2

Random Variables

The standard situation in the modeling of a random phenomenon is that the quantities of interest, rather than being defined on the underlying probability space, are *functions* from the probability space to some other (measurable) space. These functions are called *random variables*. Strictly speaking, one uses the term random variable when they are functions from the probability space to $\mathbb{R}$. If the image is in $\mathbb{R}^n$ for some $n \geq 2$ one talks about n-dimensional random variables or simply random vectors. If the image space is a general abstract one, one talks about random elements.

1 Definition and Basic Properties

Let $(\Omega, \mathcal{F}, P)$ be a probability space.

Definition 1.1. *A* random variable X *is a measurable function from the sample space Ω to $\mathbb{R}$;*

$$X : \Omega \to \mathbb{R},$$

that is, the inverse image of any Borel set is $\mathcal{F}$-measurable:

$$X^{-1}(A) = \{\omega : X(\omega) \in A\} \in \mathcal{F} \quad \text{for all} \quad A \in \mathcal{R}.$$

We call X simple *if, for some n,*

$$X = \sum_{k=1}^{n} x_k I\{A_k\},$$

where $\{x_k, 1 \leq k \leq n\}$ are real numbers, and $\{A_k, 1 \leq k \leq n\}$ is a finite *partition of Ω, that is $A_i \cap A_j = \emptyset$ if $i \neq j$ and $\cup_{k=1}^{n} A_k = \Omega$.*

We call X elementary *if*

$$X = \sum_{n=1}^{\infty} x_n I\{A_n\},$$

where $\{x_n,\, n \geq 1\}$ *are real numbers, and* $\{A_n,\, n \geq 1\}$ *is an* infinite *partition of* Ω.

If $X : \Omega \to [-\infty, +\infty]$ *we call* X *an* extended *random variable.* □

Random variables are traditionally denoted by large capitals toward the end of the alphabet; X, Y, Z, U, V, W, ... For sequences of "similar kinds" it is convenient to use indices; X_1, X_2, ..., and so on.

We do not distinguish between random variables that differ on a null set.

Definition 1.2. *Random variables which only differ on a null set are called* equivalent

The equivalence class *of a random variable* X *is the collection of random variables that differ from* X *on a null set.*

If X *and* Y *are equivalent random variables we write* $X \sim Y$. □

So far we have described the map from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{R})$. In order to complete the picture we must find a third component in the triple – the appropriate probability measure.

To each random variable X we associate an *induced probability measure*, $\mathbb{P}$, through the relation

$$\mathbb{P}(A) = P(X^{-1}(A)) = P(\{\omega : X(\omega) \in A\}) \quad \text{for all} \quad A \in \mathcal{R}. \tag{1.1}$$

In words this means that we define the probability (on $(\mathbb{R}, \mathcal{R})$) that the random variable X falls into a Borel set as the probability (on $(\Omega, \mathcal{F})$) of the inverse image of this Borel set. This is the motivation for the measurability assumption.

That the definition actually works is justified by the following result.

Theorem 1.1. *The* induced *space* $(\mathbb{R}, \mathcal{R}, \mathbb{P})$ *with* $\mathbb{P}$ *defined by (1.1) is a probability space – the* induced *probability space.*

Proof. The proof amounts to checking the Kolmogorov axioms, which, in turn, amounts to going back and forth between the two probability spaces.

1. $\mathbb{P}(A) = P(\{\omega : X(\omega) \in A\}) \geq 0$ for any $A \in \mathcal{R}$.

2. $\mathbb{P}(X) = P(\{\omega : X(\omega) \in \Omega\}) = 1$.

3. Suppose that $\{A_n,\, n \geq 1\}$ are disjoint subsets of $\mathcal{R}$. Then

$$\mathbb{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) = P\bigg(\Big\{\omega : X(\omega) \in \bigcup_{n=1}^{\infty} A_n\Big\}\bigg) = P\bigg(\bigcup_{n=1}^{\infty} \{\omega : X(\omega) \in A_n\}\bigg)$$
$$= \sum_{n=1}^{\infty} P(\{\omega : X(\omega) \in A_n\}) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

□

Remark 1.1. Once one gets used to the fact that only the random variables are of interest one need no longer worry about the exact probability space behind the random variables. In the remainder of this book we therefore refrain from

distinguishing between the probability measures P and $\mathbb{P}$, and we omit the brackets { and } to emphasize that $\{X \in A\}$ actually is a *set*. So, instead of writing $\mathbb{P}(A)$ we shall write $P(X \in A)$. There is absolutely no danger of confusion! □

Definition 1.3. *A* degenerate *random variable is constant with probability 1. Thus, X is degenerate if, for some $a \in \mathbb{R}$, $P(X = a) = 1$. A random variable that is not degenerate is called* non-degenerate. □

There are different ways to interpret the equality $X = Y$. The random variables X and Y are *equal in distribution* iff they are governed by the same probability measure:

$$X \stackrel{d}{=} Y \quad \Longleftrightarrow \quad P(X \in A) = P(Y \in A) \quad \text{for all} \quad A \in \mathcal{R}$$

and they are *point-wise equal*, iff they agree for almost all elementary events:

$$X \stackrel{a.s.}{=} Y \quad \Longleftrightarrow \quad P(\{\omega : X(\omega) = Y(\omega)\}) = 1,$$

i.e., X and Y are equivalent random variables, $X \sim Y$.

Next we provide an example to illustrate that there is a clear difference between the two equality concepts. The following example shows that two random variables, in fact, may well have the same distribution, and at the same time there is no elementary event where they agree.

Example 1.1. Toss a fair coin once and set

$$X = \begin{cases} 1, & \text{if the outcome is heads,} \\ 0, & \text{if the outcome is tails,} \end{cases}$$

and

$$Y = \begin{cases} 1, & \text{if the outcome is tails,} \\ 0, & \text{if the outcome is heads.} \end{cases}$$

Clearly, $P(X = 1) = P(X = 0) = P(Y = 1) = P(Y = 0) = 1/2$, in particular, $X \stackrel{d}{=} Y$. But $X(\omega)$ and $Y(\omega)$ differ for every ω. □

Exercise 1.1. Prove that if $X(\omega) = Y(\omega)$ for almost all ω, then $X \stackrel{d}{=} Y$. □

For X to be a random variable one has to check that the set $\{\omega : X(\omega) \in A\} \in \mathcal{F}$ for all $A \in \mathcal{R}$. However, as a consequence of Theorem 1.3.6 it suffices to check measurability for (e.g.) all sets of the form $(-\infty, x]$; why? This important fact deserves a separate statement.

Theorem 1.2. *X is a random variable iff*

$$\{\omega : X(\omega) \leq x\} \in \mathcal{F} \quad \textit{for all} \quad x \in \mathbb{R}.$$

1.1 Functions of Random Variables

A random variable is, as we have seen, a function from one space (Ω) to another space ($\mathbb{R}$). What can be said of a (real valued) function of a random variable? Since, we know from analysis that a function of a function is a function, the following result does not come to us as a surprise.

Theorem 1.3. *A Borel measurable function of a random variable is a random variable, viz., if g is a real, Borel measurable function and X a random variable, then $Y = g(X)$ is a random variable.*

Proof. The proof follows, in fact, from the verbal statement, since Y is a composite mapping from Ω "via $\mathbb{R}$" to $\mathbb{R}$. A more detailed proof of this is that, for any $A \in \mathcal{R}$,

$$\begin{aligned}\{Y \in A\} &= \{\omega : Y(\omega) \in A\} = \{\omega : g(X(\omega)) \in A\} \\ &= \{\omega : X(\omega) \in g^{-1}(A)\} \in \mathcal{F}.\end{aligned}$$

□

By taking advantage of Theorem 1.2 we can prove measurability of the following functions, that is, we can prove that the following objects are, indeed, random variables.

Proposition 1.1. *Suppose that $X_1, X_2, \ldots$ are random variables. The following quantities are random variables:*

(a) $\max\{X_1, X_2\}$ *and* $\min\{X_1, X_2\}$*;*
(b) $\sup_n X_n$ *and* $\inf_n X_n$*;*
(c) $\limsup_{n\to\infty} X_n$ *and* $\liminf_{n\to\infty} X_n$*.*
(d) *If $X_n(\omega)$ converges for every ω as $n \to \infty$, then* $\lim_{n\to\infty} X_n$ *is a random variable.*

Proof. (a): For any x,

$$\begin{aligned}\{\omega : \max\{X_1, X_2\}(\omega) \le x\} &= \{\omega : \max\{X_1(\omega), X_2(\omega)\} \le x\} \\ &= \{\omega : X_1(\omega) \le x\} \cap \{\omega : X_2(\omega) \le x\}\end{aligned}$$

and

$$\begin{aligned}\{\omega : \min\{X_1, X_2\}(\omega) \le x\} &= \{\omega : \min\{X_1(\omega), X_2(\omega)\} \le x\} \\ &= \{\omega : X_1(\omega) \le x\} \cup \{\omega : X_2(\omega) \le x\},\end{aligned}$$

which proves (a), since an intersection and a union, respectively, of two measurable sets are measurable.
(b): Similarly,

$$\{\omega : \sup_n X_n(\omega) \le x\} = \bigcap_n \{\omega : X_n(\omega) \le x\} \in \mathcal{F},$$

since a countable intersection of measurable sets is measurable, and

$$\{\omega : \inf_n X_n(\omega) < x\} = \bigcup_n \{\omega : X_n(\omega) < x\} \in \mathcal{F},$$

since a countable union of measurable sets is measurable.

(c): In this case,

$$\{\omega : \limsup_{n\to\infty} X_n(\omega) \le x\} = \{\omega : \inf_n \sup_{m\ge n} X_m(\omega) \le x\}$$
$$= \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} \{\omega : X_m(\omega) \le x\} \in \mathcal{F},$$

and

$$\{\omega : \liminf_{n\to\infty} X_n(\omega) < x\} = \{\omega : \sup_n \inf_{m\ge n} X_m(\omega) < x\}$$
$$= \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} \{\omega : X_m(\omega) < x\} \in \mathcal{F},$$

since, once again, we have only performed legal operations.

Alternatively, since $\sup_{n\ge m} X_n(\omega)$ is a random variable by (b), it follows, also by (b), that $\inf_m \left(\sup_{n\ge m} X_n(\omega)\right)$ is a random variable. Similarly for $\liminf_{n\to\infty} X_n(\omega)$.
(d): This is true, because in this case lim sup and lim inf coincide (and both are measurable). Moreover, the limit exists and is equal to the common value of lim sup and lim inf. □

Exercise 1.2. Prove that a continuous function of a random variable is a random variable. □

The usual construction procedure for properties of functions, or for constructions of new (classes of) functions, is to proceed from non-negative simple functions, sometimes via elementary functions, to non-negative functions, to the general case. This works for random variables too. The following lemma is closely related to Lemma A.9.3 which deals with the approximation of real valued functions.

Lemma 1.1. (i) *For every non-negative random variable X there exists a sequence $\{X_n,\, n \ge 1\}$ of non-negative simple variables, such that*

$$X_n(\omega) \uparrow X(\omega) \quad \text{for all} \quad \omega \in \Omega.$$

(ii) *For every random variable X there exists a sequence $\{X_n,\, n \ge 1\}$ of simple variables, such that*

$$X_n(\omega) \to X(\omega) \quad \text{for all} \quad \omega \in \Omega.$$

Proof. (i): Let $n \ge 1$, and set

$$X_n(\omega) = \begin{cases} \frac{k-1}{2^n}, & \text{for} \quad \frac{k-1}{2^n} \le X(\omega) < \frac{k}{2^n}, \quad k = 1, 2, \ldots, n2^n, \\ n, & \text{for} \quad X(\omega) \ge n. \end{cases} \tag{1.2}$$

The sequence thus constructed has the desired property because of the dyadic construction and since

$$X(\omega) - X_n(\omega) < \frac{1}{2^n} \quad \text{for } n \text{ sufficiently large.}$$

The limit is a random variable by Proposition 1.1. This proves (i).

To prove (ii) we use the mirrored approximation and the fact that $X = X^+ - X^-$. □

2 Distributions

In analogy with the arguments that preceded Theorem 1.2, the complete description of the distribution of a random variable X would require knowledge about $P(X \in A)$ for all sets $A \in \mathcal{R}$. And, once again, the fact that the intervals $(-\infty, x]$ generate $\mathcal{R}$ comes to our rescue. This fact is manifested by introducing the concept of *distribution functions*.

2.1 Distribution Functions

Definition 2.1. *Let X be a real valued random variable. The* distribution function *of X is*

$$F(x) = P(X \le x), \quad x \in \mathbb{R}.$$

The continuity set *of F is*

$$C(F) = \{x : F(x) \text{ is continuous at } x\}. \qquad \square$$

Whenever convenient we index a distribution function by the random variable it refers to; F_X, F_Y, and so on.

Definition 2.2. *The distribution function of a degenerate random variable is a* degenerate *distribution function. If F is degenerate, then, for some $a \in \mathbb{R}$,*

$$F(x) = \begin{cases} 0, & \text{for} \quad x < a, \\ 1, & \text{for} \quad x \ge a. \end{cases}$$

A distribution function that is not degenerate is called non-degenerate. □

In order to describe the properties of distribution functions it is convenient to introduce the following class of functions.

Definition 2.3. *The class D is defined as the set of right-continuous, real valued functions with left-hand limits;*

$$F(x-) = \lim_{\substack{x_n \nearrow x \\ x_n < x}} F(x_n).$$

The class D^+ is defined as the set of non-decreasing functions in D^+. □

Proposition 2.1. *Let F be a distribution function. Then*

(a) $F \in D^+$;
(b) $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to\infty} F(x) = 1$;
(c) *F has at most a countable number of discontinuities.*

Proof. (a): Let X be a random variable associated with F. Boundedness follows from the fact that $0 \le F(x) = P(X \le x) \le 1$ for all x. To see that F is non-decreasing, let $x \le y$. Then $\{X \le x\} \subset \{X \le y\}$, so that

$$F(x) = P(X \le x) \le P(X \le y) = F(y).$$

Next, let x_n, $n \ge 1$, be reals, $x_n \searrow x$ as $n \to \infty$. Then $\{X \le x_n\} \searrow \{X \le x\}$, so that by monotonicity (Theorem 1.3.1)

$$F(x_n) = P(X \le x_n) \searrow P(X \le x) = F(x),$$

which establishes right-continuity.

In order to verify left-continuity, we let y_n, $n \ge 1$, be reals, such that $y_n \nearrow x$ as $n \to \infty$. Then $\{X \le y_n\} \nearrow \{X < x\}$, so that by monotonicity (Theorem 1.3.1),

$$F(y_n) = P(X \le y_n) \nearrow P(X < x) = F(x-).$$

This concludes the proof of the fact that $F \in D^+$.
(b): This follows from the set convergences $\{X \le x\} \searrow \emptyset$ as $x \to -\infty$, and $\{X \le x\} \nearrow \Omega$ as $x \to +\infty$, respectively, together with Theorem 1.3.1.
(c): Immediate from Lemma A.9.1(i). □

Remark 2.1. We shall, at times, encounter non-negative functions in D^+ with total mass *at most* equal to 1. We shall call such functions *sub-probability distribution functions*. They can be described as distribution functions, except for the fact that the total mass need not be equal to 1. □

To complement the proposition, we find that with x_n and y_n as given there, we have

$$P(y_n < X \le x_n) \to \begin{cases} F_X(x) - F_X(x-) \\ P(X = x) \end{cases} \qquad \text{as} \quad n \to \infty.$$

Proposition 2.2. *Let X be a random variable. Then*

$$P(X = x) = F_X(x) - F_X(x-).$$

In order to prove uniqueness it sometimes suffices to check a generator or a dense set. Following are some results of this kind.

Proposition 2.3. *Suppose that F and G are distribution functions, and that $F = G$ on a dense subset of the reals. Then $F = G$ for all reals.*

Proof. Combine Proposition 2.1 and Lemma A.9.1(ii). □

In Chapter 1 we discussed probability measures. In this chapter we have introduced distribution functions of random variables. Now, to any given probability measure $\mathbb{P}$ on $(\mathbb{R}, \mathcal{R})$ we can associate a distribution function F via the relation

$$F(b) - F(a) = \mathbb{P}((a, b]) \quad \text{for all } a, b, \qquad -\infty < a \leq b < \infty,$$

and since $\mathbb{P}$ as well as F is uniquely defined by their values on the rectangles we have established the following equivalence.

Theorem 2.1. *Every probability measure on* $(\mathbb{R}, \mathcal{R})$ *corresponds uniquely to the distribution function (of some random variable(s)).*

Remark 2.2. Recall from Example 1.1 that different random variables may have coinciding distribution functions. □

Having defined distribution functions, a natural challenge is to determine how many different kinds or types there may exist. For example, the number of dots resulting after throwing dice, the number of trials until a first success, the number of customers that visit a given store during one day, all those experiments have non-negative integers as outcomes. Waiting times, durations, weights, and so on are continuous quantities. So, there are at least two kinds of random variables or distributions. Are there any others?

Well, there also exist mixtures. A simple mixture is the waiting time at a traffic light. With some given probability the waiting time is 0, namely if the light is green upon arrival. If the light is red the waiting time is some continuous random quantity. So, there exist at least two kinds of random variables and mixtures of them. Are there any others?

The main decomposition theorem states that there exist exactly three kinds of random variables, and mixtures of them. However, before we turn to that problem in Subsection 2.2.3 we need some additional terminology.

2.2 Integration: A Preview

The classical integral is the Riemann integral, which was later generalized to the Riemann-Stieltjes integral. However, it turns out that the Riemann-Stieltjes integral has certain deficiencies that are overcome by another integral, the Lebesgue integral. The problem is that we need to be able to integrate certain wild functions that the Riemann-Stieltjes integral cannot handle. After having defined the Lebesgue integral we shall exhibit a perverse example that is Lebesgue integrable, but not Riemann-Stieltjes integrable.

There exists a probabilistic analog to integration, called *expectation*, denoted by the letter E. Now, instead of describing and proving a number of properties for functions and integrals and then translating them into statements about random variables and expectations (which basically amounts to

replacing f by X and $\int$ by E), we shall develop the theory in a probabilistic framework, beginning in Section 2.4. However, since we need some (not much) terminology earlier, we present a few definitions and facts without proof in the language of mathematics already here. The reader who is eager for proofs is referred to standard books on measure theory and/or function theory. The amount of details and the choice of which statements to prove and which to "leave as exercises" varies between books.

The Riemann-Stieltjes Integral

From analysis we remember that the *Riemann integral* of a function g on a bounded interval $(a, b]$ is defined via a partition $\mathbf{\Delta}$ of the interval into disjoint subintervals;

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b,$$

and the Riemann sums

$$R(n) = \sum_{j=1}^{n} g(t_j)\Delta_j,$$

where $\Delta_j = x_j - x_{j-1}$, and $t_j \in (x_{j-1}, x_j]$. The *mesh* of the partition is $\|\mathbf{\Delta}\| = \max_{1\le k\le n}\{\Delta_k\}$.

The integral exists iff there exists a number A, such that

$$\lim_{\|\mathbf{\Delta}\|\to 0} |R(n) - A| \to 0,$$

for any partition and arbitrary intermediate points. The limit is denoted with the aid of the integral sign:

$$A = \int_a^b g(x)\,\mathrm{d}x.$$

If the integral exists we may, in particular, select the t_j's so that g always assumes its maximum in the subintervals, and also such that the minimum is attained. As a consequence the actual value A is sandwiched between those two special sums, the upper sum and the lower sum.

We also note that, for simple functions, the Riemann integral coincides with the Riemann sum (let the partition coincide with the steps).

In the definition of the *Riemann-Stieltjes integral* of a function f on a bounded interval one replaces the Δ-differences along the x-axis by differences of a function. Thus, let, in addition, γ be a real valued function on the interval $(a, b]$, let the partition be defined as before, and (or but) set $\Delta_j = \gamma(x_j) - \gamma(x_{j-1})$. The *Riemann-Stieltjes sum* is

$$RS(n) = \sum_{j=1}^{n} g(t_j)\Delta_j = \sum_{j=1}^{n} g(t_j)(\gamma(x_j) - \gamma(x_{j-1})),$$

and the Riemann-Stieltjes integral exists iff there exists a number A, such that

$$\lim_{\|\boldsymbol{\Delta}\|\to 0} |RS(n) - A| \to 0$$

for any partition and arbitrary intermediate points. The notation is

$$A = \int_a^b g(x)\,\mathrm{d}\gamma(x).$$

Once again, we may select the points of the partition in such a way that the actual value A can be sandwiched between an upper sum and a lower sum.

As for existence criteria we mention without proof that the Riemann-Stieltjes integral exists if (for example) g is continuous and γ is bounded and non-decreasing. The integral is then suitably extended to all of $\mathbb{R}$. The interesting example is that distribution functions fit this requirement for γ.

An inspection of the definition and the limiting procedure shows that

- if γ is discrete with point masses $\{x_j\}$, then

$$\int g(x)\,\mathrm{d}\gamma(x) = \sum_k g(x_k)\gamma(\{x_k\});$$

- if γ is absolutely continuous with density $f(x)$, then

$$\int g(x)\,\mathrm{d}\gamma(x) = \int g(x)f(x)\,\mathrm{d}x.$$

For the latter conclusion we also lean on the mean value theorem.

In addition, by departing from the approximating Riemann-Stieltjes sum and partial summation, one obtains a formula for partial integration:

$$\int_a^b g(x)\,\mathrm{d}\gamma(x) = g(b)\gamma(b) - g(a)\gamma(a) - \int_a^b \gamma(x)\,\mathrm{d}g(x).$$

And, needless to say, if $\gamma(x) = x$ the Riemann-Stieltjes integral reduces to the ordinary Riemann integral.

The Lebesgue Integral

Paralleling the notion of a simple random variable (Definition 1.1) we say that f is *simple* real valued function if, for some n,

$$f = \sum_{k=1}^{n} x_k I\{A_k\},$$

where $\{x_k, 1 \le k \le n\}$ are real numbers, and $\{A_k, 1 \le k \le n\}$ is a *finite* partition of $\mathbb{R}$.

We call f *elementary* if

$$f = \sum_{n=1}^{\infty} x_n I\{A_n\},$$

where $\{x_n,\, n \geq 1\}$ are real numbers, and $\{A_n,\, n \geq 1\}$ is an *infinite* partition of $\mathbb{R}$.

The following lemma is a translation of Lemma 1.1; cf. also Lemma A.9.3.

Lemma 2.1. (i) *For every non-negative function f there exists a sequence $\{f_n,\, n \geq 1\}$ of non-negative simple functions, such that*

$$f_n \uparrow f \quad \textit{point-wise.}$$

(ii) *For every function f there exists a sequence $\{f_n,\, n \geq 1\}$ of simple functions, such that*

$$f_n \to f \quad \textit{point-wise.}$$

Proof. For (i) we set

$$f_n(x) = \begin{cases} \frac{k-1}{2^n}, & \text{for} \quad \frac{k-1}{2^n} \leq f(x) < \frac{k}{2^n}, \quad k = 1, 2, \ldots, n2^n, \\ n, & \text{for} \quad f(x) \geq n, \end{cases}$$

and for (ii) we add the mirrored version, and apply $f = f^+ - f^-$. □

The Lebesgue integral is an integral with respect to Lebesgue measure.

Definition 2.4. *The* Lebesgue measure*, λ, is a measure on $(\mathbb{R}, \mathcal{R})$, satisfying*

$$\lambda((a, b]) = b - a \quad \textit{for all} \quad a < b,\ a, b \in \mathbb{R}.$$

Definition 2.5. *For the simple function $f = \sum_{k=1}^{n} x_k I\{A_k\}$ we define the* Lebesgue integral *with respect to a probability measure λ as*

$$\int f \, \mathrm{d}\lambda = \sum_{k=1}^{n} x_k \lambda(A_k). \qquad \square$$

After proving several properties such as additivity and monotonicity one defines the Lebesgue integral of arbitrary non-negative functions as the limit of the integrals of the simple functions defined in the proof of Lemma 1.1:

$$\int f \, \mathrm{d}\lambda = \lim_{n \to \infty} \sum_{k=1}^{n2^n} \frac{k-1}{2^n} \, \lambda\Big(\frac{k-1}{2^n} \leq f(x) < \frac{k}{2^n}\Big).$$

Since, as mentioned in the introduction of this section, we shall traverse the theory in the probabilistic language in a moment, we close the discussion in mathematical terms with some comments on how the Lebesgue integral and the Riemann-Stieltjes integral relate to each other.

Theorem 2.2. *If the Riemann-Stieltjes integral of a function exists, then so does the Lebesgue integral, and both agree.*

We close this preview (i) by recalling that what has been stated so far will soon be justified, and (ii) with an example of a function that is Lebesgue integrable but *not* Riemann-Stieltjes integrable.

Example 2.1. Let $f(x)$ be defined as follows on the unit interval:

$$f(x) = \begin{cases} 1, & \text{for} \quad x \in [0,1] \setminus \mathbb{Q}, \\ 0, & \text{for} \quad x \in [0,1] \cap \mathbb{Q}, \end{cases}$$

that is, f equals 1 on the irrationals and 0 on the rationals.

This function is Lebesgue integrable – the integral equals 1 – but not Riemann integrable, the reason for the latter being that the upper and lower sums equal 1 and 0, respectively, for any partition of the unit interval.

The explanation for the difference in integrability is that the "slices" in the definition of the Lebesgue integral are horizontal, whereas those of the Riemann integral are vertical. □

As for a converse we mention without proof that any Lebesgue integrable function can be arbitrarily well approximated by Riemann integrable functions, and refer the reader, once again, to specialized literature.

Theorem 2.3. *If f is Lebesgue integrable, then, for any $\varepsilon > 0$, there exists,*
(a) *a simple function g, such that $\int_{-\infty}^{\infty} |f(x) - g(x)| \, \mathrm{d}x < \varepsilon$.*
(b) *a continuous, integrable function h, such that $\int_{-\infty}^{\infty} |f(x) - h(x)| \, \mathrm{d}x < \varepsilon$.*

2.3 Decomposition of Distributions

In this subsection we show that that every distribution function can be decomposed into a convex combination of three "pure" kinds.

Definition 2.6. *A distribution function F is*

- discrete *iff for some countable set of numbers $\{x_j\}$ and point masses $\{p_j\}$,*

$$F(x) = \sum_{x_j \le x} p_j, \qquad \textit{for all} \quad x \in \mathbb{R}.$$

  The function p is called probability function.
- continuous *iff it is continuous for all x.*
- absolutely continuous *iff there exists a non-negative, Lebesgue integrable function f, such that*

$$F(b) - F(a) = \int_a^b f(x) \, \mathrm{d}x \qquad \textit{for all} \quad a < b.$$

  The function f is called the density *of F.*
- singular *iff $F \not\equiv 0$, F' exists and equals 0 a.e.* □

The ultimate goal of this subsection is to prove the following decomposition theorem.

Theorem 2.4. *Every distribution function can be decomposed into a convex combination of three pure types, a* discrete *one, an* absolutely continuous *one, and a* continuous singular *one. Thus, if F is a distribution function, then*

$$F = \alpha F_{ac} + \beta F_d + \gamma F_{cs},$$

where $\alpha, \beta, \gamma \geq 0$ and $\alpha + \beta + \gamma = 1$. This means that

- *$F_{ac}(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y$, where $f(x) = F'_{ac}(x)$ a.e.;*
- *F_d is a pure jump function with at most a countable number of jumps;*
- *F_{cs} is continuous and $F'_{cs}(x) = 0$ a.e.*

For the proof we have to accept the following (rather natural) facts. For the proof we refer the reader to his or her favourite book on measure theory or function theory.

Lemma 2.2. *Let F be a distribution function. Then*

(a) *$F'(x)$ exists a.e., and is non-negative and finite.*
(b) *$\int_a^b F'(x)\,\mathrm{d}x \leq F(b) - F(a)$ for all $a, b \in \mathbb{R}$.*
(c) *Set $F_{ac}(x) = \int_{-\infty}^{x} F'(y)\,\mathrm{d}y$, and $F_s(x) = F(x) - F_{ac}(x)$ for all $x \in \mathbb{R}$. Then $F'_{ac}(x) = F'(x)$ a.e. and $F'_s = 0$ a.e. In particular, $F_s \equiv 0$ or F_s is singular.*

Remark 2.3. The components $F_{ac}(x)$ and $F_s(x)$ in Lemma 2.2 are, in contrast to those of Theorem 2.4, sub-distribution functions in that the total mass is only *at most* equal to 1. □

Discrete distributions are obviously singular. But, as we shall see, there also exist *continuous* singular distributions. We shall exhibit one, the Cantor distribution, in Subsection 2.2.6 below, and later, in more detail, in Section 2.11.

The first step is the *Lebesgue decomposition theorem*, in which the distribution function is split into an absolutely continuous component and a singular one.

Theorem 2.5. *Every distribution function can be decomposed into a convex combination of an absolutely continuous distribution function and a singular one. Thus, if F is a distribution function, then*

$$F = \alpha F_{ac} + (1 - \alpha) F_s,$$

where $0 \leq \alpha \leq 1$.

Proof. Let $f(x) = F'(x)$, which, according to Lemma 2.2 exists a.e., and, moreover, equals $F^*_{ac}(x)$ a.e., where

$$F_{ac}^*(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y.$$

In order to see that F_{ac}^* is a distribution function, except, possibly, for the fact that $F_{ac}^*(+\infty) \le 1$, we observe that F_{ac}^* is non-decreasing since $f \ge 0$ a.e., and that $F_{ac}^*(-\infty) = 0$ since $F_{ac}^*(x) \le F(x)$. Continuity is obvious, since the integral is continuous.

Next, set $F_s^*(x) = F(x) - F_{ac}^*$. Then, F_s^* is non-decreasing by Lemma 2.2, and $F_s^*(-\infty) = 0$, since $F_s^*(x) \le F(x)$, which shows that F_s^* is also a distribution function, except, possibly, for having total mass less than one.

If $\alpha = 0$ or 1 we are done. Otherwise, set

$$F_{ac}(x) = \frac{F_{ac}^*(x)}{F_{ac}^*(+\infty)} \text{ and } F_s(x) = \frac{F_s^*(x)}{F_s^*(+\infty)}. \qquad \square$$

The following theorem provides a decomposition of a distribution function into a discrete component and a continuous component.

Theorem 2.6. *Every distribution function F can be written as a convex combination of a discrete distribution function and a continuous one:*

$$F = \beta F_d + (1-\beta)F_c,$$

where $0 \le \beta \le 1$.

Proof. By Proposition 2.1 we know that F may have at most a countable number of jumps. Let $\{x_j\}$ be those jumps (if they exist), let $p(j) = F(x_j+) - F(x_j-) = F(x_j) - F(x_j-)$ for all j (recall that F is right-continuous), and define

$$F_d^*(x) = \sum_{x_j \le x} p_j, \quad x \in \mathbb{R}.$$

By construction, F_d^*, being equal to the sum of all jumps to the left of x, is discrete, and has all properties of a distribution function, except that we only know that $\lim_{x\to\infty} F_d^*(x) \le 1$.

Next, let $F_c^*(x) = F(x) - F_d^*(x)$ for all x. Since $F_d^*(x)$ increases at x_j by p_j and stays constant between jumps, and since F is non-decreasing, it follows that $F_c^*(x)$ must be non-negative and non-decreasing. Moreover,

$$\lim_{x\to-\infty} F_c^*(x) = \lim_{x\to-\infty} (F(x) - F_d^*(x)) = 0 - 0 = 0,$$

and

$$0 \le \lim_{x\to\infty} F_c^*(x) = \lim_{x\to\infty} (F(x) - F_d^*(x)) = 1 - \lim_{x\to\infty} F_d^*(x) \le 1,$$

so that $F_c^*(x)$ also has all properties of a distribution function, except that the total mass may be less than 1. In particular, $F_c^* \in D^+$.

The next thing to prove is that $F_c^*(x)$ is continuous, which seems rather obvious, since we have reduced F by its jumps. Nevertheless,

$$\begin{aligned} F_c^*(x) - F_c^*(x-) &= F(x) - F_d^*(x) - (F(x-) - F_d^*(x-)) \\ &= F(x) - F(x-) - (F_d^*(x) - F_d^*(x-)) \\ &= \begin{cases} p_j - p_j = 0, & \text{when} \quad x = x_j \text{ for some } j, \\ 0, & \text{otherwise,} \end{cases} \end{aligned}$$

which shows that F_c^* is left-continuous. This tells us that F_c^* is continuous, since, as we have already seen, $F_c^* \in D^+$. A final rescaling, if necessary, finishes the proof. □

Proof of Theorem 2.4. Let F be a distribution function. Then, by the Lebesgue decomposition theorem, we know that

$$F = \alpha F_{ac} + (1-\alpha)F_s,$$

and by Theorem 2.6 applied to the singular part we know that

$$F_s = \beta F_d + (1-\beta)F_{cs}.$$

□

Theorem 2.7. *The decompositions are unique.*

Exercise 2.1. Prove the uniqueness (by contradiction). □

2.4 Some Standard Discrete Distributions

Following is a list of some of the most common *discrete* distributions. The domains of the parameters below are $a \in \mathbb{R}$, $0 \le p = 1 - q \le 1$, $n \in \mathbb{N}$, and $m > 0$.

Distribution	Notation	Probability function	Domain
One point	$\delta(a)$	$p(a) = 1$	
Symmetric Bernoulli		$p(-1) = p(1) = \frac{1}{2}$	
Bernoulli	$\mathrm{Be}(p)$	$p(0) = q$, $p(1) = p$	
Binomial	$\mathrm{Bin}(n,p)$	$p(k) = \binom{n}{k} p^k q^{n-k}$	$k = 0, 1, \ldots, n$
Geometric	$\mathrm{Ge}(p)$	$p(k) = pq^k$	$k \in \mathbb{N} \cup 0$
First success	$\mathrm{Fs}(p)$	$p(k) = pq^{k-1}$	$k \in \mathbb{N}$
Poisson	$\mathrm{Po}(m)$	$p(k) = \mathrm{e}^{-m} \frac{m^k}{k!}$	$k \in \mathbb{N} \cup 0$

Table 2.1. Some discrete distributions

The $\mathrm{Be}(p)$-distribution describes the outcome of one "coin-tossing" experiment, and the $\mathrm{Bin}(n,p)$-distribution the number of successes in n trials. The $\mathrm{Ge}(p)$-distribution describes the number of failures prior to the first success, and the $\mathrm{Fs}(p)$-distribution the number of trails required to succeed once. Finally, the typical experiment for a Poisson distribution is a coin-tossing experiment where the probability of success is "small". Vaguely speaking, $\mathrm{Bin}(n,p) \approx \mathrm{Po}(np)$ if p is small (typically "small" means < 0.1). This can, of course, be rigorously demonstrated.

2.5 Some Standard Absolutely Continuous Distributions

In this subsection we list some of the most common *absolutely continuous* distributions. The parameters $p, \theta, \sigma, r, s, \alpha, \beta$ below are all non-negative, and $a, b, \mu \in \mathbb{R}$.

Distribution	Notation	Density function	Domain
Uniform	$U(a,b)$	$f(x) = \frac{1}{b-a}$	$a < x < b$
	$U(0,1)$	$f(x) = 1$	$0 < x < 1$
	$U(-1,1)$	$f(x) = \frac{1}{2}$	$\lvert x\rvert < 1$
Triangular	$\mathrm{Tri}(-1,1)$	$f(x) = 1 - \lvert x\rvert$	$\lvert x\rvert < 1$
Exponential	$\mathrm{Exp}(\theta)$	$f(x) = \frac{1}{\theta}\mathrm{e}^{-x/\theta}$	$x > 0$
Gamma	$\Gamma(p,\theta)$	$f(x) = \frac{1}{\Gamma(p)}\, x^{p-1}\, \frac{1}{\theta^p}\, \mathrm{e}^{-x/\theta}$	$x > 0$
Beta	$\beta(r,s)$	$f(x) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\, x^{r-1}(1-x)^{s-1}$	$0 < x < 1$
Normal	$N(\mu,\sigma^2)$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(x-\mu)^2/\sigma^2}$	$x \in \mathbb{R}$
	$N(0,1)$	$f(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}$	$x \in \mathbb{R}$
Log-normal	$LN(\mu,\sigma^2)$	$f(x) = \frac{1}{\sigma x\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\log x-\mu)^2/\sigma^2}$	$x > 0$
Cauchy	$C(0,1)$	$f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$	$x \in \mathbb{R}$
Pareto	$\mathrm{Pa}(\beta,\alpha)$	$f(x) = \frac{\alpha k^\alpha}{x^{\alpha+1}}$	$x > \beta$

Table 2.2. Some absolutely continuous distributions

We have listed two special uniform distributions and the standard normal distribution because of their frequent occurrences, and confined ourselves to the special triangular distribution which has support on $[-1,1]$ and the standard Cauchy distribution for convenience.

Uniform distributions typically describe phenomena such as picking a point "at random" in the sense that the probability that the resulting point belongs to an interval only depends on the length of the interval and not on its position.

Exponential and gamma distributed random variables typically are used to model waiting times, life lengths, and so on, in particular in connection with the so-called Poisson process.

The normal distribution, also called the Gaussian distribution, models cumulative or average results of "many" repetitions of an experiment; the formal result is the central limit theorem, which we shall meet in Chapter 7. The *multivariate* normal distribution, that we shall encounter in Subsection 4.5.1, plays, i.a., an important role in many statistical applications.

2.6 The Cantor Distribution

The third kind of distributions, the continuous singular ones, are the most special or delicate ones. In this subsection we shall define the Cantor distribution and prove that it belongs to that class.

The standard Cantor *set* is constructed on the interval $[0,1]$ as follows. One successively removes the open middle third of each subinterval of the previous set. The Cantor set itself is the infinite intersection of all remaining sets. More precisely, let $C_0 = [0,1]$, and, successively,

$$C_1 = \Big[0,\frac{1}{3}\Big] \cup \Big[\frac{2}{3},1\Big], \quad C_2 = \Big[0,\frac{1}{9}\Big] \cup \Big[\frac{2}{9},\frac{1}{3}\Big] \cup \Big[\frac{2}{3},\frac{7}{9}\Big] \cup \Big[\frac{8}{9},1\Big],$$

and so on.

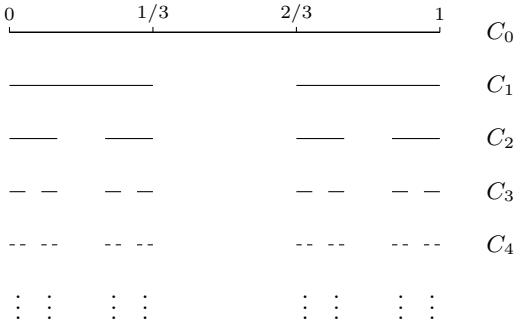

Figure 2.1. The Cantor set on $[0,1]$

The Cantor set is

$$C = \bigcap_{n=0}^{\infty} C_n,$$

and the Cantor distribution is the distribution that is uniform on the Cantor set.

Having thus defined the distribution we now show that it is continuous singular.
(i): The Lebesgue measure of the Cantor set equals 0, since $C \subset C_n$ for all n, so that

$$\lambda(C) \le \lambda(C_n) = \Big(\frac{2}{3}\Big)^n \quad \text{for every } n \quad \Longrightarrow \quad \lambda(C) = 0.$$

Alternatively, in each step we remove the middle thirds. The Lebesgue measure of the pieces we *remove* thus equals

$$\frac{1}{3} + 2\Big(\frac{1}{3}\Big)^2 + 4\Big(\frac{1}{3}\Big)^3 + \cdots = \sum_{n=1}^{\infty} 2^{n-1}\Big(\frac{1}{3}\Big)^n = 1.$$

The Cantor set is the complement, hence $\lambda(C) = 0$.
(ii): The Cantor distribution is singular, since its support is a Lebesgue null set.

(iii): The distribution function is continuous. Namely, let F_n be the distribution function corresponding to the distribution that is uniform on C_n. This means that $F_n(0) = 0$, F_n is piecewise constant with 2^n jumps of size 2^{-n} and $F_n(1) = 1$. Moreover, $F_n'(x) = 0$ for all x except for the end-points of the 2^n intervals.

The distribution function of the Cantor distribution is

$$F(x) = \lim_{n\to\infty} F_n(x).$$

Now, let $x, y \in C_n$. Every subinterval of C_n has length 3^{-n}. Therefore,

$$0 < x - y < \frac{1}{3^n} \Longrightarrow$$
$$F(y) - F(x) \begin{cases} = 0, & \text{when } x, y \text{ are in the same subinterval,} \\ \leq \frac{1}{2^n} & \text{when } x, y \text{ are in adjacent subintervals.} \end{cases}$$

This proves that F, in fact, is uniformly continuous on C.

(iv) $F'(x) = 0$ for almost all x, because $F'(x) = 0$ for all $x \in C_n$ for all n.

This finishes the proof of the fact that the Cantor distribution is continuous singular. We shall return to this distribution in Section 2.11, where an elegant representation in terms of an infinite sum will be given.

2.7 Two Perverse Examples

In Example 2.1 we met the function on the unit interval, which was equal to 1 on the irrationals and 0 on the rationals:

$$f(x) = \begin{cases} 1, & \text{for} \quad x \in [0,1] \setminus \mathbb{Q}, \\ 0, & \text{for} \quad x \in [0,1] \cap \mathbb{Q}. \end{cases}$$

Probabilistically this function can be interpreted as the density of a random variable, X, which is uniformly distributed on the irrationals in $[0, 1]$.

Note that, if $U \in U(0,1)$, then the probability that the two random variables differ equals

$$P(X \neq U) = P(X \in \mathbb{Q}) = 0,$$

so that $X \sim U$.

An extreme variation of this example is the following:

Example 2.2. Let $\{r_k,\, k \geq 1\}$ be an enumeration of the rationals in the unit interval, and define

$$p(r_k) = \begin{cases} \frac{6}{\pi^2 k^2}, & \text{for} \quad r_k \in (0,1) \cap \mathbb{Q}, \\ 0, & \text{otherwise.} \end{cases}$$

Since $\sum_{k=1}^{\infty} 1/k^2 = \pi^2/6$, this is a bona fide discrete distribution.

This may seem as a somewhat pathological distribution, since it is defined along an enumeration of the rationals, which by no means is neither unique nor "chronological". □

3 Random Vectors; Random Elements

Random vectors are the same as multivariate random variables. Random elements are "random variables" in (more) abstract spaces.

3.1 Random Vectors

Random vectors are elements in the Euclidean spaces $\mathbb{R}^n$ for some $n \in \mathbb{N}$.

Definition 3.1. *An n-dimensional* random vector $\mathbf{X}$ *is a measurable function from the sample space Ω to $\mathbb{R}^n$;*

$$\mathbf{X} : \Omega \to \mathbb{R}^n,$$

that is, the inverse image of any Borel set is $\mathcal{F}$-measurable:

$$X^{-1}(A) = \{\omega : X(\omega) \in A\} \in \mathcal{F} \quad \textit{for all} \quad A \in \mathcal{R}^n.$$

Random vectors are considered column vectors;

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)',$$

where $'$ denotes transpose (i.e., $\mathbf{X}'$ is a row vector).

The joint distribution function *of* $\mathbf{X}$ *is*

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n),$$

for $x_k \in \mathbb{R}$, $k = 1, 2, \ldots, n$. □

Remark 3.1. A more compact way to express the distribution function is

$$F_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} \le \mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n,$$

where the event $\{\mathbf{X} \le \mathbf{x}\}$ is to be interpreted component-wise, that is,

$$\{\mathbf{X} \le \mathbf{x}\} = \{X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n\} = \bigcap_{k=1}^{n} \{X_k \le x_k\}. \qquad \square$$

For discrete distributions the *joint probability function* is defined by

$$p_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n.$$

In the absolutely continuous case we have a *joint density*;

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^n F_{\mathbf{X}}(\mathbf{x})}{\partial x_1 \partial x_2 \cdots \partial x_n} \quad \mathbf{x} \in \mathbb{R}^n\,.$$

The following example illuminates a situation where a problem intrinsically is defined in a "high" dimension, but the object of interest is "low-dimensional" (in the example, high = 2 and low = 1).

Example 3.1. Let (X, Y) be a point that is uniformly distributed on the unit disc, that is,

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi}, & \text{for} \quad x^2 + y^2 \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Determine the distribution of the x-coordinate. □

In order to solve this problem we consider, as a preparation, the discrete analog, which is easier to handle. Let (X, Y) be a given two-dimensional random variable whose joint probability function is $p_{X,Y}(x, y)$ and that we are interested in finding $p_X(x)$. By the law of total probability, Proposition 1.4.1,

$$\begin{aligned} p_X(x) = P(X = x) &= P\Big(\{X = x\} \cap \Big\{\bigcup_y \{Y = y\}\Big\}\Big) \\ &= P\Big(\bigcup_y \Big\{\{X = x\} \cap \{Y = y\}\Big\}\Big) \\ &= \sum_y P(X = x, Y = y) = \sum_y p_{X,Y}(x, y). \end{aligned}$$

The distribution of one of the random variables thus is obtained by adding the joint probabilities along "the other" variable.

Distributions thus obtained are called *marginal distributions*, and the corresponding probability functions are called *marginal probability functions*.

The *marginal distribution function* of X at the point x is obtained by adding the values of the marginal probabilities to the left of x:

$$F_X(x) = \sum_{u \leq x} p_X(u) = \sum_{u \leq x} \sum_v p_{X,Y}(u, v).$$

Alternatively,

$$F_X(x) = P(X \leq x, Y < \infty) = \sum_{u \leq x} \sum_v p_{X,Y}(u, v).$$

In the absolutely continuous case we depart from the distribution function

$$F_X(x) = P(X \leq x, Y < \infty) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(u, v)\, \mathrm{d}u\mathrm{d}v,$$

and differentiate to obtain the *marginal density function*,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, \mathrm{d}y,$$

that is, this time we *integrate* along "the other" variable.

Marginal distribution functions are integrals of the marginal densities.

Analogous formulas hold in higher dimensions, and for more general distributions. Generally speaking, marginal distributions are obtained by integrating in the generalized sense ("getting rid of") those components that are not relevant for the problem at hand.

Let us now solve the problem posed in Example 3.1. The joint density was given by

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi}, & \text{for} \quad x^2+y^2 \le 1, \\ 0, & \text{otherwise}, \end{cases}$$

from which we obtain

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi}\,\mathrm{d}y = \frac{2}{\pi}\sqrt{1-x^2}, \tag{3.1}$$

for $-1 < x < 1$ (and $f_X(x) = 0$ otherwise).

Exercise 3.1. Let (X, Y, Z) be a point chosen uniformly within the three-dimensional unit sphere. Find the marginal distributions of (X, Y) and X. □

We have seen how a problem might naturally be formulated in a higher dimension than that of interest. The converse concerns to what extent the marginal distributions determine the joint distribution. Interesting applications are computer tomography and satellite pictures; in both cases one departs from two-dimensional pictures from which one wishes to make conclusions about three-dimensional objects (the brain and the Earth).

A multivariate distribution of special importance is the normal one, for which some facts and results will be presented in Chapter 4.

3.2 Random Elements

Random elements are random variables on abstract spaces.

Definition 3.2. *A* random element *is a measurable mapping from a measurable space* $(\Omega, \mathcal{F})$ *to a measurable, metric space* $(S, \mathcal{S})$*:*

$$X : (\Omega, \mathcal{F}) \to (S, \mathcal{S}). \qquad \square$$

In this setting measurability thus means that

$$X^{-1}(A) = \{\omega : X(\omega) \in A\} \in \mathcal{F} \quad \text{for all} \quad A \in \mathcal{S}.$$

The meaning thus is the same as for ordinary random variables. With a slight exaggeration one may say that the difference is "notational".

The distribution of a random element is "the usual one", namely the induced one, $\mathbb{P} = P \circ X^{-1}$;

$$\mathbb{P}(A) = P(\{\omega : X(\omega) \in A\}) \quad \text{for} \quad A \in \mathcal{S}.$$

A typical example is the space $C[0,1]$ of continuous functions on the unit interval, endowed with the uniform topology or metric

$$d(x,y) = \sup_{0\le t\le 1} |x(t) - y(t)| \quad \text{for} \quad x, y \in C[0,1].$$

For more on this and on the analog for the space $D[0,1]$ of right-continuous functions with left-hand limits on the unit interval, endowed with the Skorohod J_1- or M_1-topologies [224], see also [20, 188].

4 Expectation; Definitions and Basics

Just as random variables are "compressed versions" of events from a probability space, one might be interested in compressed versions of random variables. The typical one is the *expected value*, which is the probabilistic version of the center of gravity of a physical body. Another name for expectation is *mean*.

Mathematically, expectations are integrals with respect to distribution functions or probability measures. We must therefore develop the theory of integration, more precisely, the theory of *Lebesgue integration*. However, since this is a book on probability theory we prefer to develop the theory of Lebesgue integration in terms of expectations. We also alert the reader to the small integration preview in Subsection 2.2.2, and recommend a translation of what is to come into the traditional mathematics language – remember that rewriting is much more profitable than rereading.

Much of what follows next may seem like we are proving facts that are "completely obvious" or well known (or both). For example, the fact that the tails of convergent integrals tend to 0 just as the terms in a convergent series do. We must, however, remember that we are introducing a new integral concept, namely the Lebesgue integral, and for that concept we "do not yet know" that the results are "trivial". So, proofs and some care are required. Along the way we also obtain the promised justifications of facts from Subsection 2.2.2.

4.1 Definitions

We begin with the simple case.

Simple Random Variables

We remember from Definition 1.1 that a random variable X is *simple* if, for some n,

$$X = \sum_{k=1}^{n} x_k I\{A_k\},$$

where $\{x_k,\, 1 \le k \le n\}$ are real numbers, and $\{A_k,\, 1 \le k \le n\}$ is a *finite* partition of Ω.

Definition 4.1. *For the simple random variable* $X = \sum_{k=1}^{n} x_k I\{A_k\}$*, we define the* expected value *as*

$$E\,X = \sum_{k=1}^{n} x_k P(A_k).$$

□

Non-negative Random Variables

In the first section of this chapter we found that if X is a non-negative random variable, then the sequence of *simple* non-negative random variables X_n, $n \geq 1$, defined by

$$X_n(\omega) = \begin{cases} \frac{k-1}{2^n}, & \text{for} \quad \frac{k-1}{2^n} \leq X(\omega) < \frac{k}{2^n}, \quad k = 1, 2, \ldots, n2^n, \\ n, & \text{for} \quad X(\omega) \geq n, \end{cases}$$

converges monotonically from below to X as $n \to \infty$. With this in mind we make the following definition of the expected value of arbitrary, non-negative random variables.

Definition 4.2. *Suppose that* X *is a* non-negative *random variable. The* expected value *of* X *is defined as*

$$E\,X = \lim_{n\to\infty} \sum_{k=1}^{n2^n} \frac{k-1}{2^n} P\Big(\frac{k-1}{2^n} \leq X < \frac{k}{2^n}\Big).$$

Note that the limit may be infinite. □

The definition is particularly appealing for bounded random variables. Namely, suppose that X is a non-negative random variable, such that

$$X \leq M < \infty, \quad \text{for some} \quad M > 0,$$

and set,

$$Y_n(\omega) = \begin{cases} \frac{k}{2^n}, & \text{for} \quad \frac{k-1}{2^n} \leq X(\omega) < \frac{k}{2^n}, \quad k = 1, 2, \ldots, n2^n, \\ n, & \text{for} \quad X(\omega) \geq n, \end{cases}$$

for $n \geq M$, where we pretend, for simplicity only, that M is an integer. Then $Y_n \searrow X$ as $n \to \infty$, and moreover,

$$X_n \leq X \leq Y_n, \quad \text{and} \quad Y_n - X_n = \frac{1}{2^n}.$$

Thus, by the consistency property that we shall prove in Theorem 4.2 below,

$$E\,X_n \leq E\,X \leq E\,Y_n, \quad \text{and} \quad E(Y_n - X_n) = \frac{1}{2^n} \to 0 \quad \text{as} \quad n \to \infty.$$

The General Case

Definition 4.3. *For an arbitrary random variable X we define*

$$E\,X = E\,X^+ - E\,X^-,$$

provided at least one of $E\,X^+$ and $E\,X^-$ is finite (thus prohibiting $\infty - \infty$). We write

$$E\,X = \int_\Omega X(\omega)\,\mathrm{d}P(\omega) \quad \textit{or, simply,} \quad \int X\,\mathrm{d}P.$$

If both values are finite, that is, if $E|X| < \infty$, we say that X is integrable.□

Throughout our treatment, P is a probability measure, and assumptions about integrability are with respect to P. Recall that a.s. means almost surely, that is, if a property holds a.s. then the set where it does *not* hold is a null set. If X and Y are random variables, such that $X = Y$ a.s., this means that $P(X = Y) = 1$, or, equivalently, that $P(X \neq Y) = 0$.

During the process of constructing the concept of expectated values, we shall need the concept of almost sure convergence, which means that we shall meet situations where we consider sequences $X_1, X_2, \ldots$ of random variables such that $X_n(\omega) \to X(\omega)$ as $n \to \infty$, not for *every* ω, but for *almost all* ω. This, as it turns out, is sufficient, (due to equivalence; Definition 1.2), since integrals over sets of measure 0 are equal to 0.

As a, somewhat unfortunate, consequence, the introduction of the concept of almost sure convergence cannot wait until Chapter 5.

Definition 4.4. *Let $X, X_1, X_2, \ldots$ be random variables. We say that X_n converges* almost surely (a.s.) *to the random variable X as $n \to \infty$, $X_n \overset{a.s.}{\to} X$ as $n \to \infty$, iff*

$$P\big(\{\omega : X_n(\omega) \to X(\omega) \ \textit{as}\ n \to \infty\}\big) = 1,$$

or, equivalently, iff

$$P\big(\{\omega : X_n(\omega) \not\to X(\omega) \ \textit{as}\ n \to \infty\}\big) = 0. \qquad \square$$

4.2 Basic Properties

The first thing to prove is that the definition of expectation is consistent, after which we turn our attention to a number of properties, such as additivity, linearity, domination, and so on.

Simple Random Variables

We thus begin with a lemma proving that the expected value of a random variable is independent of the partition.

Lemma 4.1. *If* $\{A_k,\, 1 \le k \le n\}$ *and* $\{B_j,\, 1 \le j \le m\}$ *are partitions of* Ω*, such that*

$$X = \sum_{k=1}^{n} x_k I\{A_k\} \quad \textit{and} \quad X = \sum_{j=1}^{m} y_j I\{B_j\},$$

Then

$$\sum_{k=1}^{n} x_k P(A_k) = \sum_{j=1}^{m} y_j P(B_j).$$

Proof. The fact that $\{A_k,\, 1 \le k \le n\}$ and $\{B_k,\, 1 \le k \le m\}$ are partitions implies that

$$P(A_k) = \sum_{j=1}^{m} P(A_k \cap B_j) \quad \text{and} \quad P(B_j) = \sum_{k=1}^{n} P(A_k \cap B_j),$$

and, hence, that

$$\sum_{k=1}^{n} x_k P(A_k) = \sum_{k=1}^{n} \sum_{j=1}^{m} x_k P(A_k \cap B_j),$$

and

$$\sum_{j=1}^{m} y_j P(B_j) = \sum_{j=1}^{m} \sum_{k=1}^{n} y_j P(A_k \cap B_j).$$

Since the sets $\{A_k \cap B_j,\, 1 \le k \le n,\, 1 \le j \le m\}$ also form a partition of Ω it follows that $x_k = y_j$ whenever $A_k \cap B_j \neq \emptyset$, which proves the conclusion. □

Next we show that intuitively obvious operations are permitted.

Theorem 4.1. *Let* X, Y *be non-negative simple random variables. Then:*

(a) *If* $X = 0$ *a.s., then* $E\,X = 0$*;*
(b) $E\,X \ge 0$*;*
(c) *If* $E\,X = 0$*, then* $X = 0$ *a.s.;*
(d) *If* $E\,X > 0$*, then* $P(X > 0) > 0$*;*
(e) Linearity: $E(aX + bY) = aE\,X + bE\,Y$ *for any* $a, b \in \mathbb{R}^+$*;*
(f) $E\,X I\{X > 0\} = E\,X$*;*
(g) Equivalence: *If* $X = Y$ *a.s., then* $E\,Y = E\,X$*;*
(h) Domination: *If* $Y \le X$ *a.s., then* $E\,Y \le E\,X$*.*

Proof. (a): If $X(\omega) = 0$ for all $\omega \in \Omega$, then, with $A_1 = \{X = 0\}$ $(= \Omega)$, we have $X = 0 \cdot I\{A_1\}$, so that $E\,X = 0 \cdot P(A_1) = 0 \cdot 1 = 0$.

If $X = 0$ a.s., then $X = \sum_{k=1}^{n} x_k I\{A_k\}$, where $x_1 = 0$, and $x_2, x_3, \ldots, x_n$ are finite numbers, $A_1 = \{X = 0\}$, and $A_2, A_3, \ldots, A_n$ are null sets. It follows that

$$E\,X = 0 \cdot P(A_1) + \sum_{k=2}^{n} x_k \cdot 0 = 0.$$

(b): Immediate, since the sum of non-negative terms is non-negative.
(c): By assumption,

$$\sum_{k=1}^{n} x_k P(A_k) = 0.$$

The fact that the sum of non-negative terms can be equal to 0 if and only if all terms are equal to 0, forces one of x_k and $P(A_k)$ to be equal to 0 for every $k \geq 2$ ($A_1 = \{X = 0\}$ again). In particular, we must have $P(A_k) = 0$ for any nonzero x_k, which shows that $P(X = 0) = 1$.
(d): The assumption implies that at least one of the terms $x_k P(A_k)$, and therefore both factors of this term must be positive.
(e): With $X = \sum_{k=1}^{n} x_k I\{A_k\}$ and $Y = \sum_{j=1}^{m} y_j I\{B_j\}$, we have

$$X + Y = \sum_{k=1}^{n} \sum_{j=1}^{m} (x_k + y_j) I\{A_k \cap B_j\},$$

so that

$$\begin{aligned} E(aX + bY) &= \sum_{k=1}^{n} \sum_{j=1}^{m} (ax_k + by_j) P(A_k \cap B_j) \\ &= a \sum_{k=1}^{n} \sum_{j=1}^{m} x_k P(A_k \cap B_j) + b \sum_{k=1}^{n} \sum_{j=1}^{m} y_j P(A_k \cap B_j) \\ &= a \sum_{k=1}^{n} x_k P\Big(A_k \cap \Big(\bigcup_{j=1}^{m} B_j\Big)\Big) + b \sum_{j=1}^{m} y_j P\Big(\Big(\bigcup_{k=1}^{n} A_k\Big) \cap B_j\Big) \\ &= a \sum_{k=1}^{n} x_k P(A_k) + b \sum_{j=1}^{m} y_j P(B_j) = aE\,X + bE\,Y. \end{aligned}$$

(f): Joining (a) and (e) yields

$$E\,X = E\,XI\{X > 0\} + E\,XI\{X = 0\} = E\,XI\{X > 0\} + 0 = E\,XI\{X > 0\}.$$

(g): If $X = Y$ a.s., then $X - Y = 0$ a.s., so that, by (a), $E(X - Y) = 0$, and by (e),

$$E\,X = E\big((X - Y) + Y\big) = E(X - Y) + E\,Y = 0 + E\,Y.$$

(h): The proof is similar to that of (g). By assumption, $X - Y \geq 0$ a.s., so that, by (b), $E(X - Y) \geq 0$, and by linearity,

$$E\,Y = E\,X - E(X - Y) \leq E\,X. \qquad \square$$

Non-negative Random Variables

Once again, the first thing to prove is consistency.

Theorem 4.2. (Consistency)
Let X be a non-negative random variable, and suppose that $\{Y_n,\, n \geq 1\}$ and $\{Z_n,\, n \geq 1\}$ are sequences of simple *random variables, such that*

$$Y_n \nearrow X \quad \textit{and} \quad Z_n \nearrow X \quad \textit{as} \quad n \to \infty.$$

Then

$$\lim_{n\to\infty} E\, Y_n = \lim_{n\to\infty} E\, Z_n \quad (= E\, X).$$

Proof. The first remark is that *if* the limits are equal, then they must be equal to $E\, X$ because of the definition of the expected value for non-negative random variables (Definition 4.2).

To prove equality between the limits it suffices to show that if $0 \leq Y_n \nearrow X$ as $n \to \infty$, and $X \geq Z_m$, then

$$\lim_{n\to\infty} E\, Y_n \geq E\, Z_m, \tag{4.1}$$

because by switching roles between the two sequences, we similarly obtain

$$\lim_{m\to\infty} E\, Z_m \geq E\, Y_n,$$

and the desired equality follows.

To prove (4.1) we first suppose that

$$Z_m > c > 0.$$

Next we note that there exists $M < \infty$, such that $Z_m \leq M$ (because Z_m is simple, and therefore has only a finite number of supporting points).

Let $\varepsilon < M$, set $A_n = \{Y_n \geq Z_m - \varepsilon\}$, and observe that, by assumption, $A_n \nearrow \Omega$ a.s. as $n \to \infty$. Moreover,

$$Y_n \geq Y_n I\{A_n\} \geq (Z_m - \varepsilon) I\{A_n\}.$$

By domination we therefore obtain (all random variables are simple)

$$\begin{aligned} E\, Y_n &\geq E\, Y_n I\{A_n\} \geq E(Z_m - \varepsilon) I\{A_n\} = E\, Z_m I\{A_n\} - \varepsilon P(A_n) \\ &= E\, Z_m - E\, Z_m I\{A_n^c\} - \varepsilon \geq E\, Z_m - M P(A_n^c) - \varepsilon, \end{aligned}$$

so that,

$$\liminf_{n\to\infty} E\, Y_n \geq E\, Z_m - \varepsilon,$$

since $P(A_n^c) \to 0$ as $n \to \infty$. The arbitrariness of ε concludes the proof for that case. Since c was arbitrary, (4.1) has been verified for Z_m strictly positive.

If $c = 0$, then, by domination, and what has already been shown,

$$\liminf_{n\to\infty} E\, Y_n \geq \liminf_{n\to\infty} E\, Y_n I\{Z_m > 0\} \geq E\, Z_m I\{Z_m > 0\} = E\, Z_m,$$

where, to be precise, we used Theorem 4.1(f) in the last step. □

We have thus shown consistency and thereby that the definition of the expected value is in order.

A slight variation to prove consistency runs as follows.

Theorem 4.3. *Suppose that X is a non-negative random variable, and that $\{Y_n,\, n \geq 1\}$ are non-negative simple random variables, such that $0 \leq Y_n \nearrow X$ as $n \to \infty$. Suppose further, that Y is a simple random variable, such that $0 \leq Y \leq X$. Then*

$$\lim_{n\to\infty} EY_n \geq E\,Y.$$

Exercise 4.1. Prove the theorem by showing that it suffices to consider indicator functions, $Y = I\{A\}$ for $A \in \mathcal{F}$.
Hint: Think metatheorem. □

The next point in the program is to show that the basic properties we have provided for simple random variables carry over to general non-negative random variables.

Theorem 4.4. *Let X, Y be non-negative random variables. Then*

(a) *If $X = 0$ a.s., then $E\,X = 0$;*
(b) *$E\,X \geq 0$;*
(c) *If $E\,X = 0$, then $X = 0$ a.s.;*
(d) *If $E\,X > 0$, then $P(X > 0) > 0$;*
(e) Linearity: *$E(aX + bY) = aE\,X + bE\,Y$ for any $a, b \in \mathbb{R}^+$;*
(f) *$E\,X I\{X > 0\} = E\,X$;*
(g) Equivalence: *If $X = Y$ a.s., then $E\,Y = E\,X$;*
(h) Domination: *If $Y \leq X$ a.s., then $E\,Y \leq E\,X$;*
(j) *If $E\,X < \infty$, then $X < \infty$ a.s., that is, $P(X < \infty) = 1$.*

Remark 4.1. Note that infinite expected values are allowed. □

Proof. The properties are listed in the same order as for simple random variables, but verified in a different order (property (j) is new).

The basic idea is that there exist sequences $\{X_n,\, n \geq 1\}$ and $\{Y_n,\, n \geq 1\}$ of non-negative simple random variables converging monotonically to X and Y, respectively, as $n \to \infty$, and which obey the basic rules for each n. The conclusions then follow by letting $n \to \infty$.

For (a) there is nothing new to prove.

To prove linearity, we know from Theorem 4.1(e) that

$$E(aX_n + bY_n) = aE\,X_n + bE\,Y_n \quad \text{for any} \quad a, b \in \mathbb{R}^+,$$

which, by letting $n \to \infty$, shows that

$$E(aX + bY) = aE\,X + bE\,Y \quad \text{for any} \quad a, b \in \mathbb{R}^+.$$

The proof of (h), domination, follows exactly the same pattern. Next, (b) follows from (h) and (a): Since $X \geq 0$, we obtain $E\,X \geq E\,0 = 0$. In order to prove (c), let $A_n = \{\omega : X(\omega) \geq \frac{1}{n}\}$. Then

$$\frac{1}{n} I\{A_n\} \le X_n I\{A_n\} \le X,$$

so that

$$\frac{1}{n} P(A_n) \le E\, X_n I\{A_n\} \le E\, X = 0,$$

which forces $P(A_n) = 0$ for all n, that is $P(X < \frac{1}{n}) = 1$ for all n.

Moreover, (d) follows from (a), and (f) follows from (e) and (a), since

$$E\, X = E\, X I\{X = 0\} + E\, X I\{X > 0\} = 0 + E\, X I\{X > 0\}.$$

Equivalence follows as in Theorem 4.1, and (j), finally, by linearity,

$$\infty > E\, X = E\, X I\{X < \infty\} + E\, X I\{X = \infty\} \ge E\, X I\{X = \infty\},$$

from which there is no escape except $P(X = \infty) = 0$. □

The General Case

Recall that the expected value of a random variable X is defined as the difference between the expected values of the positive and negative parts, $E\, X = E\, X^+ - E\, X^-$, provided at least one of them is finite, and that the expected value is finite if and only if $E|X| < \infty$, in which case we call the random variable integrable.

By reviewing the basic properties we find that (a) remains (nothing is added), that (b) disappears, and that (c) is no longer true, since, e.g., symmetric random variables whose mean exists have mean 0 – one such example is $P(X = 1) = P(X = -1) = 1/2$. The remaining properties remain with minor modifications.

Theorem 4.5. *Let X, Y be integrable random variables. Then*

(a) *If $X = 0$ a.s., then $E\, X = 0$;*
(b) *$|X| < \infty$ a.s., that is, $P(|X| < \infty) = 1$;*
(c) *If $E\, X > 0$, then $P(X > 0) > 0$;*
(d) Linearity: *$E(aX + bY) = aE\, X + bE\, Y$ for any $a, b \in \mathbb{R}$;*
(e) *$E\, X I\{X \neq 0\} = E\, X$;*
(f) Equivalence: *If $X = Y$ a.s., then $E\, Y = E\, X$;*
(g) Domination: *If $Y \le X$ a.s., then $E\, Y \le E\, X$;*
(h) Domination: *If $|Y| \le X$ a.s., then $E|Y| \le E\, X$.*

Proof. For the proofs one considers the two tails separately. Let us illustrate this by proving linearity.

Since, by the triangle inequality, $|aX + bY| \le |a||X| + |b||Y|$ it follows, by domination and linearity for non-negative random variables, Theorem 4.4(h) and (e), that

$$E|aX + bY| \le E|a||X| + E|b||Y| = |a|E|X| + |b|E|Y| < \infty,$$

so that the sum is integrable. Next we split the sum in two different ways:

$$aX + bY = \begin{cases} (aX + bY)^+ - (aX + bY)^-, \\ (aX)^+ - (aX)^- + (bY)^+ - (bY)^-. \end{cases}$$

Because of linearity it suffices to prove additivity.

Since all random variables to the right are non-negative we use linearity to conclude that

$$E(X + Y)^+ + E(X^-) + E(Y^-) = E(X + Y)^- + E(X^+) + E(Y^+),$$

which shows that

$$\begin{aligned} E(X + Y) &= E(X + Y)^+ - E(X + Y)^- \\ &= E(X^+) - E(X^-) + E(Y^+) - E(Y^-) = E\,X + E\,Y. \end{aligned}$$ □

Exercise 4.2. Complete the proof of the theorem. □

5 Expectation; Convergence

In addition to the basic properties one is frequently faced with an infinite sequence of functions and desires information about the limit. A well-known fact is that it is not permitted in general to reverse the order of taking a limit and computing an integral; in technical probabilistic terms the problem amounts to the question

$$\lim_{n\to\infty} E\,X_n \;?\; =?\; E \lim_{n\to\infty} X_n. \tag{5.1}$$

We shall encounter this problem in greater detail in Chapter 5 which is devoted to various convergence modes. We therefore provide just one illustration here, the full impact of which will be clearer later.

Example 5.1. Let $\alpha > 0$, and set

$$P(X_n = 0) = 1 - \frac{1}{n^2} \quad \text{and} \quad P(X_n = n^\alpha) = \frac{1}{n^2}, \quad n \geq 1.$$

Taking only two different values these are certainly simple random variables, but we immediately observe that one of the points slides away toward infinity as n increases.

One can show (this will be done in Chapter 5) that $X_n(\omega) \to 0$ as $n \to \infty$ for almost every ω, which means that $P(\lim_{n\to\infty} X_n = 0) = 1$ – at this point we may at least observe that $P(X_n = 0) \to 1$ as $n \to \infty$.

As for the limit of the expected values,

$$E\,X_n = 0 \cdot \Big(1 - \frac{1}{n^2}\Big) + n^\alpha \cdot \frac{1}{n^2} = n^{\alpha-2} \to \begin{cases} 0, & \text{for} \quad 0 < \alpha < 2, \\ 1, & \text{for} \quad \alpha = 2, \\ \infty, & \text{for} \quad \alpha > 2. \end{cases}$$

The answer to the question addressed in (5.1) thus may vary. □

Typical conditions that yield positive results are uniformity, monotonicity or domination conditions. All of these are tailored in order to prevent masses to escape, "to pop up elsewhere".

A first positive result concerns random variables that converge monotonically.

Theorem 5.1. (Monotone convergence)
Let $\{X_n,\, n \geq 1\}$ *be non-negative random variables. If* $X_n \nearrow X$ *as* $n \to \infty$, *then*

$$E\, X_n \nearrow E\, X \quad \text{as} \quad n \to \infty.$$

Remark 5.1. The limit may be infinite. □

Proof. From the consistency proof we know that the theorem holds if $\{X_n,\, n \geq 1\}$ are non-negative *simple* random variables. For the general case we therefore introduce non-negative, simple random variables $\{Y_{k,n},\, n \geq 1\}$ for every k, such that

$$Y_{k,n} \nearrow X_k \quad \text{as} \quad n \to \infty.$$

Such sequences exist by definition and consistency.

In addition, we introduce the non-negative simple random variables

$$Z_n = \max_{1 \leq k \leq n} Y_{k,n}, \quad n \geq 1.$$

By construction, and domination, respectively,

$$Y_{k,n} \leq Z_n \leq X_n, \quad \text{and} \quad E\, Y_{k,n} \leq E\, Z_n \leq E\, X_n. \tag{5.2}$$

Letting $n \to \infty$ and then $k \to \infty$ in the point-wise inequality yields

$$X_k \leq \lim_{n\to\infty} Z_n \leq \lim_{n\to\infty} X_n = X \quad \text{and then} \quad X \leq \lim_{n\to\infty} Z_n \leq X,$$

respectively, so that,

$$\lim_{n\to\infty} E\, Z_n = E\, X = E \lim_{n\to\infty} Z_n, \tag{5.3}$$

where the first equality holds by definition (and consistency), and the second one by equivalence (Theorem 4.4(g)).

The same procedure in the inequality between the expectations in (5.2) yields

$$E\, X_k \leq \lim_{n\to\infty} EZ_n \leq \lim_{n\to\infty} E\, X_n,$$

and then

$$\lim_{k\to\infty} E\, X_k \leq \lim_{n\to\infty} E\, Z_n \leq \lim_{n\to\infty} E\, X_n.$$

Combining the latter one with (5.3) finally shows that

$$\lim_{n\to\infty} E\, X_n = \lim_{n\to\infty} E\, Z_n = E \lim_{n\to\infty} Z_n = E\, X.$$

□

The following variation for non-increasing sequences immediately suggests itself.

Corollary 5.1. *Let* $\{X_n,\, n \geq 1\}$ *be non-negative random variables and suppose that* X_1 *is integrable. If* $X_n \searrow X$ *as* $n \to \infty$, *then*

$$E\,X_n \searrow E\,X \quad \text{as} \quad n \to \infty.$$

Proof. Since $0 \leq 2X - X_n \nearrow X$ as $n \to \infty$, the conclusion is, indeed, a corollary of the monotone convergence theorem. □

A particular case of importance is when the random variables X_n are partial sums of other random variables. The monotone convergence theorem then translates as follows:

Corollary 5.2. *Suppose that* $\{Y_n,\, n \geq 1\}$ *are non-negative random variables. Then*

$$E\Big(\sum_{n=1}^{\infty} Y_n\Big) = \sum_{n=1}^{\infty} E\,Y_n.$$

Exercise 5.1. Please write out the details of the translation. □

In Example 5.1 we found that the limit of the expected values coincided with the expected value in some cases and was larger in others. This is a common behavior.

Theorem 5.2. (Fatou's lemma)
(i) *If* $\{X_n,\, n \geq 1\}$ *are non-negative random variables, then*

$$E\,\liminf_{n\to\infty} X_n \leq \liminf_{n\to\infty} E\,X_n.$$

(ii) *If, in addition,* Y *and* Z *are integrable random variables, such that* $Y \leq X_n \leq Z$ *a.s. for all* n, *then*

$$E\,\liminf_{n\to\infty} X_n \leq \liminf_{n\to\infty} EX_n \leq \limsup_{n\to\infty} E\,X_n \leq E\,\limsup_{n\to\infty} X_n.$$

Proof. (i): Set $Y_n = \inf_{k\geq n} X_k$, $n \geq 1$. Since

$$Y_n = \inf_{k\geq n} X_k \nearrow \liminf_{n\to\infty} X_n \quad \text{as} \quad n \to \infty,$$

the monotone convergence theorem yields

$$E\,Y_n \nearrow E\,\liminf_{n\to\infty} X_n.$$

Moreover, since $Y_n \leq X_n$, Theorem 4.4(h) tells us that

$$E\,Y_n \leq E\,X_n \quad \text{for all} \quad n.$$

Combining the two proves (i).

To prove (ii) we begin by noticing that

$$\liminf_{n\to\infty}(X_n - Y) = \liminf_{n\to\infty} X_n - Y \quad \text{and} \quad \liminf_{n\to\infty}(Z - X_n) = Z - \limsup_{n\to\infty} X_n,$$

after which (ii) follows from (i) and additivity, since $\{X_n - Y,\, n \geq 1\}$ and $\{Z - X_n,\, n \geq 1\}$ are non-negative random variables. □

Remark 5.2. The right-hand side of (i) may be infinite.

Remark 5.3. If the random variables are are indicators, the result transforms into an inequality for probabilities and we rediscover Theorem 1.3.2. Technically, if $X_n = I\{A_n\}$, $n \geq 1$, then (i) reduces to $P(\liminf_{n\to\infty} A_n) \leq \liminf_{n\to\infty} P(A_n)$, and so on.

A typical use of Fatou's lemma is in cases where one knows that a pointwise limit exists, and it is enough to assert that the expected value of the limit is finite. This situation will be commonplace in Chapter 5. However, if, in addition, the sequence of random variables is dominated by another, integrable, random variable, we obtain another celebrated result.

Theorem 5.3. (The Lebesgue dominated convergence theorem)
Suppose that $|X_n| \leq Y$, for all n, where $E\,Y < \infty$, and that $X_n \to X$ a.s. as $n \to \infty$. Then

$$E|X_n - X| \to 0 \quad \textit{as} \quad n \to \infty,$$

In particular,

$$E\,X_n \to E\,X \quad \textit{as} \quad n \to \infty.$$

Proof. Since also $|X| \leq Y$ it follows that $|X_n - X| \leq 2Y$, so that by replacing X_n by $|X_n - X|$, we find that the proof reduces to showing that if $0 \leq X_n \leq Y \in L^1$, and $X_n \to 0$ almost surely as $n \to \infty$, then $E\,X_n \to 0$ as $n \to \infty$. This, however, follows from Theorem 5.2(ii). □

Remark 5.4. If, in particular, Y is constant, that is, if the random variables are uniformly bounded, $|X_n| \leq C$, for all n and some constant C, the result is sometimes called the *bounded* convergence theorem.

Remark 5.5. In the special case when the random variables are indicators of measurable sets we rediscover the last statement in Theorem 1.3.2:

$$A_n \to A \quad \Longrightarrow \quad P(A_n) \to P(A) \quad \text{as} \quad n \to \infty.$$

□

The following corollary, the verification of which we leave as an exercise, parallels Corollary 5.2.

Corollary 5.3. *Suppose that $\{Y_n,\, n \geq 1\}$ are random variables, such that $\big|\sum_{n=1}^{\infty} Y_n\big| \leq X$, where X is integrable. If $\sum_{n=1}^{\infty} Y_n$ converges a.s. as $n \to \infty$, then $\sum_{n=1}^{\infty} Y_n$, as well as every Y_n, are integrable, and*

$$E\Big(\sum_{n=1}^{\infty} Y_n\Big) = \sum_{n=1}^{\infty} E\,Y_n.$$

This concludes our presentation of expected values. Looking back we find that the development is rather sensitive in the sense that after having traversed elementary random variables, the sequence of results, that is, extensions, convergence results, uniqueness, and so on, have to be pursued in the correct order. Although many things, such as linearity, say, are intuitively "obvious" we must remember that when the previous section began we *knew* nothing about expected values – everything had to be verified.

Let us also mention that one can define expected values in different, albeit equivalent ways. Which way one chooses is mainly a matter of taste.

Exercise 5.2. Prove that the definition

$$E\,X = \sup_{0\le Y\le X} \{E\,Y : Y \text{ is a simple random variable}\}$$

is equivalent to Definition 4.2.

Exercise 5.3. Review the last two sections in the language of Subsection 2.2.2, i.e., "translate" the results (and the proofs) into the language of mathematics. □

6 Indefinite Expectations

In mathematical terminology one integrates over sets. In probabilistic terms we suppose that X is an integrable random variable, and consider expressions of the form

$$\mu_X(A) = E\,XI\{A\} = \int_A X\,\mathrm{d}P = \int_\Omega XI\{A\}\,\mathrm{d}P, \qquad \text{where} \quad A \in \mathcal{F}.$$

In other words, $\mu_X(\,\cdot\,)$ is an "ordinary" expectation applied to the random variable $XI\{\,\cdot\,\}$. In order to justify the definition and the equalities it therefore suffices to consider indicator variables, for which the equalities reduce to equalities between probabilities – note that $\mu_{I\{\Lambda\}}(A) = P(\Lambda\cap A)$ for $\Lambda \in \mathcal{F}$ –, after which one proceeds via non-negative simple random variables, monotone convergence, and $X = X^+ - X^-$ according to the usual procedure.

The notation $\mu_X(\,\cdot\,)$ suggests that we are confronted with a *signed measure* with respect to the random variable X, that is, a measure that obeys the properties of a probability measure except that it can take negative values, and that the total mass need not be equal to 1. If X is non-negative and integrable the expression suggests that μ is a non-negative, finite measure, and if $E\,X = 1$ a probability measure.

Theorem 6.1. *Suppose that X is a non-negative, integrable random variable. Then:*

(a) $\mu_X(\emptyset) = 0$.
(b) $\mu_X(\Omega) = E\,X$.
(c) $P(A) = 0 \implies \mu_X(A) = 0$.

(d) *If* $\mu_X(A) = 0$ *for all* $A \in \mathcal{F}$*, then* $X = 0$ *a.s.*
(e) *If* $\{A_n,\, n \geq 1\}$ *are disjoint sets, then* $\mu_X(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu_X(A_n)$.
(f) *If* $\mu_X(A) = 0$ *for all* $A \in \mathcal{A}$*, where* $\mathcal{A}$ *is a* π*-system that generates* $\mathcal{F}$*, then* $X = 0$ *a.s.*
(g) *If* $\mu_X(A) = 0$ *for all* $A \in \mathcal{A}$*, where* $\mathcal{A}$ *is an algebra that generates* $\mathcal{F}$*, then* $X = 0$ *a.s.*

Proof. The conclusions follow, essentially, from the definition and the different equivalent forms of $\mu_X(\,\cdot\,)$. For (a)–(e) we also need to exploit some of the earlier results from this chapter, and for (f) and (g) we additionally need Theorems 1.2.3 and 1.2.2, respectively.

Exercise 6.1. Spell out the details. □

Remark 6.1. The theorem thus verifies that μ_X is a finite measure whenever X is a non-negative integrable random variable. □

It is now possible to extend the theorem to arbitrary integrable random variables by considering positive and negative parts separately, and to compare measures, corresponding to different random variables, by paralleling the development for ordinary expectations.

Theorem 6.2. *Suppose that* X *and* Y *are integrable random variables. Then:*

(i) *If* $\mu_X(A) = \mu_Y(A)$ *for all* $A \in \mathcal{F}$*, then* $X = Y$ *a.s.*
(ii) *If* $\mu_X(A) = \mu_Y(A)$ *for all* $A \in \mathcal{A}$*, where* $\mathcal{A}$ *is a* π*-system that generates* $\mathcal{F}$*, then* $X = Y$ *a.s.*
(iii) *If* $\mu_X(A) = \mu_Y(A)$ *for all* $A \in \mathcal{A}$*, where* $\mathcal{A}$ *is an algebra that generates* $\mathcal{F}$*, then* $X = Y$ *a.s.*

Exercise 6.2. Once again we urge the reader to fill in the proof. □

The following result is useful for integrals over tails or small, shrinking sets of integrable random variables.

Theorem 6.3. *Let* X *be a random variable with finite mean, and* A *and* A_n*,* $n \geq 1$*, be arbitrary measurable sets (events). Then:*

(i) $\big|\mu_X(\{|X| > n\})\big| \leq \mu_{|X|}(\{|X| > n\}) \to 0$ *as* $n \to \infty$.
(ii) *If* $P(A_n) \to 0$ *as* $n \to \infty$*, then* $\big|\mu_X(A_n)\big| \leq \mu_{|X|}(A_n) \to 0$ *as* $n \to \infty$.

Proof. Since the inequalities are consequences of the basic properties it suffices to prove the conclusion for non-negative random variables.

Thus, suppose that $X \geq 0$. The first claim follows from monotone convergence, Theorem 5.1(i) and linearity. Namely, since $XI\{X \leq n\} \nearrow X$, which is integrable, it follows that

$$E\,XI\{X \leq n\} \nearrow E\,X < \infty \quad \text{as} \quad n \to \infty,$$

so that

$$\mu_X(\{X > n\}) = E\,XI\{X > n\} = E\,X - E\,XI\{X \le n\} \searrow 0 \quad \text{as} \quad n \to \infty.$$

As for (ii), let $M > 0$. Then

$$\begin{aligned}\mu_X(A_n) &= E\,XI\{A_n\} = E\,XI\{A_n \cap \{X \le M\}\} + E\,XI\{A_n \cap \{X > M\}\} \\ &\le MP(A_n) + E\,XI\{X > M\},\end{aligned}$$

so that

$$\limsup_{n\to\infty} E\,XI\{A_n\} \le E\,XI\{X > M\}.$$

The conclusion now follows from (i), since $E\,XI\{X > M\}$ can be made arbitrarily small by choosing M large enough. □

Remark 6.2. Note the idea in (ii) to split the set A_n into a "nice" part which can be handled in more detail, and a "bad" part which is small. This device is used abundantly in probability theory (and in analysis in general) and will be exploited several times as we go on. □

7 A Change of Variables Formula

We have seen that random variables are functions from the sample space to the real line, and we have defined expectations of random variables in terms of integrals over the sample space. Just as the probability space behind the random variables sinks into the background, once they have been properly defined by the induced measure (Theorem 1.1), one would, in the same vein, prefer to compute an integral on the real line rather than over the probability space. Similarly, since measurable functions of random variables are new random variables (Theorem 1.3), one would also like to find the relevant integral corresponding to expectations of functions of random variables. The following theorem, which we might view as the establishing of "induced expectations", settles the problem.

Theorem 7.1. (i) *Suppose that X is integrable. Then*

$$E\,X = \int_\Omega X\,\mathrm{d}P = \int_{\mathbb{R}} x\,\mathrm{d}F_X(x).$$

(ii) *Let X be a random variable, and suppose that g is a measurable function, such that $g(X)$ is an integrable random variable. Then*

$$E\,g(X) = \int_\Omega g(X)\,\mathrm{d}P = \int_{\mathbb{R}} g(x)\,\mathrm{d}F_X(x).$$

Proof. We follow the usual procedure.
(i) If X is an indicator random variable, $X = I\{A\}$ for some $A \in \mathcal{F}$, then the three members all reduce to $P(X \in A)$. If X is a simple random variable, $X = \sum_{k=1}^{n} x_k I\{A_k\}$, where $\{A_k,\, 1 \le k \le n\}$ is a partition of Ω, then the three members reduce to $\sum_{k=1}^{n} P(A_k)$. If X is non-negative, the conclusion follows by monotone convergence, and for the general case we use $X = X^+ - X^-$ and additivity.
(ii) We proceed as in (i) with g playing the role of X. If $g(x) = I_A(x)$, then

$$\{\omega : g(X(\omega)) = 1\} = \{\omega : X(\omega) \in A\},$$

so that

$$E\, g(X) = P(X \in A) = \int_A \mathrm{d}F_X(x) = \int_{\mathbb{R}} g(x)\, \mathrm{d}F_X(x).$$

If g is simple, the conclusion follows by linearity, if g is non-negative by monotone convergence, and, finally, in the general case by decomposition into positive and negative parts. □

Exercise 7.1. As always, write out the details. □

By analyzing the proof we notice that if X is discrete, then X is, in fact, an elementary random variable (recall Definition 1.1), that is, an infinite sum $\sum_{k=1}^{\infty} x_k I\{A_k\}$. If X is non-negative, then, by monotonicity,

$$E\, X = \sum_{k=1}^{\infty} x_k P(A_k),$$

and in the general case this holds by the usual decomposition. This, and the analogous argument for $g(X)$, where g is measurable proves the following variation of the previous result in the discrete and absolutely continuous cases, respectively.

Theorem 7.2. *If X is a discrete random variable with probability function $p_X(x)$, g is a measurable function, and $E|g(X)| < \infty$, then*

$$E\, g(X) = \int_\Omega g(X)\, \mathrm{d}P = \sum_{k=1}^{\infty} g(x_k) p_X(x_k) = \sum_{k=1}^{\infty} g(x_k) P(X = x_k).$$

Proof. We use the decomposition $A_k = \{X = x_k\}$, $k = 1, 2, \ldots$, and $A_0 = \left(\bigcup_{n=1}^{\infty} A_k\right)^c$, observing that $P(A_0) = 0$. □

Theorem 7.3. *If X is an absolutely continuous random variable, with density function $f_X(x)$, g is a measurable function, and $E|g(X)| < \infty$, then*

$$E\, g(X) = \int_\Omega g(X)\, \mathrm{d}P = \int_{-\infty}^{\infty} g(x) f_X(x)\, \mathrm{d}x.$$

Proof. If $g(x) = I_A(x)$ is an indicator, of $A \in \mathcal{R}$, say, then

$$E\,g(X) = \int_\Omega I\{A\}\,\mathrm{d}P = P(A) = \int_A f_X(x)\,\mathrm{d}x$$
$$= \int_{-\infty}^{\infty} I_A(x) f_X(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x,$$

after which one proceeds along the usual scheme. □

In addition to being a computational vehicle, the formula for computing $E\,g(X)$ shows that we do not need to know the distribution of $g(X)$ in order to find its mean.

Example 7.1. Let $X \in U(0,1)$, and suppose, for example, that $g(x) = \sin x$. Then

$$E\,\sin X = \int_0^1 \sin x\,\mathrm{d}x = 1 - \cos 1,$$

whereas one has to turn to the arcsin function in order to find the density of $\sin X$. And, ironically, if one then computes $E\,\sin X$, one obtains the same integral as three lines ago after a change of variable. □

8 Moments, Mean, Variance

Expected values measure the center of gravity of a distribution; they are measures of *location*. In order to describe a distribution in brief terms there exist additional measures, such as the variance which measures the *dispersion* or spread, and moments.

Definition 8.1. *Let X be a random variable. The*

- moments *are $E\,X^n$, $n = 1, 2, \ldots$;*
- central moments *are $E(X - E\,X)^n$, $n = 1, 2, \ldots$;*
- absolute moments *are $E|X|^n$, $n = 1, 2, \ldots$;*
- absolute central moments *are $E|X - E\,X|^n$, $n = 1, 2, \ldots$.*

The first moment, $E\,X$, is the mean*. The second central moment is called* variance*:*

$$\operatorname{Var} X = E(X - E\,X)^2 \quad (= E\,X^2 - (E\,X)^2).$$

All of this, provided the relevant quantities exist. □

Following are tables which provide mean and variance for the standard discrete and absolutely continuous distributions listed earlier in this chapter. The reader is advised to check that the entries have been correctly inserted in both tables.

Mean and variance for the Cantor distribution will be given in Section 2.11 ahead.

Distribution	Notation	Mean	Variance
One point	$\delta(a)$	a	0
Symmetric Bernoulli		0	1
Bernoulli	$\mathrm{Be}(p)$	p	pq
Binomial	$\mathrm{Bin}(n,p)$	np	npq
Geometric	$\mathrm{Ge}(p)$	$\frac{q}{p}$	$\frac{q}{p^2}$
First success	$\mathrm{Fs}(p)$	$\frac{1}{p}$	$\frac{q}{p^2}$
Poisson	$\mathrm{Po}(m)$	m	m

Table 2.3. Mean and variance for some discrete distributions

Distribution	Notation	Mean	Variance
Uniform	$U(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
	$U(0,1)$	$\frac{1}{2}$	$\frac{1}{12}$
	$U(-1,1)$	0	$\frac{1}{3}$
Triangular	$\mathrm{Tri}(-1,1)$	0	$\frac{1}{6}$
Exponential	$\mathrm{Exp}(\theta)$	θ	θ^2
Gamma	$\Gamma(p,\theta)$	$p\theta$	$p\theta^2$
Beta	$\beta(r,s)$	$\frac{r}{r+s}$	$\frac{rs}{(r+s)^2(r+s+1)}$
Normal	$N(\mu,\sigma^2)$	μ	σ^2
	$N(0,1)$	0	1
Log-normal	$LN(\mu,\sigma^2)$	$\mathrm{e}^{\mu+\frac{1}{2}\sigma^2}$	$\mathrm{e}^{2\mu}(\mathrm{e}^{2\sigma^2}-\mathrm{e}^{\sigma^2})$
Cauchy	$C(0,1)$	–	–
Pareto	$\mathrm{Pa}(\beta,\alpha)$	$\frac{\alpha\beta}{\alpha-1}$	$\frac{\alpha\beta^2}{(\alpha-2)(\alpha-1)}$

Table 2.4. Mean and variance for some absolutely continuous distributions

The Cauchy distribution possesses neither mean nor variance. The expected value and variance for the Pareto distribution only exist for $\alpha > 1$ and $\alpha > 2$, respectively (as is suggested by the formulas).

If we think of the physical interpretation of mean and variance it is reasonable to expect that a linear transformation of a random variable changes the center of gravity linearly, and that a translation does not change the dispersion. The following exercise puts these observations into formulas.

Exercise 8.1. Prove the following properties for linear transformations: Let X be a random variable with $E\,X = \mu$ and $\operatorname{Var} X = \sigma^2$, and set $Y = aX + b$, where $a, b \in \mathbb{R}$. Prove that

$$E\,Y = a\mu + b \quad \text{and that} \quad \operatorname{Var} Y = a^2\sigma^2. \qquad \square$$

Two Special Examples Revisited

In Subsection 2.2.7 we presented two examples, the first of which was a random variable X which was uniformly distributed on the irrationals in $[0,1]$, that is, with density

$$f(x) = \begin{cases} 1, & \text{for} \quad x \in [0,1] \setminus \mathbb{Q}, \\ 0, & \text{for} \quad x \in [0,1] \cap \mathbb{Q}. \end{cases}$$

The random variable was there seen to be equivalent to a standard $U(0,1)$-distributed random variable, so that a direct computation shows that $E\,X = 1/2$ and that $\operatorname{Var} X = 1/12$.

The other example was a discrete random variable with probability function

$$p(r_k) = \begin{cases} \frac{6}{\pi^2 k^2}, & \text{for} \quad r_k \in (0,1) \cap \mathbb{Q}, \\ 0, & \text{otherwise}, \end{cases}$$

where $\{r_k,\, k \geq 1\}$ was an enumeration of the rationals in the unit interval. We also pointed out that this is a somewhat pathological situation, since the enumeration of the rationals is not unique. This means that all moments, in particular the expected value and the variance, are ambiguous quantities in that they depend on the actual enumeration of $\mathbb{Q}$.

9 Product Spaces; Fubini's Theorem

Expectations of functions of random vectors are defined in the natural way as the relevant multidimensional integral. The results from Section 2.7 carry over, more or less by notation, that is, by replacing appropriate roman letters by boldface ones.

For example, if $(X,Y)'$ is a random vector and g a measurable function, then

$$E\,g(X,Y) = \int_\Omega g(X,Y)\,\mathrm{d}P = \int_{\mathbb{R}^2} g(x,y)\,\mathrm{d}F_{X,Y}(x,y).$$

In the discrete case,

$$E\,g(X,Y) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} g(x_i,x_j)p_{X,Y}(x_i,x_j),$$

and in the absolutely continuous case

$$E\,g(X,Y) = \int_{\mathbb{R}^2} g(x,y)f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y.$$

In each case the proviso is absolute convergence.

Expectations of functions of random variables take special and useful forms when the probability spaces are product spaces.

9.1 Finite-dimensional Product Measures

Let $(\Omega_k, \mathcal{F}_k, P_k)$, $1 \leq k \leq n$, be probability spaces. We introduce the notation

$$\mathcal{F}_1 \times \mathcal{F}_2 \times \cdots \times \mathcal{F}_n = \sigma\{F_1 \times F_2 \times \cdots \times F_n : F_k \in \mathcal{F}_k,\, k = 1, 2, \ldots, n\}.$$

Given this setup one can now construct a product space, $(\times_{k=1}^n \Omega_k, \times_{k=1}^n \mathcal{F}_k)$, with an associated probability measure $\mathbb{P}$, such that

$$\mathbb{P}(A_1 \times A_2 \times \cdots \times A_n) = \prod_{k=1}^n P_k(A_k) \quad \text{for} \quad A_k \in \mathcal{F}_k,\, 1 \le k \le n.$$

Note that the probability measure has a built-in independence.

Moreover, the probability space $(\times_{k=1}^n \Omega_k, \times_{k=1}^n \mathcal{F}_k, \times_{k=1}^n P_k)$ thus obtained is unique. We refer to the literature on measure theory for details.

As for infinite dimensions we confine ourselves to mentioning the existence of a theory. A prominent example is the space of continuous functions on the unit interval and the associated σ-algebra – $(C[0,1], \mathcal{C}[0,1])$. For this and more we recommend [20].

9.2 Fubini's Theorem

Fubini's theorem is a result on integration, which amounts to the fact that an expectation, which in its general form is a double integral, can be evaluated as iterated single integrals.

Theorem 9.1. *Let $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$ be probability spaces, and consider the product space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, P)$, where $P = P_1 \times P_2$ is the product measure as defined above, suppose that $\mathbf{X} = (X_1, X_2)'$ is a two-dimensional random variable, and that g is $\mathcal{F}_1 \times \mathcal{F}_2$-measurable, and* (i) *non-negative or* (ii) *integrable. Then*

$$\begin{aligned} E\, g(\mathbf{X}) &= \int_\Omega g(\mathbf{X})\, \mathrm{d}P = \int_{\Omega_1 \times \Omega_2} g(X_1, X_2)\, \mathrm{d}(P_1 \times P_2) \\ &= \int_{\Omega_1} \Big(\int_{\Omega_2} g(\mathbf{X})\, \mathrm{d}P_2 \Big) \mathrm{d}P_1 = \int_{\Omega_2} \Big(\int_{\Omega_1} g(\mathbf{X})\, \mathrm{d}P_1 \Big) \mathrm{d}P_2. \end{aligned}$$

Proof. For indicators the theorem reduces to the construction of product measure, after which one proceeds via simple functions, monotone convergence and non-negative functions and the usual decomposition. We omit all details. □

A change of variables (recall Section 2.7) applied to Fubini's theorem yields the following computationally more suitable variant.

Theorem 9.2. *Suppose that $(X, Y)'$ is a two-dimensional random variable, and g is $\mathcal{R}^2 = \mathcal{R} \times \mathcal{R}$-measurable, and non-negative or integrable. Then*

$$\begin{aligned} E\, g(X, Y) &= \iint_{\mathbb{R}^2} g(x, y)\, \mathrm{d}F_X(x)\, \mathrm{d}F_Y(y) \\ &= \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} g(x, y)\, \mathrm{d}F_Y(y) \Big) \mathrm{d}F_X(x) \\ &= \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} g(x, y)\, \mathrm{d}F_X(x) \Big) \mathrm{d}F_Y(y). \end{aligned}$$

Exercise 9.1. Write down the analogous formulas in the absolutely continuous and discrete cases, respectively. □

9.3 Partial Integration

A first application of Fubini's theorem is to show that the usual formula for partial integration carries over to the present context.

Theorem 9.3. *Let $a < b \in \mathbb{R}$, and suppose that $F, G \in D^+$ have no common points of discontinuity on $(a, b]$. Then*

$$\int_a^b G(x)\,\mathrm{d}F(x) = G(b)F(b) - G(a)F(a) - \int_a^b F(x)\,\mathrm{d}G(x).$$

If, in addition, G is absolutely continuous with density g, then

$$\int_a^b G(x)\,\mathrm{d}F(x) = G(b)F(b) - G(a)F(a) - \int_a^b F(x)g(x)\,\mathrm{d}x.$$

Proof. We first note that if the formula holds for F and G, then, by linearity, it also holds for linear transformations; $\alpha F + \beta$ and $\gamma G + \delta$, since then

$$\begin{aligned}\int_a^b \gamma G(x)\,\mathrm{d}(\alpha F(x)) &= \gamma\alpha \int_a^b G(x)\mathrm{d}F(x)\\ &= \gamma\alpha\left(G(b)F(b) - G(a)F(a) - \int_a^b F(x)\,\mathrm{d}G(x)\right)\\ &= \big(\gamma G(b)\big)\big(\alpha F(b)\big) - \big(\gamma G(a)\big)\big(\alpha F(a)\big) - \int_a^b \big(\alpha F(x)\big)\,\mathrm{d}(\gamma G(x)),\end{aligned}$$

and

$$\begin{aligned}\int_a^b (G(x)+\delta)\,\mathrm{d}(F(x)+\beta) &= \int_a^b G(x)\,\mathrm{d}F(x) + \delta(F(b)-F(a))\\ &= G(b)F(b) - G(a)F(a) - \int_a^b F(x)\,\mathrm{d}G(x) + \delta(F(b)-F(a))\\ &= (G(b)+\delta)F(b) - (G(a)+\delta)F(a) - \int_a^b F(x)\,\mathrm{d}(G(x)+\delta).\end{aligned}$$

It is therefore no restriction to assume that F and G are true distribution functions, which we associate with the random variables X and Y, respectively, the point being that we can express the integrals as probabilities. Namely, by an appeal to Fubini's theorem, we obtain, on the one hand, that

$$\begin{aligned}P(a < X \le b, a < Y \le b) &= \int_a^b \int_a^b \mathrm{d}(F\times G)(x,y) = \int_a^b \int_a^b \mathrm{d}F(x)\mathrm{d}G(y)\\ &= \int_a^b \mathrm{d}F(x) \int_a^b \mathrm{d}G(y) = \big(F(b)-F(a)\big)\big(G(b)-G(a)\big),\end{aligned}$$

and, by splitting the probability that the point (X, Y) lies inside the square $(a, b] \times (a, b]$ into three pieces, on the other hand (via product measure and Fubini), that

$$\begin{aligned}
P(a < X \le b, a < Y \le b) &= P(a < X < Y \le b) + P(a < Y < X \le b) \\
&\qquad + P(a < Y = X \le b) \\
&= \int_a^b \int_a^x \mathrm{d}(F \times G)(x, y) + \int_a^b \int_a^x \mathrm{d}(G \times F)(x, y) + 0 \\
&= \int_a^b \Big(\int_a^x \mathrm{d}F(y) \Big) \mathrm{d}G(x) + \int_a^b \Big(\int_a^x \mathrm{d}G(y) \Big) \mathrm{d}F(x) \\
&= \int_a^b \big(F(x) - F(a) \big) \mathrm{d}G(x) + \int_a^b \big(G(x) - G(a) \big) \mathrm{d}F(x) \\
&= \int_a^b F(x) \, \mathrm{d}G(x) + \int_a^b G(x) \, \mathrm{d}F(x) - F(a) \big(G(b) - G(a) \big) \\
&\qquad - G(a) \big(F(b) - F(a) \big).
\end{aligned}$$

The formula for partial integration now follows by equating the two expressions for $P(a < X \le b, a < Y \le b)$.

The conclusion for the special case when G is absolutely continuous follows from the fact that

$$\int_a^b F(x) \, \mathrm{d}G(x) = \int_a^b F(x) g(x) \, \mathrm{d}x. \qquad \square$$

Remark 9.1. The interval $(a, b]$ can be replaced by infinite intervals provided enough integrability is available. $\square$

9.4 The Convolution Formula

Consider once again the usual product space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, P_1 \times P_2)$, and suppose that $(X_1, X_2)'$ is a two-dimensional random variable whose marginal distribution functions are F_1 and F_2, respectively. The convolution formula provides the distribution of $X_1 + X_2$.

Theorem 9.4. *In the above setting*

$$F_{X_1 + X_2}(u) = \int_{-\infty}^{\infty} F_1(u - y) \, \mathrm{d}F_2(y).$$

If, in addition, X_2 is absolutely continuous with density f_2, then

$$F_{X_1 + X_2}(u) = \int_{-\infty}^{\infty} F_1(u - y) f_2(y) \, \mathrm{d}y.$$

If X_1 is absolutely continuous with density f_1, the density of the sum equals

$$f_{X_1+X_2}(u) = \int_{-\infty}^{\infty} f_1(u-y)\,\mathrm{d}F_2(y).$$

If both are absolutely continuous, then

$$f_{X_1+X_2}(u) = \int_{-\infty}^{\infty} f_1(u-y) f_2(y)\,\mathrm{d}y.$$

Proof. Once again, an application of Fubini's theorem does the job for us.

$$\begin{aligned} F_{X_1+X_2}(u) = P(X_1+X_2 \le u) &= \iint_{x+y\le u} \mathrm{d}(F_1\times F_2)(x,y) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{u-y} \mathrm{d}(F_1\times F_2)(x,y) = \int_{-\infty}^{\infty}\Big(\int_{-\infty}^{u-y}\mathrm{d}F_1(x)\Big)\mathrm{d}F_2(y) \\ &= \int_{-\infty}^{\infty} F_1(u-y)\,\mathrm{d}F_2(y). \end{aligned}$$

The remaining parts are immediate. □

10 Independence

One of the central concepts in probability theory is independence. The outcomes of repeated tosses of coins and throws of dice are "independent" in a sense of normal language, meaning that coins and dice do not have a memory. The successive outcomes of draws *without replacements* of cards from a deck are not independent, since a card that has been drawn cannot be drawn again. The mathematical definition of independence differs from source to source. Luckily the two following ones are equivalent.

Definition 10.1. *The random variables X_1, X_2, …, X_n are* independent *iff, for arbitrary Borel measurable sets $A_1, A_2, \ldots, A_n$,*

$$P\Big(\bigcap_{k=1}^{n}\{X_k\in A_k\}\Big) = \prod_{k=1}^{n} P(X_k\in A_k).$$

Definition 10.2. *The random variables X_1, X_2, …, X_n or, equivalently, the components of the random vector $\mathbf{X}$ are* independent *iff*

$$F_{\mathbf{X}}(\mathbf{x}) = \prod_{k=1}^{n} F_{X_k}(x_k) \quad \textit{for all} \quad \mathbf{x}\in\mathbb{R}^n.$$ □

Independence according to the first definition thus means that all possible joint events are independent, and according to the second definition that the joint distribution function equals the product of the marginal ones.

Theorem 10.1. *The two definitions are equivalent.*

Proof. The second definition obviously is implied by the first one, since the half-open infinite sets are a subclass of all measurable sets. For the converse we note that this subclass is a π-system that generates the σ-algebra of Borel measurable sets (Theorem 1.3.6). An application of Theorem 1.3.5 finishes the proof. □

Remark 10.1. Independence implies that the joint measure is product measure (due to uniqueness). □

Exercise 10.1. Prove that it is, in fact, enough to check any class of sets that generates the Borel sets to assert independence. □

For discrete and absolutely continuous distributions independence is equivalent to the factorization of joint probability functions and joint densities, respectively.

Theorem 10.2. (i) *If X and Y are discrete, then X and Y are independent iff the joint probability function is equal to the product of the marginal ones, that is iff*

$$p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y) \quad \textit{for all} \quad x, y \in \mathbb{R}.$$

(ii) *If X and Y are absolutely continuous, then X and Y are independent iff the joint density is equal to the product of the marginal ones, that is iff*

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y) \quad \textit{for all} \quad x, y \in \mathbb{R}.$$

Proof. The discrete case follows immediately by taking differences.

As for the absolutely the continuous case, if factorization holds, then, via Fubini's Theorem, Theorem 9.1,

$$\begin{aligned} F_{X,Y}(x,y) &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\,\mathrm{d}u\mathrm{d}v = \int_{-\infty}^{x}\int_{-\infty}^{y} f_X(u)f_Y(v)\,\mathrm{d}u\mathrm{d}v \\ &= \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u \int_{-\infty}^{y} f_Y(v)\,\mathrm{d}v = F_X(x)\cdot F_Y(y). \end{aligned}$$

To prove the converse, we use the metatheorem approach. Suppose that X and Y are independent and define, for $C = A \times B$, where $A, B \in \mathcal{R}$,

$$\mathcal{E} = \left\{ C : \iint_C f_{X,Y}(u,v)\,\mathrm{d}u\mathrm{d}v = \iint_C f_X(u)f_Y(v)\,\mathrm{d}u\mathrm{d}v. \right\}$$

Let, for $x, y \in \mathbb{R}$, $A = (-\infty, x]$ and $B = (-\infty, y]$. Then, by definition, the independence assumption, and Fubini's theorem,

$$\begin{aligned} \iint_C f_{X,Y}(u,v)\,\mathrm{d}u\mathrm{d}v &= P(A \cap B) = P(A)P(B) \\ &= \int_A f_X(u)\,\mathrm{d}u \int_B f_Y(v)\,\mathrm{d}v = \iint_{A\times B} f_X(u)f_Y(v)\,\mathrm{d}u\mathrm{d}v \\ &= \iint_C f_X(u)f_Y(v)\,\mathrm{d}u\mathrm{d}v. \end{aligned}$$

This shows that $\mathcal{E}$ contains all rectangles. Since the class of rectangles constitutes a π-system and generate the Borel σ-algebra, Theorem 1.2.3 tells us that $\mathcal{E} = \mathcal{R}$. □

A more modern (but less common) definition (which we state for $n = 2$) is that X and Y are independent iff

$$E\, g(X)h(Y) = E\, g(X) \cdot E\, h(Y) \quad \text{for all} \quad g, h \in C_B,$$

where C_B is the class of bounded continuous functions. For details and equivalences, see [145], Chapter 10.

Exercise 10.2. Prove, via simple functions, non-negative functions, monotone convergence, and differences of non-negative functions, that this definition is equivalent to the other ones. □

Exercise 10.3. Prove that if X_1, X_2, ..., X_n are independent, then

$$E \prod_{k=1}^{n} |X_k|^{s_k} = \prod_{k=1}^{n} E|X_k|^{s_k},$$

where $s_1, s_2, \ldots, s_n$ are positive reals, and that

$$E \prod_{k=1}^{n} X_k^{j_k} = \prod_{k=1}^{n} E X_k^{j_k},$$

where $j_1, j_2, \ldots, j_n$ are positive integers. □

Two of the basic properties of expectations were additivity and linearity. A related question concerns variances; if X and Y are random variables with finite variances, is it true that the variance of the sum equals the sum of the variances? Do variances have the linearity property? These questions are (partially) answered next.

Theorem 10.3. *Let X and Y be* independent *random variables with finite variances, and $a, b \in \mathbb{R}$. Then*

$$\begin{aligned} \operatorname{Var} aX &= a^2 \operatorname{Var} X, \\ \operatorname{Var}(X + Y) &= \operatorname{Var} X + \operatorname{Var} Y, \\ \operatorname{Var}(aX + bY) &= a^2 \operatorname{Var} X + b^2 \operatorname{Var} Y. \end{aligned}$$

Exercise 10.4. Prove the theorem. □

Remark 10.2. Independence is sufficient for the variance of the sum to be equal to the sum of the variances, but not necessary.

Remark 10.3. Linearity should not hold, since variance is a quadratic quantity.

Remark 10.4. Note, in particular, that $\operatorname{Var}(-X) = \operatorname{Var} X$. This is as expected, since switching the sign should not alter the spread of the distribution. □

10.1 Independence of Functions of Random Variables

The following theorem puts the natural result that functions of independent random variables are independent into print.

Theorem 10.4. *Let X_1, X_2, …, X_n be random variables and $h_1, h_2, \ldots, h_n$, be measurable functions. If X_1, X_2, …, X_n are independent, then so are $h_1(X_1), h_2(X_2), \ldots, h_n(X_n)$.*

Proof. Let $A_1, A_2, \ldots, A_n$ be Borel measurable sets. Then, by turning to inverse images and the Definition 10.1, we find that

$$P\Big(\bigcap_{k=1}^{n}\{h_k(X_k) \in A_k\}\Big) = P\Big(\bigcap_{k=1}^{n}\{X_k \in h_k^{-1}(A_k)\}\Big)$$
$$= \prod_{k=1}^{n} P\big(X_k \in h_k^{-1}(A_k)\big) = \prod_{k=1}^{n} P(h_k(X_k) \in A_k). \qquad \square$$

10.2 Independence of σ-Algebras

As an analog to Theorem 1.4.1, independence of random variables implies independence of the σ-algebras generated by them.

Theorem 10.5. *If X_1, X_2, …, X_n are independent, then so are*

$$\sigma\{X_1\},\ \sigma\{X_2\},\ \ldots,\ \sigma\{X_n\}.$$

Exercise 10.5. Write out the details of the proof. $\square$

10.3 Pair-wise Independence

Recall the distinction between independence and *pair-wise* independence of sets from Section 1.4. The same distinction exists for random variables.

Definition 10.3. *The random variables X_1, X_2, …, X_n are* pair-wise independent *iff all pairs are independent.* $\square$

Independence obviously implies pair-wise independence, since there are several additional relations to check in the former case. The following example shows that there exist random variables that are pair-wise independent, but not (completely) independent.

Example 10.1. Pick one of the points $(1,0,0)$, $(0,1,0)$, $(0,0,1)$, and $(1,1,1)$ uniformly at random, and set, for $k = 1, 2, 3$,

$$X_k = \begin{cases} 1, & \text{if coordinate } k = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Then, with $A_k = \{X_k = 1\}$, we rediscover Example 1.4.1, which proves the desired assertion. In addition,

$$E\,X_k = \frac{1}{2}, \quad \text{for} \quad k = 1, 2, 3,$$
$$E(X_1X_2X_3) = \frac{1}{4} \neq E\,X_1E\,X_2E\,X_3 = \frac{1}{8}.$$

However, since $X_iX_j = 1$ if the point $(1, 1, 1)$ is chosen, and $X_iX_j = 0$ otherwise, we obtain

$$P(X_iX_j = 1) = \frac{1}{4}, \quad \text{and} \quad P(X_iX_j = 0) = \frac{3}{4}.$$

for all pairs (i, j), where $(i \neq j)$, which implies that

$$E\,X_iX_j = \frac{1}{4} = E\,X_iE\,X_j.$$

In other words, moment factorization holds for pairs but not for triplets. □

Exercise 10.6. Prove that if X_1, X_2, ..., X_n are *pair-wise* independent, then

$$\mathrm{Var}\,(X_1 + X_2 + \cdots + X_n) = \mathrm{Var}\,X_1 + \mathrm{Var}\,X_2 + \cdots + \mathrm{Var}\,X_n.$$ □

10.4 The Kolmogorov Zero-one Law Revisited

The proof of the following Kolmogorov zero-one law for random variables amounts to a translation of the proof of the zero-one law for events, Theorem 1.5.1.

Let $\{X_n, n \geq 1\}$ be arbitrary random variables, and set

$$\mathcal{F}_n = \sigma\{X_1, X_2, \ldots, X_n\} \qquad \text{for} \quad n \geq 1,$$
$$\mathcal{F}'_n = \sigma\{X_{n+1}, X_{n+2}, \ldots\} \qquad \text{for} \quad n \geq 0.$$

Then

$$\mathcal{T} = \bigcap_{n=0}^{\infty} \mathcal{F}'_n$$

is the tail-σ-field (with respect to $\{X_n, n \geq 1\}$).

Theorem 10.6. (The Kolmogorov zero-one law)
Suppose that $\{X_n, n \geq 1\}$ are independent random variables. If $A \in \mathcal{T}$, then

$$P(A) = 0 \quad \textit{or} \quad 1.$$

Exercise 10.7. Prove the theorem, that is, rewrite (e.g.) the second proof of Theorem 1.5.1 into the language of random variables. □

Corollary 10.1. *If, in the setting of the Theorem 10.6, X is a $\mathcal{T}$-measurable random variable, then X is a.s. constant.*

Proof. The event $\{X \le x\} \in \mathcal{T}$ for all $x \in \mathbb{R}$. Thus,

$$F_X(x) = P(X \le x) = 0 \quad \text{or} \quad 1 \quad \text{for all} \quad x \in \mathbb{R},$$

which, in view of the properties of distribution functions, implies that there exists $c \in \mathbb{R}$, such that

$$F_X(x) = P(X \le x) = \begin{cases} 0, & \text{for} \quad x < c, \\ 1, & \text{for} \quad x \ge c. \end{cases}$$ □

A consequence of the corollary is that random variables, such as limits, limit superior and limit inferior of sequences of independent random variables must be constant a.s. if they converge at all.

11 The Cantor Distribution

A beautiful way to describe a random variable that has the Cantor distribution on the unit interval is the following: Let X, X_1, X_2, ... be independent identically distributed random variables such that

$$P(X = 0) = P(X = 2) = \frac{1}{2}.$$

Then

$$Y = \sum_{n=1}^{\infty} \frac{X_n}{3^n} \in \text{Cantor}(0, 1).$$

Namely, the random variables X_1, X_2, ... are the successive decimals of a number whose decimals in the base 3 expansion are 0 or 2, and never 1. Moreover, since the decimals each time have a 50-50 chance of being 0 or 2, the infinite sum that constitutes Y is uniformly distributed over the Cantor set.

To compute the mean we use additivity and monotone convergence for series to obtain

$$E\,Y = E\Big(\sum_{n=1}^{\infty} \frac{X_n}{3^n}\Big) = \sum_{n=1}^{\infty} E\Big(\frac{X_n}{3^n}\Big) = \sum_{n=1}^{\infty} \frac{E\,X_n}{3^n} = \sum_{n=1}^{\infty} \frac{1}{3^n} = \frac{1}{2},$$

which coincides with intuition.

To verify the result for the variance we also need the fact that the summands are independent (and an additional argument due to the fact that we are faced with an infinite series) to obtain

$$\operatorname{Var} Y = \operatorname{Var}\left(\sum_{n=1}^{\infty} \frac{X_n}{3^n}\right) = \sum_{n=1}^{\infty} \operatorname{Var}\left(\frac{X_n}{3^n}\right) = \sum_{n=1}^{\infty} \frac{\operatorname{Var} X_n}{(3^n)^2} = \sum_{n=1}^{\infty} \frac{1}{9^n} = \frac{1}{8}.$$

By letting X be equal to 0 with probability 1/2, and equal to some other positive integer with probability 1/2, and by modifying Y accordingly, we can construct other Cantor-type distributions. For example,

$$Z = \sum_{n=1}^{\infty} \frac{X_n}{4^n}, \quad \text{where} \quad P(X=0) = P(X=3) = \frac{1}{2},$$

is a random variable corresponding to a number that is uniform over the subset of the interval $[0,1]$ which consists of the numbers whose base 4 decimal expansion contains only 0's and 3's, no 1's or 2's.

We have thus exhibited two different Cantor-type distributions.

Exercise 11.1. We have not explicitly *proved* that the base 4 example produces a continuous singular distribution. Please check that this is the case (although this seems pretty clear since the constructions is the same as that of the Cantor distribution). □

Exercise 11.2. Compute $E\,Z$ and $\operatorname{Var} Z$. □

Although Cantor sets have Lebesgue measure 0 they are, somehow, of different "sizes" in the sense that some are more "nullish" than others. After all, in the classical, first case, we delete one-third of the support in each step, whereas, in the second case we delete halves. The null set in the first case therefore seems larger than in the second case.

There exists, in fact, a means to classify such sets, namely the *Hausdorff dimension*, which can be used to measure the dimension of sets (such as fractals), whose topological dimension is not a natural number. One can show that the Hausdorff dimension of the classical Cantor set on the unit interval is $\log 2/\log 3$, and that the Hausdorff dimension pertaining to our second example is $\log 2/\log 4 = 1/2 < \log 2/\log 3 \approx 0.631$, and, hence smaller than the classical Cantor set.

We close by mentioning that the same argument with 3 (or 4) replaced by 2, and X_k being 0 or 1 with equal probabilities for all k, generates a number that is $U(0,1)$-distributed (and, hence, an absolutely continuous distribution), since it is the binary expansion of such a number. Its Hausdorff dimension is, in fact, equal to $\log 2/\log 2 = 1$, (which coincides with the topological dimension).

12 Tail Probabilities and Moments

The existence of an integral or a moment clearly depends on how quickly tails decay. It is therefore not far-fetched to guess that there exist precise results concerning this connection.

Theorem 12.1. *Let $r > 0$, and suppose that X is a non-negative random variable. Then:*

(i) $E\,X = \int_0^\infty (1 - F(x))\,\mathrm{d}x = \int_0^\infty P(X > x)\,\mathrm{d}x$,
where both members converge or diverge simultaneously;

(ii) $E\,X^r = r\int_0^\infty x^{r-1}(1 - F(x))\,\mathrm{d}x = r\int_0^\infty x^{r-1}P(X > x)\,\mathrm{d}x$,
where both members converge or diverge simultaneously;

(iii) $E\,X < \infty \iff \sum_{n=1}^\infty P(X \geq n) < \infty$.
More precisely,

$$\sum_{n=1}^\infty P(X \geq n) \leq E\,X \leq 1 + \sum_{n=1}^\infty P(X \geq n).$$

(iv) $E\,X^r < \infty \iff \sum_{n=1}^\infty n^{r-1}P(X \geq n) < \infty$.
More precisely,

$$\sum_{n=1}^\infty n^{r-1}P(X \geq n) \leq E\,X^r \leq 1 + \sum_{n=1}^\infty n^{r-1}P(X \geq n).$$

Proof. (i) and (ii): Let $A > 0$. By partial integration,

$$\begin{aligned}\int_0^A x^r\,\mathrm{d}F(x) &= -A^r(1 - F(A)) + \int_0^A rx^{r-1}(1 - F(x))\,\mathrm{d}x \\ &= -A^rP(X > A) + r\int_0^A x^{r-1}P(X > x)\,\mathrm{d}x.\end{aligned}$$

If $E\,X^r < \infty$, then

$$A^r(1 - F(A)) \leq \int_A^\infty x^r\,\mathrm{d}F(x) \to 0 \quad \text{as} \quad A \to \infty,$$

which shows that the integral on the right-hand side converges. If, on the other hand, the latter converges, then so does the integral on the left-hand side since it is smaller.

As for (iii),

$$\begin{aligned}E\,X &= \sum_{n=1}^\infty \int_{n-1}^n x\,\mathrm{d}F(x) \leq \sum_{n=1}^\infty nP(n-1 < |X| \leq n) \\ &= \sum_{n=1}^\infty \sum_{k=1}^n P(n-1 < |X| \leq n) = \sum_{k=1}^\infty \sum_{n=k}^\infty P(n-1 < |X| \leq n) \\ &= \sum_{k=1}^\infty P(X > k-1) \leq 1 + \sum_{k=1}^\infty P(X > k) \leq 1 + \sum_{k=1}^\infty P(X \geq k).\end{aligned}$$

The other half follows similarly, since

$$E\,X \geq \sum_{n=1}^{\infty}(n-1)P(n-1 < |X| \leq n),$$

after which the computations are the same as before, and (iv) follows by "slicing" the corresponding integral similarly. □

Remark 12.1. Alternatively, it suffices to prove (i), because

$$E\,X^r = \int_0^\infty P(X^r > x)\,\mathrm{d}x = \int_0^\infty P(X > x^{1/r})\,\mathrm{d}x,$$

after which the change of variable $y = x^{1/r}$ establishes the claim. □

If X is integer valued one can be a little more precise.

Theorem 12.2. *If X is a non-negative, integer valued random variable, then*

$$E\,X = \sum_{n=1}^{\infty} P(X \geq n).$$

Proof. The conclusion can be obtained from Theorem 12.1, or, else, directly:

$$\begin{aligned} E\,X &= \sum_{n=1}^{\infty} nP(X=n) = \sum_{n=1}^{\infty}\Big(\sum_{k=1}^{n} 1\Big)P(X=n) \\ &= \sum_{k=1}^{\infty}\sum_{n=k}^{\infty} P(X=n) = \sum_{k=1}^{\infty} P(X \geq k). \end{aligned}$$

Interchanging the order of summation is no problem since all terms are non-negative. □

Exercise 12.1. Let X and Y be random variables and suppose that $E|Y| < \infty$. Show that, if there exists $x_0 > 0$, such that

$$P(|X| > x) \leq P(|Y| > x) \quad \text{for all} \quad x > x_0,$$

then $E|X| < \infty$. □

By modifying the proof of Theorem 12.1 one can obtain the following more general results.

Theorem 12.3. *Let X be a non-negative random variable, and g a non-negative, strictly increasing, differentiable function. Then,*

(i) $E\,g(X) = g(0) + \int_0^\infty g'(x)P(X > x)\,\mathrm{d}x$*, where both members converge or diverge simultaneously;*
(ii) $E\,g(X) < \infty \iff \sum_{n=1}^{\infty} g'(n)P(X > n) < \infty$.

Exercise 12.2. Prove the theorem. □

Exercise 12.3. Let X be a non-negative random variable. Prove that

$$E\,\log^+ X < \infty \quad \Longleftrightarrow \quad \sum_{n=1}^{\infty} \frac{1}{n} P(X > n) < \infty;$$

$$E\,\log^+\log^+ X < \infty \quad \Longleftrightarrow \quad \sum_{n=1}^{\infty} \frac{1}{n\log n} P(X > n) < \infty;$$

$$E\,X^r(\log^+ X)^p < \infty \quad \Longleftrightarrow \quad \sum_{n=1}^{\infty} n^{r-1}(\log n)^p P(X > n) < \infty, \quad r > 1,\ p > 0;$$

$$E(\log^+ X)^p < \infty \quad \Longleftrightarrow \quad \sum_{n=1}^{\infty} \frac{(\log n)^{p-1}}{n} P(X > n) < \infty, \quad p > 1. \qquad \square$$

A common proof technique is to begin by proving a desired result for some subsequence. In such cases one sometimes runs into sums of the above kind for subsequences. The following results may then be useful.

Theorem 12.4. *Let X be a non-negative random variable, and $\lambda > 1$. Then,*

$$E\,X < \infty \iff \int_0^{\infty} \lambda^x P(X > \lambda^x)\,\mathrm{d}x < \infty \iff \sum_{n=1}^{\infty} \lambda^n P(X > \lambda^n) < \infty.$$

Proof. By a change of variable, $y = \lambda^x$,

$$\int_0^{\infty} \lambda^x P(X > \lambda^x)\,\mathrm{d}x = \log\lambda \int_0^{\infty} P(X > y)\,\mathrm{d}y,$$

which, together with Theorem 12.1 proves the conclusion. $\square$

More general subsequences can be handled as follows.

Theorem 12.5. *Suppose that $\{n_k,\ k \geq 1\}$ is a strictly increasing subsequence of the positive integers, and set*

$$m(x) = \#\{k : n_k \leq x\} \quad \text{and} \quad M(x) = \sum_{k=1}^{[x]} n_k, \quad x > 0.$$

Finally, let X be a non-negative random variable. Then

$$\sum_{k=1}^{\infty} n_k P(X \geq n_k) = E\,M(m(X)),$$

where both sides converge and diverge together.

Proof. The conclusion follows from the fact that

$$\{X \geq n_k\} = \{m(X) \geq k\},$$

partial summation and Theorem 12.1. $\square$

Exercise 12.4. Verify the following special cases:

$$E\,X^{3/2} < \infty \quad \Longleftrightarrow \quad \sum_{k=1}^{\infty} k^2 P(X \geq k^2) < \infty;$$

$$E\,X^{1+(1/d)} < \infty \quad \Longleftrightarrow \quad \sum_{k=1}^{\infty} k^d P(X \geq k^d) < \infty \quad \text{for} \quad d \in \mathbb{N}.$$

Exercise 12.5. Show that Theorem 12.5 reduces to Theorem 12.4 for $n_k = \lambda^k$ where $\lambda > 1$. □

The subsequences we have dealt with so far were at most geometrically increasing. For more rapidly increasing subsequences we have the following special case.

Theorem 12.6. *Suppose that* $\{n_k,\, k \geq 1\}$ *is a strictly increasing subsequence of the positive integers, such that*

$$\limsup_{k\to\infty} \frac{n_k}{n_{k+1}} < 1,$$

and let X *be a non-negative random variable. Then*

$$E\,X < \infty \quad \Longrightarrow \quad \sum_{k=1}^{\infty} n_k P(X \geq n_k) < \infty.$$

Proof. Set $\Sigma = \sum_{k=1}^{\infty} n_k P(X \geq n_k)$. A consequence of the growth condition is that there exists $\lambda > 1$, such that $n_{k+1} \geq \lambda n_k$ for all k, so that

$$\begin{aligned}
\Sigma &= \sum_{k=1}^{\infty} \big(n_{k-1} + (n_k - n_{k-1})\big) P(X \geq n_k) \\
&\leq \sum_{k=1}^{\infty} \lambda^{-1} n_k + \sum_{k=1}^{\infty} \sum_{j=n_{k-1}+1}^{n_k} P(X \geq j) \\
&\leq \lambda^{-1}\Sigma + \sum_{j=1}^{\infty} P(X \geq j) = \lambda^{-1}\Sigma + E\,X,
\end{aligned}$$

so that

$$\Sigma \leq \frac{\lambda}{\lambda - 1} E\,X < \infty.$$

□

Remark 12.2. Combining this with Theorem 12.5 shows that

$$E\,M(m(X)) \leq \frac{\lambda}{\lambda - 1} E\,X.$$

The last result is, in general, weaker than the previous one, although frequently sufficient. If, in particular, $M(m(x)) \geq Cx$ as $x \to \infty$, the results coincide. One such example is $n_k = 2^{2^k}$, $k \geq 1$. □

Another variation involves double sums.

Theorem 12.7. *Let X be a non-negative random variable. Then*

$$E\,X\log^+ X < \infty \quad \Longleftrightarrow \quad \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} P(X > nm) < \infty.$$

Proof. By modifying the proof of Theorem 12.1(i) we find that the double sum converges iff

$$\int_1^\infty \int_1^\infty P(X > xy)\,\mathrm{d}x\mathrm{d}y < \infty.$$

Changing variables $u = x$ and $v = xy$ transforms the double integral into

$$\int_1^\infty \int_1^v \frac{1}{u} P(X > v)\,\mathrm{d}u\mathrm{d}v = \int_1^\infty \log v P(X > v)\,\mathrm{d}v,$$

and the conclusion follows from Theorem 12.3. □

13 Conditional Distributions

Conditional distributions in their complete generality involve some rather delicate mathematical complications. In this section we introduce this concept for pairs of purely discrete and purely absolutely continuous random variables. Being an essential ingredient in the theory of martingales, *conditional expectations* will be more thoroughly discussed in Chapter 10.

Definition 13.1. *Let X and Y be discrete, jointly distributed random variables. For $P(X = x) > 0$, the* conditional probability function *of Y* given that $X = x$ *equals*

$$p_{Y|X=x}(y) = P(Y = y \mid X = x) = \frac{p_{X,Y}(x, y)}{p_X(x)}\,,$$

and the conditional distribution function *of Y* given that $X = x$ *is*

$$F_{Y|X=x}(y) = \sum_{z \le y} p_{Y|X=x}(z).$$

□

Exercise 13.1. Show that $p_{Y|X=x}(y)$ is a probability function of a true probability distribution. □

This definition presents no problems. It is validated by the definition of conditional probability; just put $A = \{X = x\}$ and $B = \{Y = y\}$. If, however, X and Y are jointly absolutely continuous, expressions like $P(Y = y \mid X = x)$ have no meaning, since they are of the form $\frac{0}{0}$. However, a glance at the previous definition suggests the following one.

Definition 13.2. *Let X and Y have a joint absolutely continuous distribution. For $f_X(x) > 0$, the* conditional density function *of Y* given that $X = x$ *equals*

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}\,,$$

and the conditional distribution function of Y *given that $X = x$ is*

$$F_{Y|X=x}(y) = \int_{-\infty}^{y} f_{Y|X=x}(z)\,dz. \qquad \square$$

Exercise 13.2. Show that $f_{Y|X=x}(y)$ is the density function of a true probability distribution

Exercise 13.3. Prove that if X and Y are independent then the conditional distributions and the unconditional distributions are the same. Explain why this is reasonable. $\square$

Remark 13.1. The definitions can (of course) be extended to situations with more than two random variables. $\square$

By combining the expression for the marginal density with the definition of conditional density we obtain the following density version of the law of total probability, Proposition 1.4.1:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X=x}(y) f_X(x)\,\mathrm{d}x. \tag{13.1}$$

We also formulate, leaving the details to the reader, the following mixed version, in which Y is discrete and X absolutely continuous:

$$P(Y = y) = \int_{-\infty}^{\infty} p_{Y|X=x}(y) f_X(x)\,\mathrm{d}x. \tag{13.2}$$

Example 13.1. In Example 3.1 a point was chosen uniformly on the unit disc. The joint density was $f_{X,Y}(x,y) = \frac{1}{\pi}$, for $x^2 + y^2 \le 1$, and 0 otherwise, and we found that the marginal densities were $f_X(x) = f_Y(x) = \frac{2}{\pi}\sqrt{1-x^2}$, for $|x| < 1$ and 0 otherwise.

Using this we find that the conditional density of the y-coordinate given the x-coordinate equals

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{1/\pi}{\frac{2}{\pi}\sqrt{1-x^2}} = \frac{1}{2\sqrt{1-x^2}} \quad \text{for} \quad |y| \le \sqrt{1-x^2},$$

and 0 otherwise. This shows that the conditional distribution is uniform on the interval $(-\sqrt{1-x^2}, \sqrt{1-x^2})$.

This should not be surprising, since we can view the joint distribution in the three-dimensional space as a homogeneous, circular cake with a thickness

equal to $1/\pi$. The conditional distributions can then be viewed as the profile of a face after a vertical cut across the cake. And this face, which is a picture of the marginal distribution is a rectangle.

Note also that the conditional density is a function of the x-coordinate, which means that the coordinates are not independent (as they would have been if the cake were a square and we make a cut parallel to one of the coordinate axes).

The conditional density of the x-coordinate given the y-coordinate is the same, by symmetry. □

A simple example involving discrete distributions is that we pick a digit randomly among $0, 1, 2, \ldots, 9$, and then a second one among those that are smaller than the first one. The corresponding continuous analog is to break a stick of length 1 randomly at some point, and then break one of the remaining pieces randomly.

14 Distributions with Random Parameters

Random variables with random parameters are very natural objects. For example, suppose that X follows a Poisson distribution, but in such a way that the parameter itself is random. An example could be a particle counter that emits particles of different kinds. For each kind the number of particles emitted during one day, say, follows a Poisson distribution, However, the parameters for the different kinds are different. Or the intensity depends on temperature or air pressure, which, in themselves, are random. Another example could be an insurance company that is subject to claims according to some distribution, the parameter of which depends on the kind of claim: is it a house on fire? a stolen bicycle? a car that has been broken into? Certainly, the intensities with which these claims occur can be expected to be different.

It could also be that the parameter is unknown. The so-called Bayesian approach is to consider the parameter as a random variable with a so-called prior distribution.

Let us for computational convenience consider the following situation:

$$X \in \text{Po}(M) \quad \text{where} \quad M \in \text{Exp}(1).$$

This is an abusive way of writing that

$$X \mid M = m \in \text{Po}(m) \quad \text{with} \quad M \in \text{Exp}(1).$$

What is the "real" (that is, the unconditional) distribution of X? Is it a Poisson distribution? Is it definitely not a Poisson distribution?

By use of the mixed version (13.2) of the law of total probability, the following computation shows tells us that X is geometric; $X \in \text{Ge}(\frac{1}{2})$. Namely, for $k = 0, 1, 2, \ldots$ we obtain

$$P(X=k) = \int_0^\infty P(X=k \mid M=x)\cdot f_M(x)\,\mathrm{d}x = \int_0^\infty \mathrm{e}^{-x}\frac{x^k}{k!}\cdot \mathrm{e}^{-x}\,\mathrm{d}x$$
$$= \int_0^\infty \frac{x^k}{k!}\mathrm{e}^{-2x}\,\mathrm{d}x = \frac{1}{2^{k+1}}\cdot\int_0^\infty \frac{1}{\Gamma(k+1)}2^{k+1}x^{k+1-1}\mathrm{e}^{-2x}\,\mathrm{d}x$$
$$= \frac{1}{2^{k+1}}\cdot 1 = \frac{1}{2}\cdot\Big(\frac{1}{2}\Big)^k,$$

which establishes the geometric distribution as claimed.

Exercise 14.1. Determine the distribution of X if

- $M \in \mathrm{Exp}(a)$;
- $M \in \Gamma(p,a)$. □

Suppose that a radioactive substance emits α-particles in such a way that the number of particles emitted during one hour, $N \in \mathrm{Po}(\lambda)$. Unfortunately, though, the particle counter is unreliable in the sense that an emitted particle is registered with probability $p \in (0,1)$, whereas it remains unregistered with probability $q = 1-p$. All particles are registered independently of each other. Let X be the number of particles that are registered during one hour.

This means that our model is

$$X \mid N = n \in \mathrm{Bin}(n,p) \quad \text{with} \quad N \in \mathrm{Po}(\lambda).$$

So, what is the unconditional distribution of X? The following computation shows that $X \in \mathrm{Po}(\lambda p)$. Namely, for $k = 0,1,2,\ldots$,

$$P(X=k) = \sum_{n=0}^\infty P(X=k \mid N=n)P(N=n) = \sum_{n=k}^\infty \binom{n}{k}p^k q^{n-k}\mathrm{e}^{-\lambda}\frac{\lambda^n}{n!}$$
$$= \mathrm{e}^{-\lambda}\frac{(\lambda p)^k}{k!}\sum_{n=k}^\infty \frac{(\lambda q)^{n-k}}{(n-k)!} = \mathrm{e}^{-\lambda}\frac{(\lambda p)^k}{k!}\mathrm{e}^{\lambda q} = \mathrm{e}^{-\lambda p}\frac{(\lambda p)^k}{k!}.$$

Note that the sum starts at $n = k$; there must be at least as many particles emitted as there are registered ones.

The following two exercises may or may not have anything to do with everyday life.

Exercise 14.2. Susan has a coin with $P(\text{head}) = p_1$ and John has a coin with $P(\text{head}) = p_2$. Susan tosses her coin m times. Each time she obtains heads, John tosses his coin (otherwise not). Find the distribution of the total number of heads obtained by John.

Exercise 14.3. Toss a coin repeatedly, and let X_n be the number of heads after n coin tosses, $n \geq 1$. Suppose now that the coin is completely unknown to us in the sense that we have no idea of whether or not it is fair. Suppose, in fact, the following, somewhat unusual situation, namely, that

$$X_n \mid P = p \in \mathrm{Bin}(n,p) \quad \text{with} \quad P \in U(0,1),$$

that is, we suppose that the probability of heads is $U(0,1)$-distributed.

- Find the distribution of X_n.
- Explain why the answer is reasonable.
- Compute $P(X_{n+1} = n+1 \mid X_n = n)$.
- Are the outcomes of the tosses independent? □

A special family of distributions is the family of *mixed normal*, or *mixed Gaussian*, distributions. These are normal distributions with a random variance, namely,

$$X \mid \Sigma^2 = y \in N(\mu, y) \quad \text{with} \quad \Sigma^2 \in F\,,$$

where F is some distribution (on $(0, \infty)$).

As an example, consider a production process where some measurement of the product is normally distributed, and that the production process is not perfect in that it is subject to rare disturbances. More specifically, the observations might be $N(0,1)$-distributed with probability 0.99 and $N(0,100)$-distributed with probability 0.01. We may write this as

$$X \in N(0, \Sigma^2), \quad \text{where} \quad P(\Sigma^2 = 1) = 0.99 \text{ and } P(\Sigma^2 = 100) = 0.01.$$

What is the "real" distribution of X? A close relative is the next section.

15 Sums of a Random Number of Random Variables

In many applications involving processes that evolve with time, one is interested in the state of affairs at some given, fixed, *time* rather than after a given, fixed, number of steps, which therefore amounts to checking the random process or sequence after a *random number of events*. With respect to what we have discussed so far this means that we are interested in the state of affairs of the sum of *a random number* of independent random variables. In this section we shall always assume that *the number of terms is independent of the summands*. More general random indices or "times" will be considered in Chapter 10.

Apart from being a theory in its own right, there are several interesting and important applications; let us, as an appetizer, mention branching processes and insurance risk theory which we shall briefly discuss in a subsection following the theory.

Thus, let X, X_1, X_2, ... be independent, identically distributed random variables with partial sums $S_n = \sum_{k=1}^{n} X_k$, $n \geq 1$, and let N be a non-negative, integer valued random variable which is independent of X_1, X_2, ... Throughout, $S_0 = 0$.

The object of interest is S_N, that is, the sum of N X's. We may thus interpret N as a random index.

For any Borel set $A \subset (-\infty, \infty)$,

$$P(S_N \in A \mid N = n) = P(S_n \in A \mid N = n) = P(S_n \in A), \qquad (15.1)$$

where the last equality, being a consequence of the additional independence, is the crucial one.

Here is an example in which the index is *not* independent of the summands.

Example 15.1. Let $N = \min\{n : S_n > 0\}$. Clearly, $P(S_N > 0) = 1$. This implies that if the summands are allowed to assume negative values (with positive probability) then so does S_n, whereas S_N is always positive. Hence, N is not independent of the summands, on the contrary, N is, in fact, defined in terms of the summands. □

By (15.1) and the law of total probability, Proposition 1.4.1, it follows that

$$\begin{aligned} P(S_N \in A) &= \sum_{n=1}^{\infty} P(S_N \in A \mid N = n) P(N = n) \\ &= \sum_{n=1}^{\infty} P(S_n \in A) P(N = n), \end{aligned} \tag{15.2}$$

in particular,

$$P(S_N \le x) = \sum_{n=1}^{\infty} P(S_n \le x) P(N = n), \quad -\infty < x < \infty, \tag{15.3}$$

so that, by changing the order of integration and summation,

$$E\, h(S_N) = \sum_{n=1}^{\infty} E\big(h(S_n)\big) P(N = n), \tag{15.4}$$

provided the integrals are absolutely convergent.

By letting $h(x) = x$ and $h(x) = x^2$ we obtain expressions for the mean and variance of S_N.

Theorem 15.1. *Suppose that $X, X_1, X_2, \ldots$ are independent, identically distributed random variables with partial sums $S_n = \sum_{k=1}^{n} X_k$, $n \ge 1$, and that N is a non-negative, integer valued random variable which is independent of $X_1, X_2, \ldots$.*

(i) *If*

$$E\, N < \infty \quad \text{and} \quad E\,|X| < \infty,$$

then

$$E\, S_N = E\, N \cdot E\, X.$$

(ii) *If, in addition,*

$$\operatorname{Var} N < \infty \quad \text{and} \quad \operatorname{Var} X < \infty,$$

then

$$\operatorname{Var} S_N = E\, N \cdot \operatorname{Var} X + (E\, X)^2 \cdot \operatorname{Var} N.$$

Proof. (i): From (15.4) we know that

$$\begin{aligned} E\,S_N &= \sum_{n=1}^{\infty} E\,S_n P(N=n) = \sum_{n=1}^{\infty} nE\,XP(N=n) \\ &= E\,X \sum_{n=1}^{\infty} nP(N=n) = E\,XE\,N. \end{aligned}$$

(ii): Similarly

$$\begin{aligned} E\big(S_N^2\big) &= \sum_{n=1}^{\infty} E\big(S_n^2\big)P(N=n) = \sum_{n=1}^{\infty} \big(\operatorname{Var} S_n + (E\,S_n)^2\big)P(N=n) \\ &= \sum_{n=1}^{\infty} \big(n \operatorname{Var} X + n^2 (E\,X)^2\big)P(N=n) \\ &= \operatorname{Var} X \sum_{n=1}^{\infty} nP(N=n) + (E\,X)^2 \sum_{n=1}^{\infty} n^2 P(N=n) \\ &= \operatorname{Var} X E\,N + (E\,X)^2 E\,N^2. \end{aligned}$$

By inserting the conclusion from (i) we find that

$$\begin{aligned} \operatorname{Var} S_N &= E\big(S_N^2\big) - (E\,S_N)^2 = E\,N \operatorname{Var} X + (E\,X)^2 E\,N^2 - \big(E\,NE\,X\big)^2 \\ &= E\,N \operatorname{Var} X + (E\,X)^2 \operatorname{Var} N. \end{aligned}$$

□

15.1 Applications

Applications of this model are ubiquitous. In this subsection we first illustrate the theory with what might be called a toy example, after which we mention a few more serious applications. It should also be mentioned that in some of the latter examples the random index is not necessarily independent of the summands (but this is of no significance in the present context).

A "Toy" Example

Example 15.2. Suppose that the number of customers that arrive at a store during one day is $\operatorname{Po}(\lambda)$-distributed and that the probability that a customer buys something is p and just browses around without buying is $q = 1 - p$. Then the number of customers that buy something can be described as S_N, where $N \in \operatorname{Po}(\lambda)$, and $X_k = 1$ if customer k shops and 0 otherwise.

Theorem 15.1 then tells us that

$$E\,S_N = E\,NE\,X = \lambda \cdot p,$$

and that

$$\operatorname{Var} S_N = E\,N \operatorname{Var} X + (E\,X)^2 \operatorname{Var} N = \lambda \cdot pq + p^2\lambda = \lambda p.$$

We have thus found that $E\,S_N = \operatorname{Var} S_N = \lambda p$, which makes it tempting to guess that, in fact $S_N \in \operatorname{Po}(\lambda p)$. This may seem bold, but knowing that the Poisson process has many "nice" features, this may seem reasonable. After all, the new process can be viewed as the old process after having run through a "filter", which makes it seem like a thinner version of the old one. And, in fact, there is a concept, the *thinned Poisson process*, which is precisely this, and which is Poisson distributed with a parameter that is the product of the old one and the thinning probability.

And, in fact, by (15.2), we have, for $k = 0, 1, 2, \ldots$,

$$\begin{aligned} P(S_N = k) &= \sum_{n=1}^{\infty} P(S_n = k) P(N = n) = \sum_{n=k}^{\infty} \binom{n}{k} p^k q^{n-k} \mathrm{e}^{-\lambda} \frac{\lambda^n}{n!} \\ &= \mathrm{e}^{-\lambda} \frac{(\lambda p)^k}{k!} \sum_{n=k}^{\infty} \frac{(\lambda q)^{n-k}}{(n-k)!} = \mathrm{e}^{-\lambda} \frac{(\lambda p)^k}{k!} \mathrm{e}^{\lambda q} = \mathrm{e}^{-\lambda p} \frac{(\lambda p)^k}{k!}, \end{aligned}$$

so that, indeed $S_n \in \operatorname{Po}(\lambda p)$.

Remark 15.1. The computations for determining the distribution of S_N are the same as in the previous section. The reason for this is that, instead of introducing an indicator random variable to each customer, we may consider the total number of customers as some random variable X, say, and note that $X \mid N = n \in \operatorname{Bin}(n, p)$, after which we proceed as before. This is no surprise, since if we identify every customer with a particle, then a shopping customer is identified with a registered particle. So, they are conceptually the same problem, just modeled or interpreted somewhat differently. □

More generally, let Y_k be the amount spent by the kth customer. The sum $S_N = \sum_{k=1}^{N} Y_k$ then describes the total amount spent by the customers during one day.

If, for example, $Y_1, Y_2, \ldots \in \operatorname{Exp}(\theta)$, and $N \in \operatorname{Fs}(p)$, then

$$E\,S_N = \frac{1}{p} \cdot \theta \quad \text{and} \quad \operatorname{Var} S_N = \frac{1}{p}\theta^2 + \theta^2 \frac{q}{p^2} = \frac{\theta^2}{p^2}. \qquad \square$$

Exercise 15.1. Find the distribution of S_N and check that mean and variance agree with the above ones. □

Branching Processes

The most basic kind of *branching processes*, the *Galton-Watson process*, can be described as follows:

At time $t = 0$ there exists one (or many) founding members $X(0)$. During its life span, every individual gives birth to a random number of children, who during their life spans give birth to a random number of children, who during their life spans ….

The reproduction rules in this model are the same for all individuals:

- all individuals give birth according to the same probability law, independently of each other;
- the number of children produced by an individual is independent of the number of individuals in his or her generation.

Let, for $n \geq 0$, $X(n) = \#$ individuals in generation n, and $\{Y_k,\ k \geq 1\}$ and Y be generic random variables denoting the number of children obtained by individuals. We also suppose that $X(0) = 1$, and exclude the degenerate case $P(Y = 1) = 1$.

It follows from the assumptions that

$$X(2) = Y_1 + \cdots + Y_{X(1)},$$

and, recursively, that

$$X(n+1) = Y_1 + \cdots + Y_{X(n)}.$$

Thus, by identifying $Y_1, Y_2, \ldots$ with $X_1, X_2, \ldots$ and $X(n)$ with N it follows that $X(n+1)$ is an "S_N-sum".

One simple example is cells that split or die, in other words, with probability p they get two children and with probability $1-p$ they die. What happens after many generations? Will the cells spread all over the universe or is the cell culture going to die out? If the cells are antrax cells, say, this question may be of some interest.

Insurance Risk Theory

Consider an insurance company whose business runs as follows:

- Claims arrive at random time points according to some random process;
- Claim sizes are (can be considered as being) independent, identically distributed random variables;
- The gross premium rate, that is, the premium paid by the policy holders, arrive at a constant rate β/month (which is probably not realistic since people pay their bills at the end of the month, just after payday).

Let us denote the number of claims during one year by N, and the successive claims by $X_1, X_2, \ldots$. If the initial capital, called the risk reserve, is v, then the capital at the end of the first year equals

$$v + 12\beta - \sum_{k=1}^{N} X_k.$$

Relevant questions are probabilities of ruin, of ruin in 5 years, and so on. Another important issue is the deciding of premiums, which means that one wishes to estimate parameters from given data, and, for example, investigate if parameters have changed or not.

A Simple Queueing Model

Consider a store to which customers arrive, one at a time, according to some random process (and that the service times, which are irrelevant here, are, say, i.i.d. exponentially distributed random variables). If $X_1, X_2, \ldots$ denotes the amount of money spent by the customers and there are M customers during one day, then

$$\sum_{k=1}^{M} X_k$$

depicts the amount of money in the cash register at the end of the day. The toy example above falls into this category.

16 Random Walks; Renewal Theory

An important assumption in Theorem 15.1 was the *independence of the random index N and the random summands $X_1, X_2, \ldots$*. There obviously exist many situations where such an assumption is unrealistic. It suffices to imagine examples where a process is observed until something "special" occurs. The number of summands at that moment is random and, by construction, defined via the summands. In this section we present some applications where more general random indices are involved.

16.1 Random Walks

A *random walk* $\{S_n, n \geq 0\}$ is a sequence of random variables, starting at $S_0 = 0$, with independent, identically distributed increments $X_1, X_2, \ldots$.

The classical example is the *simple random walk*, for which the increments, or steps, assume the values $+1$ or -1. The standard notation is

$$P(X = 1) = p, \quad P(X = -1) = q, \quad \text{where} \quad 0 \leq p, q \leq 1, \quad p + q = 1,$$

and where X is a generic random variable.

The following figure illustrates the situation.

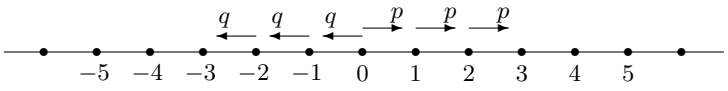

Figure 2.2. The simple random walk

If the values are assumed with equal probabilities, $p = q = 1/2$, we call it a *symmetric simple random walk*. Another example is the *Bernoulli random walk*, where the steps are $+1$ or 0 with probabilities p and q, respectively.

Random walk theory is a classical topic. For an introduction and background we refer to the second edition of Spitzer's legendary 1964 book, [234]. Applications are abundant: Sequential analysis, insurance risk theory, queueing theory, reliability theory, just to name a few.

16.2 Renewal Theory

Renewal processes are random walks with non-negative increments. The canonical application is a light bulb that fails after a random time and is instantly replaced by a new, identical one, which, upon failure is replaced by another one, which, in turn, ... The central object of interest is the number of replacements during a given time.

In order to model a renewal process we let X_1, X_2, ... be the individual life times and set $S_n = \sum_{k=1}^{n} X_k$, $n \geq 1$. The number of replacements in the time interval $(0, t]$ then becomes

$$N(t) = \max\{n : S_n \leq t\}.$$

The following figure depicts a typical realization of a renewal process.

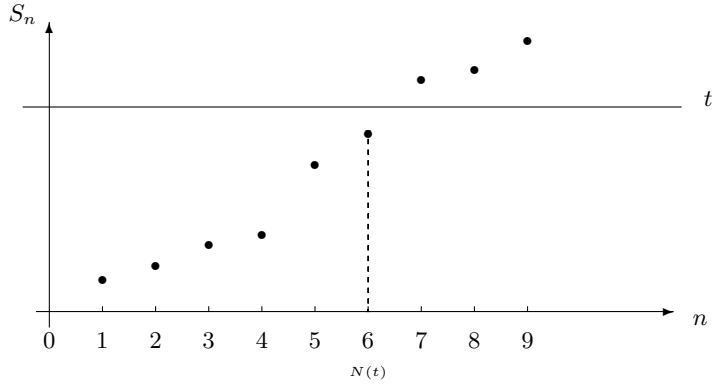

Figure 2.3. The realization of a renewal process

The main process of interest is the *renewal counting process*,

$$\{N(t),\, t \geq 0\}.$$

Some classical references are [53, 65, 201, 229, 230]. A summary of results can be found in [110], Chapter II. A discrete version called *recurrent events* dates back to [85] see also [87]. If, in particular, the life times are exponential, then $\{N(t),\, t \geq 0\}$, is a *Poisson process*.

A more general model which allows for repair times is the *alternating renewal process*, a generalization of which is a two-dimensional random walk, stopped when the second component reaches a given level after which the first component is evaluated at that time point. For more on this, see [118] and/or [110], Chapter IV (and Problem 7.8.17).

Classical proofs for the renewal counting process are based on the inversion

$$\{N(t) \geq n\} = \{S_n \leq t\}. \tag{16.1}$$

The idea is that a limit theorem for one of the processes may be derived from the corresponding limit theorem for the other one via inversion by letting t and n tend to infinity jointly in a suitable manner.

16.3 Renewal Theory for Random Walks

Instead of considering a random walk after a fixed number of steps, that is, at a random time point, one would rather inspect or observe the process at fixed time points, which means after *a random number of steps*. For example, the closing time of a store is fixed, but the number of customers during a day is random. The number of items produced by a machine during an 8-hour day is random, and so on. A typical random index is "the first n, such that ...". With reference to renewal theory in the previous subsection, we also note that it seems more natural to consider a random process at the *first* occurrence of some kind rather than the *last* one, defined by the counting process, let alone, how does one know that a given occurrence really is the last one before having information about the future of the process?

For this model we let $X, X_1, X_2, \ldots$ be independent, identically distributed random variables, with positive, finite, mean $EX = \mu$, and set $S_n = \sum_{k=1}^{n} X_k$, $n \geq 1$. However, instead of the counting process we shall devote ourselves to *the first passage time process*, $\{\tau(t), t \geq 0\}$, defined by

$$\tau(t) = \min\{n : S_n > t\}, \quad t \geq 0.$$

Although the counting process and the first passage time process are close on average, they have somewhat different behaviors in other respects. In addition, first passage times have, somewhat vaguely stated, "better" mathematical properties than last exit times. Some of this vagueness will be clarified in Section 10.14. A more extensive source is [110], Section III.3. Here we confine ourselves by remarking that

- whereas $N(t) + 1 = \tau(t)$ for renewal processes, this is not necessarily the case for random walks;
- the inversion relation (16.1) does not hold for random walks, since the random walk may well fall below the level t after having crossed it.

Both facts may be observed in the next figure.

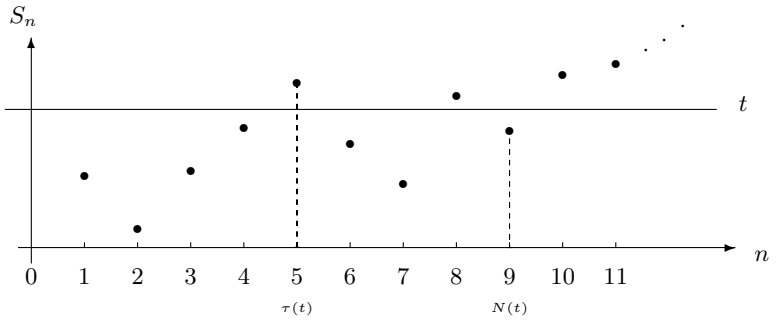

Figure 2.4. First passage times of a random walk

Proofs of subsequent limit theorems for first passage time processes will be based on limit theorems for randomly indexed random walks, $\{S_{N(t)}, t \geq 0\}$.

A special feature is that those proofs cover renewal processes as well as random walks. In addition, no distinction is necessary between the continuous cases and the discrete ones. A specialized reference on this topic is [110].

16.4 The Likelihood Ratio Test

Let $X_1, X_2, \ldots, X_n$ be a sample from an absolutely continuous distribution with a characterizing parameter θ of interest, and suppose that we wish to test the null hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$. The Neyman-Pearson lemma in statistics tells us that such a test should be based on the likelihood ratio statistic

$$L_n = \prod_{k=1}^{n} \frac{f(X_k;\theta_1)}{f(X_k;\theta_0)},$$

where f_{θ_0} and f_{θ_1} are the densities under the null and alternative hypotheses, respectively.

The factors $\frac{f(X_k;\theta_1)}{f(X_k;\theta_0)}$ are independent, identically distributed random variables, and, under the null hypothesis, the mean equals 1;

$$E_0\Big(\frac{f(X_k;\theta_1)}{f(X_k;\theta_0)}\Big) = \int_{-\infty}^{\infty} \frac{f(x;\theta_1)}{f(x;\theta_0)} f(x;\theta_0)\,\mathrm{d}x = \int_{-\infty}^{\infty} f(x;\theta_1)\,\mathrm{d}x = 1,$$

so that L_n equals a product of independent, identically distributed random variables with mean 1.

For technical reasons it is sometimes more convenient to investigate the log-likelihood, $\log L_n$, which is a *sum* of independent, identically distributed random variables, however, *not* with mean $\log 1 = 0$.

16.5 Sequential Analysis

This is one of the most important statistical applications within the renewal theoretic framework. The idea is that, instead of basing a log-likelihood test on a sample of a *fixed* predetermined size, one performs the test sequentially.

The typical sequential procedure then would be to continue sampling until, depending on the circumstances, the likelihood ratio L_n or the log-likelihood ratio, $\log L_n$, falls outside a given strip, at which time point one takes a decision. Technically, this means that one defines

$$\tau_{a,b} = \min\{n : L_n \notin (a,b)\}, \quad \text{where} \quad 0 < a < b < \infty,$$

or, equivalently,

$$\tau_{A,B} = \min\{n : \log L_n \notin (A,B)\}, \quad \text{where} \quad -\infty < A < B < \infty.$$

and continues sampling until the likelihood ratio (or, equivalently, the log-likelihood ratio) escapes from the interval and rejects the null hypothesis

if $L_{\tau_{a,b}} > b$ ($\log L_{\tau_{A,B}} > B$), and accepts the null hypothesis if $L_{\tau_{a,b}} < a$ ($\log L_{\tau_{A,B}} < A$).

Although one can show that the procedure stops after a finite number of steps, that is, that the sample size will be finite almost surely, one may introduce a "time horizon", m, and stop sampling at $\min\{\tau, m\}$ (and accept H_0 if the (log)likelihood-ratio has not escaped from the strip at time m).

The classic here is the famous book by Wald [250]. A more recent one is [223].

16.6 Replacement Based on Age

Let $X_1, X_2, \ldots$ be the independent, identically distributed lifetimes of some component in a larger machine. The simplest replacement policy is to change a component as soon as it fails. In this case it may be necessary to call a repairman at night, which might be costly. Another policy, called *replacement based on age*, is to replace at failure or at some given age, a, say, whichever comes first. The inter-replacement times are

$$W_n = \min\{X_n, a\}, \quad n \geq 1,$$

in this case. A quantity of interest would be the number of replacements due to failure during some given time unit.

In order to describe this quantity we define

$$\tau(t) = \min\Big\{n : \sum_{k=1}^{n} W_k > t\Big\}, \quad t > 0.$$

The quantity $\tau(t)$ equals the number of components that have been in action at time t.

Next, let

$$Z_n = I\{X_n \leq a\}, \quad n \geq 1,$$

that is, $Z_n = 1$ if the nth component is replaced because of failure, and $Z_n = 0$ if replacement is due to age. The number of components that have been replaced because of failure during the time span $(0, t]$ is then described by

$$\sum_{k=1}^{\tau(t)} Z_k.$$

If we attach a cost c_1 to replacements due to failure and a cost c_2 to replacements due to age, then

$$\sum_{k=1}^{\tau(t)} \big(c_1 I\{X_k \leq a\} + c_2 I\{X_k > a\}\big)$$

provides information about the replacement cost during the time span $(0, t]$.

For detailed results on this model, see [110, 118].

Remark 16.1. Replacement based on age applied to humans is called retirement, where a is the retirement age. □

17 Extremes; Records

The central results in probability theory are limit theorems for sums. However, in many applications, such as strength of materials, fatigue, flooding, oceanography, and "shocks" of various kinds, *extremes* rather than sums are of importance. A flooding is the result of one single extreme wave, rather than the cumulative effect of many small ones.

In this section we provide a brief introduction to the concept of extremes – "the largest observation so far" and a more extensive one to the theory of records – "the extreme observations at their first appearance".

17.1 Extremes

Let $X_1, X_2, \ldots$ be independent, identically distributed random variables. The quantities in focus are the *partial maxima*

$$Y_n = \max_{1\le k\le n} X_k \quad \text{or, at times,} \quad \max_{1\le k\le n} |X_k|.$$

Typical results are analogs to the law of large numbers and the central limit theorem for sums. For the latter this means that we wish to find normalizing sequences $\{a_n > 0,\, n \ge 1\}$, $\{b_n \in \mathbb{R},\, n \ge 1\}$, such that

$$\frac{Y_n - b_n}{a_n} \quad \text{possesses a limit distribution,}$$

a problem that will be dealt with in Chapter 9.

17.2 Records

Let $X, X_1, X_2, \ldots$ be independent, identically distributed, continuous random variables. The *record times* are $L(1) = 1$ and, recursively,

$$L(n) = \min\{k : X_k > X_{L(n-1)}\}, \quad n \ge 2,$$

and the *record values* are

$$X_{L(n)}, \quad n \ge 1.$$

The associated *counting process* $\{\mu(n),\, n \ge 1\}$ is defined by

$$\mu(n) = \#\,\text{records among } X_1, X_2, \ldots, X_n = \max\{k : L(k) \le n\}.$$

The reason for assuming continuity is that we wish to avoid ties. And, indeed, in this case we obtain, by monotonicity (Lemma 1.3.1),

$$P\Big(\bigcup_{\substack{i,j=1\\ i\neq j}}^{\infty}\{X_i = X_j\}\Big) = \lim_{n\to\infty} P\Big(\bigcup_{\substack{i,j=1\\ i\neq j}}^{n}\{X_i = X_j\}\Big)$$
$$\leq \lim_{n\to\infty} \sum_{\substack{i,j=1\\ i\neq j}}^{n} P(X_i = X_j) = 0.$$

The pioneering paper in the area is [205]. For a more recent introduction and survey of results, see [187, 207, 208].

Whereas the sequence of partial maxima, Y_n, $n \geq 1$, describe "the largest value so far", the record values pick these values the first time they appear. The sequence of record values thus constitutes a subsequence of the partial maxima. Otherwise put, the sequence of record values behaves like a compressed sequence of partial maxima, as is depicted in the following figure.

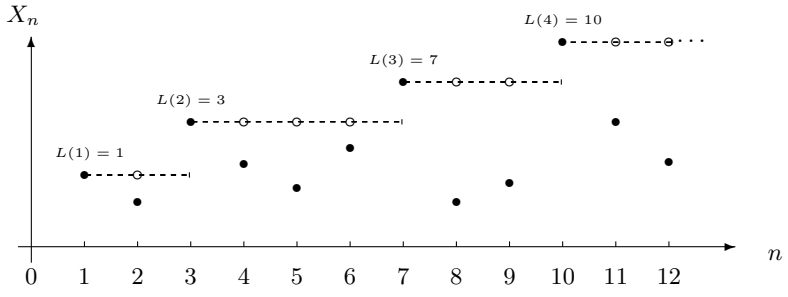

Figure 2.5. Partial maxima ∘

A preliminary observation is that the record times and the number of records are distribution independent. This is a consequence of the fact that given X with distribution function F, then $F(X)$ is $U(0,1)$-distributed, so that there is a $1-1$ map from every (absolutely continuous) random variable to every other one. And, by monotonicity, record times are preserved under this transformation – however, not the record *values*.

Next, set

$$I_k = \begin{cases} 1, & \text{if } X_k \text{ is a record,} \\ 0, & \text{otherwise,} \end{cases}$$

so that $\mu(n) = \sum_{k=1}^{n} I_k$, $n \geq 1$.

By symmetry, all permutations between $X_1, X_2, \ldots, X_n$ are equally likely. Taking advantage of this fact, we introduce *ranks*, so that X_n has rank j if X_n is the jth largest among $X_1, X_2, \ldots, X_n$. Notationally, $R_n = j$. This means, in particular, that if X_n is the largest among them, then $R_n = 1$, and if X_n is the smallest, then $R_n = n$. Moreover,

$$P(R_1 = r_1, R_2 = r_2, \ldots, R_n = r_n) = \frac{1}{n!},$$

in particular,

$$P(I_k = 1) = 1 - P(I_k = 0) = \frac{1}{k}, \quad k = 1, 2, \ldots, n.$$

The marginal probabilities are

$$P(R_n = r_n) = \sum_{\{r_1, r_2, \ldots, r_{n-1}\}} P(R_1 = r_1, R_2 = r_2, \ldots, R_n = r_n),$$

where the summation thus extends over all possible values of $r_1, r_2, \ldots, r_{n-1}$. By symmetry, the summation involves $(n-1)!$ terms, all of which are the same, namely $1/n!$, so that

$$P(R_n = r_n) = \frac{(n-1)!}{n!} = \frac{1}{n}.$$

Since the same argument is valid for all n, we have, in fact, shown that

$$P(R_1 = r_1, R_2 = r_2, \ldots, R_n = r_n) = \frac{1}{n!} = \prod_{k=1}^{n} \frac{1}{k} = \prod_{k=1}^{n} P(R_k = r_k),$$

which proves the independence of the ranks. Moreover, since $\{I_n = 1\} = \{R_n = 1\}$ it follows, in particular, that $\{I_k,\, k \geq 1\}$ are independent random variables.

Joining the above conclusions yields the following result.

Theorem 17.1. *Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed, absolutely continuous, random variables, $n \geq 1$. Then*

(i) *The ranks $R_1, R_2, \ldots, R_n$ are independent, and $P(R_k = j) = 1/k$ for $j = 1, 2, \ldots, k$, where $k = 1, 2, \ldots, n$;*
(ii) *The indicators $I_1, I_2, \ldots, I_n$ are independent, and $P(I_k = 1) = 1/k$ for $k = 1, 2, \ldots, n$.*

As a corollary it is now a simple task to compute the mean and the variance of $\mu(n)$, and their asymptotics.

Theorem 17.2. *Let $\gamma = 0.5772\ldots$ denote Euler's constant. We have*

$$m_n = E\,\mu(n) = \sum_{k=1}^{n} \frac{1}{k} = \log n + \gamma + o(1) \qquad \textit{as} \quad n \to \infty;$$

$$\operatorname{Var} \mu(n) = \sum_{k=1}^{n} \frac{1}{k}\Big(1 - \frac{1}{k}\Big) = \log n + \gamma - \frac{\pi^2}{6} + o(1) \qquad \textit{as} \quad n \to \infty.$$

Proof. That $E\,\mu(n) = \sum_{k=1}^{n} \frac{1}{k}$, and that $\operatorname{Var} \mu(n) = \sum_{k=1}^{n} \frac{1}{k}(1 - \frac{1}{k})$, is clear. The remaining claims follow from Remark A.3.1, and the (well-known) fact that $\sum_{n=1}^{\infty} 1/n^2 = \pi^2/6$. □

18 Borel-Cantelli Lemmas

This section is devoted to an important tool frequently used in connection with questions concerning almost sure convergence – a concept that we shall meet in detail in Chapter 5 – the *Borel-Cantelli lemmas* [26].

We begin by recalling the definitions of limsup and liminf of sets from Chapter 1 and by interpreting them in somewhat greater detail.

Let $\{A_n,\, n \geq 1\}$ be a sequence of events, that is, measurable subsets of Ω. Then, recalling Definition 1.2.1,

$$A_* = \liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m, \quad \text{and} \quad A^* = \limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m.$$

Thus, if $\omega \in \Omega$ belongs to the set $\liminf_{n\to\infty} A_n$, then ω belongs to $\bigcap_{m=n}^{\infty} A_m$ for some n, that is, there exists an n such that $\omega \in A_m$ for *all* $m \geq n$. In particular, if A_n is the event that something special occurs at "time" n, then $\liminf_{n\to\infty} A_n^c$ means that from some n on this property never occurs.

Similarly, if $\omega \in \Omega$ belongs to the set $\limsup_{n\to\infty} A_n$, then ω belongs to $\bigcup_{m=n}^{\infty} A_m$ for every n, that is, no matter how large we choose n there is always some $m \geq n$ such that $\omega \in A_m$, or, equivalently, $\omega \in A_m$ for infinitely many values of m or, equivalently, for arbitrarily large values of m. A convenient way to express this is

$$\omega \in A^* \quad \Longleftrightarrow \quad \omega \in \{A_n \text{ i.o.}\} = \{A_n \text{ infinitely often}\}.$$

If the upper and lower limits coincide the limit exists, and

$$A = A^* = A_* = \lim_{n\to\infty} A_n.$$

18.1 The Borel-Cantelli Lemmas 1 and 2

We now present the standard Borel-Cantelli lemmas, after which we prove a zero-one law and provide an example to illustrate the applicability of the results.

Theorem 18.1. (The first Borel-Cantelli lemma)
Let $\{A_n,\, n \geq 1\}$ *be arbitrary events. Then*

$$\sum_{n=1}^{\infty} P(A_n) < \infty \quad \Longrightarrow \quad P(A_n \text{ i.o.}) = 0.$$

Proof. We have

$$\begin{aligned} P(A_n \text{ i.o.}) = P(\limsup_{n\to\infty} A_n) &= P\Big(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m\Big) \\ &\leq P\Big(\bigcup_{m=n}^{\infty} A_m\Big) \leq \sum_{m=n}^{\infty} P(A_m) \to 0 \quad \text{as} \quad n \to \infty. \end{aligned}$$

□

The converse does not hold in general. The easiest accessible one is obtained under the additional assumption of independence.

Theorem 18.2. (The second Borel-Cantelli lemma)
Let $\{A_n,\, n \geq 1\}$ *be* independent *events. Then*

$$\sum_{n=1}^{\infty} P(A_n) = \infty \quad \Longrightarrow \quad P(A_n \text{ i.o.}) = 1.$$

Proof. By independence,

$$\begin{aligned} P(A_n \text{ i.o.}) &= P\Big(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m \Big) = 1 - P\Big(\bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m^c \Big) \\ &= 1 - \lim_{n\to\infty} P\Big(\bigcap_{m=n}^{\infty} A_m^c \Big) = 1 - \lim_{n\to\infty} \prod_{m=n}^{\infty} P(A_m^c) \\ &= 1 - \lim_{n\to\infty} \prod_{m=n}^{\infty} \big(1 - P(A_m)\big) = 1 - 0 = 1, \end{aligned}$$

since, by Lemma A.4.1, the divergence of $\sum_{n=1}^{\infty} P(A_n)$ is equivalent to the divergence of $\prod_{m=1}^{\infty}(1 - P(A_m))$. □

By combining the two results we note, in particular, that if the events $\{A_n,\, n \geq 1\}$ are independent, then $P(A_n \text{ i.o.})$ can only assume the values 0 or 1, and that the convergence or divergence of $\sum_{n=1}^{\infty} P(A_n)$ is the decisive factor.

Theorem 18.3. (A zero-one law)
If the events $\{A_n,\, n \geq 1\}$ *are independent, then*

$$P(A_n \text{ i.o.}) = \begin{cases} 0, & \textit{when} \quad \sum_{n=1}^{\infty} P(A_n) < \infty, \\ 1, & \textit{when} \quad \sum_{n=1}^{\infty} P(A_n) = \infty. \end{cases}$$ □

A consequence of this zero-one law is that it suffices to prove that $P(A_n \text{ i.o.}) > 0$ in order to conclude that the probability equals 1 (and that $P(A_n \text{ i.o.}) < 1$ in order to conclude that it equals 0).

Here is an example to illuminate the results.

Example 18.1. Let $X_1, X_2, \ldots$ be a sequence of arbitrary random variables and let $A_n = \{|X_n| > \varepsilon\}$, $n \geq 1$, $\varepsilon > 0$. Then $\omega \in \liminf_{n\to\infty} A_n^c$ means that ω is such that $|X_n(\omega)| \leq \varepsilon$, for all sufficiently large n, and $\omega \in \limsup_{n\to\infty} A_n$ means that ω is such that there exist arbitrarily large values of n such that $|X_n(\omega)| > \varepsilon$. In particular, every ω for which $X_n(\omega) \to 0$ as $n \to \infty$ must be such that, for every $\varepsilon > 0$, only finitely many of the real numbers $X_n(\omega)$ exceed ε in absolute value. Hence,

$$P(\{\omega : \lim_{n\to\infty} X_n(\omega) = 0\}) = 1 \quad \Longleftrightarrow \quad P(|X_n| > \varepsilon \text{ i.o.}) = 0 \text{ for all } \varepsilon > 0.$$

If convergence holds as in the left-hand side we recall from Definition 4.4 that the conclusion may be rephrased as

$$X_n \stackrel{a.s.}{\to} 0 \text{ as } n \to \infty \quad \Longleftrightarrow \quad P(|X_n| > \varepsilon \text{ i.o.}) = 0 \text{ for all } \varepsilon > 0. \qquad \square$$

Summarizing our findings so far, we have seen that the first Borel-Cantelli lemma tells us that if $\sum_{n=1}^{\infty} P(|X_n| > \varepsilon) < \infty$, then $X_n \stackrel{a.s.}{\to} 0$ as $n \to \infty$, and the second Borel-Cantelli lemma tells us that the converse holds if, in addition, $X_1, X_2, \ldots$ are independent random variables. In the latter case we obtain the following zero-one law, which we state for easy reference.

Corollary 18.1. *Suppose that $X_1, X_2, \ldots$ are independent random variables. Then*

$$X_n \stackrel{a.s.}{\to} 0 \text{ as } n \to \infty \quad \Longleftrightarrow \quad \sum_{n=1}^{\infty} P(|X_n| > \varepsilon) < \infty \text{ for all } \varepsilon > 0.$$

Remark 18.1. Convergence is a tail event, since convergence or not is independent of $X_1, X_2, \ldots, X_n$ for any n. The zero-one law therefore is also a consequence of the Kolmogorov zero-one law, Theorem 1.5.1. However, the present, alternative, derivation is more elementary and direct. $\square$

A common method in probability theory is to begin by considering subsequences. A typical case in the present context is when one wishes to prove that $P(A_n \text{ i.o.}) = 1$ and the events are not independent, but a suitable subsequence consists of independent events. In such cases the following rather immediate result may be helpful.

Theorem 18.4. *Let $\{A_n, n \geq 1\}$ be arbitrary events. If $\{A_{n_k}, k \geq 1\}$ are independent events for some subsequence $\{n_k, k \geq 1\}$, and*

$$\sum_{k=1}^{\infty} P(A_{n_k}) = \infty,$$

then $P(A_n \text{ i.o.}) = 1$.

Proof. This is immediate from the fact that $\{A_n \text{ i.o.}\} \supset \{A_{n_k} \text{ i.o.}\}$, and the second Borel-Cantelli lemma:

$$P(A_n \text{ i.o.}) \geq P(A_{n_k} \text{ i.o.}) = 1 \qquad \square$$

18.2 Some (Very) Elementary Examples

We first present a simple coin-tossing example, which is then expanded via a monkey and a typewriter to the more serious problem of the so-called Bible code, where *serious* is not to be interpreted mathematically, but as an example of the dangerous impact of what is believed to be paranormal phenomena on society. In a following subsection we provide examples related to records and random walks.

Coin Tossing

Toss a *fair* coin repeatedly (independent tosses) and let

$$A_n = \{\text{the } n\text{th toss yields a head}\}, \quad n \geq 1.$$

Then

$$P(A_n \text{ i.o.}) = 1.$$

To prove this we note that $\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} \frac{1}{2} = \infty$, and the conclusion follows from Theorem 18.2.

For an arbitrary coin, one could imagine that if the probability of obtaining heads is "very small," then it might happen that, with some "very small" probability, only finitely many heads appear. However, set $P(\text{heads}) = p$, where $0 < p < 1$. Then $\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} p = \infty$, and we conclude, once again, that $P(A_n \text{ i.o.}) = 1$.

Finally, suppose that the tosses are performed with different coins, let A_n be defined as before, and set $p_n = P(A_n)$. Then

$$P(A_n \text{ i.o.}) = 1 \quad \Longleftrightarrow \quad \sum_{n=1}^{\infty} p_n = +\infty.$$

The following exercises can be solved similarly, but a little more care is required, since the corresponding events are no longer independent.

Exercise 18.1. Toss a coin repeatedly as before and let

$$A_n = \{\text{the } (n-1)\text{th and the } n\text{th toss both yield a head}\}, \quad n \geq 2.$$

Show that

$$P(A_n \text{ i.o.}) = 1.$$

In other words, the event "two heads in a row" will occur infinitely often with probability 1. (Remember Theorem 18.4.)

Exercise 18.2. Toss another coin. Show that any finite pattern occurs infinitely often with probability 1.

Exercise 18.3. Toss a fair die with one face for every letter from A to Z repeatedly. Show that any finite word will appear infinitely often with probability 1. □

The Monkey and the Typewriter

A classical, more humorous, example states that if one puts a monkey at a typewriter he (or she) will "some day" all of a sudden have produced the complete works of Shakespeare, and, in fact, repeat this endeavor infinitely many times. In between successes the monkey will also complete the Uppsala telephone directory and lots of other texts.

Let us prove that this is indeed the case. Suppose that the letters the monkey produces constitute an independent sequence of identically distributed random variables. Then, by what we have just shown for coins, and extended in the exercises, every finite sequence of letters will occur (infinitely often!) with probability 1. And since the complete works of Shakespeare (as well as the Uppsala telephone directory) are exactly that, a finite sequence of letters, the proof is complete – under these model assumptions, which, of course, can be debated. After all, it is not quite obvious that the letters the monkey will produce are independent of each other ....

Finally, by the same argument it follows that the same texts also will appear if we spell out only every second letter or every 25th letter or every 37,658th letter.

The Bible Code

Paranormal or supernatural phenomena and superstition have always been an important ingredient in the lives of many persons. Unfortunately a lot of people are fooled and conned by this kind of mumbo-jumbo or by others who exploit their fellow human beings.

In 1997 there appeared a book, *The Bible Code* [67], which to a large extent is based on the paper [254]. In the book it is claimed that the Hebrew Bible contains a code that reveals events that will occur thousands of years later. The idea is that one writes the 304,805 letters in an array, after which one reads along lines backward or forward, up or down, and looks for a given word. It is also permitted to follow every nth letter for any n. By doing so one finds all sorts of future events. One example is that by checking every 4772nd letter one finds the name of Yitzhak Rabin, which shows that one could already in the Bible find a hint concerning his murder in November 1995. An additional comment is that it is claimed that only the Hebrew version contains the code, no translation of it.

Although the "problem" is not exactly the same as the problem with the monkey and the typewriter, the probabilistic parallel is that one faces a (random) very long list of letters, among which one looks for a given word. Here we do not have an infinite sequence, but, on the other hand, we do not require a given word to appear infinitely often either.

If we look for a word of, say, k letters in an alphabet of, say, N letters, the probability of this word appearing at any given spot is $p = 1/N^k$, under the assumption that letters occur independently of each other and with the same distribution at every site. Barring all model discussions, starting at letters $m(k+1)$, for $m = 1, 2, \ldots$ (in order to make occurrences independent of each other), the number of repetitions before a hit is geometric with mean $1/p = N^k$, which is a finite number.

With the Borel-Cantelli lemmas in our mind it is thus not surprising that one can find almost anything one wishes with this program. More about the book can be found in the article [244], where, among other things, results

from the same search method applied to translations of the bible as well as to other books are reported.

Apart from all of this one might wonder: If G-d really has put a code into the Bible, wouldn't one expect a more sophisticated one? And if the code really is a code, why did nobody discover the WTC attack on September 11, 2001, ahead of time? And the subway bombing in Madrid 2-1/2 years later?

Admittedly these examples may seem a bit elementary. On the other hand, they illustrate to what extent such examples are abundant in our daily lives; one may wonder how many fewer copies of books of this kind would be sold if everybody knew the Borel-Cantelli lemmas ....

18.3 Records

Recall the setting from Subsection 2.17.2: $X_1, X_2, \ldots$ are independent, identically distributed, continuous random variables; the record times are

$$L(n) = \min\{k : X_k > X_{L(n-1)}\}, \quad n \geq 2, \quad L(1) = 1;$$

and the associated counting variables are

$$\mu(n) = \#\,\text{records among } X_1, X_2, \ldots, X_n = \sum_{k=1}^{n} I_k, \quad n \geq 1,$$

where $P(I_k = 1) = P(X_k \text{ is a record}) = 1 - P(I_k = 0) = 1/k$, and the indicators are independent.

Our concern for now is the "intuitively obvious(?)" fact that, one should obtain infinitely many records if we continue sampling indefinitely, the reason being that there is always room for a larger value than the largest one so far. But, intuition is not enough; we require a proof.

Mathematically we thus wish to prove that

$$P(I_n = 1 \text{ i.o.}) = 1.$$

Now,

$$\sum_{n=1}^{\infty} P(I_n = 1) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty,$$

so that, the second Borel-Cantelli lemma tells us that our intuition was, indeed, a good one. Note that independence was important.

Let us also consider the number of *double records*, that is, two records in a row. What about our intuition? Is it equally obvious that there will be infinitely many double records? If there are infinitely many records, why not infinitely many times two of them following immediately after each other?

Let $D_n = 1$ if X_n produces a double record, that is, if X_{n-1} and X_n both are records. Let $D_n = 0$ otherwise. Then, for $n \geq 2$,

$$P(D_n = 1) = P(I_n = 1, I_{n-1} = 1) = P(I_n = 1) \cdot P(I_{n-1} = 1) = \frac{1}{n} \cdot \frac{1}{n-1},$$

because of the independence of the indicators. Alternatively, by symmetry and combinatorics, $D_n = 1$ precisely when X_n is the largest and X_{n-1} is the second largest among the first n observations. Thus,

$$\sum_{n=2}^{\infty} P(D_n = 1) = \sum_{n=2}^{\infty} \frac{1}{n(n-1)} = \lim_{m\to\infty} \sum_{n=2}^{m} \Big(\frac{1}{n-1} - \frac{1}{n}\Big) = \lim_{m\to\infty} (1 - \frac{1}{m}) = 1,$$

so that by the first Borel-Cantelli lemma

$$P(D_n = 1 \text{ i.o.}) = 0,$$

that is, the probability of infinitely many double records is 0. Note that $\{D_n,\, n \geq 2\}$ are *not* independent, which, however, is no problem since the sum was convergent.

The expected number of double records equals

$$E \sum_{n=2}^{\infty} D_n = \sum_{n=2}^{\infty} E\, D_n = \sum_{n=2}^{\infty} P(D_n = 1) = 1,$$

in other words, we can expect *one* double record.

Moreover, since double records seem to be rare events, one might guess that the total number of double records, $\sum_{n=2}^{\infty} D_n$, has a Poisson distribution, and if so, with parameter 1. That this is a correct guess has been proved independently in [125] and [42], Theorem 1.

18.4 Recurrence and Transience of Simple Random Walks

Consider a simple random walk, $\{S_n,\, n \geq 1\}$, starting at 0, and the probabilities that

- the random walk eventually returns to 0;
- doing so infinitely often.

A return can only occur after an even number of steps, equally many to the left and the right. It follows that

$$P(S_{2n} = 0) = \binom{2n}{n} p^n q^n \sim \begin{cases} \frac{1}{\sqrt{\pi n}} (4pq)^n, & \text{for} \quad p \neq q, \\ \frac{1}{\sqrt{\pi n}}, & \text{for} \quad p = q, \end{cases}$$

so that

$$\sum_{n=1}^{\infty} P(S_n = 0) \begin{cases} < +\infty, & \text{for} \quad p \neq q, \\ = +\infty, & \text{for} \quad p = q. \end{cases}$$

The first Borel-Cantelli lemma therefore tells us that

$$P(S_n = 0 \text{ i.o.}) = 0 \quad \text{for} \quad p \neq q.$$

One can, in fact, show that the probability of returning eventually equals $\min\{p, q\}/\max\{p, q\}$, when $p \neq q$. In this case the random walk is called *transient*.

The case $p = q = 1/2$ is called the *recurrent* case, since the probability of eventually returning to 0 equals 1. However, this is *not* a consequence of the second Borel-Cantelli lemma, since the events $\{S_n = 0\}$ are not independent. So, in order to prove this we must use different arguments.

Thus, suppose that $p = q = 1/2$, let x be the probability we seek, namely, that a random walk starting at 0 eventually returns to 0, and let y be the probability that a random walk starting at 0 eventually reaches the point $+1$. By symmetry, y also equals the probability that a random walk starting at 0 eventually reaches the point -1, and by translation invariance, y also equals the probability of eventually being one step to the left (or right) of the current state. Conditioning on the first step we obtain, with the aid of these properties,

$$\begin{aligned} x &= \frac{1}{2}y + \frac{1}{2}y, \\ y &= \frac{1}{2} + \frac{1}{2}y^2, \end{aligned}$$

which has the solution $x = y = 1$.

We have thus shown that the probability of eventually returning to 0 equals 1. Now, having returned once, the probability of returning again equals 1, and so on, so that the probability of returning infinitely often equals 1, as claimed.

Remark 18.2. Note that the hard part is to show that the random walk returns *once*; that it returns infinitely often follows as an immediate consequence! □

Exercise 18.4. If $p \neq q$ an analogous argument also requires $z =$ the probability that a random walk starting at 0 eventually reaches the point -1. In the symmetric case $y = z$, but not here. Find the analogous system of (three) equations. □

Remark 18.3. A natural extension would be to consider the two-dimensional variant, in which the random walk is performed in the plane in such a way that transitions occur with probability $1/4$ in each of the four directions. The answer is that the probability of eventually returning to 0 equals 1 also in this case. So, what about three dimensions? Well, in this case even the symmetric random walk is transient. This is true for any dimension $d \geq 3$. The mathematical reason is that

$$\sum_{n=1}^{\infty} \left(\frac{1}{\sqrt{n}}\right)^d \begin{cases} = +\infty, & \text{for} \quad d = 1, 2, \\ < +\infty, & \text{for} \quad d \geq 3. \end{cases}$$

Note that for $d \geq 3$ transience is a consequence of this and the first Borel-Cantelli lemma. □

18.5 $\sum_{n=1}^{\infty} P(A_n) = \infty$ and $P(A_n \text{ i.o.}) = 0$

In the previous example with $p = q$ we found that $P(A_n \text{ i.o.}) = 1$, but not because the Borel-Cantelli sum was divergent; the events were not independent, so we had to use a different argument. In the following example the Borel-Cantelli sum diverges too, but in this case the conclusion is that $P(A_n \text{ i.o.}) = 0$. In other words, anything can happen for dependent events.

We ask the reader to trust the following claim, and be patient until Chapter 6 where everything will be verified.

Example 18.2. Let X, X_1, X_2, ... be a sequence of independent, identically distributed random variables and set $S_n = X_1 + X_2 + \cdots + X_n$, $n \geq 1$.

The two facts we shall prove in Chapter 6 are that

$$P\Big(\big|\frac{S_n}{n} - \mu\big| > \varepsilon \text{ i.o.}\Big) = 0 \text{ for all } \varepsilon > 0 \quad \Longleftrightarrow \quad E|X| < \infty \text{ and } E\,X = \mu,$$

and that

$$\sum_{n=1}^{\infty} P\Big(\big|\frac{S_n}{n} - \mu\big| > \varepsilon\Big) < \infty \text{ for all } \varepsilon > 0 \quad \Longleftrightarrow \quad E\,X = \mu \text{ and } \operatorname{Var} X < \infty.$$

This means that if the mean is finite, but the variance is infinite, then the Borel-Cantelli sum diverges, and, yet, $P(|\frac{S_n}{n} - \mu| > \varepsilon \text{ i.o.}) = 0$. □

The remainder of this section deals with how to handle cases without (total) independence.

18.6 Pair-wise Independence

We know from Subsection 2.10.3 that independence is a more restrictive assumption than *pair-wise* independence. However, if sums of random variables are involved it frequently suffices to assume pair-wise independence; for example, because the variance of a sum is equal to the sum of the variances.

Our first generalization is, basically, a consequence of that fact.

Theorem 18.5. *Let $\{A_n,\, n \geq 1\}$ be pair-wise independent events. Then*

$$\sum_{n=1}^{\infty} P(A_n) = \infty \quad \Longrightarrow \quad P(A_n \text{ i.o.}) = 1.$$

Proof. It is convenient to introduce indicator random variables. Let

$$I_n = I\{A_n\}, \quad n \geq 1.$$

Then $E\,I_n = P(A_n)$, $\operatorname{Var} I_n = P(A_n)(1 - P(A_n))$, the pair-wise independence translates into

$$E(I_i I_j) = E\,I_i \cdot E\,I_j \quad \text{for} \quad i \neq j,$$

and the statement of the theorem into

$$\sum_{n=1}^{\infty} E\, I_n = \infty \quad \Longrightarrow \quad P\Big(\sum_{n=1}^{\infty} I_n = \infty \Big) = 1.$$

Now, by Chebyshev's inequality,

$$\begin{aligned} P\Big(\Big| \sum_{k=1}^{n} (I_k - E\, I_k) \Big| > \frac{1}{2} \sum_{k=1}^{n} P(A_k) \Big) &\le \frac{\operatorname{Var}\left(\sum_{k=1}^{n} I_k \right)}{(\frac{1}{2} \sum_{k=1}^{n} P(A_k))^2} \\ &= \frac{4 \sum_{k=1}^{n} P(A_k)(1 - P(A_k))}{\left(\sum_{k=1}^{n} P(A_k) \right)^2} \le \frac{4}{\sum_{k=1}^{n} P(A_k)} \to 0 \quad \text{as} \quad n \to \infty. \end{aligned}$$

Recalling that $E\, I_k = P(A_k)$, it follows, in particular, that

$$P\Big(\sum_{k=1}^{n} I_k > \frac{1}{2} \sum_{k=1}^{n} E\, I_k \Big) \to 1 \quad \text{as} \quad n \to \infty.$$

Since both sums increase with n we may let n tend to infinity in $\sum_{k=1}^{n} I_k$ and then in $\sum_{k=1}^{n} E\, I_k$, to conclude that

$$P\Big(\sum_{n=1}^{\infty} I_n = \infty \Big) = 1. \qquad \square$$

An immediate consequence is that the zero-one law, Theorem 18.3, remains true for pair-wise independent random variables. For convenience we state this fact as a theorem of its own.

Theorem 18.6. (A second zero-one law)
If $\{A_n,\, n \ge 1\}$ *are* pair-wise *independent events, then*

$$P(A_n \text{ i.o.}) = \begin{cases} 0, & \textit{when} \quad \sum_{n=1}^{\infty} P(A_n) < \infty, \\ 1, & \textit{when} \quad \sum_{n=1}^{\infty} P(A_n) = \infty. \end{cases}$$

18.7 Generalizations Without Independence

The following result is due to Barndorff-Nielsen, [10].

Theorem 18.7. *Let* $\{A_n,\, n \ge 1\}$ *be arbitrary events satisfying*

$$P(A_n) \to 0 \quad \textit{as} \quad n \to \infty, \tag{18.1}$$

and

$$\sum_{n=1}^{\infty} P(A_n \cap A_{n+1}^c) < \infty. \tag{18.2}$$

Then

$$P(A_n \text{ i.o.}) = 0.$$

Remark 18.4. Note that $\sum_{n=1}^{\infty} P(A_n)$ may be convergent as well as divergent under the present assumptions. In particular, the convergence of the sum is not *necessary* in order for $P(A_n \text{ i.o.})$ to equal 0. □

Proof. A glance at Theorem 18.1 shows that the second assumption alone implies that $P(A_n \cap A_{n+1}^c \text{ i.o.}) = 0$, that is, that there are almost surely only a finite number of switches between the sequences $\{A_n\}$ and $\{A_n^c\}$, so that one of them occurs only a finite number of times, after which the other one takes over for ever. To prove the theorem it therefore suffices to prove that

$$P(A_n^c \text{ i.o.}) = 1.$$

Now,

$$P(A_n^c \text{ i.o.}) = \lim_{m\to\infty} P\Big(\bigcup_{n\ge m} A_n^c\Big) \ge \lim_{m\to\infty} P(A_m^c) \to 1 \quad \text{as} \quad m \to \infty,$$

where the convergence to 1 follows from the first assumption. □

Continuing the discussion at the beginning of the proof we note that if $\{A_n,\, n \ge 1\}$ are independent events, we may, in addition, conclude that one of $\{A_n \text{ i.o.}\}$ and $\{A_n^c \text{ i.o.}\}$ has probability 1 and the other one has probability 0, since by the zero-one law in Theorem 18.3, the probabilities of these events can only assume the values 0 or 1. For ease of future reference we collect these facts separately. Note also that the conclusions are true whether (18.1) holds or not.

Theorem 18.8. *Let $\{A_n,\, n \ge 1\}$ be arbitrary events, and suppose (18.2) holds.*
(i) *Then*

$$P(A_n \cap A_{n+1}^c \text{ i.o.}) = 0.$$

(ii) *If, in addition, $\{A_n,\, n \ge 1\}$ are independent, then*

$$P(A_n \text{ i.o.}) = 0 \quad \textit{and} \quad P(A_n^c \text{ i.o.}) = 1 \quad \textit{or vice versa.}$$

To exploit the crossing concept further we formulate the following result.

Theorem 18.9. *Let $\{A_n,\, n \ge 1\}$ and $\{B_n,\, n \ge 1\}$ be arbitrary events, and suppose that the pairs A_n and B_{n+1} are independent for all n. If*

$$\sum_{n=1}^{\infty} P(A_n \cap B_{n+1}) < \infty,$$

then

$$P(A_n \text{ i.o.}) = 0 \quad \textit{and} \quad P(B_n \text{ i.o.}) = 1 \quad \textit{or vice versa.}$$

Proof. The arguments for Theorem 18.8 were given prior to its statement, and those for Theorem 18.9 are the same. □

In his paper Barndorff-Nielsen applied this result in order to prove a theorem on the rate of growth of partial maxima of independent, identically distributed random variables. In order to illustrate the efficiency of his result we apply the idea to the partial maxima of standard exponentials. The computations are based on a more general result in [124].

18.8 Extremes

Suppose that $X_1, X_2, \ldots$ are independent, standard exponential random variables, and set $Y_n = \max\{X_1, X_2, \ldots, X_n\}$, $n \geq 1$.

We begin by considering the original sequence, after which we turn our attention to the sequence of partial maxima.

Since

$$P(X_n > \varepsilon \log n) = \frac{1}{n^\varepsilon},$$

it follows that

$$\sum_{n=1}^{\infty} P(X_n > \varepsilon \log n) \begin{cases} < +\infty & \text{for} \quad \varepsilon > 1, \\ = +\infty & \text{for} \quad \varepsilon \leq 1. \end{cases}$$

An appeal to the Borel-Cantelli lemmas asserts that

$$P(\{X_n > \varepsilon \log n\} \text{ i.o.}) = \begin{cases} 0 & \text{for} \quad \varepsilon > 1, \\ 1 & \text{for} \quad \varepsilon \leq 1, \end{cases} \tag{18.3}$$

and, consequently, that

$$\limsup_{n\to\infty} \frac{X_n}{\log n} = 1 \quad \text{a.s.}$$

Moreover, since

$$\sum_{n=1}^{\infty} P(X_n < \varepsilon \log n) = \sum_{n=1}^{\infty} \left(1 - \frac{1}{n^\varepsilon}\right) = +\infty \quad \text{for all} \quad \varepsilon > 0,$$

the second Borel-Cantelli lemma yields

$$\liminf_{n\to\infty} \frac{X_n}{\log n} = 0 \quad \text{a.s.}$$

This means, roughly speaking, that the sequence $\{X_n/\log n,\, n \geq 1\}$ oscillates between 0 and 1.

Since $Y_n = \max\{X_1, X_2, \ldots, X_n\}$ is non-decreasing in n (and, hence cannot oscillate) it is tempting to guess that

$$\lim_{n\to\infty} \frac{Y_n}{\log n} = 1 \quad \text{a.s.}$$

This is not only a guess as we shall show next.

The crucial observation is that

$$\{Y_n > \varepsilon \log n \text{ i.o.}\} \quad \Longleftrightarrow \quad \{X_n > \varepsilon \log n \text{ i.o.}\},$$

since $\log n$ is increasing in n; although Y_n exceeds $\varepsilon \log n$ more often that X_n, the whole sequences do so *infinitely often* simultaneously. It follows that

$$P(Y_n > \varepsilon \log n \text{ i.o.}) = 1 \quad \text{for} \quad \varepsilon < 1,$$

and that

$$P\Big(\limsup_{n\to\infty} \frac{Y_n}{\log n} = 1 \Big) = 1. \tag{18.4}$$

In order to show that the limit actually equals 1 we have a problem, since Y_n, $n \geq 1$, are not independent, and this is where Theorem 18.9 comes to our rescue.

Let $0 < \varepsilon < 1$, and set

$$A_n = \{Y_n \leq \varepsilon \log n\} \quad \text{and} \quad B_n = \{X_n > \varepsilon \log n\}, \quad n \geq 1.$$

Then

$$\begin{aligned}
\sum_{n=1}^{\infty} P(A_n \cap A_{n+1}^c) &= \sum_{n=1}^{\infty} P(A_n \cap B_{n+1}) = \sum_{n=1}^{\infty} P(A_n) \cdot P(B_{n+1}) \\
&= \sum_{n=1}^{\infty} \Big(1 - \frac{1}{n^{\varepsilon}}\Big)^n \cdot \frac{1}{(n+1)^{\varepsilon}} \leq \sum_{n=1}^{\infty} \exp\{-n^{1-\varepsilon}\} \cdot \frac{1}{n^{\varepsilon}} \\
&= \sum_{n=1}^{\infty} \int_{n-1}^{n} \exp\{-n^{1-\varepsilon}\} \cdot \frac{1}{n^{\varepsilon}} \, \mathrm{d}x \leq \sum_{n=1}^{\infty} \int_{n-1}^{n} \exp\{-x^{1-\varepsilon}\} \cdot \frac{1}{x^{\varepsilon}} \, \mathrm{d}x \\
&= \int_0^{\infty} \exp\{-x^{1-\varepsilon}\} \cdot \frac{1}{x^{\varepsilon}} \, \mathrm{d}x = \Big[\frac{-\exp\{-x^{1-\varepsilon}\}}{1-\varepsilon} \Big]_0^{\infty} = \frac{1}{1-\varepsilon} < \infty.
\end{aligned}$$

Since $P(B_n \text{ i.o.}) = 1$ by (18.3), Theorem 18.9 tells us that we must have

$$P(A_n \text{ i.o.}) = 0 \quad \text{for} \quad \varepsilon < 1,$$

which implies that

$$P\Big(\liminf_{n\to\infty} \frac{Y_n}{\log n} \geq 1 \Big) = 1. \tag{18.5}$$

Joining this with (18.4) establishes that

$$P\Big(\lim_{n\to\infty} \frac{Y_n}{\log n} = 1 \Big) = 1,$$

or, equivalently, that $\frac{Y_n}{\log n} \stackrel{a.s.}{\to} 1$ as $n \to \infty$, as desired.

If, instead, the random variables have a standard normal distribution, Mill's ratio, Lemma A.2.1, yields

$$P(X > x) \sim \frac{1}{x\sqrt{2\pi}} \exp\{-x^2/2\} \quad \text{as} \quad x \to \infty,$$

so that, for N large,

$$\sum_{n \geq N} P(X_n > \varepsilon\sqrt{2\log n}) \sim \sum_{n \geq N} \frac{1}{\varepsilon\sqrt{2\pi \log n}} \cdot \frac{1}{n^{\varepsilon^2}} \begin{cases} < +\infty & \text{for} \quad \varepsilon > 1, \\ = +\infty & \text{for} \quad \varepsilon \leq 1, \end{cases}$$

from which it similarly follows that

$$P(\{X_n > \varepsilon\sqrt{2\log n}\} \text{ i.o.}) = \begin{cases} 0 & \text{for} \quad \varepsilon > 1, \\ 1 & \text{for} \quad \varepsilon \leq 1, \end{cases}$$

and that

$$\limsup_{n\to\infty} \frac{X_n}{\sqrt{2\log n}} = 1 \quad \text{a.s.}$$

Since the standard normal distribution is symmetric around 0, it follows, by considering the sequence $\{-X_n,\, n \geq 1\}$, that

$$\liminf_{n\to\infty} \frac{X_n}{\sqrt{2\log n}} = -1 \quad \text{a.s.}$$

Exercise 18.5. Prove the analog for partial maxima of independent standard normal random variables. □

18.9 Further Generalizations

For notational convenience we set, throughout the remainder of this section,

$$p_k = P(A_k) \quad \text{and} \quad p_{ij} = P(A_i \cap A_j), \quad \text{for all} \quad k, i, j,$$

in particular, $p_{kk} = p_k$.

Inspecting the proof of Theorem 18.5, we find that the variance of the sum of the indicator becomes

$$\begin{aligned} \operatorname{Var}\left(\sum_{k=1}^n I_k\right) &= \sum_{k=1}^n p_k(1-p_k) + \sum_{\substack{i=1 \\ i\neq j}}^n \sum_{j=1}^n (p_{ij} - p_i p_j) \\ &= \sum_{i=1}^n \sum_{j=1}^n p_{ij} - \sum_{i=1}^n \sum_{j=1}^n p_i p_j = \sum_{i=1}^n \sum_{j=1}^n p_{ij} - \left(\sum_{k=1}^n p_k\right)^2, \end{aligned}$$

so that, in this case, the computation turns into

$$P\Big(\Big|\sum_{k=1}^{n}(I_k - E\,I_k)\Big| > \frac{1}{2}\sum_{k=1}^{n} p_k\Big) \le \frac{\mathrm{Var}\left(\sum_{k=1}^{n} I_k\right)}{(\frac{1}{2}\sum_{k=1}^{n} p_k)^2} = 4\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}}{\left(\sum_{k=1}^{n} p_k\right)^2} - 1\right),$$

which suggests the following strengthening.

Theorem 18.10. *Let $\{A_n,\, n \ge 1\}$ be arbitrary events, such that*

$$\liminf_{n\to\infty} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} P(A_i \cap A_j)}{\left(\sum_{k=1}^{n} P(A_k)\right)^2} = 1$$

Then

$$\sum_{n=1}^{\infty} P(A_n) = \infty \quad \Longrightarrow \quad P(A_n \text{ i.o.}) = 1.$$

Proof. By arguing as in the proof of Theorem 18.5, it follows from the computations preceding the statement of Theorem 18.10 that

$$\liminf_{n\to\infty} P\Big(\sum_{k=1}^{n} I_k \le \frac{1}{2}\sum_{k=1}^{n} E\,I_k\Big) = 0.$$

We may therefore select a subsequence $\{n_j,\, j \ge 1\}$ of the integers in such a way that

$$\sum_{j=1}^{\infty} P\Big(\sum_{k=1}^{n_j} I_k \le \frac{1}{2}\sum_{k=1}^{n_j} E\,I_k\Big) < \infty,$$

which, by the first Borel-Cantelli lemma, shows that

$$P\Big(\sum_{k=1}^{n_j} I_k \le \frac{1}{2}\sum_{k=1}^{n_j} E\,I_k \text{ i.o.}\Big) = 0,$$

so that,

$$P\Big(\sum_{k=1}^{n_j} I_k > \frac{1}{2}\sum_{k=1}^{n_j} E\,I_k \text{ i.o.}\Big) = 1.$$

Finally, since this is true for any j and the sum of the expectations diverges, we may, as in the proof of Theorem 18.5, let j tend to infinity in the sum of the indicators, and then in the sum of the expectations to conclude that

$$P\Big(\sum_{k=1}^{\infty} I_k = \infty\Big) = 1. \qquad \square$$

With some additional work one can prove the following, stronger, result.

Theorem 18.11. *Let $\{A_n,\, n \geq 1\}$ be arbitrary events, such that*

$$\limsup_{n\to\infty} \frac{\big(\sum_{k=m+1}^{n} P(A_k)\big)^2}{\sum_{i=m+1}^{n}\sum_{j=m+1}^{n} P(A_i \cap A_j)} \geq \alpha,$$

for some $\alpha > 0$ and m large. Then

$$\sum_{n=1}^{\infty} P(A_n) = \infty \quad \Longrightarrow \quad P(A_n \text{ i.o.}) \geq \alpha.$$

An early related paper is [49], from which we borrow the following lemma, which, in turn, is instrumental for the proof of the theorem. We also refer to [234], P3, p. 317, where the result is used in connection with a three-dimensional random walk, and to [195], Section 6.1 where also necessary and sufficient conditions for ensuring that $P(A_n \text{ i.o.}) = \alpha$ are given.

Lemma 18.1. *Let $\{A_n,\, n \geq 1\}$ be arbitrary events. For $m \geq 1$,*

$$P\Big(\bigcup_{k=m+1}^{n} A_k\Big) \geq \frac{\big(\sum_{k=m+1}^{n} P(A_k)\big)^2}{\sum_{i=m+1}^{n}\sum_{j=m+1}^{n} P(A_i \cap A_j)}.$$

Proof. Set $I_n = I\{A_n\}$, $n \geq 1$. Then

$$\begin{aligned} E\Big(\sum_{k=m+1}^{n} I_k\Big)^2 &= \sum_{i,j=m+1}^{n} E\, I_i I_j = \sum_{k=m+1}^{n} E\, I_k + \sum_{\substack{i,j=m+1 \\ i\neq j}}^{n} E\, I_i I_j \\ &= \sum_{k=m+1}^{n} p_k + \sum_{\substack{i,j=m+1 \\ i\neq j}}^{n} p_{ij} = \sum_{i,j=m+1}^{n} p_{ij}. \end{aligned}$$

Secondly, via Cauchy's inequality,

$$\begin{aligned} &\Big(\sum_{k=m+1}^{n} p_k\Big)^2 = \Big(E\sum_{k=m+1}^{n} I_k\Big)^2 = \Big(E\sum_{k=m+1}^{n} I_k \cdot I\{\sum_{k=m+1}^{n} I_k > 0\}\Big)^2 \\ &\leq E\Big(\sum_{k=m+1}^{n} I_k\Big)^2 \cdot E\Big(I\Big\{\sum_{k=m+1}^{n} I_k > 0\Big\}\Big)^2 \\ &= E\Big(\sum_{k=m+1}^{n} I_k\Big)^2 E\Big(I\Big\{\sum_{k=m+1}^{n} I_k > 0\Big\}\Big) = E\left(\sum_{k=m+1}^{n} I_k\right)^2 P\Big(\bigcup_{k=m+1}^{n} A_k\Big). \end{aligned}$$

The conclusion follows by joining the extreme members from the two calculations. □

Proof of Theorem 18.11. Choosing m sufficiently large, and applying the lemma, we obtain

$$P(A_n \text{ i.o.}) = \lim_{m\to\infty} P\Big(\bigcup_{k=m+1}^{\infty} A_k\Big) \geq \lim_{m\to\infty} \limsup_{n\to\infty} P\Big(\bigcup_{k=m+1}^{n} A_k\Big)$$
$$\geq \lim_{m\to\infty} \limsup_{n\to\infty} \frac{\big(\sum_{k=m+1}^{n} p_k\big)^2}{\sum_{i=m+1}^{n}\sum_{j=m+1}^{n} p_{ij}} \geq \alpha. \qquad \square$$

Our final extension is a recent result in which the assumption about the ratio of the sums is replaced by the same condition applied to the individual terms; see [196]. For a further generalization we refer to [197].

Theorem 18.12. *Let $\{A_n,\, n \geq 1\}$ be arbitrary events, such that, for some $\alpha \geq 1$,*

$$P(A_i \cap A_j) \leq \alpha P(A_i)P(A_j) \quad \textit{for all} \quad i,j > m,\ i \neq j. \tag{18.6}$$

Then

$$\sum_{n=1}^{\infty} P(A_n) = \infty \quad \Longrightarrow \quad P(A_n \text{ i.o.}) \geq 1/\alpha.$$

Proof. We first consider the denominator in right-hand side of Lemma 18.1. Using the factorizing assumption and the fact that $\alpha \geq 1$, we obtain

$$\sum_{i,j=m+1}^{n} p_{ij} = \sum_{k=m+1}^{n} p_k + \sum_{\substack{i,j=m+1\\ i\neq j}}^{n} p_{ij} \leq \sum_{k=m+1}^{n} p_k + \alpha \sum_{\substack{i,j=m+1\\ i\neq j}}^{n} p_{ij}$$
$$\leq \sum_{k=m+1}^{n} p_k - \alpha \sum_{k=m+1}^{n} p_k^2 + \alpha \left(\sum_{k=m+1}^{n} p_k\right)^2$$
$$\leq \alpha \sum_{k=m+1}^{n} p_k \Big(1 + \sum_{k=m+1}^{n} p_k\Big),$$

so that, by the lemma,

$$P\Big(\bigcup_{k=m+1}^{n} A_k\Big) \geq \frac{\big(\sum_{k=m+1}^{n} p_k\big)^2}{\sum_{i,j=m+1}^{n} p_{ij}} \geq \frac{\sum_{k=m+1}^{n} p_k}{\alpha\Big(1 + \sum_{k=m+1}^{n} p_k\Big)}.$$

The divergence of the sum, finally, yields

$$P(A_n \text{ i.o.}) = \lim_{m\to\infty} P\Big(\bigcup_{k=m+1}^{\infty} A_k\Big) \geq \lim_{m\to\infty} \limsup_{n\to\infty} P\Big(\bigcup_{k=m+1}^{n} A_k\Big)$$
$$\geq \lim_{m\to\infty} \limsup_{n\to\infty} \frac{\sum_{k=m+1}^{n} p_k}{\alpha\Big(1 + \sum_{k=m+1}^{n} p_k\Big)} = 1/\alpha. \qquad \square$$

We close by connecting Theorem 18.12 to some of the earlier results.

- Lamperti [167] proves that $P(A_n \text{ i.o.}) > 0$ under the assumption (18.6).
- If there exists some kind of zero-one law, the conclusion of the theorem (and of that of Lamperti) becomes $P(A_n \text{ i.o.}) = 1$; cf. Theorem 18.3.
- If (18.6) holds with $\alpha = 1$, then $P(A_n \text{ i.o.}) = 1$. One such case is when $\{A_n, n \geq 1\}$ are (pair-wise) independent, in which case we rediscover Theorems 18.2 and 18.5, respectively.

19 A Convolution Table

Let X and Y be independent random variables and set $Z = X+Y$. What type of distribution does Z have if X and Y are absolutely continuous, discrete, or continuous singular, respectively? If both are discrete one would guess that so is Z. But what if X has a density and Y is continuous singular? What if X is continuous singular and Y is discrete?

The convolution formula for densities, cf. Subsection 2.9.4, immediately tells us that if both distributions are absolutely continuous, then so is the sum. However, by inspecting the more general result, Theorem 9.4, we realize that it suffices for one of them to be absolutely continuous.

If both distributions are discrete, then so is the sum; the support of the sum is the direct sum of the respective supports:

$$\text{supp}\,(F_{X+Y}) = \{x + y : x \in \text{supp}\,(F_X) \text{ and } y \in \text{supp}\,(F_Y)\}.$$

If X is discrete and Y is continuous singular, the support of the sum is a Lebesgue null set, so that the distribution is singular. The derivative of the distribution function remains 0 almost everywhere, the new exceptional points are contained in the support of X, which, by Proposition 2.1(iii), is at most countable.

Our findings so far may be collected in the following diagram:

X \ Y	AC	D	CS
AC	AC	AC	AC
D	AC	D	CS
CS	AC	CS	??

Figure 2.6. The distribution of $X + Y$

It remains to investigate the case when X and Y both are continuous singular. However, this is a more sophisticated one. We shall return to that slot in Subsection 4.2.2 with the aid of (rescaled versions of) the Cantor type random variables Y and Z from Subsection 2.2.6.

20 Problems

1. Let X and Y be random variables and suppose that $A \in \mathcal{F}$. Prove that
$$Z = XI\{A\} + YI\{A^c\} \quad \text{is a random variable.}$$
2. Show that if X is a random variable, then, for every $\varepsilon > 0$, there exists a bounded random variable X_ε, such that
$$P(X \neq X_\varepsilon) < \varepsilon.$$
♣ Observe the difference between a *finite* random variable and a *bounded* random variable.
3. Show that
(a) if X is a random variable, then so is $|X|$;
(b) the converse does not necessarily hold.
♠ Don't forget that there exist non-measurable sets.
4. Let X be a random variable with distribution function F.
(a) Show that
$$\lim_{h \searrow 0} P(x-h < X \le x+h) = \lim_{h \searrow 0} \big(F(x+h) - F(x-h)\big) = \begin{cases} P(X=x), & \text{if} \quad x \in \mathbb{J}_F, \\ 0, & \text{otherwise.} \end{cases}$$
(b) A point $x \in \operatorname{supp}(F)$ if and only if
$$F(x+h) - F(x-h) > 0 \quad \text{for every} \quad h > 0.$$
Prove that
$$x \in \mathbb{J}_F \quad \Longrightarrow \quad x \in \operatorname{supp}(F).$$
(c) Prove that the converse holds for isolated points.
(d) Prove that the support of any distribution function is closed.
5. Suppose that X is an integer valued random variable, and let $m \in \mathbb{N}$. Show that
$$\sum_{n=1}^{\infty} P(n < X \le n+m) = m.$$
6. Show that, for any random variable, X, and $a \in \mathbb{R}$,
$$\int_{-\infty}^{\infty} P(x < X \le x+a)\,\mathrm{d}x = a.$$
♣ An extension to two random variables will be given in Problem 20.15.

7. Let $(\Omega, \mathcal{F}, P)$ be the Lebesgue measure space $([0,1], \mathcal{B}([0,1]), \lambda)$, and let $\{X_t,\, 0 \le t \le 1\}$ be a *family* of random variables defined as

$$X_t(\omega) = \begin{cases} 1, & \text{for} \quad \omega = t, \\ 0, & \text{otherwise.} \end{cases}$$

Show that

$$P(X_t = 0) = 1 \text{ for all } t \qquad \text{and/but} \qquad P(\sup_{0 \le t \le 1} X_t = 1) = 1.$$

♣ Note that there is no contradiction, since the supremum is taken over an uncountable set of t values.

8. Show that, if $\{X_n,\, n \ge 1\}$ are independent random variables, then

$$\sup_n X_n < \infty \quad \text{a.s} \qquad \Longleftrightarrow \qquad \sum_{n=1}^{\infty} P(X_n > A) < \infty \quad \text{for some } A.$$

9. The name of the log-normal distribution comes from the fact that its logarithm is a normal random variable. Prove that the name is adequate, that is, let $X \in N(\mu, \sigma^2)$, and set $Y = \mathrm{e}^X$. Compute the distribution function of Y, differentiate, and compare with the entry in Table 2.2.
10. Let $X \in U(0,1)$, and $\theta > 0$. Verify, by direct computation, that

$$Y = -\theta \log X \in \mathrm{Exp}(\theta).$$

♠ This is useful for generating exponential random numbers, which are needed in simulations related to the Poisson process.

11. Compute the expected number of trials needed in order for all faces of a symmetric die to have appeared at least once.
12. *The coupon collector's problem.* Each time one buys a bag of cheese doodles one obtains as a bonus a picture (hidden inside the package) of a soccer player. Suppose there are n different pictures which are equally likely to be inside every package. Find the expected number of packages one has to buy in order to get a complete collection of players.

♣ For $n = 100$ the numerical answer is 519, i.e., "a lot" more than 100.

13. Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables with $E\, X^4 < \infty$, and set $\mu = E\, X$, $\sigma^2 = \mathrm{Var}\, X$, and $\mu_4 = E(X - \mu)^4$. Furthermore, set

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^{n} X_k \quad \text{and} \quad m_n^2 = \frac{1}{n} \sum_{k=1}^{n} (X_k - \bar{X}_n)^2.$$

Prove that

$$E(m_n^2) = \sigma^2 \frac{n-1}{n},$$
$$\mathrm{Var}\,(m_n^2) = \frac{\mu_4 - \sigma^4}{n} - \frac{2\mu_4 - 4\sigma^4}{n^2} + \frac{\mu_4 - 3\sigma^4}{n^3}.$$

♣ Observe that the *sample variance*, $s_n^2 = \frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X}_n)^2 = \frac{n-1}{n} m_n^2$, is *unbiased*, which means that $E\, s_n^2 = \sigma^2$. On the other hand, the expression for $\operatorname{Var} s_n^2$ is more involved.

14. Let, for $k \geq 1$, $\mu_k = E(X - E\,X)^k$ be the kth central moment of the random variable X. Prove that the matrix

$$\begin{pmatrix} 1 & 0 & \mu_2 \\ 0 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{pmatrix}$$

(a) has a non-negative determinant;
(b) is non-negative definite.

♠ Assume w.l.o.g. that $E\,X = 0$ and investigate $E(a_0 + a_1 X + a_2 X^2)^2$, where $a_0, a_1, a_2 \in \mathbb{R}$.

(c) Generalize to higher dimensions.

15. This problem extends Problem 20.6. Let X, Y be random variables with finite mean. Show that

$$\int_{-\infty}^{\infty} \big(P(X < x \leq Y) - P(Y < x \leq X)\big)\mathrm{d}x = E\,Y - E\,X.$$

16. Show that, if X and Y are independent random variables, such that $E|X| < \infty$, and B is an arbitrary Borel set, then

$$E\,X I\{Y \in B\} = E\,X \cdot P(Y \in B).$$

17. Suppose that $X_1, X_2, \ldots, X_n$ are random variables, such that $E|X_k| < \infty$ for all k, and set $Y_n = \max_{1\leq k\leq n} X_k$.
(a) Prove that $E\,Y_n < \infty$.
(b) Prove that $E\,X_k \leq E\,Y_n$ for all k.
(c) Prove that $E|Y_n| < \infty$.
(d) Show that the analog of (b) for absolute values (i.e. $E|X_k| \leq E|Y_n|$ for all k) need *not* be true.

♣ Note the distinction between the random variables $|\max_{1\leq k\leq n} X_k|$ and $\max_{1\leq k\leq n} |X_k|$.

18. Let $X_1, X_2, \ldots$ be random variables, and set $Y = \sup_n |X_n|$. Show that

$$E|Y|^r < \infty \quad \Longleftrightarrow \quad |Y| \leq Z \quad \text{for some} \quad Z \in L^r, \quad r > 0.$$

19. Let X be a non-negative random variable. Show that

$$\lim_{n\to\infty} nE\Big(\frac{1}{X} I\{X > n\}\Big) = 0\,,$$
$$\lim_{n\to\infty} \frac{1}{n} E\Big(\frac{1}{X} I\{X > \frac{1}{n}\}\Big) = 0\,.$$

♠ A little more care is necessary for the second statement.

20. Let $\{A_n,\, n \geq 1\}$ be independent events, and suppose that $P(A_n) < 1$ for all n. Prove that

$$P(A_n \text{ i.o.}) = 1 \quad \Longleftrightarrow \quad P\Big(\bigcup_{n=1}^{\infty} A_n\Big) = 1.$$

Why is $P(A_n) = 1$ forbidden?

21. Consider the dyadic expansion of $X \in U(0,1)$, and let l_n be the *run length of zeroes* from the nth decimal and onward. This means that $l_n = k$ if decimals $n, n+1, \ldots, n+k-1$ are all zeroes. In particular, $l_n = 0$ if the nth decimal equals 1.

(a) Prove that $P(l_n = k) = \frac{1}{2^{k+1}}$ for all $k \geq 0$.

(b) Prove that $P(l_n = k \text{ i.o.}) = 1$ for all k.

♠ Note that the events $\{l_n = k,\, n \geq 1\}$ are *not* independent (unless $k = 0$).

♣ The result in (b) means that, with probability 1, there will be infinitely many arbitrarily long stretches of zeroes in the decimal expansion of X.

(c) Prove that $P(l_n = n \text{ i.o.}) = 0$.

♣ This means that if we require the run of zeroes that starts at n to have length n, then, almost surely, this will happen only finitely many times. (There exist stronger statements.)

22. Let $X,\, X_1,\, X_2,\, \ldots$ be independent, identically distributed random variables, such that $P(X = 0) = P(X = 1) = 1/2$.

(a) Let N_1 be the number of 0's and 1's until the first appearance of the pattern 10. Find $E\, N_1$.

(b) Let N_2 be the number of 0's and 1's until the first appearance of the pattern 11. Find $E\, N_2$.

(c) Let N_3 be the number of 0's and 1's until the first appearance of the pattern 100. Find $E\, N_3$.

(d) Let N_4 be the number of 0's and 1's until the first appearance of the pattern 101. Find $E\, N_4$.

(e) Let N_5 be the number of 0's and 1's until the first appearance of the pattern 111. Find $E\, N_5$.

(f) Solve the same problem if $X \in \text{Be}(p)$, for $0 < p < 1$.

♣ No two answers are the same (as one might think concerning (a) and (b)).

http://www.springer.com/978-0-387-22833-4

Probability: A Graduate Course
Gut, A.
2005, XXIV, 608 p., Hardcover
ISBN: 978-0-387-22833-4