MATH/STAT 235A — Probability Theory
Lecture Notes, Fall 2013

Dan Romik
Department of Mathematics, UC Davis

December 30, 2013

Contents

Chapter 1: Introduction 6

1.1 What is probability theory? 6

1.2 The algebra of events 6

Chapter 2: Probability spaces 11

2.1 Basic definitions 11

2.2 Properties and examples 12

Chapter 3: Random variables 17

3.1 Random variables and their distributions 17

3.2 Distribution functions 18

3.3 Examples 21

Chapter 4: Random vectors and independence 24

4.1 Random vectors 24

4.2 Multi-dimensional distribution functions 25

4.3 Independence 27

Chapter 5: The Borel-Cantelli lemmas 30

Chapter 6: A brief excursion into measure theory 32

Chapter 7: Expected values 34

7.1 Construction of the expectation operator 34

7.2 Properties 36

7.3 Convergence theorems for expectations 38

7.4 Computing expected values 40

7.5 Expectation and independent random variables 41

7.6 Moments 43

Chapter 8: Special distributions and their properties 45

8.1 The Bernoulli distribution 45

8.2 The binomial distribution 45

8.3 The geometric distribution 46
8.4 The negative binomial distribution 47
8.5 The Poisson distribution 47
8.6 The uniform distribution 48
8.7 The normal distribution 48
8.8 The exponential distribution 50
8.9 The gamma distribution 51
8.10 The beta distribution 52
8.11 The Cauchy distribution 53

Chapter 9: Laws of large numbers 55
9.1 Weak laws of large numbers 55
9.2 Strong laws of large numbers 56

Chapter 10: Applications and further examples 63
10.1 The Weierstrass approximation theorem 63
10.2 Infinite expectations and triangular arrays 64
10.3 Random series of independent samples 68

Chapter 11: The Central Limit Theorem, Stirling's formula and the de Moivre-Laplace theorem 72

Chapter 12: Convergence in distribution 76
12.1 Definition 76
12.2 Examples 77
12.3 Compactness and tightness 78

Chapter 13: Characteristic functions 82
13.1 Definition and basic properties 82
13.2 Examples 83
13.3 The inversion formula 85
13.4 The continuity theorem 88
13.5 Moments 89

Chapter 14: Central limit theorems **91**

14.1 The case of i.i.d. r.v.'s 91

14.2 Generalizations 91

14.2.1 Triangular arrays 92

14.2.2 Erdös-Kac theorem 93

14.2.3 The Euclidean algorithm 94

Chapter 15: Random number generation **95**

15.1 Unbiasing a source of randomness: simulating an unbiased coin using biased coin tosses 95

15.2 Simulating a biased coin using unbiased coin tosses 96

15.3 Simulating an arbitrary discrete distribution using unbiased coin tosses . . . 97

15.4 Simulating a general r.v. using a uniform r.v. 98

15.5 Simulating an exponential r.v. 98

15.6 Simulating a normal r.v. 99

15.7 Simulating a Poisson r.v. 99

Chapter 16: Additional topics **101**

16.1 The Kolmogorov 0-1 law 101

16.2 The Kolmogorov three-series theorem 102

16.3 The Poisson limit law 103

Exercises **106**

Note to the reader. These notes are based to a large extent on Chapters 1–3 in the textbook *Probability: Theory and Examples, 4th Ed.* by Rick Durrett. References to Durrett's book appear throughout the text as [Dur2010]. References to the earlier 3rd edition appear as [Dur2004].

Chapter 1: Introduction

1.1 What is probability theory?

In this course we'll learn about probability theory. But what exactly *is* probability theory? Like some other mathematical fields (but unlike some others), it has a dual role:

- It is a **rigorous mathematical theory** – with definitions, lemmas, theorems, proofs etc.

- It is a **mathematical model** that purports to explain or model real-life phenomena.

We will concentrate on the rigorous mathematical aspects, but we will try not to forget the connections to the intuitive notion of real-life probability. These connections will enhance our intuition, and they make probability an extremely useful tool in all the sciences. And they make the study of probability much more *fun*, too! A note of caution is in order, though: mathematical models are only as good as the assumptions they are based on. So probability can be *used*, and it can be (and quite frequently is) *abused*...

Example 1.1. The **theory of differential equations** is another mathematical theory which has the dual role of a rigorous theory and an applied mathematical model. On the other hand, **number theory**, **complex analysis** and **algebraic topology** are examples of fields which are not normally used to model real-life phenomena.

1.2 The algebra of events

A central notion in probability is that of the **algebra of events** (we'll clarify later what the word "algebra" means here). We begin with an informal discussion. We imagine that probability is a function, denoted $\mathbf{P}$, that takes as its argument an "event" (i.e., occurrence of something in a real-life situation involving uncertainty) and returns a real number in $[0,1]$ representing how likely this event is to occur. For example, if a fair coin is tossed 10 times and we denote the results of the tosses by $X_1, X_2, \ldots, X_{10}$ (where each of X_i is 0 or 1, signifying "tails" or "heads"), then we can write statements like

$$\mathbf{P}(X_i = 0) = 1/2, \qquad (1 \le i \le 10),$$

$$\mathbf{P}\left(\sum_{i=1}^{10} X_i = 4\right) = \frac{\binom{10}{4}}{2^{10}}.$$

Note that if A and B represent events (meaning, for the purposes of the present informal discussion, objects that have a well-defined probability), then we expect that the phrases "A did not occur", "A and B both occurred" and "at least one of A and B occurred" also represent events. We can use notation borrowed from mathematical logic and denote these new events by $\neg A$, $A \wedge B$, and $A \vee B$, respectively. Thus, the set of events is not just a set – it is a set with some extra structure, namely the ability to perform **negation**, **conjunction** and **disjunction** operations on its elements. Such a set is called an **algebra** in some contexts.

But what if the coin is tossed an *infinite* number of times? In other words, we now imagine an infinite sequence $X_1, X_2, X_3, \ldots$ of (independent) coin toss results. We want to be able to ask questions such as

$$\begin{aligned}
\mathbf{P}(\text{infinitely many of the } X_i\text{'s are } 0) &= ? \\
\mathbf{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n} X_k = \frac{1}{2}\right) &= ? \\
\mathbf{P}\left(\sum_{k=1}^{\infty} \frac{2X_k - 1}{k} \text{ converges}\right) &= ?
\end{aligned}$$

Do such questions make sense? (And if they do, can you guess what the answers are?) Maybe it is not enough to have an *informal* discussion to answer this...

Example 1.2. (a) An urn initially contains a white ball and a black ball. A ball is drawn out at random from the urn, then added back and another white ball is added to the urn. This procedure is repeated infinitely many times, so that after step n the urn contains 1 black ball and $n+1$ white balls. For each $n \geq 1$, let A_n denote the event that at step n the black ball was drawn. Now let A_∞ denote the event

$$A_\infty = \text{"in total, the black ball was selected infinitely many times"},$$

(i.e., the event that infinitely many of the events A_n occurred).

(b) While this experiment takes place, an identical copy of the experiment is taking place in the next room. The random selections in the two neighboring rooms have no connection to each other, i.e., they are "independent". For each $n \ge 1$, let B_n be the event that at step n the black ball was drawn out of the "copy" experiment urn. Now let B_∞ denote the event

$$B_\infty = \text{"in total, the black ball was selected infinitely many times in the second copy of the experiment"},$$

(in other words, the event that infinitely many of the events B_n occurred).

(c) For each $n \ge 1$, let C_n be the event that both A_n and B_n occurred, i.e.

$$C_n = \text{"at step } n\text{, the black ball was selected simultaneously in both experiments"},$$

and let C_∞ denote the event "C_n occurred for infinitely many values of n".

Theorem 1.3. *We have*

$$\mathbf{P}(A_\infty) = \mathbf{P}(B_\infty) = 1, \qquad \mathbf{P}(C_\infty) = 0.$$

Proof. These claims are consequences of the **Borel-Cantelli lemmas** which we will learn about later in the course. Here is a sketch of the proof that $P(C_\infty) = 0$ (remember, this is still an "informal discussion", so our "proof" is really more of an exploration of what formal assumptions are needed to make the claim hold). For each n we have

$$\mathbf{P}(A_n) = \mathbf{P}(B_n) = \frac{1}{n+1},$$

since at time n each of the urns contains $n+1$ balls, only one of which is black. Moreover, the choices in both rooms are made independently, so we have

$$\mathbf{P}(C_n) = \mathbf{P}(A_n \wedge B_n) = \mathbf{P}(A_n)\mathbf{P}(B_n) = \frac{1}{(n+1)^2}.$$

It turns out that to prove that $\mathbf{P}(C_\infty) = 0$, the only relevant bit of information is that the infinite series $\sum_{n=1}^\infty \mathbf{P}(C_n)$ is a convergent series; the precise values of the probabilities are irrelevant. Indeed, we can try to do various manipulations on the definition of the event C, as follows:

$$\begin{aligned} C_\infty &= \text{"infinitely many of the } C_n\text{'s occurred"} \\ &= \text{"for all } N \ge 1\text{, the event } C_n \text{ occurred for some } n \ge N\text{"}. \end{aligned}$$

For any $N \geq 1$, denote the event "C_n occurred for some $n \geq N$" by D_N. Then

$$\begin{aligned} C_\infty &= \text{"for all } N \geq 1,\ D_N \text{ occurred"} \\ &= D_1 \wedge D_2 \wedge D_3 \wedge \ldots \quad \text{(infinite conjunction...?!)} \\ &= \bigwedge_{N=1}^{\infty} D_N \qquad\qquad \text{(shorthand notation for infinite conjunction).} \end{aligned}$$

In particular, in order for the event C_∞ to happen, D_N must happen for any fixed value of N (for example, D_{100} must happen, D_{101} must happen, etc.). It follows that C_∞ is at most as likely to happen as any of the D_N's; in other words we have

$$\mathbf{P}(C_\infty) \leq \mathbf{P}(D_N), \qquad (N \geq 1).$$

Now, what can we say about $\mathbf{P}(D_N)$? Looking at the definition of D_N, we see that it too can be written as an infinite *disjunction* of events, namely

$$\begin{aligned} D_N &= C_N \vee C_{N+1} \vee C_{N+2} \vee \ldots \quad \text{(infinite disjunction)} \\ &= \bigvee_{n=N}^{\infty} C_n \qquad\qquad \text{(shorthand for infinite disjunction).} \end{aligned}$$

If this were a finite disjunction, we could say that the likelihood for at least one of the events to happen is at most the sum of the likelihoods (for example, the probability that it will rain next weekend is at most the probability that it will rain next Saturday, *plus* the probability that it will rain next Sunday; of course it might rain on both days, so the sum of the probabilities can be strictly greater than the probability of the disjunction). What can we say for an *infinite* disjunction? Since this is an informal discussion, it is impossible to answer this without being more formal about the precise mathematical model and its assumptions. As it turns out, the correct thing to do (in the sense that it leads to the most interesting and natural mathematical theory) is to *assume* that this fact that holds for finite disjunctions also holds for infinite ones. Whether this has any relevance to real life is a different question! If we make this assumption, we get for each $N \geq 1$ the bound

$$\mathbf{P}(D_N) \leq \sum_{n=N}^{\infty} \mathbf{P}(C_n).$$

But now recall that the infinite series of probabilities $\sum_{n=1}^{\infty} \mathbf{P}(C_n)$ converges. Therefore, for any $\epsilon > 0$, we can find an N for which the tail $\sum_{n=N}^{\infty} \mathbf{P}(C_n)$ of the series is less than ϵ. For such an N, we get that $\mathbf{P}(D_N) < \epsilon$, and therefore that $\mathbf{P}(C_\infty) < \epsilon$. This is true for any $\epsilon > 0$, so it follows that $\mathbf{P}(C_\infty) = 0$. □

Chapter 2: Probability spaces

2.1 Basic definitions

We now formalize the concepts introduced in the previous lecture. It turns out that it's easiest to deal with events as subsets of a large set called the **probability space**, instead of as abstract logical statements. The logical operations of negation, conjunction and disjunction are replaced by the set-theoretic operations of taking the complement, intersection or union, but the intuitive meaning attached to those operations is the same.

Definition 2.1 (Algebra)**.** *If Ω is a set, an **algebra** of subsets of Ω is a collection $\mathcal{F}$ of subsets of Ω that satisfies the following axioms:*

$$\emptyset \in \mathcal{F}, \tag{A1}$$

$$A \in \mathcal{F} \implies \Omega \setminus A \in \mathcal{F}, \tag{A2}$$

$$A, B \in \mathcal{F} \implies A \cup B \in \mathcal{F}. \tag{A3}$$

*A word synonymous with algebra in this context is **field**.*

Definition 2.2 (σ-algebra)**.** *A σ**-algebra** (also called a σ**-field**) is an algebra $\mathcal{F}$ that satisfies the additional axiom*

$$A_1, A_2, A_3, \ldots \in \mathcal{F} \implies \cup_{n=1}^{\infty} A_n \in \mathcal{F}. \tag{A4}$$

Example 2.3. If Ω is any set, then $\{\emptyset, \Omega\}$ is a σ-algebra – in fact it is the smallest possible σ-algebra of subsets of Ω. Similarly, the power set $\mathcal{P}(\Omega)$ of all subsets of Ω is a σ-algebra, and is (obviously) the largest σ-algebra of subsets of Ω.

Definition 2.4 (Measurable space)**.** *A **measurable space** is a pair $(\Omega, \mathcal{F})$ where Ω is a set and $\mathcal{F}$ is a σ-algebra of subsets of Ω.*

Definition 2.5 (Probability measure)**.** *Given a measurable space $(\Omega, \mathcal{F})$, a **probability measure** on $(\Omega, \mathcal{F})$ is a function $\mathbf{P} : \mathcal{F} \to [0,1]$ that satisfies the properties:*

$$\mathbf{P}(\emptyset) = 0, \quad \mathbf{P}(\Omega) = 1, \tag{P1}$$

$$A_1, A_2, \ldots \in \mathcal{F} \textit{ are pairwise disjoint} \implies \mathbf{P}(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbf{P}(A_n). \tag{P2}$$

Definition 2.6 (Probability space). *A **probability space** is a triple* $(\Omega, \mathcal{F}, \mathbf{P})$*, where* $(\Omega, \mathcal{F})$ *is a measurable space, and* $\mathbf{P}$ *is a probability measure on* $(\Omega, \mathcal{F})$*.*

Intuitively, we think of Ω as representing the set of possible **outcomes** of a probabilistic experiment, and refer to it as the **sample space**. The σ-algebra $\mathcal{F}$ is the **σ-algebra of events**, namely those subsets of Ω which have a well-defined probability (as we shall see later, it is not always possible to assign well-defined probabilities to *all* sets of outcomes). And $\mathbf{P}$ is the "notion" or "measure" of probability on our sample space.

Probability theory can be described loosely as the study of probability spaces (this is of course a gross oversimplification...). A more general mathematical theory called **measure theory** studies **measure spaces**, which are like probability spaces except that the measures can take values in $[0, \infty]$ instead of $[0, 1]$, and the total measure of the space is not necessarily equal to 1 (such measures are referred to as **σ-additive nonnegative measures**). Measure theory is an important and non-trivial theory, and studying it requires a separate concentrated effort. We shall content ourselves with citing and using some of its most basic results. For proofs and more details, refer to Chapter 1 and the measure theory appendix in [Dur2010] or to a measure theory textbook.

2.2 Properties and examples

Lemma 2.7. *If* $(\Omega, \mathcal{F}, \mathbf{P})$ *is a probability space, then we have:*

(i) ***Monotonicity:*** *If* $A, B \in \mathcal{F}$, $A \subset B$ *then* $\mathbf{P}(A) \le \mathbf{P}(B)$.

(ii) ***Sub-additivity:*** *If* $A_1, A_2, \ldots \in \mathcal{F}$ *then* $\mathbf{P}(\cup_{n=1}^{\infty} A_n) \le \sum_{n=1}^{\infty} \mathbf{P}(A_n)$.

(iii) ***Continuity from below:*** *If* $A_1, A_2, \ldots \in \mathcal{F}$ *such that* $A_1 \subset A_2 \subset A_3 \subset \ldots$*, then*

$$\mathbf{P}(\cup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \mathbf{P}(A_n).$$

(iv) ***Continuity from above:*** *If* $A_1, A_2, \ldots \in \mathcal{F}$ *such that* $A_1 \supset A_2 \supset A_3 \supset \ldots$*, then*

$$\mathbf{P}(\cap_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \mathbf{P}(A_n).$$

Exercise 2.8. *Prove Lemma 2.7.*

Example 2.9. Discrete probability spaces. Let Ω be a countable set and let $p : \Omega \to [0,1]$ be a function such that

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

This corresponds to the intuitive notion of a probabilistic experiment with a finite or countably infinite number of outcomes, where each individual outcome ω has a probability $p(\omega)$ of occurring. We can put such an "elementary" or "discrete" experiment in our more general framework by defining the σ-algebra of events $\mathcal{F}$ to be the set of subsets of Ω, and defining the probability measure $\mathbf{P}$ by

$$\mathbf{P}(A) = \sum_{\omega \in A} p(\omega), \qquad A \in \mathcal{F}.$$

If Ω is a finite set, a natural probability measure to consider is the **uniform measure**, defined by

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|}.$$

Example 2.10. Choosing a random number uniformly in $(0,1)$. The archetypical example of a "non-elementary" probability space (i.e., one which does not fall within the scope of the previous example) is the experiment of choosing a random number uniformly in the interval $(0,1)$. How do we know that it makes sense to speak of such an experiment? We don't, *yet*. But let us imagine what constructing such an experiment might entail. We are looking for a hypothetical probability space $(\Omega, \mathcal{F}, \mathbf{P})$, in which the sample space Ω is simply $(0,1)$, $\mathcal{F}$ is some σ-algebra of subsets of $(0,1)$, and $\mathbf{P}$ is a probability measure that corresponds to our notion of a "uniform" choice of a random number. One plausible way to formalize this is to require that intervals of the form $(a,b) \subset (0,1)$ be considered as events, and that the probability for our "uniform" number to fall in such an interval should be equal to its length $b-a$. In other words, we shall require that

$$(a,b) \in \mathcal{F}, \qquad (0 \le a < b \le 1),$$

and that

$$\mathbf{P}\big((a,b)\big) = b - a, \qquad (0 \le a < b \le 1). \tag{1}$$

How do we generate a σ-algebra of subsets of $(0,1)$ that contains all the intervals? We already saw that the set of *all* subsets of $(0,1)$ will work. But that is too large! If we take all subsets, we will see in an exercise later that it will be impossible to construct the probability measure $\mathbf{P}$ to satisfy our requirements. So let's try to build the *smallest possible* σ-algebra. One way (which can perhaps be described as the **bottom-up** approach) would be to start with the intervals, then take all countable unions of such and add them to our collection of sets, then add all countable intersections of such sets, then add all countable unions, etc. Will this work? In principle it can be made to work, but is a bit difficult and requires knowing something about **transfinite induction**. Fortunately there is a more elegant way (but somewhat more abstract and less intuitive) of constructing the minimal σ-algebra, that is outlined in the next exercise below, and can be thought of as the **top-down** approach. The resulting σ-algebra of subsets of $(0,1)$ is called the **Borel σ-algebra**; its elements are called **Borel sets**.

What about the probability measure $\mathbf{P}$? Here we will simply cite a result from measure theory that says that the measure we are looking for exists, and is unique. This is not too difficult to prove, but doing so would take us a bit too far off course.

Theorem 2.11.. *Let $\mathcal{B}$ be the σ-algebra of Borel sets on $(0,1)$, the minimal σ-algebra containing all the sub-intervals of $(0,1)$, proved to exist in the exercise below. There exists a unique measure $\mathbf{P}$ on the measure space satisfying (1), called* ***Lebesgue measure*** *on $(0,1)$.*

Exercise 2.12 (The σ-algebra generated by a set of subsets of Ω)**.** *(i) Let Ω be a set, and let $\{\mathcal{F}_i\}_{i\in I}$ be some collection of σ-algebras of subsets of Ω, indexed by some index set I. Prove that the intersection of all the $\mathcal{F}_i$'s (i.e., the collection of subsets of Ω that are elements of all the $\mathcal{F}_i$'s) is also a σ-algebra.*
(ii) Let Ω be a set, and let $\mathcal{A}$ be a collection of subsets of Ω. Prove that there exists a unique σ-algebra $\sigma(\mathcal{A})$ of subsets of Ω that satisfies the following two properties:

1. *$\mathcal{A} \subset \sigma(\mathcal{A})$ (in words, $\sigma(\mathcal{A})$ contains all the elements of $\mathcal{A}$).*

2. *$\sigma(\mathcal{A})$ is the minimal σ-algebra satisfying property 1 above, in the sense that if $\mathcal{F}$ is any other σ-algebra that contains all the elements of $\mathcal{A}$, then $\sigma(\mathcal{A}) \subset \mathcal{F}$.*

Hint for (ii): Let $(\mathcal{F}_i)_{i\in I}$ be the collection of all σ-algebras of subsets of Ω that contain $\mathcal{A}$. This is a non-empty collection, since it contains for example $\mathcal{P}(\Omega)$, the set of all subsets of Ω. Any σ-algebra $\sigma(\mathcal{A})$ that satisfies the two properties above is necessarily a subset of any of the $\mathcal{F}_i$'s, hence it is also contained in the intersection of all the $\mathcal{F}_i$'s, which is a σ-algebra by part (i) of the exercise.

Definition 2.13. *If $\mathcal{A}$ is a collection of subsets of a set Ω, the σ-algebra $\sigma(\mathcal{A})$ discussed above is called **the σ-algebra generated by $\mathcal{A}$**.*

Example 2.14. The space of infinite coin toss sequences. Another archetypical experiment in probability theory is that of a sequence of independent fair coin tosses, so let's try to model this experiment with a suitable probability space. If for convenience we represent the result of each coin as a binary value of 0 or 1, then the sample space Ω is simply the set of infinite sequences of 0's and 1's, namely

$$\Omega = \big\{(x_1, x_2, x_3, \ldots) : x_i \in \{0,1\}, i = 1, 2, \ldots\big\} = \{0,1\}^{\mathbb{N}}.$$

What about the σ-algebra $\mathcal{F}$? We will take the same approach as we did in the previous example, which is to require certain natural sets to be events, and to take as our σ-algebra the σ-algebra generated by these "elementary" events. In this case, surely, for each $n \ge 1$, we would like the set

$$A_n(1) := \{\mathbf{x} = (x_1, x_2, \ldots) \in \Omega : x_n = 1\} \tag{2}$$

to be an event (in words, this represents the event "the coin toss x_n came out Heads"). Therefore we take $\mathcal{F}$ to be the σ-algebra generated by the collection of sets of this form.

Finally, the probability measure $\mathbf{P}$ should conform to our notion of a sequence of independent fair coin tosses. Generalizing the notation in (2), for $a \in \{0,1\}$ define

$$A_n(a) = \{\mathbf{x} = (x_1, x_2, \ldots) \in \Omega : x_n = a\}.$$

Then $\mathbf{P}$ should satisfy

$$\mathbf{P}(A_n(a)) = \frac{1}{2},$$

representing the fact that the n-th coin toss is unbiased. But more generally, for any $n \ge 1$ and $(a_1, a_2, \ldots, a_n) \in \{0,1\}^n$, since the first n coin tosses are independent, $\mathbf{P}$ should satisfy

$$\mathbf{P}\big(A_1(a_1) \cap A_2(a_2) \cap \ldots \cap A_n(a_n)\big) = \frac{1}{2^n}. \tag{3}$$

As in the example of Lebesgue measure discussed above, the fact that a probability measure $\mathbf{P}$ on $(\Omega, \mathcal{F})$ that satisfies (3) exists and is unique follows from some slightly non-trivial facts from measure theory, and we will take it on faith for the time being. Below we quote the relevant theorem from measure theory, which generalizes the setting discussed in this example to the more general situation of a **product** of probability spaces.

Theorem 2.15 (Products of probability spaces)**.** *Let* $\big((\Omega_n, \mathcal{F}_n, \mathbf{P}_n)\big)_{n=1}^{\infty}$ *be a sequence of probability spaces. Denote* $\Omega = \prod_{n=1}^{\infty} \Omega_n$ *(the cartesian product of the outcome sets), and let* $\mathcal{F}$ *be the* σ*-algebra of subsets of* Ω *generated by sets which are of the form*

$$\{(x_1, x_2, \ldots) \in \Omega : x_n \in A\}$$

for some $n \geq 1$ *and set* $A \in \mathcal{F}_n$*. Then there exists a unique probability measure* $\mathbf{P}$ *on* $(\Omega, \mathcal{F})$ *such that for any* $n \geq 1$ *and any finite sequence*

$$(A_1, A_2, \ldots, A_n) \in \mathcal{F}_1 \times \mathcal{F}_2 \times \ldots \times \mathcal{F}_n$$

the equation

$$\mathbf{P}\Big(\big\{(x_1, x_2, \ldots) \in \Omega : x_1 \in A_1, x_2 \in A_2, \ldots, x_n \in A_n\big\}\Big) = \prod_{k=1}^{n} \mathbf{P}_k(A_k)$$

holds.

Exercise 2.16. *Explain why the "infinite sequence of coin tosses" experiment is a special case of a product of probability spaces, and why the existence and uniqueness of a probability measure satisfying* (3) *follows from Theorem 2.15.*

In an upcoming homework exercise we will show an alternative way of proving the existence of the probability space of infinite coin toss sequences using Lebesgue measure on $(0, 1)$.

Chapter 3: Random variables

3.1 Random variables and their distributions

As we have seen, a probability space is an abstract concept that represents our intuitive notion of a probabilistic experiment. Such an experiment however can be very long (even infinite) and contain a lot of information. To make things more manageable, we consider numerical-valued functions on probability spaces, which we call **random variables**. However, not any function will do: a random variable has to relate in a nice way to the measurable space structure, so that we will be able to ask questions like "what is the probability that this random variable takes a value less than 8", etc. This leads us to the following definitions.

Definition 3.1. *If $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ are two measurable spaces, a function $X : \Omega_1 \to \Omega_2$ is called **measurable** if for any set $E \in \mathcal{F}_2$, the set*

$$X^{-1}(E) = \{\omega \in \Omega_1 : X(\omega) \in E\}$$

is in $\mathcal{F}_1$.

Definition 3.2. *If $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space, a real-valued function $X : \Omega \to \mathbb{R}$ is called a **random variable** if it is a measurable function when considered as a function from the measurable space $(\Omega, \mathcal{F})$ to the measurable space $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}$, namely the σ-algebra generated by the intervals.*

Exercise 3.3. *Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be two measurable spaces such that $\mathcal{F}_2$ is the σ-algebra generated by a collection $\mathcal{A}$ of subsets of Ω_2. Prove that a function $X : \Omega_1 \to \Omega_2$ is measurable if and only if $X^{-1}(A) \in \mathcal{F}_1$ for all $A \in \mathcal{A}$.*

It follows that the random variables are exactly those real-valued functions on Ω for which the question

"What is the probability that $a < X < b$?"

has a well-defined answer for all $a < b$. This observation makes it easier in practice to check if a given function is a random variable or not, since working with intervals is much easier than with the rather unwieldy (and mysterious, until you get used to them) Borel sets.

What can we say about the behavior of a random variable X defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$? All the information is contained in a new probability measure μ_X on the measurable space $(\mathbb{R}, \mathcal{B})$ that is induced by X, defined by

$$\mu_X(A) = \mathbf{P}(X^{-1}(A)) = \mathbf{P}(\omega \in \Omega : X(\omega) \in A).$$

The number $\mu_X(A)$ is the probability that X "falls in A" (or "takes its value in A").

Exercise 3.4. *Verify that μ_X is a probability measure on $(\mathbb{R}, \mathcal{B})$. This measure is called the **distribution** of X, or sometimes referred to more fancifully as the **law** of X. In some textbooks it is denoted $\mathcal{L}_X$.*

Definition 3.5. *If X and Y are two random variables (possibly defined on different probability spaces), we say that X and Y are **identically distributed** (or **equal in distribution**) if $\mu_X = \mu_Y$ (meaning that $\mu_X(A) = \mu_Y(A)$ for any Borel set $A \subset \mathbb{R}$). We denote this*

$$X \overset{d}{=} Y.$$

How can we check if two random variables are identically distributed? Once again, working with Borel sets can be difficult, but since the Borel sets are generated by the intervals, a simpler criterion involving just this generating family of sets exists. The following lemma is a consequence of basic facts in measure theory, which can be found in the Measure Theory appendix in [Dur2010].

Lemma 3.6. *Two probability measures μ_1, μ_2 on the measurable space $(\mathbb{R}, \mathcal{B})$ are equal if only if they are equal on the generating set of intervals, namely if*

$$\mu_1\big((a, b)\big) = \mu_2\big((a, b)\big)$$

for all $-\infty < a < b < \infty$.

3.2 Distribution functions

Instead of working with distributions of random variables (which are probability measure on the measurable space $(\mathbb{R}, \mathcal{B})$ and themselves quite unwieldy objects), we will encode them in a simpler object called a **distribution function** (sometimes referred to as a **cumulative distribution function**, or **c.d.f.**).

Definition 3.7. *The* ***cumulative distribution function*** *(or* ***c.d.f.****, or just* ***distribution function****) of a random variable X defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ is the function $F_X : \mathbb{R} \to [0, 1]$ defined by*

$$F_X(x) = \mathbf{P}(X \le x) = \mathbf{P}(X^{-1}((-\infty, x])) = \mathbf{P}(\{\omega \in \Omega : X(\omega) \le x\}), \quad (x \in \mathbb{R}).$$

Note that we have introduced here a useful notational device that will be used again many times in the following sections: if A is a Borel set, we will often write $\{X \in A\}$ as shorthand for the set $\{\omega \in \Omega : X(\omega) \in A\}$. In words, we may refer to this as "the event that X falls in A". When discussing its probability, we may omit the curly braces and simply write $P(X \in A)$. Of course, one should always remember that on the formal level this is just the set-theoretic inverse image of a set by a function!

Theorem 3.8 (Properties of distribution functions)**.** *If $F = F_X$ is a distribution function, then it has the following properties:*

(i) *F is nondecreasing.*

(ii) $\lim_{x\to\infty} F(x) = 1, \quad \lim_{x\to-\infty} F(x) = 0.$

(iii) *F is right-continuous, i.e., $F(x+) := \lim_{y\downarrow x} F(y) = F(x)$.*

(iv) $F(x-) := \lim_{y\uparrow x} F(y) = \mathbf{P}(X < x)$.

(v) $\mathbf{P}(X = x) = F(x) - F(x-)$.

(vi) *If $G = F_Y$ is another distribution function of a random variable Y, then X and Y are equal in distribution if and only if $F \equiv G$.*

Proof. Exercise (recommended), or see page 9 of [Dur2010]. □

Definition 3.9. *A function $F : \mathbb{R} \to [0, 1]$ satisfying properties (i)–(iii) in the previous theorem is called a* ***cumulative distribution function****, or just* ***distribution function****.*

Theorem 3.10. *If F is a distribution function, then there exists a random variable X such that $F = F_X$.*

This fact has a measure-theoretic proof similar to the proof of Theorem 2.11, but fortunately in this case, there is a more probabilistic proof that relies only on the existence of Lebesgue measure. (This is one of many examples of probabilistic ideas turning out to be useful to prove facts in analysis and measure theory.) This involves the probabilistic concept of a **quantile** (a generalization of the concepts of percentile and median that we frequently hear about in news reports).

Definition 3.11. *If X is a random variable on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and $0 < p < 1$ is a number, then a real number x is called a p-**quantile** of X if the inequalities*

$$\begin{aligned} \mathbf{P}(X \le x) &\ge p, \\ \mathbf{P}(X \ge x) &\ge 1 - p \end{aligned}$$

hold.

Note that the question of whether t is a p-quantile of X can be answered just by knowing the distribution function F_X of X: since $\mathbf{P}(X \le x) = F_X(x)$ and $\mathbf{P}(X \ge x) = 1 - F(x-)$, we can write the conditions above as

$$F_X(x-) \le p \le F_X(x).$$

Lemma 3.12. *A p-quantile for X always exists. Moreover, the set of p-quantiles of X is equal to the (possibly degenerate) closed interval $[a_p, b_p]$, where*

$$\begin{aligned} a_p &= \sup\{x : F_X(x) < p\}, \\ b_p &= \inf\{x : F_X(x) > p\}. \end{aligned}$$

Exercise 3.13. *Prove Lemma 3.12.*

Proof of Theorem 3.10. Let $((0,1), \mathcal{B}, \mathbf{P})$ be the unit interval with Lebesgue measure, representing the experiment of drawing a uniform random number in $(0,1)$. We shall construct our random variable X on this space. Inspired by the discussion of quantiles above, we define

$$X(p) = \sup\{y : F(y) < p\}, \qquad (0 < p < 1).$$

If F were the distribution function of a random variable, then $X(p)$ would be its (minimal) p-quantile.

Note that properties (i) and (ii) of F imply that $X(p)$ is defined and finite for any p and that it is a monotone nondecreasing function on $(0, 1)$. In particular, it is measurable, so it is in fact a random variable on the probability space $((0, 1), \mathcal{B}, \mathbf{P})$. We need to show that F is its distribution function. We will show that for each $p \in (0, 1)$ and $x \in \mathbb{R}$, we have that $X(p) \le x$ if and only if $p \le F(x)$. This will imply that for every $x \in \mathbb{R}$ we have the equality of sets

$$\{p : X(p) \le x\} = \{p : p \le F(x)\},$$

and, applying $\mathbf{P}$ to both sides of this equation we will get

$$F_X(x) = \mathbf{P}(X \le x) = \mathbf{P}\big(p : p \le F(x)\big) = \mathbf{P}\big((0, F(x)]\big) = F(x)$$

(since $\mathbf{P}$ is Lebesgue measure).

To prove the claim, note that if $p \le F(x)$ then all elements of the set $\{y : F(y) < p\}$ satisfy $y \le x$, and therefore the supremum $X(p)$ of this set also satisfies $X(p) \le x$. Conversely, if $p > F(x)$, then, since F is right-continuous, we have $p > F(x + \epsilon)$ for some $\epsilon > 0$. It follows that $X(p) \ge x + \epsilon > x$ (since $x + \epsilon$ is in the set $\{y : F(y) < p\}$). □

The function X defined in the proof above is sometimes referred to as the (lower) quantile function of the distribution F. Note that if F is a strictly increasing function then X is simply its ordinary (set-theoretic) inverse function.

3.3 Examples

Example 3.14. Indicator random variables If A is an event in a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, its **indicator random variable** is the r.v. $\mathbf{1}_A$ defined by

$$\mathbf{1}_A(\omega) = \begin{cases} 0 & \omega \notin A, \\ 1 & \omega \in A. \end{cases}$$

The above discussion shows that to specify the behavior of a random variable, it is enough to specify its distribution function. Another useful concept is that of a **density function**.

If $F = F_X$ is a distribution function such that for some nonnegative function $f : \mathbb{R} \to \mathbb{R}$ we have

$$F(x) = \int_{-\infty}^{x} f(y)\, dy, \qquad (y \in \mathbb{R}), \tag{4}$$

then we say that X **has a density function** f. Note that f determines F but is itself only determined by F up to "small" changes that do not affect the integral in (4) (in measure-theoretic terminology we say that f is determined "up to a set of measure 0"). For example, changing the value$f$ in a finite number of points results in a density function that is equally valid for computing F.

Example 3.15. Uniform random variables. We say that X **is a uniform random variable on** $(0, 1)$ if it has the distribution function

$$F(x) = \begin{cases} 0 & x \le 0, \\ x & 0 \le x \le 1, \\ 1 & x \ge 1. \end{cases}$$

Such a r.v. has as its density function the function

$$f(x) = \begin{cases} 1 & 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

More generally, if $a < b$ we say that X is a **uniform random variable in the interval** (a, b) if it has the (respective) distribution and density functions

$$F(x) = \begin{cases} 0 & x \le a, \\ \frac{x-a}{b-a} & a \le x \le b, \\ 1 & x \ge b, \end{cases} \qquad f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b, \\ 0 & \text{otherwise.} \end{cases}$$

Example 3.16. Exponential distribution.

$$F(x) = \begin{cases} 0 & x \le 0 \\ 1 - e^{-x} & x \ge 0, \end{cases} \qquad f(x) = \begin{cases} 0 & x < 0 \\ e^{-x} & x \ge 0. \end{cases}$$

Example 3.17. Standard normal distribution. The normal (or gaussian) distribution is given in terms of its density function

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

The cumulative distribution function is denoted by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy.$$

This integral cannot be evaluated explicitly in terms of more familiar functions, but Φ is an important special function of mathematics nonetheless.

Chapter 4: Random vectors and independence

4.1 Random vectors

A random variable is a real-valued measurable function defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ (when we talk of $\mathbb{R}$ as a measurable space, it will always be taken to mean with the σ-algebra of Borel sets). Similarly, we now wish to talk about *vector*-valued measurable functions on a probability space, i.e., functions taking values in $\mathbb{R}^d$. First, we need to identify a good σ-algebra of subsets of $\mathbb{R}^d$. Risking some confusion, we will still call it the Borel σ-algebra and denote it by $\mathcal{B}$, or sometimes by $\mathcal{B}(\mathbb{R}^d)$.

Definition 4.1. *The Borel σ-algebra on $\mathbb{R}^d$ is defined in one of the following equivalent ways:*

(i) It is the σ-algebra generated by boxes of the form

$$(a_1, b_1) \times (a_2, b_2) \times \ldots \times (a_d, b_d).$$

(ii) It is the σ-algebra generated by the balls in $\mathbb{R}^d$.

(iii) It is the σ-algebra generated by the open sets in $\mathbb{R}^d$.

(iv) It is the minimal σ-algebra of subsets of $\mathbb{R}^d$ such that the coordinate functions $\pi_i : \mathbb{R}^d \to \mathbb{R}$ defined by

$$\pi_i(\mathbf{x}) = x_i, \qquad i = 1, 2, \ldots, d$$

are all measurable (where measurability is respect to the Borel σ-algebra on the target space $\mathbb{R}$).

Exercise 4.2. *Check that the definitions above are indeed all equivalent.*

Definition 4.3. *A* ***random (d-dimensional) vector*** *(or* ***vector random variable****) $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ is a function $\mathbf{X} : \Omega \to \mathbb{R}^d$ that is measurable (as a function between the measurable spaces $(\Omega, \mathcal{F})$ and $(\mathbb{R}^d, \mathcal{B})$.*

Lemma 4.4. $\mathbf{X} = (X_1, \ldots, X_d)$ *is a random vector if and only if X_i is a random variable for each $i = 1, \ldots, d$.*

Proof. If $\mathbf{X}$ is a random vector then each of its coordinates $X_i = \pi_i \circ \mathbf{X}$ is a composition of two measurable functions and therefore (check!) measurable. Conversely, if $X_1, \ldots, X_d$ are random variables then for any box $E = (a_1, b_1) \times (a_2, b_2) \times \ldots \times (a_d, b_d) \subset \mathbb{R}^d$ we have

$$\mathbf{X}^{-1}(E) = \cap_{k=1}^{d} X_i^{-1}((a_i, b_i)) \in \mathcal{F}.$$

Therefore by Definition 4.1 and Exercise 3.3, $\mathbf{X}$ is a random vector. □

Exercise 4.5. *(i) Prove that any continuous function $f : \mathbb{R}^m \to \mathbb{R}^n$ is measurable (when each of the spaces is equipped with the respective Borel σ-algebra).*
(ii) Prove that the composition $g \circ f$ of measurable functions $f : (\Omega_1, \mathcal{F}_1) \to (\Omega_2, \mathcal{F}_2)$ and $g : (\Omega_2, \mathcal{F}_2) \to (\Omega_3, \mathcal{F}_3)$ (where $(\Omega_i, \mathcal{F}_i)$ are measurable spaces for $i = 1, 2, 3$) is a measurable function.
(iii) Deduce that the sum $X_1 + \ldots + X_d$ of random variables is a random variable.

Exercise 4.6. *Prove that if $X_1, X_2, \ldots$ is a sequence of random variables (all defined on the same probability space, then the functions*

$$\inf_n X_n, \quad \sup_n X_n, \quad \limsup_n X_n, \quad \liminf_n X_n$$

are all random variables. Note: Part of the question is to generalize the notion of random variable to a function taking values in $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$, or you can solve it first with the additional assumption that all the X_i's are uniformly bounded by some constant M.

4.2 Multi-dimensional distribution functions

If $\mathbf{X} = (X_1, \ldots, X_d)$ is a d-dimensional random vector, we define its **distribution** to be the probability measure

$$\mu_{\mathbf{X}}(A) = \mathbf{P}(X^{-1}(A)) = \mathbf{P}(\omega \in \Omega : X(\omega) \in A), \qquad A \in \mathcal{B}(\mathbb{R}^d),$$

similarly to the one-dimensional case. The measure μ_X is also called the **joint distribution** (or **joint law**) of the random variables $X_1, \ldots, X_d$.

Once again, to avoid having to work with measures, we introduce the concept of a d-**dimensional distribution function**.

Definition 4.7. *The d-dimensional distribution function of a d-dimensional random vector* $\mathbf{X} = (X_1, \ldots, X_d)$ *(also called the joint distribution function of* $X_1, \ldots, X_d$*) is the function* $F_{\mathbf{X}} : \mathbb{R}^d \to [0,1]$ *defined by*

$$\begin{aligned} F_{\mathbf{X}}(x_1, x_2, \ldots, x_d) &= \mathbf{P}(X_1 \le x_1, X_2 \le x_2, \ldots, X_d \le x_d) \\ &= \mu_{\mathbf{X}}\Big((-\infty, x_1] \times (-\infty, x_2] \times \ldots \times (-\infty, x_d]\Big) \end{aligned}$$

Theorem 4.8 (Properties of distribution functions)**.** *If* $F = F_{\mathbf{X}}$ *is a distribution function of a d-dimensional random vector, then it has the following properties:*

(i) F *is nondecreasing in each coordinate.*

(ii) For any $1 \le i \le d$, $\lim_{x_i \to -\infty} F(\mathbf{x}) = 0$.

(iii) $\lim_{\mathbf{x} \to (\infty, \ldots, \infty)} F(\mathbf{x}) = 1$.

(iv) F *is right-continuous, i.e.,* $F(\mathbf{x}+) := \lim_{\mathbf{y} \downarrow \mathbf{x}} F(\mathbf{y}) = F(\mathbf{x})$, *where here* $\mathbf{y} \downarrow \mathbf{x}$ *means that* $y_i \downarrow x_i$ *in each coordinate.*

(v) For $1 \le i \le d$ *and* $a < b$, *denote by* $\Delta^x_{a,b}$ *the* differencing operator in the variable x, *which takes a function* f *of the real variable* x *(and possibly also dependent on other variables) and returns the value*

$$\Delta^x_{a,b} f = f(b) - f(a)$$

Then, for any real numbers $a_1 < b_1$, $a_2 < b_2$, $\ldots$, $a_d < b_d$, *we have*

$$\Delta^{x_1}_{a_1,b_1} \Delta^{x_2}_{a_d,b_d} \ldots \Delta^{x_d}_{a_d,b_d} F \ge 0.$$

Proof. See Chapter 1 in [Dur2010] or Appendix A.2 in [Dur2004]. □

Theorem 4.9. *Any function* F *satisfying the properties in Theorem 4.8 above is a distribution function of some random vector* $\mathbf{X}$.

Proof. See Chapter 1 in [Dur2010] or Appendix A.2 in [Dur2004]. □

4.3 Independence

Definition 4.10. *Events $A, B \in \mathcal{F}$ in a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ are called independent if*

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B).$$

More generally, a family $\mathcal{A} = (A_i)_{i \in \mathcal{I}}$ of events in a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ is called an independent family if for any finite subset $A_{i_1}, A_{i_1}, \ldots, A_{i_k} \in \mathcal{A}$ of distinct events in the family we have that

$$\mathbf{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = \prod_{j=1}^{k} \mathbf{P}(A_{i_j}).$$

Definition 4.11. *Random variables X, Y on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ are called independent if*

$$\mathbf{P}(X \in E, Y \in F) = \mathbf{P}(X \in E)\mathbf{P}(Y \in F)$$

for any Borel sets $E, F \subset \mathbb{R}$. In other words any two events representing possible statements about the behaviors of X and Y are independent events.

Definition 4.12. *If Ω is a set and $X : \Omega \to \mathbb{R}$ is a function, the family of subsets of Ω defined by*

$$\sigma(X) = \left\{ X^{-1}(A) : A \in \mathcal{B}(\mathbb{R}) \right\}$$

*is a σ-algebra (check!) called the **σ-algebra generated by** X. It is easy to check that it is the minimal σ-algebra with which Ω can be equipped so as to make X into a random variable.*

Definition 4.13. *If $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space, two σ-algebras $\mathcal{A}, \mathcal{B} \subset \mathcal{F}$ are called independent if any two events $A \in \mathcal{A}, B \in \mathcal{B}$ are independent events.*

It follows from the above definitions that r.v.'s X, Y are independent if and only if the σ-algebras $\sigma(X), \sigma(Y)$ generated by them are independent σ-algebras.

Definition 4.14. *If $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space($\mathcal{F}$ and $(\mathcal{F}_i)_{i \in I}$ is some family of sub-σ-algebras of $\mathcal{F}$ (i.e., σ-algebras that are subsets of $\mathcal{F}$, we say that $(\mathcal{F}_i)_{i \in I}$ is an independent family of σ-algebras if for any $i_1, i_2, \ldots i_k \in I$ and events $A_1 \in \mathcal{F}_{i_1}, A_2 \in \mathcal{F}_{i_2}, \ldots, A_k \in \mathcal{F}_{i_k}$, the events $A_1, \ldots, A_k$ are independent.*

Definition 4.15. *A family $(X_i)_{i\in I}$ of random variables defined on some common probability space $(\Omega, \mathcal{F}, \mathbf{P})$ is called an* ***independent family of random variables*** *if the σ-algebras $\{\sigma(X_i)\}_{i\in I}$ form an independent family of σ-algebras.*

Unraveling these somewhat abstract definitions, we see that $(X_i)_{i\in I}$ is an independent family of r.v.'s if and only if we have

$$\mathbf{P}(X_{i_1} \in A_1, \dots X_{i_k} \in A_k) = \prod_{j=1}^{k} \mathbf{P}(X_{i_j} \in A_j)$$

for all indices $i_1, \dots, i_k \in I$ and Borel sets $A_1, \dots, A_k \in \mathcal{B}(\mathbb{R})$.

Theorem 4.16. *If $(\mathcal{F}_i)_{i\in I}$ are a family of sub-σ-algebras of the σ-algebra of events $\mathcal{F}$ in a probability space, and for each $i \in I$, the σ-algebra $\mathcal{F}_i$ is generated by a family $\mathcal{A}_i$ of subsets of Ω, and each family $\mathcal{A}_i$ is closed under taking the intersection of two sets (such a family is called a* π***-system****), then the family $(\mathcal{F}_i)_{i\in I}$ is independent if and only if for each $i_1, \dots, i_k \in I$, any finite sequence of events $A_1 \in \mathcal{A}_{i_1}, A_2 \in \mathcal{A}_{i_2}, \dots, A_k \in \mathcal{A}_{i_k}$ is independent.*

Proof. This uses Dynkin's $\pi - \lambda$ theorem from measure theory. See [Dur2010], Theorem 2.1.3, p. 39 or [Dur2004], Theorem (4.2), p. 24. □

As a corollary, we get a convenient criterion for checking when the coordinates of a random vector are independent.

Lemma 4.17. *If $X_1, \dots, X_d$ are random variables defined on a common probability space, then they are independent if and only if for all $x_1, \dots, x_d \in \mathbb{R}$ we have that*

$$F_{X_1,\dots,X_d}(x_1, \dots, x_d) = F_{X_1}(x_1) F_{X_2}(x_2) \dots F_{X_d}(x_d).$$

Exercise 4.18. *(i) We say that a Riemann-integrable function $f : \mathbb{R}^d \to [0, \infty)$ is a (d-dimensional, or joint) density function for a random vector $\mathbf{X} = (X_1, \dots, X_d)$ if*

$$F_X(x_1, \dots, x_d) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_d} f(u_1, \dots, u_d)\, du_d \dots du_1 \qquad \forall x_1, \dots, x_d \in \mathbb{R}.$$

Show that if f is a density for $\mathbf{X}$ and can be written in the form

$$f(x_1, \dots, x_d) = f_1(x_1) f_2(x_2) \dots f_d(x_d),$$

then $X_1, \ldots, X_d$ are independent.

(ii) Show that if $X_1, \ldots, X_d$ are random variables taking values in a countable set S, then in order for $X_1, \ldots, X_d$ to be independent it is enough that for all $x_1, \ldots, x_d \in S$ we have

$$\mathbf{P}(X_1 = x_1, \ldots, X_d = x_d) = \mathbf{P}(X_1 = x_1) \ldots \mathbf{P}(X_d = x_d).$$

Chapter 5: The Borel-Cantelli lemmas

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $A_1, A_2, A_3, \ldots \in \mathcal{F}$ be a sequence of events. We define the following events derived from the sequence $(A_n)_n$:

$$\begin{aligned} \limsup A_n &= \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n, \\ \liminf A_n &= \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n. \end{aligned}$$

What is the meaning of these events? If we think of the sequence A_n as representing a sequence of (not necessarily independent) probabilistic experiments, then we can translate the first event into words as

$$\begin{aligned} \limsup A_n &= \text{“the event that for all } N \geq 1 \text{ there is an } n \geq N \text{ such} \\ &\quad\ \text{that } A_n \text{ occurred”} \\ &= \text{“the event that infinitely many of the } A_n\text{'s occurred”}. \end{aligned}$$

For this reason, the event $\limsup A_n$ is often denoted by $\{A_n \text{ infinitely often}\}$ or $\{A_n \text{ i.o.}\}$.

The definition of the event $\liminf A_n$ can similarly be given meaning by writing

$$\begin{aligned} \liminf A_n &= \text{“the event that there exists an } N \geq 1 \text{ such that } A_n \\ &\quad\ \text{occurred for all } n \geq N\text{”} \\ &= \text{“the event that all but finitely many of the } A_n\text{'s occurred”}. \end{aligned}$$

Exercise 5.1. *Prove that for any $\omega \in \Omega$ we have*

$$\begin{aligned} \mathbf{1}_{\limsup A_n}(\omega) &= \limsup_{n\to\infty} \mathbf{1}_{A_n}(\omega), \\ \mathbf{1}_{\liminf A_n}(\omega) &= \liminf_{n\to\infty} \mathbf{1}_{A_n}(\omega). \end{aligned}$$

Theorem 5.2 (Borel-Cantelli lemmas)**.**

(i) If $\sum_{n=1}^{\infty} \mathbf{P}(A_n) < \infty$ then $\mathbf{P}(A_n \text{ i.o.}) = 0$.

(ii) If $\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \infty$ and $(A_n)_{n=1}^{\infty}$ are independent then $\mathbf{P}(A_n \text{ i.o.}) = 1$.

Proof. We essentially already proved part (i) in the first lecture, but here is a more general repetition of the same argument.

$$\mathbf{P}(A_n \text{ i.o.}) = \mathbf{P}\left(\bigcap_{N=1}^{\infty}\bigcup_{n=N}^{\infty} A_n\right) \le \inf_{N\ge 1} \mathbf{P}\left(\bigcup_{n=N}^{\infty} A_n\right) \le \inf_{N\ge 1} \sum_{n=N}^{\infty} \mathbf{P}(A_n).$$

Since we assumed that $\sum_{n=1}^{\infty} \mathbf{P}(A_n)$, converges, this last expression is equal to 0.

Proof of (ii): Consider the complementary event that the A_n's did *not* occur for infinitely many values of n. Using De-Morgan's laws, we get

$$\begin{aligned} \mathbf{P}(\{A_n \text{ i.o.}\}^c) &= \mathbf{P}\left(\left(\bigcap_{N=1}^{\infty}\bigcup_{n=N}^{\infty} A_n\right)^c\right) = \mathbf{P}\left(\bigcup_{N=1}^{\infty}\bigcap_{n=N}^{\infty} A_n^c\right) \\ &\le \sum_{N=1}^{\infty} \mathbf{P}\left(\bigcap_{n=N}^{\infty} A_n^c\right). \end{aligned}$$

So, to show that this is 0 (under the assumptions that $\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \infty$ and that the events are independent), we show that $\mathbf{P}\left(\cap_{n=N}^{\infty} A_n^c\right) = 0$ for all $N \ge 1$. Since the events are independent, the probability of the intersection is the product of the probabilities, so we need to show that

$$\prod_{n=N}^{\infty} (1 - \mathbf{P}(A_n)) = 0,$$

or equivalently (taking minus the logarithm) that

$$-\sum_{n=N}^{\infty} \log(1 - \mathbf{P}(A_n)) = \infty.$$

But $-\log(1-x) \ge x$ for all $x > 0$, so this follows from the assumption that the series of probabilities diverges. □

Chapter 6: A brief excursion into measure theory

Here we briefly mention some of the basic definitions and results from measure theory, and point out how we used them in the previous lectures. The relevant material is covered in Appendix A.1 in [Dur2010] (Appendix A.1 and A.2 in [Dur2004]). It is not required reading, but if you read it you are perhaps more likely to attain a good understanding of the material that *is* required...

Definition 6.1. *(i) A π-system is a collection $\mathcal{P}$ of subsets of a set Ω that is closed under intersection of two sets, i.e., if $A, B \in \mathcal{P}$ then $A \cap B \in \mathcal{P}$.*
(ii) A λ-system is a collection $\mathcal{L}$ of subsets of a set Ω such that: 1. $\Omega \in \mathcal{L}$; 2. If $A, B \in \mathcal{L}$ and $A \subset B$ then $B \setminus A \in \mathcal{L}$; 3. If $(A_n)_{n=1}^{\infty}$ are all in $\mathcal{L}$ and $A_n \uparrow A$ then $A \in \mathcal{L}$.

The following is a somewhat technical result that turns out to be quite useful:

Theorem 6.2 (Dynkin's $\pi - \lambda$ theorem)**.** *If $\mathcal{P}$ is a π-system and $\mathcal{L}$ is a λ-system that contains $\mathcal{P}$ then $\sigma(\mathcal{P}) \subset \mathcal{L}$.*

Lemma 6.3 (Uniqueness theorem)**.** *If the values of two probability measures μ_1 and μ_2 coincide on a collection $\mathcal{P}$ of sets, and $\mathcal{P}$ is a π-system (closed under finite intersection), then μ_1 and μ_2 coincide on the generated σ-algebra $\sigma(\mathcal{P})$.*

The uniqueness theorem implies for example that to check if random variables X and Y are equal in distribution, it is enough to check that they have the same distribution functions.

Both of the above results are used in the proof of the following important theorem in measure theory:

Theorem 6.4 (Carathéodory's extension theorem)**.** *Let μ be an almost-probability-measure defined on an algebra $\mathcal{A}$ of subsets of a set Ω. That is, it satisfies all the axioms of a probability measure except it is defined on an algebra and not a σ-algebra; σ-additivity is satisfied whenever the countable union of disjoint sets in the algebra is also an element of the algebra. Then μ has a unique extension to a probability measure on the σ-algebra $\sigma(\mathcal{A})$ generated by $\mathcal{A}$.*

Carathéodory's extension theorem is the main tool used in measure theory for constructing measures: one always starts out by defining the measure on some relatively small family of sets and then extending to the generated σ-algebra (after verifying σ-additivity, which often requires using topological arguments, e.g., involving compactness). Applications include:

- Existence and uniqueness of Lebesgue measure in $(0, 1)$, $\mathbb{R}$ and $\mathbb{R}^d$.
- Existence and uniqueness of probability measures associated with a given distribution function in $\mathbb{R}$ (sometimes called "Lebesgue-Stieltjes measures in $\mathbb{R}$"). We proved existence instead by starting with Lebesgue measure and using quantile functions.
- Existence and uniqueness of Lebesgue-Stieltjes measures in $\mathbb{R}^d$ (i.e., measures associated with a d-dimensional joint distribution function). Here there is no good concept analogous to quantile functions, although there are other ways to construct such measures explicitly using ordinary Lebesgue measures.
- Product measures – this corresponds to probabilistic experiments which consist of several independent smaller experiments.

Note that Durrett's book also talks about measures that are not probability measures, i.e., the total measure of the space is not 1 and may even be infinite. In this setting, the theorems above can be formulated in greater generality.

Chapter 7: Expected values

7.1 Construction of the expectation operator

We wish to define the notion of the **expected value**, or **expectation**, of a random variable X, which will be denoted $\mathbf{E}X$ (or $\mathbf{E}(X)$). In measure theory this is denoted $\int X dP$ and is called the "Lebesgue integral". It is one of the most important concepts in all of mathematical analysis! So time invested in understanding it is time well-spent.

The idea is simple. For bounded random variables, we want the expectation to satisfy three properties: First, the expectation of an indicator variable 1_A, where A is an event, should be equal to $\mathbf{P}(A)$. Second, the expectation operator should be linear i.e., should satisfy $\mathbf{E}(aX+bY) = a\mathbf{E}X + b\mathbf{E}Y$ for real numbers a, b and r.v.'s X, Y. Third, it should be monotone, i.e., if $X \le Y$ (meaning $X(\omega) \le Y(\omega)$ for all $\omega \in \Omega$) then $\mathbf{E}X \le \mathbf{E}Y$.

For unbounded random variables, we will also require some kind of continuity, but let's treat the case of bounded case first. It turns out that these properties determine the expectation/Lebesgue integral operator uniquely. Different textbooks may have some variation in how they construct it, but the existence and uniqueness are really the essential facts.

Theorem 7.1. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Let B_Ω denote the class of bounded random variables. There exists a unique operator $\mathbf{E}$ that takes a r.v. $X \in B_\Omega$ and returns a number in $\mathbb{R}$, and satisfies:*

1. *If $A \in \mathcal{F}$ then $\mathbf{E}(1_A) = \mathbf{P}(A)$.*
2. *If $X, Y \in B_\Omega$, $a, b \in \mathbb{R}$ then $\mathbf{E}(aX+bY) = a\mathbf{E}(X) + b\mathbf{E}(Y)$.*
3. *If $X, Y \in B_\Omega$ and $X \ge Y$ then $\mathbf{E}(X) \ge \mathbf{E}(Y)$.*

Sketch of proof. Call X a **simple** function if it is of the form $X = \sum_{i=1}^n a_i 1_{B_i}$, where $a_1, \ldots, a_n \in \mathbb{R}$ and $B_1, \ldots, B_n$ are disjoint events. For such r.v.'s define $\mathbf{E}(X) = \sum a_i \mathbf{P}(B_i)$. Show that the linearity and monotonicity properties hold, and so far uniqueness clearly holds since we had no choice in how to define $\mathbf{E}(X)$ for such functions if we wanted the properties above to hold. Now for a general bounded r.v. X with $|X| \le M$, for any $\epsilon > 0$ it is possible to approximate X from below and above by simple functions $Y \le X \le Z$ such that

$\mathbf{E}(Z-Y)<\epsilon$. This suggests defining

$$\mathbf{E}(X)=\sup\{\mathbf{E}(Y): Y \text{ is a simple function such that } Y\le X\}. \tag{5}$$

By approximation, the construction is shown to still satisfy the properties in the Theorem and to be unique, since $\mathbf{E}(X)$ is squeezed between $\mathbf{E}(Y)$ and $\mathbf{E}(Z)$, and these can be made arbitrarily close to each other. □

We can now extend the definition of the expectation operator to non-negative random variables. In that case we still define $\mathbf{E}X$ by eq. (5). This can be thought of as a kind of "continuity from below" axiom that is added to the properties 1–3 above, although we shall see that it can be reformulated in several equivalent ways. Note that now $\mathbf{E}X$ may sometimes be infinite.

Finally, for a general random variable X, we decompose X as a difference of two non-negative r.v.'s by writing

$$X=X_+-X_-,$$

where $X_+=\max(X,0)$ is called the **positive part of** X and $X_-=\max(-X,0)$ is called the **negative part of** X.

We say that X **has an expectation** if the two numbers $\mathbf{E}X_-,\mathbf{E}X_+$ are not both ∞. In this case we define

$$\mathbf{E}X=\mathbf{E}X_+-\mathbf{E}X_-.$$

This is a number in $\mathbb{R}\cup\{-\infty,\infty\}$. If both $\mathbf{E}X_-,\mathbf{E}X_+$ are $<\infty$, or in other words if $\mathbf{E}|X|<\infty$ (since $|X|=X_++X_-$), we say that X **has finite expectation** or is **integrable**.

Theorem 7.2. *Suppose $X,Y\ge 0$ or $X,Y\le 0$ or $\mathbf{E}|X|,\mathbf{E}|Y|<\infty$. Then:*

1. *If X is a simple function then the definition* (5) *coincides with the original definition, namely $\mathbf{E}(\sum_i a_i 1_{B_i})=\sum_i a_i\mathbf{P}(B_i)$.*

2. *$\mathbf{E}(aX+bY+c)=a\mathbf{E}X+b\mathbf{E}Y+c$ for any real numbers a,b,c, where in the case where $\mathbf{E}(X)=\mathbf{E}(Y)=\pm\infty$, we require a,b to have the same sign in order for the right-hand side of this identity to be well-defined.*

3. *If $X\ge Y$ then $\mathbf{E}X\ge\mathbf{E}Y$.*

Proof. See [Dur2010], section 1.4. □

Remark. See
`http://en.wikipedia.org/wiki/Lebesgue_integration#Intuitive_interpretation`
for a nice description of the difference in approaches between the more familiar Riemann integral and the Lebesgue integral.

7.2 Properties

1. Expectation is invariant under "almost-sure equivalence": If $X \le Y$ *almost surely*, meaning $\mathbf{P}(X \le Y) = 1$, then by the definition we have $\mathbf{E}X \le \mathbf{E}Y$, since any simple function Z such that $Z \le X$ can be replaced with another simple function Z' such that $Z' \le Y$ and $Z = Z'$ almost surely. It follows also that if $X = Y$ almost surely then $\mathbf{E}X = \mathbf{E}Y$.

2. Triangle inequality: $|\mathbf{E}X| \le \mathbf{E}|X|$.

Proof.

$$|\mathbf{E}X| = |\mathbf{E}X_+ - \mathbf{E}X_-| \le \mathbf{E}X_+ + \mathbf{E}X_- = \mathbf{E}|X|.$$

□

3. Markov's inequality (Called Chebyshev's inequality in [Dur2010]):

$$\mathbf{P}(X \ge t) \le \frac{\mathbf{E}X}{t}.$$

Proof. Use monotonicity twice to deduce:

$$\mathbf{P}(X \ge t) = \mathbf{E}(1_{\{X\ge t\}}) \le \mathbf{E}\left[\frac{1}{t}X1_{\{X\ge t\}}\right] \le \frac{\mathbf{E}X}{t}.$$

□

4. Variance: If X has finite expectation, we define its **variance** to be

$$\mathbf{V}(X) = \mathbf{E}(X - \mathbf{E}X)^2.$$

If $\mathbf{V}(X) < \infty$, by expanding the square it is easy to rewrite the variance as

$$\mathbf{V}(X) = \mathbf{E}(X^2) - (\mathbf{E}X)^2.$$

We denote $\sigma(X) = \sqrt{\mathbf{V}(X)}$ and call this quantity the **standard deviation of** X. Note that if $a \in \mathbb{R}$ then $\mathbf{V}(aX) = a^2\mathbf{V}(X)$ and $\sigma(aX) = |a|\sigma(X)$.

5. Chebyshev's inequality:

$$\mathbf{P}(|X - \mathbf{E}X| \geq t) \leq \frac{\mathbf{V}(X)}{t^2}.$$

Proof. Apply Markov's inequality to $Y = (X - \mathbf{E}X)^2$. □

6. Cauchy-Schwartz inequality:

$$\mathbf{E}|XY| \leq \left(\mathbf{E}X^2 \mathbf{E}Y^2\right)^{1/2}.$$

Equality holds if and only if X and Y are linearly dependent, i.e. $aX + bY \equiv 0$ holds for some $a, b \in \mathbb{R}$.

Proof. Consider the function

$$p(t) = \mathbf{E}(|X| + t|Y|)^2 = t^2\mathbf{E}Y^2 + 2t\mathbf{E}|XY| + \mathbf{E}X^2.$$

Since $p(t) = at^2 + bt + c$ is a quadratic polynomial in t that satisfies $p(t) \geq 0$ for all t, its discriminant $b^2 - 4ac$ must be non-positive. This gives

$$(\mathbf{E}|XY|)^2 - \mathbf{E}X^2\mathbf{E}Y^2 \leq 0,$$

as claimed. The condition for equality is left as an exercise. □

7. Jensen's inequality: A function $\varphi : \mathbb{R} \to \mathbb{R}$ is called convex if it satisfies

$$\varphi(\alpha x + (1-\alpha)y) \leq \alpha\varphi(x) + (1-\alpha)\varphi(y)$$

for all $x, y \in \mathbb{R}$ and $\alpha \in [0, 1]$. If φ is convex then

$$\varphi(\mathbf{E}X) \leq \mathbf{E}(\varphi(X)).$$

Proof. See homework. □

8. L_p-norm monotonicity: If $0 < r \leq s$ then

$$(\mathbf{E}|X|^r)^{1/r} \leq (\mathbf{E}|X|^s)^{1/s}. \tag{6}$$

Proof. Apply Jensen's inequality to the r.v. $|X|^r$ with the convex function $\varphi(x) = x^{s/r}$. □

7.3 Convergence theorems for expectations

We want to study notions of continuity for the expectation operator. If $X_n \to X$ as $n \to \infty$, under what conditions do we have that $\mathbf{E}(X_n) \to \mathbf{E}X$? First we have to decide what "$X_n \to X$" actually *means.* We define two notions of convergence of a sequence of random variables to a limit.

Definition 7.3. *Let $X, X_1, X_2, \ldots$ be random variables all defined on the same probability space. We say that X_n **converges in probability to** X, and denote $X_n \xrightarrow[n\to\infty]{\mathbf{P}} X$, if for all $\epsilon > 0$ we have that*

$$\mathbf{P}(|X_n - X| > \epsilon) \xrightarrow[n\to\infty]{} 0.$$

Definition 7.4. *With $X, X_1, X_2, \ldots$ as before, we say that X_n **converges almost surely to** X (or **converges to** X **with probability 1**), and denote $X_n \xrightarrow[n\to\infty]{a.s.} X$, if*

$$\mathbf{P}(X_n \to X) = \mathbf{P}\left(\left\{\omega \in \Omega : X(\omega) = \lim_{n\to\infty} X_n(\omega)\right\}\right) = 1.$$

Exercise 7.5. *Show that $\{\omega \in \Omega : X(\omega) = \lim_{n\to\infty} X_n(\omega)\}$ is an event and therefore has a well-defined probability. In other words, represent it in terms of countable union, intersection and complementation operations on simple sets that are known to be events. Hint: Use the $\epsilon - \delta$ definition of a limit.*

Lemma 7.6. *Almost sure convergence is a stronger notion of convergence than convergence in probability. In other words, if $X_n \xrightarrow[n\to\infty]{a.s.} X$ then $X_n \xrightarrow[n\to\infty]{\mathbf{P}} X$, but the converse is not true.*

Exercise 7.7. *Prove Lemma 7.6. For the counterexample showing that convergence in probability does not imply almost sure convergence, consider the following sequence of random variables defined on the space $((0,1), \mathcal{B}$, Lebesgue measure):*

$$\begin{aligned}
&1_{(0,1)},\\
&1_{(0,1/2)}, 1_{(1/2,1)},\\
&1_{(0,1/4)}, 1_{(1/4,2/4)}, 1_{(2/4,3/4)}, 1_{(3/4,1)},\\
&1_{(0,1/8)}, 1_{(1/8,2/8)}, 1_{(2/8,3/8)}, 1_{(3/8,4/8)}, 1_{(4/8,5/8)}, 1_{(5/8,6/8)}, 1_{(6/8,7/8)}, 1_{(7/8,1)},\\
&\ldots
\end{aligned}$$

Lemma 7.8. *If $(X_n)_{n=1}^\infty$ is a sequence of r.v.s such that $X_n \xrightarrow[n\to\infty]{\mathbf{P}} X$ then there exists a subsequence $(X_{n_k})_{k=1}^\infty$ such that $X_n \xrightarrow[k\to\infty]{a.s.} X$.*

Exercise 7.9. *Prove lemma 7.8.*

We can now formulate the fundamental convergence theorems for Lebesgue integration.

Theorem 7.10 (Bounded convergence theorem)**.** *If X_n is a sequence of r.v.'s such that $|X_n| \le M$ for all n, and $X_n \to X$ in probability, then $\mathbf{E}X_n \to \mathbf{E}X$.*

Proof. Fix $\epsilon > 0$. Then

$$\begin{aligned} |\mathbf{E}X_n - \mathbf{E}X| &\le \mathbf{E}|X_n - X| = \mathbf{E}|X_n - X|1_{\{|X_n - X| > \epsilon\}} + \mathbf{E}|X_n - X|1_{\{|X_n - X| \le \epsilon\}} \\ &\le 2M\mathbf{P}(|X_n - X| > \epsilon) + \epsilon \xrightarrow[n\to\infty]{} \epsilon. \end{aligned}$$

Since ϵ was an arbitrary positive number, this implies that $|\mathbf{E}X_n - \mathbf{E}X| \to 0$, as claimed. □

Theorem 7.11 (Fatou's lemma)**.** *If $X_n \ge 0$ then* $\liminf_{n\to\infty} \mathbf{E}X_n \ge \mathbf{E}(\liminf_{n\to\infty} X_n)$.

To see that the inequality in the lemma can fail to be an equality, let $U \sim U(0,1)$, and define $X_n = n1_{\{U \le 1/n\}}$. Clearly $\liminf_{n\to\infty} X_n = \lim_{n\to\infty} X_n \equiv 0$, but $\mathbf{E}(X_n) = 1$ for all n.

Proof. Let $Y = \liminf_{n\to\infty} X_n$. Note that Y can be written as

$$Y = \sup_{n\ge1} \inf_{m\ge n} X_m$$

(this is a general fact about the lim inf of a sequence of real numbers), or $Y = \sup_n Y_n$, where we denote

$$Y_n = \inf_{m\ge n} X_m.$$

We have $Y_n \le X_n$, and as $n \to \infty$, $Y_n \to Y$ (in fact $Y_n \uparrow Y$) almost surely. Therefore $\mathbf{E}Y_n \le \mathbf{E}X_n$, so $\liminf_{n\to\infty} \mathbf{E}Y_n \le \liminf_{n\to\infty} \mathbf{E}X_n$, and therefore it is enough to show that

$$\liminf_{n\to\infty} \mathbf{E}Y_n \ge \mathbf{E}Y.$$

But for any M we have that

$$Y_n \wedge M \xrightarrow[n\to\infty]{\text{a.s.}} Y \wedge M,$$

and this is a sequence of uniformly bounded r.v.'s, therefore by the bounded convergence theorem we get that

$$\mathbf{E}(Y_n) \geq \mathbf{E}(Y_n \wedge M) \xrightarrow[n\to\infty]{} \mathbf{E}(Y \wedge M).$$

We therefore get that $\liminf_{n\to\infty} \mathbf{E}(Y_n) \geq \mathbf{E}(Y \wedge M)$ for any $M > 0$, which implies the result because of the following exercise. □

Exercise 7.12. *Let $Y \geq 0$ be a random variable. Prove that*

$$\mathbf{E}(Y) = \sup_{M>0} \mathbf{E}(Y \wedge M).$$

Theorem 7.13 (Monotone convergence theorem)**.** *If $0 \leq X_n \uparrow X$ as $n \to \infty$ then $\mathbf{E}X_n \uparrow \mathbf{E}X$.*

Proof.

$$\mathbf{E}X = \mathbf{E}[\liminf_{n\to\infty} X_n] \leq \liminf_{n\to\infty} \mathbf{E}X_n \leq \limsup_{n\to\infty} \mathbf{E}X_n \leq \limsup_{n\to\infty} \mathbf{E}X = \mathbf{E}X.$$

□

Theorem 7.14 (Dominated convergence theorem)**.** *If $X_n \to X$ almost surely, $|X_n| \leq Y$ for all $n \geq 1$ and $\mathbf{E}Y < \infty$, then $\mathbf{E}X_n \to \mathbf{E}X$.*

Proof. Apply Fatou's lemma separately to $Y + X_n$ and to $Y - X_n$. □

7.4 Computing expected values

Lemma 7.15. *If X is a discrete r.v., that is, takes values in some countable set S, then*

$$\mathbf{E}X = \sum_{s\in S} s\,\mathbf{P}(X = s)$$

when the right-hand side is well-defined, i.e., when at least one of the numbers

$$\mathbf{E}(X_-) = \sum_{s\in S, s<0} (-s)\,\mathbf{P}(X = s), \qquad \mathbf{E}(X_+) = \sum_{s\in S, s>0} s\,\mathbf{P}(X = s)$$

is finite. It follows that for any function $g : \mathbb{R} \to \mathbb{R}$, we also have

$$\mathbf{E}(g(X)) = \sum_{s\in S} g(s)\mathbf{P}(X = s).$$

Proof. If S is finite then X is a simple function, and can be written $X = \sum_{s\in S} s\, 1_{\{X=s\}}$, so this follows from the definition of $\mathbf{E}(\cdot)$ for simple functions. If S is infinite this follows (check!) from the convergence theorems in the previous section by considering approximations to X of the form $\sum_{s\in S, |s|<M} s\, 1_{\{X=s\}}$. □

Lemma 7.16. *If X is a r.v. with a density function f_X, then*

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

when the right-hand side is well-defined, i.e., when at least one of the numbers

$$\mathbf{E}(X_-) = -\int_{-\infty}^{0} x f_X(x)\, dx, \qquad \mathbf{E}(X_+) = \int_{0}^{\infty} x f_X(x)\, dx$$

is finite. Similarly, for any "reasonable" function g we have

$$\mathbf{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx.$$

Proof. Fix $\epsilon > 0$, and approximate X by a discrete r.v. Y, e.g.,

$$Y = \sum_{k=-\infty}^{\infty} k\epsilon 1_{\{k\epsilon < X \leq (k+1)\epsilon\}}.$$

Then $|\mathbf{E}(X) - \mathbf{E}(Y)| \leq \mathbf{E}|X - Y| \leq \epsilon$. By the previous lemma we have

$$\mathbf{E}(Y) = \sum_{k=-\infty}^{\infty} k\epsilon \mathbf{P}(k\epsilon < X < (k+1)\epsilon) = \sum_{k=-\infty}^{\infty} k\epsilon \int_{k\epsilon}^{(k+1)\epsilon} f_X(x)\, dx,$$

so the result for $\mathbf{E}X$ follows by letting $\epsilon \to 0$. For general functions $g(X)$ repeat this argument, and invoke the relevant convergence theorem to deduce that $\mathbf{E}(g(Y)) \to \mathbf{E}g(Y)$ as $\epsilon \to 0$. □

7.5 Expectation and independent random variables

Theorem 7.17. *(i) If X, Y are independent r.v.'s then $\mathbf{E}(XY) = \mathbf{E}X\mathbf{E}Y$.*
(ii) X, Y are independent if and only if $\mathbf{E}[g(X)h(Y)] = \mathbf{E}g(X)\mathbf{E}h(Y)$ for all bounded measurable functions $g, h : \mathbb{R} \to \mathbb{R}$.

Sketch of proof. Part (i) follows either by approximation of X, Y using simple functions, or using Fubini's theorem, which you can read about in section 1.7 of [Dur2010] (Note that Fubini's theorem in turn is proved by approximation using simple functions, so these two seemingly different approaches are really equivalent).

For part (ii), the "only if" follows from part (i) together with the observation that if X, Y are independent then so are $g(X), h(Y)$. For the "if" part, observe that the function $1_{(a,b)}$ is a bounded measurable function, so in particular the condition $\mathbf{E}[g(X)h(Y)] = \mathbf{E}g(X)\mathbf{E}h(Y)$ includes the information that $\mathbf{P}(X \in I, Y \in J) = \mathbf{P}(X \in I)\mathbf{P}(Y \in J)$ for any two finite intervals I, J, which we already know is enough to imply independence. $\square$

Theorem 7.18. *(i) If $X_1, X_2, \ldots, X_n$ are independent r.v.'s then $\mathbf{E}(X_1 \ldots X_n) = \prod_{k=1}^n \mathbf{E}X_k$.*
(ii) $X_1, \ldots, X_n$ are independent if and only if $\mathbf{E}\left(\prod_{k=1}^n g_k(X_k)\right) = \prod_{k=1}^n \mathbf{E}g_k(X_k)$ for all bounded measurable functions $g_1, \ldots, g_n : \mathbb{R} \to \mathbb{R}$.

The fact that expectation is multiplicative for independent random variables implies an important fact about the variance of a sum of independent r.v.'s. Let X, Y be independent r.v.'s with finite variance. Then we get immediately that

$$\begin{aligned} \mathbf{V}(X+Y) &= \mathbf{E}\left[(X - \mathbf{E}X) + (Y - \mathbf{E}Y)\right]^2 \\ &= \mathbf{E}(X-\mathbf{E}X)^2 + \mathbf{E}(Y-\mathbf{E}Y)^2 + 2\mathbf{E}\left[(X-\mathbf{E}X)(Y-\mathbf{E}Y)\right] \\ &= \mathbf{V}(X) + \mathbf{V}(Y) + 0 = \mathbf{V}(X) + \mathbf{V}(Y). \end{aligned}$$

More generally, if X, Y are not necessarily independent, then we can define the covariance of X and Y by

$$\mathrm{Cov}(X,Y) = \mathbf{E}\left[(X-\mathbf{E}X)(Y-\mathbf{E}Y)\right].$$

We then get the more general formula

$$\mathbf{V}(X+Y) = \mathbf{V}(X) + \mathbf{V}(Y) + 2\mathrm{Cov}(X,Y).$$

Repeating this computation with a sum of n variables instead of just two, we get the following formula for the variance of a sum of r.v.'s.

Lemma 7.19. *If $X_1, \ldots, X_n$ are r.v.'s with finite variance, then*

$$\mathbf{V}\left(\sum_{k=1}^n X_k\right) = \sum_{k=1}^n \mathbf{V}(X_k) + 2\sum_{1 \le i < j \le n} Cov(X_i, X_j).$$

Lemma 7.20 (Properties of the covariance)**.** *If X, Y are r.v.'s with finite variance, then:*

1. $Cov(X, Y) = Cov(Y, X) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)$.
2. $Cov(X, X) = \mathbf{V}(X)$.
3. $Cov(aX_1 + bX_2, Y) = a\,Cov(X_1, Y) + b\,Cov(X_2, Y)$
4. $Cov(X, aY_1 + bY_2) = a\,Cov(X, Y_1) + b\,Cov(X, Y_2)$
5. *If X, Y are independent then* $Cov(X, Y) = 0$.
6. $|Cov(X, Y)| \leq \sigma(X)\sigma(Y)$, *with equality if and only if X and Y are linearly dependent.*

Proof. Properties 1–5 are obvious. Property 6 follows by applying the Cauchy-Schwartz inequality to the r.v.'s $X - \mathbf{E}X$ and $Y - \mathbf{E}Y$. □

If $\text{Cov}(X, Y) = 0$ we say that X and Y are **uncorrelated**, or **orthogonal**. This is a weaker condition than being independent, but because of the way the variance of a sum of r.v.'s behaves, it is still often useful for deriving bounds, as we shall see.

Define the **correlation coefficient of X and Y** by

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma(X)\sigma(Y)}.$$

This measures the correlation in units of the standard deviation of X and Y so does not depend on the choice of scale. From property 6 in the above lemma, we get that

$$-1 \leq \rho(X, Y) \leq 1,$$

with equality on either side if and only if X and Y are linearly dependent.

7.6 Moments

For an integer $k \geq 0$, the **k-th moment** of a random variable X is the number $\mathbf{E}(X^k)$. The **k-th moment around a point** $c \in \mathbb{R}$ is the number $\mathbf{E}(X - c)^k$. If c is not mentioned it is understood that the moment is around 0. The **k-th central moment** is the k-th moment around $\mathbf{E}X$ (when it exists), i.e., $\mathbf{E}(X - \mathbf{E}X)^k$. In this terminology, the variance is the second central moment.

The sequence of moments (usually around 0, or around $\mathbf{E}X$) often contains important information about the behavior of X and is an important computational and theoretical tool. Important special distributions often turn out to have interesting sequences of moments. Also note that by the monotonicity of the L_p norms (inequality (6) in Section 7.2), the set of values $r \geq 0$ such that $\mathbf{E}(X^r)$ exists (one can also talk about r-th moments for non-integer r, but that is much less commonly discussed) is an interval containing 0.

A nice characterization of the variance is that it is the minimal second moment. To compute the second moment around a point t, we can write

$$\begin{aligned} \mathbf{E}(X-t)^2 &= \mathbf{E}[(X-\mathbf{E}X)-(t-\mathbf{E}X)]^2 \\ &= \mathbf{E}(X-\mathbf{E}X)^2+\mathbf{E}(t-\mathbf{E}X)^2+2(t-\mathbf{E}X)\mathbf{E}(X-\mathbf{E}X) \\ &= \mathbf{V}(X)+(t-\mathbf{E}X)^2 \geq \mathbf{V}(X). \end{aligned}$$

So the function $t \longrightarrow \mathbf{E}(X-t)^2$ is a quadratic polynomial that attains its minimum at $t=\mathbf{E}X$, and the value of the minimum is $\mathbf{V}(X)$. In words, the identity

$$\mathbf{E}(X-t)^2 = \mathbf{V}(X)+(t-\mathbf{E}X)^2$$

says that "the second moment around t is equal to the second moment around the mean $\mathbf{E}X$ plus the square of the distance between t and the mean". Note that this is analogous to (and mathematically equivalent to) the Huygens-Steiner theorem (also called the Parallel Axis theorem, see Wikipedia) from mechanics, which says that "the moment of inertia of a body with unit mass around a given axis L is equal to the moment of inertia around the line parallel to L passing through the center of mass of the body, plus the square of the distance between the two lines". Indeed, the "moment" terminology seems to originate in this physical context.

Chapter 8: Special distributions and their properties

A small number of distributions come up again and again in real life applications of probability theory and as answers to natural theoretical questions. These are the so-called **special distributions**. In this chapter we survey the most important of the special distributions and some of their properties. The properties are either obvious identities, restatements of known results or are left as (strongly recommended) exercises.

Below we use the following notation: if D_1 and D_2 are probability distributions, $D_1 \overset{d}{=} D_2$ denotes that they are equal; $D_1 \boxplus D_2$ denotes the distribution of the random variable $X+Y$ where $X \sim D_1$, $Y \sim D_2$ and X, Y are independent. Similarly, $\underset{k=1}{\overset{n}{\boxplus}} D_k$ denotes the distribution of the random variable $\sum_{k=1}^n X_k$ where $X_1, \ldots, X_n$ are independent random variables such that $X_k \sim D_k$ for $k = 1, \ldots, n$.

8.1 The Bernoulli distribution

The Bernoulli distribution models the probabilistic experiment of a single coin toss. We say that X has the Bernoulli distribution with parameter $0 < p < 1$, and denote $X \sim \text{Ber}(p)$, if X satisfies

$$\mathbf{P}(X = 0) = p = 1 - \mathbf{P}(X = 1).$$

Properties:

1. $\mathbf{E}X = p$, $\mathbf{V}(X) = p(1-p)$.

2. $\text{Ber}(1/2)$ is the distribution that maximizes the variance $\mathbf{V}(X)$ subject to the constraint that $0 \leq X \leq 1$.

8.2 The binomial distribution

We say that X has the binomial distribution with parameters $n \geq 1$ and $0 < p < 1$, and denote $X \sim \text{Bin}(n, p)$, if X satisfies

$$\mathbf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad (0 \leq k \leq n).$$

Properties:

1. $\mathrm{Bin}(1,p) \overset{d}{=} \mathrm{Ber}(p)$.

2. $\mathrm{Bin}(n,p) \boxplus \mathrm{Bin}(m,p) \overset{d}{=} \mathrm{Bin}(n+m,p)$.

3. $\mathrm{Bin}(n,p) \overset{d}{=} \underset{k=1}{\overset{n}{\boxplus}} \mathrm{Ber}(p)$. That is, the binomial distribution models the number of successes when n identical experiments are performed independently, where each experiment has probability p of success.

4. If $X \sim \mathrm{Bin}(n,p)$ then $\mathbf{E}X = np$, $\mathbf{V}(X) = np(1-p)$.

8.3 The geometric distribution

We say that X has the geometric distribution with parameter $0 < p < 1$, and denote $X \sim \mathrm{Geom}(p)$, if X satisfies

$$\mathbf{P}(X = k) = p(1-p)^{k-1} \qquad (k \geq 1).$$

Some authors prefer a slightly different convention whereby the geometric random variables take nonnegative values (including 0) rather than only positive values. Thus, denote $X' \sim \mathrm{Geom}_0(p)$, and say that X' has the geometric distribution starting from 0, if it satisfies

$$\mathbf{P}(X' = k) = p(1-p)^{k} \qquad (k \geq).$$

Properties:

1. $\mathrm{Geom}(p) \overset{d}{=} \mathrm{Geom}_0(p) + 1$.

2. If $W_1, W_2, W_3, \ldots$ is a sequence of i.i.d. r.v.'s with distribution $\mathrm{Ber}(p)$, then

$$X = \min\{k \geq 1 \,:\, W_k = 1\} \sim \mathrm{Geom}(p).$$

   That is, the distribution $\mathrm{Geom}(p)$ models the number of identical independent experiments we had to perform to get the first successful outcome, when each experiment has probability p of success. The variant $\mathrm{Geom}_0(p)$ corresponds to the number of *failed* experiments before the first success.

3. The geometric distribution has the (discrete) **lack of memory property**. More precisely, if $X \sim \mathrm{Geom}(p)$ then

$$\mathbf{P}(X \geq n+k \mid X \geq k) = \mathbf{P}(X \geq n) \qquad \text{for all } n, k \geq 1.$$

4. If $X \sim \mathrm{Geom}(p)$ then $\mathbf{E}X = \frac{1}{p}$, $\mathbf{V}(X) = \frac{1-p}{p^2}$.

5. If $X' \sim \mathrm{Geom}(p)$ then $\mathbf{E}X' = \frac{1-p}{p}$, $\mathbf{V}(X') = \frac{1-p}{p^2}$.

8.4 The negative binomial distribution

We say that X has the negative binomial distribution with parameters $m \geq 1$ and $0 < p < 1$, and denote $X \sim \mathrm{NB}(m,p)$, if X satisfies

$$\mathbf{P}(X = k) = \binom{k+m-1}{k} p^m (1-p)^k \qquad (k \geq 0).$$

Properties:

1. $\mathrm{NB}(1,p) \overset{d}{=} \mathrm{Geom}_0(1-p)$.

2. $\mathrm{NB}(m,p) \boxplus \mathrm{NB}(n,p) = \mathrm{NB}(n+m,p)$.

3. If $W_1, W_2, W_3, \ldots$ is a sequence of i.i.d. r.v.'s with distribution $\mathrm{Ber}(p)$, then

$$X = \min\left\{ k \geq 0 \,:\, \sum_{j=1}^{k+m} (1 - W_j) = m \right\} \sim \mathrm{NB}(m,p).$$

   In words, when performing a sequence of identical experiments, each with probability p of success, the number of successes observed before the mth failure is distributed according to $\mathrm{NB}(m,p)$.

4. $\mathrm{NB}(m,p) = \mathop{\boxplus}_{k=1}^{m} \mathrm{Geom}_0(1-p)$.

8.5 The Poisson distribution

We say that X has the Poisson distribution with parameter $\lambda > 0$, and denote $X \sim \mathrm{Poi}(\lambda)$, if X satisfies

$$\mathbf{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \qquad (k \geq 0).$$

Properties:

1. $\mathrm{Poi}(\lambda) \boxplus \mathrm{Poi}(\mu) = \mathrm{Poi}(\lambda + \mu)$.

2. The Poisson distribution is the limit of the binomial distributions $\mathrm{Bin}(n,p)$ where the number n of experiments tends to infinity and the probability p of success in each individual experiment goes to 0 in such a way that the mean number np of successes stays fixed. More precisely, if $X \sim \mathrm{Poi}(\lambda)$ and for each n, W_n is a r.v. with distribution $\mathrm{Bin}(n, \lambda/n)$, then
$$\mathbf{P}(W_n = k) \xrightarrow[n\to\infty]{} \mathbf{P}(X = k) \qquad (k \ge 0).$$
(This is known as the **law of rare events**; see section 16.3 for the proof of a similar result that holds in much greater generality.)

3. If $X \sim \mathrm{Poi}(\lambda)$ then $\mathbf{E}X = \lambda$, $\mathbf{V}(X) = \lambda$.

8.6 The uniform distribution

We say that X has the uniform distribution in the interval $[a,b]$, and denote $X \sim U[a,b]$, if X has density function
$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b, \\ 0 & \text{otherwise.} \end{cases},$$
or equivalently if the c.d.f. of X is given by
$$F_X(x) = \begin{cases} 0 & \text{if } x < a, \\ \frac{x-a}{b-a} & \text{if } a \le x \le b, \\ 1 & \text{if } x > b. \end{cases}.$$

Properties:

1. $\mathbf{E}(X) = \frac{a+b}{2}$, $\mathbf{V}(X) = \frac{(b-a)^2}{12}$.

8.7 The normal distribution

We say that X has the normal (a.k.a. gaussian) distribution with mean μ and variance σ^2, and denote $X \sim N(\mu, \sigma^2)$, if X has density function
$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \qquad (x \in \mathbb{R}).$$

In particular, the standard normal distribution is the distribution $N(0,1)$, whose density function is given by

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

Properties:

1. If $X \sim N(\mu, \sigma^2)$ then $\mathbf{E}(X) = \mu$, $\mathbf{V}(X) = \sigma^2$.

2. $N(\mu_1, \sigma_1^2) \boxplus N(\mu_2, \sigma_2^2) = N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

3. If $X, Y \sim N(0,1)$ are independent and standard normal then $\frac{1}{\sqrt{2}}(X+Y) \sim N(0,1)$.

4. More generally, if $X_1, \ldots, X_n \sim N(0,1)$ are independent standard normal r.v.s then

   $$\frac{1}{\sqrt{n}}(X_1 + \ldots + X_n) \sim N(0,1),$$

   and also

   $$\sum_{j=1}^n \alpha_j X_j \sim N(0,1)$$

   if $\alpha_1, \ldots, \alpha_n$ are real numbers such that $\sum_j \alpha_j^2 = 1$. (Geometrically, $\boldsymbol{\alpha} \cdot \mathbf{X} = \sum_{j=1}^n \alpha_j X_j$ can be interpreted as the projection of the random vector $(X_1, \ldots, X_n)$ in $\mathbb{R}^n$ in the direction of the unit vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$.)

5. If $X \sim N(0,1)$ then $\mathbf{E}(X^k) = \begin{cases} 0 & \text{if } k \text{ is odd,} \\ 1 \cdot 3 \cdot 5 \cdot \ldots \cdot (k-1) & \text{if } k \text{ is even.} \end{cases}$

6. The normal distribution is the single most important distribution in probability! The theoretical reason for this is the **Central Limit Theorem**, a result we will discuss in detail in chapters 11–14.

7. The polar decomposition of a bivariate standard normal vector: given a pair (X,Y) of random variables which in the polar representation are written as $X = R\cos\Theta$, $Y = R\sin\Theta$, where $R > 0$ and $0 \le \Theta < 2\pi$, we have

   $$\begin{aligned} &X, Y \sim N(0,1),\ X, Y \text{ are independent} \\ &\qquad\qquad \iff R^2 \sim \mathrm{Exp}(1/2),\ \Theta \sim U[0, 2\pi] \text{ and } R, \Theta \text{ are independent.} \end{aligned}$$

8.8 The exponential distribution

We say that X has the exponential distribution with parameter λ, and denote $X \sim \text{Exp}(\lambda)$, if X has density function

$$f_X(x) = \lambda e^{-\lambda x} \qquad (x \geq 0)$$

and the associated c.d.f.

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 - e^{-\lambda x} & \text{if } x \geq 0. \end{cases}$$

Properties:

1. λ has the role of an (inverse) **scale parameter**, in the sense that for $c > 0$ we have that $c\,\text{Exp}(\lambda) \overset{d}{=} \text{Exp}(\lambda/a)$; i.e., scaling an exponential r.v. by a factor c gives a new exponential r.v. where the scale parameter is divided by c.

2. The exponential distribution satisfies the **lack of memory property**. More precisely, if $X \sim \text{Exp}(\lambda)$ then

$$\mathbf{P}(X > t + s \,|\, X > t) = \mathbf{P}(X > s) \qquad (t, s > 0).$$

   Furthermore, it is not hard to show that the exponential distribution is the unique distribution on $[0, \infty)$ satisfying this property.

3. $\mathbf{E}(X) = \lambda$, $\mathbf{V}(X) = \lambda^2$.

4. The exponential distribution can be thought of as a scaling limit of geometric random variables, when the geometric distribution is interpreted as measuring *time* rather than the number of experiments, and time is scaled so that the i.i.d. Bernoulli experiments are performed more and more frequently, but are becoming less and less probable to succeed, in such a way that the mean number of successful experiments per unit of time remains constant. More precisely, if $\lambda > 0$ is fixed, $X \sim \text{Exp}(\lambda)$, and for each n (larger than λ) we let W_n denote a random variable with distribution $\text{Geom}(\lambda/n)$, then we have

$$\mathbf{P}\left(\frac{1}{n}W_n > t\right) \xrightarrow[n\to\infty]{} \mathbf{P}(X > t) \qquad (t > 0).$$

5. If $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\mu)$ are independent r.v.s then $\min(X, Y) \sim \text{Exp}(\lambda + \mu)$.

6. If $X_1, X_2, \ldots$ are i.i.d. Exp(1) random variables, and we define the cumulative sums $S_0 = 0$, $S_n = \sum_{k=1}^n X_k$, then for each $\lambda > 0$, the random variable

$$N(\lambda) = \max\{n \geq 0 \,:\, S_n \leq \lambda\},$$

then $N(\lambda) \sim \text{Poi}(\lambda)$.

Note. One can consider $N(\lambda)$ not just for a single value of λ but the entire family $N(t)$, where $t > 0$ is a parameter denoting time. When considered as such a family, $(N(t))_{t>0}$ is called a **Poisson process**.

8.9 The gamma distribution

To define the gamma distribution, first we define the **Euler gamma function** (also called the **generalized factorial function**), an important special function of mathematical analysis, denoted $\Gamma(t)$, by

$$\Gamma(t) = \int_0^\infty e^{-x} x^{t-1}\, dt \qquad (t > 0).$$

Properties of the gamma function:

1. $\Gamma(n) = (n-1)!$ for integer $n \geq 1$.

2. $\Gamma(t+1) = t\,\Gamma(t)$ for all $t > 0$.

3. $\Gamma(1/2) = \sqrt{\pi}$.

Next, we say that X has the gamma distribution with parameters $\alpha, \lambda > 0$, and denote $X \sim \text{Gamma}(\lambda)$, if X has density function

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1} \qquad (x > 0).$$

Properties of the gamma distribution:

1. $\text{Exp}(\lambda) \overset{d}{=} \text{Gamma}(1, \lambda)$.

2. The parameter λ has the role of a scale parameter in the same sense as for the exponential distribution: for $c > 0$ we have $c\,\text{Gamma}(\alpha, \lambda) \overset{d}{=} \text{Gamma}(\alpha, \lambda/c)$.

3. $\text{Gamma}(\alpha, \lambda) \boxplus \text{Gamma}(\beta, \lambda) = \text{Gamma}(\alpha + \beta, \lambda)$.

4. $\text{Gamma}(\alpha, \lambda) = \overset{n}{\underset{k=1}{\boxplus}} \text{Exp}(\lambda)$.

5. If $X \sim \text{Gamma}(\alpha, \lambda)$ then $\mathbf{E}X = \frac{\alpha}{\lambda}$, $\mathbf{V}(X) = \frac{\alpha}{\lambda^2}$.

8.10 The beta distribution

Define the **Euler beta function** (which is closely related to the Euler gamma function) by

$$B(a, b) = \int_0^1 u^{a-1}(1-u)^{b-1}\, du \qquad (a, b > 0).$$

Properties of the beta function:

1. $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

2. For integer $m, n \geq 1$, $B(m, n) = \frac{(m-1)!(n-1)!}{(m+n-1)!}$.

We say that X has the beta distribution with parameters $a, b > 0$, and denote $X \sim \text{Beta}(a, b)$, if X has density function

$$f_X(x) = \frac{1}{B(a, b)} x^{a-1}(1-x)^{b-1} \qquad (0 < x < 1).$$

Properties of the beta distribution:

1. $U[0, 1] \overset{d}{=} \text{Beta}(1, 1)$.

2. If $X \sim \text{Gamma}(\alpha, \lambda)$ and $Y \sim \text{Gamma}(\beta, \lambda)$ are independent, then $U = \frac{X}{X+Y}$ has distribution $\text{Beta}(\alpha, \beta)$, and is independent of $X + Y$.

3. If $X_1, X_2, \ldots$ is a sequence of i.i.d. r.v.s with distribution $\text{Exp}(\lambda)$, and $S_m = \sum_{k=1}^m X_k$ are the cumulative sums of the sequence, then for all $n \geq k \geq 1$, $S_k/S_n \sim \text{Beta}(k, n-k)$, and S_k/S_n is independent of S_n.

4. If $X \sim \text{Beta}(a, b)$ then $\mathbf{E}X = \frac{a}{a+b}$, $\mathbf{V}(X) = \frac{ab}{(a+b)^2(a+b+1)}$.

5. If $X_1, \ldots, X_n$ are i.i.d. $U[0, 1]$ random variables, and $X^{(1)} < X^{(2)} < \ldots < X^{(n)}$ are their order statistics, i.e., $X^{(k)}$ is defined as the kth smallest among the numbers $X_1, \ldots, X_n$, then $X^{(k)} \sim \text{Beta}(k, n+1-k)$.

8.11 The Cauchy distribution

We say that X has the Cauchy distribution, and denote $X \sim \text{Cauchy}$, if X has the density

$$f_X(x) = \frac{1}{\pi} \frac{1}{1+x^2}.$$

Properties:

1. $\mathbf{E}|X| = \infty$, i.e., the Cauchy distribution has no expectation.

2. If $X, Y \sim \text{Cauchy}$ are independent then their average $\frac{1}{2}(X+Y)$ is also distributed according to the Cauchy distribution.

3. More generally, if $X_1, \ldots, X_n \sim \text{Cauchy}$ and $\alpha_1, \ldots, \alpha_n \geq 0$ are numbers such that $\sum_j \alpha_j = 1$, then the weighted average

$$\sum_{j=1}^{n} \alpha_j X_j \sim \text{Cauchy}.$$

4. If $\Theta \sim U[-\pi/2, \pi/2]$ then $X = \tan \Theta \sim \text{Cauchy}$.

Summary: Special distributions

Name	Notation	Formula		$\mathbf{E}(X)$	$\mathbf{V}(X)$	$\mathbf{E}(X^k)$
Discrete uniform	$X \sim U\{1,\dots,n\}$	$\mathbf{P}(X=k)=\frac{1}{n}$	$(1 \le k \le n)$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	
Bernoulli	$X \sim \mathrm{Bernoulli}(p)$	$\mathbf{P}(X=0)=1-p,\ \mathbf{P}(X=1)=p$		p	$p(1-p)$	p
Binomial	$X \sim \mathrm{Binomial}(n,p)$	$\mathbf{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$	$(0 \le k \le n)$	np	$np(1-p)$	
Geometric (from 0)	$X \sim \mathrm{Geom}_0(p)$	$\mathbf{P}(X=k)=p(1-p)^k$	$(k \ge 0)$	$\frac{1}{p}-1$	$\frac{1-p}{p^2}$	
Geometric (from 1)	$X \sim \mathrm{Geom}(p)$	$\mathbf{P}(X=k)=p(1-p)^{k-1}$	$(k \ge 1)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	
Poisson	$X \sim \mathrm{Poisson}(\lambda)$	$\mathbf{P}(X=k)=e^{-\lambda}\frac{\lambda^k}{k!}$	$(k \ge 0)$	λ	λ	Bell numbers (for $\lambda=1$)
Negative binomial	$X \sim \mathrm{NB}(m,p)$	$\mathbf{P}(X=k)=\binom{k+m-1}{m-1}p^m(1-p)^k$	$(k \ge 0)$	$\frac{m(1-p)}{p}$	$\frac{m(1-p)}{p^2}$	
Uniform	$X \sim U(a,b)$	$f_X(x)=\frac{1}{b-a}$	$(a<x<b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{b^{k+1}-a^{k+1}}{(k+1)(b-a)}$
Exponential	$X \sim \mathrm{Exp}(\lambda)$	$f_X(x)=\lambda e^{-\lambda x}$	$(x>0)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\lambda^{-k}k!$
Standard normal	$X \sim N(0,1)$	$f_X(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$	$(x \in \mathbb{R})$	0	1	$\begin{cases} \frac{k!}{(k/2)!2^{k/2}} & k \text{ even} \\ 0 & k \text{ odd} \end{cases}$
Normal	$X \sim N(\mu,\sigma^2)$	$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}$	$(x \in \mathbb{R})$	μ	σ^2	
Gamma	$X \sim \mathrm{Gamma}(\alpha,\lambda)$	$f_X(x)=\frac{\lambda^\alpha}{\Gamma(\alpha)}e^{-\lambda x}x^{\alpha-1}$	$(x>0)$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$	$\lambda^{-k}\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}$
Cauchy	$X \sim \mathrm{Cauchy}$	$f_X(x)=\frac{1}{\pi}\frac{1}{1+x^2}$	$(x \in \mathbb{R})$	N/A	N/A	N/A
Beta	$X \sim \mathrm{Beta}(a,b)$	$f_X(x)=\frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$	$(0<x<1)$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	$\frac{B(a+k,b)}{B(a,b)}$
Chi-squared	$X \sim \chi^2_{(n)}$	$f_X(x)=\frac{1}{2^{n/2}\Gamma(n/2)}e^{-x/2}x^{\frac{n}{2}-1}$	$(x>0)$	n	$2n$	

Useful facts: ("$\boxplus$" denotes convolution, i.e., sum of independent samples; "$\overset{d}{=}$" denotes equality of distributions)

$\mathrm{Binomial}(n,p) \boxplus \mathrm{Binomial}(m,p) \overset{d}{=} \mathrm{Binomial}(n+m,p)$

$\mathrm{Poisson}(\lambda) \boxplus \mathrm{Poisson}(\mu) \overset{d}{=} \mathrm{Poisson}(\lambda+\mu)$

$\mathrm{Geom}_0(p) \overset{d}{=} \mathrm{NB}(1,1-p)$

$\mathrm{NB}(n,p) \boxplus \mathrm{NB}(m,p) \overset{d}{=} \mathrm{NB}(n+m,p)$

$N(0,1)^2 \overset{d}{=} \mathrm{Gamma}(1/2,1/2) \overset{d}{=} \chi^2_{(1)}$

$\mathrm{Gamma}(\alpha,\lambda) \boxplus \mathrm{Gamma}(\beta,\lambda) \overset{d}{=} \mathrm{Gamma}(\alpha+\beta,\lambda)$

$N(\mu_1,\sigma_1^2) \boxplus N(\mu_2,\sigma_2^2) \overset{d}{=} N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$

$\mathrm{Exp}(\lambda) \overset{d}{=} \mathrm{Gamma}(1,\lambda)$

$\big(\alpha\,\mathrm{Cauchy}\big) \boxplus \big((1-\alpha)\,\mathrm{Cauchy}\big) \overset{d}{=} \mathrm{Cauchy} \qquad (0 \le \alpha \le 1)$

$\chi^2_{(n)} \overset{d}{=} \mathrm{Gamma}(n/2,1/2)$

Chapter 9: Laws of large numbers

Let $X_1, X_2, X_3, \dots$ be a sequence of independent and identically distributed random variables. A common abbreviation for "independent and identically distributed" is **i.i.d.**. What this can mean is that we are taking repeated independent samples from some distribution, for example when doing a poll or an experiment in quantum mechanics (when doing a poll, it might be better to have the samples not be independent – after all, does it really make sense to call up the same person twice? But if the population is large the effect of having the samples be independent is negligible, and can make the analysis of the results a bit simpler).

Assume that $\mathbf{E}X_1$, the mean of X_1 (therefore of every X_n), is defined and finite, and denote it by μ. In a real-life situation, we might not know what the mean is, and wish to estimate it. So we look at the sum of the first n samples,

$$S_n = \sum_{k=1}^{n} X_k,$$

and use it to form the **empirical average**,

$$\overline{X}_n = \frac{1}{n} S_n = \frac{1}{n} \sum_{k=1}^{n} X_k.$$

The natural question is whether we can expect the empirical average to be close to the true mean $\mu = \mathbf{E}X_1$ as $n \to \infty$, and in what sense of "close". A theorem or statement to that effect, making some assumptions, is called a **law of large numbers**. This can be generalized to a large extent to cases when the X_n's are not identically distributed, or not independent, or both, etc. Thus, the "Law of Large Numbers" is not a single theorem in the normal sense of the word but rather a class of theorems, or even a "principle" or "meta-theorem" that you might hear a probabilist refer to in a somewhat vague, metaphorical way.

9.1 Weak laws of large numbers

Theorem 9.1 (Weak Law of Large Numbers (WLLN)). *If* $\mathbf{E}|X_1| < \infty$ *then*

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{\mathbf{P}} \mu.$$

Proof in the case of finite variance. In the most general case, proving this requires some work – we shall deduce it from its stronger cousin, the Strong Law of Large Numbers. However, if we assume that $\sigma^2 = \mathbf{V}(X_1) < \infty$, the proof is extremely easy! It suffices to note that in this case $\mathbf{V}(S_n) = n\sigma^2$ (this is true even when the X_n's are not independent but only uncorrelated), or $\mathbf{V}(S_n/n) = \sigma^2/n$. From Chebyshev's inequality we then have that for any $\epsilon > 0$,

$$\mathbf{P}(|n^{-1}S_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \xrightarrow[n\to\infty]{} 0. \tag{7}$$

□

9.2 Strong laws of large numbers

Our goal in this section will be to prove:

Theorem 9.2 (Strong Law of Large Numbers (SLLN))**.** *If* $\mathbf{E}|X_1| < \infty$ *then*

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{a.s.} \mu.$$

As in the case of the weak law, it turns out that this is easier to prove when making more restrictive assumptions about the distribution of X, and specifically about the existence of moments. So we will prove it several times, successively weakening our assumptions until we get to the most general (but least easy to prove) result.

Proof in the case of a finite fourth moment. To prove that $S_n/n \to \mu$, we want to prove that $\mathbf{P}(|S_n/n - \mu| > \epsilon \text{ i.o.}) = 0$. Looking at the bound (7), we see that if the bound $\sigma^2/n\epsilon^2$ were to form the n-th general term of a convergent series, then by the Borel-Cantelli lemma this would be enough to get the desired consequence. This is not the case, but if we assume that X_1 has a finite fourth moment, then we could do a similar trick that *will* give a convergent series. In that case, using Markov's inequality we can get that

$$\mathbf{P}(|S_n/n - \mu| > \epsilon) = \mathbf{P}\left(\left(\sum_{k=1}^{n}(X_k - \mu)\right)^4 > n^4\epsilon^4\right) \leq \frac{\mathbf{E}(S_n - n\mu)^4}{n^4\epsilon^4}. \tag{8}$$

Denote $\hat{X}_k = X_k - \mu$, and $T_n = S_n - n\mu = \sum_{k=1}^n \hat{X}_k$. To bound $\mathbf{E}(T_n^4)$, note that we can write

$$\begin{aligned} T_n^4 &= \sum_{k=1}^n \hat{X}_k{}^4 + \binom{4}{1} \sum_{1\le i\ne j\le n} \hat{X}_i{}^3 \hat{X}_j + \frac{1}{2}\binom{4}{2} \sum_{1\le i\ne j\le n} \hat{X}_i{}^2 \hat{X}_j{}^2 \\ &\quad + \frac{1}{2}\binom{4}{2,1,1} \sum_{1\le i,j,k\le n \text{ distinct}} \hat{X}_i{}^2 \hat{X}_j \hat{X}_k + \frac{1}{4!}\binom{4}{1,1,1,1} \sum_{1\le i,j,k,\ell\le n \text{ distinct}} \hat{X}_i \hat{X}_j \hat{X}_k \hat{X}_\ell \end{aligned}$$

(where $\binom{4}{2,1,1} = 4!/2!1!1!$, $\binom{4}{1,1,1,1} = 4!/1!1!1!1!$ are multinomial coefficients). Now take the expectations on both sides, and use the fact that $\mathbf{E}(\hat{X}_k) = 0$ and that the $\hat{X}_k$'s are independent, to get

$$\mathbf{E}(T_n^4) = nm_4 + 3n(n-1)m_2^2,$$

where we denote $m_2 = \mathbf{E}(\hat{X}_k^2) = \mathbf{V}(X_k)$, $m_4 = \mathbf{E}(\hat{X}_k^4) < \infty$. This gives us the bound we wanted! In fact, all that matters is that $\mathbf{E}(T_n^4) \le Cn^2$ for some constant $C > 0$. Combining this with (8), we get that

$$\mathbf{P}(|S_n/n - \mu| > \epsilon) \le \frac{C}{n^2\epsilon^4}.$$

In particular, we have that $\sum_{n=1}^\infty \mathbf{P}(|S_n/n - \mu| > \epsilon) < \infty$, so by the Borel-Cantelli lemma, the probability of the event

$$A_\epsilon := \{|S_n/n - \mu| > \epsilon \text{ i.o.}\}$$

is 0. This is true for all $\epsilon > 0$, therefore the probability of

$$\{S_n/n \nrightarrow \mu\} \subset \bigcup_{\epsilon>0} A_\epsilon = \bigcup_{n=1}^\infty A_{1/n}$$

is also 0, since it is contained in a countable union of events of probability 0. □

Proof in the case of finite variance. We are slowly refining our techniques to weaken the assumptions required to prove the SLLN. The following nice proof manages to harness the Chebyshev variance bound (7) after all to deduce the theorem in the case of r.v.'s with finite variance. Observe that while the series of terms on the right-hand side of (7) diverges, if we restrict it to a subsequence $n_k = k^2$, it will become a *convergent* series: $\sum_{k=1}^\infty \sigma^2/k^2\epsilon^2 < \infty$. This implies, again by the Borel-Cantelli lemma, that

$$\mathbf{P}(|S_{n_k}/n_k - \mu| > \epsilon \text{ i.o.}) = 0$$

for all $\epsilon > 0$ (the "i.o." here refers to the "running index" k, not n). This implies as before that

$$\mathbf{P}\left(\frac{S_{n_k}}{n_k} \xrightarrow[k\to\infty]{\text{a.s.}} \mu\right) = 1.$$

So, while we have not shown almost sure convergence of the empirical averages to the true mean for all values of n, at least we have done so for the subsequence $n_k = k^2$. But this subsequence is relatively dense. In particular, if we show that in between elements of the subsequence $n_k = k^2$, the empirical average cannot fluctuate too wildly, then the theorem would follow. This will again follow from a combination of variance-based (i.e., Chebyshev) bounds and the Borel-Cantelli lemma. Fix some k, and take an integer n satisfying $n_k \le n < n_{k+1}$. How likely is the n-th empirical average to deviate significantly from the n_k-th empirical average? Using the notation $T_n = S_n - n\mu$ as before, we have

$$\begin{aligned}
\mathbf{P}\left(\left|\frac{T_n}{n} - \frac{T_{n_k}}{n_k}\right| > \epsilon\right) &= \mathbf{P}\left(\left|\frac{n_k T_n - n T_{n_k}}{n \cdot n_k}\right| > \epsilon\right) \\
\text{(by triangle ineq.)} \quad &\le \mathbf{P}\left(\left|\frac{T_n - T_{n_k}}{n_k}\right| > \epsilon/2\right) + \mathbf{P}\left(|T_n| \frac{n - n_k}{n \cdot n_k} > \epsilon/2\right) \\
\text{(by Chebyshev's ineq.)} \quad &\le \frac{4\mathrm{Var}(T_n - T_{n_k})}{\epsilon^2 n_k^2} + \frac{4\mathrm{Var}(T_n)(n-n_k)^2}{n^2 n_k^2} \\
&\le \frac{10k\sigma^2}{\epsilon^2 k^4} + \frac{20k^4\sigma^2}{k^8} < \frac{20\sigma^2}{k^3},
\end{aligned}$$

where we have denoted $\sigma^2 = \mathbf{V}(X_1)$.

The estimate that we obtained is valid for a single n. We are actually interested in the *maximal* fluctuation of the empirical averages S_n/n from S_{n_k}/n_k, when n ranges in the interval $[n_k, n_{k+1})$. By a simple subadditivity argument, usually referred to as a **union bound** (i.e., bounding the probability of a union of events by the sum of the probabilities, which is the most naive bound you can imagine), we get easily that

$$\begin{aligned}
\mathbf{P}\left(\max_{n_k \le n < n_{k+1}} \left|\frac{T_n}{n} - \frac{T_{n_k}}{n_k}\right| > \epsilon\right) &= \mathbf{P}\left(\bigcup_{n_k \le n < n_{k+1}} \left\{\left|\frac{T_n}{n} - \frac{T_{n_k}}{n_k}\right| > \epsilon\right\}\right) \\
&\le \sum_{n=n_k}^{n_{k+1}-1} \mathbf{P}\left(\left|\frac{T_n}{n} - \frac{T_{n_k}}{n_k}\right| > \epsilon\right) < \frac{20\sigma^2 \cdot 2k}{k^3} = \frac{40\sigma^2}{k^2}.
\end{aligned}$$

Once again, we have obtained as our bound the general term of a convergent series! Denoting

$$A_{k,\epsilon} = \left\{\max_{n_k \le n < n_{k+1}} \left|\frac{T_n}{n} - \frac{T_{n_k}}{n_k}\right| > \epsilon\right\},$$

by the Borel-Cantelli lemma this implies that for any $\epsilon > 0$, $\mathbf{P}(A_{k,\epsilon} \text{ i.o.}) = 0$ (with the "i.o." qualifier again referring to the running index k). What about the chance that this will happen for some $\epsilon > 0$? After all, there are so many ϵ's out there... But no, we also get that

$$\mathbf{P}\left(\bigcup_{\epsilon>0} \{A_{k,\epsilon} \text{ i.o.}\}\right) = \mathbf{P}\left(\bigcup_{m=1}^{\infty} \left\{A_{k,1/m} \text{ i.o. (w.r.t. } k)\right\}\right) = 0,$$

by subadditivity for countable unions and the fact that $A_{k,\epsilon} \subset A_{k,\epsilon'}$ if $\epsilon > \epsilon'$.

Finally, combining our two main results, namely that the two events

$$\begin{aligned} E_1 &= \left\{ \frac{S_{n_k}}{n_k} \xrightarrow[n\to\infty]{} \mu \right\}, \\ E_2 &= \left(\bigcup_{m=1}^{\infty} \left\{ \max_{n_k \le n < n_{k+1}} \left| \frac{T_n}{n} - \frac{T_{n_k}}{n_k} \right| > \frac{1}{m} \text{ for inifinitely many } k\text{'s} \right\} \right)^c \end{aligned}$$

both have probability 1, we get that their intersection $E_1 \cap E_2$ also has probability 1. But we have the event inclusion

$$E_1 \cap E_2 \subset \left\{ \frac{S_n}{n} \xrightarrow[n\to\infty]{} \mu \right\}$$

(if the conditions in both the events E_1, E_2 occur, then S_n/n must converge to μ), so this latter event also has probability 1. □

The above proof was a little involved, but is based on a few simple ideas: 1. Use Chebyshev bounds together with the Borel-Cantelli lemma. 2. Break up the event $\{S_n/n \to \mu\}$ in a clever way into events which can be managed using this technique, namely (in this case) the weaker convergence along the subsequence $n_k = k^2$, and separately the control of the fluctuations of the empirical averages in each range $[n_k, n_{k+1})$ between successive elements of the subsequence.

An extra benefit of the above proof is that it doesn't use the full power of the assumption that the X_n's are independent. In fact, everything is based on variance computations of sums of the X_k's, so the proof works equally well for *uncorrelated* random variables!

Theorem 9.3 (SLLN for uncorrelated r.v.'s with finite variance)**.** *If $X_1, X_2, \ldots$ are a sequence of uncorrelated and identically distributed r.v.'s with finite variance, then*

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow[n\to\infty]{a.s.} \mathbf{E}X_1.$$

Finally, we turn to the proof of the full SLLN for an i.i.d. sequence (Theorem 9.2), assuming only a finite first moment. Etemadi's 1981 proof presented in [Dur2010] is based on a similar idea of proving convergence first for a subsequence, and introduces another useful technique, that of **truncation**.

Proof of Theorem 9.2. First, observe that it is enough to prove the theorem in the case where $X_1 \geq 0$, since in the general case we may decompose each X_n into its positive and negative parts, and this gives two i.i.d. sequences of nonnegative r.v.'s. The validity of the SLLN for each of the two sequences implies the result for their difference.

Second, for each $n \geq 1$ let $Y_n = X_n \mathbf{1}_{\{X_n \leq n\}}$ ("X_n truncated at n"), and denote $T_n = Y_1 + Y_2 + \ldots + Y_n$. Since $\mathbf{E}X_1 < \infty$, by a homework exercise we know that $\sum_{n=1}^{\infty} \mathbf{P}(X_n > n) < \infty$, which by the Borel-Cantelli lemma implies that $\mathbf{P}(X_n > n \text{ i.o.}) = \mathbf{P}(X_n \neq Y_n \text{ i.o.}) = 0$. It follows that the event

$$\left\{ \sup_{n\to\infty} |S_n - T_n| = \infty \right\}$$

has probability 0, and therefore the even smaller event

$$\left\{ \limsup_{n\to\infty} \left| \frac{S_n}{n} - \frac{T_n}{n} \right| > 0 \right\} = \left\{ \limsup_{n\to\infty} \left| \left(\frac{S_n}{n} - \mu \right) - \frac{T_n - \mathbf{E}(T_n)}{n} \right| > 0 \right\},$$

has probability 0 (the equivalence between these last two events follows from the observation that $\mathbf{E}(Y_n) \uparrow \mu$ by the monotone/dominated convergence theorems, and therefore also $\mathbf{E}(T_n)/n \to \mu$ as $n \to \infty$ since $(n^{-1}\mathbf{E}(T_n))_{n=1}^{\infty}$ is the sequence of arithmetic averages of the $\mathbf{E}(Y_n)$'s). So we have shown that to prove the theorem, it is enough to prove that

$$\mathbf{P}\left(\frac{T_n - \mathbf{E}T_n}{n} \to 0 \right) = 1.$$

Now, to prove this, as before we establish a.s. convergence first along a subsequence n_k, this time taking $n_k = \lfloor \alpha^k \rfloor$ where $\alpha > 1$. (Here, $\lfloor x \rfloor$ denotes the "floor" function, namely the largest integer $\leq x$). This is done by again combining the Borel-Cantelli lemma with a Chebyshev variance bound, except that showing that the sum of the bounds gives a convergent series requires more work than before. We have

$$\mathbf{P}\left(\left| \frac{T_{n_k}}{n_k} - \frac{\mathbf{E}(T_{n_k})}{n_k} \right| > \epsilon \right) \leq \frac{\mathbf{V}(T_{n_k})}{\epsilon^2 n_k^2} = \frac{1}{\epsilon^2 n_k^2} \sum_{m=1}^{n_k} \mathbf{V}(Y_m).$$

Therefore

$$\begin{aligned}
\sum_{k=1}^{\infty} \mathbf{P}\left(\left|\frac{T_{n_k}}{n_k} - \frac{\mathbf{E}(T_{n_k})}{n_k}\right| > \epsilon\right) &\le \sum_{k=1}^{\infty} \frac{1}{\epsilon^2 n_k^2} \sum_{m=1}^{n_k} \mathbf{V}(Y_m) = \frac{1}{\epsilon^2} \sum_{m=1}^{\infty} \mathbf{V}(Y_m) \sum_{n_k \ge m} \frac{1}{n_k^2} \\
&\le \frac{1}{\epsilon^2} \sum_{m=1}^{\infty} \mathbf{V}(Y_m) \frac{1}{m^2} (1 + \alpha^{-2} + \alpha^{-4} + \alpha^{-6} + \ldots) \\
&= \frac{1}{\epsilon^2 (1-\alpha^{-2})} \sum_{m=1}^{\infty} \frac{\mathbf{V}(Y_m)}{m^2} \qquad (9)
\end{aligned}$$

Lemma 9.4. *We have*

$$\sum_{m=1}^{\infty} \frac{\mathbf{V}(Y_m)}{m^2} \le 4\mathbf{E}X_1 < \infty.$$

Proof. Using the formula $\mathbf{E}(Z) = \int_0^\infty \mathbf{P}(Z > x)\,dx$ that's valid for any nonnegative r.v. Z, we have that

$$\begin{aligned}
\sum_{m=1}^{\infty} \frac{\mathbf{V}(Y_m)}{m^2} &\le \sum_{m=1}^{\infty} \frac{1}{m^2} \mathbf{E}(Y_m^2) = \sum_{m=1}^{\infty} \frac{1}{m^2} \mathbf{E}\left(X_m^2 \mathbf{1}_{\{X_m \le m\}}\right) \\
&= \sum_{m=1}^{\infty} \frac{1}{m^2} \int_0^\infty \mathbf{P}\left(X_m^2 \mathbf{1}_{\{X_m \le m\}} > t\right) dt \\
&= \sum_{m=1}^{\infty} \frac{1}{m^2} \int_0^\infty \mathbf{P}\left(\sqrt{t} \le X_m \le m\right) dt \\
&= \sum_{m=1}^{\infty} \frac{1}{m^2} \int_0^{m^2} \mathbf{P}\left(\sqrt{t} \le X_1 \le m\right) dt \\
&\le \sum_{m=1}^{\infty} \frac{1}{m^2} \int_0^{m^2} \mathbf{P}\left(X_1 \ge \sqrt{t}\right) dt \\
&= \int_0^\infty \left(\sum_{m \ge \sqrt{t}} \frac{1}{m^2}\right) \mathbf{P}\left(X_1 \ge \sqrt{t}\right) dt \\
&\le \int_0^\infty \frac{2}{\sqrt{t}} \mathbf{P}\left(X_1 \ge \sqrt{t}\right) dt = 4 \int_0^\infty \mathbf{P}(X_1 \ge u) du = 4\mathbf{E}(X_1). \quad \square
\end{aligned}$$

It follows that the infinite sum on the left-hand side of (9) converges, and therefore by the Borel-Cantelli lemma, we have that

$$\mathbf{P}\left(\left|\frac{T_{n_k}}{n_k} - \frac{\mathbf{E}(T_{n_k})}{n_k}\right| > \epsilon \text{ infinitely often (w.r.t. } k)\right) = 0.$$

Since this is true for all $\epsilon > 0$, using the standard trick we get that

$$\mathbf{P}\left(\frac{T_{n_k}}{n_k} - \frac{\mathbf{E}(T_{n_k})}{n_k} \xrightarrow[k\to\infty]{} 0\right) = \mathbf{P}\left(\frac{T_{n_k}}{n_k} \xrightarrow[k\to\infty]{} \mu\right) = 1.$$

The last step is now to show that convergence along this subsequence forces $(T_n - \mathbf{E}T_n)/n$ to behave well between successive n_k's. Observe that if $n_k \le n < n_{k+1}$ then

$$\frac{T_{n_k}}{n_{k+1}} \le \frac{T_n}{n} \le \frac{T_{n_{k+1}}}{n_k}.$$

Since $n_k = \lfloor \alpha^k \rfloor$, in the limit this gives that almost surely the bounds

$$\frac{1}{\alpha}\mu \le \liminf_{n\to\infty} \frac{T_n}{n} \le \limsup_{n\to\infty} \frac{T_n}{n} \le \alpha\mu$$

hold. Since $\alpha > 1$ was arbitrary, the intersection of these events for $\alpha = 1+1/d$, $d = 1, 2, \ldots$, implies that

$$\mathbf{P}\left(\frac{T_n}{n} \to \mu\right) = 1,$$

which finishes the proof of Theorem 9.2. □

Chapter 10: Applications and further examples

10.1 The Weierstrass approximation theorem

As an application of WLLN (or, rather, of Chebyshev's inequality), we prove the following theorem in analysis, which seems to have no connection to probability whatsoever.

Theorem 10.1 (The Weierstrass approximation theorem)**.** *If $f : [0,1] \to \mathbb{R}$ is a continuous function, then f can be uniformly approximated by polynomials. That is, for any $\epsilon > 0$ there exists a polynomial p such that $||f - p|| := \max_{x\in[0,1]} |f(x) - p(x)| < \epsilon$.*

Proof. Let $f : [0,1] \to \mathbb{R}$ be a continuous function. Define the sequence of **Bernstein polynomials of** f by

$$B_n(x) = B_n^f(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k} x^k (1-x)^{n-k}.$$

We claim that $||B_n - f|| \to 0$ as $n \to \infty$, which will prove the theorem. We will prove this by giving a probabilistic interpretation to B_n. Indeed, fix $x \in [0,1]$, and let $X_1, X_2, \ldots$ be a sequence of i.i.d. r.v.'s with the Bernoulli distribution Binom$(1,x)$. Denote $S_n = \sum_{k=1}^{\infty} X_k$. Then it is easy to see that

$$B_n(x) = \mathbf{E}_x\left[f\left(\frac{S_n}{n}\right)\right],$$

where the notation "$\mathbf{E}_x$" just means taking the expectation, while at the same time highlighting the fact that the distribution of S_n depends on the parameter x.

Now the idea is that since the law of large numbers implies that S_n/n is with high probability close to its mean x, and f is a continuous function, then $f(S_n/n)$ should with high probability be close to $f(x)$, and therefore also the average value of $f(S_n/n)$, namely $B_n(x)$, should be close to $f(x)$, which is what we want. However, we want to make this claim uniformly in x, so instead of invoking the WLLN (Theorem 9.1) we have to go back and look "under the hood" at Chebyshev's inequality which we used to prove it. We estimate $|B_n(x) - f(x)|$ as follows. Fix some arbitrary $\epsilon > 0$. Let $\delta > 0$ be such that for any $u, v \in [0,1]$, if $|u - v| < \delta$ then $|f(u) - f(v)| < \epsilon$ (this is guaranteed to exist because f is uniformly continuous on $[0,1]$). Then

$$\begin{aligned}|B_n(x)-f(x)| &= \left|\mathbf{E}_x f\left(\frac{S_n}{n}\right)-f(x)\right| \le \mathbf{E}_x\left|f\left(\frac{S_n}{n}\right)-f(x)\right| \\ &= \mathbf{E}_x\left[\left|f\left(\frac{S_n}{n}\right)-f(x)\right|\mathbf{1}_{\{|\frac{S_n}{n}-x|>\delta\}}\right]+\mathbf{E}_x\left[\left|f\left(\frac{S_n}{n}\right)-f(x)\right|\mathbf{1}_{\{|\frac{S_n}{n}-x|\le\delta\}}\right]\end{aligned}$$

In this last expression, each of the two expectations is small for a different reason. The second expectation is bounded by ϵ, since on the event that $|S_n/n-x|\le\delta$, we have that $|f(S_n/n)-f(x)|<\epsilon$. To bound the first expectation, denote $M=||f||:=\max_{x\in[0,1]}|f(x)|<\infty$. Then, by bounding the difference of f-values by $2M$ and then using Chebyshev's inequality, we get

$$\mathbf{E}_x\left[\left|f\left(\frac{S_n}{n}\right)-f(x)\right|\mathbf{1}_{\{|\frac{S_n}{n}-x|>\epsilon\}}\right]\le 2M\cdot\mathbf{P}_x\left(\left|\frac{S_n}{n}-x\right|>\epsilon\right)\le\frac{2M\sigma^2(X_1)}{n\epsilon^2}=\frac{2Mx(1-x)}{n\epsilon^2}$$

This bound converges to 0, not just for a single x but (fortunately for us) uniformly in $x\in[0,1]$, since it is bounded from above by $M/(2n\epsilon^2)$. So we have shown that

$$||B_n-f||=\max_{x\in[0,1]}|B_n(x)-f(x)|\le\epsilon+\frac{M}{2n\epsilon^2}.$$

It follows that $\limsup_{n\to\infty}||B_n-f||\le\epsilon$, and since ϵ was an arbitrary positive number the result follows. □

10.2 Infinite expectations and triangular arrays

After treating the "classical" case of an i.i.d. sequence with finite expectations, let's turn to slightly more exotic situations. First, what happens if the expectation is infinite? For the strong law, the following result shows that we have no hope of having convergence in a meaningful sense.

Theorem 10.2 (Converse to SLLN)**.** *If $X_1, X_2, \ldots,$ is an i.i.d. sequence of r.v.'s with $\mathbf{E}|X_1|=\infty$, then*

$$\mathbf{P}\left(\exists\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^n X_k\right)=0.$$

Proof. See homework. □

What about the weak law? It turns out that a weak law can still hold in certain situations with infinite expectations, although one has to consider a generalized empirical average where the sum of the first n samples is divided by a quantity growing faster than n. Rather than develop a complete theory, we will consider the particular example of the **St. Petersburg Lottery**. In this example, the winning in a single round of lottery is an integer-valued random variable with the following distribution:

$$\mathbf{P}(X = 2^k) = 2^{-k}, \qquad k = 1, 2, 3, \ldots$$

Let $X_1, X_2, \ldots,$ be an i.i.d. sequence with the same distribution, and let $S_n = \sum_{k=1}^n X_k$. How much should you agree to pay to be allowed to play this lottery n times? (Even the seemingly simple case $n = 1$ of this question has been a subject of quite some debate by economists! See `http://en.wikipedia.org/wiki/St._Petersburg_paradox`). In other words, how big should we expect S_n to be, *with probability close to 1*?

Theorem 10.3.

$$\frac{S_n}{n \log_2 n} \xrightarrow[n\to\infty]{\mathbf{P}} 1.$$

In other words, to be allowed to pay the game n times when n is large (and assuming the payoff is in dollars), it may be considered reasonable to pay exactly (or even better, slightly less than) $\log_2 n$ dollars *per round played.* For example, if $n = 1024$ you would be paying \$10 per round.

Proof. The proof uses a truncation idea similar to the one we saw in the proof of SLLN, except that for each n we will truncate the first n variables at a level which is a function of n. Denote $b_n = n \log_2 n$, $Y_{n,k} = X_k \mathbf{1}_{\{X_k < b_n\}}$, $T_n = \sum_{k=1}^n Y_{n,k}$, and $a_n = \mathbf{E}(T_n)$. We will prove that

$$\frac{S_n - a_n}{b_n} \xrightarrow[n\to\infty]{\mathbf{P}} 0. \tag{10}$$

First we check that this is enough, by estimating a_n:

$$a_n \;=\; \sum_{k=1}^n \mathbf{E}(Y_{n,k}) = \sum_{k=1}^n \mathbf{E}(X_k \mathbf{1}_{\{X_k < b_n\}}) = \sum_{k=1}^n \left(\frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 4 + \frac{1}{8} \cdot 8 + \ldots + \frac{1}{2^{m_n}} \cdot 2^{m_n} \right)$$

where m_n is the largest integer such that $2^{m_n} \le b_n$, or in other words $m_n = \lfloor \log_2 n + \log_2 \log_2 n \rfloor$, which gives

$$a_n = \sum_{k=1}^n \lfloor \log_2 n + \log_2 \log_2 n \rfloor = n \log_2 n + O(n \log_2 \log_2 n),$$

so a_n indeed behaves like $n \log_2 n$ up to first-order asymptotics. Now to prove (10), note that for any $\epsilon > 0$,

$$\mathbf{P}\left(\left|\frac{S_n - a_n}{b_n}\right| > \epsilon\right) \le \mathbf{P}(T_n \ne S_n) + \mathbf{P}\left(\left|\frac{T_n - a_n}{b_n}\right| > \epsilon\right). \tag{11}$$

In this bound, the first term is bounded by

$$\sum_{k=1}^{n} \mathbf{P}(Y_{n,k} \ne X_k) = \sum_{k=1}^{n} \mathbf{P}(X_k > b_n) \le \sum_{k=1}^{n} \frac{2}{b_n} \le \frac{2}{\log_2 n} \xrightarrow[n\to\infty]{} 0.$$

To bound the second term, use Chebyshev's inequality and the fact that $\mathbf{V}(Y_{n,k}) \le \mathbf{E}(Y_{n,k}^2)$ to write

$$\begin{aligned} \mathbf{P}\left(\left|\frac{T_n - a_n}{b_n}\right| > \epsilon\right) &\le \frac{\mathbf{V}(T_n)}{\epsilon^2 b_n^2} \le \frac{1}{\epsilon^2 b_n^2} \sum_{k=1}^{n} \mathbf{E}(Y_{n,k}^2) \\ &\le \frac{1}{\epsilon^2 b_n^2} \sum_{k=1}^{n} \left(\frac{1}{2} \cdot 2^2 + \frac{1}{4} \cdot 4^2 + \frac{1}{8} \cdot 8^2 + \ldots + \frac{1}{2^{m_n}} \cdot 2^{2m_n}\right) \\ &\le \frac{1}{\epsilon^2 b_n^2} \sum_{k=1}^{n} 2 \cdot 2^{m_n} \le \frac{2}{\epsilon^2 b_n^2} \sum_{k=1}^{n} b_n = \frac{2}{\epsilon^2 \log_2 n} \xrightarrow[n\to\infty]{} 0. \end{aligned}$$

We conclude that the left-hand side of (11) converges to 0, as $n \to \infty$, which finishes the proof. □

Another twist on laws of large numbers comes when we replace the notion of an i.i.d. sequence by a more general notion of a **triangular array**. In this case, for each n we have a sequence $X_{n,1}, X_{n,2}, \ldots, X_{n,n}$ of independent, but not necessarily identically distributed, and we denote $S_n = \sum_{k=1}^{n} X_{n,k}$ – this is the sum of the samples in the n-th experiment. Here, for each n there could be a separate experiment involving n different r.v.'s, and the r.v.'s $X_{n,k}$ and $X_{m,j}$ for $n \ne m$ are not even assumed to be defined on the same probability space, let alone to be independent of each other.

Again, instead of giving general conditions for a law of large numbers to hold, consider the following example of the so-called **coupon collector's problem**: A brand of breakfast cereals comes with a small toy chosen uniformly at random from a set of n possible kinds of toys. A collector will buy more boxes of cereals until she has collected all n different toys. Denote by T_n the number of boxes she ends up buying. What can we say about the size of T_n? Fortunately, we can represent it as a sum of independent r.v.'s, by writing

$$T_n = X_{n,1} + X_{n,2} + X_{n,3} + \ldots + X_{n,n},$$

where

$$\begin{aligned}
X_{n,1} &= \text{number of boxes purchased to get one kind of toy} \;=\; 1, \\
X_{n,2} &= \text{number of boxes purchased after having one toy to get a different kind}, \\
X_{n,3} &= \text{number of boxes purchased after having two kinds of toys to get a third kind}, \\
&\vdots \\
X_{n,n} &= \text{number of boxes purchased after having } n-1 \text{ kinds of toys to get the last kind.}
\end{aligned}$$

Clearly these r.v.'s are independent. Furthermore, $X_{n,k}$ is a geometric r.v. with parameter $p_{n,k} = (n-k+1)/n$. This gives us that

$$\mathbf{E}(T_n) = \sum_{k=1}^{n} \mathbf{E}(X_{n,k}) = \sum_{k=1}^{n} \frac{n}{n-k+1} = n\left(\frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{2} + \frac{1}{1}\right) = nH_n,$$

where $H_n = \sum_{k=1}^{n} 1/k$ is the *n-th harmonic number*, and

$$\mathbf{V}(T_n) = \sum_{k=1}^{n} \mathbf{V}(X_{n,k}) = \sum_{k=1}^{n} \frac{k-1}{n}\left(\frac{n}{n-k+1}\right)^2 \le n^2 \sum_{k=1}^{n} \frac{1}{k^2} \le 2n^2$$

(in this example, we only need a bound for $\mathbf{V}(T_n)$, but it is possible also to get more precise asymptotics for this quantity). It follows using Chebyshev's inequality that for each $\epsilon > 0$,

$$\mathbf{P}\left(\left|\frac{T_n - nH_n}{n\log n}\right| > \epsilon\right) \le \frac{\mathbf{V}(T_n)}{\epsilon^2 n^2 (\log n)^2} \le \frac{2}{\epsilon^2 (\log n)^2} \xrightarrow[n\to\infty]{} 0,$$

so $(T_n - nH_n)/(n\log n)$ converges in probability to 0, and therefore we get that

$$\frac{T_n}{n\log n} \xrightarrow[n\to\infty]{\mathbf{P}} 1.$$

The lesson to be learned from the above examples is that some naturally-occurring problems in real life lead to more complicated situations than can be modeled with an i.i.d. sequence with finite mean; but often such problems can be analyzed anyway using the same ideas and techniques that we developed. In probability textbooks you can find a general treatment of various conditions under which a triangular array of independent random variables satisfies a (strong or weak) law of large numbers.

10.3 Random series of independent samples

We now look at the related topic of infinite series of independent r.v.'s. When can such a series be said to converge? A key technical result that will help us answer this question in some cases is the following beautiful inequality due to Kolmogorov.

Theorem 10.4 (Kolmogorov's maximal inequality)**.** *Assume that $X_1, X_2, \ldots, X_n$ are independent r.v.'s with finite variances, and let $S_k = \sum_{j=1}^k X_j$. Then*

$$\mathbf{P}\left(\max_{1\le k\le n} |S_k - \mathbf{E}(S_k)| > t\right) \le \frac{\mathbf{V}(S_n)}{t^2} = \frac{\sum_{k=1}^n \sigma^2(X_k)}{t^2}.$$

Before we start the proof, note that the bound on the right-hand side is the usual variance bound that follows from Chebyshev's inequality; except that the event on the left-hand side whose probability this quantity bounds is a *much bigger* event than the usual deviation event $\{|S_n - \mathbf{E}(S_n)| > t\}$ for which we know the Chebyshev bound holds!

Proof. We may assume without loss of generality that $\mathbf{E}(X_k) = 0$ for all k. Denote

$$A = \left\{\max_{1\le k\le n} |S_k| > t\right\},$$

and define events $A_1, A_2, \ldots, A_n$ by

$$A_k = \left\{|S_k| \ge t, \quad \max_{1\le j<k} |S_k| < t\right\}.$$

In words, A_k is the event that the sequence of cumulative sums $(S_j)_{j=1}^n$ exceeded t in absolute value for the first time at time k. Note that these events are disjoint and their union is the event A whose probability we are trying to bound. As a consequence, we can lower-bound the variance $\mathbf{V}(S_n) = \mathbf{E}(S_n^2)$ of S_n, as follows:

$$\begin{aligned}\mathbf{E}(S_n^2) &\ge \mathbf{E}(S_n^2 1_A) = \sum_{k=1}^n \mathbf{E}\left(S_n^2 1_{A_k}\right) = \sum_{k=1}^n \mathbf{E}\left[(S_k + (S_n - S_k))^2 1_{A_k}\right] \\ &= \sum_{k=1}^n \Big[\mathbf{E}(S_k^2 1_{A_k}) + \mathbf{E}\left[(S_n - S_k)^2 1_{A_k}\right] + 2\mathbf{E}\left[(S_k 1_{A_k})(S_n - S - k)\right]\Big].\end{aligned}$$

In this last expression, the terms $\mathbf{E}\left[(S_k 1_{A_k})(S_n - S - k)\right]$ are equal to 0, since $S_k 1_{A_k}$ is a random variable that depends only on $X_1, \ldots, X_k$ and hence independent of $S_n - S_k$, which depends only on the values of $X_{k+1}, \ldots, X_n$, which causes the expectation of their product to be equal to the product of the expectations, which is 0. Furthermore, the middle terms $\mathbf{E}\left[(S_n - S_k)^2 1_{A_k}\right]$ are all nonnegative, and each of the first terms $\mathbf{E}(S_k^2 1_{A_k})$ satisfies

$$\mathbf{E}(S_k^2 1_{A_k}) \geq \mathbf{E}(t^2 1_{A_k}) = t^2 \mathbf{P}(A_k),$$

since on the event A_k we know that S_k^2 is at least t^2 (look again at the definition of A_k). Combining these observations, we get that

$$\mathbf{V}(S_n) \geq t^2 \sum_{k=1}^{n} \mathbf{P}(A_k) = t^2 \mathbf{P}(A),$$

which is exactly the claim that was to be proved. □

As a corollary, we get a result on convergence of random series.

Theorem 10.5. *Let $X_1, X_2, \ldots$ be a sequence of independent r.v.'s such that $\mathbf{E}(X_n) = 0$ for all n, and assume that $\sum_{n=1}^{\infty} \mathbf{V}(X_n) < \infty$. Then the random series $\sum_{n=1}^{\infty} X_n$ converges almost surely.*

Proof. Denote as usual $S_n = \sum_{k=1}^{n} X_k$. We have the following equality of events:

$$\begin{aligned} \left\{ \sum_{n=1}^{\infty} X_n \text{ converges} \right\} &= \left\{ \left(\sum_{n=1}^{N} X_n \right)_{N \geq 1} \text{ is a Cauchy sequence} \right\} \\ &= \bigcap_{\epsilon > 0} \bigcup_{N \geq 1} \bigcap_{n \geq N} \Big\{ |S_n - S_N| < \epsilon \Big\}. \end{aligned}$$

Or, put differently, we can look at the complement of this event and represent it as

$$\begin{aligned} \left\{ \sum_{n=1}^{\infty} X_n \text{ does not converge} \right\} &= \bigcup_{\epsilon > 0} \bigcap_{N \geq 1} \bigcup_{n \geq N} \Big\{ |S_n - S_N| \geq \epsilon \Big\} \\ &= \bigcup_{\epsilon > 0} \bigcap_{N \geq 1} \left\{ \sup_{n \geq N} |S_n - S_N| \geq \epsilon \right\}. \end{aligned}$$

This form is exactly suitable for an application of the Kolmogorov's maximal inequality, except that here we have an infinite sequence of partial sums instead of a finite maximum.

However, by the "continuity from below" property of probability measures, we see that it does not matter. More precisely, for any $\epsilon > 0$ and $N \geq 1$, we have

$$\begin{aligned}\mathbf{P}\left(\sup_{n\geq N}|S_n - S_N| \geq \epsilon\right) &= \lim_{M\to\infty}\mathbf{P}\left(\sup_{N\leq n\leq M}|S_n - S_N| \geq \epsilon\right) \leq \lim_{M\to\infty}\frac{\mathbf{V}(S_M - S_N)}{\epsilon^2} \\ &= \frac{1}{\epsilon^2}\sum_{n=N}^{\infty}\mathbf{V}(X_n).\end{aligned}$$

Therefore also

$$\begin{aligned}\mathbf{P}\left(\bigcap_{N\geq 1}\left\{\sup_{n\geq N}|S_n - S_N| \geq \epsilon\right\}\right) &\leq \inf_{N\geq 1}\mathbf{P}\left(\sup_{n\geq N}|S_n - S_N| \geq \epsilon\right) \\ &\leq \inf_{N\geq 1}\frac{1}{\epsilon^2}\sum_{n=N}^{\infty}\mathbf{V}(X_n) = 0,\end{aligned}$$

because of our assumption that the sum of the variances converges. Finally, this is true for all $\epsilon > 0$, so by the usual trick of replacing an uncountably-infinite intersection by a countable one (provided that the particular form of the event in question warrants this!), we get the claim that

$$\mathbf{P}\left(\sum_{n=1}^{\infty}X_n \text{ does not converge}\right) = 0.$$

□

One could ask whether the sufficient condition given by the theorem above is also necessary (it is). More generally, what happens for random variables with non-zero expectations? What happens for r.v.'s with infinite expectations, or infinite variances? Kolmogorov formulated a general theorem that gives a necessary and sufficient condition for a series of independent random variables to converge almost surely.

Theorem 10.6 (The Kolmogorov three-series theorem)**.** *If $X_1, X_2, \ldots$ is a sequence of independent random variables, then the random series $\sum_{n=1}^{\infty} X_n$ converges almost surely if and only if the following three conditions hold:*

1. $\sum_{n=1}^{\infty}\mathbf{P}(|X_n| > 1) < \infty$.

2. *The series* $\sum_{n=1}^{\infty}\mathbf{E}(X_n \mathbf{1}_{\{|X_n|\leq 1\}})$ *converges.*

3. $\sum_{n=1}^{\infty}\mathbf{V}(X_n \mathbf{1}_{\{|X_n|\leq 1\}}) < \infty$.

If one of the conditions does not hold, then series $\sum X_n$ diverges almost surely.

We postpone the proof of Theorem 10.6 until later; we will prove it in Section 16.2 as an application of a generalized version of the central limit theorem. Note that since the convergence of the series $\sum X_n$ is equivalent to the convergence of $\sum aX_n$ for any constant a, the value 1 chosen for the truncation of X_n in the theorem is arbitrary and can be replaced by any other constant.

Example 10.7. Let $(c_n)_{n=1}^{\infty}$ be a sequence of real numbers. Consider the **series with random signs** associated with the sequence (c_n), which we denote $\sum_{n=1}^{\infty} \pm c_n$, and which more precisely represents the series

$$\sum_{n=1}^{\infty} c_n X_n,$$

where $X_1, X_2, \ldots$ is a sequence of i.i.d. r.v.'s taking the values $-1, +1$ with respective probabilities $1/2, 1/2$. By Theorem 10.5 it follows that if $\sum_{n=1}^{\infty} c_n^2 < \infty$ then the series $\sum_{n=1}^{\infty} \pm c_n$ converges almost surely. From Theorem 10.6, one can check easily that this condition is also necessary, in other words that the series with random series converges a.s. if and only if the series of squares $\sum c_n^2$ converges. Thus, for example, the **harmonic series with random signs** $\sum \pm \frac{1}{n}$ converges a.s., but the analogous series of square root reciprocals $\sum_n \pm \frac{1}{\sqrt{n}}$ diverges a.s. Compare this to the series with *alternating* signs $\sum_n \frac{(-1)^n}{n}$ and $\sum_n \frac{(-1)^n}{\sqrt{n}}$, both of which are known to converge! Try to develop your intuition by thinking about the reasons why the behavior for series with alternating signs and that for random signs is not the same.

Chapter 11: The Central Limit Theorem, Stirling's formula and the de Moivre-Laplace theorem

Our goal in the next few chapters will be to formulate and prove one of the fundamental results of probability theory, known as the Central Limit Theorem. Roughly speaking, this theorem establishes the normal distribution as the universal limiting law for the distribution of sums of independent and identically distributed random variables, and therefore explains the central role that the normal distribution plays in probability theory and statistics, and why it appears in virtually all applied sciences and is applicable to the study of many real-life phenomena.

We start with a motivating example that was also historically the first instance in which the phenomenon that came to be known as the Central Limit Theorem was observed. Let $X_1, X_2, \ldots$ be an i.i.d. of $\text{Binom}(1, p)$ random variables, and let $S_n = \sum_{k=1}^n X_k$, a r.v. with distribution $\text{Binom}(n, p)$.

Theorem 11.1 (The de Moivre-Laplace theorem). *For any $t \in \mathbb{R}$,*

$$\mathbf{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \le t\right) \xrightarrow[n\to\infty]{} \Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-x^2/2}\, dx.$$

Since this is such a concrete example, the proof will simply require us to estimate a sum of the form $\sum_{0 \le k \le t} \binom{n}{k} p^k (1-p)^{n-k}$. Knowing how to estimate such sums is a useful skill in its own right. Since the binomial coefficients are involved, we also need some preparation related to Stirling's formula.

Lemma 11.2. *The limit $C = \lim_{n\to\infty} \frac{n!}{\sqrt{n}(n/e)^n}$ exists.*

Proof.

$$\begin{aligned} \log n! &= \sum_{k=1}^n \log k = \sum_{=1}^n \int_1^k \frac{dx}{x} = \int_1^n \frac{n - \lfloor x \rfloor}{x}\, dx \\ &= \int_1^n \frac{n + \frac{1}{2} + (\{x\} - \frac{1}{2}) - x}{x}\, dx = (n + 1/2)\log n - n + 1 + \int_1^n \frac{\{x\} - \frac{1}{2}}{x}\, dx \\ &= (n+1/2)\log n - n + 1 + \int_1^\infty \frac{\{x\} - \frac{1}{2}}{x}\, dx + o(1), \end{aligned}$$

where the last integral converges because $\int_1^t(\{x\}-\frac{1}{2})dx$ is bounded and $1/x$ decreates monotonically to 0 as $x \to \infty$. □

Note that an easy consequence of Lemma 11.2 is that $\binom{2n}{n} = (1+o(1))2^{2n}/C\sqrt{n/2}$. We shall now use this to find the value of C.

Lemma 11.3. *Let $f : \mathbb{R} \to \mathbb{R}$ be an $n+1$ times continuously-differentiable function. Then for all $x \in \mathbb{R}$, we have*

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2}x^2 + \ldots + \frac{f^{(n)}(0)}{n!}x^n + R_n(x),$$

where

$$R_n(x) = \frac{1}{n!}\int_0^x f^{(n+1)}(t)(x-t)^n\,dt.$$

Proof. This follows by induction on n, using integration by parts. □

Lemma 11.4. $C = \sqrt{2\pi}$.

Proof. Apply Lemma 11.3 with $f(x) = (1+x)^{2n+1}$ to compute $R_n(1)$:

$$\begin{aligned}\frac{1}{2^{2n+1}}R_n(1) &= \frac{1}{2^{2n+1}}\cdot\frac{1}{n!}\int_0^1 (2n+1)(2n)\cdots(n+1)(1+t)^n(1-t)^n\,dt \\ &= \frac{2\binom{2n}{n}}{2^{2n+1}}(n+\frac{1}{2})\int_0^1(1-t^2)^n\,dt = \frac{\binom{2n}{n}\sqrt{n}}{2^{2n}}(1+\frac{1}{2n})\int_0^{\sqrt{n}}\left(1-\frac{u^2}{n}\right)^n du \\ &\xrightarrow[n\to\infty]{} \frac{\sqrt{2}}{C}\int_0^\infty e^{-u^2}\,du = \frac{\sqrt{2}}{C}\cdot\frac{\sqrt{\pi}}{2}.\end{aligned}$$

The convergence of the integrals is justified by the fact that $(1-u^2/n)^n \le e^{-u^2}$ for all $0 \le u \le \sqrt{n}$, and $(1-u^2/n)^n \to e^{-u^2}$ as $n \to \infty$, uniformly on compact intervals. To finish the proof, note that

$$\frac{1}{2^{2n+1}}R_n(1) = \sum_{n<k\le 2n+1}\frac{\binom{2n+1}{k}}{2^{2n+1}} = \frac{1}{2}$$

(this is the probability that a Binom$(2n+1, 1/2)$ random variable takes a value $> n$). Therefore $C = \sqrt{2\pi}$, as claimed. □

Corollary 11.5 (Stirling's formula)**.** $\lim_{n\to\infty}\frac{n!}{\sqrt{2\pi n}(n/e)^n} = 1$.

Note that the proof is based on computing $\mathbf{P}(S_{2n+1} > n)$ in two different ways, when $S_{2n+1} \sim \mathrm{Binom}(2n+1, 1/2)$. This is just the special case $p = 1/2, t = 0$ of Theorem 11.1. In this very special case, by symmetry the probability is equal to 1/2; on the other hand, Lemma 11.3 enables us to relate this to the asymptotic behavior of $n!$ and to (half of) the gaussian integral $\int_{-\infty}^{\infty} e^{-x^2}\,dx$. The evaluation of the constant C in Stirling's formula is the part that is attributed to James Stirling. The form that appears in Lemma 11.2 is due to Abraham de Moivre (1733).

With this preparation, it is now possible to apply the same technique to prove Theorem 11.1. Instead of the function $f(x) = (1+x)^{2n+1}$, take the function $g(x) = ((1-p)+px)^n = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} x^k$, and compute the remainder $R_k(1)$ of the Taylor expansion of g, where $k \approx np + t\sqrt{np(1-p)}$. This should converge to $1 - \Phi(t)$, and indeed, this follows without too much difficulty from Lemma 11.3. The computation is left as an exercise. We also sketch another way of proving Theorem 11.1 by directly approximating the probabilities $\binom{n}{k} p^k (1-p)^{n-k}$ by Gaussian densities.

Sketch of Proof of Theorem 11.1. Denote $q = 1 - p$. For a large n, let k be approximately equal to $np + t\sqrt{npq}$, and use Stirling's formula to estimate the probability $\mathbf{P}(S_n = k)$, as follows:

$$\begin{aligned}
\mathbf{P}(S_n = k) &= \binom{n}{k} p^k q^{n-k} = (1+o(1)) \frac{\sqrt{2\pi n}(n/e)^n p^k q^{n-k}}{\sqrt{2\pi k}(k/e)^k \sqrt{2\pi(n-k)}((n-k)/e)^{n-k}} \\
&= \frac{1+o(1)}{\sqrt{2\pi npq}} \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k} \\
&= \frac{1+o(1)}{\sqrt{2\pi npq}} \left(1 + \frac{t\sqrt{q}}{\sqrt{np}}\right)^{-k} \left(1 - \frac{t\sqrt{p}}{\sqrt{nq}}\right)^{-(n-k)}.
\end{aligned}$$

Taking the logarithm of the product of the last two factors, using the facts that $k \approx np + t\sqrt{npq}$, $n - k \approx nq - t\sqrt{npq}$, and that $\log(1+x) = x - x^2/2 + O(x^3)$ when $x \to 0$, we see that

$$\log\left[\left(1 + \frac{t\sqrt{q}}{\sqrt{np}}\right)^{-k} \left(1 - \frac{t\sqrt{p}}{\sqrt{nq}}\right)^{-(n-k)}\right]$$

$$\begin{aligned}
&= -(np + t\sqrt{npq})\log\left(1+\frac{t\sqrt{q}}{\sqrt{np}}\right) - (nq - t\sqrt{npq})\log\left(1-\frac{t\sqrt{p}}{nq}\right)\\
&= -(np + t\sqrt{npq})\left(\frac{t\sqrt{q}}{\sqrt{np}} - \frac{t^2 q}{2np}\right) - (nq - t\sqrt{npq})\left(-\frac{t\sqrt{p}}{\sqrt{nq}} - \frac{t^2 p}{2nq}\right) + O\left(\frac{t^3}{\sqrt{n}}\right)\\
&= -t\sqrt{npq} - t^2 q + \frac{t^2 q}{2} + t\sqrt{npq} - t^2 p + \frac{t^2 p}{2} + O\left(\frac{t^3}{\sqrt{n}}\right)\\
&= -\frac{t^2}{2} + O\left(\frac{t^3}{\sqrt{n}}\right).
\end{aligned}$$

It follows that

$$\mathbf{P}(S_n = k) = \frac{1+o(1)}{\sqrt{2\pi npq}} e^{-t^2/2}$$

In other words, the individual probabilities for S_n approximate a normal density! From here, it is not too hard to show that the probability

$$\mathbf{P}\left(a \le \frac{S_n - np}{\sqrt{npq}} \le b\right) = \sum_{np + a\sqrt{npq} \le k \le np + b\sqrt{npq}} \mathbf{P}(S_n = k)$$

is approximately a Riemann sum for the integral $(2\pi)^{-1/2}\int_a^b e^{-x^2/2}\,dx = \Phi(b) - \Phi(a)$. In fact, this is true since for a, b fixed and k ranging between $np + a\sqrt{npq}$ and $np + b\sqrt{npq}$, the error concealed by the $o(1)$ term is uniformly small (smaller than any $\epsilon > 0$, say, when n is sufficiently large), since this error term originates with three applications of Stirling's approximation formula (for $n!$, for $k!$ and for $(n-k)!$) followed by the log function second-order Taylor expansion above. □

One lesson that can be learned from this proof is that doing computations for specific distributions can be *messy*! So we might be better off looking for more general, and therefore more conceptual, techniques for proving convergence to the normal distribution, that require less explicit computations; fortunately such techniques exist, and will lead us to the much more general central limit theorem.

Chapter 12: Convergence in distribution

12.1 Definition

Since we will be talking about convergence of the distribution of random variables to the normal distribution, it makes sense to develop the general theory of convergence of distributions to a limiting distribution.

Definition 12.1. *Let $(F_n)_{n=1}^{\infty}$ be a sequence of distribution functions. We say that F_n* ***converges to a limiting distribution function*** *F, and denote this by $F_n \implies F$, if $F_n(x) \to F(x)$ as $n \to \infty$ for any $x \in \mathbb{R}$ which is a continuity point of F. If $X, (X_n)_{n=1}^{\infty}$ are random variables, we say that X_n* ***converges in distribution to*** *X (or, interchangeably,* ***converges in distribution to*** *F_X) if $F_{X_n} \implies F_X$.*

This definition, which may seem unnatural at first sight, will become more reasonable after we prove the following lemma.

Lemma 12.2. *The following are equivalent:*

1. *$X_n \implies X$.*
2. *$\mathbf{E}f(X_n) \xrightarrow[n\to\infty]{} \mathbf{E}f(X)$ for any bounded continuous function $f : \mathbb{R} \to \mathbb{R}$.*
3. *There exists a r.v. Y and a sequence $(Y_n)_{n=1}^{\infty}$ of r.v.'s, all defined on some probability space $(\Omega, \mathcal{F}, \mathbf{P})$ such that $Y_n \to Y$ a.s., Y is equal in distribution to X, and each Y_n is equal in distribution to the respective X_n.*

Proof. Proof that $2 \implies 1$: Assume that $\mathbf{E}f(X_n) \xrightarrow[n\to\infty]{} \mathbf{E}f(X)$ for any bounded continuous function $f : \mathbb{R} \to \mathbb{R}$, and fix $x \in \mathbb{R}$. For any $t \in \mathbb{R}$ and $\epsilon > 0$, define a function $g_{t,\epsilon} : \mathbb{R} \to \mathbb{R}$ by

$$g_{t,\epsilon}(u) = \begin{cases} 1 & u < t, \\ \frac{t-u+\epsilon}{\epsilon} & u \le t \le t+\epsilon, \\ 0 & u > t+\epsilon. \end{cases}$$

Then we have that

$$\mathbf{E}(g_{x-\epsilon,\epsilon}(X_n)) \le F_{X_n}(x) = \mathbf{E}(\mathbf{1}_{(-\infty,x]}(X_n)) \le \mathbf{E}(g_{x,\epsilon}(X_n))$$

Letting $n \to \infty$ gives the chain of inequalities

$$F_X(x-\epsilon) \le \mathbf{E}(g_{x-\epsilon,x}(X)) \le \liminf_{n\to\infty} F_{X_n}(x) \le \limsup_{n\to\infty} F_{X_n}(x) \le \mathbf{E}(g_{x,\epsilon}(X)) \le F_X(x+\epsilon).$$

Now if x is a point of continuity of F_X, letting $\epsilon \downarrow 0$ gives that $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$.

Proof that 3 $\implies$ 2: this follows immediately by applying the bounded convergence theorem to the sequence $g(Y_n)$.

Proof that 1 $\implies$ 3: Take $(\Omega, \mathcal{F}, \mathbf{P}) = ((0,1), \mathcal{B}(0,1), \text{Leb})$. For each $n \ge 1$, let $Y_n(x) = \sup\{y : F_{X_n}(y) < x\}$ be the **lower quantile function** of X_n, as discussed in a previous lecture, and similarly let $Y(x) = \sup\{y : F_X(y) < x\}$ be the lower quantile function of X. Then as we previously showed, we have $F_Y \equiv F_X$ and $F_{Y_n} \equiv F_{X_n}$ for all n. It remains to show that $Y_n(x) \to Y(x)$ for almost all $x \in (0,1)$. In fact, we show that this is true for all but a countable set of x's. Denote $Y^*(x) = \inf\{y : F_X(y) > x\}$ (the **upper quantile function** of X). As we have seen, we always have $Y(x) \le Y^*(x)$, and $Y(x) = Y^*(x)$ for all $x \in (0,1)$ except on a countable set of x's (the exceptional x's correspond to intervals where F_X is constant; these intervals are disjoint and each one contains a rational point).

Let $x \in (0,1)$ be such that $Y(x) = Y^*(x)$. This means that for any $y < Y(x)$ we have $F_X(y) < x$, and for any $z > Y(x)$ we have $F_X(z) > x$. Now, take a $y < Y(x)$ which is a continuity point of F_X. Then $F_{X_n}(y) \to F_X(y)$ as $n \to \infty$, so also $F_{X_n}(y) < x$ for sufficiently large n, which means (by the definition of Y_n) that $Y_n(x) \ge y$ for such large n. This establishes that $\liminf_{n\to\infty} Y_n(x) \ge y$, and therefore that $\liminf_{n\to\infty} Y_n(x) \ge Y(x)$, since we have continuity points $y < Y(x)$ that are arbitrarily close to $Y(x)$.

Similarly, take a $z > Y(x)$ which is a continuity point of F_X. Then $F_{X_n}(z) \to F_x(z)$ as $n \to \infty$, so also $F_{X_n}(z) > x$ for large n, which implies that $Y_n(x) \le z$. Again, by taking continuity points $z > Y(x)$ that are arbitrarily close to $Y(x)$ we get that $\limsup_{n\to\infty} Y_n(x) \le Y(x)$. Combining these last two results shows that $Y_n(x) \to Y(x)$ which was what we wanted. □

12.2 Examples

1. **Normal convergence:** We showed that if $X_1, X_2, \ldots$ are i.i.d. $\text{Binom}(1,p)$ r.v.'s and $S_n = \sum_{k=1}^n X_k$, then
$$\frac{S_n - n\mathbf{E}(X_1)}{\sqrt{n}\sigma(X_1)} \implies N(0,1).$$

Similarly, using explicit computations (see the homework) it is not too difficult to see that this is also true when $X_1 \sim \text{Poisson}(1)$, $X_1 \sim \text{Exp}(1)$, and in other specific examples. The central limit theorem generalizes this claim to any i.i.d. sequence with finite variance.

2. **Waiting for rare events:** If for each $0 < p < 1$ we have a r.v. $X_p \sim \text{Geom}_0(p)$, then $\mathbf{P}(X_p \geq n) = (1-p)^{n-1}$. It follows that
$$\mathbf{P}(pX_p > x) = (1-p)^{\lfloor x/p \rfloor} \xrightarrow[p\downarrow 0]{} e^{-x}, \qquad (x > 0),$$
so
$$pX_p \implies \text{Exp}(1) \qquad \text{as } p \downarrow 0.$$

3. **Pólya's urn:** Let X_n be the number of white balls in the Pólya urn experiment after starting with one white ball and one black ball and performing the experiment for n steps (so that there are $n+2$ balls). In a homework exercise we showed that X_n is a discrete uniform r.v. on $\{1, 2, \ldots, n+1\}$. It follows easily that the proportion of white balls in the urn converges in distribution:
$$\frac{X_n}{n+2} \implies U(0,1).$$

4. **Gumbel distribution:** If $X_1, X_2, \ldots$ are i.i.d. Exp(1) random variables, and $M_n = \max(X_1, \ldots, X_n)$, we showed in a homework exercise that
$$\mathbf{P}(M_n - \log n \leq x) \xrightarrow[n\to\infty]{} e^{-e^{-x}}, \qquad x \to \infty$$
It follows that
$$M_n - \log n \implies F$$
where $F(x) = \exp\left(-e^{-x}\right)$ is called the Gumbel distribution.

12.3 Compactness and tightness

Theorem 12.3 (Helly's selection theorem). *If $(F_n)_{n=1}^\infty$ is a sequence of distribution functions, then there is a subsequence F_{n_k} and a right-continuous, nondecreasing function $H : \mathbb{R} \to [0,1]$ such that*
$$F_{n_k}(x) \xrightarrow[n\to\infty]{} H(x)$$
holds for any $x \in \mathbb{R}$ which is a continuity point of H.

Note. The subsequential limit H need not be a distribution function, since it may not satisfy the properties $\lim_{x\to-\infty} H(x) = 0$ or $\lim_{x\to\infty} H(x) = 1$. For example, taking $F_n = F_{X_n}$, where $X_n \sim U[-n,n]$, we see that $F_n(x) \to 1/2$ for all $x \in \mathbb{R}$. For a more interesting example, take $G_n = (F_n + F_{Z_n})/2$ where F_n are as in the previous example, and Z_n is some sequence of r.v.'s that converges in distribution.

Proof. First, note that we can find a subsequence $(n_k)_{k=1}^{\infty}$ such that $F_{n_k}(r)$ converges to a limit $G(r)$ at least for any *rational* number r. This is done by combining the compactness of the interval $[0,1]$ (which implies that for any specific $a \in \mathbb{R}$ we can always take a subsequence to make the sequence of numbers $F_n(a)$ converge to a limit) with a diagonal argument (for some enumeration $r_1, r_2, r_3, \ldots$ of the rationals, first take a subsequence to force convergence at r_1; then take a subsequence of that subsequence to force convergence at r_2, etc.; now form a subsequence whose k-th term is the k-th term of the k-th subsequence in this series).

Now, use $G(\cdot)$, which is defined only on the rationals and not necessarily right-continuous (but is nondecreasing), to define a function $H : \mathbb{R} \to \mathbb{R}$ by

$$H(x) = \inf\{G(r) : r \in \mathbb{Q}, r > x\}.$$

This function is clearly nondecreasing, and is also right-continuous, since we have

$$\lim_{x_n \downarrow x} H(x_n) = \inf\{G(r) : r \in \mathbb{Q}, r > x_n \text{ for some } n\} = \inf\{G(r) : r \in \mathbb{Q}, r > x\} = H(x).$$

Finally, let x be a continuity point of H. To show that $F_{n_k}(x) \to H(x)$, fix some $\epsilon > 0$ and let r_1, r_2, s be rationals such that $r_1 < r_2 < x < s$ and

$$H(x) - \epsilon < H(r_1) \le H(r_2) \le H(x) \le H(s) < H(x) + \epsilon.$$

Then since $F_{n_k}(r_2) \to G(r_2) \ge H(r_1)$, and $F_{n_k}(s) \to G(s) \le H(s)$, it follows that for sufficiently large k we have

$$H(x) - \epsilon < F_{n_k}(r_2) \le F_{n_k}(x) \le F_{n_k}(s) < H(x) + \epsilon.$$

Therefore

$$H(x) - \epsilon \le \liminf_{n\to\infty} F_{n_k}(x) \le \limsup_{n\to\infty} F_{n_k}(x) \le H(x) + \epsilon,$$

and since ϵ was arbitrary this proves the claim. $\square$

Theorem 12.3 can be thought of as a kind of compactness property for probability distributions, except that the subsequential limit guaranteed to exist by the theorem is not a distribution function. To ensure that we get a distribution function, it turns out that a certain property called **tightness** has to hold.

Definition 12.4. *A sequence $(\mu_n)_{n=1}^\infty$ of probability measures on $(\mathbb{R}, \mathcal{B})$ is called **tight** if for any $\epsilon > 0$ there exists an $M > 0$ such that*

$$\liminf_{n\to\infty} \mu_n([-M, M]) \geq 1 - \epsilon.$$

A sequence of distribution functions $(F_n)_{n=1}^\infty$ is called tight if the associated probability measures determined by F_n form a tight sequence, or, more explicitly, if for any $\epsilon > 0$ there exists an $M > 0$ such that

$$\limsup_{n\to\infty} (1 - F_n(M) + F_n(-M)) < \epsilon.$$

A sequence of random variables is called tight if the sequence of their distribution functions is tight.

Theorem 12.5. *If $(F_n)_{n=1}^\infty$ is a tight sequence of distribution functions, then there exists a subsequence $(F_{n_k})_{k=1}^\infty$ and a distribution function F such that $F_{n_k} \implies F$. In fact, any subsequential limit H as guaranteed to exist in the previous theorem is a distribution function.*

Exercise 12.6. *Prove that the converse is also true, i.e., if a sequence is not tight then it must have at least one subsequential limit H (in the sense of the subsequence converging to H at any continuity point of H) that is not a proper distribution function. In particular, it is worth noting that a sequence that converges in distribution is tight.*

Proof. Let H be a nondecreasing, right-continuous function that arises as a subsequential limit-in-distribution of a subsequence F_{n_k}, that we know exists by Theorem 12.3. To show that H is a distribution function, fix $\epsilon > 0$, and let $M > 0$ be the constant guaranteed to exist in the definition of tightness. Let $x < -M$ be a continuity point of H. We have

$$H(x) = \lim_{k\to\infty} F_{n_k}(x) \leq \limsup_{k\to\infty} F_{n_k}(-M) \leq \limsup_{k\to\infty} (F_{n_k}(-M) + (1 - F_{n_k}(M))) < \epsilon,$$

so this shows that $\lim_{x\to-\infty} H(x) = 0$. Similarly, let $x > M$ be a continuity point of H. Then

$$H(x) = \lim_{k\to\infty} F_{n_k}(x) \geq \liminf_{k\to\infty} F_{n_k}(M) \geq \liminf_{k\to\infty} (F_{n_k}(M)) - F_{n_k}(-M)) > 1 - \epsilon,$$

which shows that $\lim_{x\to\infty} H(x) = 1$. $\square$

The condition of tightness is not very restrictive, and in practical situations it is usually quite easy to verify. The following lemma gives an example that is relevant for our purposes.

Lemma 12.7. *If $X_1, X_2, \ldots$ are r.v.'s such that $\mathbf{E}X_n = 0$ and $\mathbf{V}(X_n) < C$ for all n, then $(X_n)_n$ is a tight sequence.*

Proof. Use Chebyshev's inequality:

$$\mathbf{P}(|X_n| > M) \le \frac{\mathbf{V}(X_n)}{M^2} \le \frac{C}{M^2},$$

so, if $\epsilon > 0$ is given, taking $M = \sqrt{C/\epsilon}$ ensures that the left-hand side is bounded by ϵ. □

Chapter 13: Characteristic functions

13.1 Definition and basic properties

A main tool in our proof of the central limit theorem will be that of characteristic functions. The basic idea will be to show that

$$\mathbf{E}\left[g\left(\frac{S_n - n\mu}{\sqrt{n}\sigma}\right)\right] \xrightarrow[n\to\infty]{} \mathbf{E}g(N(0,1))$$

for a sufficiently large family of functions g. It turns out that the family of functions of the form

$$g_t(x) = e^{itx}, \qquad (t \in \mathbb{R}),$$

is ideally suited for this purpose. (Here and throughout, $i = \sqrt{-1}$).

Definition 13.1. *The characteristic function of a r.v. X, denoted φ_X, is defined by*

$$\varphi_X(t) = \mathbf{E}\left(e^{itX}\right) = \mathbf{E}(\cos(tX)) + i\mathbf{E}(\sin(tX)), \qquad (t \in \mathbb{R}).$$

Note that we are taking the expectation of a *complex-valued random variable* (which is a kind of two-dimensional random vector, really). However, the main properties of the expectation operator (linearity, the triangle inequality etc.) that hold for real-valued random variables also hold for complex-valued ones, so this will not pose too much of a problem.

Here are some simple properties of characteristic functions. For simplicity we denote $\varphi = \varphi_X$ where there is no risk of confusion.

1. $\varphi(0) = \mathbf{E}e^{i\cdot 0\cdot X} = 1$.

2. $\varphi(-t) = \mathbf{E}e^{-itX} = \mathbf{E}\left(\overline{e^{itX}}\right) = \overline{\varphi(t)}$ (where $\overline{z}$ denotes the complex conjugate of a complex number z).

3. $|\varphi(t)| \le \mathbf{E}\left|e^{itX}\right| = 1$ by the triangle inequality.

4. $|\varphi(t) - \varphi(s)| \le \mathbf{E}\left|e^{itX} - e^{isX}\right| = \mathbf{E}\left|e^{isX}\left(e^{i(t-s)X} - 1\right)\right| = \mathbf{E}\left|e^{i(t-s)X} - 1\right|$. Note also that $\mathbf{E}\left|e^{iuX} - 1\right| \to 0$ as $u \downarrow 0$ by the bounded convergence theorem. It follows that φ is a uniformly continuous function on $\mathbb{R}$.

5. $\varphi_{aX}(t) = \mathbf{E}e^{iatX} = \varphi_X(at)$, $(a \in \mathbb{R})$.

6. $\varphi_{X+b}(t) = \mathbf{E}e^{it(X+b)} = e^{ibt}\varphi_X(t), \ \ (b \in \mathbb{R})$.

7. **Important:** If X, Y are independent then

$$\varphi_{X+Y}(t) = \mathbf{E}\left(e^{it(X+Y)}\right) = \mathbf{E}\left(e^{itX}e^{itY}\right) = \mathbf{E}\left(e^{itX}\right)\mathbf{E}\left(e^{itY}\right) = \varphi_X(t)\varphi_Y(t).$$

   Note that this is the main reason why characteristic functions are such a useful tool for studying the distribution of a sum of independent random variables.

A note on terminology. If X has a density function f, then the characteristic function can be computed as

$$\varphi_X(t) = \int_{-\infty}^{\infty} f_X(x)e^{itx}\,dx.$$

In all other branches of mathematics, this would be called the **Fourier transform**[1] **of** f. So the concept of a characteristic function generalizes the Fourier transform. If μ is the distribution measure of X, some authors write

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itx}d\mu(x)$$

(which is an example of a **Lebesgue-Stieltjes integral**) and call this the **Fourier-Stieltjes transform** (or just the Fourier transform) of the measure μ.

13.2 Examples

No study of characteristic functions is complete without "dirtying your hands" a little to compute the characteristic function for some important cases. The following exercise is highly recommended

Exercise 13.2. *Compute the characteristic functions for the following distributions.*

1. ***Coin flips:*** *Compute* φ_X *when* $\mathbf{P}(X = -1) = \mathbf{P}(X = 1) = 1/2$ *(this comes out slightly more symmetrical than the usual Bernoulli r.v. for which* $\mathbf{P}(X = 0) = \mathbf{P}(X = 1) = 1/2$*).*

[1] Well, more or less – it is really the inverse Fourier transform; but it will be the Fourier transform if we replace t by $-t$, so that is almost the same thing

2. ***Symmetric random walk:*** *Compute φ_{S_n} where $S_n = \sum_{k=1}^n X_k$ is the sum of n i.i.d. copies of the coin flip distribution above.*

3. ***Poisson distribution:*** *$X \sim Poisson(\lambda)$.*

4. ***Uniform distribution:*** *$X \sim U[a,b]$, and in particular $X \sim [-1,1]$ which is especially symmetric and useful in applications.*

5. ***Exponential distribution:*** *$X \sim Exp(\lambda)$.*

6. ***Symmetrized exponential:*** *A r.v. Z with density function $f_Z(x) = e^{-|x|}$. Note that this is the distribution of the exponential distribution after being "symmetrized" in either of two ways: (i) We showed that if $X, Y \sim Exp(1)$ are independent then $X - Y$ has density $e^{-|x|}$; (ii) alternatively, it is the distribution of an "exponential variable with random sign", namely $\varepsilon \cdot X$ where $X \sim Exp(1)$ and ε is a random sign (same as the coin flip distribution mentioned above) that is independent of X.*

The normal distribution has the nice property that its characteristic function is equal, up to a constant, to its density function.

Lemma 13.3. *If $Z \sim N(0,1)$ then*

$$\varphi_Z(t) = e^{-t^2/2}.$$

Proof.

$$\begin{aligned} \varphi_Z(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-x^2/2}\,dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2} e^{(x-it)^2/2}\,dx \\ &= e^{-t^2/2} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{(x-it)^2/2}\,dx \right). \end{aligned}$$

As Durrett suggests in his "physics proof" (p. 92 in [Dur2010], 91 in [Dur2004]), the expression in parentheses is 1, since it is the integral of a normal density with mean it and variance 1. This is a nonsensical argument, of course (it being an imaginary number), but the claim is true, easy and is proved in any complex analysis course using contour integration.

Alternatively, let $S_n = \sum_{k=1}^n X_k$ where $X_1, X_2, \ldots$ are i.i.d. coin flips with $\mathbf{P}(X_k) = -1 = \mathbf{P}(X_k) = 1 = 1/2$. We know from the de Moivre-Laplace theorem (Theorem 11.1) that

$$S_n/\sqrt{n} \implies N(0,1),$$

so that

$$\varphi_{S_n/\sqrt{n}}(t) = \mathbf{E}\left(e^{itS_n/\sqrt{n}}\right) \xrightarrow[n\to\infty]{} \varphi_Z(t), \qquad (t \in \mathbb{R}),$$

since the function $x \to e^{itx}$ is bounded and continuous. On the other hand, from the exercise above it is easy to compute that $\varphi_{S_n}(t) = \cos^n(t)$, which implies that

$$\varphi_{S_n/\sqrt{n}}(t) = \cos^n\left(\frac{t}{\sqrt{n}}\right) = \left(1 - \frac{t^2}{2n} + O\left(\frac{t^4}{n^2}\right)\right)^n \xrightarrow[n\to\infty]{} e^{-t^2/2}.$$

□

As a consequence, let $X \sim N(0, \sigma_1^2)$ and $Y \sim N(0, \sigma_2^2)$ be independent, and let $Z = X+Y$. Then

$$\varphi_X(t) = e^{-\sigma_1^2 t^2/2}, \qquad \varphi_Y(t) = e^{-\sigma_2^2 t^2/2},$$

so $\varphi_Z(t) = e^{-(\sigma_1^2+\sigma_2^2)/2}$. This is the same as $\varphi_W(t)$, where $W \sim N(0, \sigma_1^2 + \sigma_2^2)$. It would be nice if we could deduce from this that $Z \sim N(0, \sigma_1^2 + \sigma_2^2)$ (we already proved this fact in a homework exercise, but it's always nice to have several proofs of a result, especially an important one like this one). This naturally leads us to an important question about characteristic functions, which we consider in the next section.

13.3 The inversion formula

A fundamental question about characteristic functions is whether they contain all the information about a distribution, or in other words whether knowing the characteristic function determines the distribution uniquely. This question is answered (affirmatively) by the following theorem, which is a close cousin of the standard inversion formula from analysis for the Fourier transform.

Theorem 13.4 (The inversion formula)**.** *If X is a r.v. with distribution μ_X, then for any $a < b$ we have*

$$\begin{aligned} \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-iat} - e^{-ibt}}{it} \varphi_X(t)\,dt &= \mu_X((a,b)) + \frac{1}{2}\mu_X(\{a,b\}) \\ &= \mathbf{P}(a < X < b) + \frac{1}{2}\mathbf{P}(X = a) + \frac{1}{2}\mathbf{P}(X = b). \end{aligned}$$

Corollary 13.5. *If X, Y are r.v.s such that $\varphi_X(t) \equiv \varphi_Y(t)$ for all $t \in \mathbb{R}$ then $X \overset{d}{=} Y$.*

Exercise 13.6. *Explain why Corollary 13.5 follows from the inversion formula.*

Proof of Theorem 13.4. Throughout the proof, denote $\varphi(t) = \varphi_X(t)$ and $\mu = \mu_X$. For convenience, we use the notation of Lebesgue-Stieltjes integration with respect to the measure μ, remembering that this really means taking the expectation of some function of the r.v. X. Denote

$$I_T = \int_{-T}^{T} \frac{e^{-iat} - e^{-ibt}}{it} \varphi(t)\,dt = \int_{-T}^{T} \int_{-\infty}^{\infty} \frac{e^{-iat} - e^{-ibt}}{it} e^{itx} d\mu(x)\,dt. \tag{12}$$

Since $\frac{e^{-iat}-e^{-ibt}}{it} = \int_a^b e^{-ity}\,dy$ is a bounded function of t (it is bounded in absolute value by $b-a$), it follows by Fubini's theorem that we can change the order of integration, so

$$\begin{aligned} I_T &= \int_{-\infty}^{\infty} \int_{-T}^{T} \frac{e^{-iat} - e^{-ibt}}{it} e^{itx} dt\,d\mu(x) \\ &= \int_{-\infty}^{\infty} \left[\int_{-T}^{T} \frac{\sin(t(x-a))}{t}\,dt - \int_{-T}^{T} \frac{\sin(t(x-b))}{t}\,dt \right] d\mu(x) \\ &= \int_{-\infty}^{\infty} (R(x-a,T) - R(x-b,T))\,d\mu(x), \end{aligned}$$

where we denote $R(\theta,T) = \int_{-T}^{T} \sin(\theta t)/t\,dt$. Note that in the notation of expectations this can be written as $I_T = \mathbf{E}\left(R(X-a,T) - R(X-b,T)\ \right)$. This can be simplified somewhat; in fact, observe also that

$$R(\theta,T) = 2\mathrm{sgn}(\theta) \int_0^{|\theta|T} \frac{\sin x}{x}\,dx = 2\mathrm{sgn}(\theta) S(|\theta|T),$$

where we denote $S(x) = \int_0^x \frac{\sin(u)}{u}\,du$ and $\mathrm{sgn}(\theta)$ is 1 if $\theta > 0$, -1 if $\theta < 0$ and 0 if $\theta = 0$. By a standard convergence test for integrals, the improper integral $\int_0^\infty \frac{\sin u}{u}\,du = \lim_{x\to\infty} S(x)$ converges; denote its value by $C/4$. Thus, we have shown that $R(\theta,T) \to \frac{1}{2}\mathrm{sgn}(\theta)C$ as $T \to \infty$, hence that

$$R(x-a,T) - R(x-b,T) \xrightarrow[T\to\infty]{} \begin{cases} C & a < x < b, \\ C/2 & x = a \text{ or } x = b, \\ 0 & x < a \text{ or } x > b. \end{cases}$$

Furthermore, the function $R(x-a,T) - R(x-b,T)$ is bounded in absolute value by $2\sup_{x\ge 0} S(x)$. It follows that we can apply the bounded convergence theorem in (12) to get that

$$I_T \xrightarrow[T\to\infty]{} C\mathbf{E}(1_{a<X<b}) + (C/2)\mathbf{E}(\mathbf{1}_{\{X=a\}} + \mathbf{1}_{\{X=b\}}) = C\mu((a,b)) + (C/2)\mu(\{a,b\}). \tag{13}$$

This is just what we claimed, minus the fact that $C = 2\pi$. This fact is a well-known integral evaluation from complex analysis. We can also deduce it in a self-contained manner, by applying what we proved to a specific measure μ and specific values of a and b for which we can evaluate the limit in (12) directly. This is not entirely easy to do, but one possibility, involving an additional limiting argument, is outlined in the next exercise; see also Exercise 1.7.5 on p. 35 in [Dur2010], (Exercise 6.6, p. 470 in Appendix A.6 of [Dur2004]) for a different approach to finding the value of C. □

Exercise 13.7. ***(Recommended for aspiring analysts...)*** *For each $\sigma > 0$, let X_σ be a r.v. with distribution $N(0, \sigma^2)$ and therefore with density $f_X(x) = (\sqrt{2\pi}\sigma)^{-1} e^{-x^2/2\sigma^2}$ and characteristic function $\varphi_X(t) = e^{-\sigma^2 t^2/2}$. For fixed σ, apply Theorem 13.4 in its weak form given by* (13) *(that is, without the knowledge of the value of C), with parameters $X = X_\sigma$, $a = -1$ and $b = 1$, to deduce the identity*

$$\frac{C}{\sqrt{2\pi}\sigma} \int_{-1}^{1} e^{-x^2/2\sigma^2}\, dx = \int_{-\infty}^{\infty} \frac{2 \sin t}{t} e^{-\sigma^2 t^2/2}\, dt.$$

Now multiply both sides by σ and take the limit as $\sigma \to \infty$. For the left-hand side this should give in the limit (why?) the value $(2C)/\sqrt{2\pi}$. For the right-hand side this should give $2\sqrt{2\pi}$. Justify these claims and compare the two numbers to deduce that $C = 2\pi$.

The following theorem shows that the inversion formula can be written as a simpler connection between the characteristic function and the density function of a random variable, in the case when the characteristic function is integrable.

Theorem 13.8. *If $\int_{-\infty}^{\infty} |\varphi_X(t)|\, dt < \infty$, then X has a bounded and continuous density function f_X, and the density and characteristic function are related by*

$$\begin{aligned} \varphi_X(t) &= \int_{-\infty}^{\infty} f_X(x) e^{itx}\, dx, \\ f_X(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi_X(t) e^{-itx}\, dt. \end{aligned}$$

In the lingo of Fourier analysis, this is known as the **inversion formula for Fourier transforms**.

Proof. This is a straightforward corollary of Theorem 13.4. See p. 95 in either [Dur2010] or [Dur2004]. □

13.4 The continuity theorem

Theorem 13.9. *Let $(X_n)_{n=1}^\infty$ be r.v.'s. Then:*

(i) If $X_n \implies X$ for some r.v. X, then $\varphi_{X_n}(t) \to \varphi_X(t)$ for all $t \in \mathbb{R}$.

(ii) If the limit $\varphi(t) = \lim_{n\to\infty} \varphi_{X_n}(t)$ exists for all $t \in \mathbb{R}$, and φ is continuous at 0, then $\varphi \equiv \varphi_X$ for some r.v. X, and $X_n \implies X$.

Proof. Part (i) follows immediately from the fact that convergence in distribution implies that $\mathbf{E}g(X_n) \to \mathbf{E}g(X)$ for any bounded continuous function. It remains to prove the less trivial claim in part (ii). Assume that $\varphi_{X_n}(t) \to \varphi(t)$ for all $t \in \mathbb{R}$ and that φ is continuous at 0. First, we show that the sequence $(X_n)_{n=1}^\infty$ is tight. Fixing an $M > 0$, we can bound the probability $\mathbf{P}(|X_n| > M)$, as follows:

$$\begin{aligned}\mathbf{P}(|X_n| > M) &= \mathbf{E}\left(\mathbf{1}_{\{|X_n|>M\}}\right) \le \mathbf{E}\left[2\left(1 - \frac{M}{2|X_n|}\right)\mathbf{1}_{\{|X_n|>M\}}\right] \\ &\le \mathbf{E}\left[2\left(1 - \frac{\sin(2X_n/M)}{2X_n/M}\right)\mathbf{1}_{\{|X_n|>M\}}\right].\end{aligned}$$

But this last expression can be related to the behavior of the characteristic function near 0. Denote $\delta = 2/M$. Reverting again to the Lebesgue-Stieltjes integral notation, we have

$$\begin{aligned}\mathbf{E}\left[2\left(1 - \frac{\sin(2X_n/M)}{2X_n/M}\right)\mathbf{1}_{\{|X_n|>M\}}\right] &= 2\int_{|x|>2/\delta}\left(1 - \frac{\sin(\delta x)}{\delta x}\right) d\mu_{X_n}(x) \\ \le 2\int_{-\infty}^{\infty}\left(1 - \frac{\sin(\delta x)}{\delta x}\right) d\mu_{X_n}(x) &= \int_{-\infty}^{\infty}\frac{1}{\delta}\left(\int_{-\delta}^{\delta}(1 - e^{itx})\,dt\right) d\mu_{X_n}(x).\end{aligned}$$

Now use Fubini's theorem to get that this bound can be written as

$$\frac{1}{\delta}\int_{-\delta}^{\delta}\int_{-\infty}^{\infty}(1 - e^{itx})\,d\mu_{X_n}(x)\,dt = \frac{1}{\delta}\int_{-\delta}^{\delta}(1 - \varphi_{X_n}(t))\,dt \xrightarrow[n\to\infty]{} \frac{1}{\delta}\int_{-\delta}^{\delta}(1 - \varphi(t))\,dt$$

(the convergence follows from the bounded convergence theorem). So we have shown that

$$\limsup_{n\to\infty} \mathbf{P}(|X_n| > M) \le \frac{1}{\delta}\int_{-\delta}^{\delta}(1 - \varphi(t))\,dt.$$

But, because of the assumption that $\varphi(t) \to \varphi(0) = 1$ as $t \to 0$, it follows that if δ is sufficiently small then $\delta^{-1}\int_{-\delta}^{\delta}(1 - \varphi(t))\,dt < \epsilon$, where $\epsilon > 0$ is arbitrary; so this establishes the tightness claim.

Finally, to finish the proof, let $(n_k)_{k=1}^\infty$ be a subsequence (guaranteed to exist by tightness) such that $X_{n_k} \implies Y$ for some r.v. Y. Then $\varphi_{X_{n_k}}(t) \to \varphi_Y(t) = \varphi(t)$ as $k \to \infty$ for all $t \in \mathbb{R}$, so $\varphi \equiv \varphi_Y$. This determines the distribution of Y, which means that the limit in distribution is the same no matter what convergent in distribution subsequence of the sequence $(X_n)_n$ we take. But this implies that $X_n \implies Y$ (why? The reader is invited to verify this last claim; it is best to use the definition of convergence in distribution in terms of expectations of bounded continuous functions). □

13.5 Moments

The final step in our lengthy preparation for the proof of the central limit theorem will be to tie the behavior of the characteristic function $\varphi_X(t)$ near $t = 0$ to the moments of X. Note that, computing formally without regards to rigor, we can write

$$\varphi_X(t) = \mathbf{E}(e^{itX}) = \mathbf{E}\left[\sum_{n=0}^\infty \frac{i^n t^n X^n}{n!}\right] = \sum_{n=0}^\infty \frac{i^n \mathbf{E} X^n}{n!} t^n.$$

So it appears that the moments of X appear as (roughly) the coefficients in the Taylor expansion of φ_X around $t = 0$. However, for CLT we don't want to assume anything beyond the existence of the second moment, so a (slightly) more delicate estimate is required.

Lemma 13.10. $\left| e^{ix} - \sum_{m=0}^n \frac{(ix)^m}{m!} \right| \le \min\left(\frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!} \right)$.

Proof. Start with the identity

$$R_n(x) := e^{ix} - \sum_{m=0}^n \frac{(ix)^m}{m!} = \frac{i^{n+1}}{n!} \int_0^x (x-s)^n e^{is}\, ds,$$

which follows from Lemma 11.3 that we used in the proof of Stirling's formula. Taking the absolute value and using the fact that $|e^{is}| = 1$ gives

$$|R_n(x)| \le \frac{1}{n!} \left| \int_0^x |x-s|^n \, ds \right| = \frac{|x|^{n+1}}{n!}. \tag{14}$$

To get a bound that is better-behaved for large x, note that

$$\begin{aligned} R_n(x) &= R_{n-1}(x) - \frac{(ix)^n}{n!} = R_{n-1}(x) - \frac{i^n}{(n-1)!} \int_0^x (x-s)^{n-1}\, ds \\ &= \frac{i^n}{(n-1)!} \int_0^x (x-s)^{n-1} (e^{is} - 1)\, ds. \end{aligned}$$

So, since $|e^{is} - 1| \leq 2$, we get that

$$|R_n(x)| \leq \frac{2}{(n-1)!} \left| \int_0^x |x-s|^{n-1}\, ds \right| = \frac{2|x|^n}{(n-1)!}. \tag{15}$$

Combining (14) and (15) gives the claim. □

Now let X be a r.v. with $\mathbf{E}|X|^n < \infty$. Letting $x = tX$ in Lemma 13.10, taking expectations and using the triangle inequality, we get that

$$\left| \varphi_X(t) - \sum_{m=0}^{n} \frac{i^m \mathbf{E} X^m}{m!} t^m \right| \leq \mathbf{E} \left[\min \left(\frac{|t|^{n+1}|X|^{n+1}}{(n+1)!}, \frac{2|t|^n |X|^n}{n!} \right) \right]. \tag{16}$$

Note that in this minimum of two terms, when t is very small the first term gives a better bound, but when taking expectations we need the second term to ensure that the expectation is finite if X is only assumed to have a finite n-th moment.

Theorem 13.11. *If X is a r.v. with mean $\mu = \mathbf{E}X$ and $\mathbf{V}(X) < \infty$ then*

$$\varphi_X(t) = 1 + i\mu t - \frac{\mathbf{E}X^2}{2} t^2 + o(t^2) \qquad \text{as } t \to 0.$$

Proof. By (16) above, we have

$$\frac{1}{t^2} \left| \varphi_X(t) - \left(1 + i\mu t - \frac{\mathbf{E}X^2}{2} t^2 \right) \right| \leq \mathbf{E} \left[\min \left(|t| \cdot |X|^3/6, X^2 \right) \right].$$

As $t \to 0$, the right-hand side converges to 0 by the dominated convergence theorem. □

Chapter 14: Central limit theorems

14.1 The case of i.i.d. r.v.'s

We are now ready to prove:

Theorem 14.1 (The central limit theorem)**.** *Let $X_1, X_2, \ldots$ be an i.i.d. sequence of r.v.'s with finite variance. Denote $\mu = \mathbf{E}X_1$, $\sigma = \sigma(X_1)$ and $S_n = \sum_{k=0}^{n} X_k$. Then as $n \to \infty$ we have the convergence in distribution*

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \implies N(0,1).$$

Proof. For convenience, denote $\hat{X}_k = (X_k - \mu)/\sigma$ and $\hat{S}_n = \sum_{k=0}^{n} \hat{X}_k$. Then

$$\varphi_{\hat{S}_n/\sqrt{n}}(t) = \varphi_{\hat{S}_n}(t/\sqrt{n}) = \prod_{k=1}^{n} \varphi_{\hat{X}_k}(t/\sqrt{n}) = \left(\varphi_{\hat{X}_1}(t/\sqrt{n})\right)^n .$$

Note that $\mathbf{E}\hat{X}_1 = 0$ and $\mathbf{V}(\hat{X}_1) = \mathbf{E}\hat{X}_1^2 = 1$. Therefore by Theorem 13.11, $\varphi_{\hat{X}_1}$ satisfies

$$\varphi_{\hat{X}_1}(u) = 1 - \frac{u^2}{2} + o(u^2)$$

as $u \to 0$. It follows that

$$\varphi_{\hat{S}_n/\sqrt{n}}(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \xrightarrow[n\to\infty]{} e^{-t^2/2}$$

for any $t \in \mathbb{R}$. Using the continuity theorem (Theorem 13.9) and our previous computations, it follows that $\hat{S}_n \implies N(0,1)$, as claimed. □

14.2 Generalizations

The CLT can be generalized in many ways. None of the assumptions (independence, identical distributions, even finite variance) are entirely necessary. A central paradigm of probability theory is that any random quantity that arises as a sum of many small contributions that are either independent or not too strongly dependent, will converge to the normal distribution in some asymptotic limit. Thousands of examples exist, but there is no single all-encompassing theorem that includes all of them as a special case. Rather, probabilists have a toolbox of tricks and techniques that they try to apply in order to prove normal convergence in any given situation. characteristic functions are among the more useful techniques. Another important technique, the so-called **moment method**, involves the direct use of moments: If we can show that $\mathbf{E}(W_n^k) \to \mathbf{E}(Z^k)$, where $(W_n)_{n=1}^\infty$ is the (normalized) sequence being studied, and $Z \sim N(0,1)$, then by Theorem 3.3.12 in [Dur2010] (p. 105) or Theorem (3.12) in [Dur2004] (p. 109), that implies that $W_n \implies N(0,1)$.

We now discuss several examples of interesting generalizations of CLT.

14.2.1 Triangular arrays

Theorem 14.2 (Lindeberg-Feller CLT for triangular arrays)**.** *Let $(X_{n,k})_{1\le k\le n<\infty}$ be a triangular array of r.v.'s. Denote $S_n = \sum_{k=1}^n X_{n,k}$ (the sum of the n-th row). Assume that:*

1. *For each n, the r.v.'s $(X_{n,k})_{k=1}^n$ are independent.*
2. *$\mathbf{E}X_{n,k} = 0$ for all n, k.*
3. *$\mathbf{V}(S_n) = \sigma_n^2 \to \sigma^2 < \infty$ as $n \to \infty$.*
4. *For all $\epsilon > 0$, $\lim\limits_{n\to\infty} \sum_{k=1}^n \mathbf{E}\left(X_{n,k}^2 \mathbf{1}_{\{|X_{n,k}|>\epsilon\}}\right) = 0$.*

Then $S_n \implies N(0,\sigma^2)$ as $n \to \infty$.

Proof. See [Dur2010], p. 110–111 or [Dur2004], p. 115–116. The proof uses the characteristic function technique and is a straightforward extension of the proof for the i.i.d. case. □

Example 14.3. (Record times and cycles in permutations). Let $X_1, X_2, \ldots$ be i.i.d. $U(0,1)$ r.v.'s. Let A_n be the event that $\{X_n = \max(X_1,\ldots,X_n)\}$ (in this case, we say that n is a **record time**). Let $S_n = \sum_{k=1}^n \mathbf{1}_{A_k}$ be the number of record times up to time n. We saw in a homework exercise that the A_k's are independent events and $\mathbf{P}(A_k) = 1/k$. This implies that $\mathbf{E}(S_n) = \sum_{k=1}^n \frac{1}{k} = H_n$ (the n-th **harmonic number**) and $\mathbf{V}(S_n) = \sum_{k=1}^n \frac{k-1}{k^2}$. Note that both $\mathbf{E}(S_n)$ and $\mathbf{V}(S_n)$ are approximately equal to $\log n$, with an error term that is $O(1)$. Now taking $X_{n,k} = (\mathbf{1}_{A_k} - k^{-1})/\sqrt{\mathbf{V}(S_n)}$ in Theorem 14.2, it is easy to check that the assumptions of the theorem hold. It follows that

$$\frac{S_n - H_n}{\sigma(S_n)} \implies N(0,1).$$

Equivalently, because of the asymptotic behavior of $\mathbf{E}(S_n)$ and $\mathbf{V}(S_n)$ it is also true that

$$\frac{S_n - \log n}{\sqrt{\log n}} \implies N(0,1).$$

Note: S_n describes the distribution of another interesting statistic on random permutations. It is not too difficult to show by induction (using an amusing construction often referred to as the **Chinese restaurant process**) that if $\sigma \in S_n$ is a uniformly random permutation on n elements, then the number of cycles in σ is a random variable which is equal in distribution to S_n.

14.2.2 Erdös-Kac theorem

Theorem 14.4 (Erdös-Kac theorem (1940))**.** *Let $g(m)$ denote the number of prime divisors of an integer k (for example, $g(28) = 2$). For each $n \geq 1$, let X_n be a uniformly random integer chosen in $\{1, 2, \ldots, n\}$, and let $Y_n = g(X_n)$ be the number of prime divisors of X_n. Then we have*

$$\frac{Y_n - \log\log n}{\sqrt{\log\log n}} \implies N(0,1).$$

In other words, for any $x \in \mathbb{R}$ we have

$$\frac{1}{n}\#\left\{1 \leq k \leq n : g(k) \leq \log\log n + k\sqrt{\log\log n}\right\} \xrightarrow[n\to\infty]{} \Phi(x).$$

Proof. See [Dur2010], p. 114–117 or [Dur2004], p. 119–124. The proof uses the moment method. □

Note that Y_n can be written in the form $\sum_{p \leq n} \mathbf{1}_{\{p|X_n\}}$, namely the sum over all primes $p \leq n$ of the indicator of the event that X_n is divisible by p. The probability that X_n is divisible by p is roughly $1/p$, at least if p is significantly smaller than n. Therefore we can expect Y_n to be on the average around

$$\sum_{\text{prime } p \leq n} \frac{1}{p},$$

a sum that is known (thanks to Euler) to behave roughly like $\log\log n$. The Erdös-Kac theorem is intuitively related to the observation that these indicators $\mathbf{1}_{\{p|X_n\}}$ for different p's are also close to being independent (a fact which follows from the Chinese remainder theorem). Of course, they are only approximately independent, and making these observations precise is the challenge to proving the theorem. In fact, many famous open problems in number theory (even the Riemann Hypothesis, widely considered to be the most important open problem in mathematics) can be formulated in terms of a statement about approximate independence (in some loose sense) of some arithmetic sequence relating to the prime numbers.

14.2.3 The Euclidean algorithm

As a final example from number theory, consider the following problem: For some $n \geq 1$, choose X_n and Y_n independently and uniformly at random in $\{1, 2, \ldots, n\}$, and compute their greatest common divisor (g.c.d.) using the Euclidean algorithm. Let N_n be the number of division (with remainder) steps that were required. For example, if $X = 58$ and $Y = 24$ then the application of the Euclidean algorithm would result in the sequence of steps

$$(58, 24) \to (24, 10) \to (10, 4) \to (4, 2) \to (2, 0),$$

so 4 division operations were required (and the g.c.d. is 2).

Theorem 14.5 (CLT for the number of steps in the Euclidean algorithm; D. Hensley (1992))**.** *There exists a constant σ_∞ (which has a very complicated definition) such that*

$$\frac{N_n - \frac{12 \log 2}{\pi^2} \log n}{\sigma_\infty \sqrt{\log n}} \implies N(0, 1).$$

Hensley's theorem was in recent years significantly generalized and the techniques extended by Brigitte Vallée, a French mathematician. The fact that the average value of N_n is approximately $(12 \log 2/\pi^2) \log n$ was previously known from work of Heilbronn and Dixon in 1969–1970, using ideas dating back to Gauss, who discovered the probability distribution now called the "Gauss measure". This is the probability distribution on $(0, 1)$ with density $\frac{1}{\log 2(1+x)}$, which Gauss found (but did not prove!) describes the limiting distribution of the ratio of a pair of independent $U(0, 1)$ random variables after many iterations of the division-with-remainder step in the Euclidean algorithm.

Chapter 15: Random number generation

An important problem in applied probability theory is to produce one or more random variables distributed according to some specified distribution, starting from a source of random information whose distribution is fixed. This is referred to as **random number generation** and is also known as **statistical simulation** or **sampling**. In this chapter we survey several interesting methods for random number generation

Note that we assume that our source of randomness provides us with truly random information as the input for the computation. In the absence of such random information, the problem of producing information that *appears* to be random using a deterministic computation applied to a non-random input, is a very different problem, known as **pseudorandom number generation**, which we will not discuss (although it is also very interesting!).

15.1 Unbiasing a source of randomness: simulating an unbiased coin using biased coin tosses

Assume that your source of randomness produces a sequence $X_1, X_2, \ldots$ of Bernoulli random variables with some bias $0 < p < 1$. Such random bits are much more practical for use in applications if they are unbiased, that is, if $p = 1/2$. Can we use the p-biased coin tosses to produce an *unbiased* coin toss? Yes: the famous mathematician John von Neumann suggested the following simple method in 1951.

Simulation Method 1.

1. Sample a pair X, Y of the p-biased coin tosses.
2. If $(X, Y) = (1, 0)$, output "0".
3. If $(X, Y) = (0, 1)$, output "1".
4. If $(X, Y) = (0, 0)$ or $(X, Y) = (1, 1)$, go back to step 1.

Exercise 15.1. *Show that the method works; that is, that with probability* 1 *the method eventually outputs a random variable* $Z \sim \mathrm{Ber}(1/2)$.

An interesting and useful feature of Von Neumann's method is that it does not even require knowing the bias p of the source; that is, it is a "universal unbiasing method."

What about the efficiency of the method? It seems interesting to ask how many samples of the biased sequence we will need to produce our unbiased random bit. This number is random, and unbounded — we may have to wait an arbitrarily long time for the simulation to finish — so perhaps it makes more sense to ask about the *mean* number of samples required.

Exercise 15.2. *Show that on average the method requires* $\frac{1}{p(1-p)}$ *samples of the sequence* $(X_n)_{n=1}^{\infty}$*. (In particular, when p is very close to* 0 *or* 1 *the average wait becomes very long.)*

15.2 Simulating a biased coin using unbiased coin tosses

Let us now consider the reverse problem of simulating a biased coin toss with bias p when our source of randomness produces independent *unbiased* bits $X_1, X_2, \ldots \sim \text{Ber}(1/2)$. Of course, here we assume that the value p of the desired bias is known, otherwise the question makes no sense. It turns out that in this case too there is a simple method, and moreover the method works extremely well even if p is a very complicated number such as $1/\sqrt{2}$ or $\pi - 3$.

The method is based on the easy-to-prove observation that the random variable $U = \sum_{n=1}^{\infty} \frac{X_n}{2^n}$ has distribution $U[0,1]$. Then the indicator random variable $Z = 1_{\{U \le p\}}$ is a $\text{Ber}(p)$ random variable. It only remains to note that Z can be computed efficiently by uncovering the random bits $X_1, X_2, \ldots$ one at a time and stopping the computation as soon as it becomes apparent whether the event $\{U \le p\}$ occurred or its complement $\{U > p\}$. The way this question is settled is described as follows.

Simulation Method 2. Let $p = (0.\alpha_1\alpha_2\alpha_3\ldots)_2$ be the binary expansion of p, that is, $\alpha_n \in \{0,1\}$ and $p = \sum_{n=1}^{\infty} \alpha_n/2^n$.

1. Sample the random bits $X_1, X_2, \ldots$ one at a time until the first time $m \ge 1$ such that $X_m \ne \alpha_m$.

2. If $(X_m, \alpha_m) = (0,1)$, output "1".

3. If $(X_m, \alpha_m) = (1,0)$, output "0".

Exercise 15.3. *Prove that the output is* 1 *if* $U < p$*;* 0 *if* $U > p$*; and the algorithm never terminates in the event (which has probability 0) that* $U = p$*.*

Exercise 15.4. *Show that the number* N *of bits which had to be sampled to obtain an answer satisfies* $\mathbf{E}N = 2$*. What is the distribution of* N*?*

15.3 Simulating an arbitrary discrete distribution using unbiased coin tosses

We can generalize Simulation Method 2 described above to get a method for simulating an arbitrary discrete random variable taking values $\alpha_1, \ldots, \alpha_k$ with respective probabilities $p_1, \ldots, p_k$, using a sequence of independent unbiased coin tosses $X_1, X_2, \ldots \sim \text{Ber}(1/2)$. Let $U = \sum_{n=1}^{\infty} X_n/2^n$ as before, and for $m \geq 1$ let $U_m = \sum_{k=1}^{m} X_k/2^k$ be the partial sums of the binary expansion of U.

Simulation Method 3. Denote $c_0 = 0, c_j = p_1 + \ldots + p_j$. Note that the intervals $(c_0, c_1), (c_1, c_2), \ldots, (c_{k-1}, c_k)$ form a partition of the interval $(0, 1)$ into subintervals of lengths $p_1, \ldots, p_k$. To produce a sample from the discrete distribution with atoms of size p_j at α_j for $j = 1, \ldots, k$:

1. Sample $X_1, X_2, \ldots$ one at a time until the first time m for which there is a j such that the condition
$$c_{j-1} < U_m < U_m + \frac{1}{2^m} < c_j$$
holds. Note that this condition can be checked by comparing the binary string $(X_1, \ldots, X_m)$ against the first m digits in the binary expansions of the numbers $c_0, \ldots, c_k$.

2. Output α_j.

Exercise 15.5. *Show that the condition* $c_{j-1} < U_m < U_m + \frac{1}{2^m} < c_j$ *is equivalent to* $c_{j-1} < U < c_j$*, and that therefore for each* $1 \leq j \leq k$ *the output* α_j *is obtained with probability* p_j*.*

The average number of samples from $X_1, X_2, \ldots$ required for the simulation looks like a highly nontrivial function of the probabilities $p_1, \ldots, p_k$. Amazingly, to a good approximation it is equal to the **entropy** of the distribution $p_1, \ldots, p_k$, a well-known function from information theory that is known to measure the information content of a random variable. The precise result to that effect is as follows.

Theorem 15.6. *The **entropy** of the discrete probability distribution $p_1, \ldots, p_k$ is defined by $H(p_1, \ldots, p_j) = -\sum_{j=1}^{k} p_j \log_2 p_j$. The number of samples N required in the simulation satisfies*

$$H(p_1, \ldots, p_k) \leq \mathbf{E}N \leq H(p_1, \ldots, p_k) + 4.$$

For the proof, see my paper *Sharp entropy bounds for discrete statistical simulation (Stat. Prob. Lett. 42 (1999), 219–227)*, or section 5.11 of the book *Elements of Information Theory, 2nd Ed.* by T. M. Cover and J. A. Thomas, where a similar result is proved for a different simulation method for which $\mathbf{E}N$ can be bounded from above by the slightly better bound $H(p_1, \ldots, p_k) + 2$.

15.4 Simulating a general r.v. using a uniform r.v.

Theorem 3.10 from section 3.2, which we proved earlier for its theoretical importance, can be interpreted about a statement about how to simulate a random variable with a specified distribution F using a $U[0,1]$ random variable.

Simulation Method 4. Given a c.d.f. F and a $U[0,1]$-distributed random variable U, set $X = g(U)$, where

$$g(p) = \sup\{x \in \mathbb{R} : F(x) < p\} \qquad (0 < p < 1).$$

(this is the lower quantile function associated with F; see section 3.2).

Exercise 15.7. *Explain why the above method works and its connection to Theorem 3.10.*

15.5 Simulating an exponential r.v.

As an illustration of Simulation Method 4 described above, in the case where F is the $\text{Exp}(\lambda)$ distribution, the function g is easily computed to be $g(p) = -\frac{1}{\lambda}\log(1-p)$, so taking $X = -\frac{1}{\lambda}\log(1-U)$ produces an $\text{Exp}(\lambda)$ r.v. We can simplify this slightly by noting that if $U \sim U[0,1]$ then also $1-U \sim U[0,1]$, so one can use the simpler function $g(1-p) = -\frac{1}{\lambda}\log p$.

Simulation Method 5. Given a random variable $U \sim U[0,1]$, the random variable $X = -\frac{1}{\lambda}\log U$ has distribution $\text{Exp}(\lambda)$.

15.6 Simulating a normal r.v.

Although Simulation Method 4 is very general, in many practical cases it is hard or annoying to compute the associated quantile function $g(p)$ (which in the case of an absolutely continuous distribution amounts to inverting the c.d.f.). The normal distribution is an example where there exists a more practical method that is based instead on the polar decomposition of a standard bivariate normal vector (section 8.7).

Simulation Method 6. Given two independent r.v.s $U_1, U_2 \sim U[0,1]$, define

$$\begin{aligned}
\Theta &= 2\pi U_1, \\
R &= \sqrt{-2\log U_2}, \\
X &= R\cos\Theta, \\
Y &= R\sin\Theta.
\end{aligned}$$

Then $\Theta \sim U[0,2\pi]$, $R^2 \sim \text{Exp}(1/2)$, and R and Θ are independent. Therefore X, Y are independent and have the standard normal distribution $N(0,1)$.

15.7 Simulating a Poisson r.v.

Given a sequence $U_1, U_2, \ldots$ of i.i.d. $U[0,1]$ random variables, the following method simulates a $\text{Poi}(\lambda)$ r.v.

Simulation Method 7.

1. Sample the U_k's one at a time until the first time m when $M_m = \prod_{k=1}^m U_k$ satisfies $M_m \le e^{-\lambda}$.

2. Output $m-1$.

Exercise 15.8. *Explain why the output Z of this algorithm has distribution* $\text{Poi}(\lambda)$*, using the connection between the Poisson distribution and the cumulative sums of i.i.d. exponential random variables (see section 8.8).*

Note that the number of samples this method requires is equal to one plus the random variable Z being simulated. In particular, the mean number of samples required is $\mathbf{E}Z = \lambda + 1$.

Chapter 16: Additional topics

16.1 The Kolmogorov 0-1 law

In this section we prove a well-known and easy result due to Kolmogorov, the Kolmogorov 0-1 law. This will be useful in the next section.

Theorem 16.1 (Kolmogorov 0-1 law)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent random variables in some probability space $(\Omega, \mathcal{F}, \mathbf{P})$. An event $A \in \mathcal{F}$ is called a **tail event** if*

$$A \in \sigma(X_n, X_{n+1}, \ldots)$$

*for all n. That is, one can tell whether A occurred or not by looking at the sequence $X_1, X_2, \ldots$, and the occurrence of A is unaffected by changing a finite number of X_k's. (Note that the set of tail events is a σ-algebra, called the **tail σ-algebra** of the sequence, and denoted $\mathcal{T}$; formally, $\mathcal{T} = \cap_{n\geq 1}\sigma(X_n, X_{n+1}, \ldots)$). Then for any tail event A we have that $\mathbf{P}(A) = 0$ or $\mathbf{P}(A) = 1$.*

Proof. It will be enough to prove the somewhat strange statement that the tail event A is independent of itself, since in that case we'll have

$$\mathbf{P}(A) = \mathbf{P}(A \cap A) = \mathbf{P}(A)\mathbf{P}(A) = \mathbf{P}(A)^2,$$

which would imply the claim, since the only solutions to the equation $x^2 = x$ are $x = 0$ and $x = 1$. The fact that a tail event is independent of itself is a consequence of the following two claims: 1. If $A \in \sigma(X_1, \ldots, X_n)$ and $B \in \sigma(X_{n+1}, X_{n+2}, \ldots)$ then A and B are independent. This follows by observing that $\cup_{k\geq 1}\sigma(X_{n+1}, X_{n+2}, \ldots, X_{n+k})$ is a π-system generating $\sigma(X_{n+1}, X_{n+2}, \ldots)$, and for B in the π-system the claim is trivial. 2. If $A \in \sigma(X_1, X_2, \ldots)$ and $B \in \mathcal{T}$ then A and B are independent. The fact that is true when $A \in \sigma(X_1, \ldots, X_n)$ for some n follows from claim 1. But $\cup_{n\geq 1}\sigma(X_1, \ldots, X_n)$ is a π-system generating $\sigma(X_1, X_2, \ldots)$, so the same is true also for A in the generated σ-algebra. □

Example 16.2. The events $\{\lim_{n\to\infty} X_n \text{ exist}\}$ and $\{\sum_{n=1}^{\infty} X_n \text{ converges}\}$ are tail events.

16.2 The Kolmogorov three-series theorem

As a further application of the Lindeberg-Feller central limit theorem, we prove the Kolmogorov three-series theorem (Theorem 10.6).

Proof of Theorem 10.6. Denote $Y_n = X_n \mathbf{1}_{\{|X_n|\le 1\}}$. Assume that the eponymous three series

$$\sum_{n=1}^{\infty} \mathbf{P}(|X_n| > 1) = \sum_{n=1}^{\infty} \mathbf{P}(X_n \neq Y_n), \tag{17}$$

$$\sum_{n=1}^{\infty} \mathbf{E}(Y_n), \tag{18}$$

$$\sum_{n=1}^{\infty} \mathbf{V}(Y_n) \tag{19}$$

all converge. Applying Theorem 10.5 to the random variables $Y_n - \mathbf{E}(Y_n)$, from the assumption that (19) converges we infer that the series $\sum_{n=1}^{\infty} \big(Y_n - \mathbf{E}(Y_n)\big)$ converges almost surely. Since (18) converges, this also means that $\sum_n Y_n$ converges almost surely. Since (17) converges we get using the first Borel-Cantelli lemma that the event $\{X_n \neq Y_n \text{ i.o.}\}$ has probability 0. Outside this event, the series $\sum_n X_n$ converges if and only if $\sum_n Y_n$ converges, so $\sum_n X_n$ also converges almost surely.

Next, for the converse claim, assume that $\sum_n X_n$ converges with positive probability. By the Kolmogorov 0-1 law, it therefore converges a.s. We need to show that the series (17), (18), (19) converge.

First, note that (17) must converge, since if it doesn't, the second Borel-Cantelli lemma implies that the event $\{|X_n| > 1 \text{ i.o.}\}$ has probability one,which would imply that $\sum X_n$ diverges almost surely.

Second, assume by contradiction that (19) diverges, or equivalently that

$$v_n := \sum_{k=1}^{n} \mathbf{V}(Y_n) \to \infty \quad \text{as } n \to \infty.$$

We apply Theorem 14.2 to the triangular array

$$X_{n,k} = v_n^{-1/2} \left(Y_k - \mathbf{E}Y_k\right).$$

Assumptions 1, 2 and 3 of the theorem are trivially satisfied (with $\sigma^2 = \sum_{k=1}^{n} \mathbf{V}(X_{n,k}) = 1$ for all n). For assumption 4, note that $|X_{n,k}| \le 2v_n^{-1/2}$, so $X_{n,k}^2 \mathbf{1}_{\{|X_{n,k}|>\epsilon\}} = 0$ for all k if

$v_n > 4/\epsilon^2$, which holds for n large enough. Thus, the assumptions of the theorem are valid and hence we obtain the conclusion that $S_n = \sum_{k=1}^n X_{n,k} \implies N(0,1)$ as $n \to \infty$. Define a sequence (of numbers, not random variables)

$$t_n = v_n^{-1/2} \sum_{k=1}^n \mathbf{E} Y_k = S_n - v_n^{-1/2} \sum_{k=1}^n Y_k,$$

and note that on the event $\{\sum_n X_n \text{ converges}\}$ (in which case clearly $\sum_n Y_n$ also converges), which we assumed has probability 1, $S_n - t_n = v_n^{-1/2} \sum_{k=1}^n Y_k \to 0$ as $n \to \infty$. This is easily seen to be in contradiction to the convergence in distribution of S_n to a limiting distribution of a non-constant random variable. So we have shown that (19) converges.

Finally, having shown that (19) converges, we conclude from Theorem 10.5 that the series $\sum_n (Y_n - \mathbf{E}(Y_n))$ converges almost surely. We already saw that on the event where $\{\sum_n X_n \text{ converges}\}$, which we assumed has positive probability, also $\sum_n Y_n$ converges, and therefore the series $\sum_n \mathbf{E}(Y_n) = \sum \big(Y_n - (Y_n - \mathbf{E}(Y_n))\big)$ also converges.

□

16.3 The Poisson limit law

The Poisson distribution is the one-parameter family of distributions Poisson(λ), $\lambda > 0$, where $X \sim \text{Poisson}(\lambda)$ is a discrete random variable satisfying

$$\mathbf{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

This distribution arises naturally as measuring the number of successful coin tossing experiments when one performs a large number of independent coin tosses with equal bias, as long as the expected number of successes remains bounded away from 0 and ∞. That is, if $X_1, X_2, \ldots, X_n$ denote i.i.d. Bernoulli random variables with bias $p = p(n)$ (which is a function of $n0$, and $S_n = \sum_{k=1}^n$, then if $n \cdot p(n) \to \lambda$ as $n \to \infty$, it is a simple exercise to check that

$$S_n \implies \text{Poisson}(\lambda). \tag{20}$$

As in the case of the Central Limit Theorem, it turns out that the i.i.d. condition can be significantly weakened, so in fact the Poisson distribution arises almost universally as the limit law for counting numbers of “rare events,”, i.e., sums of indicator functions of many independent or weakly correlated events each of which has a small chance of success, in the asymptotic regime in which the expected number of successes converges to a constant. For this reason, the Poisson distribution is considered an excellent model for many real-life phenomena, e.g.:

- The number of typographical errors in a web page or book chapter.
- The number of radioactive particles emitted from a chunk of radioactive material in a unit of time.
- The number of cars passing a point on a remote desert road in an hour.
- The number of incandescent light-bulbs burning out in your house each month.
- Etc. (try to think of more examples from science or your daily life...)

Our goal is to prove the following generalization of (20), analogous to the Lindeberg-Feller extension of the Central Limit Theorem. The technique used in the proof is also similar and uses characteristic functions.

Theorem 16.3. *Let $(X_{n,m})_{1\le m\le n}$ be a triangular array of random variables such that for each $n \ge 1$, the random variables $X_{n,1}, \ldots, X_{n,n}$ are independent, and each $X_{n,m}$ satisfies $\mathbf{P}(X_{n,m} = 1) = p_{n,m}, \mathbf{P}(X_{n,m} = 0) = 1 - p_{n,m}$. Denote $S_n = \sum_{m=1}^n X_m$. Assume that the biases $(p_{n,m})_{n,m}$ satisfy the following conditions:*

(i) $\mathbf{E}S_n = \sum_{m=1}^n p_{n,m} \to \lambda \in (0,\infty)$ *as* $n \to \infty$.

(ii) $\max_{1\le m\le n} p_{n,m} \to 0$ *as* $n \to \infty$.

Then $S_n \implies \text{Poisson}(\lambda)$.

Proof. Let $\varphi_{n,m}(t) = \varphi_{X_{n,m}}(t) = \mathbf{E}(\exp(itX_{n,m})) = (1 - p_{n,m}) + p_{n,m}e^{it}$ denote the characteristic function of $X_{n,m}$. The ch. f. of S_n is then

$$\varphi_{S_n} = \prod_{k=1}^n \varphi_{n,m}(t) = \prod_{k=1}^n \left((1 - p_{n,m}) + p_{n,m}e^{it}\right).$$

Our goal is to show that $\varphi_{S_n}(t) \to \varphi_Z(t)$ where $Z \sim \text{Poisson}(\lambda)$. From a past homework computation we know that

$$\varphi_Z(t) = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} e^{ikt} = \exp(\lambda(e^{it}-1)) = \lim_{n\to\infty} \prod_{m=1}^{n} \exp(p_{n,m}(e^{it}-1)) \;\; \text{(by assumption (i))},$$

so let us investigate how closely $\prod_{m=1}^n \varphi_{n,m}(t)$ is approximated by $\prod_{m=1}^n \exp(p_{n,m}(e^{it}-1))$. Note that for any $0 \le p \le 1$, we have

$$\begin{aligned} |\exp(p(e^{it}-1))| &= \exp(p\,\text{Re}(e^{it}-1)) \le \exp(0) = 1, \\ |1-p+pe^{it}| &= |1+p(e^{it}-1)| \le 1 \;\; \text{(a convex combination of 1 and } e^{it}\text{)}. \end{aligned}$$

We therefore get using the exercise below that

$$\begin{aligned} &\left| \prod_{m=1}^{n} \exp\big(p_{n,m}(e^{it}-1)\big) - \prod_{m=1}^{n} (1+p_{n,m}(e^{it}-1)) \right| \\ &\qquad \le \sum_{m=1}^{n} |\exp\big(p_{n,m}(e^{it}-1)\big) - (1+p_{n,m}(e^{it}-1))| \\ &\qquad \le 10 \sum_{m=1}^{n} p_{n,m}^2 |e^{it}-1|^2 \\ &\qquad \le 40 \left(\max_{1\le m\le n} p_{n,m} \right) \sum_{m=1}^{n} p_{n,m} \xrightarrow[n\to\infty]{} 0. \end{aligned}$$

This is exactly what was needed to finish the proof. □

Exercise 16.4. *(i) Let $z_1, \ldots, z_n, w_1, \ldots, w_n$ be complex numbers such that $|z_m|, |w_m| \le 1$ for all $1 \le m \le n$. Prove that*

$$\left| \prod_{m=1}^{n} z_m - \prod_{m=1}^{n} w_m \right| \le \sum_{m=1}^{n} |z_m - w_m|.$$

(ii) Prove that if z is a complex number with $|z| \le 2$ then

$$|\exp(z) - (1+z)| \le 10|z|^2.$$

Exercises

1. (a) If $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space and $A, B \in \mathcal{F}$ are events such that $\mathbf{P}(B) \neq 0$, the **conditional probability of** A **given** B is denoted $\mathbf{P}(A|B)$ and defined by

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}.$$

Prove the **total probability formula:** if $A, B_1, B_2, \ldots, B_k \in \mathcal{F}$ such that Ω is the disjoint union of $B_1, \ldots, B_k$ and $\mathbf{P}(B_i) \neq 0$ for $1 \leq i \leq k$, then

$$\mathbf{P}(A) = \sum_{i=1}^{k} \mathbf{P}(B_i)\mathbf{P}(A|B_i). \tag{TPF}$$

(b) An urn initially contains one white ball and one black ball. At each step of the experiment, a ball is drawn at random from the urn, then put back and another ball of the same color is added. Prove that the number of white balls that are in the urn after N steps is a uniform random number in $\{1, 2, \ldots, N+1\}$. That is, the event that the number of white balls after step N is equal to k has probability $1/(N+1)$ for each $1 \leq k \leq N+1$. (Note: The idea is to use (TPF), but there is no need to be too formal about constructing the relevant probability space — you can assume an intuitive notion of probabilities.)

2. If $\Omega = \{1, 2, 3\}$, list all the possible σ-algebras of subsets of Ω.

3. Let $(\Omega, \mathcal{F})$ be a measurable space. A **pre-probability measure** is a function $\mathbf{P} : \mathcal{F} \to [0, 1]$ that satisfies

$$\mathbf{P}(\emptyset) = 0, \qquad \mathbf{P}(\Omega) = 1. \tag{P1}$$

If a pre-probability measure $\mathbf{P}$ satisfies

$$A_1, A_2, \ldots \in \mathcal{F} \text{ are pairwise disjoint } \implies \mathbf{P}(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbf{P}(A_n). \tag{P2}$$

then we say that it is **σ-additive**. If it satisfies

$$A_1, \ldots, A_n \in \mathcal{F} \text{ are pairwise disjoint } \implies \mathbf{P}(\cup_{k=1}^{n} A_k) = \sum_{k=1}^{n} \mathbf{P}(A_k). \tag{P3}$$

then we say that it is **additive**. Recall that we defined a **probability measure** to be a pre-probability measure that is σ-additive.

We say that a pre-probability measure satisfies the **continuity** properties if it satisfies

$$(A_n)_{n=1}^\infty \subset \mathcal{F}, \quad A_n \subset A_{n+1} \ \forall n \implies \mathbf{P}(\cup_{n=1}^\infty A_n) = \lim_{n\to\infty} \mathbf{P}(A_n), \tag{CONT1}$$

$$(A_n)_{n=1}^\infty \subset \mathcal{F}, \quad A_n \supset A_{n+1} \ \forall n \implies \mathbf{P}(\cap_{n=1}^\infty A_n) = \lim_{n\to\infty} \mathbf{P}(A_n). \tag{CONT2}$$

Prove that a probability measure satisfies the continuity properties, and that an additive pre-probability measure that satisfies the first continuity property (CONT1) ("continuity from below") is σ-additive, and is therefore a probability measure.

4. (a) A coin has some bias $p \in (0,1)$, so when tossed it comes up Heads with probability p, or Tails with probability $1-p$. Suppose the coin is tossed N times independently, and let $A_{N,k}$ denote the event that the result came up Heads exactly k times. Refresh your memory concerning why the Binomial Distribution Formula, which says that

$$\mathbf{P}(A_{N,k}) = \binom{N}{k} p^k (1-p)^{N-k}, \tag{BINOM}$$

is true. You may submit a short written explanation to test your understanding, but it is not required.

(b) A group of N prisoners is locked up in a prison, each in a separate cell with no ability to communicate with the other prisoners. Each cell contains a mysterious on/off electrical switch. One evening the warden visits each of the prisoners and presents them with the following dilemma: During the night each prisoner must choose whether to leave his switch in the on or off position. If at midnight *exactly one* of the switches is in the on position, all the prisoners will be set free in the morning; otherwise they will all be executed!

The prisoners cannot coordinate their actions, but they are all rational, know calculus and probability theory, and each is equipped with a random number generator. Find the strategy that the prisoners will take to maximize their chance of survival, and compute what that chance is, as a function of N and in the limit when N is very large. For extra fun, try to guess in advance how big or small you expect the survival likelihood to be, and see how your guess measures up to the actual result.

5. (a) Let Ω be a set, and let $\mathcal{S} = \{\mathcal{F}_i\}_{i \in I}$ be some collection of σ-algebras of subsets of Ω, indexed by some index set I (note that $\mathcal{S}$ is a *set* of *subsets* of *subsets* of Ω - try to avoid dizziness!). Prove that the intersection of all the $\mathcal{F}_i$'s (i.e., the collection of subsets of Ω that are elements of all the $\mathcal{F}_i$'s) is also a σ-algebra.

   (b) Let Ω be a set, and let $\mathcal{A}$ be a collection of subsets of Ω. Prove that there exists a unique σ-algebra $\sigma(\mathcal{A})$ of subsets of Ω that satisfies the following two properties:

   (i) $\mathcal{A} \subset \sigma(\mathcal{A})$ (in words, $\sigma(\mathcal{A})$ contains all the elements of $\mathcal{A}$).

   (ii) $\sigma(\mathcal{A})$ is the minimal σ-algebra satisfying property 1 above, in the sense that if $\mathcal{F}$ is any other σ-algebra that contains all the elements of $\mathcal{A}$, then $\sigma(\mathcal{A}) \subset \mathcal{F}$.

   The σ-algebra $\sigma(\mathcal{A})$ is called the **σ-algebra generated by $\mathcal{A}$**.

   Hint for (b). Let $(\mathcal{F}_i)_{i \in I}$ be the collection of all σ-algebras of subsets of Ω that contain $\mathcal{A}$. This is a non-empty collection, since it contains for example $\mathcal{P}(\Omega)$, the set of all subsets of Ω. Any σ-algebra $\sigma(\mathcal{A})$ that satisfies the two properties above is necessarily a subset of any of the $\mathcal{F}_i$'s, hence it is also contained in the intersection of all the $\mathcal{F}_i$'s, which is a σ-algebra by part (a) of the question.

6. (a) Let X be a random variable with distribution function F_X and piecewise continuous density function f_X. Let $[a,b] \subset \mathbb{R}$ be an interval (possibly infinite) such that

$$\mathbf{P}(X \in [a,b]) = 1,$$

   and let $g : [a,b] \to \mathbb{R}$ be a monotone (strictly) increasing and differentiable function. Prove that the random variable $Y = g(X)$ (this is the function on Ω defined by $Y(\omega) = g(X(\omega))$, in other words the composition of the two functions g and X) has density function

$$f_Y(x) = \begin{cases} \frac{f_X(g^{-1}(x))}{g'(g^{-1}(x))} & x \in (g(a), g(b)), \\ 0 & \text{otherwise.} \end{cases}$$

   (b) If $\lambda > 0$, we say that a random variable has the exponential distribution with parameter λ if

$$F_X(x) = \begin{cases} 0 & x < 0, \\ 1 - e^{-\lambda x} & x \geq 0, \end{cases}$$

and denote this $X \sim \mathrm{Exp}(\lambda)$. Find an algorithm to produce a random variable with $\mathrm{Exp}(\lambda)$ distribution using a random number generator that produces uniform random numbers in $(0, 1)$. In other words, if $U \sim U(0, 1)$, find a function $g : (0, 1) \to \mathbb{R}$ such that the random variable $X = g(U)$ has distribution $\mathrm{Exp}(\lambda)$.

(c) We say that a non-negative random variable $X \geq 0$ has the **lack of memory property** if it satisfies that

$$\mathbf{P}(X \geq t \mid X \geq s) = \mathbf{P}(X \geq t - s) \qquad \text{for all } 0 < s < t.$$

Prove that exponential random variables have the lack of memory property.

(d) Prove that any non-negative random variable that has the lack of memory property has the exponential distribution with some parameter $\lambda > 0$. (This is easier if one assumes that the function $G(x) = \mathbf{P}(X \geq x)$ is differentiable on $[0, \infty)$, so you can make this assumption if you fail to find a more general argument).

7. (a) Prove the **inclusion-exclusion principle**: If $A_1, \ldots, A_n$ are events in a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, then

$$\mathbf{P}\left(\bigcup_{k=1}^{n} A_k\right) = s_1 - s_2 + s_3 - s_4 + s_5 - \ldots + (-1)^{n-1} s_n,$$

where

$$\begin{aligned}
s_1 &= \mathbf{P}(A_1) + \mathbf{P}(A_2) + \ldots + \mathbf{P}(A_n) = \sum_{k=1}^{n} \mathbf{P}(A_k), \\
s_2 &= \sum_{1 \leq k_1 < k_2 \leq n} \mathbf{P}(A_{k_1} \cap A_{k_2}), \\
s_3 &= \sum_{1 \leq k_1 < k_2 < k_3 \leq n} \mathbf{P}(A_{k_1} \cap A_{k_2} \cap A_{k_3}), \\
&\ \vdots \\
s_d &= \sum_{1 \leq k_1 < \ldots < k_d \leq n} \mathbf{P}(A_{k_1} \cap A_{k_2} \cap \ldots \cap A_{k_d}), \\
&\ \vdots \\
s_n &= \mathbf{P}(A_1 \cap A_2 \cap \ldots \cap A_n).
\end{aligned}$$

(b) N letters addressed to different people are inserted at random into N envelopes that are labelled with the names and addresses of the N recipients, such that all $N!$ possible matchings between the letters and envelopes are equally likely. What is the probability of the event that no letter will arrive at its intended destination? Compute this probability for any N, and in the limit when $N \to \infty$.

8. Let F be the distribution function

$$F(x) = \begin{cases} 0 & x < 0, \\ \frac{1}{3} + \frac{1}{6}x & 0 \le x < 1, \\ \frac{1}{2} & 1 \le x < 2, \\ 1 - \frac{1}{4}e^{2-x} & x \ge 2. \end{cases}$$

Compute the lower and upper quantile functions of F, defined by

$$\begin{aligned} X_*(p) &= \sup\{x : F(x) < p\}, \\ X^*(p) &= \inf\{x : F(x) > p\}, \end{aligned} \qquad (0 < p < 1).$$

A recommended way is to plot F on paper and then figure out the quantiles by "eyeballing". Of course, the answer should be spelled out in precise formulas.

9. A drunken archer shoots at a target hanging on a wall 1 unit of distance away. Since he is drunk, his arrow ends up going in a random direction at an angle chosen uniformly in $(-\pi/2, \pi/2)$ (an angle of 0 means he will hit the target precisely) until it hits the wall. Ignoring gravity and the third dimension, compute the distribution function (and density function if it exists) of the random distance from the hitting point of the arrow to the target.

10. (a) Let $(\Omega_1, \mathcal{F}_1, \mathbf{P}_1)$ be a probability space, let $(\Omega_2, \mathcal{F}_2)$ be a measurable space, and let $f : \Omega_1 \to \Omega_2$ be a measurable function. Verify that the function $\mathbf{P}_2 : \mathcal{F}_2 \to [0, 1]$ defined by

$$\mathbf{P}_2(A) = \mathbf{P}_1(f^{-1}(A))$$

is a probability measure. This probability measure is called the **push-forward measure of $\mathbf{P}_1$ under f.**

(b) For a real number x, denote the **integer part of** x by

$$\lfloor x \rfloor = \sup\{n \in \mathbb{Z} : n \le x\},$$

and denote the **fractional part of** x by

$$\{x\} = x - \lfloor x \rfloor.$$

Let $((0,1), \mathcal{B}, \mathbf{P})$ be the unit interval with the σ-algebra of Borel subsets and the Lebesgue probability measure, corresponding to the experiment of choosing a uniform random number in $(0,1)$. Define a sequence of functions $R_1, R_2, \ldots : (0,1) \to \mathbb{R}$ by

$$R_n(x) = \begin{cases} 0 & 0 \le \{2^{n-1}x\} < 1/2, \\ 1 & 1/2 \le \{2^{n-1}x\} < 1. \end{cases}$$

For any $n \in \mathbb{N}$ and $a_1, a_2, \ldots, a_n \in \{0,1\}$, denote by $B_n(a_1, \ldots, a_n)$ the set

$$B_n(a_1, \ldots, a_n) = \{x \in (0,1) : R_1(x) = a_1, R_2(x) = a_2, \ldots, R_n(x) = a_n\}.$$

Find a good explicit description for this set ("the set of all x's such that ..."), and deduce from it that

$$\mathbf{P}(B_n(a_1, \ldots, a_n)) = \frac{1}{2^n}.$$

(c) Define a function $f : (0,1) \to \{0,1\}^{\mathbb{N}}$ by

$$f(x) = (R_1(x), R_2(x), R_3(x), \ldots).$$

Prove that f is a measurable function when the space $\{0,1\}^{\mathbb{N}}$ is equipped with the σ-algebra generated by the sets

$$A_n(1) = \{(x_1, x_2, \ldots) \in \{0,1\}^{\mathbb{N}} : x_n = 1\}.$$

(d) Prove that the push-forward of Lebesgue measure under f is the probability measure corresponding to the random experiment of an infinite sequence of fair coin tosses.

11. Let X be an exponential r.v. with parameter λ, i.e., $F_X(x) = (1-e^{-\lambda x})1_{[0,\infty)}(x)$. Define random variables

$$\begin{aligned} Y &= \lfloor X \rfloor := \sup\{n \in \mathbb{Z} : n \le x\} && (\text{“the integer part of } X\text{”}), \\ Z &= \{X\} := X - \lfloor X \rfloor && (\text{“the fractional part of } X\text{”}). \end{aligned}$$

(a) Compute the (1-dimensional) distributions of Y and Z (in the case of Y, since it's a discrete random variable it is most convenient to describe the distribution by giving the individual probabilities $\mathbf{P}(Y = n), n = 0, 1, 2, \ldots$; for Z one should compute either the distribution function or density function).

(b) Show that Y and Z are independent. (**Hint:** check that $\mathbf{P}(Y = n, Z \le t) = \mathbf{P}(Y = n)\mathbf{P}(Z \le t)$ for all n and t.)

12. (a) Let X, Y be independent r.v.'s. Define $U = \min(X, Y)$, $V = \max(X, Y)$. Find expressions for the distribution functions F_U and F_V in terms of the distribution functions of X and Y.

(b) Assume that $X \sim \text{Exp}(\lambda), Y \sim \text{Exp}(\mu)$ (and are independent as before). Prove that $\min(X, Y)$ has distribution $\text{Exp}(\lambda + \mu)$. Try to give an intuitive explanation in terms of the kind of real-life phenomena that the exponential distribution is intended to model (e.g., measuring the time for a light-bulb to burn out, or for a radioactive particle to be emitted from a chunk of radioactive material).

(c) Let $X_1, X_2, \ldots$ be a sequence of independent r.v.'s, all of them having distribution $\text{Exp}(1)$. For each $n \ge 1$ denote

$$M_n = \max(X_1, X_2, \ldots, X_n) - \log n.$$

Compute for each n the distribution function of M_n, and find the limit (if it exists)

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

13. If X, Y are r.v.'s with a joint density $f_{X,Y}$, the identity

$$\mathbf{P}((X, Y) \in A) = \iint_A f_{X,Y}(x, y)\, dx\, dy$$

holds for all "reasonable" sets $A \subset \mathbb{R}^2$ (in fact, for all Borel-measurable sets, but that requires knowing what that integral means for a set such as $\mathbb{R}^2 \setminus Q^2$...). In particular, if X, Y are independent and have respective densities f_X and f_Y, so $f_{X,Y}(x,y) = f_X(x) f_Y(y)$, then

$$F_{X+Y}(t) = \mathbf{P}(X+Y \le t) = \int_{-\infty}^{\infty} \int_{-\infty}^{t-x} f_X(x) f_Y(y)\, dy\, dx.$$

Differentiating with respect to t gives (assuming without justification that it is allowed to differentiate under the integral):

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(x) f_Y(t-x)\, dx.$$

Use this formula to compute the distribution of $X+Y$ when X and Y are independent r.v.'s with the following (pairs of) distributions:

(a) $X \sim U[0,1],\ Y \sim U[0,2]$.

(b) $X \sim \text{Exp}(1),\ Y \sim \text{Exp}(1)$.

(c) $X \sim \text{Exp}(1),\ -Y \sim \text{Exp}(1)$.

14. (a) Let $(A_n)_{n=1}^{\infty}$ be a sequence of events in a probability space. Show that

$$1_{\limsup A_n} = \limsup_n 1_{A_n}.$$

(The lim-sup on the left refers to the lim-sup operation on events; on the right it refers to the lim-sup of a sequence of functions; the identity is an identity of real-valued functions on Ω, i.e., should be satisfied for each individual point $\omega \in \Omega$ in the sample space). Similarly, show (either separately or by relying on the first claim) that

$$1_{\liminf A_n} = \liminf_n 1_{A_n}.$$

(b) Let U be a uniform random variable in $(0,1)$. For each $n \ge 1$ define an event A_n by

$$A_n = \{U < 1/n\}.$$

Note that $\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \infty$. However, compute $\mathbf{P}(A_n \text{ i.o.})$ and show that the conclusion of the second Borel-Cantelli lemma does not hold (of course, one of the assumptions of the lemma also doesn't hold, so there's no contradiction).

15. If P, Q are two probability measures on a measurable space $(\Omega, \mathcal{F})$, we say that P **is absolutely continuous with respect to** Q, and denote this $P << Q$, if for any $A \in \mathcal{F}$, if $Q(A) = 0$ then $P(A) = 0$.

    Prove that $P << Q$ if and only if for any $\epsilon > 0$ there exists a $\delta > 0$ such that if $A \in \mathcal{F}$ and $Q(A) < \delta$ then $P(A) < \epsilon$.

    Hint. Apply a certain famous lemma.

    Note. The intuitive meaning of the relation $P << Q$ is as follows: suppose there is a probabilistic experiment, and we are told that one of the measures P or Q governs the statistical behavior of the outcome, but we don't know which one. (This is a situation that arises frequently in real-life applications of probability and statistics.) All we can do is perform the experiment, observe the result, and make a guess. If $P << Q$, any event which is observable with positive probability according to P also has positive Q-probability, so we can never rule out Q as the correct measure, although we may get an event with $Q(A) > 0$ and $P(A) = 0$ that enables us to rule out P. If we also have the symmetric relation $Q << P$, then we can't rule out either of the measures.

16. A function $\varphi : (a, b) \to \mathbb{R}$ is called **convex** if for any $x, y \in (a, b)$ and $\alpha \in [0, 1]$ we have

$$\varphi(\alpha x + (1 - \alpha)y) \leq \alpha\varphi(x) + (1 - \alpha)\varphi(y).$$

    (a) Prove that an equivalent condition for φ to be convex is that for any $x < z < y$ in (a, b) we have

$$\frac{\varphi(z) - \varphi(x)}{z - x} \leq \frac{\varphi(y) - \varphi(z)}{y - z}.$$

    Deduce using the mean value theorem that if φ is twice continuously differentiable and satisfies $\varphi'' \geq 0$ then it is convex.

    (b) Prove **Jensen's inequality**, which says that if X is a random variable such that $\mathbf{P}(X \in (a, b)) = 1$ and $\varphi : (a, b) \to \mathbb{R}$ is convex, then

$$\varphi(\mathbf{E}X) \leq \mathbf{E}(\varphi(X)).$$

    Hint. Start by proving the following property of a convex function: If φ is convex then at any point $x_0 \in (a, b)$, φ has a **supporting line**, that is, a linear function

$y(x) = ax + b$ such that $y(x_0) = \varphi(x_0)$ and such that $\varphi(x) \geq y(x)$ for all $x \in (a, b)$ (to prove its existence, use the characterization of convexity from part (a) to show that the left-sided derivative of φ at x_0 is less than or equal to the right-sided derivative at x_0; the supporting line is a line passing through the point $(x_0, \varphi(x_0))$ whose slope lies between these two numbers). Now take the supporting line function at $x_0 = \mathbf{E}X$ and see what happens.

17. If X is a random variable satisfying $a \leq X \leq b$, prove that

$$\mathbf{V}(X) \leq \frac{(b-a)^2}{4},$$

and identify when equality holds.

18. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. (independent and identically distributed) random variables with distribution $U(0, 1)$. Define events $A_1, A_2, \ldots$ by

$$A_n = \{X_n = \max(X_1, X_2, \ldots, X_n)\}$$

(if A_n occurred, we say that n is a **record time**).

(a) Prove that $A_1, A_2, \ldots$ are independent events.

Hint. for each $n \geq 1$, let π_n be the random permutation of $(1, 2, \ldots, n)$ obtained by forgetting the values of $(X_1, \ldots, X_n)$ and only retaining their respective order. In other words, define

$$\pi_n(k) = \#\{1 \leq j \leq n : X_j \leq X_k\}.$$

By considering the joint density $f_{X_1,\ldots,X_n}$ (a uniform density on the n-dimensional unit cube), show that π_n is a uniformly random permutation of n elements, i.e. $\mathbf{P}(\pi_n = \sigma) = 1/n!$ for any permutation $\sigma \in S_n$. Deduce that the event $A_n = \{\pi_n(n) = n\}$ is independent of π_{n-1} and therefore is independent of the previous events $(A_1, \ldots, A_{n-1})$, which are all determined by π_{n-1}.

(b) Define

$$R_n = \sum_{k=1}^{n} 1_{A_k} = \#\{1 \leq k \leq n : k \text{ is a record time}\}, \qquad (n = 1, 2, \ldots).$$

Compute $\mathbf{E}(R_n)$ and $\mathbf{V}(R_n)$. Deduce that if $(m_n)_{n=1}^{\infty}$ is a sequence of positive numbers such that $m_n \uparrow \infty$, however slowly, then the number R_n of record times up to time n satisfies

$$\mathbf{P}\left(|R_n - \log n| > m_n\sqrt{\log n}\right) \xrightarrow[n\to\infty]{} 0.$$

19. Compute $\mathbf{E}(X)$ and $\mathbf{V}(X)$ when X is a random variable having each of the following distributions:

    (a) $X \sim \text{Binomial}(n,p)$.

    (b) $X \sim \text{Poisson}(\lambda)$, i.e., $\mathbf{P}(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}$, $(k = 0,1,2,\ldots)$.

    (c) $X \sim \text{Geom}(p)$, i.e,. $\mathbf{P}(X=k) = p(1-p)^{k-1}$, $(k = 1,2,\ldots)$.

    (d) $X \sim U\{1,2,\ldots,n\}$ (the discrete uniform distribution on $\{1,2,\ldots,n\}$).

    (e) $X \sim U(a,b)$ (the uniform distribution on the interval (a,b)).

    (f) $X \sim \text{Exp}(\lambda)$

20. (a) If X, Y are independent r.v.'s taking values in $\mathbb{Z}$, show that

$$\mathbf{P}(X+Y=n) = \sum_{k=-\infty}^{\infty} \mathbf{P}(X=k)\mathbf{P}(Y=n-k) \qquad (n \in \mathbb{Z})$$

(compare this formula with the convolution formula in the case of r.v.'s with density).

(b) Use this to show that if $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ are independent then $X+Y \sim \text{Poisson}(\lambda+\mu)$. (Recall that for a parameter $\lambda > 0$, we say that $X \sim \text{Poisson}(\lambda)$ if $\mathbf{P}(X=k) = e^{-\lambda}\lambda^k/k!$ for $k = 0,1,2,\ldots$).

(c) Use the same "discrete convolution" formula to prove directly that if $X \sim \text{Bin}(n,p)$ and $Y \sim \text{Bin}(m,p)$ are independent then $X+Y \sim \text{Bin}(n+m,p)$. You may make use of the combinatorial identity (known as the Vandermonde identity or Chu-Vandermonde identity)

$$\sum_{j=0}^{k} \binom{n}{j}\binom{m}{k-j} = \binom{n+m}{k}, \qquad (n,m \geq 0,\ 0 \leq k \leq n+m).$$

As a bonus, try to find a direct combinatorial proof for this identity. An amusing version of the answer can be found at:

`http://en.wikipedia.org/wiki/Vandermonde's_identity`.

21. Prove that if X is a random variable that is independent of itself, then X is a.s. constant, i.e., there is a constant $c \in \mathbb{R}$ such that $\mathbf{P}(X = c) = 1$.

22. (a) If $X \ge 0$ is a nonnegative r.v. with distribution function F, show that

$$\mathbf{E}(X) = \int_0^\infty \mathbf{P}(X \ge x)\, dx.$$

(b) Prove that if $X_1, X_2, \ldots,$ is a sequence of independent and identically distributed ("i.i.d.") r.v.'s, then

$$\mathbf{P}(|X_n| \ge n \text{ i.o.}) = \begin{cases} 0 & \text{if } \mathbf{E}|X_1| < \infty, \\ 1 & \text{if } \mathbf{E}|X_1| = \infty. \end{cases}$$

(c) Deduce the following converse to the Strong Law of Large Numbers in the case of undefined expectations: If $X_1, X_2, \ldots$ are i.i.d. and $\mathbf{E}X_1$ is undefined (meaning that $\mathbf{E}X_{1+} = \mathbf{E}X_{1-} = \infty$) then

$$\mathbf{P}\left(\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n X_k \text{ does not exist} \right) = 1.$$

23. Let X be a r.v. with finite variance, and define a function $M(t) = \mathbf{E}|X - t|$, the "mean absolute deviation of X from t". The goal of this question is to show that the function $M(t)$, like its easier to understand and better-behaved cousin, $\mathbf{E}(X-t)^2$ (the "moment of inertia" around t, which by the Huygens-Steiner theorem is simply a parabola in t, taking its minimum value of $\mathbf{V}(X)$ at $t = \mathbf{E}X$), also has some unexpectedly nice propreties.

(a) Prove that $M(t) \ge |t - \mathbf{E}X|$.

(b) Prove that $M(t)$ is a convex function.

(c) Prove that

$$\int_{-\infty}^\infty \Big(M(t) - |t - \mathbf{E}X| \Big) dt = \mathbf{V}(X)$$

(see hints below). Deduce in particular that $M(t) - |t - \mathbf{E}X| \xrightarrow[t\to\pm\infty]{} 0$ (again under the assumption that $\mathbf{V}(X) < \infty$). If it helps, you may assume that X has a density f_X.

(d) Prove that if t_0 is a (not necessarily unique) minimum point of $M(t)$, then t_0 is a median (that is, a 0.5-quantile) of X.

(e) Optionally, draw (or, at least, imagine) a diagram showing the graphs of the two functions $M(t)$ and $|t - \mathbf{E}X|$ illustrating schematically the facts (a)–(d) above.

Hints. For (c), assume first (without loss of generality - why?) that $\mathbf{E}X = 0$. Divide the integral into two integrals, on the positive real axis and the negative real axis. For each of the two integrals, by decomposing $|X - t|$ into a sum of its positive and negative parts and using the fact that $\mathbf{E}X = 0$ in a clever way, show that one may replace the integrand $(\mathbf{E}|X - t| - |t|)$ by a constant multiple of either $\mathbf{E}(X - t)_+$ or $\mathbf{E}(X - t)_-$, and proceed from there.

For (d), first, develop your intuition by plotting the function $M(t)$ in a couple of cases, for example when $X \sim \text{Binom}(1, 1/2)$ and when $X \sim \text{Binom}(2, 1/2)$. Second, if $t_0 < t_1$, plot the graph of the function $x \to \frac{|x-t_1|-|x-t_0|}{t_1-t_0}$, and deduce from this a formula for $M'(t_0+)$ and (by considering $t_1 < t_0$ instead) a similar formula for $M'(t_0-)$, the right- and left-sided derivatives of M at t_0, respectively. On the other hand, think how the condition that t_0 is a minimum point of $M(t)$ can be expressed in terms of these one-sided derivatives.

24. (a) Let $\Gamma(t)$ denote the *Euler gamma function*, defined by

$$\Gamma(t) = \int_0^\infty e^{-x} x^{t-1}\, dx, \qquad (t > 0).$$

Show that the special value $\Gamma(1/2) = \sqrt{\pi}$ of the gamma function is equivalent to the integral evaluation $\sqrt{2\pi} = \int_{-\infty}^{\infty} e^{-x^2/2}\, dx$ (which is equivalent to the standard normal density being a density function).

(b) Prove that the Euler gamma function satisfies for all $t > 0$ the identity

$$\Gamma(t+1) = t\,\Gamma(t).$$

(This identity immediately implies the fact that $\Gamma(n+1) = n!$ for integer $n \geq 0$.)

(c) Find a formula for the values of $\Gamma(\cdot)$ at half-integers, that is,

$$\Gamma\left(n+\tfrac{1}{2}\right) = ?, \qquad (n \geq 0).$$

25. Compute $\mathbf{E}X^n$ when $n \geq 0$ is an integer and X has each of the following distributions:

    (a) $X \sim U(a,b)$

    (b) $X \sim \text{Exp}(\lambda)$

    (c) $X \sim \text{Gamma}(\alpha, \lambda)$, i.e. $f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1}, \ \ (x>0)$.

    (d) $X \sim \text{Beta}(a,b)$, i.e. $f_X(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \ \ (0<x<1)$, where

    $$B(a,b) = \int_0^1 u^{a-1}(1-u)^{b-1}\,du = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

    is the *Euler beta function.*

    (e) $X \sim N(0,1)$. In this case, identify $\mathbf{E}X^n$ combinatorially as the number of **matchings** of a set of size n into pairs (for example, if a university dorm has only 2-person housing units, then when n is even this is the number of ways to divide n students into pairs of roommates; no importance is given to the ordering of the pairs).

    (f) $X \sim N(1,1)$. In this case, identify $\mathbf{E}X^n$ combinatorially as the number of **involutions** (permutations which are self-inverse) of a set of n elements. To count the involutions, it is a good idea to divide them into classes according to how many fixed points they have. (Note: the expression for $\mathbf{E}(X^n)$ may not have a very simple form.)

26. Let $f : [0,1] \to \mathbb{R}$ be a continuous function. Prove that

    $$\int_0^1 \int_0^1 \ldots \int_0^1 f\left(\frac{x_1+x_2+\ldots+x_n}{n}\right) dx_1\,dx_2 \ldots dx_n \xrightarrow[n\to\infty]{} f(1/2).$$

    Hint. Interpret the left-hand side as an expected value; use the laws of large numbers.

27. A bowl contains n spaghetti noodles arranged in a chaotic fashion. Bob performs the following experiment: he picks two random ends of noodles from the bowl (chosen

uniformly from the $2n$ possible ends), ties them together, and places them back in the bowl. Then he picks at random two more ends (from the remaining $2n-2$), ties them together and puts them back, and so on until no more loose ends are left.

Let L_n denote the number of **spaghetti loops** at the end of this process (a loop is a chain of one or more spaghettis whose ends are tied to each other to form a cycle). Compute $\mathbf{E}(L_n)$ and $\mathbf{V}(L_n)$. Find a sequence of numbers $(b_n)_{n=1}^{\infty}$ such that

$$\frac{L_n}{b_n} \xrightarrow[n\to\infty]{\mathbf{P}} 1,$$

if such a sequence exists.

28. Martians communicate in a binary language with two symbols, 0 and 1. A text of length n symbols written in the Martian language looks like a sequence $X_1, X_2, \ldots, X_n$ of i.i.d. random symbols, each of which is 1 with probability p and 0 with probability $1-p$. Here, $p \in (0,1)$ is a parameter (the "Martian bias").

Define the **entropy function** $H(p)$ by

$$H(p) = -p\log_2 p - (1-p)\log_2(1-p).$$

Prove the following result that effectively says that if n is large, then with high probability a Martian text of length n can be encoded into an ordinary (man-made) computer file of length approximately $n \cdot H(p)$ computer bits (note that if $p \neq 1/2$ then this is smaller than n, meaning that the text can be compressed by a linear factor):

Theorem. *Let $X_1, X_2, X_3, \ldots$ be a sequence of i.i.d. Martian symbols (i.e., Bernoulli variables with bias p). Denote by $\boldsymbol{T}_n = (X_1, \ldots, X_n)$ the Martian text comprising the first n symbols. For any $\epsilon > 0$, if n is sufficiently large, the set $\{0,1\}^n$ of possible texts of length n can be partitioned into two disjoint sets,*

$$\{0,1\}^n = A_n \cup B_n,$$

such that the following statements hold:

(a) $\mathbf{P}(\boldsymbol{T}_n \in B_n) < \epsilon$

(b) $2^{n(H(p)-\epsilon)} \leq |A_n| \leq 2^{n(H(p)+\epsilon)}$.

Notes: The texts in B_n can be thought of as the "exceptional sequences" – they are the Martian texts of length n that are rarely observed. The texts in A_n are called "typical sequences". Because of the two-sided bounds the theorem gives on the number of typical sequences, it follows that we can encode them in a computer file of size approximately $nH(p)$ bits, provided we prepare in advance a "code" that translates the typical sequences to computer files of the appropriate size (this can be done algorithmically, for example by making a list of all the typical sequences sorted in lexicographic order, and matching them to successive binary strings of length $(H(p)+\epsilon)n)$.

Hint. To prove the theorem, let P_n be the random variable given by

$$P_n = \prod_{k=1}^{n} \left(p^{X_k}(1-p)^{1-X_k} \right).$$

Note that P_n measures the probability of the sequence that was observed up to time n. (Somewhat unusually, in this problem the probability itself is thought of as a random variable). Try to represent P_n in terms of cumulative sums of a sequence of i.i.d. random variables. Apply the Weak Law of Large Numbers to that sequence, and see where that gets you.

29. Prove the following one-sided version of Chebyshev's inequality: For any r.v. X and $t \geq 0$,

$$\mathbf{P}(X - \mathbf{E}X \geq t) \leq \frac{\sigma^2(X)}{t^2 + \sigma^2(X)}.$$

Hint. Assume without loss of generality that $\mathbf{E}X = 0$. For any $a > 0$, we have that $\mathbf{P}(X \geq t) \leq \mathbf{P}((X+a)^2 \geq (a+t)^2)$. Bound this using known methods and then look for the value of a that gives the best bound.

30. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. r.v.'s with distribution Exp(1). Prove that

$$\mathbf{P}\left(\limsup_{n\to\infty} \frac{X_n}{\log n} = 1 \right) = 1.$$

31. Let $A = (X_{i,j})_{i,j=1}^n$ be a random $n \times n$ matrix of i.i.d. random signs (i.e., random variables such that $\mathbf{P}(X_{i,j} = -1) = \mathbf{P}(X_{i,j} = 1) = 1/2$). Compute Var(det($A$)).

32. (a) Read, in Durrett's book (p. 63 in the 3rd edition) or on Wikipedia, the statement and proof of **Kronecker's lemma**.

    (b) Deduce from this lemma, using results we learned in class, the following rate of convergence result for the Strong Law of Large Numbers in the case of a finite variance: If $X_1, X_2, \ldots$ is an i.i.d. sequence such that $\mathbf{E}X_1 = 0$, $\mathbf{V}(X_1) < \infty$, and $S_n = \sum_{k=1}^n X_k$, then for any $\epsilon > 0$,

$$\frac{S_n}{n^{1/2+\epsilon}} \xrightarrow[n\to\infty]{\text{a.s.}} 0.$$

    Notes. When X_1 is a "random sign", i.e., a random variable that takes the values $-1, +1$ with respective probabilities $1/2, 1/2$, the sequence of cumulative sums $(S_n)_{n=1}^\infty$ is often called a **(symmetric) random walk on** $\mathbb{Z}$, since it represents the trajectory of a walker starting from 0 and taking a sequence of independent jumps in a random (positive or negative) direction. An interesting question concerns the rate at which the random walk can drift away from its starting point. By the SLLN, it follows that almost surely, $S_n = o(n)$, so the distance of the random walk from the origin almost surely has sub-linear growth. By the exercise above, the stronger result $S_n = o(n^{1/2+\epsilon})$ also holds for all ϵ. This is close to optimal, since by the Central Limit Theorem which we will discuss soon, one cannot hope to show that $S_n = o(n^{1/2})$. In fact, the "true" rate of growth is given by the following famous theorem, whose proof is a (somewhat complicated) elaboration on the techniques we have discussed.

    Theorem (The Law of the Iterated Logarithm (A. Y. Khinchin, 1924))**.**

$$\mathbf{P}\left(\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = 1\right) = 1.$$

    Therefore, by symmetry, also

$$\mathbf{P}\left(\liminf_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = -1\right) = 1.$$

    It follows in particular that, almost surely, the random walk will cross the origin infinitely many times.

33. Prove that if F and $(F_n)_{n=1}^\infty$ are distribution functions, F is continuous, and for any $t \in \mathbb{R}$ we have $F_n(t) \to F(t)$ as $n \to \infty$, then the convergence is uniform in t.

34. Let $\varphi(x) = (2\pi)^{-1/2}e^{-x^2/2}$ be the standard normal density function.

(a) If $X_1, X_2, \ldots$ are i.i.d. Poisson(1) random variables and $S_n = \sum_{k=1}^n X_k$ (so $S_n \sim$ Poisson(n)), show that if n is large and k is an integer such that $k \approx n + x\sqrt{n}$ then

$$\mathbf{P}(S_n = k) \approx \frac{1}{\sqrt{n}}\varphi(x).$$

Hint. Use the fact that $\log(1+u) = u - u^2/2 + O(u^3)$ as $u \to 0$.

(b) Find $\lim_{n\to\infty} e^{-n} \sum_{k=0}^n \frac{n^k}{k!}$.

(c) If $X_1, X_2, \ldots$ are i.i.d. Exp(1) random variables and denote $S_n = \sum_{k=1}^n X_k$ (so $S_n \sim$ Gamma$(n, 1)$), $\hat{S}_n = (S_n - n)/\sqrt{n}$. Show that if n is large and $x \in \mathbb{R}$ is fixed then the density of $\hat{S}_n$ satisfies

$$f_{\hat{S}_n}(x) \approx \varphi(x).$$

4. Prove that if $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are independent r.v.'s, then $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Hint. First, show why it is enough to prove the following statement: If $U, V \sim N(0,1)$ are independent and $a^2 + b^2 = 1$, then $W = aU + bV \sim N(0,1)$. Then, to prove this, introduce another auxiliary variable $Z = -bU + aV$, and consider the two-dimensional transformation $(U, V) \to (W, Z)$. Apply the formula

$$f_{\phi(U,V)}(w, z) = \frac{1}{|J_\phi(\phi^{-1}(w, z))|} f_{U,V}(\phi^{-1}(w, z))$$

for the density of a transformed random 2-d vector to get the joint density of W, Z.

35. **(a)** Prove that if $X, (X_n)_{n=1}^\infty$ are random variables such that $X_n \to X$ in probability then $X_n \implies X$.

(b) Prove that if $X_n \implies c$ where $c \in \mathbb{R}$ is a constant, then $X_n \to c$ in probability.

(c) Prove that if $Z, (X_n)_{n=1}^\infty, (Y_n)_{n=1}^\infty$ are random variables such that $X_n \implies Z$ and $X_n - Y_n \to 0$ in probability, then $Y_n \implies Z$.

36. **(a)** Let $X, (X_n)_{n=1}^{\infty}$ be integer-valued r.v.'s. Show that $X_n \implies X$ if and only if $\mathbf{P}(X_n = k) \to \mathbf{P}(X = k)$ for any $k \in \mathbb{Z}$.

(b) If $\lambda > 0$ is a fixed number, and for each n, Z_n is a r.v. with distribution Binomial$(n, \lambda/n)$, show that

$$Z_n \implies \text{Poisson}(\lambda).$$

37. Let $f(x) = (2\pi)^{-1/2}e^{-x^2/2}$ be the density function of the standard normal distribution, and let $\Phi(x) = \int_{-\infty}^{x} f(u)\,du$ be its c.d.f. Prove the inequalities

$$\frac{1}{x + x^{-1}} f(x) \le 1 - \Phi(x) \le \frac{1}{x} f(x), \qquad (x > 0). \tag{21}$$

Note that for large x this gives a very accurate two-sided bound for the tail of the normal distribution. In fact, it can be shown that

$$1 - \Phi(x) = f(x) \cdot \cfrac{1}{x + \cfrac{1}{x + \cfrac{2}{x + \cfrac{3}{x + \cfrac{4}{x + \ldots}}}}}$$

which gives a relatively efficient method of estimating $\Phi(x)$.

Hint. To prove the upper bound in (21), use the fact that for $t > x$ we have $e^{-t^2/2} \le (t/x)e^{-t^2/2}$. For the lower bound, use the identity

$$\frac{d}{dx}\left(\frac{e^{-x^2/2}}{x}\right) = -\left(1 + \frac{1}{x^2}\right)e^{-x^2/2}$$

to compute $\int_x^\infty (1+u^{-2})e^{-u^2/2}\,du$. On the other hand, show that this integral is bounded from above by $(1 + x^{-2})\int_x^\infty e^{-u^2/2}\,du$.

38. (a) Let $X_1, X_2, \ldots$ be a sequence of independent r.v.'s that are uniformly distributed on $\{1, \ldots, n\}$. Define

$$T_n = \min\{k : X_k = X_m \text{ for some } m < k\}.$$

If the X_j's represent the birthdays of some sequence of people on a planet in which the calendar year has n days, then T_n represents the number of people in the list who have to declare their birthdays before two people are found to have the same birthday. Show that

$$\mathbf{P}(T_n > k) = \prod_{m=1}^{k-1} \left(1 - \frac{m}{n}\right), \qquad (k \geq 2),$$

and use this to prove that

$$\frac{T_n}{\sqrt{n}} \implies F_{\text{birthday}},$$

where F_{birthday} is the distribution function defined by

$$F_{\text{birthday}}(x) = \begin{cases} 0 & x < 0, \\ 1 - e^{-x^2/2} & x \geq 0 \end{cases}$$

(note: this is not the same as the normal distribution!)

(b) Take $n = 365$. Assuming that the approximation $F_{T_n/\sqrt{n}} \approx F_{\text{birthday}}$ is good for such a value of n, estimate what is the minimal number of students that have to be put into a classroom so that the probability that two of them have the same birthday exceeds 50%. (Ignore leap years, and assume for simplicity that birthdays are distributed uniformly throughout the year; in practice this is not entirely true.)

39. Consider the following two-step experiment: First, we choose a uniform random variable $U \sim U(0,1)$. Then, conditioned on the event $U = u$, we perform a sequence of n coin tosses with bias u, i.e., we have a sequence $X_1, X_2, \ldots, X_n$ such that conditioned on the event $U = u$, the X_k's are independent and have distribution $\text{Binom}(1, u)$. (Note: without this conditioning, the X_k's are not independent!)

Let $S_n = \sum_{k=1}^n X_k$. Assume that we know that $S_n = k$, but don't know the value of U. What is our subjective estimate of the probability distribution of U given this information? Show that the conditional distribution of U given that $S_n = k$ is the beta distribution $\text{Beta}(k+1, n-k+1)$. In other words, show that

$$\mathbf{P}(U \leq x \mid S_n = k) = \frac{1}{B(k, n-k)} \int_0^x u^k (1-u)^{n-k}\, du, \qquad (0 \leq x \leq 1).$$

Note: This problem has been whimsically suggested by Laplace in the 18th century as a way to estimate the probability that the sun will rise tomorrow, given the knowledge that it has risen in the last n days. (Of course, this assumes the unlikely theological scenario whereby at the dawn of history, a $U(0,1)$ random number U was drawn, and that subsequently, every day an independent experiment was performed with probability U of success, such that if the experiment is successful then the sun rises.)

Hint. Use the following density version of the total probability formula: If A is an event and X is a random variable with density f_X, then

$$\mathbf{P}(A) = \int_{\mathbb{R}} f_X(u)\mathbf{P}(A \mid X = u)\, du.$$

Note that we have not defined what it means to condition on a 0-probability event (this is a somewhat delicate subject that we will not discuss in this quarter) — but don't worry about it, it is possible to use the formula in computations anyway and get results.

40. Let $Z_1, Z_2, \ldots$ be a sequence of i.i.d. random variables with the standard normal $N(0,1)$ distribution. For each n, define the random vector

$$\mathbf{X}_n = (X_{n,1}, \ldots, X_{n,n}) = \frac{1}{\left(\sum_{i=1}^n Z_i^2\right)^{1/2}}(Z_1, \ldots, Z_n)$$

(a) The distribution of the random vector $\mathbf{X}_n$ is called the *uniform distribution on the* $(n-1)$*-dimensional sphere* $S^{n-1} = \{x \in \mathbb{R}^n : ||x|| = 1\}$. Explain why this makes intuitive sense, and if possible explain rigorously what conditions this distribution satisfies that justifies describing it by this name.

(b) Show that $\sqrt{n}X_{n,1} \implies N(0,1)$ as $n \to \infty$.

Hint. Use the law of large numbers.

(c) For each $n \geq 1$, find the density function of the coordinate $X_{n,1}$. Optionally, use this to give an alternative solution to part (b) above.

Hint. Do it first for $n = 2$ and $n = 3$, and generalize using ideas from multivariate calculus. For $n = 3$, you should find that $X_{3,1} \sim U[-1,1]$, a geometric fact which was known to Archimedes.

41. Compute the characteristic functions for the following distributions.

(a) **Poisson distribution:** $X \sim \text{Poisson}(\lambda)$.

(b) **Geometric distribution:** $X \sim \text{Geom}(p)$ (assume a geometric that starts at 1).

(c) **Uniform distribution:** $X \sim U[a,b]$, and in particular $X \sim [-1,1]$ which is especially symmetric and useful in applications.

(d) **Exponential distribution:** $X \sim \text{Exp}(\lambda)$.

(e) **Symmetrized exponential:** A r.v. Z with density function $f_Z(x) = \frac{1}{2}e^{-|x|}$. Note that this is the distribution of the exponential distribution after being "symmetrized" in either of two ways: (i) We showed that if $X, Y \sim \text{Exp}(1)$ are independent then $X - Y$ has density $\frac{1}{2}e^{-|x|}$; (ii) alternatively, it is the distribution of an "exponential variable with random sign", namely $\varepsilon \cdot X$ where $X \sim \text{Exp}(1)$ and ε is a random sign (same as the coin flip distribution mentioned above) that is independent of X.

42. **(a)** If X is a r.v., show that $\text{Re}(\varphi_X)$ (the real part of φ_X) and $|\varphi_X|^2 = \varphi_X\overline{\varphi_X}$ are also characteristic functions (i.e., construct r.v.'s Y and Z such that $\varphi_Y(t) = \text{Re}(\varphi_X(t))$, $\varphi_Z(t) = |\varphi_X(t)|^2$).

(b) Show that X is equal in distribution to $-X$ if and only if φ_X is a real-valued function.

43. **(a)** Let $Z_1, Z_2, \ldots$ be a sequence of independent r.v.'s such that the random series $X = \sum_{n=1}^\infty Z_n$ converges a.s. Prove that

$$\varphi_X(t) = \prod_{n=1}^\infty \varphi_{Z_n}(t), \qquad (t \in \mathbb{R}).$$

(b) Let X be a uniform r.v. in $(0,1)$, and let $Y_1, Y_2, \ldots$ be the (random) bits in its binary expansion, i.e. each Y_n is either 0 or 1, and the equation

$$X = \sum_{n=1}^\infty \frac{Y_n}{2^n} \tag{22}$$

holds. Show that $Y_1, Y_2, \ldots$ are i.i.d. unbiased coin tosses (i.e., taking values $0, 1$ with probabilities $1/2, 1/2$).

(c) Compute the characteristic function φ_Z of $Z = 2X-1$ (which is uniform in $(-1,1)$). Use (22) to represent this in terms of the characteristic functions of the Y_n's (note that the series (22) converges absolutely, so here there is no need to worry about almost sure convergence). Deduce the infinite product identity

$$\frac{\sin(t)}{t} = \prod_{n=1}^{\infty} \cos\left(\frac{t}{2^n}\right), \qquad (t \in \mathbb{R}). \tag{23}$$

(d) Substitute $t = \pi/2$ in (23) to get the identity

$$\frac{2}{\pi} = \frac{\sqrt{2}}{2} \cdot \frac{\sqrt{2+\sqrt{2}}}{2} \cdot \frac{\sqrt{2+\sqrt{2+\sqrt{2}}}}{2} \cdot \ldots$$

44. Let X be a r.v. From the inversion formula, it follows without much difficulty (see p. 95 in Durrett's book, 3rd or 4th eds.), that if φ_X is integrable, then X has a density f_X, and the density and characteristic function are related by

$$\begin{aligned} \varphi_X(t) &= \int_{-\infty}^{\infty} f_X(x) e^{itx}\, dx, \\ f_X(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi_X(t) e^{-itx}\, dt \end{aligned}$$

(this shows the duality between the Fourier transform and its inverse). Use this and the answer to a previous exercise to conclude that if X is a r.v. with the Cauchy distribution (i.e., X has density $f_X(x) = 1/\pi(1+x^2)$) then its characteristic function is given by

$$\varphi_X(t) = e^{-|t|}.$$

Deduce from this that if X, Y are independent Cauchy r.v.'s then any weighted average $\lambda X + (1-\lambda)Y$, where $0 \le \lambda \le 1$, is also a Cauchy r.v. (As a special case, it follows by induction that if $X_1, \ldots, X_n$ are i.i.d. Cauchy r.v.'s, then their average $(X_1 + \ldots + X_n)/n$ is also a Cauchy r.v., which was a claim we made without proof earlier in the course.)