n}(x_n).\end{aligned}$$

For the converse, (c) implies that for every n

$$\mathbb{F}_{Z_n} = \mathbb{F}_{X_1} \times \cdots \times \mathbb{F}_{X_n}.$$

The condition (1) for independence of random variables can be generalized into a statement about expectations of products of random variables. The generalization is a simple consequence of the product form of the joint law of independent random variables.

Proposition D.2 If $Z = (X_1, \dots, X_n)$ is a vector of independent random variables, and if h_1, \dots, h_n are Borel functions such that $E[|h_i(X_i)|] < \infty$, for $1 \leq i \leq n$, then

$$(2) \quad E[h_1(X_1) \cdots h_n(X_n)] = E[h_1(X_1)] \cdots E[h_n(X_n)]$$

Conversely, if (2) is true for all bound, Borel functions h_1, \dots, h_n , then X_1, \dots, X_n are independent.

Proof: By formula (1) and the identity $\mathbb{F}_Z = \mathbb{F}_{X_1} \times \cdots \times \mathbb{F}_{X_n}$, the left hand side of (2) is

$$\int_{\mathbf{R}^n} h_1(x_1) \cdots h_n(x_n) \mathbb{F}_{X_1}(dx_1) \cdots \mathbb{F}_{X_n}(dx_n).$$

By Fubini's theorem this equals

$$\int_{\mathbf{R}} h_1(x_1) \mathbb{F}(dx_1) \cdots \int_{\mathbf{R}} h_n(x_n) \mathbb{F}(dx_n),$$

which, by (1), is the right hand side of (2).

To prove the converse statement, observe that if we set $h_1(x) = \mathbf{1}_{A_1}, \dots, h_n(x) = \mathbf{1}_{A_n}$, for any Borel sets A_1, \dots, A_n in \mathbb{R} , then (2) implies

$$\mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n).$$

Thus X_1, \dots, X_n are independent. \diamond

Formula (2) of the previous proposition allows us to relate independence to the notion of correlation.

Corollary D.3 Independent random variables with finite variance are uncorrelated.

Proof: Let X and Y be independent with finite variance. By (2)

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[X - \mu_X]E[Y - \mu_Y] = 0. \quad \diamond$$

Here is another simple, but basic application of Propostion D.2.

Corollary D.4 Let X_1, \dots, X_n be independent, and let $S = X_1 + \dots + X_n$. Then the moment generating function $m_S(t)$ is the product of the moment generating functions $m_{X_i}(t)$ of the summands:

$$(3) \quad m_S(t) = m_{X_1}(t)m_{X_2}(t) \cdots m_{X_n}(t).$$

Likewise, the characteristic function $\phi_S(\lambda)$ of S is the product of characteristic functions:

$$(4) \quad \phi_S(t) = \phi_{X_1}(t)\phi_{X_2}(t) \cdots \phi_{X_n}(t).$$

Proof: Using (2) and the definition of moment generating function,

$$\begin{aligned} m_S(t) &= E[\exp\{t(X_1 + \dots + X_n)\}] = E\left[\prod_1^n e^{tX_i}\right] \\ &= \prod_1^n E[e^{tX_i}] = \prod_1^n m_{X_i}(t). \quad \diamond \end{aligned}$$

Example D.2 Let X_1, X_2, \dots be independent random variables, all with the same distribution, $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = 1 - p$. We think of the X_i random variables as recording the results of independent tosses of a coin with probability p of heads. Let $S_n = X_1 + \dots + X_n$ count the number of heads observed in the first n tosses. We know that $E[X_i] = p$ for each i , and hence $E[S_n] = np$. It is not hard to calculate that $\text{Var}(X_i) = p(1 - p)$, so, using (5) of Proposition C.4, $\text{Var}(S_n) = np(1 - p)$. Since for each i , $m_{X_i}(t) = (1 - p) + pe^t$, we obtain

$$m_{S_n}(t) = (1 - p + pe^t)^n.$$

Now consider the random variable

$$Z_n = \frac{S_n - np}{\sqrt{np(1 - p)}}.$$

Z_n is a scaling and centering of S_n , defined so that $\mu_Z = 0$ and $\text{Var}(Z_n) = 1$ for any n . Now we calculate from the formulas derived so far that

$$m_{Z_n}(t) = \left(e^{-pt/\sqrt{np(1-p)}} (1 - p + pe^{t/\sqrt{np(1-p)}}) \right)^n.$$

In order to understand how the law of Z_n behaves as $n \rightarrow \infty$, one can ask for the limit of $m_{Z_n}(t)$. An exercise shows that this limit is

$$(5) \quad \lim_{n \rightarrow \infty} m_{Z_n}(t) = \exp t^2/2.$$

In section E we shall show that $\exp t^2/2$ is the moment generating function of a probability distribution called the normal distribution, and we shall relate (5), at least at the heuristic level, to the Central Limit Theorem. \diamond

The Fourier transform of the convolution of two functions is the product of the Fourier transforms of the individual functions. Since equation (4) shows that the Fourier transform (characteristic function) of the distribution of a sum of independent random variables is the product of their individual Fourier transforms, it follows conversely that the distribution of the sum is a convolution of the individual probability distributions. We state this here as another basic theorem about independent random variables.

Given two probability measures \mathbb{F}_1 and \mathbb{F}_2 on the Borel sets of \mathbb{R} , define their convolution as

$$\mathbb{F}_1 * \mathbb{F}_2(A) := \int_{\mathbb{R}} \mathbb{F}_1(A - x) \mathbb{F}_2(dx).$$

It is easy to check directly that $\mathbb{F}_1 * \mathbb{F}_2$ is a probability measure and that convolution is commutative and associative.

Proposition D.5 If X_1, \dots, X_n are independent random variables,

$$\mathbb{F}_{X_1 + \dots + X_n} = \mathbb{F}_1 * \dots * \mathbb{F}_n.$$

PROOF: We prove the case $n = 2$; the proof for general n follows by induction. If X_1 and X_2 are independent, the law of their sum is the product of the law of X_1 and the law of X_2 . Thus, by Fubini's Theorem and the identity $\mathbf{1}_A(x_1 + x_2) = \mathbf{1}_{A-x_2}(x_1)$,

$$\begin{aligned} \mathbb{F}_{X_1 + X_2}(A) &= \int \int \mathbf{1}_A(x_1 + x_2) \mathbb{F}_{X_1}(dx_1) \mathbb{F}_{X_2}(dx_2) \\ &= \int \left(\int \mathbf{1}_{A-x_2}(x_1) \mathbb{F}_{X_1}(dx_1) \right) \mathbb{F}_2(dx_2) = \int \mathbb{F}_{X_1}(A - x_2) \mathbb{F}_2(dx_2). \quad \diamond \end{aligned}$$

E. Some basic probability laws

In this section we list a few basic probability laws. Our aim is to show how these laws all arise from a study of the simple model of independent coin flipping. On the way, we shall introduce some of the major limit themes of probability.

A random variable that has only two possible values is called a Bernoulli random variable. We shall say that a random variable X has the Bernoulli(p) distribution in the specific case, $\mathbb{P}(X = 0) = 1 - p$, $\mathbb{P}(X = 1) = p$. We shall also refer to a random variable with law, $\mathbb{P}(X = -1) = 1 - p$ and $\mathbb{P}(X = 1) = p$, as Bernoulli. The expectation of a Bernoulli(p) random variable X is simply p , and its variance is $p(1 - p)$.

Let $S_n = X_1 + \cdots + X_n$ be the sum of n independent Bernoulli(p) random variables. Then the probability mass function of S_n is

$$(1) \quad P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad 0 \leq k \leq n.$$

Any random variable Z with the same law is said to have the binomial(n, p) distribution. By taking expectation and variance of S_n , using Proposition C.4, we find that the mean and variance of a binomial(n, p) random variable are, respectively, np and $np(1-p)$.

Fix a p between 0 and 1. Early probabilists were interested in what the binomial distribution looks like for large n . Let X_1, X_2, \dots be independent Bernoulli(p), and define S_n as above. We have already stated the strong law of large numbers:

$$(2) \quad \mathbb{P}(\lim_{n \rightarrow \infty} S_n/n = p) = 1.$$

We remarked above that the strong law is a statement about an infinite product measure. We have not yet spoken about the *weak law of large numbers*. This says:

$$(3) \quad \lim_{n \rightarrow \infty} \mathbb{P}(|S_n/n - p| > \epsilon) = 0 \quad \forall \epsilon > 0.$$

Since $\mathbb{P}(|S_n/n - p| > \epsilon) = E[\mathbf{1}_{\{|S_n/n - p| > \epsilon\}}]$ and since, by (1), the indicator function in the expectation converges to 0, \mathbb{P} -almost surely, the weak law (3) follows from the strong law (2) by dominated convergence. However, a simple direct argument also shows the weak law. Indeed, observe that

$$\text{Var}(S_n/n) = \text{Var}(S_n)/n^2 = p(1-p)/n.$$

Thus from Tchebysheff's inequality, inequality (9) in section C,

$$\mathbb{P}(|S_n/n - p| \geq \epsilon) \leq \frac{p(1-p)}{n\epsilon^2},$$

which tends to 0 as $n \rightarrow \infty$, thus proving (3).

A more sophisticated argument, using the moment generating function, leads to an exponential decay rate in (3). From the calculation of $M_{S_n}(t)$ in Example D.2, the moment generating function of $Y_n := S_n/n - p$ is

$$m_{Y_n}(t) = e^{-pt} m_{S_n}(t/n) = e^{-pt} (1-p + pe^{t/n})^n = ((1-p)e^{-pt/n} + pe^{(1-p)t/n})^n.$$

A bit of calculus will show that

$$M_{Y_n}(t) \leq e^{t^2/8n},$$

and, hence following the argument of Proposition C.7 (see (11) of section C), for $\epsilon > 0$,

$$\mathbb{P}(S_n/n - p \geq \epsilon) \leq \exp\{-\epsilon t + t^2/8n\} \quad \forall t > 0.$$

Optimizing over t ,

$$(4) \quad \mathbb{P}(S/n - p \geq \epsilon) \leq e^{-2n\epsilon^2}.$$

Since $S_n - p \leq -\epsilon$ if and only if $1 - S_n/n - (1 - p) \geq \epsilon$, and since $1 - S_n/n$ is just binomial($n, 1 - p$), we also obtain from (4) that

$$\mathbb{P}(S/n - p \leq -\epsilon) \leq e^{-2n\epsilon^2}.$$

Putting the two together gives Chernoff's inequality:

$$(5) \quad \mathbb{P}(|S_n/n - p| \geq \epsilon) \leq 2 \exp\{-2n\epsilon^2\}.$$

The exponential decay rate in (5) is a big improvement over that derived from Tchebysheff's inequality using only the existence of second moments. Note that (5) is really a statement about the binomial(n, p) distribution, as opposed to (2), which is a theorem about an event in an infinite product probability space.

The scaling of S_n by a factor of $1/n$ produces a sequence of random variables whose variance tends to 0. Suppose now we form the random variable $\frac{S_n - np}{\sqrt{np(1-p)}}$. It is easy to check that this random variable has a mean of zero and a variance of 1 for every n . A major discovery of early probability theory (now about 200 years old!) was the Laplace-DeMoivre Central Limit Theorem: for every $-\infty \leq a < b \leq \infty$,

$$(6) \quad \lim_{n \rightarrow \infty} \mathbb{P}\left(a < \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right) = \int_a^b e^{-x^2/2} \frac{dx}{\sqrt{2\pi}}.$$

We shall use this fact to motivate the introduction of *normal* random variables. A random variable X has the standard normal distribution if the density of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

It is easy to check that if X is standard normal, it has mean 0 and variance 1, consistent with the Central Limit Theorem (6). We say that Z has the normal distribution $N(\mu, \sigma^2)$ if

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

A simple calculation says that if X is a standard normal, then $\sigma X + \mu$ is a $N(\mu, \sigma^2)$ normal; hence the mean and variance of a $N(\mu, \sigma^2)$ r.v. are μ and σ^2 , respectively. Normal random variables are also called Gaussian random variables. Note that the Central Limit Theorem is consistent with (5).

The central limit theorem (6) can be proved by direct calculation, using the formula (1) for the probability mass function of S_n and Stirling's formula for the asymptotic form of $n!$ for large n . We shall not undertake this proof. Rather we give only a heuristic argument,

showing why one can expect (6) to be true. Let Z be a standard normal random variable. Then one can show easily that its moment generating function is

$$m_X(t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} e^{tx} e^{-x^2/2} dx = e^{(t^2/2)}.$$

(Simply complete squares in the exponent of the integrand and pull out the constant term.) However, we saw in example D.2, that $e^{t^2/2}$ is precisely the limit of the moment generating functions of the normalized random variables $Z_n = \frac{S_n - np}{\sqrt{np(1-p)}}$. Alternatively expressed, if F_n is the cdf of Z_n , (6) says that

$$(7) \quad \lim_{n \rightarrow \infty} \int e^{tx} dF_n(x) = \frac{1}{\sqrt{2\pi}} \int e^{tx} e^{-x^2/2} dx$$

for all real t . The class of functions $x \rightarrow e^{tx}$ is sufficiently rich so that we expect (7) must imply (6). With enough care, these observations can be turned into a proof.

The Laplace-DeMoivre Central Limit Theorem, admits generalizations of very broad validity. In a rough sense, it turns out that the sum of a large number of small, independent random variables, none of which dominates the sum, will be approximately normal. Precise statements of this fact and further generalizations of it are a major branch of probability theory. The generality of the central limit phenomenon also supports the popularity of normal distributions in statistical modelling, when a random outcome is clearly the aggregate of a large number of small, roughly independent, competing random inputs.

We shall introduce one more distribution, the *Poisson* distribution, again as a limit of binomial probabilities. Suppose we are interested in modelling the number of arrivals of jobs to a queue, or customers to a service station, or the number of radioactive decay events in a mass of radioactive material, that take place in a given time interval. To fix ideas, let Z denote the number of arrivals of customers to a service station in the time interval $[0, 1]$, and set $\lambda = E[Z]$; to avoid triviality, we assume $0 < \lambda < 1$. Our goal is to derive a physically reasonable probability law for Z . If we imagine a large pool of potential customers, each deciding independently of the others whether to go to the service station, we see that the number of arrivals in disjoint time intervals should be independent. Also, assuming that customer behavior does not change over our unit time interval $[0, 1]$, the distribution of the number of arrivals in an interval should depend only on the length of the interval and not whether it occurs sooner or later. Thus, if we divide $[0, 1]$ into n equal subintervals and let X_i , $1 \leq i \leq n$ denote the number of arrivals in subinterval i , $\tilde{X}_1, \dots, \tilde{X}_n$ are independent and identically distributed random variables, taking values in the non-negative integers. Moreover, $E[\tilde{X}_i] = E[\tilde{X}_1] = \lambda/n$, because, $\lambda = E[Z] = E[\sum_1^n \tilde{X}_i] = nE[\tilde{X}_1]$. When n is large, so that the time intervals are very small, it is reasonable to suppose that the probability of two or more arrivals in any subinterval is very rare. In other words, each \tilde{X}_i is approximately a Bernoulli random variable X_i with mean λ/n . It follows that

$$Z = \tilde{X}_1 + \dots + \tilde{X}_n \approx X_1 + \dots + X_n$$

is approximately a binomial($n, \lambda/n$) random variable. Reasoning in this way, we propose to take as a model for Z the distribution

$$\begin{aligned}
 \mathbb{P}(Z = k) &= \lim_{n \rightarrow \infty} \mathbb{P}(X_1 + \cdots + X_n = k) \\
 (8) \qquad &= \lim_{n \rightarrow \infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
 &= \frac{\lambda^k}{k!} e^{-\lambda}, \quad \text{for non-negative integers } k.
 \end{aligned}$$

The derivation of the limit in the last line is left as an exercise. Any random variable with the distribution given in (8) is called a Poisson(λ) random variable.

F. Convergence concepts for random variables

Most major theorems of probability are limit theorems that describe the convergence properties of a sequence of random variables. In this section we introduce the definitions of convergence in probability, of almost sure convergence and of convergence of moments. We shall leave the notion of convergence in distribution, in which one asks for convergence of the distribution functions of the random variables, for later.

Definition F.1 Let $\{X_n\}$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

- (i) We say that X_n converges in probability to a random variable X as $n \rightarrow \infty$, written $(P)\lim_{n \rightarrow \infty} \{X_n\} = X$, if, for every $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

- (ii) We say that X_n converges almost-surely to X as $n \rightarrow \infty$, written (a.s.) $\lim_{n \rightarrow \infty} X_n = X$ if

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} X_n = X\right) = 1.$$

- (iii) Let $r > 0$. We say that X_n converges in r^{th} moment to X as $n \rightarrow \infty$, written $(L^r)\lim_{n \rightarrow \infty} X_n = X$, if

$$\lim_{n \rightarrow \infty} E[|X_n - X|^r] = 0.$$

Proposition F.1 (a) Almost sure convergence implies convergence in probability.

(b) Convergence in r^{th} moment implies convergence in probability.

Proof: (a) follows from the dominated convergence theorem and the identity

$$\mathbb{P}(|X_n - X| > \epsilon) = E[1_{|X_n - X| > \epsilon}].$$

Statement (b) is a consequence of Markov's inequality, which implies

$$\mathbb{P}(|X_n - X| > \epsilon) \leq \frac{1}{\epsilon^r} E[|X_n - X|^r].$$

Without further conditions, there are no other implications among the different types of convergence.

Excercise: Give examples of sequences which (a) converge in probability but not a.s. (almost surely) or in 1st moment, (b) converge a.s. but not in 1st moment, (c) converge in 1st moment but not a.s.

We shall develop further relationships between the different types of convergence as the need and the necessary tools of proof arise. In analysis, convergence in probability is called convergence in measure, and almost sure convergence is called almost everywhere convergence.

G. Stochastic processes and Kolmogorov's consistency theorem

A *stochastic process* is a collection of random variables, $\{X_\alpha ; \alpha \in \mathcal{A}\}$ on a common probability space, where \mathcal{A} is some index set. In practice, \mathcal{A} is usually a subset of the integers or of the real line, with \mathbb{N} and $[0, \infty)$ being particularly common choices in which the index represents, respectively, the discrete or the continuous flow of time. To any stochastic process $\{X_\alpha ; \alpha \in \mathcal{A}\}$, we associate its *family of finite-dimensional distributions*,

$$\{F_{X_{\alpha_1}, \dots, X_{\alpha_m}} ; \alpha_1, \dots, \alpha_m \in \mathcal{A}, m \geq 1\}$$

Our purpose in this section is to show that this family gives a complete statistical description of the stochastic process, just as a single distribution function provides a complete statistical description of a random vector. In particular, we pose the following converse problem. For an index set \mathcal{A} and a family

$$\mathcal{J} = \{F_{\alpha_1, \dots, \alpha_m} ; \alpha_1, \dots, \alpha_m \in \mathcal{A}, m \geq 1\}$$

of distribution functions, when does there exist a stochastic process corresponding to this family in the sense that,

$$F_{X_{\alpha_1}, \dots, X_{\alpha_m}} = F_{\alpha_1, \dots, \alpha_m}$$

for every finite subset $\{\alpha_1, \dots, \alpha_m\} \subset \mathcal{A}$?

Let us first answer this question for the case $\mathcal{A} = \mathbb{N}$, which, it turns out, is enough to treat the general case. We therefore suppose given a family $\mathcal{J} = \{F_n ; n \geq 1\}$ of distribution functions, where F_n is a distribution function on \mathbb{R}^n for each n , and we ask for a process $\{X_n ; n \geq 1\}$ such that F_n is the distribution of (X_1, \dots, X_n) for each n . We begin by noting that there is a simple consistency condition that must necessarily be satisfied if an associated stochastic process exists; namely, for every n ,

$$(1) \quad F_n(x_1, \dots, x_n) = F_{n+1}(x_1, \dots, x_n, \infty) := \lim_{z \rightarrow \infty} F_{n+1}(x_1, \dots, x_n, z).$$

This must be so because the right-hand side must represent $\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n, X_{n+1} < \infty)$, which is the same as $\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n)$, which in turn is equal to

the left-hand side. We shall say that $\{F_n ; n \geq 1\}$ is a consistent family if (1) is true for every n . The main theorem says that consistency is in fact sufficient for the existence of an associated stochastic process.

Theorem G.1 (Kolmogorov) Let \mathcal{J} be a consistent family of probability distribution functions. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a stochastic process $\{X_n\}$ on it such that the finite dimensional distributions of $\{X_n\}$ are given by \mathcal{J} .

To sketch the proof we first develop a little machinery. We shall let \overline{R} denote the two point compactification of the real line R obtained by adding to R the points at $\{\infty\}$ and $\{-\infty\}$. We shall work with \overline{R} rather than R for technical reasons; namely, it will allow us to use Tychonov's theorem as described later. However it is not strictly necessary. A set A is in the Borel σ -algebra of \overline{R} if and only if A is either a Borel set of R or A is the union of one or both of $\{\infty\}$ and $\{-\infty\}$ with a Borel set of R . We shall denote the Borel σ -algebra by $\mathcal{B}(\overline{R})$. Further, we shall let $\mathcal{B}(\overline{R})^n$ denote the Borel σ -algebra of \overline{R}^n .

We shall want our distribution measures to be defined on \overline{R} rather than R . Clearly, any probability measure \mathbb{F} on $(R^n, \mathcal{B}(R^n))$ can be extended to a probability measure on $(\overline{R}^n, \mathcal{B}(\overline{R}^n))$ by the simple formula

$$(2) \quad \mathbb{F}(A) = \mathbb{F}(A \cap R^n) \quad \text{for any Borel set } A \in \mathcal{B}(\overline{R})^n.$$

We shall think of a stochastic process $\{X_n\}$ as an infinite random vector (X_1, X_2, \dots) , taking values in the infinite product space

$$\overline{R}^\infty := \{\omega = (\omega_1, \omega_2, \dots) \mid \omega_i \in \overline{R}, \forall i \geq 1\} = \overline{R} \times \overline{R} \times \dots$$

In the proof, this product space shall carry the infinite dimensional analogue of a distribution measure for the stochastic process. Next let \mathcal{A} denote the class of all subsets A of \overline{R}^∞ for which there exist an integer n and a Borel set of \overline{R}^n so that

$$(3) \quad A = \{\omega = (\omega_1, \omega_2, \dots) \mid (\omega_1, \dots, \omega_n) \in B\}.$$

It is not hard to check that \mathcal{A} is an algebra. $\sigma(\mathcal{A})$ shall provide the σ -algebra for \overline{R}^∞ .

Finally, we review briefly Tychonov's theorem. The finite-dimensional projection of \overline{R}^∞ on its first n components is the map

$$\pi_n(\omega) = (\omega_1, \dots, \omega_n).$$

The *product topology* on \overline{R}^∞ is the minimal topology that makes the projection map π_n continuous for every n . The Borel σ -algebra of \overline{R}^∞ with this topology is the minimal σ -algebra containing the open sets, and we shall denote it by $\mathcal{B}(\overline{R})^\infty$.

Proof of Theorem G.1 : From section B we know that to the distribution function $F_{1, \dots, n}$ in the family \mathcal{J} there corresponds a unique distribution measure \mathbb{F}_n on $(R^n, \mathcal{B}(R^n))$. Extend this to $(\overline{R}^n, \mathcal{B}(\overline{R})^n)$ as in equation (2). Now define a measure $\underline{\mathbb{F}}$ on \mathcal{A} by the rule

$$\underline{\mathbb{F}}(A) = \mathbb{F}_n(B),$$

when A is defined in terms of $B \in \mathcal{B}(\overline{R}^n)$ as in (3). This rule leads to a well-defined, finitely additive measure on \mathcal{A} because the family \mathcal{J} is consistent.

We want to prove that $\underline{I}F$ is continuous from above at \emptyset . First note that if B is any Borel set in \overline{R}^n and ϵ is any positive constant, there is a compact set C contained in B such that $\underline{I}F_n(B - C) < \epsilon$. This comes from a general measure theoretic fact about Baire measures, of which $\underline{I}F_n$ is an example; see, Halmos, *Measure Theory*, see section 52. You can prove it much more simply (exercise!) for $\underline{I}F_n$ by the monotone class theorem, by first proving it is true on the algebra of finite disjoint unions of rectangles in \overline{R}^n .

Now we shall show that, given any A in \mathcal{A} and any positive ϵ , there exists a compact (in the product topology) E in \mathcal{A} such that $\underline{I}F(A - E) < \epsilon$. Indeed, let $A \in \mathcal{A}$ be given as in equation (8) where $B \subset \overline{R}^n$. Let $C \subset B$ be compact and satisfy $F_n(B - C) < \epsilon$, and set

$$\tilde{C} = \{ \omega \mid (\omega_1, \dots, \omega_n) \in C \} = C \times \overline{R}^\infty.$$

Clearly $\tilde{C} \subset A$, $\tilde{C} \in \mathcal{A}$, and $\underline{I}F(A - \tilde{C}) < \epsilon$. By Tychonov's theorem, \tilde{C} is compact in the product topology as a subset of \overline{R}^∞ .

Now we can complete the proof. Fix any $\epsilon > 0$. Let $\{A_n\}$ be a sequence of sets in \mathcal{A} decreasing to \emptyset , and let $\{B_n\}$ be a sequence of compact sets in \mathcal{A} , satisfying $B_n \subset A_n$ and $\underline{I}F(A_n - B_n) < \epsilon 2^{-n}$. Then

$$\bigcap_{n=1}^{\infty} B_n \subset \bigcap_{n=1}^{\infty} A_n = \emptyset,$$

and hence, by the finite intersection property, there exists an N such that $\bigcap_1^N B_n = \emptyset$. We that

$$\underline{I}F(A_N) = \underline{I}F(A_N - \bigcap_1^N B_n) < \epsilon.$$

Since $\epsilon > 0$ is arbitrary $\underline{I}F(A_n) \downarrow 0$.

To complete the proof, note that Carathéodory's theorem implies that there exists a unique extension I F of $\underline{I}F$ to $\mathcal{B}(\overline{R})^\infty$. $(\overline{R}^\infty, \mathcal{B}(\overline{R})^\infty, I$ F) is a probability space. Let $\{X_n\}$ be the stochastic process on this probability space defined by $X_n(\omega) = \omega_n$ for each n ; this is called the *canonical process*. Then the finite dimensional distributions of $\{X_n\}$ are given by the family \mathcal{J} . \diamond

Remark: A second proof of the continuity from above at the empty set can be patterned on the proof of continuity from above at the empty set in the construction of the product measure for infinite, independent tosses of a fair coin in Chapter 1, Example C.3.