Lecture Notes in Probability

Raz Kupferman
Institute of Mathematics
The Hebrew University

April 5, 2009

Contents

Foreword

These lecture notes were written while teaching the course “Probability 1” at the Hebrew University. Most of the material was compiled from a number of textbooks, such that *A first course in probability* by Sheldon Ross, *An introduction to probability theory and its applications* by William Feller, and *Weighing the odds* by David Williams. These notes are by no means meant to replace a textbook in probability. By construction, they are limited to the amount of material that can be taught in a 14 week course of 3 hours. I am grateful to the many students who have spotted mistakes and helped makes these notes more coherent.

Chapter 1

Basic Concepts

Discussion: Why is probability theory so often a subject of confusion?

In every mathematical theory there are three distinct aspects:

① A formal set of rules.

② An intuitive background, which assigns a *meaning* to certain concepts.

③ Applications: when and how can the formal framework be applied to solve a practical problem.

Confusion may arise when these three aspects intermingle.

Discussion: Intuitive background: what do we mean by "the probability of a die throw resulting in "5" is $1/6$?" Discuss the ***frequentist*** versus ***Bayesian*** points of view; explain why the frequentist point of view cannot be used as a fundamental definition of probability (but can certainly guide our intuition).

1.1 The Sample Space

The intuitive meaning of probability is always related to some ***experiment***, whether real or conceptual (e.g., winning the lottery, that a newborn be a boy, a person's height). We assign probabilities to ***possible outcomes*** of the experiment. We first need to develop an ***abstract model for an experiment***. In probability theory an experiment (real or conceptual) is modeled by ***all its possible outcomes***, i.e., by a ***set***, which we call the ***sample space***. Of course, a set is a mathematical entity, independent of any intuitive background.

Notation: We will usually denote the sample space by Ω and its elements by ω.

Examples:

① Tossing a coin: $\Omega = \{H, T\}$ (but what if the coin falls on its side or runs away?).

② Tossing a coin three times:

$$\Omega = \{(a_1, a_2, a_3) : a_i \in \{H, T\}\} = \{H, T\}^3.$$

(Is this the only possibility? for the same experiment we could only observe the majority.)

③ Throwing two *distinguishable* dice: $\Omega = \{1, \dots, 6\}^2$.

④ Throwing two *indistinguishable* dice: $\Omega = \{(i, j) : 1 \le i \le j \le 6\}$.

⑤ A person's lifetime (in years): $\Omega = \mathbb{R}^+$ (what about an age limitation?).

⑥ Throwing a dart into a unit circle: if we only measure the radius, $\Omega = [0, 1]$. If we measure position, we could have

$$\Omega = \{(r, \theta) : 0 \le r \le 1, 0 \le \theta < 2\pi\} = [0, 1] \times [0, 2\pi),$$

but also

$$\Omega = \left\{(x, y) : x^2 + y^2 \le 1\right\}.$$

Does it make a difference? What about missing the circle?

⑦ An infinite sequence of coin tosses: $\Omega = \{H, T\}^{\aleph_0}$ (which is isomorphic to the segment $(0, 1)$).

⑧ Brownian motion: $\Omega = C([0, 1]; \mathbb{R}^3)$.

⑨ A person throws a coin: if the result is Head he takes an exam in probability, which he either passes or fails; if the result is Tail he goes to sleep and we measure the duration of his sleep (in hours):

$$\Omega = \{H\} \times \{0, 1\} \cup \{T\} \times \mathbb{R}^+.$$

The sample space is the primitive notion of probability theory. It provides a model of an experiment in the sense that every thinkable outcome (even if extremely unlikely) is completely described by one, and only one, sample point.

1.2 Events

Suppose that we throw a die. The set of all possible outcomes (the sample space) is $\Omega = \{1, \dots, 6\}$. What about the result "the outcome is even"? Even outcome is *not* an element of Ω. It is a property shared by several points in the sample space. In other words, it corresponds to a ***subset*** of Ω ($\{2, 4, 6\}$). "The outcome is even" is therefore not an ***elementary*** outcome of the experiment. It is an aggregate of elementary outcomes, which we will call an ***event***.

Definition 1.1 An event is a property for which it is clear for every $\omega \in \Omega$ whether it has occurred or not. Mathematically, an event is a subset of the sample space.

The intuitive terms of "outcome" and "event" have been incorporated within an abstract framework of a set and its subsets. As such, we can perform on events set-theoretic operations of union, intersection and complementation. All set-theoretic relations apply as they are to events.

Let Ω be a sample space which corresponds to a certain experiment. What is the collection of all possible events? The immediate answer is $\mathscr{P}(\Omega)$, which is the set of all subsets (denoted also by 2^{Ω}). It turns out that in many cases it is advisable to restrict the collection of subsets to which probabilistic questions apply. In other words, the collection of events is only a subset of 2^{Ω}. While we leave the reasons to a more advanced course, there are certain requirements that the set of events has to fulfill:

① If $A \subseteq \Omega$ is an event so is A^c (if we are allowed to ask whether A has occurred, we are allowed to ask whether it has not occurred).

② If $A, B \subseteq \Omega$ are events, so is $A \cap B$.

③ Ω is an event (we can always ask "has any outcome occurred?").

A collection of events satisfying these requirements is called an ***algebra*** of events.

Definition 1.2 Two events A, B are called **disjoint** *if their intersection is empty. A collection of events is called* **mutually disjoint** *if every pair is disjoint. Let A, B be events, then*

$$A \cap B^c = \{\omega \in \Omega : (\omega \in A) \text{ and } (\omega \notin B)\} \equiv A \setminus B.$$

Unions of disjoint sets are denoted by $\uplus$.

Proposition 1.1 Let Ω be a sample space and $\mathscr{C}$ be an algebra of events. Then,

① $\emptyset \in \mathscr{C}$.

② *If $A_1, \dots, A_n \in \mathscr{C}$ then $\cup_{i=1}^n A_i \in \mathscr{C}$.*

③ *If $A_1, \dots, A_n \in \mathscr{C}$ then $\cap_{i=1}^n A_i \in \mathscr{C}$.*

Proof: Easy. Use de Morgan and induction. ■

Probability theory is often concerned with infinite sequences of events. A collection of events $\mathscr{C}$ is called a ***σ-algebra*** of events, if it is an algebra, and in addition, $(A_n)_{n=1}^\infty \subset \mathscr{C}$ implies that

$$\bigcap_{n=1}^\infty A_n \in \mathscr{C}.$$

That is, *a σ-algebra of events is closed with respect to countably many set-theoretic operations*. We will usually denote the σ-algebra by $\mathscr{F}$.

Discussion: Historic background on the countably-many issue.

Examples:

① A tautological remark: for every event A,

$$A = \{\omega \in \Omega : \omega \in A\}.$$

② For every $\omega \in \Omega$ the singleton $\{\omega\}$ is an event.

③ The experiment is tossing three coins and the event is “second toss was Head”.

④ The experiment is “waiting for the fish to bite” (in hours), and the event is “waited more than an hour and less than two”.

⑤ For every event A, $\{\emptyset, A, A^c, \Omega\}$ is a σ-algebra of events.

⑥ For every collection of events we can construct the σ-algebra generated by this collection.

⑦ The cylinder sets in Brownian motion.

✎ *Exercise 1.1* Construct a sample space corresponding to filling a single column in the Toto. Define two events that are disjoint. Define three events that are mutually disjoint, and whose union is Ω.

Infinite sequences of events Let $(\Omega, \mathscr{F})$ be a sample space together with a σ-algebra of events (such a pair is called a ***measurable space***), and let $(A_n)_{n=1}^{\infty} \subset \mathscr{F}$. We have

$$\bigcup_{n=1}^{\infty} A_n = \{\omega \in \Omega : \omega \in A_n \text{ for at least one } n\}$$

$$\bigcap_{n=1}^{\infty} A_n = \{\omega \in \Omega : \omega \in A_n \text{ for all } n\}.$$

Also,

$$\bigcup_{k=n}^{\infty} A_k = \{\omega \in \Omega : \omega \in A_k \text{ for at least one } k \geq n\}$$

(there exists a $k \geq n$ for which $\omega \in A_k$), so that

$$\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k = \{\omega \in \Omega : \omega \in A_k \text{ for infinitely many } k\}$$

(for every n there exists a $k \geq n$ for which $\omega \in A_k$). We denote,

$$\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k = \{\omega \in \Omega : \omega \in A_k \text{ i.o.}\} \equiv \limsup_n A_n.$$

Similarly,

$$\bigcap_{k=n}^{\infty} A_k = \{\omega \in \Omega : \omega \in A_k \text{ for all } k \geq n\}$$

($\omega \in A_k$ for all $k \geq n$), so that

$$\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k = \{\omega \in \Omega : \omega \in A_k \text{ eventually}\} \equiv \liminf_n A_n$$

(there exists an n such that $\omega \in A_k$ for all $k \geq n$). Clearly,

$$\liminf_n A_n \subseteq \limsup_n A_n.$$

Definition 1.3 *If* $(A_n)_{n=1}^\infty$ *is a sequence of events such that* $\liminf_n A_n = \limsup_n A_n$*, then we say that this sequence has a limit, and set*

$$\lim_n A_n = \liminf_n A_n = \limsup_n A_n.$$

In other words, all elements that occur ***infinitely often*** *also occur* ***eventually always****.*

Example: Let Ω be the set of integers, $\mathbb{N}$, and let

$$A_k = \{\text{all the evens/odds for } k \text{ even/odd}\}.$$

Then,

$$\limsup_k A_k = \Omega \qquad \text{and} \qquad \liminf_k A_k = \emptyset.$$

▲▲▲

Example: Let again $\Omega = \mathbb{N}$ and let

$$A_k = \left\{ k^j : \ j = 0, 1, 2, \dots \right\}.$$

Then,

$$\lim_k A_k = \{1\}.$$

▲▲▲

Definition 1.4 *A sequence* $(A_n)_{n=1}^\infty$ *is called* ***increasing*** *if* $A_1 \subseteq A_2 \subseteq \dots$*, and* ***decreasing*** *if* $A_1 \supseteq A_2 \supseteq \dots$*.*

Proposition 1.2 *If* $(A_n)_{n=1}^\infty$ *is an increasing sequence of events, then it has a limit given by*

$$\lim_n A_n = \bigcup_{n=1}^\infty A_n.$$

Proof: An increasing sequence has the property that

$$\bigcup_{k=n}^{\infty} A_k = \bigcup_{k=1}^{\infty} A_k, \qquad \text{and} \qquad \bigcap_{k=n}^{\infty} A_k = A_n,$$

and the rest is trivial. ■

✎ *Exercise 1.2* Prove that if $(A_n)_{n=1}^{\infty}$ is a decreasing sequence of events, then it has a limit given by

$$\lim_n A_n = \bigcap_{n=1}^{\infty} A_n.$$

1.3 Probability

Let $(\Omega, \mathscr{F})$ be a ***measurable space***. A probability is a function P which assigns a number to every event in $\mathscr{F}$ (the probability that this event has occurred). The function P has to satisfy the following properties:

① For every event $A \in \mathscr{F}$, $0 \le P(A) \le 1$.

② $P(\Omega) = 1$ (the probability that some result has occurred is one).

③ Let (A_n) be a sequence of mutually disjoint events, then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

This property is called ***countable additivity***.

The triple $(\Omega, \mathscr{F}, P)$ is called a ***probability space***.

The following results are immediate:

Proposition 1.3

① $P(\emptyset) = 0$.

② *For every* finite *sequence of N disjoint events* (A_n)

$$P\left(\bigcup_{n=1}^{N} A_n\right) = \sum_{n=1}^{N} P(A_n).$$

Proof: The first claim is proved by noting that for every $A \in \mathscr{F}$, $A = A \dot\cup \emptyset$. For the second claim we take $A_k = \emptyset$ for $k > n$. ■

Examples:

① Tossing a coin, and choosing $P(\{H\}) = P(\{T\}) = 1/2$.

② Throwing a die, and choosing $P(\{i\}) = 1/6$ for $i \in \{1, \dots, 6\}$. Explain why this defines uniquely a probability function.

Proposition 1.4 For every event $A \in \mathscr{F}$,

$$P(A^c) = 1 - P(A).$$

If A, B are events such that $A \subseteq B$, then

$$P(A) \leq P(B).$$

Proof: The first result follows from the fact that $A \dot\cup A^c = \Omega$. The second result follows from $B = A \dot\cup (B \setminus A)$. ■

Proposition 1.5 (Probability of a union) Let $(\Omega, \mathscr{F}, P)$ be a probability space. For every two events $A, B \in \mathscr{F}$,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Proof: We have

$$\begin{aligned} A \cup B &= (A \setminus B) \dot\cup (B \setminus A) \dot\cup (A \cap B) \\ A &= (A \setminus B) \dot\cup (A \cap B) \\ B &= (B \setminus A) \dot\cup (A \cap B), \end{aligned}$$

and it remains to use the additivity of the probability function to calculate $P(A \cup B) - P(A) - P(B)$. ■

Proposition 1.6 *Let $(\Omega, \mathscr{F}, P)$ be a probability space. For every three events $A, B, C \in \mathscr{F}$,*

$$\begin{aligned} P(A \cup B \cup C) = P(A) + P(B) + P(C) \\ - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ + P(A \cap B \cap C). \end{aligned}$$

Proof: Using the binary relation we have

$$\begin{aligned} P(A \cup B \cup C) &= P(A \cup B) + P(C) - P((A \cup B) \cap C) \\ &= P(A) + P(B) - P(A \cap B) + P(C) - P((A \cap C) \cup (B \cap C)), \end{aligned}$$

and it remains to apply the binary relation for the last expression. ■

Proposition 1.7 (Inclusion-exclusion principle) *For n events $(A_i)_{i=1}^n$ we have*

$$\begin{aligned} P(A_1 \cup \cdots \cup A_n) = \sum_{i=1}^n P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) \\ + (-1)^{n+1} P(A_1 \cap \cdots \cap A_n) \end{aligned}$$

✎ *Exercise 1.3* Prove the inclusion-exclusion principle.

✎ *Exercise 1.4* Let A, B be two events in a probability space. Show that the probability that either A or B has occurred, but not both, is $P(A)+P(B)-2\,P(A\cap B)$.

Lemma 1.1 (Boole's inequality) *Let $(\Omega, \mathscr{F}, P)$ be a probability space. Probability is* ***sub-additive*** *in the sense that*

$$P\left(\cup_{k=1}^\infty A_k\right) \le \sum_{k=1}^\infty P(A_k)$$

for every sequence (A_n) of events.

Proof: Define the following sequence of events,

$$B_1 = A_1 \quad B_2 = A_2 \setminus A_1, \quad , \dots, \quad B_n = A_n \setminus \left(\cup_{k=1}^{n-1} A_k \right).$$

The B_n are disjoint and their union equals to the union of the A_n. Also, $B_n \subseteq A_n$ for every n. Now,

$$P\left(\cup_{k=1}^{\infty} A_k\right) = P\left(\cup_{k=1}^{\infty} B_k\right) = \sum_{k=1}^{\infty} P(B_k) \leq \sum_{k=1}^{\infty} P(A_k).$$

■

1.4 Discrete Probability Spaces

The simplest sample spaces to work with are such whose sample spaces include countably many points. Let Ω be a countable set,

$$\Omega = \{a_1, a_2, \dots\},$$

and let $\mathscr{F} = \mathscr{P}(\Omega)$. Then, a probability function, P, on $\mathscr{F}$ is fully determined by its value for every singleton, $\{a_j\}$, i.e., by the probability assigned to every elementary event. Indeed, let $P(\{a_j\}) \equiv p_j$ be given, then since every event A can be expressed as a finite, or countable union of disjoint singletons,

$$A = \uplus_{\ a_j \in A} \{a_j\},$$

it follows from the additivity property that

$$P(A) = \sum_{a_j \in A} p_j.$$

A particular case which often arises in applications is when the sample space is finite (we denote by $|\Omega|$ the size of the sample space), and where every elementary event $\{\omega\}$ has equal probability, p. By the properties of the probability function,

$$1 = P(\Omega) = \sum_{a_j \in \Omega} P(\{a_j\}) = p|\Omega|,$$

i.e., $p = 1/|\Omega|$. The probability of every event A is then

$$P(A) = \sum_{a_j \in A} P(\{a_j\}) = p|A| = \frac{|A|}{|\Omega|}.$$

Comment: The probability space, i.e., the sample space, the set of events and the probability function, are a model of an experiment whose outcome is a priori unknown. There is no a priori reason why all outcomes should be equally probable. It is an assumption that has to be made only when believed to be applicable.

Examples:

① Two dice are rolled. What is the probability that the sum is 7? The sample space is $\Omega = \{(i, j) : 1 \leq i, j \leq 6\}$, and it is natural to assume that each of the $|\Omega| = 36$ outcomes is equally likely. The event ”the sum is 7” corresponds to

$$A = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\},$$

so that $P(A) = 6/36$.

② There are 11 balls in a jar, 6 white and 5 black. Two balls are taken at random. What is the probability of having one white and one black.

To solve this problem, it is helpful to imagine that the balls are numbered from 1 to 11. The sample space consists of all possible pairs,

$$\Omega = \{(i, j) : 1 \leq i < j \leq 11\},$$

and its size is $|\Omega| = \binom{11}{2}$; we assume that all pairs are equally likely. The event A = one black and one white corresponds to a number of states equal to the number of possibility to choose one white ball out of six, and one black ball out of five, i.e.,

$$P(A) = \frac{\binom{6}{1}\binom{5}{1}}{\binom{11}{2}} = \frac{6 \cdot 5}{10 \cdot 11 : 2} = \frac{6}{11}.$$

③ A deck of 52 cards is distributed between four players. What is the probability that one of the players received all 13 spades?

The sample space Ω is the set of all possible partitions, the number of different partitions being

$$|\Omega| = \frac{52!}{13!\, 13!\, 13!\, 13!}$$

(the number of possibilities to order 52 cards divided by the number of internal orders). Let A_i be the event that the i-th player has all spades, and A be the event that *some* player has all spades; clearly,

$$A = A_1 \mathbin{\dot\cup} A_2 \mathbin{\dot\cup} A_3 \mathbin{\dot\cup} A_4.$$

For each i,

$$|A_i| = \frac{39!}{13!\,13!\,13!},$$

hence,

$$P(A) = 4\,P(A_1) = 4 \times \frac{39!\,13!\,13!\,13!\,13!}{52!\,13!\,13!\,13!} \approx 6.3 \times 10^{-12}.$$

✎ *Exercise 1.5* Prove that it is not possible to construct a probability function on the sample space of integers, such that $P(\{i\}) = P(\{j\})$ for all i, j.

✎ *Exercise 1.6* A fair coin is tossed five times. What is the probability that there was at least one instance of two Heads in a row? Start by building the probability space.

We are now going to cover a number of examples, all concerning finite probability spaces with equally likely outcomes. The importance of these examples stems from the fact that they are representatives of classes of problems which recur in many applications.

Example: ***(The birthday paradox)*** In a random assembly of n people, what is the probability that none of them share the same date of birth?

We assume that $n < 365$ and ignore leap years. The sample space is the set of all possible date-of-birth assignments to n people. This is a sample space of size $|\Omega| = 365^n$. Let A_n be the event that all dates-of-birth are different. Then,

$$|A_n| = 365 \times 364 \times \cdots \times (365 - n + 1),$$

hence

$$P(A_n) = \frac{|A_n|}{|\Omega|} = 1 \times \frac{364}{365} \times \cdots \times \frac{365 - n + 1}{365}.$$

The results for various n are

$$P(A_{23}) < 0.5 \qquad P(A_{50}) < 0.03 \qquad P(A_{100}) < 3.3 \times 10^{-7}.$$

▲▲▲

Comment: Many would have guessed that $P(A_{50}) \approx 1 - 50/365$. This is a "selfish" thought.

Example: ***(The inattentive secretary, or the matching problem)*** A secretary places randomly n letters into n envelopes. What is the probability that no letter reaches its destination? What is the probability that exactly k letters reach their destination?

As usual, we start by setting the sample space. Assume that the letters and envelopes are all numbered from one to n. The sample space consists of all possible assignments of letters to envelopes, i.e., the set of all permutations of the numbers 1-to-n. The first question can be reformulated as follows: take a random one-to-one function from $\{1, \dots, n\}$ to itself; what is the probability that it has no fixed points?

If A is the event that no letter has reached its destination, then its complement, A^c is the event that at least one letter has reached its destination (at least one fixed point). Let furthermore B_i be the event that the i-th letter reached its destination, then

$$A^c = \cup_{i=1}^n B_i.$$

We apply the inclusion-exclusion principle:

$$\begin{aligned} P(A^c) &= \sum_{i=1}^n P(B_i) - \sum_{i<j} P(B_i \cap B_j) + \sum_{i<j<k} P(B_i \cap B_j \cap B_k) - \dots \\ &= n\,P(B_1) - \binom{n}{2} P(B_1 \cap B_2) + \binom{n}{3} P(B_1 \cap B_2 \cap B_3) - \dots, \end{aligned}$$

where we have used the symmetry of the problem. Now,

$$\begin{aligned} P(B_1) &= \frac{|B_1|}{|\Omega|} = \frac{(n-1)!}{n!} \\ P(B_1 \cap B_2) &= \frac{|B_1 \cap B_2|}{|\Omega|} = \frac{(n-2)!}{n!}, \end{aligned}$$

etc. It follows that

$$\begin{aligned} P(A^c) &= n\,\frac{1}{n} - \binom{n}{2}\frac{(n-2)!}{n!} + \binom{n}{3}\frac{(n-3)!}{n!} - \dots + (-1)^{n+1}\binom{n}{n}\frac{0!}{n!} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \dots + (-1)^{n+1}\frac{1}{n!} \\ &= \sum_{k=1}^n \frac{(-1)^{k+1}}{k!}. \end{aligned}$$

For large n,

$$P(A) = 1 - P(A^c) = \sum_{k=0}^{n} \frac{(-1)^k}{k!} \approx e^{-1},$$

from which we deduce that for large n, the probability that no letter has reached its destination is 0.37. The fact that the limit is finite may sound surprising (one could have argued equally well that the limit should be either 0 or 1). Note that as a side result, the number of permutations that have no fixed points is

$$|A| = |\Omega|\, P(A) = n! \sum_{k=0}^{n} \frac{(-1)^k}{k!}.$$

Now to the second part of this question. Before we answer what is the number of permutations that have exactly k fixed points, let's compute the number of permutations that have only k *specific* fixed points. This number coincides with the number of permutations of $n-k$ elements without fixed points,

$$(n-k)! \sum_{\ell=0}^{n-k} \frac{(-1)^\ell}{\ell!}.$$

The choice of k fixed points is exclusive, so to find the total number of permutations that have exactly k fixed points, we need to multiply this number by the number of ways to choose k elements out of n. Thus, if C denotes the event that there are exactly k fixed points, then

$$|C| = \binom{n}{k}(n-k)! \sum_{\ell=0}^{n-k} \frac{(-1)^\ell}{\ell!},$$

and

$$P(C) = \frac{1}{k!} \sum_{\ell=0}^{n-k} \frac{(-1)^\ell}{\ell!}.$$

For large n and fixed k we have

$$P(C) \approx \frac{e^{-1}}{k!}.$$

We will return to such expressions later on, in the context of the ***Poisson distribution***. ▲▲▲

✎ *Exercise 1.7* Seventeen men attend a party (there were also women, but it is irrelevant). At the end of it, these drunk men collect at random a hat from the hat hanger. What is the probability that

① At least someone got his own hat.
② John Doe got his own hat.
③ Exactly 3 men got their own hats.
④ Exactly 3 men got their own hats, one of which is John Doe.

As usual, start by specifying what is your sample space.

✎ *Exercise 1.8* A deck of cards is dealt out. What is the probability that the fourteenth card dealt is an ace? What is the probability that the first ace occurs on the fourteenth card?

1.5 Probability is a Continuous Function

This section could be omitted in a first course on probability. I decided to include it only in order to give some flavor of the analytical aspects of probability theory.

An important property of the probability function is its ***continuity***, in the sense that if a sequence of events (A_n) has a limit, then $P(\lim A_n) = \lim P(A_n)$.

We start with a "soft" version:

Theorem 1.1 (Continuity for increasing sequences) Let (A_n) be an increasing sequence of events, then

$$P(\lim_n A_n) = \lim_n P(A_n).$$

Comment: We have already shown that the limit of an increasing sequence of events exists,

$$\lim_n A_n = \cup_{k=1}^{\infty} A_k.$$

Note that for every n,

$$P(\cup_{k=1}^{n} A_k) = P(A_n),$$

but we can't just replace n by ∞. Moreover, since $P(A_n)$ is an increasing function we have

$$P(A_n) \nearrow P(\lim_n A_n).$$

Proof: Recall that for an increasing sequence of events, $\lim_n A_n = \cup_{n=1}^{\infty} A_n$. Construct now the following sequence of disjoint events,

$$\begin{aligned} B_1 &= A_1 \\ B_2 &= A_2 \setminus A_1 \\ B_3 &= A_3 \setminus A_2, \end{aligned}$$

etc. Clearly,

$$\dot{\cup}\ _{i=1}^{n} B_i = \cup_{i=1}^{n} A_i = A_n \quad \text{and hence} \quad \cup_{i=1}^{\infty} A_i = \dot{\cup}\ _{i=1}^{\infty} B_i.$$

Now,

$$\begin{aligned} P(\lim_n A_n) &= P(\cup_{i=1}^{\infty} A_i) = P(\dot{\cup}\ _{i=1}^{\infty} B_i) \\ &= \sum_{i=1}^{\infty} P(B_i) = \lim_n \sum_{i=1}^{n} P(B_i) \\ &= \lim_n P(\dot{\cup}\ _{i=1}^{n} B_i) = \lim_n P(A_n). \end{aligned}$$

■

✎ *Exercise 1.9 (Continuity for decreasing sequences)* Prove that if (A_n) is a decreasing sequence of events, then

$$P(\lim_n A_n) = \lim_n P(A_n),$$

and more precisely, $P(A_n) \searrow P(\lim_n A_n)$.

Now the next two lemmas are for arbitrary sequences of events, without assuming the existence of a limit.

Lemma 1.2 (Fatou) Let (A_n) be a sequence of events, then

$$P(\liminf_n A_n) \leq \liminf_n P(A_n).$$

Proof: Recall that

$$\liminf_n A_n = \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k \equiv \cup_{n=1}^{\infty} G_n$$

is the set of outcomes that occur "eventually always". The sequence (G_n) is increasing, and therefore

$$\lim_n P(G_n) = P(\lim_n G_n) = P(\liminf_n A_n).$$

On the other hand, since $G_n = \cap_{k=n}^{\infty} A_k$, it follows that

$$P(G_n) \leq \inf_{k \geq n} P(A_k).$$

The left hand side converges to $P(\liminf_n A_n)$ whereas the right hand side converges to $\liminf_n P(A_n)$, which concludes the proof. ■

Lemma 1.3 (Reverse Fatou) Let (A_n) be a sequence of events, then

$$\limsup_n P(A_n) \leq P(\limsup_n A_n).$$

Proof: Recall that

$$\limsup_n A_n = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k \equiv \cap_{n=1}^{\infty} G_n$$

is the set of outcomes that occur "infinitely often". The sequence (G_n) is decreasing, and therefore

$$\lim_n P(G_n) = P(\lim_n G_n) = P(\limsup_n A_n).$$

On the other hand, since $G_n = \cup_{k=n}^{\infty} A_k$, it follows that

$$P(G_n) \geq \sup_{k \geq n} P(A_k).$$

The left hand side converges to $P(\limsup_n A_n)$ whereas the right hand side converges to $\limsup_n P(A_n)$, which concludes the proof. ■

Theorem 1.2 (Continuity of probability) If a sequence of events (A_n) *has a limit, then*

$$P(\lim_n A_n) = \lim_n P(A_n).$$

Proof: This is an immediate consequence of the two Fatou lemmas, for

$$\limsup_n P(A_n) \le P(\limsup_n A_n) = P(\liminf_n A_n) \le \liminf_n P(A_n).$$

■

Lemma 1.4 (First Borel-Cantelli) Let (A_n) *be a sequence of events such that* $\sum_n P(A_n) < \infty$*. Then,*

$$P(\limsup_n A_n) = P(\{A_n \; i.o.\}) = 0.$$

Proof: Let as before $G_n = \cup_{k=n}^{\infty} A_k$. Since $P(G_n) \searrow P(\limsup_n A_n)$, then for all m

$$P(\limsup_n A_n) \le P(G_m) \le \sum_{k \ge m} P(A_k).$$

Letting $m \to \infty$ and using the fact that the right hand side is the tail of a converging series we get the desired result. ■

✎ *Exercise 1.10* Does the “reverse Borel-Cantelli” hold? Namely, is it true that if $\sum_n P(A_n) = \infty$ then

$$P(\limsup_n A_n) > 0.$$

Well, no. Construct a counter example.

Example: Consider the following scenario. At a minute to noon we insert into an urn balls numbered 1-to-10 and remove the ball numbered “10”. At half a minute to noon we insert balls numbered 11-to-20 and remove the ball numbered “20”, and so on. Which balls are inside the urn at noon? Clearly all integers except for the “10n”.

Now we vary the situation, except that the first time we remove the ball numbered “1”, next time the ball numbered “2”, etc. Which balls are inside the urn at noon? none.

In the third variation we remove each time a ball at random (from those inside the urn). Are there any balls left at noon? If this question is too bizarre, here is a more sensible picture. Our sample space consists of random sequences of numbers, whose elements are distinct, and whose first element is in the range 1-to-10, its second element is in the range 1-to-20, and so on. We are asking what is the probability that such a sequence contains all integers?

Let’s focus on ball number “1” and denote by E_n then event that it is still inside the urn after n steps. We have

$$P(E_n) = \frac{9}{10} \times \frac{18}{19} \times \cdots \times \frac{9n}{9n+1} = \prod_{k=1}^{n} \frac{9k}{9k+1}.$$

The events (E_n) form a decreasing sequence, whose countable intersection corresponds to the event that the first ball was not ever removed. Now,

$$\begin{aligned} P(\lim_n E_n) &= \lim_n P(E_n) = \lim_n \prod_{k=1}^{n} \frac{9k}{9k+1} \\ &= \lim_n \left[\prod_{k=1}^{n} \frac{9k+1}{9k} \right]^{-1} = \lim_n \left[\prod_{k=1}^{n} \left(1 + \frac{1}{9k}\right) \right]^{-1} \\ &= \lim_n \left[\left(1 + \frac{1}{9}\right)\left(1 + \frac{1}{18}\right) \dots \right]^{-1} \\ &\le \lim_n \left(1 + \frac{1}{9} + \frac{1}{18} + \dots \right)^{-1} = 0. \end{aligned}$$

Thus, there is zero probability that the ball numbered “1” is inside the urn after infinitely many steps. The same holds ball number “2”, “3”, etc. If F_n denotes the event that the n-th ball has remained inside the box at noon, then

$$P(\cup_{n=1}^{\infty} F_n) \le \sum_{n=1}^{\infty} P(F_n) = 0.$$

▲▲▲

Chapter 2

Conditional Probability and Independence

2.1 Conditional Probability

Example: Two dice are tossed. What is the probability that the sum is 8? This is an easy exercise: we have a sample space Ω that comprises 36 equally-probable outcomes. The event "the sum is 8" is given by

$$A = \{(2,6),(3,5),(4,4),(5,3),(6,2)\},$$

and therefore $P(A) = |A|/|\Omega| = 5/36$.

But now, suppose someone reveals that the first die resulted in "3". How does this change our predictions? This piece of information tells us that "with certainty" the outcome of the experiment lies in set

$$F = \{(3,i) : 1 \le i \le 6\} \subset \Omega.$$

As this point, outcomes that are not in F have to be ruled out. The sample space can be restricted to F (F becomes the certain event). The event A (sum was "8") has to be restricted to its intersection with F. It seems reasonable that "the probability of A knowing that F has occurred" be defined as

$$\frac{|A \cap F|}{|F|} = \frac{|A \cap F|/|\Omega|}{|F|/|\Omega|} = \frac{P(A \cap F)}{P(F)},$$

which in the present case is $1/6$. ▲▲▲

This example motivates the following definition:

Definition 2.1 Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $F \in \mathscr{F}$ be an event for which $P(F) \neq 0$. For every $A \in \mathscr{F}$ the ***conditional probability of** A **given that** F **has occurred*** *(or simply, given F) is defined (and denoted) by*

$$P(A|F) := \frac{P(A \cap F)}{P(F)}.$$

Comment: Note that conditional probability is defined only if the conditioning event has finite probability.

Discussion: Like probability itself, conditional probability also has different interpretations depending on wether you are a frequentist or Bayesian. In the frequentist interpretation, we have in mind a large set of n repeated experiments. Let n_B denote the number of times event B occurred, and $n_{A,B}$ denote the number of times that both events A and B occurred. Then in the frequentist's world,

$$P(A|B) = \lim_{n\to\infty} \frac{n_{A,B}}{n_B}.$$

In the Bayesian interpretation, this conditional probability is that belief that A has occurred after we learned that B has occurred.

Example: There are 10 white balls, 5 yellow balls and 10 black balls in an urn. A ball is drawn at random, what is the probability that it is yellow (answer: $5/25$)? What is the probability that it is yellow given that it is not black (answer: $5/15$)? Note how the additional information *restricts the sample space to a subset*. ▲▲▲

Example: Jeremy can't decide whether to study probability theory or literature. If he takes literature, he will pass with probability $1/2$; if he takes probability, he will pass with probability $1/3$. He made his decision based on a coin toss. What is the probability that he passed the probability exam?

This is an example where the main task is to set up the probabilistic model and interpret the data. First, the sample space. We can set it to be the product of the two sets

$$\{\text{prob.}, \text{lit.}\} \times \{\text{pass}, \text{fail}\}.$$

If we define the following events:

$$A = \{\text{passed}\} = \{\text{prob., lit.}\} \times \{\text{pass}\}$$
$$B = \{\text{probability}\} = \{\text{prob.}\} \times \{\text{pass, fail}\},$$

then we interpret the data as follows:

$$P(B) = P(B^c) = \frac{1}{2} \qquad P(A|B) = \frac{1}{3} \qquad P(A|B^c) = \frac{1}{2}.$$

The quantity to be calculated is $P(A \cap B)$, and this is obtained as follows:

$$P(A \cap B) = P(A|B)P(B) = \frac{1}{6}.$$

▲▲▲

Example: There are 8 red balls and 4 white balls in an urn. Two are drawn at random. What is the probability that the second was red given that the first was red?

Answer: it is the probability that both were red divided by the probability that the first was red. The result is however 7/11, which illuminates the fact that having drawn the first ball red, we can think of a new initiated experiment. ▲▲▲

The next theorem justifies the term conditional probability:

Theorem 2.1 Let $(\Omega, \mathscr{F}, P)$ *be a probability space and* F *be an event such that* $P(F) \neq 0$. *Define the set function* $Q(A) = P(A|F)$. *Then,* Q *is a probability function over* $(\Omega, \mathscr{F})$.

Proof: We need to show that the three axioms are met. Clearly,

$$Q(A) = \frac{P(A \cap F)}{P(F)} \leq \frac{P(F)}{P(F)} = 1.$$

Also,

$$Q(\Omega) = \frac{P(\Omega \cap F)}{P(F)} = \frac{P(F)}{P(F)} = 1.$$

Finally, let (A_n) be a sequence of mutually disjoint events. Then the events $(A_n \cap F)$ are also mutually disjoint, and

$$\begin{aligned} Q(\cup_n A_n) &= \frac{P((\cup_n A_n) \cap F)}{P(F)} = \frac{P(\cup_n (A_n \cap F))}{P(F)} \\ &= \frac{1}{P(F)} \sum_n P(A_n \cap F) = \sum_n Q(A_n). \end{aligned}$$

■

In fact, the function Q is a probability function on the smaller space F, with the σ-algebra

$$\mathscr{F}|_F := \{F \cap A : \ A \in \mathscr{F}\}.$$

✎ *Exercise 2.1* Let A, B, C be three events of positive probability. We say that "A favors B" if $P(B|A) > P(B)$. Is it generally true that if A favors B and B favors C, then A favors C?

✎ *Exercise 2.2* Prove the ***general multiplication rule***

$$P(A \cap B \cap C) = P(A)\, P(B|A)\, P(C|A \cap B),$$

with the obvious generalization for more events. Reconsider the "birthday paradox" in the light of this formula.

2.2 Bayes' Rule and the Law of Total Probability

Let $(A_i)_{i=1}^n$ be a partition of Ω. By that we mean that the A_i are mutually disjoint and that their union equals Ω (every $\omega \in \Omega$ is in one and only one A_i). Let B be an event. Then, we can write

$$B = \uplus_{i=1}^{n} (B \cap A_i),$$

and by additivity,

$$\boxed{P(B) = \sum_{i=1}^n P(B \cap A_i) = \sum_{i=1}^n P(B|A_i) P(A_i).}$$

This law is known as the law of ***total probability***; it is very intuitive.

The next rule is known as ***Bayes' law***: let A, B be two events (such that $P(A), P(B) > 0$) then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{and} \qquad P(B|A) = \frac{P(B \cap A)}{P(A)},$$

from which we readily deduce

$$\boxed{P(A|B) = P(B|A)\frac{P(A)}{P(B)}.}$$

Bayes' rule is easily generalized as follows: if $(A_i)_{i=1}^n$ is a partition of Ω, then

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^n P(B|A_j)P(A_j)},$$

where we have used the law of total probability.

Example: A lab screen for the HIV virus. A person that carries the virus is screened positive in only 95% of the cases. A person who does not carry the virus is screened positive in 1% of the cases. Given that 0.5% of the population carries the virus, what is the probability that a person who has been screened positive is actually a carrier?

Again, we start by setting the sample space,

$$\Omega = \{\text{carrier, not carrier}\} \times \{+, -\}.$$

Note that the sample space is not a sample of people! If we define the events,

$$A = \{\text{the person is a carrier}\} \qquad B = \{\text{the person was screened positive}\},$$

it is given that

$$P(A) = 0.005 \qquad P(B|A) = 0.95 \qquad P(B|A^c) = 0.01.$$

Now,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$
$$= \frac{0.95 \cdot 0.005}{0.95 \cdot 0.005 + 0.01 \cdot 0.995} \approx \frac{1}{3}.$$

This is a nice example where fractions fool our intuition. ▲▲▲

✎ *Exercise 2.3* The following problem is known as ***Pólya's urn***. At time $t = 0$ an urn contains two balls, one black and one red. At time $t = 1$ you draw one ball at random and replace it together with a new ball of the same color. You repeat this procedure at every integer time (so that at time $t = n$ there are $n + 2$ balls. Calculate

$$p_{n,r} = P(\text{there are } r \text{ red balls at time } n)$$

for $n = 1, 2, \dots$ and $r = 1, 2, \dots, n + 1$. What can you say about the proportion of red balls as $n \to \infty$.

Solution: In order to have r red balls at time n there must be either r or $r - 1$ red balls at time $n - 1$. By the law of total probability, we have the recursive formula

$$p_{n,r} = \left(1 - \frac{r}{n+1}\right) p_{n-1,r} + \frac{r-1}{n+1} p_{n-1,r-1},$$

with "initial conditions" $p_{0,1} = 1$. If we define $q_{n.r} = (n + 1)! p_{n,r}$, then

$$q_{n,r} = (n + 1 - r)\, q_{n-1,r} + (r - 1)\, q_{n-1,r-1}.$$

You can easily check that $q_{n,r} = n!$ so that the solution to our problem is $p_{n,r} = 1/(n + 1)$. At any time all the outcomes for the number of red balls are equally likely!

2.3 Compound experiments

So far we have only "worked" with a restricted class of probability spaces—finite probability space in which all outcomes have the same probability. The concept of conditional probabilities is also a mean to define a class of probability spaces, representing compound experiments where certain parts of the experiment rely on the outcome of other parts. The simplest way to get insight into it is through examples.

Example: Consider the following statement: "the probability that a family has k children is p_k (with $\sum_k p_k = 1$), and for any family size, all sex distributions have equal probabilities". What is the probability space corresponding to such a statement?

Since there is no a-priori limit on the number of children (although every family has a finite number of children), we should take our sample space to be the set of all finite sequences of type ”bggbg”:

$$\Omega = \left\{ a_1 a_2 \dots a_n : \; a_j \in \{b, g\}, n \in \mathbb{N} \right\}.$$

This is a countable space so the σ-algebra can include all subsets. What is then the probability of a point $\omega \in \Omega$? Suppose that ω is a string of length n, and let A_n be the event “the family has n children”, then by the law of total probability

$$P(\{\omega\}) = \sum_{m=1}^{\infty} P(\{\omega\}|A_m)P(A_m) = \frac{p_n}{2^n}.$$

Having specified the probability of all singletons of a countable space, the probability space is fully specified.

We can then ask, for example, what is the probability that a family with no girls has exactly one child? If B denotes the event “no girls”, then

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{p_1/2}{p_1/2 + p_2/4 + p_3/8 + \dots}.$$

▲▲▲

Example: Consider two dice: die A has four red and two white faces and die B has two red and four white faces. One throws a coin: if it falls Head then die A is tossed sequentially, otherwise die B is used.

What is the probability space?

$$\Omega = \{H, T\} \times \left\{ a_1 a_2 \cdots : \; a_j \in \{R, W\} \right\}.$$

It is a ***product*** of two subspaces. What we are really given is a probability on the first space and a *conditional* probability on the second space. If A_H and A_T represent the events “head has occurred” and “tail has occurred”, then we know that

$$P(A_H) = P(A_T) = \frac{1}{2},$$

and facts like

$$\begin{aligned} P(\{RRWR\}\,|A_H) &= \frac{4}{6} \cdot \frac{4}{6} \cdot \frac{2}{6} \cdot \frac{4}{6} \\ P(\{RRWR\}\,|A_T) &= \frac{2}{6} \cdot \frac{2}{6} \cdot \frac{4}{6} \cdot \frac{2}{6}. \end{aligned}$$

(No matter for the moment where these numbers come from....) ▲▲▲

The following well-known “paradox” demonstrates the confusion that can arise where the boundaries between formalism and applications are fuzzy.

Example: ***The sibling paradox.*** Suppose that in all families with two children all the four combinations $\{bb, bg, gb, gg\}$ are equally probable. Given that a family with two children has at least one boy, what is the probability that it also has a girl? The easy answer is $2/3$.

Suppose now that one knocks on the door of a family that has two children, and a boy opens the door and says "I am the oldest child". What is the probability that he has a sister? The answer is one half. Repeat the same scenario but this time the boy says "I am the youngest child". The answer remains the same. Finally, a boy opens the door and says nothing. What is the probability that he has a sister: a half or two thirds???

The resolution of this paradox is that the experiment is not well defined. We could think of two different scenarios: (i) all families decide that boys, when available, should open the door. In this case if a boy opens the door he just rules out the possibility of gg, and the likelihood of a girl is $2/3$. (ii) When the family hears knocks on the door, the two siblings toss a coin to decide who opens. In this case, the sample space is

$$\Omega = \{bb, bg, gb, gg\} \times \{1, 2\},$$

and all 8 outcomes are equally likely. When a boy opens, he gives us the knowledge that the outcome is in the set

$$A = \{(bb, 1), (bb, 2), (bg, 1), (gb, 2)\}.$$

If B is the event that there is a girl, then

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(\{(bg, 1), (gb, 2)\})}{P(A)} = \frac{2/8}{4/8} = \frac{1}{2}.$$

▲▲▲

✎ *Exercise 2.4* Consider the following generic compound experiment: one performs a first experiment to which corresponds a probability space $(\Omega_0, \mathscr{F}_0, P_0)$, where Ω_0 is a finite set of size n. Depending on the outcome of the first experiment, the person conducts a second experiment. If the outcome was $\omega_j \in \Omega_0$ (with $1 \le j \le n$), he conducts an experiment to which corresponds a probability space $(\Omega_j, \mathscr{F}_j, P_j)$. Construct a probability space that corresponds to the compound experiment.

NEEDED: exercises on compound experiments.

2.4 Independence

Definition 2.2 Let A and B be two events in a probability space $(\Omega, \mathscr{F}, P)$. *We say that A* ***is independent of*** *B if the knowledge that B has occurred does not alter the probability that A has occurred. That is,*

$$P(A|B) = P(A).$$

By the definition of conditional probability, this condition is equivalent to

$$P(A \cap B) = P(A)P(B),$$

which may be taken as an alternative definition of independence. Also, the latter condition makes sense also if $P(A)$ or $P(B)$ are zero. By the symmetry of this condition we immediately conclude:

Corollary 2.1 If A is independent of B then B is also independent of A. Independence is a mutual property.

Example: A card is randomly drawn from a deck of 52 cards. Let

$$A = \{\text{the card is an Ace}\}$$
$$B = \{\text{the card is a spade}\}.$$

Are these events independent? Answer: yes. ▲▲▲

Example: Two dice are tossed. Let

$$A = \{\text{the first die is a 4}\}$$
$$B = \{\text{the sum is 6}\}.$$

Are these events independent (answer: no)? What if the sum was 7 (answer: yes)? ▲▲▲

Proposition 2.1 Every event is independent of Ω and $\emptyset$.

Proof: Immediate. ■

Proposition 2.2 If B is independent of A then it is independent of A^c.

Proof: Since $B = (B \cap A^c) \uplus (A \cap B)$,

$$P(B \cap A^c) = P(B) - P(A \cap B) = P(B) - P(A)P(B) = P(B)(1 - P(A)) = P(B)P(A^c).$$

■

Thus, if B is independent of A, it is independent of the collection of sets $\{\Omega, A, A^c, \emptyset\}$, which is ***the σ-algebra generated by*** A. This innocent distinction will gain meaning in a moment.

Consider then three events A, B, C. What does it mean that they are independent? What does it mean for A to be independent of B and C? A first, natural guess would be to say that the knowledge that B has occurred does not affect the probability of A, as does the knowledge that C has occurred? Does it imply that the probability of A is indeed independent of *any* information regarding whether B and C occurred?

Example: Consider again the toss of two dice and

$$\begin{aligned} A &= \{\text{the sum is 7}\} \\ B &= \{\text{the first die is a 4}\} \\ C &= \{\text{the first die is a 2}\}. \end{aligned}$$

Clearly, A is independent of B and it is independent of C, but it is not true that B and C are independent.

But now, what if instead C was that the *second* die was a 2? It is still true that A is independent of B and independent of C, but can we claim that it is independent of B ***and*** C? Suppose we knew that ***both*** B and C took place. This would certainly change the probability that A has occurred (it would be zero). This example calls for a modified definition of independence between multiple events. ▲▲▲

Definition 2.3 *The event A is said to be independent of the* **pair** *of events B and C if it is independent of every event in the σ-algebra generated by B and C. That is, if it is independent of the collection*

$$\sigma(B,C) = \{B, C, B^c, C^c, B\cap C, B\cup C, B\cap C^c, B\cup C^c, \ldots, \Omega, \emptyset\}.$$

Proposition 2.3 *A is independent of B and C if and only iff it is independent of B, C, and B ∩ C, that is, if and only if*

$$P(A\cap B) = P(A)P(B) \qquad P(A\cap C) = P(A)P(C)$$
$$P(A\cap B\cap C) = P(A)P(B\cap C).$$

Proof: The "only if" part is obvious. Now to the "if" part. We need to show that A is independent of each element in the σ-algebra generated by B and C. What we already know is that A is independent of B, C, $B\cap C$, B^c, C^c, $B^c\cup C^c$, Ω, and $\emptyset$ (not too bad!). Take for example the event $B\cup C$:

$$\begin{aligned}
P(A\cap(B\cup C)) &= P(A\cap(B^c\cap C^c)^c)\\
&= P(A) - P(A\cap B^c\cap C^c)\\
&= P(A) - P(A\cap B^c) + P(A\cap B^c\cap C)\\
&= P(A) - P(A)P(B^c) + P(A\cap C) - P(A\cap C\cap B)\\
&= P(A) - P(A)(1-P(B)) + P(A)P(C) - P(A)P(B\cap C)\\
&= P(A)\,[P(B) + P(C) - P(B\cap C)]\\
&= P(A)P(B\cup C).
\end{aligned}$$

The same method applies to all remaining elements of the σ-algebra. ■

✎ *Exercise 2.5* Prove directly that if A is independent of B, C, and $B\cap C$, then it is independent of $B\setminus C$.

Corollary 2.2 The events A, B, C are mutually independent in the sense that each one is independent of the remaining pair if and only if

$$P(A \cap B) = P(A)P(B) \qquad P(A \cap C) = P(A)P(C)$$
$$P(B \cap C) = P(B)P(C) \qquad P(A \cap B \cap C) = P(A)P(B \cap C).$$

More generally,

Definition 2.4 A collection of events (A_n) is said to consist of ***mutually independent events*** *if for every subset $A_{n_1}, \dots, A_{n_k}$,*

$$P(A_{n_1} \cap \cdots \cap A_{n_k}) = \prod_{j=1}^{k} P(A_{n_j}).$$

2.5 Repeated Trials

Only now that we have defined the notion of independence we can consider the situation of an experiment being repeated again and again ***under identical conditions***—a situation underlying the very notion of probability.

Consider an experiment, i.e., a probability space $(\Omega_0, \mathscr{F}_0, P_0)$. We want to use this probability space to construct a compound probability space corresponding to the idea of repeating the same experiment sequentially n times, the outcome of each trial being independent of all other trials. For simplicity, we assume that the single experiment corresponds to a discrete probability space, $\Omega_0 = \{a_1, a_2, \dots\}$ with atomistic probabilities $P_0(\{a_j\}) = p_j$.

Consider now the compound experiment of repeating the same experiment n times. The sample space consists of n-tuples,

$$\Omega = \Omega_0^n = \left\{(a_{j_1}, a_{j_2}, \dots, a_{j_n}) : \; a_{j_k} \in \Omega_0\right\}.$$

Since this is a discrete space, the probability is fully determined by its value for all singletons. Each singleton,

$$\omega = (a_{j_1}, a_{j_2}, \dots, a_{j_n}),$$

corresponds to an event, which is the intersection of the n events: “first outcome was a_{j_1}” and “second outcome was a_{j_2}”, etc. Since we assume statistical independence between trials, its probability should be the product of the individual probabilities. I.e., it seems reasonable to take

$$P(\{(a_{j_1}, a_{j_2}, \dots, a_{j_n})\}) = p_{j_1} p_{j_2} \dots p_{j_n}. \tag{2.1}$$

Note that this is ***not*** the only possible probability that one can define on Ω_0^n!

Proposition 2.4 Definition (2.1) *defines a probability function on* Ω_0^n.

Proof: Immediate. ■

The following proposition shows that (2.1) does indeed correspond to a situation where different trials do not influence each other’s statistics:

Proposition 2.5 Let $A_1, A_2, \dots, A_n$ *be a sequence of events such that the j-th trial alone determines whether* A_j *has occurred; that is, there exists a* $B_j \subseteq \Omega_0$, *such that*

$$A_j = \Omega_0^{j-1} \times B_j \times \Omega_0^{n-j}.$$

If the probability is defined by (2.1), *then the* A_j *are mutually independent.*

Proof: Consider a pair of such events A_j, A_k, say, $j < k$. Then,

$$A_j \cap A_k = \Omega_0^{j-1} \times B_j \times \Omega_0^{k-j-1} \times B_k \times \Omega_0^{n-k},$$

which can be written as

$$A_j \cap A_k = \uplus_{\, b_1 \in \Omega_0} \dots \uplus_{b_j \in B_j} \dots \uplus_{b_k \in B_k} \dots \uplus_{\, b_n \in \Omega_0} \{(b_1, b_2, \dots, b_n)\}.$$

Using the additivity of the probability,

$$\begin{aligned} P(A_j \cap A_k) &= \sum_{b_1 \in \Omega_0} \cdots \sum_{b_j \in B_j} \cdots \sum_{b_k \in B_k} \cdots \sum_{b_n \in \Omega_0} P_0(\{b_1\}) P_0(\{b_2\}) \dots P_0(\{b_n\}) \\ &= P_0(B_j) P_0(B_k). \end{aligned}$$

It is easy, by a similar construction to show that in fact

$$P(A_j) = P_0(B_j)$$

for all j, so that the binary relation has been proved. Similarly, we can take all triples, quadruples, etc. ■

Example: Consider an experiment with two possible outcomes: ”Success” with probability p and ”Failure” with probability $q = 1-p$ (such an experiment is called a ***Bernoulli trial***). Consider now an infinite sequence of independent repetitions of this basic experiment. While we have not formally defined such a probability space (it is uncountable), we do have a precise probabilistic model for any finite subset of trials.

(1) What is the probability of at least one success in the first n trials? (2) What is the probability of exactly k successes in the first n trials? (3) What is the probability of an infinite sequence of successes?

Let A_j denote the event “the j-th trial was a success”. What we know is that for all distinct natural numbers $j_1, \dots, j_n$,

$$P(A_{j_1} \cap \cdots \cap A_{j_n}) = p^n.$$

To answer the first question, we note that the probability of having only failures in the first n trials is q^n, hence the answer is $1 - q^n$. To answer the second question, we note that exactly k successes out of n trials is a disjoint unions of n-choose-k singletons, the probability of each being $p^k q^{n-k}$. Finally, to answer the third question, we use the continuity of the probability function,

$$P(\cap_{j=1}^{\infty} A_j) = P(\cap_{n=1}^{\infty} \cap_{j=1}^{n} A_j) = P(\lim_{n\to\infty} \cap_{j=1}^{n} A_j) = \lim_{n\to\infty} P(\cap_{j=1}^{n} A_j) = \lim_{n\to\infty} p^n,$$

which equals 1 if $p = 1$ and zero otherwise. ▲▲▲

Example: ***(The gambler’s ruin problem, Bernoulli 1713)*** Consider the following game involving two players, which we call Player A and Player B. Player A starts the game owning i NIS while Player B owns $N - i$ NIS. The game is a game of zero-sum, where each turn a coin is tossed. The coin has probability p to fall on Head, in which case Player B pays Player A one NIS; it has probability $q = 1 - p$ to fall on Tail, in which case Player A pays Player B one NIS. The game ends when one of the players is broke. What is the probability for Player A to win?

While the game may end after a finite time, the simplest sample space is that of an infinite sequence of tosses, $\Omega = \{H, T\}^{\mathbb{N}}$. The event

$$E = \{\text{“Player A wins” }\},$$

consists of all sequences in which the number of Heads exceeds the number of Tails by $N - i$ before the number of Tails has exceeded the number of Heads by i. If

$$F = \{\text{“first toss was Head”}\},$$

then by the law of total probability,

$$P(E) = P(E|F)P(F) + P(E|F^c)P(F^c) = pP(E|F) + qP(E|F^c).$$

If the first toss was a Head, then by our assumption of mutual independence, we can think of the game starting anew with Player A having $i + 1$ NIS (and $i - 1$ if the first toss was Tail). Thus, if α_i denote the probability that Player A wins if he starts with i NIS, then

$$\alpha_i = p\,\alpha_{i+1} + q\,\alpha_{i-1},$$

or equivalently,

$$\alpha_{i+1} - \alpha_i = \frac{q}{p}(\alpha_i - \alpha_{i-1}).$$

The “boundary conditions” are $\alpha_0 = 0$ and $\alpha_N = 1$.

This system of equations is easily solved. We have

$$\begin{aligned}
\alpha_2 - \alpha_1 &= \frac{q}{p}\alpha_1 \\
\alpha_3 - \alpha_2 &= \frac{q}{p}(\alpha_2 - \alpha_1) = \left(\frac{q}{p}\right)^2 \alpha_1 \\
\vdots &= \vdots \\
1 - \alpha_{N-1} &= \frac{q}{p}(\alpha_{N-1} - \alpha_{N-2}) = \left(\frac{q}{p}\right)^{N-1} \alpha_1.
\end{aligned}$$

Summing up,

$$1 - \alpha_1 = \left[1 + \frac{q}{p} + \cdots + \left(\frac{q}{p}\right)^{N-1}\right]\alpha_1 - \alpha_1,$$

i.e.,

$$1 = \frac{(q/p)^N - 1}{q/p - 1}\alpha_1,$$

from which we get that

$$\alpha_i = \frac{(q/p)^i - 1}{q/p - 1}\alpha_1 = \frac{(q/p)^i - 1}{(q/p)^N - 1}.$$

What is the probability that Player B wins? Exchange i with $N - i$ and p with q. What is the probability that either of them wins? The answer turns out to be 1! ▲▲▲

2.6 On Zero-One Laws

Events that have probability either zero or one are often very interesting. We will demonstrate such a situation with a funny example, which is representative of a class of problems that have been classified by Kolmogorov as 0-1 laws. The general theory is beyond the scope of this course.

Consider a monkey typing on a typing machine, each second typing a character (a letter, number, or a space). Each character is typed at random, independent of past characters. The sample space consists thus of infinite strings of typing-machine characters. The question that interests us is how many copies of the Collected Work of Shakespeare (WS) did the monkey produce. We define the following events:

$$\begin{aligned} H &= \{\text{the monkey produces infinitely many copies of WS}\} \\ H_k &= \{\text{the monkey produces at least } k \text{ copies of WS}\} \\ H_{m,k} &= \{\text{the monkey produces at least } k \text{ copies of WS by time } m\} \\ H^m &= \{\text{the monkey produces infinitely many copies of WS after time } m+1\}. \end{aligned}$$

Of course, $H^m = H$, i.e., the event of producing infinitely many copies is not affected by any finite prefix (it is a ***tail event!***).

Now, because the first m characters are independent of the characters from $m + 1$ on, we have for all m, k,

$$P(H_{m,k} \cap H^m) = P(H_{m,k})P(H^m).$$

and since $H^m = H$,

$$P(H_{m,k} \cap H) = P(H_{m,k})P(H).$$

Take now $m \to \infty$. Clearly, $\lim_{m\to\infty} H_{m,k} = H_k$, and

$$\lim_{m\to\infty} (H_{m,k} \cap H) = H_k \cap H = H.$$

By the continuity of the probability function,

$$P(H) = P(H_k)P(H).$$

Finally, taking $k \to \infty$, we have $\lim_{k\to\infty} H_k = H$, and by the continuity of the probability function,

$$P(H) = P(H)P(H),$$

from which we conclude that $P(H)$ is either zero or one.

2.7 Further examples

In this section we examine more applications of conditional probabilities.

Example: The following example is actually counter-intuitive. Consider an infinite sequence of tosses of a fair coin. There are eight possible outcomes for three consecutive tosses, which are HHH, HHT, HTH, HTT, THH, THT, TTH, and TTT. It turns out that for any of those triples, there exists another triple, which is likely to occur first with probability strictly greater than one half.

Take for example s_1 = HHH and s_2 = THH, then

$$P(s_2 \text{ before } s_1) = 1 - P(\text{first three tosses are } s_1) = \frac{7}{8}.$$

Take s_3 = TTH, then

$$P(s_3 \text{ before } s_2) = P(\text{TT before } s_2) > P(\text{TT before HH}) = \frac{1}{2}.$$

where the last equality follows by symmetry. ▲▲▲

✎ *Exercise 2.6* Convince yourself that the above statement is indeed correct by examining all cases.

Chapter 3

Random Variables (Discrete Case)

3.1 Basic Definitions

Consider a probability space $(\Omega, \mathscr{F}, P)$, which corresponds to an "experiment". The points $\omega \in \Omega$ represent all possible outcomes of the experiment. In many cases, we are not necessarily interested in the point ω itself, but rather in some property (function) of it. Consider the following pedagogical example: in a coin toss, a perfectly legitimate sample space is the set of initial conditions of the toss (position, velocity and angular velocity of the toss, complemented perhaps with wind conditions). Yet, all we are interested in is a very complicated function of this sample space: whether the coin ended up with Head or Tail showing up. The following is a "preliminary" version of a definition that will be refined further below:

Definition 3.1 Let $(\Omega, \mathscr{F}, P)$ *be a probability space. A function* $X : \Omega \to S$ *(where* $S \subset \mathbb{R}$ *is a set) is called a* ***real-valued random variable****.*

Example: Two dice are tossed and the random variable X is the sum, i.e.,

$$X((i, j)) = i + j.$$

Note that the set S (the range of X) can be chosen to be $\{2, \dots, 12\}$. Suppose now that all our probabilistic interest is in the value of X, rather than the outcome of the individual dice. In such case, it seems reasonable to construct a new probability space in which the sample space is S. Since it is a discrete space, the events related to X can be taken to be all subsets of S. But now we need a probability function on $(S, 2^S)$, which will be compatible with the experiment. If $A \in 2^S$ (e.g., the sum was greater than 5), the probability that A has occurred is given by

$$P(\{\omega \in \Omega : \; X(\omega) \in A\}) = P(\left\{\omega \in \Omega : \; \omega \in X^{-1}(A)\right\}) = P(X^{-1}(A)).$$

That is, the probability function associated with the experiment $(S, 2^S)$ is $P \circ X^{-1}$. We call it the ***distribution*** of the random variable X and denote it by P_X. ▲▲▲

Generalization These notions need to be formalized and generalized. In probability theory, a space (the sample space) comes with a structure (a σ-algebra of events). Thus, when we consider a function from the sample space Ω to some other space S, this other space must come with its own structure—its own σ-algebra of events, which we denote by $\mathscr{F}_S$.

The function $X : \Omega \to S$ is not necessarily one-to-one (although it can always be made onto by restricting S to be the range of X), therefore X is not necessarily invertible. Yet, the inverse function X^{-1} can acquire a well-defined meaning if we define it on subsets of S,

$$X^{-1}(A) = \{\omega \in \Omega : \; X(\omega) \in A\}, \qquad \forall A \in \mathscr{F}_S.$$

There is however nothing to guarantee that for every event $A \in \mathscr{F}_S$ the set $X^{-1}(A) \subset \Omega$ is an event in $\mathscr{F}$. This is something we want to avoid otherwise it will make no sense to ask "what is the probability that $X(\omega) \in A$?".

Definition 3.2 Let $(\Omega, \mathscr{F})$ and $(S, \mathscr{F}_S)$ be two measurable spaces (a set and a σ-algebra of subsets). A function $X : \Omega \to S$ is called a ***random variable*** *if $X^{-1}(A) \in \mathscr{F}$ for all $A \in \mathscr{F}_S$. (In the context of measure theory it is called a* ***measurable function****.)*[1]

An important property of the inverse function X^{-1} is that it preserves (commutes with) set-theoretic operations:

Proposition 3.1 Let X be a random variable mapping a measurable space $(\Omega, \mathscr{F})$ into a measurable space $(S, \mathscr{F}_S)$. Then,

[1] Note the analogy with the definition of continuous functions between topological spaces.

① *For every event* $A \in \mathscr{F}_S$

$$(X^{-1}(A))^c = X^{-1}(A^c).$$

② *If* $A, B \in \mathscr{F}_S$ *are disjoint so are* $X^{-1}(A), X^{-1}(B) \in \mathscr{F}$.

③ $X^{-1}(S) = \Omega$.

④ *If* $(A_n) \subset \mathscr{F}_S$ *is a sequence of events, then*

$$X^{-1}(\cap_{n=1}^{\infty} A_n) = \cap_{n=1}^{\infty} X^{-1}(A_n).$$

Proof: Just follow the definitions. ■

Example: Let A be an event in a measurable space $(\Omega, \mathscr{F})$. An event is not a random variable, however, we can always form from an event a binary random variable (a ***Bernoulli variable***), as follows:

$$I_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \text{otherwise} \end{cases}.$$

▲▲▲

So far, we completely ignored probabilities and only concentrated on the structure that the function X induces on the measurable spaces that are its domain and range. Now, we remember that a probability function is defined on $(\Omega, \mathscr{F})$. We want to define the probability function that it induces on $(S, \mathscr{F}_S)$.

Definition 3.3 Let X *be an* $(S, \mathscr{F}_S)$*-valued random variable on a probability space* $(\Omega, \mathscr{F}, P)$*. Its* **distribution** P_X *is a function* $\mathscr{F}_S \to \mathbb{R}$ *defined by*

$$P_X(A) = P(X^{-1}(A)).$$

Proposition 3.2 The distribution P_X *is a probability function on* $(S, \mathscr{F}_S)$.

Proof: The range of P_X is obviously $[0,1]$. Also

$$P_X(S) = P(X^{-1}(S)) = P(\Omega) = 1.$$

Finally, let $(A_n) \subset \mathscr{F}_S$ be a sequence of disjoint events, then

$$\begin{aligned} P_X(\uplus_{n=1}^{\infty} A_n) = P(X^{-1}(\uplus_{n=1}^{\infty} A_n)) &= P(\uplus_{n=1}^{\infty} X^{-1}(A_n)) \\ = \sum_{n=1}^{\infty} P(X^{-1}(A_n)) &= \sum_{n=1}^{\infty} P_X(A_n). \end{aligned}$$

■

Comment: The distribution is defined such that the following diagram commutes

$$\begin{tikzcd} \Omega \arrow[r, "X"] \arrow[d, "\in"'] & S \arrow[d, "\in"'] \\ \mathscr{F} \arrow[d, "P"'] & \mathscr{F}_S \arrow[l, "X^{-1}"'] \arrow[d, "P_X"'] \\ {[0,1]} \arrow[r, equal] & {[0,1]} \end{tikzcd}$$

In this chapter, we restrict our attention to random variables whose ranges S are discrete spaces, and take $\mathscr{F}_S = 2^S$. Then the distribution is fully specified by its value for all singletons,

$$P_X(\{s\}) =: p_X(s), \qquad s \in S.$$

We call the function p_X the ***atomistic distribution of the random variable*** X. Note the following identity,

$$p_X(s) = P_X(\{s\}) = P(X^{-1}(\{s\}) = P\left(\{\omega \in \Omega : \; X(\omega) = s\}\right),$$

where

$$P : \mathscr{F} \to [0,1] \qquad P_X : \mathscr{F}_S \to [0,1] \qquad p_X : S \to [0,1]$$

are the probability, the distribution of X, and the atomistic distribution of X, respectively. The function p_X is also called the ***probability mass function*** (PMF) of the random variable X.

Notation: We will often have expressions of the form

$$P(\{\omega : X(\omega) \in A\}),$$

which we will write in short-hand notation, $P(X \in A)$ (to be read as "the probability that the random variable X assumes a value in the set A").

Example: Three balls are extracted from an urn containing 20 balls numbered from one to twenty. What is the probability that at least one of the three has a number 17 or higher.

The sample space is

$$\Omega = \{(i, j, k) : 1 \le i < j < k \le 20\},$$

and for every $\omega \in \Omega$, $P(\{\omega\}) = 1/\binom{20}{3}$. We define the random variable

$$X((i, j, k)) = k.$$

It maps every point $\omega \in \Omega$ into a point in the set $S = \{3, \dots, 20\}$. To every $k \in S$ corresponds an event in $\mathscr{F}$,

$$X^{-1}(\{k\}) = \{(i, j, k) : \ 1 \le i < j < k\}.$$

The atomistic distribution of X is

$$p_X(k) = P_X(\{k\}) = P(X = k) = \frac{\binom{k-1}{2}}{\binom{20}{3}}.$$

Then,

$$\begin{aligned} P_X(\{17, \dots, 20\}) &= p_X(17) + p_X(18) + p_X(19) + p_X(20) \\ &= \binom{20}{3}^{-1} \left\{ \binom{16}{2} + \binom{17}{2} + \binom{18}{2} + \binom{19}{2} \right\} \approx 0.508. \end{aligned}$$

▲▲▲

Example: Let A be an event in a probability space $(\Omega, \mathscr{F}, P)$. We have already defined the random variables $I_A : \Omega \to \{0, 1\}$. The distribution of I_A is determined by its value for the two singletons $\{0\}, \{1\}$. Now,

$$P_{I_A}(\{1\}) = P(I_A^{-1}(\{1\})) = P(\{\omega : I_A(\omega) = 1\}) = P(A).$$

▲▲▲

Example: **The coupon collector problem.** Consider the following situation: there are N types of coupons. A coupon collector gets each time unit a coupon at random. The probability of getting each time a specific coupon is $1/N$, independently of prior selections. Thus, our sample space consists of infinite sequences of coupon selections, $\Omega = \{1, \dots, N\}^{\mathbb{N}}$, and for every finite sub-sequence the corresponding probability space is that of equal probability.

A random variable of particular interest is the number of time units T until the coupon collector has gathered at least one coupon of each sort. This random variable takes values in the set $S = \{N, N+1, \dots\} \cup \{\infty\}$. Our goal is to compute its atomistic distribution $p_T(k)$.

Fix an integer $n \geq N$, and define the events $A_1, A_2, \dots, A_N$ such that A_j is the event that no type-j coupon is among the first n coupons. By the inclusion-exclusion principle,

$$\begin{aligned} P_T(\{n+1, n+2, \dots\}) &= P\left(\cup_{j=1}^N A_j\right) \\ &= \sum_j P(A_j) - \sum_{j<k} P(A_j \cap A_k) + \dots . \end{aligned}$$

Now, by the independence of selections, $P(A_j) = [(N-1)/N]^n$, $P(A_j \cap A_k) = [(N-2)/N]^n$, and so on, so that

$$\begin{aligned} P_T(\{n+1, n+2, \dots\}) &= N\left(\frac{N-1}{N}\right)^n - \binom{N}{2}\left(\frac{N-2}{N}\right)^n + \dots \\ &= \sum_{j=1}^N \binom{N}{j}(-1)^{j+1}\left(\frac{N-j}{N}\right)^n . \end{aligned}$$

Finally,

$$p_T(n) = P_T(\{n, n+1, \dots\}) - P_T(\{n+1, n+2, \dots\}).$$

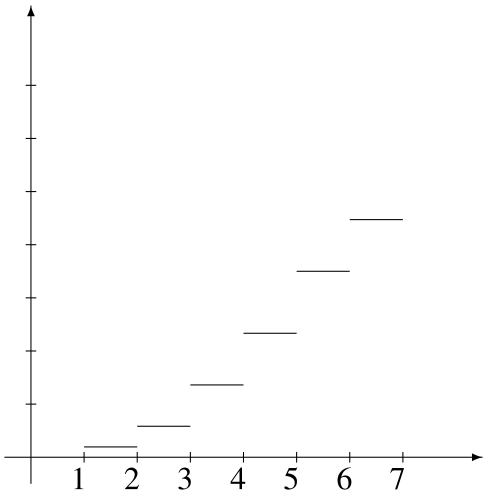
▲▲▲

3.2 The Distribution Function

*Definition 3.4 Let $X : \Omega \to S$ be a real-valued random variable ($S \subseteq \mathbb{R}$). Its **distribution function** F_X is a real-valued function $\mathbb{R} \to \mathbb{R}$ defined by*

$$F_X(x) = P(\{\omega : X(\omega) \leq x\}) = P(X \leq x) = P_X((-\infty, x]).$$

Example: Consider the experiment of tossing two dice and the random variable $X(i, j) = i + j$. Then, $F_X(x)$ is of the form



▲▲▲

Proposition 3.3 The distribution function F_X of any random variable X satisfies the following properties:

1. *F_X is non-decreasing.*
2. *$F_X(x)$ tends to zero when $x \to -\infty$.*
3. *$F_X(x)$ tends to one when $x \to \infty$.*
4. *F_x is right-continuous.*

Proof:

1. Let $a \le b$, then $(-\infty, a] \subseteq (-\infty, b]$ and since P_X is a probability function,

$$F_X(a) = P_X((-\infty, a]) \le P_X((-\infty, b]) = F_X(b).$$

2. Let (x_n) be a sequence that converges to $-\infty$. Then,

$$\lim_n F_X(x_n) = \lim_n P_X((-\infty, x_n]) = P_X(\lim_n(-\infty, x_n]) = P_X(\emptyset) = 0.$$

3. The other limit is treated along the same lines.

4. Same for right continuity: if the sequence (h_n) converges to zero from the right, then

$$\begin{aligned}\lim_n F_X(x + h_n) &= \lim_n P_X((-\infty, x + h_n]) \\ &= P_X(\lim_n(-\infty, x + h_n]) \\ &= P_X((-\infty, x]) = F_X(x).\end{aligned}$$

■

✎ *Exercise 3.1* Explain why F_X is not necessarily left-continuous.

What is the importance of the distribution function? A distribution is a complicated object, as it has to assign a number to any set in the range of X (for the moment, let's forget that we deal with discrete variables and consider the more general case where S may be a continuous subset of $\mathbb{R}$). The distribution function is a real-valued function (much simpler object) which embodies the same information. That is, the distribution function defines uniquely the distribution of any (measurable) set in $\mathbb{R}$. For example, the distribution of semi-open segments is

$$P_X((a, b]) = P_X((-\infty, b] \setminus (-\infty, a]) = F_X(b) - F_X(a).$$

What about open segments?

$$P_X((a, b)) = P_X(\lim_n(a, b - 1/n]) = \lim_n P_X((a, b - 1/n]) = F_X(b^-) - F_X(a).$$

Since every measurable set in $\mathbb{R}$ is a countable union of closed, open, or semi-open disjoint segments, the probability of any such set is fully determined by F_X.

3.3 The binomial distribution

Definition 3.5 A random variable over a probability space is called a ***Bernoulli variable*** *if its range is the set* $\{0, 1\}$*. The distribution of a Bernoulli variable* X *is determined by a single parameter* $p_X(1) := p$*. In fact, a Bernoulli variable can be identified with a two-state probability space.*

Definition 3.6 *A* ***Bernoulli process*** *is a compound experiment whose constituents are n independent Bernoulli trials. It is a probability space with sample space*

$$\Omega = \{0, 1\}^n,$$

and probability defined on singletons,

$$P(\{(a_1, \dots, a_n)\}) = p^{\text{number of ones}}(1 - p)^{\text{number of zeros}}.$$

Consider a Bernoulli process (this defines a probability space), and set the random variable X to be the number of "ones" in the sequence (the number of successes out of n repeated Bernoulli trials). The range of X is $\{0, \dots, n\}$, and its atomistic distribution is

$$p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}. \tag{3.1}$$

(Note that it sums up to one.)

Definition 3.7 *A random variable X over a probability space* $(\Omega, \mathscr{F}, P)$ *is called a* ***binomial variable*** *with parameters* (n, p) *if it takes integer values between zero and n and its atomistic distribution is* (3.1). *We write* $X \sim \mathscr{B}(n, p)$.

Discussion: A really important point: one often encounters problems starting with a statement "X is a binomial random variable, what is the probability that bla bla bla.." without any mention of the underlying probability space $(\Omega, \mathscr{F}, P)$. It this legitimate? There are two answers to this point: (i) if the question only addresses the random variable X, then it can be fully solved knowing just the distribution P_X; the fact that there exists an underlying probability space is irrelevant for the sake of answering this kind of questions. (ii) The triple $(\Omega = \{0, 1, \dots, n\}, \mathscr{F} = 2^\Omega, P = P_X)$ is a perfectly legitimate probability space. In this context the random variable X is the trivial map $X(x) = x$.

Example: Diapers manufactured by Pamp-ggies are defective with probability 0.01. Each diaper is defective or not independently of other diapers. The company sells diapers in packs of 10. The customer gets his/her money back only if *more than one* diaper in a pack is defective. What is the probability for that to happen?

Every time the customer takes a diaper out of the pack, he faces a Bernoulli trial. The sample space is $\{0, 1\}$ (1 is defective) with $p(1) = 0.01$ and $p(0) = 0.99$. The number of defective diapers X in a pack of ten is a binomial variable $\mathscr{B}(10, 0.01)$. The probability that X be larger than one is

$$\begin{aligned} P_X(\{2,3,\dots,10\}) &= 1 - P_X(\{0,1\}) \\ &= 1 - p_X(0) - p_X(1) \\ &= 1 - \binom{10}{0}(0.01)^0(0.99)^{10} - \binom{10}{1}(0.01)^1(0.99)^9 \\ &\approx 0.07. \end{aligned}$$

▲▲▲

Example: An airplane engine breaks down during a flight with probability $1 - p$. An airplane lands safely only if *at least half* of its engines are functioning upon landing. What is preferable: a two-engine airplane or a four-engine airplane (or perhaps you'd better walk)?

Here again, the number of functioning engines is a binomial variable, in one case $X_1 \sim \mathscr{B}(2, p)$ and in the second case $X_2 \sim \mathscr{B}(4, p)$. The question is whether $P_{X_1}(\{1,2\})$ is larger than $P_{X_2}(\{2,3,4\})$ or the other way around. Now,

$$\begin{aligned} P_{X_1}(\{1,2\}) &= \binom{2}{1}p^1(1-p)^1 + \binom{2}{2}p^2(1-p)^0 \\ P_{X_2}(\{2,3,4\}) &= \binom{4}{2}p^2(1-p)^2 + \binom{4}{3}p^3(1-p)^1 + \binom{4}{4}p^4(1-p)^0. \end{aligned}$$

Opening the brackets,

$$\begin{aligned} P_{X_1}(\{1,2\}) &= 2p(1-p) + p^2 = 2p - p^2 \\ P_{X_2}(\{2,3,4\}) &= 6p^2(1-p)^2 + 4p^3(1-p) + p^4 = 3p^4 - 8p^3 + 6p^2. \end{aligned}$$

One should prefer the four-engine airplane if

$$p(3p^3 - 8p^2 + 7p - 2) > 0,$$

which factors into

$$p(p-1)^2(3p-2) > 0,$$

and this holds only if $p > 2/3$. That is, the higher the probability for a defective engine, less engines should be used. ▲▲▲

Everybody knows that when you toss a fair coin 100 times it will fall Head 50 times... well, at least we know that 50 is the most probable outcome. How probable is in fact this outcome?

Example: A fair coin is tossed $2n$ times, with $n \gg 1$. What is the probability that the number of Heads equals exactly n?

The number of Heads is a binomial variable $X \sim \mathscr{B}\left(2n, \frac{1}{2}\right)$. The probability that X equals n is given by

$$p_X(n) = \binom{2n}{n}\left(\frac{1}{2}\right)^n\left(\frac{1}{2}\right)^n = \frac{(2n)!}{(n!)^2\, 2^{2n}}.$$

To evaluate this expression we use Stirling's formula, $n! \sim \sqrt{2\pi} n^{n+1/2} e^{-n}$, thus,

$$p_X(n) \sim \frac{\sqrt{2\pi}(2n)^{2n+1/2} e^{-2n}}{2^{2n}\, 2\pi n^{2n+1} e^{-2n}} = \frac{1}{\sqrt{\pi\, n}}$$

For example, with a hundred tosses ($n = 50$) the probability that exactly half are Heads is approximately $1/\sqrt{50\pi} \approx 0.08$. ▲▲▲

We conclude this section with a simple fact about the atomistic distribution of a Binomial variable:

Proposition 3.4 Let $X \sim \mathscr{B}(n, p)$, then $p_X(k)$ increases until it reaches a maximum at $k = \lfloor (n+1)p \rfloor$, and then decreases.

Proof: Consider the ratio $p_X(k)/p_X(k-1)$,

$$\frac{p_X(k)}{p_X(k-1)} = \frac{n!(k-1)!(n-k+1)!\, p^k(1-p)^{n-k}}{k!(n-k)!n! p^{k-1}(1-p)^{n-k+1}} = \frac{(n-k+1)p}{k(1-p)}.$$

$p_X(k)$ is increasing if

$$(n-k+1)p > k(1-p) \qquad \Rightarrow \qquad (n+1)p - k > 0.$$

■

✎ *Exercise 3.2* In a sequence of Bernoulli trials with probability p for success, what is the probability that a successes will occur before b failures? (Hint: the issue is decided after at most $a + b - 1$ trials).

✎ *Exercise 3.3* Show that the probability of getting *exactly n* Heads in $2n$ tosses of a fair coin satisfies the asymptotic relation

$$P(n \text{ Heads}) \sim \frac{1}{\sqrt{\pi n}}.$$

Remind yourself what we mean by the ~ sign.

3.4 The Poisson distribution

Definition 3.8 A random variable X is said to have a ***Poisson distribution*** *with parameter λ, if it takes values $S = \{0, 1, 2, \dots, \}$, and its atomistic distribution is*

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

(Prove that this defines a probability distribution.) We write $X \sim \mathcal{P}oi(\lambda)$.

The first question any honorable person should ask is "why"? After all, we can define infinitely many such distributions, and give them fancy names. The answer is that certain distributions are important because they frequently occur is real life. The Poisson distribution appears abundantly in life, for example, when we measure the number of radio-active decays in a unit of time. In fact, the following analysis reveals the origins of this distribution.

Comment: Remember the inattentive secretary. When the number of letters n is large, we saw that the probability that exactly k letters reach their destination is approximately a Poisson variable with parameter $\lambda = 1$.

Consider the following model for radio-active decay. Every ϵ seconds (a very short time) a single decay occurs with probability proportional to the length of the time interval: $\lambda\epsilon$. With probability $1 - \lambda\epsilon$ no decay occurs. Physics tells us that this probability is independent of history. The number of decays in one second is therefore a binomial variable $X \sim \mathcal{B}(n = 1/\epsilon, p = \lambda\epsilon)$. Note how as $\epsilon \to 0$, n goes to infinity and p goes to zero, but their product remains finite. The probability of observing k decays in one second is

$$\begin{aligned} p_X(k) &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!}\frac{n(n-1)\dots(n-k+1)}{n^k}\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!}\left(1-\frac{1}{n}\right)\dots\left(1-\frac{k-1}{n}\right)\frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^k}. \end{aligned}$$

Taking the limit $n \to \infty$ we get

$$\lim_{n\to\infty} p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Thus *the Poisson distribution arises from a Binomial distribution when the probability for success in a single trial is very small but the number of trials is very large such that their product is finite.*

Example: Suppose that the number of typographical errors in a page is a Poisson variable with parameter $1/2$. What is the probability that there is at least one error?

This exercise is here mainly for didactic purposes. As always, we need to start by constructing a probability space. The data tells us that the natural space to take is the sample space $\Omega = \mathbb{N}$ with a probability $P(\{k\}) = e^{-1/2}/(2^k k!)$. Then the answer is

$$P(\{k \in \mathbb{N} : k \geq 1\}) = 1 - P(\{0\}) = 1 - e^{-1/2} \approx 0.395.$$

While this is a very easy exercise, note that we converted the data about a "Poisson variable" into a probability space over the natural numbers with a Poisson distribution. Indeed, a random variable *is* a probability space. ▲▲▲

✎ *Exercise 3.4* Assume that the number of eggs laid by an insect is a Poisson variable with parameter λ. Assume, furthermore, that every egg has a probability p to develop into an insect. What is the probability that exactly k insects will survive? If we denote the number of survivors by X, what kind of random variable is X? (Hint: construct first a probability space as a compound experiment).

3.5 The Geometric distribution

Consider an infinite sequence of Bernoulli trials with parameter p, i.e., $\Omega = \{0,1\}^{\mathbb{N}}$, and define the random variable X to be the number of trials until the first success is met. This random variables takes values in the set $S = \{1,2,\dots\}$. The probability that X equals k is the probability of having first $(k-1)$ failures followed by a success:

$$p_X(k) = P_X(\{k\}) = P(X = k) = (1-p)^{k-1}p.$$

A random variable having such an atomistic distribution is said to have a ***geometric distribution*** with parameter p; we write $X \sim \mathcal{G}eo(p)$.

Comment: The number of failures until the success is met, i.e., $X-1$, is also called a geometric random variable. We will stick to the above definition.

Example: There are N white balls and M black balls in an urn. Each time, we take out one ball (with replacement) until we have a black ball. (1) What is the probability that we need k trials? (2) What is the probability that we need at least n trials.

The number of trials X is distributed $\mathcal{G}eo(M/(M+N))$. (1) The answer is simply

$$\left(\frac{N}{M+N}\right)^{k-1}\frac{M}{M+N} = \frac{N^{k-1}M}{(M+N)^k}.$$

(2) The answer is

$$\frac{M}{M+N}\sum_{k=n}^{\infty}\left(\frac{N}{M+N}\right)^{k-1} = \frac{M}{M+N}\frac{\left(\frac{N}{M+N}\right)^{n-1}}{1-\frac{N}{M+N}} = \left(\frac{N}{M+N}\right)^{n-1},$$

which is obviously the probability of failing the first $n-1$ times.

▲▲▲

An important property of the geometric distribution is its lack of memory. That is, the probability that $X = n$ given that $X > k$ is the same as the probability that $X = n-k$ (if we know that we failed the first k times, it does not imply that we will succeed earlier when we start the $k+1$-st trial, that is

$$P_X(\{n\}|\{k+1,\dots\}) = p_X(n-k).$$

This makes sense even if $n \leq k$, provided we extend P_X to all $\mathbb{Z}$. To prove this claim we follow the definitions. For $n > k$,

$$\begin{aligned} P_X(\{n\}|\{k+1, k+2\dots\}) &= \frac{P_X(\{n\} \cap \{ k+1, \dots\})}{P_X(\{k+1, k+2, \dots\})} \\ &= \frac{P_X(\{n\})}{P_X(\{k+1, k+2, \dots\})} \\ &= \frac{(1-p)^{n-1}p}{(1-p)^k} = (1-p)^{n-k-1}p = p_X(n-k). \end{aligned}$$

3.6 The negative-binomial distribution

A coin with probability p for Heads is tossed until a total of n Heads is obtained. Let X be the number of failures until n successes were met. We say that X has the ***negative-binomial distribution*** with parameters (n, p). What is $p_X(k)$ for $k = 0, 1, 2\dots$? The answer is simply

$$p_X(k) = \binom{n+k-1}{k} p^n (1-p)^k.$$

This is is a special instance of negative-binomial distribution, which can be extended to non-integer n. To allow for non-integer n we note that for integers $\Gamma(n) = (n-1)!$. Thus, the general negative-binomial distribution with parameters $0 < p < 1$ and $r > 0$ has the atomistic distribution,

$$p_X(k) = \frac{\Gamma(r+k)}{k!\,\Gamma(r)} p^r (1-p)^k.$$

We write $X \sim \mathcal{nBin}(r, p)$.

3.7 Other examples

Example: Here is a number theoretic result derived by probabilistic means. Let $s > 1$ and let X be a random variable taking values in $\{1, 2\dots,\}$ with atomistic distribution,

$$p_X(k) = \frac{k^{-s}}{\zeta(s)},$$

where ζ is the ***Riemann zeta-function***,

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s}.$$

Let $A_m \subset \mathbb{N}$ be the set of integers that are divisible by m. Clearly,

$$P_X(A_m) = \frac{\sum_{k \text{ divisible by } m} k^{-s}}{\sum_{n=1}^{\infty} n^{-s}} = \frac{\sum_{k=1}^{\infty} (mk)^{-s}}{\sum_{n=1}^{\infty} n^{-s}} = m^{-s}.$$

Next, we claim that the events $E_p = \{X \in A_p\}$, with p primes are independent. Indeed, $A_p \cap A_q = A_{pq}$, so that

$$P_X(A_p \cap A_q) = P_X(A_{pq}) = (pq)^{-s} = p^{-s} q^{-s} = P_X(A_p) P_X(A_q).$$

The same consideration holds for all collections of A_p.

Next, we note that

$$\bigcap_{p \text{ prime}} A_p^c = \{1\},$$

from which, together with the independence of the A_p, follows that

$$p_X(1) = \prod_{p \text{ prime}} P_X(A_p^c),$$

i.e.,

$$\frac{1}{\zeta(s)} = \prod_{p \text{ prime}} \left(1 - \frac{1}{p^s}\right),$$

an identity known as ***Euler's formula***.

A consequence from this formula is obtained by letting $s \to 1$,

$$\prod_{p \text{ prime}} \left(1 - \frac{1}{p}\right) = 0.$$

Taking the logarithm we get that

$$\sum_{p \text{ prime}} \log\left(1 - \frac{1}{p}\right) = -\infty.$$

Since for $0 < x < 0.6$ we have $\log(1 - x) \geq -2x$ it follows that

$$-\infty = \sum_{p \text{ prime}} \log\left(1 - \frac{1}{p}\right) \geq -2 \sum_{p \text{ prime}} \frac{1}{p},$$

i.e., the ***harmonic prime series*** diverges. ▲▲▲

3.8 Jointly-distributed random variables

Consider a probability space $(\Omega, \mathscr{F}, P)$ and a pair of random variables, X and Y. That is, we have two maps between probability spaces:

$$(\Omega, \mathscr{F}, P) \xrightarrow{X} (S_X, \mathscr{F}_X, P_X)$$
$$(\Omega, \mathscr{F}, P) \xrightarrow{Y} (S_Y, \mathscr{F}_Y, P_Y).$$

Recall that the probability that X be in a set $A \in \mathscr{F}_X$ is fully determined by the distribution P_X. Now, try to answer the following question: suppose that we are only given the distributions P_X and P_Y (i.e., we don't know P). What is the probability that $X(\omega) \in A$ *and* $Y(\omega) \in B$, where $A \in \mathscr{F}_X$ and $B \in \mathscr{F}_Y$? We cannot answer this question. The knowledge of the *separate* distributions of X and Y is insufficient to answer questions about events that are *joint* to X and Y.

The correct way to think about a pair of random variables, is as a mapping $\Omega \to S_X \times S_Y$, i.e.,

$$\omega \mapsto (X(\omega), Y(\omega)).$$

As always, we need to equip $S_X \times S_y$ with a σ-algebra of events $\mathscr{F}_{X,Y}$ and we require that every set $A \in \mathscr{F}_{X,Y}$ has a pre-image in $\mathscr{F}$. In fact, given the σ-algebra $\mathscr{F}_{X,Y}$, the σ-algebra $\mathscr{F}_X$ is a restriction of $\mathscr{F}_{X,Y}$,

$$\mathscr{F}_X = \{A \subseteq S_X : A \times S_Y \in \mathscr{F}_{X,Y}\},$$

and similarly for $\mathscr{F}_Y$.

The ***joint distribution*** of the pair X, Y is defined naturally as

$$P_{X,Y}(A) := P(\{\omega \in \Omega : (X(\omega), Y(\omega)) \in A\}).$$

Note that one can infer the individual (***marginal***) distributions of X and Y from this joint distribution, as

$$\begin{aligned} P_X(A) &= P_{X,Y}(A \times S_Y) \qquad && A \in \mathscr{F}_X \\ P_Y(B) &= P_{X,Y}(S_X \times B) \qquad && B \in \mathscr{F}_Y. \end{aligned}$$

When both S_X and S_Y are countable spaces, we define the ***atomistic joint distribution***,

$$p_{X,Y}(x, y) := P_{X,Y}(\{(x, y)\}) = P(X = x, Y = y).$$

Obviously,

$$p_X(x) = P_{X,Y}(\{x\} \times S_Y) = \sum_{y \in S_Y} p_{X,Y}(x,y)$$
$$p_Y(y) = P_{X,Y}(S_X \times \{y\}) = \sum_{x \in S_X} p_{X,Y}(x,y).$$

Finally, we define the ***joint distribution function***,

$$F_{X,Y}(x,y) := P_{X,Y}((-\infty,x] \times (-\infty,y]) = P(X \le x, Y \le y).$$

Example: There are three red balls, four white balls and five blue balls in an urn. We extract three balls. Let X be the number of red balls and Y the number of white balls. What is the joint distribution of X and Y?

The natural probability space here is the set of triples out of twelve elements. We have

$$(X,Y) : \Omega \to \{(i,j) : i,j \ge 0, i+j \le 3\}.$$

For example,

$$p_{X,Y}(0,0) = \frac{\binom{5}{3}}{\binom{3}{12}} \quad p_{X,Y}(1,1) = \frac{\binom{3}{1}\binom{4}{1}\binom{5}{1}}{\binom{12}{3}},$$

etc. ▲▲▲

✎ *Exercise 3.5* Construct two probability spaces, and on each define two random variables, X, Y, such that two P_X are the same and the two P_Y are the same, but the $P_{X,Y}$ differ.

These notions can be easily generalized to n random variables. $X_1, \dots, X_n$ are viewed as a function from Ω to the product set $S_1 \times \cdots \times S_n$, with joint distribution

$$P_{X_1,\dots,X_n}(A) = P(\{\omega : (X_1(\omega), \dots, X_n(\omega)) \in A\}),$$

where $A \in S_1 \times \cdots \times S_n$. The ***marginal distributions*** of subsets of variables, are obtained, as for example,

$$P_{X_1,\dots,X_{n-1}}(A) = P_{X_1,\dots,X_n}(A \times S_n),$$

with $A \subseteq S_1 \times S_2 \times \cdots \times S_{n-1}$.

3.9 Independence of random variables

Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $X : \Omega \to S$ be a random variable, where the set S is equipped with its σ-algebra of events $\mathscr{F}_S$. By the very definition of a random variable, for every event $A \in \mathscr{F}_S$, the event $X^{-1}(A)$ is an element of $\mathscr{F}$. That is,

$$X^{-1}(\mathscr{F}_S) = \left\{X^{-1}(A) : A \in \mathscr{F}_S\right\} \subseteq \mathscr{F}.$$

It is easily verified (given as an exercise) that $X^{-1}(\mathscr{F}_S)$ is a σ-algebra, i.e., a sub-σ-algebra of $\mathscr{F}$. We call it ***the σ-algebra generated by the random variable*** X, and denote it by $\sigma(X)$. Events in $\sigma(X)$ are subsets of Ω (not of S) that characterize the outcome of $X(\omega)$.

Similarly, when we have a pair of random variables X, Y with a σ-algebra $\mathscr{F}_{X,Y}$, they generate (together!) a σ-algebra, $\sigma(X, Y)$, which consists of all events of the form

$$\{\omega \in \Omega : (X(\omega), Y(\omega)) \in A\},$$

with $A \in \mathscr{F}_{X,Y}$. Note that in general $\mathscr{F}_{X,Y} \supseteq \mathscr{F}_X \times \mathscr{F}_Y$, from which follows that $\sigma(X, Y) \supseteq \sigma(X) \cup \sigma(Y)$. Indeed, $\sigma(X) \cup \sigma(Y)$ comprises events of the form

$$\{\omega \in \Omega : \; X(\omega) \in A, \; Y(\omega) \in B\},$$

whereas $\sigma(X, Y)$ comprises a larger family of events of the form

$$\{\omega \in \Omega : \; (X(\omega), Y(\omega)) \in C\}.$$

✎ *Exercise 3.6* Let X be a random variable (=a measurable mapping) from $(\Omega, \mathscr{F}, P)$ to the space $(S, \mathscr{F}_S, P_X)$. Consider the collection of events,

$$\{X^{-1}(A) : \; A \in \mathscr{F}_S\},$$

which is by assumption a subset of $\mathscr{F}$. Prove that this collection is a σ-algebra.

We are now ready to define the independence of two random variables. Recall that we already have a definition for the independence of events:

Definition 3.9 Two random variables X, Y over a probability space $(\Omega, \mathscr{F}, P)$ are said to be independent if every event in $\sigma(X)$ is independent of every event in $\sigma(Y)$. In other words, they are independent if every information associated with the value of X does not affect the (conditional) probability of events regarding the random variable Y.

Example: Consider the probability space associated with tossing two dice, and let $X(\omega)$ be the sum of the dice and $Y(\omega)$ be the value of the first die, i.e.,

$$X((i, j)) = i + j \qquad Y((i, j)) = i.$$

The ranges of X and Y are $S_X = \{2, \dots, 12\}$ and $S_Y = \{1, \dots, 6\}$, respectively. The σ-algebra generated by X is the collection of events of the form

$$X^{-1}(A) = \{(i, j) \in \Omega : \; i + j \in A, A \subseteq \{2, \dots, 12\}\} \in \sigma(X),$$

whereas the σ-algebra generated by Y is the collection of events of the form

$$Y^{-1}(B) = \{(i, j) \in \Omega : \; i \in B \subseteq \{1, \dots, 6\}\} \in \sigma(Y).$$

Recall that the events $X^{-1}(\{7\})$ and $Y^{-1}(\{3\})$ are independent. Does it mean that X and Y are independent variables? No, for example $X^{-1}(\{6\})$ and $Y^{-1}(\{3\})$ are dependent. It is not true that *any* information on the outcome of X does not change the probability of the outcome of Y. ▲▲▲

While the definition of independence may seem hard to work with, it is easily translated into simpler terms. Let $A \times B$ be an event in $\mathscr{F}_{X,Y}$ with $A \in \mathscr{F}_X$ and $B \in \mathscr{F}_Y$. If X and Y are independent, then

$$\begin{aligned} P_{X,Y}(A \times B) &= P(X \in A, Y \in B) \\ &= P(\{\omega : X(\omega) \in A\} \cap \{\omega : Y(\omega) \in B\}) \\ &= P(X \in A)P(Y \in B) \\ &= P_X(A)P_Y(B). \end{aligned}$$

In particular, if $A = \{x\}$ and $B = \{y\}$ are singletons, then

$$p_{X,Y}(x, y) = p_X(x)p_Y(y).$$

Finally, if $A = (-\infty, x]$ and $B = (-\infty, y]$, then

$$F_{X,Y}(x, y) = F_X(x)F_Y(y).$$

Thus, two random variables are independent only if their joint distribution (atomic joint distribution, joint distribution function) factors into a product of distributions.

✎ *Exercise 3.7* Prove that two random variables X, Y are independent *if and only if*

$$P_{X,Y}(A \times B) = P_X(A) P_Y(B)$$

for every $A \in \mathscr{F}_X$ and $B \in \mathscr{F}_Y$.

These definitions are easily generalized to n random variables. The random variables $X_1, \dots, X_n$ have a joint distribution $P_{X_1,\dots,X_n}$ defined on a σ-algebra of events of the form

$$\{\omega \in \Omega : (X_1(\omega), \dots, X_n(\omega)) \in A\}, \qquad A \in S_1 \times \dots \times S_n.$$

These variables are mutually independent if for all $A_1 \in \mathscr{F}_1, \dots, A_n \in \mathscr{F}_n$,

$$P_{X_1,\dots,X_n}(A_1 \times \dots \times A_n) = P_{X_1}(A_1) \dots P_{X_n}(A_n).$$

We further extend the definition to a countable number of random variables. An infinite sequence of random variables is said to me mutually independent if every finite subset is independent.

We will see now a strong use of independence. But first an important lemma, which really is the "second half" of a lemma whose first part we have already seen. Recall the ***first Borel-Cantelli lemma*** that states that if an infinite sequence of events (A_n) has the property that $\sum P(A_n) < \infty$, then

$$P(\limsup_n A_n) = P(A_n; \text{ i.o.}) = 0.$$

There is also a converse lemma, which however requires the independence of the events:

Lemma 3.1 (Second Borel-Cantelli) Let (A_n) be a collection of ***mutually independent*** *events in a probability space $(\Omega, \mathscr{F}, P)$. If $\sum P(A_n) = \infty$, then*

$$P(\limsup_n A_n) = P(A_n; \textit{ i.o.}) = 1.$$

Proof: Note that

$$(\limsup_n A_n)^c = (\cap_{n=1}^\infty \cup_{k=n}^\infty A_k)^c = \cup_{n=1}^\infty \cap_{k=n}^\infty A_k^c = \liminf_n A_n^c.$$

Fix n. Because the events are independent we have

$$P(\cap_{k=n}^{\infty} A_k^c) = \prod_{k=n}^{\infty} (1 - P(A_k)).$$

Using the inequality $1 - x \le e^{-x}$ we have

$$P(\cap_{k=n}^{\infty} A_k^c) \le \prod_{k=n}^{\infty} e^{-P(A_k)} = \exp\left(- \sum_{k=n}^{\infty} P(A_k) \right) = 0,$$

where we have used the divergence of the series. Thus, the event $(\limsup_n A_n)^c$ is a countable union of events that have zero probability, and therefore also has zero probability. It follows that its complement has probability one. ■

✎ *Exercise 3.8* Show, by means of a counter example, why does the second Borel-Cantelli lemma require the independence of the random variables.

Example: Here is simple application of the second Borel-Cantelli lemma. Consider an infinite sequence of Bernoulli trials with probability $0 < p < 1$ for "success". What is the probability that the sequence SFS appears infinitely many times? Let A_j be the event that the sub-sequence $a_j a_{j+1} a_{j+2}$ equals SFS, i.e.,

$$A_j = \left\{ (a_n) \in \{S, F\}^{\mathbb{N}} : \ a_j = S, a_{j+1} = F, a_{j+2} = S \right\}.$$

The events $A_1, A_4, A_7, \ldots$ are independent. Since they have an equal finite probability, $p^2(1 - p)$,

$$\sum_{n=1}^{\infty} P(A_{3n}) = \infty \qquad \Rightarrow \qquad P(\limsup_n A_{3n}) = 1.$$

▲▲▲

Example: Here is a more subtle application of the second Borel-Cantelli lemma. Let (X_n) be an infinite sequence of independent random variables assuming real positive values, and having the following distribution function,

$$F_X(x) = \begin{cases} 0 & x \le 0 \\ 1 - e^{-x} & x > 0 \end{cases}.$$

(Such random variables are called exponential; we shall study them later on). Thus, for any positive x,

$$P(X_j > x) = e^{-x}.$$

In particular, we may ask about the probability that the n-th variable exceeds $\alpha \log n$,

$$P(X_n > \alpha \log n) = e^{-\alpha \log n} = n^{-\alpha}.$$

It follows from the two Borel-Cantelli lemmas that

$$P(X_n > \alpha \log n \quad \text{i.o. }) = \begin{cases} 0 & \alpha > 1 \\ 1 & \alpha \leq 1 \end{cases}.$$

By the same method, we can obtain refined estimates, such as

$$P(X_n > \log n + \alpha \log\log n \quad \text{i.o. }) = \begin{cases} 0 & \alpha > 1 \\ 1 & \alpha \leq 1 \end{cases},$$

and so on. ▲▲▲

3.10 Sums of random variables

Let X, Y be two real-valued random variables (i.e., $S_X \times S_Y \subset R^2$) with joint distribution $P_{X,Y}$. Let $Z = X + Y$. What is the distribution of Z? To answer this question we examine the distribution function of Z, and write it as follows:

$$\begin{aligned} F_Z(z) &= P(X + Y \leq z) \\ &= P(\cup_{x\in S_X}\{\omega : X(\omega) = x, Y(\omega) \leq z - x\}) \\ &= \sum_{x\in S_X} P(\{\omega : X(\omega) = x, Y(\omega) \leq z - x\}) \\ &= \sum_{x\in S_X} \sum_{S_Y \ni y \leq z-x} p_{X,Y}(x, y). \end{aligned}$$

Similarly, we can derive the atomistic distribution of Z,

$$p_Z(z) = p_{X+Y}(z) = \sum_{x\in S_X} \sum_{S_Y \ni y = z-x} p_{X,Y}(x, y) = \sum_{x\in S_X} p_{X,Y}(x, z - x),$$

where the sum may be null if z does not belong to the set

$$S_X + S_Y := \{z : \exists(x,y) \in S_X \times S_Y, z = x + y\}.$$

For the particular case where X and Y are independent we have

$$F_{X+Y}(z) = \sum_{x \in S_X} \sum_{S_Y \ni y \le z-x} p_X(x) p_Y(y) = \sum_{x \in S_X} p_X(x) F_Y(z - x),$$

and

$$p_{X+Y}(z) = \sum_{x \in S_X} p_X(x) p_Y(z - x),$$

the last expression being the ***discrete convolution*** of p_X and p_Y evaluated at the point z.

Example: Let $X \sim \mathcal{Poi}(\lambda_1)$ and $Y \sim \mathcal{Poi}(\lambda_2)$ be independent random variables. What is the distribution of $X + Y$?

Using the convolution formula, and the fact that Poisson variables assume non-negative integer values,

$$\begin{aligned} p_{X+Y}(k) &= \sum_{j=0}^{k} p_X(j) p_Y(k-j) \\ &= \sum_{j=0}^{k} e^{-\lambda_1} \frac{\lambda_1^j}{j!} e^{-\lambda_2} \frac{\lambda_2^{k-j}}{(k-j)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \sum_{j=0}^{k} \binom{k}{j} \lambda_1^j \lambda_2^{k-j} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} (\lambda_1 + \lambda_2)^k, \end{aligned}$$

i.e., the sum of two *independent* Poisson variables is a Poisson variable, whose parameter is the sum of the two parameters. ▲▲▲

✎ *Exercise 3.9* Let $X \sim \mathcal{B}(n,p)$ and $Y \sim \mathcal{B}(m,p)$. Prove that $X + Y \sim \mathcal{B}(n+m,p)$. Give an intuitive explanation for why this must hold.

3.11 Conditional distributions

Recall our definition of the conditional probability: if A and B are events, then

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

This definition embodies the notion of prediction that A has occurred given that B has occurred. We now extend the notion of conditioning to random variables:

Definition 3.10 Let X, Y be (discrete) random variables over a probability space $(\Omega, \mathscr{F}, P)$. We denote their atomistic joint distribution by $p_{X,Y}$; it is a function $S_X \times S_Y \to [0, 1]$. The ***atomistic conditional distribution*** *of X given Y is defined as*

$$p_{X|Y}(x|y) := P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

(It is defined only for values of y for which $p_Y(y) > 0$.)

Example: Let $p_{X,Y}$ be defined by the following table:

Y, X	0	1
0	0.4	0.1
1	0.2	0.3

What is the conditional distribution of X given Y?

Answer:

$$\begin{aligned}
p_{X|Y}(0|0) &= \frac{p_{X,Y}(0, 0)}{p_Y(0)} = \frac{0.4}{0.4 + 0.1} \\
p_{X|Y}(1|0) &= \frac{p_{X,Y}(1, 0)}{p_Y(0)} = \frac{0.1}{0.4 + 0.1} \\
p_{X|Y}(0|1) &= \frac{p_{X,Y}(0, 1)}{p_Y(1)} = \frac{0.2}{0.2 + 0.3} \\
p_{X|Y}(1|1) &= \frac{p_{X,Y}(1, 1)}{p_Y(1)} = \frac{0.3}{0.2 + 0.3}.
\end{aligned}$$

▲▲▲

Note that we always have

$$p_{X,Y}(x, y) = p_{X|Y}(x|y) p_Y(y).$$

Summing over all $y \in S_Y$,

$$p_X(x) = \sum_{y \in S_Y} p_{X,Y}(x,y) = \sum_{y \in S_Y} p_{X|Y}(x|y) p_Y(y),$$

which can be identified as the ***law of total probability*** formulated in terms of random variables.

✎ *Exercise 3.10* True or false: every two random variables X, Y satisfy

$$\sum_{x \in S_X} p_{X|Y}(x|y) = 1$$
$$\sum_{y \in S_Y} p_{X|Y}(x|y) = 1.$$

Example: Assume that the number of eggs laid by an insect is a Poisson variable with parameter λ. Assume, furthermore, that every eggs has a probability p to develop into an insect. What is the probability that exactly k insects will survive?

This problem has been previously given as an exercise. We will solve it now in terms of conditional distributions. Let X be the number of eggs laid by the insect, and Y the number of survivors. We don't even bother to (explicitly) write the probability space, because we have all the needed data as distributions and conditional distributions. We know that X has a Poisson distribution with parameter λ, i.e.,

$$p_X(n) = e^{-\lambda} \frac{\lambda^n}{n!} \qquad n = 0, 1, \dots,$$

whereas the distribution of Y conditional on X is binomial,

$$p_{Y|X}(k|n) = \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0, 1, \dots, n.$$

The distribution of the number of survivors Y is then

$$\begin{aligned} p_Y(k) &= \sum_{n=0}^{\infty} p_{Y|X}(k|n) p_X(n) = \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} e^{-\lambda} \frac{\lambda^n}{n!} \\ &= e^{-\lambda} \frac{(\lambda p)^k}{k!} \sum_{n=k}^{\infty} \frac{[\lambda(1-p)]^{n-k}}{(n-k)!} \\ &= e^{-\lambda} \frac{(\lambda p)^k}{k!} e^{\lambda(1-p)} = e^{-\lambda p} \frac{(\lambda p)^k}{k!}. \end{aligned}$$

Thus, $Y \sim \mathcal{P}oi(\lambda p)$. ▲▲▲

Example: Let $X \sim \mathcal{P}oi(\lambda_1)$ and $Y \sim \mathcal{P}oi(\lambda_2)$ be *independent* random variables. What is the conditional distribution of X given that $X + Y = n$?

We start by writing things explicitly,

$$\begin{aligned} p_{X|X+Y}(k|n) &= P(X = k|X + Y = n) \\ &= \frac{P(X = k, X + Y = n)}{P(X + Y = n)} \\ &= \frac{p_{X,Y}(k, n-k)}{\sum_{j=0}^{n} p_{X,Y}(j, n-j)}. \end{aligned}$$

At this point we use the fact that the variables are independent and their distributions are known:

$$\begin{aligned} p_{X|X+Y}(k|n) &= \frac{e^{-\lambda_1}\frac{\lambda_1^k}{k!}e^{-\lambda_2}\frac{\lambda_2^{n-k}}{(n-k)!}}{\sum_{j=0}^{n} e^{-\lambda_1}\frac{\lambda_1^j}{j!}e^{-\lambda_2}\frac{\lambda_2^{n-j}}{(n-j)!}} \\ &= \frac{\binom{n}{k}\lambda_1^k\lambda_2^{n-k}}{\sum_{j=0}^{n}\binom{n}{j}\lambda_1^j\lambda_2^{n-j}} \\ &= \frac{\binom{n}{k}\lambda_1^k\lambda_2^{n-k}}{(\lambda_1+\lambda_2)^n}. \end{aligned}$$

Thus, it is a binomial distribution with parameters n and $\lambda_1/(\lambda_1 + \lambda_2)$, which me may write as

$$[X \text{ conditional on } X + Y = n] \sim \mathscr{B}\left(n, \frac{\lambda_1}{\lambda_1 + \lambda_2}\right).$$

▲▲▲

Conditional probabilities can be generalized to multiple variables. For example,

$$\begin{aligned} p_{X,Y|Z}(x, y|z) &:= P(X = x, Y = y|Z = z) = \frac{p_{X,Y,Z}(x, y, z)}{p_Z(z)} \\ p_{X|Y,Z}(x|y, z) &:= P(X = x|Y = y, Z = z) = \frac{p_{X,Y,Z}(x, y, z)}{p_{Y,Z}(y, z)}, \end{aligned}$$

and so on.

Proposition 3.5 *Every three random variables X, Y, Z satisfy*

$$p_{X,Y,Z}(x, y, z) = p_{X|Y,Z}(x|y, z)p_{Y|Z}(y|z)p_Z(z).$$

Proof: Immediate. Just follow the definitions. ■

Example: Consider a sequence of random variables $(X_k)_{k=0}^n$, each assuming values in a finite alphabet $\mathscr{A} = \{1, \dots, s\}$. Their joint distribution can be expressed as follows:

$$p(x_0, x_1, \dots, x_n) = p(x_n|x_0, \dots, x_{n-1})p(x_{n-1}|x_0, \dots, x_{n-2}) \dots p(x_1|x_0)p(x_0),$$

where we have omitted the subscripts to simplify notations. There exists a class of such sequences called ***Markov chains***. In a Markov chain,

$$p(x_n|x_0, \dots, x_{n-1}) = p(x_n|x_{n-1}),$$

i.e., the distribution of X_n "depends on its history only through its predecessor"; if X_{n-1} is known, then the knowledge of its predecessors is superfluous for the sake of predicting X_n. Note that this does not mean that X_n is independent of X_{n-2}! Moreover, the function $p(x_k|x_{k-1})$ is the same for all k, i.e., it can be represented by an s-by-s matrix, M.

Thus, for a Markov chain,

$$\begin{aligned} p(x_0, x_1, \dots, x_n) &= p(x_n|x_{n-1})p(x_{n-1}|x_{n-2}) \dots p(x_1|x_0)p(x_0) \;. \\ &= M_{x_n,x_{n-1}} M_{x_{n-1},x_{n-2}} \dots M_{x_1,x_0} p(x_0). \end{aligned}$$

If we now sum over all values that X_0 through X_{n-1} can assume, then

$$p(x_n) = \sum_{x_{n-1}\in\mathscr{A}} \cdots \sum_{x_0\in\mathscr{A}} M_{x_n,x_{n-1}} M_{x_{n-1},x_{n-2}} \dots M_{x_1,x_0} p(x_0) = \sum_{x_0\in\mathscr{A}} M^n_{x_n,x_0} p(x_0).$$

Thus, the distribution on X_n is related to the distribution of X_0 through the application of the n-power of a matrix (the ***transition matrix***). Situations of interest are when the distribution of X_n tends to a limit, which does not depend on the initial distribution of X_0. Such Markov chains are said to be ***ergodic***. When the rate of approach to this limit is exponential, the Markov chain is said to be ***exponentially mixing***. The study of such systems has many applications, which are unfortunately beyond the scope of this course. ▲▲▲

Chapter 4

Expectation

שקיעה ורודה על סף הרחוב
ורחוב כמנהרה של תכלת
מי שיגיע עד הסוף
ירצה לבכות מרב תוחלת

4.1 Basic definitions

Definition 4.1 Let X be a real-valued random variable over a discrete *probability space $(\Omega, \mathscr{F}, P)$. We denote the atomistic probability by $p(\omega) = P(\{\omega\})$. The* ***expectation*** *or* ***expected value*** *of X is a real number denoted by $\mathbb{E}[X]$, and defined by*

$$\mathbb{E}[X] := \sum_{\omega \in \Omega} X(\omega)\, p(\omega).$$

It is an ***average*** *over $X(\omega)$,* ***weighted*** *by the probability $p(\omega)$.*

Comment: The expected value is only defined for random variables for which the sum converges *absolutely*. In the context of measure theory, the expectation is the ***integral*** of X over the measure space $(\Omega, \mathscr{F}, P)$.

The expected value of X can be rewritten in terms of the distribution of X:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{\omega\in\Omega} X(\omega)\, p(\omega) \\
&= \sum_{x\in S_X} \sum_{\omega\in X^{-1}(x)} X(\omega)\, p(\omega) \\
&= \sum_{x\in S_X} x \sum_{\omega\in X^{-1}(x)} p(\omega) \\
&= \sum_{x\in S_X} x\, p_X(x).
\end{aligned}$$

Thus, $\mathbb{E}[X]$ is the expected value of the identity function, $X(x) = x$, with respect to the probability space $(S_X, \mathscr{F}_X, P_X)$.

Example: Let X be the outcome of a tossed die, what is the expected value of X?

In this case $S = \{1, \dots, 6\}$ and $p_X(k) = \frac{1}{6}$, thus

$$\mathbb{E}[X] = \sum_{k=1}^{6} k \cdot \frac{1}{6} = \frac{21}{6}.$$

▲▲▲

Example: What is the expected value of X, which is a Bernoulli variable with $p_X(1) = p$? Answer: p. ▲▲▲

Example: What is the expected value of $X \sim \mathscr{B}(n, p)$?

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} \frac{n!}{(n-k)!(k-1)!} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n-1} \frac{n!}{(n-k-1)!k!} p^{k+1} (1-p)^{n-k-1} \\
&= np \sum_{k=0}^{n-1} \frac{(n-1)!}{(n-k-1)!k!} p^k (1-p)^{n-k-1} \\
&= np.
\end{aligned}$$

▲▲▲

Example: What is the expected value of $X \sim \mathcal{P}oi(\lambda)$?

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{k!} = \lambda.$$

▲▲▲

Example: What is the expected value of $X \sim \mathcal{G}eo(p)$?

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k q^{k-1} p = p \frac{d}{dq} \sum_{k=1}^{\infty} q^k = p \frac{d}{dq} \left(\frac{q}{1-q} \right) = \frac{p}{(1-q)^2} = \frac{1}{p}.$$

If you don't like this method, you can obtain the same result by noting that

$$\mathbb{E}[X] = p + \sum_{k=2}^{\infty} k q^{k-1} p = p + \sum_{k=1}^{\infty} (k+1) q^k p = p + \sum_{k=1}^{\infty} q^k p + q\,\mathbb{E}[X],$$

i.e.,

$$p\,\mathbb{E}[X] = p + \frac{pq}{1-q} = 1.$$

▲▲▲

✎ *Exercise 4.1* What is the expected value of the number of times one has to toss a die until getting a "3"? What is the expected value of the number of times one has to toss two dice until getting a Shesh-Besh $(6,5)$?

Example: Let X be a random variable assuming integer values and having an atomistic distribution of the form $p_X(k) = a/k^2$, where a is a constant. What is a? What is the expected value of X? ▲▲▲

What is the intuitive (until we prove some theorems) meaning of the expected value? Suppose we repeat the same experiment many times and thus obtain a sequence (X_k) of random variables that are mutually independent and have the same distribution. Consider then the statistical average

$$Y = \frac{1}{n} \sum_{k=1}^{n} X_k = \sum_{a \in S} a \frac{\text{number of times the outcome was } a}{n}.$$

As n goes to infinity, this ratio tends to $p_X(a)$, and Y tends to $\mathbb{E}[X]$. This heuristic argument lacks rigor (e.g., does it hold when S is an infinite set?), but should give more insight into the definition of the expected value.

4.2 The expected value of a function of a random variable

Example: Consider a random variable X assuming the values $\{0, 1, 2\}$ and having a distribution

x	0	1	2
$p_X(x)$	1/2	1/3	1/6

What is the expected value of the random variable X^2?

We need to follow the definition, construct the distribution p_Y of the random variable $Y = X^2$, and then

$$\mathbb{E}[X^2] = \mathbb{E}[Y] = \sum_y y\, p_Y(y).$$

This is easy to do, because the distribution of Y is readily inferred from the distribution of X,

y	0	1	4
$p_Y(y)$	1/2	1/3	1/6

thus

$$\mathbb{E}[Y] = \frac{1}{2} \cdot 0 + \frac{1}{3} \cdot 1 + \frac{1}{6} \cdot 4 = 1.$$

Note then that the arithmetic operation we do is equivalent to

$$\mathbb{E}[X^2] = \sum_x x^2\, p_X(x).$$

The question is whether it is generally true that for any function $g : \mathbb{R} \to \mathbb{R}$,

$$E[g(x)] = \sum_x g(x)\, p_X(x).$$

While this may seem intuitive, note that by definition,

$$E[g(x)] = \sum_y y\, p_{g(X)}(y).$$

▲▲▲

Theorem 4.1 (The unconscious statistician) Let X be a random variable with range S_X and atomistic distribution p_X. Then, for any real valued function g,

$$\mathbb{E}[g(X)] = \sum_{x \in S_X} g(x)\, p_X(x),$$

provided that the right-hand side is finite.

Proof: Let $Y = g(X)$ and set $S_Y = g(S_X)$ be the range set of Y. We need to calculate $\mathbb{E}[Y]$, therefore we need to express the atomistic distribution of Y. Let $y \in S_Y$, then

$$\begin{aligned} p_Y(y) &= P(\{\omega : Y(\omega) = y\}) = P(\{\omega : g(X(\omega)) = y\}) \\ &= P\left(\{\omega : X(\omega) \in g^{-1}(y)\}\right), \end{aligned}$$

where $g^{-1}(y)$ may be a subset of S_X if g is not one-to-one. Thus,

$$p_Y(y) = \sum_{S_X \ni x \in g^{-1}(y)} p_X(x).$$

The expected value of Y is then obtained by

$$\begin{aligned} \mathbb{E}[Y] &= \sum_{y \in S_Y} y\, p_Y(y) = \sum_{y \in S_Y} y \sum_{S_X \ni x \in g^{-1}(y)} p_X(x) \\ &= \sum_{y \in S_Y} \sum_{S_X \ni x \in g^{-1}(y)} g(x) p_X(x) = \sum_{x \in S_X} g(x) p_X(x). \end{aligned}$$

■

Comment: In a sense, this theorem is trivial. Had we followed the original definition of the expected value, we would have had,

$$\begin{aligned}\mathbb{E}[g(X)] &= \sum_{\omega\in\Omega} g(X(\omega))\, p(\omega) \\ &= \sum_{x\in S_X} \sum_{\omega\in X^{-1}(x)} g(X(\omega))\, p(\omega) \\ &= \sum_{x\in S_X} g(x) \sum_{\omega\in X^{-1}(x)} p(\omega) \\ &= \sum_{x\in S_X} g(x)\, p_X(x).\end{aligned}$$

(This is the way I will prove it next time I teach this course...)

✎ *Exercise 4.2* Let X be a random variable and f, g be two real valued functions. Prove that

$$\mathbb{E}[f(X)g(X)] \le \left(\mathbb{E}[f^2(X)]\right)^{1/2} \left(\mathbb{E}[g^2(X)]\right)^{1/2}.$$

Hint: use the Cauchy inequality.

Example: The soccer club of Maccabbi Tel-Aviv plans to sell jerseys carrying the name of their star Eyal Berkovic. They must place their order at the beginning of the year. For every sold jersey they gain b Sheqels, but for every jersey that remains unsold they lose ℓ Sheqels. Suppose that the demand is a random variable with atomistic distribution $p(j)$, $j = 0, 1, \dots$. How many jerseys they need to order to maximize their expected profit?

Let a be the number of jerseys ordered by the club, and X be the demand. The net profit is then

$$g(X) = \begin{cases} Xb - (a - X)\ell & X = 0, 1, \dots, a \\ ab & X > a \end{cases}$$

The expected gain is deduced using the law of the unconscious statistician,

$$\begin{aligned}
\mathbb{E}[g(X)] &= \sum_{j=1}^{a} [jb - (a-j)\ell]\, p(j) + \sum_{j=a+1}^{\infty} abp(j) \\
&= -a\ell \sum_{j=0}^{a} p(j) + ab \sum_{j=a+1}^{\infty} p(j) + (b+\ell) \sum_{j=0}^{a} jp(j) \\
&= ab \sum_{j=0}^{\infty} p(j) + (b+\ell) \sum_{j=0}^{a} (j-a)p(j) \\
&= ab + (b+\ell) \sum_{j=0}^{a} (j-a)p(j) =: G(a).
\end{aligned}$$

We need to maximize this expression with respect to a. The simplest way to do it is to check what happens when we go from a to $a+1$:

$$G(a+1) = G(a) + b - (b+\ell) \sum_{j=0}^{a} p(j).$$

That is, it is profitable to increase a as long as

$$P(X \le a) = \sum_{j=0}^{a} p(j) < \frac{b}{b+\ell}.$$

▲▲▲

Comment: Consider a probability space $(\Omega, \mathscr{F}, P)$, and let $a \in \mathbb{R}$ be a constant. Then $\mathbb{E}[a] = a$. To justify this identity, we consider a to be a constant random variable $X(\omega) = a$. Then,

$$p_X(a) = P(\{\omega : X(\omega) = a\}) = P(\Omega) = 1,$$

and the identity follows.

The calculation of the expected value of a function of a random variable is easily generalized to multiple random variables. Consider a probability space $(\Omega, \mathscr{F}, P)$ on which two random variables X, Y are defined, and let $g : \mathbb{R}^2 \to \mathbb{R}$. The theorem of the unconscious statistician generalizes into

$$\mathbb{E}[g(X,Y)] = \sum_{x,y} g(x,y) p_{X,Y}(x,y).$$

✎ *Exercise 4.3* Prove it.

Corollary 4.1 The expectation is a linear functional in the vector space of random variables: if X, Y are random variables over a probability space $(\Omega, \mathscr{F}, P)$ and $a, b \in \mathbb{R}$, then

$$\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y].$$

Proof: By the theorem of the unconscious statistician,

$$\begin{aligned}\mathbb{E}[aX + bY] &= \sum_{x,y}(ax + by)p_{X,Y}(x, y)\\ &= a\sum_x x\sum_y p_{X,Y}(x, y) + b\sum_y y\sum_x p_{X,Y}(x, y)\\ &= a\,\mathbb{E}[X] + b\,\mathbb{E}[Y].\end{aligned}$$

■

This simple fact will be used extensively later on.

4.3 Moments

Definition 4.2 Let X be a random variable over a probability space. The ***n-th moment*** *of X is defined by*

$$M_n[X] := \mathbb{E}[X^n].$$

If we denote the expected value of X by μ, then the ***n-th central moment*** *of X is defined by*

$$C_n[X] := \mathbb{E}[(X - \mu)^n].$$

The second central moment of a random variable is called its ***variance****, and it is denoted by*

$$\begin{aligned}\mathrm{Var}[X] &:= \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2]\\ &= \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.\end{aligned}$$

(We have used the linearity of the expectation.) The square root of the variance is called the ***standard deviation****, and is denoted by*

$$\sigma_X = \sqrt{\mathrm{Var}[X]}.$$

The standard deviation is a measure of the (absolute) distance of the random variable from its expected value (or mean). It provides a measure of how “spread” the distribution of X is.

Proposition 4.1 If $\text{Var}[X] = 0$*, then* $X(\omega) = \mathbb{E}[X]$ *with probability one.*

Proof: Let $\mu = \mathbb{E}[X]$. By definition,

$$\text{Var}[X] = \sum_x (x - \mu)^2 p_X(x).$$

This is a sum of non-negative terms. It can only be zero if $p_X(\mu) = 1$. ■

Example: The second moment of a random variable $X \sim \mathscr{B}(n, p)$ is calculated as follows:

$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{k=0}^{n} k(k-1)\binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=2}^{n} \frac{n!}{(n-k)!(k-2)!} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n-2} \frac{n!}{(n-k-2)!k!} p^{k+2} (1-p)^{n-k-2} \\ &= n(n-1)p^2.\end{aligned}$$

Therefore,

$$\mathbb{E}[X^2] = n(n-1)p^2 + \mathbb{E}[X] = n(n-1)p^2 + np.$$

The variance of X is

$$\text{Var}[X] = n(n-1)p^2 + np - (np)^2 = np(1-p).$$

▲▲▲

Example: What is the variance of a Poisson variable $X \sim \mathcal{P}oi(\lambda)$? Recalling that a Poisson variable is the limit of a binomial variable with $n \to \infty$, $p \to 0$, and $np = \lambda$, we deduce that $\text{Var}[X] = \lambda$. ▲▲▲

✎ *Exercise 4.4* Calculate the variance of a Poisson variable directly, without using the limit of a binomial variable.

Proposition 4.2 *For any random variable,*

$$\operatorname{Var}[aX + b] = a^2 \operatorname{Var}[X].$$

Proof:

$$\operatorname{Var}[aX + b] = \mathbb{E}[(aX + b - \mathbb{E}[aX + b])^2] = \mathbb{E}[a^2(X - \mathbb{E}[X])^2] = a^2 \operatorname{Var}[X].$$

■

Definition 4.3 *Let X, Y be a pair of random variables. Their* ***covariance*** *is defined by*

$$\operatorname{Cov}[X, Y] := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Two random variables whose covariance vanishes are said to be ***uncorrelated****. The* ***correlation coefficient*** *of a pair of random variables is defined by*

$$\rho[X, Y] = \frac{\operatorname{Cov}[X, Y]}{\sigma_X \sigma_Y}.$$

The covariance of two variables is a measure of their tendency to be larger than their expected value together. A negative covariance means that when one of the variables is larger than its mean, the other is more likely to be less than its mean.

✎ *Exercise 4.5* Prove that the correlation coefficient of a pair of random variables assumes values between -1 and 1 (Hint: use the Cauchy-Schwarz inequality).

Proposition 4.3 *If X, Y are independent random variables, and g, h are real valued functions, then*

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\,\mathbb{E}[h(Y)].$$

Proof: One only needs to apply the law of the unconscious statistician and use the fact that the joint distribution is the product of the marginal distributions,

$$\begin{aligned}\mathbb{E}[g(X)h(Y)] &= \sum_{x,y} g(x)h(y)p_X(x)p_Y(y) \\ &= \left(\sum_x g(x)p_X(x)\right)\left(\sum_y h(y)p_Y(y)\right) \\ &= \mathbb{E}[g(X)]\,\mathbb{E}[h(Y)].\end{aligned}$$

■

Corollary 4.2 If X, Y are independent then they are uncorrelated.

Proof: Obvious. ■

Is the opposite statement true? Are uncorrelated random variables necessarily independent? Consider the following joint distribution:

X/Y	-1	0	1
0	1/3	0	1/3
1	0	1/3	0

X and Y are not independent, because, for example, knowing that $X = 1$ implies that $Y = 0$. On the other hand,

$$\text{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - \frac{1}{3} \cdot 0 = 0.$$

That is, zero correlation does not imply independence.

Proposition 4.4 For any two random variables X, Y,

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\,\text{Cov}[X, Y].$$

Proof: Just do it! ■

✎ *Exercise 4.6* Show that for any collection of random variables $X_1, \dots, X_n$,

$$\operatorname{Var}\left[\sum_{k=1}^{n} X_k\right] = \sum_{k=1}^{n} \operatorname{Var}[X_k] + 2\sum_{i<j} \operatorname{Cov}[X_i, X_j].$$

4.4 Using the linearity of the expectation

In this section we will examine a number of examples that make use of the additive property of the expectation.

Example: Recall that we calculated the expected value of a binomial variable, $X \sim \mathscr{B}(n, p)$, and that we obtained $\mathbb{E}[X] = np$. There is an easy way to obtain this result. A binomial variable can be represented as a sum of independent Bernoulli variables,

$$X = \sum_{k=1}^{n} X_k, \qquad X_k\text{'s Bernoulli with success probability } p.$$

By the additivity of the expectation,

$$\mathbb{E}[X] = \sum_{k=1}^{n} \mathbb{E}[X_k] = n \cdot p.$$

The variance can be obtained by the same method. We have

$$\operatorname{Var}[X] = n \times \operatorname{Var}[X_1],$$

and it remains to verify that

$$\operatorname{Var}[X_1] = p(1-p)^2 + (1-p)(0-p)^2 = (1-p)(p(1-p) + p^2) = p(1-p).$$

Note that the calculation of the expected value does not use the independence property, whereas the calculation of the variance does. ▲▲▲

Example: A hundred dice are tossed. What is the expected value of their sum X? Let X_k be the outcome of the k-th die. Since $\mathbb{E}[X_k] = 21/6 = 7/2$, we have by additivity, $\mathbb{E}[X] = 100 \times 7/2 = 350$. ▲▲▲

Example: Consider again the problem of the inattentive secretary who puts n letters randomly into n envelopes. What is the expected number of letters that reach their destination?

Define for $k = 1, \dots, n$ the Bernoulli variables,

$$X_k = \begin{cases} 1 & \text{the } k\text{-th letter reached its destination} \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, $p_{X_k}(1) = 1/n$. If X is the number of letters that reached their destination, then $X = \sum_{k=1}^n X_k$, and by additivity,

$$\mathbb{E}[X] = n \times \frac{1}{n} = 1.$$

We then proceed to calculate the variance. We've already seen that for the Bernoulli variable with parameter p the variance equals $p(1-p)$. In this case, the X_k are not independent therefore we need to calculate their covariance. The variable X_1X_2 is also a Bernoulli variable, with parameter $1/n(n-1)$, so that

$$\operatorname{Cov}[X_1, X_2] = \frac{1}{n(n-1)} - \frac{1}{n^2}.$$

Putting things together,

$$\begin{aligned} \operatorname{Var}[X] &= n \times \frac{1}{n}\left(1 - \frac{1}{n}\right) + 2\binom{n}{2} \times \left(\frac{1}{n(n-1)} - \frac{1}{n^2}\right) \\ &= \left(1 - \frac{1}{n}\right) + n(n-1)\left(\frac{1}{n(n-1)} - \frac{1}{n^2}\right) = 1. \end{aligned}$$

Should we be surprised? Recall that X tends as $n \to \infty$ to a Poisson variable with parameter $\lambda = 1$, so that we expect that in this limit $\mathbb{E}[X] = \operatorname{Var}[X] = 1$. It turns out that this result holds exactly for every finite n. ▲▲▲

✎ *Exercise 4.7* In an urn are N white balls and M black balls. n balls are drawn randomly. What is the expected value of the number of white balls that were drawn? (Solve this problem by using the additivity of the expectation.)

Example: Consider a randomized deck of $2n$ cards, two of type "1", two of type "2", and so on. m cards are randomly drawn. What is the expected value of the number of pairs that will remain intact? (This problem was solved by Daniel

Bernoulli in the context of the number of married couples remaining intact after m deaths.)

We define X_k to be a Bernoulli variable taking the value 1 if the k-th couple remains intact. We have

$$\mathbb{E}[X_k] = p_{X_k}(1) = \frac{\binom{2n-2}{m}}{\binom{2n}{m}} = \frac{(2n-m)(2n-m-1)}{2n(2n-1)}.$$

The desired result is n times this number. ▲▲▲

Example: Recall the coupon collector: there are n different coupons, and each turn there is an equal probability to obtain any coupon. What is the expected value of the number of coupons that need to be collected before obtaining a complete set?

Let X be the number of coupons that need to be collected and X_k be the number of coupons that need to be collected from the moment that we had k different coupons to the moment we have $k+1$ different coupons. Clearly,

$$X = \sum_{k=0}^{n-1} X_k.$$

Now, suppose we have k different coupons. Every new coupon can be viewed as a Bernoulli experiment with success probability $(n-k)/n$. Thus, X_k is a geometric variable with parameter $(n-k)/n$, and $\mathbb{E}[X_k] = n/(n-k)$. Summing up,

$$\mathbb{E}[X] = \sum_{k=0}^{n-1} \frac{n}{n-k} = n \sum_{k=1}^{n} \frac{1}{k} \approx n \log n.$$

▲▲▲

✎ *Exercise 4.8* Let $X_1, \dots, X_n$ be a sequence of independent random variables that have the same distribution. We denote $\mathbb{E}[X_1] = \mu$ and $\text{Var}[X_1] = \sigma^2$. Find the expected value and the variance of the empirical mean

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

We conclude this section with a remark about infinite sums. First a simple lemma:

Lemma 4.1 *Let X be a random variable. If $X(\omega) \ge a$ (with probability one) then $\mathbb{E}[X] \ge a$. Also,*

$$\mathbb{E}[|X|] \ge |\mathbb{E}[X]|.$$

Proof: The first result follows from the definition of the expectation. The second result follows from the inequality

$$\left| \sum_x x\, p_X(x) \right| \le \sum_x |x|\, p_X(x).$$

■

Theorem 4.2 *Let (X_n) be an infinite sequence of random variables such that*

$$\sum_{n=1}^{\infty} \mathbb{E}[|X_n|] < \infty.$$

Then,

$$\mathbb{E}\left[\sum_{n=1}^{\infty} X_n \right] = \sum_{n=1}^{\infty} \mathbb{E}[X_n].$$

Proof: TO BE COMPLETED. ■

The following is an application of the above theorem. Let X be a random variable assuming positive integer values and having a finite expectation. Define for every natural i,

$$X_i(\omega) = \begin{cases} 1 & i \le X(\omega) \\ 0 & \text{otherwise} \end{cases}$$

Then,

$$\sum_{i=1}^{\infty} X_i(\omega) = \sum_{i \le X(\omega)} 1 = X(\omega).$$

Now,

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{E}[X_i] = \sum_{i=1}^{\infty} P(\{\omega : X(\omega) \geq i\}).$$

4.5 Conditional expectation

Definition 4.4 Let X, Y be random variables over a probability space $(\Omega, \mathscr{F}, P)$. The ***conditional expectation of** X **given that*** *$Y = y$ is defined by*

$$\mathbb{E}[X|Y = y] := \sum_{x} x \, p_{X|Y}(x|y).$$

Note that this definition makes sense because $p_{X|Y}(\cdot|y)$ is an atomistic distribution on S_X.

Example: Let $X, Y \sim \mathscr{B}(n, p)$ be independent. What is the conditional expectation of X given that $X + Y = m$?

To answer this question we need to calculate the conditional distribution $p_{X|X+Y}$. Now,

$$p_{X|X+Y}(k|m) = \frac{P(X = k, X + Y = m)}{P(X + Y = m)} = \frac{P(X = k, Y = m - k)}{P(X + Y = m)},$$

with $k \leq m, n$. We know what the numerator is. For the denominator, we realize that the sum of two binomial variables with parameters (n, p) is a binomial variable with parameters $(2n, p)$ (think of two independent sequences of Bernoulli trials added up). Thus,

$$p_{X|X+Y}(k|m) = \frac{\binom{n}{k} p^k (1-p)^{n-k} \binom{n}{m-k} p^{m-k} (1-p)^{n-m+k}}{\binom{2n}{m} p^m (1-p)^{2n-m}} = \frac{\binom{n}{k}\binom{n}{m-k}}{\binom{2n}{m}}.$$

The desired result is

$$\mathbb{E}[X|X + Y = m] = \sum_{k=0}^{\min(m,n)} k \frac{\binom{n}{k}\binom{n}{m-k}}{\binom{2n}{m}}.$$

It is not clear how to simplify this expression. A useful trick is to observe that $p_{X|X+Y}(k|m)$ with m fixed is the probability of obtaining k white balls when one draws m balls from an urn containing n white balls and n black balls. Since every ball is white with probability $1/2$, by the additivity of the expectation, the expected number of white balls is $m/2$.

Now that we know the result, we may see that we could have reached it much more easily. By symmetry,

$$\mathbb{E}[X|X+Y=m] = \mathbb{E}[Y|X+Y=m],$$

hence, by the linearity of the expectation,

$$\mathbb{E}[X|X+Y=m] = \frac{1}{2}\mathbb{E}[X+Y|X+Y=m] = \frac{m}{2}.$$

In particular, this result holds whatever is the distribution of X, Y (as long as it is the same). ▲▲▲

We now refine our definition of the conditional expectation:

Definition 4.5 Let $X(\omega), Y(\omega)$ be random variables over a probability space $(\Omega, \mathscr{F}, P)$. The ***conditional expectation of** X **given*** *Y is a random variable $Z(\omega)$, which is a composite function of $Y(\omega)$, i.e., $Z(\omega) = \varphi(Y(\omega))$, and*

$$\varphi(y) := \mathbb{E}[X|Y=y].$$

Another way to write it is:

$$\mathbb{E}[X|Y](\omega) := \mathbb{E}[X|Y=y]_{y=Y(\omega)}.$$

That is, having performed the experiment, we are given only $Y(\omega)$, and the random variable $\mathbb{E}[X|Y](\omega)$ is the expected value of $X(\omega)$ now that we know $Y(\omega)$.

Proposition 4.5 For every two random variables X, Y,

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X].$$

Proof: What does the proposition say? That

$$\sum_{\omega\in\Omega} \mathbb{E}[X|Y](\omega)\, p(\omega) = \sum_{\omega\in\Omega} X(\omega)\, p(\omega).$$

Since $\mathbb{E}[X|Y](\omega)$ is a composite function of $Y(\omega)$ we can use the law of the unconscious statistician to rewrite this as

$$\sum_y \mathbb{E}[X|Y=y]\, p_Y(y) = \sum_x x\, p_X(x).$$

Indeed,

$$\begin{aligned}\sum_y \mathbb{E}[X|Y=y]\, p_Y(y) &= \sum_y \sum_x x\, p_{X|Y}(x|y) p_Y(y) \\ &= \sum_y \sum_x x\, p_{X,Y}(x,y) = \mathbb{E}[X].\end{aligned}$$

■

This simple proposition is quite useful. It states that the expected value of X can be computed by averaging over its expectation conditioned over another variable.

Example: A miner is inside a mine, and doesn't know which of three possible tunnels will lead him out. If he takes tunnel A he will be out within 3 hours. If he takes tunnel B he will be back to the same spot after 5 hours. If he takes tunnel C he will be back to the same spot after 7 hours. He chooses the tunnel at random with equal probability for each tunnel. If he happens to return to the same spot, the poor thing is totally disoriented, and has to redraw his choice again with equal probabilities. What is the expected time until he finds the exit?

The sample space consists of infinite sequences of "BCACCBA...", with the standard probability of independent repeated trials. Let $X(\omega)$ be the exit time and $Y(\omega)$ be the label of the first door he chooses. By the above proposition,

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[\mathbb{E}[X|Y]] \\ &= \mathbb{E}[X|Y=A]\, p_Y(A) + \mathbb{E}[X|Y=B]\, p_Y(B) + \mathbb{E}[X|Y=C]\, p_Y(C) \\ &= \frac{1}{3}\left(3 + \mathbb{E}[X|Y=B] + \mathbb{E}[X|Y=C]\right).\end{aligned}$$

What is $\mathbb{E}[X|Y=B]$? If the miner chose tunnel B, then he wandered for 5 hours, and then faced again the original problem, independently of his first choice. Thus,

$$\mathbb{E}[X|Y=B] = 5 + \mathbb{E}[X] \qquad \text{and similarly} \qquad \mathbb{E}[X|Y=C] = 7 + \mathbb{E}[X].$$

Substituting, we get

$$\mathbb{E}[X] = 1 + \frac{1}{3}(5 + \mathbb{E}[X]) + \frac{1}{3}(7 + \mathbb{E}[X]).$$

This equation is easily solved, $\mathbb{E}[X] = 15$. ▲▲▲

Example: Consider a sequence of independent Bernoulli trials with success probability p. What is the expected number of trials until one obtains two 1's in a row?

Let $X(\omega)$ be the number of trials until two 1's in a row, and let $Y_j(\omega)$ be the outcome of the j-th trial. We start by writing

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y_1]] = p\,\mathbb{E}[X|Y_1 = 1] + (1-p)\,\mathbb{E}[X|Y_1 = 0].$$

By the same argument as above,

$$\mathbb{E}[X|Y_1 = 0] = 1 + \mathbb{E}[X].$$

Next, we use a simple generalization of the conditioning method,

$$\mathbb{E}[X|Y_1 = 1] = p\,\mathbb{E}[X|Y_1 = 1, Y_2 = 1] + (1-p)\,E[X|Y_1 = 1, Y_2 = 0].$$

Using the fact that

$$\mathbb{E}[X|Y_1 = 1, Y_2 = 1] = 2 \qquad \text{and} \qquad E[X|Y_1 = 1, Y_2 = 0] = 2 + \mathbb{E}[X],$$

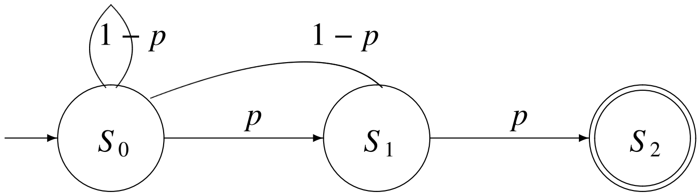
we finally obtain an implicit equation for $\mathbb{E}[X]$:

$$\mathbb{E}[X] = p\,\left[2p + (1-p)(2 + \mathbb{E}[X])\right] + (1-p)(1 + \mathbb{E}[X]),$$

from which we readily obtain

$$\mathbb{E}[X] = \frac{1+p}{p^2}.$$

We can solve this same problem differently. We view the problem in terms of a three-state space: one can be in the initial state (having to produce two 1's in a row), be in state after a single 1, or be in the terminal state after two 1's in a row. We label these states S_0, S_1, and S_2. Now every sequence of successes and failures implies a trajectory on the state space. That is, we can replace the original sample space of sequences of zero-ones by a sample space of sequences of states S_j. This defined a new compound experiment, with transition probabilities that can be represented as a graph:



Let now $X(\omega)$ be the number of steps until reaching state S_2. The expected value of X depends on the initial state. The graph suggests the following relations,

$$\begin{aligned}
\mathbb{E}[X|S_0] &= 1 + p\,\mathbb{E}[X|S_1] + (1-p)\,\mathbb{E}[X|S_0] \\
\mathbb{E}[X|S_1] &= 1 + p\,\mathbb{E}[X|S_2] + (1-p)\,\mathbb{E}[X|S_0] \\
\mathbb{E}[X|S_2] &= 0
\end{aligned}$$

It is easily checked that $\mathbb{E}[X|S_0] = (1+p)/p^2$. ▲▲▲

I AM NOT SATISFIED WITH THE WAY THIS IS EXPLAINED. REQUIRES ELABORATION.

✎ *Exercise 4.9* Consider a sequence of independent Bernoulli trials with success probability p. What is the expected number of trials until one obtains three 1's in a row? four 1's in a row?

✎ *Exercise 4.10* A monkey types randomly on a typing machine. Each character has a probability of $1/26$ of being each of the letters of the alphabet, independently of the other. What is the expected number of characters that the monkey will type until generating the string ”ABCD”? What about the string ”ABAB”?

The following paragraphs are provided for those who want to know more.

The conditional expectation $\mathbb{E}[X|Y](\omega)$ plays a very important role in probability theory. Its formal definition, which remains valid in the general case (i.e., uncountable spaces), is somewhat more involved than that presented in this section, but we do have all the necessary background to formulate it. Recall that a random variable $Y(\omega)$ generates a σ-algebra of events (a sub-σ-algebra of $\mathscr{F}$),

$$\mathscr{F} \supseteq \sigma(Y) := \left\{ Y^{-1}(A) : A \in \mathscr{F}_Y \right\}.$$

Let φ be a real valued function defined on S_Y, and define a random variable $Z(\omega) = \varphi(Y(\omega))$. The σ-algebra generated by Z is

$$\sigma(Z) := \left\{ Z^{-1}(B) : B \in \mathscr{F}_Z \right\} = \left\{ Y^{-1}(\varphi^{-1}(B)) : B \in \mathscr{F}_Z \right\} \subseteq \sigma(Y).$$

That is, the σ-algebra generated by a function of a random variable is contained in the σ-algebra generated by this random variable. In fact, it can be shown that the opposite is true. If Y, Z are random variables and $\sigma(Z) \subseteq \sigma(Y)$, then Z can be expressed as a composite function of Y.

Recall now our definition of the conditional expectation,

$$\mathbb{E}[X|Y](\omega) = \mathbb{E}[X|Y=y]_{y=Y(\omega)} = \sum_x x\, p_{X|Y}(x|Y(\omega)) = \sum_x x\, \frac{p_{X,Y}(x, Y(\omega))}{p_Y(Y(\omega))}.$$

Let $A \in \mathscr{F}$ be any event in $\sigma(Y)$, that is, there exists a $B \in \mathscr{F}_Y$ for which $Y^{-1}(B) = A$. Now,

$$\begin{aligned}
\sum_{\omega \in A} \mathbb{E}[X|Y](\omega)\, p(\omega) &= \sum_{\omega \in A} \sum_{x} x \, \frac{p_{X,Y}(x, Y(\omega))}{p_Y(Y(\omega))}\, p(\omega) \\
&= \sum_{x} x \sum_{\omega \in A} \frac{p_{X,Y}(x, Y(\omega))}{p_Y(Y(\omega))}\, p(\omega) \\
&= \sum_{x} x \sum_{y \in B} \sum_{\omega \in Y^{-1}(y)} \frac{p_{X,Y}(x, y)}{p_Y(y)}\, p(\omega) \\
&= \sum_{x} x \sum_{y \in B} \frac{p_{X,Y}(x, y)}{p_Y(y)} \sum_{\omega \in Y^{-1}(y)} p(\omega) \\
&= \sum_{x} x \sum_{y \in B} p_{X,Y}(x, y).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\sum_{\omega \in A} X(\omega)\, p(\omega) &= \sum_{y \in B} \sum_{\omega \in Y^{-1}(y)} X(\omega)\, p(\omega) \\
&= \sum_{x} \sum_{y \in B} \sum_{\{\omega : (X(\omega), Y(\omega)) = (x,y)\}} X(\omega)\, p(\omega) \\
&= \sum_{x} x \sum_{y \in B} \sum_{\{\omega : (X(\omega), Y(\omega)) = (x,y)\}} p(\omega) \\
&= \sum_{x} x \sum_{y \in B} p_{X,Y}(x, y).
\end{aligned}$$

That is, for every $A \in \sigma(Y)$,

$$\sum_{\omega \in A} \mathbb{E}[X|Y](\omega)\, p(\omega) = \sum_{\omega \in A} X(\omega)\, p(\omega).$$

This property is in fact the standard definition of the conditional expectation:

Definition 4.6 *Let X, Y be random variables over a probability space $(\Omega, \mathscr{F}, P)$. The conditional expectation of X given Y is a random variable Z satisfying the following two properties: (1) $\sigma(Z) \subseteq \sigma(Y)$, (2) For every $A \in \sigma(Y)$*

$$\sum_{\omega \in A} Z(\omega)\, p(\omega) = \sum_{\omega \in A} X(\omega)\, p(\omega).$$

It can be proved that there exists a unique random variable satisfying these properties.

4.6 The moment generating function

Definition 4.7 Let X be a discrete random variable. Its ***moment generating*** *function $M_X(t)$ is a real-valued function defined by*

$$M_X(t) := \mathbb{E}[e^{tX}] = \sum_x e^{tx} p_X(x).$$

Example: What is the moment generating function of a binomial variable $X \sim \mathscr{B}(n,p)$?

$$M_X(t) = \sum_{k=1}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} = (1-p+e^t p)^n.$$

▲▲▲

Example: What is the moment generating function of a Poisson variable $X \sim \mathcal{P}oi(\lambda)$?

$$M_X(t) = \sum_{k=0}^{\infty} e^{tk} e^{-\lambda} \frac{\lambda^k}{k!} = e^{\lambda(e^t-1)}.$$

▲▲▲

What are the uses of the moment generating function? Note that

$$\begin{aligned} M_X(0) &= 1 \\ M_X'(0) &= \mathbb{E}[X] \\ M_X''(0) &= \mathbb{E}[X^2], \end{aligned}$$

and in general, the k-th derivative evaluated at zero equals to the k-th moment.

Example: Verify that we get the correct moments for the binomial and Poisson variables ▲▲▲

Comment: The moment-generating function is the ***Laplace transform*** of the atomistic distribution. It has many uses, which are however beyond the scope of this course.

Chapter 5

Random Variables (Continuous Case)

So far, we have purposely limited our consideration to random variables whose ranges are countable, or discrete. The reason for that is that distributions on countable spaces are easily specified by means of the atomistic distribution. The construction of a distribution on an uncountable space is only done rigorously within the framework of measure theory. Here, we will only provide limited tools that will allow us to operate with such variables.

5.1 Basic definitions

Definition 5.1 Let $(\Omega, \mathscr{F}, P)$ *be a probability space. A real-valued function* $\Omega \to \mathbb{R}$ *is called a continuous random variable, if there exists a non-negative real-valued integrable function* $f_X(x)$*, such that*

$$P(\{\omega : X(\omega) \leq a\}) = F_X(a) = \int_{-\infty}^{a} f_X(x)\, dx.$$

The function f_X *is called the* ***probability density function*** *(PDF) of* X*.*

Comment: Recall that a random variable has a σ-algebra of events $\mathscr{F}_X$ associated with its range (here $\mathbb{R}$), and we need $X^{-1}(A) \in \mathscr{F}$ for all $A \in \mathscr{F}_X$. What is a suitable σ-algebra for $\mathbb{R}$? These are precisely the issues that we sweep under the carpet (well, I can still tell you that it is the σ-algebra generated by all open subsets of the line).

Thus, a continuous random variable is defined by its PDF. Since a random variable is by definition defined by its distribution P_X, we need to show that the PDF defines the distribution uniquely. Since we don't really know how to define distributions when we don't even know the set of events, this cannot really be achieved. Yet, we can at least show the following properties:

(1) For every segment $(a,b]$,

$$P_X((a,b]) = F_X(b) - F_X(a) = \int_a^b f_X(x)\,dx.$$

(2) For every A expressible as a countable union of disjoint $(a_j,b_j]$,

$$P_X(A) = \int_A f_X(x)\,dx.$$

(3) Normalization,

$$\int_{\mathbb{R}} f_X(x)\,dx = 1.$$

(4) The distribution of closed sets follows from the continuity of the probability,

$$P_X([a,b]) = P\left(\lim_{n\to\infty}(a-\tfrac{1}{n},b]\right) = \lim_{n\to\infty}\int_{a-1/n}^b f_X(x)\,dx = P_X((a,b]).$$

(5) As a result, $P_X(\{a\}) = 0$ for every $a\in\mathbb{R}$, as

$$P(\{a\}) = P_X([a,b]) - P_X((a,b]).$$

Comment: We may consider discrete random variables as having a PDF which is a sum of δ-functions.

Example: The random variable X has a PDF of the form

$$f_X(x) = \begin{cases} 2C(2x-x^2) & 0\le x\le 2 \\ 0 & \text{otherwise}\end{cases}.$$

What is the value of the constant C and what is the probability that $X(\omega) > 1$?

The constant is obtained by normalization,

$$1 = 2C \int_0^2 (2x - x^2)\, dx = 2C \left(4 - \frac{8}{3} \right) = \frac{8C}{3}.$$

Then,

$$P(X > 1) = 2C \int_1^2 (2x - x^2)\, dx = \frac{1}{2}.$$

▲▲▲

5.2 The uniform distribution

Definition 5.2 A random variable X is called ***uniformly distributed in*** $[a, b]$, *denoted $X \sim \mathcal{U}(a, b)$, if its* PDF *is given by*

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & otherwise \end{cases}.$$

Example: Buses are arriving at a station every 15 minutes. A person arrives at the station at a random time, uniformly distributed between 7:00 and 7:30. What is the probability that he has to wait less than 5 minutes?

Let $X(\omega)$ be the arrival time (in minutes past 7:00), and $Y(\omega)$ the time he has to wait. We know that $X \sim \mathcal{U}(0, 30)$. Now,

$$P(Y < 5) = P(\{X = 0\} \cup \{10 \le X < 15\} \cup \{25 \le X < 30\}) = 0 + \frac{5}{30} + \frac{5}{30} = \frac{1}{3}.$$

▲▲▲

Example: ***Bertrand's paradox***: consider a random chord of a circle. What is the probability that the chord is longer than the side of an equilateral triangle inscribed in that circle?

The "paradox" stems from the fact that the answer depends on the way the random chord is selected. One possibility is to take the distance of the chord from the center of the circle r to be $\mathcal{U}(0, R)$. Since the chord is longer than the side of the equilateral triangle when $r < R/2$, the answer is $1/2$. A second possibility is to take the angle θ between the chord and the tangent to the circle to be $\mathcal{U}(0, \pi)$. The chord is longer than the side of the triangle when $\pi/3 < \theta < 2\pi/3$, in which case the answer is $1/3$.

▲▲▲

5.3 The normal distribution

Definition 5.3 A random variable X is said to be **normally distributed** *with parameters μ, σ^2, denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$, if its* PDF *is*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

X is called a **standard normal variable** *if $\mu = 0$ and $\sigma^2 = 1$.*

✎ *Exercise 5.1* Show that the PDF of a normal variable $\mathcal{N}(\mu, \sigma^2)$ is normalized.

Proposition 5.1 Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and set $Y = aX + b$, where $a > 0$. Then

$$Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2).$$

Proof: At this point, where we don't know how to change variables, we simply operate on the distribution function of Y,

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(X \le a^{-1}(y-b)) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{a^{-1}(y-b)} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\ &= \frac{1/a}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{y} \exp\left(-\frac{(a^{-1}(u-b)-\mu)^2}{2\sigma^2}\right) du \\ &= \frac{1}{\sqrt{2\pi a^2\sigma^2}} \int_{-\infty}^{y} \exp\left(-\frac{(u-b-a\mu)^2}{2a^2\sigma^2}\right) du, \end{aligned}$$

where we have changed variables, $x = a^{-1}(u-b)$. ■

Corollary 5.1 If $X \sim \mathcal{N}(\mu, \sigma^2)$ then $(X-\mu)/\sigma$ is a standard normal variable.

Notation: The distribution function of a standard normal variable will be denoted by $\Phi(x)$,

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

(This function is closely related to ***Gauss' error function***).

The importance of the normal distribution stems from the ***central limit theorem***, which we will encounter later on. The following theorem is an instance of the central limit theorem for a particular case:

Theorem 5.1 (DeMoivre-Laplace) Let (X_n) be a sequence of independent Bernoulli variables with parameter p and set

$$Y_n := \frac{X_n - p}{\sqrt{p(1-p)}}.$$

(The variables Y_n have zero expectation and unit variance.) Set then

$$S_n := \frac{1}{\sqrt{n}} \sum_{k=1}^{n} Y_k.$$

Then S_n tends, as $n \to \infty$, to a standard normal variable in the sense that

$$\lim_{n\to\infty} P(a \le S_n \le b) = \Phi(b) - \Phi(a).$$

Comment: This theorem states that the sequence of random variables S_n converges to a standard normal variable ***in distribution***, or ***in law***.

Proof: The event $\{a \le S_n \le b\}$ can be written as

$$\left\{ np + \sqrt{np(1-p)}\, a \le \sum_{k=1}^{n} X_k \le np + \sqrt{np(1-p)}\, b \right\},$$

and the sum over X_k is a binomial variable $\mathscr{B}(n,p)$. Thus, the probability of this event is

$$P(a \le S_n \le b) = \sum_{k=np+\sqrt{np(1-p)}\,a}^{np+\sqrt{np(1-p)}\,b} \binom{n}{k} p^k (1-p)^{n-k}.$$

(We will ignore the fact that limits are integer as the correction is negligible as $n \to \infty$.) As n becomes large (while p remains fixed), n, k, and $n-k$ become large, hence we use Stirling's approximation,

$$\binom{n}{k} \sim \frac{\sqrt{2\pi n}\, n^n e^{-n}}{\sqrt{2\pi k}\, k^k e^{-k}\ \sqrt{2\pi(n-k)}\,(n-k)^{n-k} e^{-(n-k)}},$$

and so

$$\binom{n}{k} p^k (1-p)^{n-k} \sim \sqrt{\frac{n}{2\pi k(n-k)}} \left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k},$$

where, as usual, the $\sim$ relation means that the ratio between the two sides tends to one as $n \to \infty$. The summation variable k takes values that are of order $O(\sqrt{n})$ around np. This suggests a change of variables, $k = np + \sqrt{np(1-p)}\,m$, where m varies from a to b in units of $\Delta m = [np(1-p)]^{-1/2}$.

Consider the first term in the above product. As $n \to \infty$,

$$\lim_{n\to\infty} \frac{1}{\Delta m} \sqrt{\frac{n}{2\pi k(n-k)}} = \lim_{n\to\infty} \frac{\sqrt{n}}{\sqrt{2\pi n}} \frac{\sqrt{p(1-p)}}{\sqrt{(k/n)(1-k/n)}} = \frac{1}{\sqrt{2\pi}}.$$

Consider the second term, which we can rewrite as

$$\left(\frac{np}{k}\right)^k = \left(\frac{np}{np + r\sqrt{n}}\right)^{np+r\sqrt{n}},$$

where $r = \sqrt{p(1-p)}\,m$. To evaluate the $n \to \infty$ limit it is easier to look at the logarithm of this expression, whose limit we evaluate using Taylor's expansion,

$$\begin{aligned}
\log\left(\frac{np}{k}\right)^k &= (np + r\sqrt{n}) \log\left(1 + \frac{r}{p} n^{-1/2}\right)^{-1} \\
&= (np + r\sqrt{n}) \log\left(1 - \frac{r}{p} n^{-1/2} + \frac{r^2}{p^2} n^{-1}\right) + \text{l.o.t} \\
&= (np + r\sqrt{n}) \left(-\frac{r}{p} n^{-1/2} + \frac{r^2}{2p^2} n^{-1}\right) + \text{l.o.t} \\
&= -r\sqrt{n} - \frac{r^2}{2p} = -r\sqrt{n} - \frac{1}{2}(1-p)m^2 + \text{l.o.t.}
\end{aligned}$$

Similarly,

$$\log\left(\frac{n(1-p)}{n-k}\right)^{n-k} = r\sqrt{n} - \frac{1}{2}pm^2 + \text{l.o.t.}$$

Combining the two together we have

$$\lim_{n\to\infty} \log\left[\left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k}\right] = -\frac{1}{2}m^2.$$

Thus, as $n \to \infty$ we have

$$\lim_{n\to\infty} P(a \le S_n \le b) = \lim_{n\to\infty} \frac{1}{\sqrt{2\pi}} \sum_{m=a}^{b} e^{-m^2/2}\,\Delta m = \Phi(b) - \Phi(a),$$

which concludes the proof. ■

A table of the values of $\Phi(x)$ is all that is needed to compute probabilities for general normal variables. Indeed, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$F_X(x) = P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Example: The duration of a normal pregnancy (in days) is a normal variable $\mathcal{N}(270, 100)$. A sailor's wife gave birth to a baby. It turns out that her husband was on the go during a period that started 290 days before the delivery and ended 240 days before the delivery. What is the probability that the baby was conceived while he was at home?

Let X be the actual duration of the pregnancy. The question is

$$P(\{X > 290\} \cup \{X < 240\}) = ?,$$

which we solve as follows,

$$\begin{aligned} P(\{X > 290\} \cup \{X < 240\}) &= P\left(\left\{\frac{X-270}{10} > 2\right\} \cup \left\{\frac{X-270}{10} < -3\right\}\right) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-3} e^{-y^2/2}\,dy + \frac{1}{\sqrt{2\pi}} \int_{2}^{\infty} e^{-y^2/2}\,dy \\ &= \Phi(-3) + [1 - \Phi(2)] = 0.241. \end{aligned}$$

(It is with great relief that we learn that after having completed this calculation, the sailor decided not to slaughter his wife.) ▲▲▲

Example: A fair coin is tossed 40 times. What is the probability that the number of Heads equals exactly 20?

Since the number of heads is a binomial variable, the answer is

$$\binom{40}{20}\left(\frac{1}{2}\right)^{20}\left(\frac{1}{2}\right)^{20} = 0.1268...$$

We can also approximate the answer using the DeMoivre-Laplace theorem. Indeed, the random variable

$$\frac{X - 40 \times \frac{1}{2}}{\sqrt{40 \times \frac{1}{2} \times (1 - \frac{1}{2})}} \approx \mathcal{N}(0, 1)\,.$$

The number of heads is a discrete variable, whereas the normal distribution refers to a continuous one. We will approximate the probability that the number of heads be 20 by the probability that it is, in a continuous context, between 19.5 and 20.5, i.e., that

$$\frac{|X - 20|}{\sqrt{10}} \le \frac{1}{2\sqrt{10}},$$

and

$$P\left(\frac{1}{2\sqrt{10}} \le \frac{X - 20}{\sqrt{10}} \le \frac{1}{2\sqrt{10}}\right) \approx 2\left(\Phi\left(\frac{1}{2\sqrt{10}}\right) - \Phi(0)\right) = 0.127...$$

▲▲▲

5.4 The exponential distribution

Definition 5.4 A random variable X is said to be ***exponentially distributed*** *with parameter λ, denoted X ∼ Exp(λ), if its* PDF *is*

$$f_X(x) = \begin{cases} \lambda\, e^{-\lambda x} & x \ge 0 \\ 0 & x < 0 \end{cases}.$$

The corresponding distribution function is

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

An exponential distribution is a suitable model in many situations, like the time until the next earthquake.

Example: Suppose that the duration of a phone call (in minutes) is a random variable $\mathcal{E}xp(1/10)$. What is the probability that a given phone call lasts more than 10 minutes? The answer is

$$P(X > 10) = 1 - F_x(10) = e^{-10/10} \approx 0.368.$$

Suppose we know that a phone call has already lasted 10 minutes. What is the probability that it will last at least 10 more minutes. The perhaps surprising answer is

$$P(X > 20|X > 10) = \frac{P(X > 20, X > 10)}{P(X > 10)} = \frac{e^{-2}}{e^{-1}} = e^{-1}.$$

More generally, we can show that for every $t > s$,

$$P(X > t|X > s) = P(X > t - s).$$

A random variable satisfying this property is called ***memoryless***. ▲▲▲

Proposition 5.2 A random variable that satisfies

$$P(X > t|X > s) = P(X > t - s) \qquad \textit{for all } t > s > 0$$

is exponentially distributed.

Proof: It is given that

$$\frac{P(X > t, X > s)}{P(X > s)} = P(X > t - s),$$

or in terms of the distribution function,

$$\frac{1 - F_X(t)}{1 - F_X(s)} = 1 - F_X(t - s).$$

Let $g(t) = 1 - F_X(t)$, then for all $t > s$,

$$g(t) = g(s)g(t - s),$$

and the only family of functions satisfying this property is the exponentials. ■

5.5 The Gamma distribution

The Gamma function is defined by

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

for $x > 0$. This function is closely related to factorials since $\Gamma(1) = 1$ and by integration by parts,

$$\Gamma(n+1) = \int_0^\infty t^n e^{-t}\, dt = n \int_0^\infty t^{n-1} e^{-t}\, dt = n\,\Gamma(n),$$

hence $\Gamma(n+1) = n!$ for integer n. A random variable X is said to be Gamma-distributed with parameters r, λ if it assumes positive values and

$$f_X(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}.$$

We denote it by $X \sim \mathcal{G}amma(r, \lambda)$. This is a normalized PDF since

$$\int_0^\infty f_X(x)\, dx = \frac{1}{\Gamma(r)} \int_0^\infty (\lambda x)^{r-1} e^{-\lambda x}\, d(\lambda x) = 1.$$

Note that for $r = 1$ we get the PDF of an exponential distribution, i.e.,

$$\mathcal{G}amma(1, \lambda) \sim \mathcal{E}xp(\lambda)\,.$$

The significance of the Gamma distribution will be seen later on in this chapter.

5.6 The Beta distribution

A random variable assuming values in $[0, 1]$ is said to have the Beta distribution with parameters $K, L > 0$, i.e., $X \sim \mathcal{B}eta(K, L)$, if it has the PDF

$$f_X(x) = \frac{\Gamma(K+L)}{\Gamma(K)\Gamma(L)} x^{K-1} (1-x)^{L-1}.$$

5.7 Functions of random variables

In this section we consider the following problem: let X be a continuous random variable with PDF $f_X(x)$. Let g be a real-valued function and $Y(\omega) = g(X(\omega))$. What is the distribution of Y?

Example: Let $X \sim \mathcal{U}(0,1)$. What is the distribution of $Y = X^n$?

The random variable Y, like X, assumes values in the interval $[0,1]$. Now,

$$F_Y(y) = P(Y \le y) = P(X^n \le y) = P(X \le y^{1/n}) = F_X(y^{1/n}),$$

where we used the monotonicity of the power function for positive arguments. In the case of a uniform distribution,

$$F_X(x) = \int_{-\infty}^{x} f_X(x) = \begin{cases} 0 & x < 0 \\ x & 0 \le x \le 1 \\ 1 & x > 1 \end{cases}.$$

Thus,

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ y^{1/n} & 0 \le y \le 1 \\ 1 & y > 1 \end{cases}.$$

Differentiating,

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \begin{cases} \frac{1}{n}\, y^{1/n-1} & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}.$$

▲▲▲

Example: Let X be a continuous random variable with PDF $f_X(x)$. What is the distribution of $Y = X^2$.

The main difference with the previous exercise is that X may possibly assume both positive and negative values, in which case the square function is non-monotonic. Thus, we need to proceed with more care,

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(X^2 \le y) \\ &= P(X \ge -\sqrt{y}, X \le \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$

Differentiating, we get the PDF

$$f_Y(y) = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}}.$$

▲▲▲

With these preliminaries, we can formulate the general theorem:

Theorem 5.2 Let X be a continuous random variable with PDF $f_X(x)$*. Let g be a strictly monotonic, differentiable function and set* $Y(\omega) = g(X(\omega))$*. Then the random variable Y has a* PDF

$$f_Y(y) = \begin{cases} \left|(g^{-1})'(y)\right| \, f_X(g^{-1}(y)) & y \text{ is in the range of } g(X) \\ 0 & \text{otherwise} \end{cases}.$$

Comment: If g is non-monotonic then $g^{-1}(y)$ may be set-valued and the above expression has to be replaced by a sum over all "branches" of the inverse function:

$$\sum_{g^{-1}(y)} \left|(g^{-1})'(y)\right| \, f_X(g^{-1}(y)).$$

Proof: Consider the case where g is strictly increasing. Then, g^{-1} exists, and

$$F_Y(y) = P(Y \le y) = P(g(X) \le y) = P(X \le g^{-1}(y)) = F_X(g^{-1}(y)),$$

and upon differentiation,

$$f_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = (g^{-1})'(y) \, f_X(g^{-1}(y)).$$

The case where g is strictly decreasing is handled similarly. ■

The inverse transformation method An application of this formula is the following. Suppose that you have a computer program that generates a random variable $X \sim \mathcal{U}(0,1)$. How can we use it to generate a random variable with distribution function F? The following method is known as the ***inverse transformation method***.

If F is strictly increasing (we know that it is at least non-decreasing), then we can define

$$Y(\omega) = F^{-1}(X(\omega)).$$

Note that F^{-1} maps $[0,1]$ onto the entire real line, while X has range $[0,1]$. Moreover, $F_X(x)$ is the identity on $[0,1]$. By the above formula,

$$F_Y(y) = F_X(F(y)) = F(y).$$

Example: Suppose we want to generate an exponential variable $Y \sim \mathcal{Exp}(\lambda)$, in which case $F(y) = 1 - e^{-\lambda y}$. The inverse function is $F^{-1}(x) = -\frac{1}{\lambda}\log(1-x)$, i.e., an exponential variable is generated by setting

$$Y(\omega) = -\frac{1}{\lambda}\log(1 - X(\omega)).$$

In fact, since $1 - X$ has the same distribution as X, we may equally well take $Y = -\lambda^{-1}\log X$.

▲▲▲

5.8 Multivariate distributions

We proceed to consider joint distributions of multiple random variables. The treatment is fully analogous to that for discrete variables.

*Definition 5.5 A pair of random variables X, Y over a probability space $(\Omega, \mathscr{F}, P)$ is said to have a **continuous joint distribution** if there exists an integrable non-negative bi-variate function $f_{X,Y}(x,y)$ (the **joint** PDF) such that for every (measurable) set $A \subseteq \mathbb{R}^2$,*

$$P_{X,Y}(A) = P(\{\omega : (X(\omega), Y(\omega)) \in A\}) = \iint_A f_{X,Y}(x,y)\,dxdy.$$

Note that in particular,

$$F_{X,Y}(x,y) = P(\{\omega : X(\omega) \le x, Y(\omega) \le y\}) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(x,y)\,dxdy,$$

and consequently,

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x\,\partial y} F_{X,Y}(x,y).$$

Furthermore, if X, Y are jointly continuous, then each random variable is continuous as a single variable. Indeed, for all $A \subseteq \mathbb{R}$,

$$P_X(A) = P_{X,Y}(A \times \mathbb{R}) = \int_A \left[\int_{\mathbb{R}} f_{X,Y}(x,y)\,dy \right] dx,$$

from which we identify the ***marginal*** PDF of X,

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)\,dy,$$

with an analogous expression for $f_Y(y)$. The generalization to ***multivariate distributions*** is straightforward.

Example: Consider a uniform distribution inside a circle of radius R,

$$f_{X,Y}(x,y) = \begin{cases} C & x^2 + y^2 \le R^2 \\ 0 & \text{otherwise} \end{cases}.$$

(1) What is C? (2) What is the marginal distribution of X? (3) What is the probability that the Euclidean norm of (X, Y) is less than a?

(1) The normalization condition is

$$\int_{x^2+y^2 \le R^2} C\,dxdy = \pi R^2 C = 1.$$

(2) For $|x| \le R$ the marginal PDF of X is given by

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)\,dy = \frac{1}{\pi R^2} \int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} dy = \frac{2\sqrt{R^2 - x^2}}{\pi R^2}.$$

Finally,

$$P(X^2 + Y^2 \le a^2) = \frac{a^2}{R^2}.$$

▲▲▲

We next consider how does independence affect the joint PDF. Recall that X, Y are said to be independent if for all $A, B \subseteq \mathbb{R}$,

$$P_{X,Y}(A \times B) = P_X(A) P_Y(B).$$

For continuous distributions, this means that for all A, B,

$$\int_A \int_B f_{X,Y}(x,y)\, dxdy = \int_A f_X(x)\, dx \int_B f_Y(y)\, dy,$$

from which we conclude that

$$f_{X,Y}(x,y) = f_X(x) f_Y(y).$$

Similarly, n random variables with continuous joint distribution are independent if their joint PDF equals to the product of their marginal PDFS.

Example: Let X, Y, Z be independent variables all being $\mathcal{U}(0,1)$. What is the probability that $X > YZ$?

The joint distribution of X, Y, Z is

$$f_{X,Y,Z}(x,y,z) = f_X(x) f_Y(y) f_Z(z) = \begin{cases} 1 & x,y,z \in [0,1] \\ 0 & \text{otherwise} \end{cases}.$$

Now,

$$\begin{aligned} P(X > YZ) &= \iint_{x>yz} dxdydz = \int_0^1 \int_0^1 \left(\int_{yz}^1 dx \right) dydz \\ &= \int_0^1 \int_0^1 (1 - yz)\, dydz = \int_0^1 \left(1 - \tfrac{z}{2}\right) dz = \tfrac{3}{4}. \end{aligned}$$

▲▲▲

Sums of independent random variables Let X, Y be independent continuous random variables. What is the distribution of $X + Y$?

We proceed as usual,

$$\begin{aligned}F_{X+Y}(z) = P(X+Y \le z) &= \iint_{x+y\le z} f_X(x) f_Y(y)\, dxdy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f_X(x) f_Y(y)\, dxdy \\ &= \int_{-\infty}^{\infty} F_X(z-y) f_Y(y)\, dy.\end{aligned}$$

Differentiating, we obtain,

$$f_{X+Y}(z) = \frac{d}{dz} F_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\, dy,$$

i.e., the PDF of a sum is the ***convolution*** of the PDFS, $f_{X+Y} = f_X * f_Y$.

Example: What is the distribution of $X+Y$ when $X, Y \sim \mathcal{U}(0,1)$ are independent? We have

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\, dy = \int_{z-1}^{z} f_X(w)\, dw.$$

The integral vanishes if $z < 0$ and if $z > 2$. Otherwise,

$$f_{X+Y}(z) = \begin{cases} z & 0 \le z \le 1 \\ 2-z & 1 < z \le 2 \end{cases}.$$

▲▲▲

We conclude this section with a general formula for variable transformations. Let $\mathbf{X} = (X_1, X_2)$ be two random variables with joint PDF $f_{\mathbf{X}}(\mathbf{x})$, and set

$$\boldsymbol{Y} = \boldsymbol{g}(\mathbf{X}).$$

What is the joint PDF of $\boldsymbol{Y} = (Y_1, Y_2)$? We will assume that these relations are invertible, i.e., that

$$\mathbf{X} = \boldsymbol{g}^{-1}(\boldsymbol{Y}).$$

Furthermore, we assume that $\boldsymbol{g}$ is differentiable. Then,

$$F_{\boldsymbol{Y}}(\boldsymbol{y}) = \iint_{\boldsymbol{g}(\mathbf{x})\le \boldsymbol{y}} f_{\mathbf{X}}(\mathbf{x})\, dx_1 dx_2 = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} f_{\mathbf{X}}(\boldsymbol{g}^{-1}(\boldsymbol{u}))\, |J(\boldsymbol{u})|\; du_1 du_2,$$

where $J(\boldsymbol{y}) = \partial(\mathbf{x})/\partial(\boldsymbol{y})$ is the Jacobian of the transformation. Differentiating twice with respect to y_1, y_2 we obtain the joint PDF,

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = |J(\boldsymbol{y})|\; f_{\mathbf{X}}(\boldsymbol{g}^{-1}(\boldsymbol{u})).$$

✎ *Exercise 5.2* Let X_1, X_2 be two independent random variables with distribution $\mathcal{U}(0,1)$ (i.e., the variables that two subsequent calls of the `rand()` function on a computer would return). Define,

$$Y_1 = \sqrt{-2\log X_1}\,\cos(2\pi X_2)$$
$$Y_2 = \sqrt{-2\log X_1}\,\sin(2\pi X_2).$$

Show that Y_1 and Y_2 are independent and distributed $\mathcal{N}(0,1)$. This is the standard way of generating normally-distributed random variables on a computer. This change of variables is called the ***Box-Muller transformation*** (G.E.P. Box and M.E. Muller, 1958).

Example: Suppose that $X \sim \mathcal{G}amma(K,1)$ and $Y \sim \mathcal{G}amma(L,1)$ are independent, and consider the variables

$$V = \frac{X}{X+Y} \qquad \text{and} \qquad W = X + Y.$$

The reverse transformation is

$$X = VW \qquad \text{and} \qquad Y = W(1-V).$$

Since $X, Y \in [0,\infty)$ it follows that $V \in [0,1]$ and $W \in [0,\infty)$.

The Jacobian is

$$|J(v,w)| = \begin{vmatrix} w & -w \\ v & 1-v \end{vmatrix} = w.$$

Thus,

$$\begin{aligned} f_{V,W}(v,w) &= \frac{(vw)^{K-1}e^{-vw}}{\Gamma(K)}\frac{[w(1-v)]^{L-1}e^{-w(1-v)}}{\Gamma(L)}w \\ &= \frac{w^{K+L-1}e^{-w}}{\Gamma(K+L)} \times \frac{\Gamma(K+L)}{\Gamma(K)\Gamma(L)}v^{K-1}(1-v)^{L-1}. \end{aligned}$$

This means that

$$V \sim \mathcal{B}eta(K,L) \qquad \text{and} \qquad W \sim \mathcal{G}amma(K+L,1)\,.$$

Moreover, they are independent. ▲▲▲

Example: We now develop a general formula for the PDF ***of ratios***. Let X, Y be random variables, not necessarily independent, and set

$$V = X \qquad \text{and} \qquad W = X/Y.$$

The inverse transformation is

$$X = V \qquad \text{and} \qquad Y = V/W.$$

The Jacobian is

$$|J(v,w)| = \begin{vmatrix} 1 & 1/w \\ 0 & -v/w^2 \end{vmatrix} = \left| \frac{v}{w^2} \right|.$$

Thus,

$$f_{V,W}(v,w) = f_{X,Y}\left(v, \frac{v}{w}\right) \left| \frac{v}{w^2} \right|,$$

and the uni-variate distribution of W is given by

$$f_W(w) = \int f_{X,Y}\left(v, \frac{v}{w}\right) \left| \frac{v}{w^2} \right| \, dv.$$

▲▲▲

✎ *Exercise 5.3* Find the distribution of X/Y when $X, Y \sim \mathcal{E}xp(1)$ are independent.

Example: Let $X \sim \mathcal{U}(0,1)$ and let Y be any (continuous) random variable independent of X. Define

$$W = X + Y \mod 1.$$

What is the distribution of W?

Clearly, W assumes value in $[0,1]$. We need to express the set $\{W \le c\}$ in terms of X, Y. If we decompose $Y = N + Z$, where $Z = Y \mod 1$, then

$$\begin{aligned} \{W \le c\} &= \{Z \le c\} \cap \{\, 0 \le X \le c - Z\} \\ &\cup \{Z \le c\} \cap \{\, 1 - Z \le X \le 1\} \\ &\cup \{Z > c\} \cap \{\, 1 - Z \le X \le 1 - (Z - c)\} \end{aligned}$$

It follows that

$$P(W \le c) = \sum_{n=-\infty}^{\infty} \int_0^1 f_Y(n+y) \left[c\, I_{z \le c} + c\, I_{z > c} \right] dz = c,$$

i.e., no matter what Y is, $W \sim \mathcal{U}(0,1)$. ▲▲▲

5.9 Conditional distributions and conditional densities

Remember that in the case of *discrete* random variables we defined

$$p_{X|Y}(x|y) := P(X = x|Y = y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}.$$

Since the PDF is, in a sense, the continuous counterpart of the atomistic distribution, the following definition seems most appropriate:

Definition 5.6 The ***conditional probability density function*** *(CPDF) of X given Y is*

$$f_{X|Y}(x|y) := \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

The question is what is the *meaning* is this conditional density? First, we note that viewed as a function of x, with y fixed, it is a density, as it is non-negative, and

$$\int_{\mathbb{R}} f_{X|Y}(x|y)\,dx = \frac{\int_{\mathbb{R}} f_{X,Y}(x,y)\,dx}{f_Y(y)} = 1.$$

Thus, it seems natural to speculate that the integral of the CPDF over a set A is the probability that $X \in A$ given that $Y = y$,

$$\int_A f_{X|Y}(x|y)\,dx \overset{?}{=} P(X \in A|Y = y).$$

The problem is that the right hand side is not defined, since the condition $(Y = y)$ has probability zero!

A heuristic way to resolve the problem is the following (for a rigorous way we need again measure theory): construct a sequence of sets $B_n \subset \mathbb{R}$, such that $B_n \to \{y\}$ and each of the B_n has finite measure (for example, $B_n = (y - 1/n, y + 1/n)$), and define

$$P(X \in A|Y = y) = \lim_{n\to\infty} P(X \in A|Y \in B_n).$$

Now, the right-hand side is well-defined, provided the limit exists. Thus,

$$\begin{aligned} P(X \in A|Y = y) &= \lim_{n\to\infty} \frac{P(X \in A, Y \in B_n)}{P(Y \in B_n)} \\ &= \lim_{n\to\infty} \frac{\int_A \int_{B_n} f_{X,Y}(x,u)\, dudx}{\int_{B_n} f_Y(u)\, du} \\ &= \int_A \lim_{n\to\infty} \frac{\int_{B_n} f_{X,Y}(x,u)\, du}{\int_{B_n} f_Y(u)\, du}\, dx \\ &= \int_A \frac{f_{X,Y}(x,y)}{f_Y(y)}\, dx, \end{aligned}$$

where we have used something analogous to l'Hopital's rule in taking the limit. This is precisely the identity we wanted to obtain.

What is the CPDF good for? We have the identity

$$f_{X,Y}(x,y) = f_{X|Y}(x|y) f_Y(y).$$

In many cases, it is more natural to define models in terms of conditional densities, and our formalism tells us how to convert this data into joint distributions.

Example: Let the joint PDF of X, Y be given by

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{y} e^{-x/y} e^{-y} & x, y \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

What is the CPDF of X given Y, and what is the probability that $X(\omega) > 1$ given that $Y = y$?

For $x, y \geq 0$ the CPDF is

$$f_{X|Y}(x|y) = \frac{\frac{1}{y} e^{-x/y} e^{-y}}{\int_0^\infty \frac{1}{y} e^{-x/y} e^{-y}\, dx} = \frac{1}{y} e^{-x/y},$$

and

$$P(X > 1|Y = y) = \int_1^\infty f_{X|Y}(x|y)\, dx = \frac{1}{y} \int_1^\infty e^{-x/y}\, dx = e^{-1/y}.$$

▲▲▲

5.10 Expectation

Recall our definition of the expectation for discrete probability spaces,

$$\mathbb{E}[X] = \sum_{\omega\in\Omega} X(\omega)\, p(\omega),$$

where $p(\omega)$ is the atomistic probability in Ω, i.e., $p(\omega) = P(\{\omega\})$. We saw that an equivalent definition was

$$\mathbb{E}[X] = \sum_{x\in S_x} x\, p_X(x).$$

In a more general context, the first expression is the integral of the function $X(\omega)$ over the probability space $(\Omega, \mathscr{F}, P)$, whereas the second equation is the integral of the identity function $X(x) = x$ over the probability space $(S_x, \mathscr{F}_X, P_X)$. We now want to generalize these definitions for uncountable spaces.

The definition of the expectation in the general case relies unfortunately on integration theory, which is part of measure theory. The expectation of X is defined as

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

but this is not supposed to make much sense to us. On the other hand, the equivalent definition,

$$\mathbb{E}[X] = \int_{\mathbb{R}} x\, P_X(dx),$$

does make sense if we identify $P_X(dx)$ with $f_X(x)\, dx$. That is, our definition of the expectation for continuous random variables is

$$\mathbb{E}[X] = \int_R x\, f_X(x)\, dx.$$

Example: For $X \sim \mathcal{U}(a,b)$,

$$\mathbb{E}[X] = \frac{1}{b-a}\int_a^b x\, dx = \frac{a+b}{2}.$$

▲▲▲

Example: For $X \sim \mathcal{E}xp(\lambda)$,

$$\mathbb{E}[X] = \int_0^\infty x\,\lambda e^{-\lambda x}\,dx = \frac{1}{\lambda}.$$

▲▲▲

Example: For $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$\mathbb{E}[X] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} x\,e^{-(x-\mu)^2/2\sigma^2}\,dx = \mu.$$

▲▲▲

Lemma 5.1 Let Y be a continuous random variable with PDF $f_Y(y)$. *Then*

$$\mathbb{E}[Y] = \int_0^\infty \left[1 - F_Y(y) - F_Y(-y)\right]\,dy.$$

Proof: Note that the lemma states that

$$\mathbb{E}[Y] = \int_0^\infty \left[P(Y > y) - P(Y \le -y)\right]\,dy.$$

We start with the first expression

$$\begin{aligned}\int_0^\infty P(Y > y)\,dy &= \int_0^\infty \int_y^\infty f_Y(u)\,dudy \\ &= \int_0^\infty \int_0^u f_Y(u)\,dydu \\ &= \int_0^\infty u\,f_Y(u)\,du,\end{aligned}$$

where the passage from the first to the second line involves a change in the order of integration, with the corresponding change in the limits of integration. Similarly,

$$\begin{aligned}\int_0^\infty P(Y \le -y)\,dy &= \int_0^\infty \int_{-\infty}^{-y} f_Y(u)\,dudy \\ &= \int_{-\infty}^0 \int_0^{-u} f_Y(u)\,dydu \\ &= -\int_{-\infty}^0 u\,f_Y(u)\,du.\end{aligned}$$

Subtracting the two expressions we get the desired result. ■

Theorem 5.3 (The unconscious statistician) Let X be a continuous random variable and let g : $\mathbb{R} \to \mathbb{R}$. *Then,*

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) f_X(x)\, dx.$$

Proof: In principle, we could write the PDF of $g(X)$ and follow the definition of its expected value. The fact that g does not necessarily have a unique inverse complicates the task. Thus, we use instead the previous lemma,

$$\begin{aligned}\mathbb{E}[g(X)] &= \int_0^\infty P(g(X) > y)\, dy - \int_0^\infty P(g(X) \le -y)\, dy \\ &= \int_0^\infty \int_{g(x)>y} f_X(x)\, dxdy - \int_0^\infty \int_{g(x)\le -y} f_X(x)\, dxdy.\end{aligned}$$

We now exchange the order of integration. Note that for the first integral,

$$\{0 < y < \infty, g(x) > y\} \qquad \text{can be written as} \qquad \{x \in \mathbb{R}, 0 < y < g(x)\}$$

whereas for the second integral,

$$\{0 < y < \infty, g(x) < -y\} \qquad \text{can be written as} \qquad \{x \in \mathbb{R}, 0 < y < -g(x)\}$$

Thus,

$$\begin{aligned}\mathbb{E}[g(X)] &= \int_{\mathbb{R}} \int_0^{\max(0,g(x))} f_X(x)\, dydx - \int_{\mathbb{R}} \int_0^{\max(0,-g(x))} f_X(x)\, dydx \\ &= \int_{\mathbb{R}} [\max(0, g(x)) - \max(0, -g(x))]\, f_X(x)\, dx \\ &= \int_{\mathbb{R}} g(x) f_X(x)\, dx.\end{aligned}$$

■

Example: What is the variance of $X \sim \mathcal{N}(\mu, \sigma^2)$?

$$\begin{aligned}\operatorname{Var}[X] &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} (x-\mu)^2 e^{-(x-\mu)^2/2\sigma^2}\, dx \\ &= \frac{\sigma^3}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} u^2 e^{-u^2/2}\, du = \sigma^2.\end{aligned}$$

▲▲▲

✎ *Exercise 5.4* Let $X \sim \mathcal{N}(0, \sigma^2)$. Calculate the moments $\mathbb{E}[X^k]$ (hint: consider separately the cases of odd and even k's).

The law of the unconscious statistician is readily generalized to multiple random variables, for example,

$$\mathbb{E}[g(X,Y)] = \iint_{\mathbb{R}^2} g(x,y) f_{X,Y}(x,y)\, dxdy.$$

✎ *Exercise 5.5* Show that if X and Y are independent continuous random variables, then for every two functions f, g,

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\, \mathbb{E}[g(Y)].$$

5.11 The moment generating function

As for discrete variables the moment generating function is defined as

$$M_X(t) := \mathbb{E}[e^{tX}] = \int_{\mathbb{R}} e^{tx} f_X(x)\, dx,$$

that is, it is the ***Laplace transform*** of the PDF. Without providing a proof, we state that the transformation $f_X \mapsto M_X$ is invertible (it is one-to-one), although the formula for the inverse is complicated and relies on complex analysis.

Comment: A number of other generating functions are commonly defined: first the ***characteristic function***,

$$\varphi_X(t) = \mathbb{E}[e^{\imath tX}] = \int_{\mathbb{R}} e^{\imath tx} f_X(x)\, dx,$$

which unlike the moment generating function is always well defined for every t. Since its use relies on complex analysis we do not use it in this course. Another used generating function is the ***probability generating function***

$$g_X(t) = \mathbb{E}[t^X] = \sum_x t^x p_X(x).$$

Example: What is the moment generating function of $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$?

$$\begin{aligned}
M_X(t) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{tx} e^{-(x-\mu)^2/2\sigma^2}\, dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} \exp\left[-\frac{x^2 - 2\mu x + \mu^2 - 2\sigma^2 tx}{2\sigma^2}\right] dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\mu^2/2\sigma^2} e^{(\mu+\sigma^2 t)^2/2\sigma^2} \int_{\mathbb{R}} \exp\left[-\frac{(x-\mu-\sigma^2 t)^2}{2\sigma^2}\right] dx \\
&= \exp\left[\mu t + \frac{\sigma^2}{2} t^2\right].
\end{aligned}$$

From this we readily obtain, say, the first two moments,

$$\mathbb{E}[X] = M_X'(0) = (\mu + \sigma^2 t) e^{\mu t + \frac{1}{2}\sigma^2 t^2}\Big|_{t=0} = \mu,$$

and

$$\mathbb{E}[X^2] = M_X''(0) = \left[(\mu + \sigma^2 t)^2 + \sigma^2\right] e^{\mu t + \frac{1}{2}\sigma^2 t^2}\Big|_{t=0} = \sigma^2 + \mu^2,$$

as expected. ▲▲▲

Example: Recall the Gamma-distribution whose PDF is

$$f_X(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}.$$

To calculate its moments it is best to use the moment generating function,

$$M_X(t) = \frac{\lambda^r}{\Gamma(r)} \int_0^\infty e^{tx} x^{r-1} e^{-\lambda x}\, dx = \frac{\lambda^r}{(\lambda - t)^r},$$

defined only for $t < \lambda$. We can then calculate the moment, e.g.,

$$\mathbb{E}[X] = M_X'(0) = \lambda^r\, r(\lambda - t)^{-(r+1)}|_{t=0} = \frac{r}{\lambda},$$

and

$$\mathbb{E}[X^2] = M_X''(0) = \lambda^r\, r(r+1)(\lambda - t)^{-(r+2)}|_{t=0} = \frac{r(r+1)}{\lambda^2},$$

from which we conclude that

$$\text{Var}[X] = \frac{r}{\lambda^2}.$$

▲▲▲

From the above discussion it follows that the moment generating function embodies the same information as the PDF. A nice property of the moment generating function is that it converts convolutions into products. Specifically,

Proposition 5.3 Let f_X and f_Y be probability densities functions and let $f = f_X * f_Y$ be their convolution. If M_X, M_Y and M are the moment generating functions associated with f_X, f_Y and f, respectively, then $M = M_X M_Y$.

Proof: By definition,

$$\begin{aligned} M(t) &= \int_{\mathbb{R}} e^{tx} f(x)\,dx = \int_{\mathbb{R}} e^{tx} \int_{\mathbb{R}} f_X(y) f_Y(x-y)\,dy\,dx \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} e^{ty} f_X(y) e^{t(x-y)} f_Y(x-y)\,dy\,d(x-y) \\ &= \int_{\mathbb{R}} e^{ty} f_X(y)\,dy \int_{\mathbb{R}} e^{tu} f_Y(u)\,du = M_X(t) M_Y(t). \end{aligned}$$

■

Example: Here is an application of the above proposition. Let $X \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right)$ and $Y \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right)$ be independent variables. We have already calculated their moment generating function,

$$\begin{aligned} M_X(t) &= \exp\left[\mu_1 t + \frac{\sigma_1^2}{2} t^2\right] \\ M_Y(t) &= \exp\left[\mu_2 t + \frac{\sigma_2^2}{2} t^2\right]. \end{aligned}$$

By the above proposition, the generating function of their sum is the product of the generating functions,

$$M_{X+Y}(t) = \exp\left[(\mu_1 + \mu_2)t + \frac{\sigma_1^2 + \sigma_2^2}{2}t^2\right],$$

from which we conclude at once that

$$X + Y \sim \mathcal{N}\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right),$$

i.e., sums of independent normal variables are normal. ▲▲▲

Example: Consider now the sum of n independent exponential random variables $X_i \sim \mathcal{E}xp(\lambda)$. Since $\mathcal{E}xp(\lambda) \sim \mathcal{G}amma(1,\lambda)$ we know that

$$M_{X_i}(t) = \frac{\lambda}{\lambda - t}.$$

The PDF of the sum of n independent random variables,

$$Y = \sum_{i=1}^{n} X_i$$

is the n-fold convolution of their PDFS, and its generating function is the product of their generating functions,

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) = \frac{\lambda^n}{(\lambda - t)^n},$$

which we identify as the generating function of the $\mathcal{G}amma(n,\lambda)$ distribution. Thus ***the Gamma distribution with parameters*** (n,λ) ***characterizes the sum of*** n ***independent exponential variables with parameter*** λ***.*** ▲▲▲

✎ *Exercise 5.6* What is the distribution of $X_1 + X_2$ where $X_1 \sim \mathcal{G}amma(r_1,\lambda)$ and $X_2 \sim \mathcal{G}amma(r_2,\lambda)$ are independent?

Example: A family of distributions that have an important role in statistics are the χ_ν^2 distributions with $\nu = 1, 2 \ldots$. A random variable Y has the χ_ν^2-distribution if it is distributed like

$$Y \sim X_1^2 + X_2^2 + \cdots + X_\nu^2,$$

where $X_i \sim \mathcal{N}(0,1)$ are independent.

The distribution of X_1^2 is obtained by the change of variable formula,

$$f_{X_1^2}(x) = \frac{f_{X_1}(\sqrt{x}) + f_{X_1}(-\sqrt{x})}{2\sqrt{x}} = 2\frac{\frac{1}{\sqrt{2\pi}}e^{-x/2}}{2\sqrt{x}} = \frac{1}{\sqrt{2\pi x}}e^{-x/2}.$$

The moment generating function is

$$M_{X_1^2}(t) = \int_0^\infty e^{tx}\frac{1}{\sqrt{2\pi x}}e^{-x/2}\,dx = \frac{2}{\sqrt{2\pi}}\int_0^\infty e^{-\frac{1}{2}(1-2t)y^2}\,dy = (1-2t)^{-1/2},$$

and by the addition rule, the moment generating function of the χ_ν^2-distribution is

$$M_Y(t) = (1-2t)^{-\nu/2} = \frac{(1/2)^{\nu/2}}{(1/2-t)^{\nu/2}}.$$

We identify this moment generating function as that of $\mathcal{Gamma}(\nu/2, 1/2)$. ▲▲▲

5.12 Other distributions

We conclude this section with two distributions that have major roles in statistics. Except for the additional exercise in the change of variable formula, the goal is to know the definition of these very useful distributions.

Definition 5.7 Let $X \sim \chi_r^2$ and $Y \sim \chi_s^2$ be independent. A random variable that has the same distribution as

$$W = \frac{X/r}{Y/s}$$

is said to have the ***Fischer*** *$F_{r,s}$* ***distribution****.*

Since, by the previous section

$$f_{X/r}(x) = \frac{(r/2)^r\,(rx)^{r/2-1}e^{-\frac{1}{2}rx}}{\Gamma(\frac{r}{2})}\,r$$

$$f_{Y/s}(y) = \frac{(s/2)^s\,(sy)^{s/2-1}e^{-\frac{1}{2}sy}}{\Gamma(\frac{s}{2})}\,s,$$

it follows from the distribution of ratios formula that

$$\begin{aligned} f_W(w) &= \int_0^\infty \frac{\left(\frac{1}{2}\right)^{r/2} (rv)^{r/2-1} e^{-\frac{1}{2}rv}}{\Gamma(\frac{r}{2})}\, r \frac{\left(\frac{1}{2}\right)^{s/2} (sv/w^2)^{s/2-1} e^{-\frac{1}{2}sv/w^2}}{\Gamma(\frac{s}{2})}\, s\frac{v}{w^2}\, dv \\ &= \frac{1}{\Gamma(\frac{r}{2})\Gamma(\frac{s}{2})} \frac{(\frac{1}{2}r)^{r/2}(\frac{1}{2}s)^{s/2}}{w^s} \int_0^\infty v^{r/2+s/2-1} e^{-\frac{1}{2}v(r+s/w^2)}\, dv. \end{aligned}$$

Changing variables we get

$$\begin{aligned} f_W(w) &= \frac{1}{\Gamma(\frac{r}{2})\Gamma(\frac{s}{2})} \frac{(\frac{1}{2}r)^{r/2}(\frac{1}{2}s)^{s/2}}{w^s} \left[\frac{1}{2}(r+s/w^2)\right]^{-(r/2+s/2)} \int_0^\infty \xi^{r/2+s/2-1} e^{-\xi}\, d\xi \\ &= \frac{\Gamma(\frac{r}{2}+\frac{s}{2})}{\Gamma(\frac{r}{2})\Gamma(\frac{s}{2})} \frac{(\frac{1}{2}r)^{r/2}(\frac{1}{2}s)^{s/2}}{w^s} \left[\frac{1}{2}(r+s/w^2)\right]^{-(r/2+s/2)} \\ &= \frac{\Gamma(\frac{r}{2}+\frac{s}{2})}{\Gamma(\frac{r}{2})\Gamma(\frac{s}{2})} \frac{r^{r/2}s^{s/2}}{w^s(r+s/w^2)^{\frac{1}{2}(r+s)}}. \end{aligned}$$

Definition 5.8 *Let $X \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_\nu$ be independent. A random variable that has the same distribution as*

$$W = \frac{X}{\sqrt{Y/\nu}}$$

is said to have the ***Student's*** t_ν ***distribution****.*

✎ *Exercise 5.7* Find the PDF of the Student's t_ν distribution.

Chapter 6

Inequalities

6.1 The Markov and Chebyshev inequalities

As you've probably seen in today's front page: the upper tenth percentile earns 12 times more than the average salary. The following theorem will show that this is not possible.

Theorem 6.1 (Markov inequality) Let X be a random variable assuming non-negative values. Then for every $a \geq 0$,

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Comment: Note first that this is a vacuous statement for $a < \mathbb{E}[X]$. For $a > \mathbb{E}[X]$ this inequality limits the probability that X assumes values larger than its mean. This is the first time in this course that we derive an inequality. Inequalities, in general, are an important tool for analysis, where estimates (rather than exact identities) are needed.

Proof: We will assume a continuous variable. A similar proof holds in the discrete case.

$$\mathbb{E}[X] = \int_0^\infty x f_X(x)\, dx \geq \int_a^\infty x f_X(x)\, dx \geq a \int_a^\infty f_X(x)\, dx = a\, P(X > a).$$

■

Theorem 6.2 (Chebyshev inequality) Let X be a random variable with mean value μ and variance σ^2. Then, for every $a > 0$

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}.$$

Comment: If we write $a = k\sigma$, this theorem states that the probability that a random variable assumes a value whose absolute distance from its mean is more than k times its standard deviation is less that $1/k^2$.

The notable thing about these inequalities is that they make no assumption about the distribution. As a result, of course, they may provide very loose estimates.

Proof: Since $|X - \mu|^2$ is a positive variable we may apply the Markov inequality,

$$P(|X - \mu| \geq a) = P(|X - \mu|^2 \geq a^2) \leq \frac{\mathbb{E}[|X - \mu|^2]}{a^2} = \frac{\sigma^2}{a^2}.$$

■

Example: On average, an alcoholic drinks 25 liters of wine every week. What is the probability that he drinks more than 50 liters of wine on a given week?

Here we apply the Markov inequality. If X is the amount of wine he drinks on a given week, then

$$P(X > 50) \leq \frac{\mathbb{E}[X]}{50} = \frac{1}{2}.$$

▲▲▲

Example: Let $X \sim \mathcal{U}(0, 10)$ what is the probability that $|X - \mathbb{E}[X]| > 4$?

Since $\mathbb{E}[X] = 5$, the answer is 0.2. The Chebyshev inequality, on the other hand gives,

$$P(|X - \mathbb{E}[X]| > 4) \leq \frac{\sigma^2}{16} = \frac{25/3}{16} \approx 0.52.$$

▲▲▲

6.2 Jensen's inequality

Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is called ***convex*** if it is always below its secants, i.e., if for every x, y and $0 < \lambda < 1$,

$$f(\lambda x + (1 - \lambda)y) \le \lambda f(x) + (1 - \lambda)f(y).$$

Proposition 6.1 (Jensen's inequality) If g is a convex real valued function and X is a real valued random variable, then

$$g(\mathbb{E}[X]) \le \mathbb{E}[g(X)],$$

provided that the expectations exist.

Proof: Let's start with an easy proof for a particular case. Consider a continuous random variable and assume that g is twice differentiable (therefore $g''(x) \ge 0$). Taylor expanding g about $\mu = \mathbb{E}[X]$,

$$g(x) = g(\mu) + g'(\mu)\,(x - \mu) + \frac{1}{2}g''(\xi)(x - \mu)^2 \ge g(\mu) + g'(\mu)\,(x - \mu).$$

Multiplying both sides by the non-negative functions f_X and integrating over x we get

$$\int_{\mathbb{R}} g(x) f_X(x)\,dx \ge \int_{\mathbb{R}} g(\mu) f_X(x)\,dx + \int_{\mathbb{R}} g'(\mu)\,(x - \mu) f_X(x)\,dx = g(\mu),$$

which is precisely what we need to show.

What about the more general case? Any convex function is continuous, and has one-sided derivatives with

$$g'_-(x) := \lim_{y \uparrow x} \frac{g(x) - g(y)}{x - y} \le \lim_{y \downarrow x} \frac{g(x) - g(y)}{x - y} =: g'_+(x).$$

For every $m \in [g'_-(\mu), g'_+(\mu)]$

$$g(x) \ge g(\mu) + m(x - \mu),$$

so the same proof holds with m replacing $g'(\mu)$. If X is a discrete variable, we use summation instead of integration. ■

Example: Since exp is a convex function,

$$\exp(t\,\mathbb{E}[X]) \le \mathbb{E}[e^{tX}] = M_X(t),$$

or,

$$\mathbb{E}[X] \le \frac{1}{t}\log M_X(t),$$

for all $t > 0$. ▲▲▲

Example: Consider a discrete random variable assuming the positive values $x_1, \dots, x_n$ with equal probability $1/n$. Jensen's inequality for $g(x) = -\log(x)$ gives,

$$-\log\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \le -\frac{1}{n}\sum_{i=1}^{n}\log(x_i) = -\log\left(\prod_{i=1}^{n} x_i^{1/n}\right).$$

Reversing signs, and exponentiating, we get

$$\frac{1}{n}\sum_{i=1}^{n} x_i \ge \left(\prod_{i=1}^{n} x_i\right)^{1/n},$$

which is the classical ***arithmetic mean-geometric mean inequality***. In fact, this inequality can be generalized for arbitrary distributions, $p_X(x_i) = p_i$, yielding

$$\sum_{i=1}^{n} p_i x_i \ge \prod_{i=1}^{n} x_i^{p_i}.$$

▲▲▲

6.3 Kolmogorov's inequality

The Kolmogorov inequality may first seem to be of similar flavor as Chebyshev's inequality, but it is considerably stronger. I have decided to include it here because its proof involves some interesting subtleties. First, a lemma:

Lemma 6.1 If X, Y, Z are random variables such that Y is independent of X and Z, then

$$\mathbb{E}[XY|Z] = \mathbb{E}[X|Z]\,\mathbb{E}[Y].$$

Proof: The fact that Y is independent of both X, Z implies that (in the case of discrete variables),

$$p_{X,Y,Z}(x,y,z) = p_{X,Z}(x,z)p_Y(y).$$

Now, for every z,

$$\begin{aligned}
\mathbb{E}[XY|Z=z] &= \sum_{x,y} xy\, p_{X,Y|Z}(x,y|z) \\
&= \sum_{x,y} xy\, \frac{p_{X,Y,Z}(x,y,z)}{p_Z(z)} \\
&= \sum_{x,y} xy\, \frac{p_{X,Z}(x,z)p_Y(y)}{p_Z(z)} \\
&= \sum_{y} y\, p_Y(y) \sum_x x\, \frac{p_{X,Z}(x,z)}{p_Z(z)} \\
&= \mathbb{E}[Y]\,\mathbb{E}[X|Z=z].
\end{aligned}$$

■

Theorem 6.3 (Kolmogorov's inequality) Let $X_1, \dots, X_n$ *be independent random variables such that* $\mathbb{E}[X_k] = 0$ *and* $\mathrm{Var}[X_k] = \sigma_k^2 < \infty$*. Then, for all* $a > 0$*,*

$$P\left(\max_{1\le k\le n} |X_1 + \cdots + X_k| \ge a\right) \le \frac{1}{a^2}\sum_{i=1}^n \sigma_i^2.$$

Comment: For $n = 1$ this is nothing but the Chebyshev inequality. For $n > 1$ it would still be Chebyshev's inequality if the maximum over $1 \le k \le n$ was replaced by $k = n$, since by independence

$$\mathrm{Var}[X_1 + \cdots + X_n] = \sum_{i=1}^n \sigma_i^2.$$

Proof: We introduce the notation $S_k = X_1 + \cdots + X_k$. This theorem is concerned with the probability that $|S_k| > a$ for *some* k. We define the random variable $N(\omega)$

to be the smallest integer k for which $|S_k| > a$; if there is no such number we set $N(\omega) = n$. We observe the equivalence of events,

$$\left\{\omega : \max_{1\le k\le n} |S_k| > a\right\} = \left\{\omega : S_{N(\omega)}^2 > a^2\right\},$$

and from the Markov inequality

$$P\left(\max_{1\le k\le n} |S_k| > a\right) \le \frac{1}{a^2}\mathbb{E}[S_N^2].$$

We need to estimate the right hand side. If we could replace

$$\mathbb{E}[S_N^2] \qquad \text{by} \qquad \mathbb{E}[S_n^2] = \text{Var}[S_n] = \sum_{i=1}^n \sigma_i^2,$$

then we would be done.

The trick is to show that $\mathbb{E}[S_N^2] \le \mathbb{E}[S_n^2]$ by using conditional expectations. If

$$\mathbb{E}[S_N^2|N = k] \le \mathbb{E}[S_n^2|N = k]$$

for all $1 \le k \le n$ then the inequality holds, since we have then an inequality between *random variables* $\mathbb{E}[S_N^2|N] \le \mathbb{E}[S_n^2|N]$, and applying expectations on both sides gives the desired result.

For $k = n$, the identity

$$\mathbb{E}[S_N^2|N = n] = \mathbb{E}[S_n^2|N = n],$$

hold trivially. Otherwise, we write

$$\begin{aligned}\mathbb{E}[S_n^2|N = k] &= \mathbb{E}[S_k^2|N = k] + \mathbb{E}[(X_{k+1} + \cdots + X_n)^2|N = k] \\ &\quad + 2\mathbb{E}[S_k(X_{k+1} + \cdots + X_n)|N = k]\end{aligned}$$

The first term on the right hand side equals $\mathbb{E}[S_N^2|N = k]$, whereas the second terms is non-negative. Remains the third term for which we remark that $X_{k+1} + \cdots + X_n$ is independent of both S_k and N, and by the previous lemma,

$$\mathbb{E}[S_k(X_{k+1} + \cdots + X_n)|N = k] = \mathbb{E}[S_k|N = k]\,\mathbb{E}[X_{k+1} + \cdots + X_n] = 0.$$

Putting it all together,

$$\mathbb{E}[S_n^2|N = k] \ge \mathbb{E}[S_N^2|N = k].$$

Since this holds for all k's we have thus shown that

$$\mathbb{E}[S_N^2] \le \mathbb{E}[S_n^2] = \sum_{i=1}^n \sigma_i^2,$$

which completes the proof. ■

Chapter 7

Limit theorems

Throughout this section we will assume a probability space $(\Omega, \mathscr{F}, P)$, in which is defined an infinite sequence of random variables (X_n) and a random variable X. The fact that for every infinite sequence of distributions it is possible to construct a probability space with a corresponding sequence of random variables is a non-trivial fact, whose proof is due to Kolmogorov (see for example Billingsley).

7.1 Convergence of sequences of random variables

Definition 7.1 The sequence (X_n) is said to converge to X ***almost-surely*** *(or, w.p. 1) if*

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

We write $X_n \xrightarrow{a.s} X$.

It is often easier to express this mode of convergence using its complement. $X_n(\omega)$ fails to converge to $X(\omega)$ if there exists an $\epsilon > 0$ such that $|X_n(\omega) - X(\omega)| \geq \epsilon$ holds for infinitely many values of n. Let us denote the following family of events,

$$B_n^\epsilon = \{\omega : |X_n(\omega) - X(\omega)| \geq \epsilon\}.$$

Thus, $X_n(\omega)$ does not converge almost-surely to $X(\omega)$ is there exists an $\epsilon > 0$ such that

$$P(B_n^\epsilon \text{ i.o.}) > 0,$$

and inversely, it does converge almost-surely to $X(\omega)$ if for all $\epsilon > 0$

$$P(B_n^\epsilon \text{ i.o.}) = P\left(\limsup_n B_n^\epsilon\right) = 0.$$

Definition 7.2 The sequence (X_n) is said to converge to X ***in the mean-square*** *if*

$$\lim_{n\to\infty} \mathbb{E}\left[|X_n - X|^2\right] = 0.$$

We write $X_n \xrightarrow{m.s} X$.

Definition 7.3 The sequence (X_n) is said to converge to X ***in probability*** *if for every $\epsilon > 0$,*

$$\lim_{n\to\infty} P\left(\{\omega : \ |X_n(\omega) - X(\omega)| > \epsilon\}\right) = 0.$$

We write $X_n \xrightarrow{\text{Pr}} X$.

Definition 7.4 The sequence (X_n) is said to converge to X ***in distribution*** *if for every $a \in \mathbb{R}$,*

$$\lim_{n\to\infty} F_{X_n}(a) = F_X(a),$$

i.e., if the sequence of distribution functions of the X_n converges point-wise to the distribution function of X. We write $X_n \xrightarrow{D} X$.

Comment: Note that the first three modes of convergence require that the sequence (X_n) and X are all defined on a joint probability space. Since convergence in distribution only refers to distribution, each variable could, in principle, belong to a “separate world”.

The first question to be addressed is whether there exists a hierarchy of modes of convergence. We want to know which modes of convergence imply which. The answer is that both almost-sure and mean-square convergence imply convergence in probability, which in turn implies convergence in distribution. On the other hand, almost-sure and mean-square convergence do not imply each other.

Proposition 7.1 Almost-sure convergence implies convergence in probability.

Proof: As we have seen, almost sure convergence means that for every $\epsilon > 0$

$$P\left(\{\omega : \ |X_n(\omega) - X(\omega)| > \epsilon \text{ i.o.}\}\right) = 0.$$

Define the family of events

$$B_n^\epsilon = \{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}.$$

$X_n \to X$ almost-surely if for every $\epsilon > 0$

$$P\left(\limsup_{n\to\infty} B_n^\epsilon\right) = 0.$$

By the Fatou lemma,

$$\limsup_{n\to\infty} P(B_n^\epsilon) \le P\left(\limsup_{n\to\infty} B_n^\epsilon\right) = 0,$$

from which we deduce that

$$\lim_{n\to\infty} P(B_n^\epsilon) = \lim_{n\to\infty} P\left(\{\omega : \ |X_n(\omega) - X(\omega)| > \epsilon\}\right) = 0,$$

i.e., $X_n \to X$ in probability. ■

Proposition 7.2 Mean-square convergence implies convergence in probability.

Proof: This is an immediate consequence of the Markov inequality, for

$$P(|X_n - X| > \epsilon) = P(|X_n - X|^2 > \epsilon^2) \le \frac{\mathbb{E}|X_n - X|^2}{\epsilon^2},$$

and the right-hand side converges to zero. ■

Proposition 7.3 Mean-square convergence does not imply almost sure convergence.

Proof: All we need is a counter example. Consider a family of independent Bernoulli variables X_n with atomistic distributions,

$$p_{X_n}(x) = \begin{cases} 1/n & x = 1 \\ 1 - 1/n & x = 0 \end{cases},$$

and set $X = 0$. We claim that $X_n \to X$ in the mean square, as

$$\mathbb{E}|X_n - X|^2 = \mathbb{E}[X_n] = \frac{1}{n} \to 0.$$

On the other hand, it does not converge to X almost surely, as for $\epsilon = 1/2$,

$$\sum_{n=1}^{\infty} P(|X_n - X| > \epsilon) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty,$$

and by the second lemma of Borel-Cantelli,

$$P(|X_n - X| > \epsilon \ i.o.) = 1.$$

■

Proposition 7.4 Almost-sure convergence does not imply mean square convergence.

Proof: Again we construct a counter example, with

$$p_{X_n}(x) = \begin{cases} 1/n^2 & x = n^3 \\ 1 - 1/n^2 & x = 0 \end{cases},$$

and again $X = 0$. We immediately see that X_n does not converge to X in the mean square, since

$$\mathbb{E}|X_n - X|^2 = \mathbb{E}[X_n^2] = \frac{n^6}{n^2} = \infty.$$

It remains to show that $X_n \to X$ almost-surely. For every $\epsilon > 0$, and n sufficiently large, $P(|X_n| > \epsilon) = 1/n^2$, i.e., for every $\epsilon > 0$,

$$\sum_{n=1}^{\infty} P(|X_n - X| > \epsilon) < \infty,$$

and by the first lemma of Borel-Cantelli,

$$P(|X_n - X| > \epsilon \ i.o.) = 0.$$

■

Comment: In the above example $X_n \to X$ in probability, so that the latter does not imply convergence in the mean square either.

Proposition 7.5 Convergence in probability implies convergence in distribution.

Proof: Let $a \in R$ be given, and set $\epsilon > 0$. On the one hand

$$\begin{aligned} F_{X_n}(a) &= P\left(X_n \le a, X \le a+\epsilon\right) + P\left(X_n \le a, X > a+\epsilon\right) \\ &= P\left(X_n \le a | X \le a+\epsilon\right) P\left(X \le a+\epsilon\right) + P\left(X_n \le a, X > a+\epsilon\right) \\ &\le P\left(X \le a+\epsilon\right) + P\left(X_n < X-\epsilon\right) \\ &\le F_X(a+\epsilon) + P\left(|X_n - X| > \epsilon\right), \end{aligned}$$

where we have used the fact that if A implies B then $P(A) \le P(B)$). By a similar argument

$$\begin{aligned} F_X(a-\epsilon) &= P\left(X \le a-\epsilon, X_n \le a\right) + P\left(X \le a-\epsilon, X_n > a\right) \\ &= P\left(X \le a-\epsilon | X_n \le a\right) P\left(X_n \le a\right) + P\left(X \le a-\epsilon, X_n > a\right) \\ &\le P\left(X_n \le a\right) + P\left(X < X_n - \epsilon\right) \\ &\le F_{X_n}(a) + P\left(|X_n - X| > \epsilon\right), \end{aligned}$$

Thus, we have obtained that

$$F_X(a-\epsilon) - P\left(|X_n - X| > \epsilon\right) \le F_{X_n}(a) \le F_X(a+\epsilon) + P\left(|X_n - X| > \epsilon\right).$$

Taking now $n \to \infty$ we have

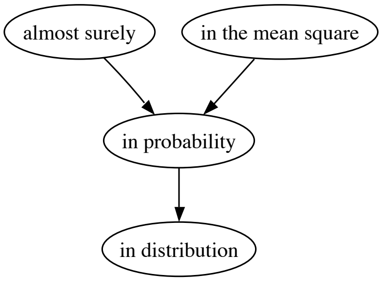
$$F_X(a-\epsilon) \le \liminf_{n\to\infty} F_{X_n}(a) \le \limsup_{n\to\infty} F_{X_n}(a) \le F_X(a+\epsilon).$$

Finally, since this inequality holds for any $\epsilon > 0$ we conclude that

$$\lim_{n\to\infty} F_{X_n}(a) = F_X(a).$$

■

To conclude, the various modes of convergence satisfy the following scheme:



✎ *Exercise 7.1* Prove that if X_n converges in distribution to a constant c, then X_n converges in probability to c.

✎ *Exercise 7.2* Prove that if X_n converges to X in probability then it has a subsequence that converges to X almost-surely.

7.2 The weak law of large numbers

Theorem 7.1 (Weak law of large numbers) Let X_n be a sequence of independent identically distributed random variables on a probability space $(\Omega, \mathscr{F}, P)$. Set $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$. Define the sequence of cummulative averages,

$$S_n = \frac{X_1 + \cdots + X_n}{n}.$$

Then, S_n converges to μ in probability, i.e., for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|S_n - \mu| > \epsilon) = 0.$$

Comments:

① The assumption that the variance is finite is not required; it only simplifies the proof.

② Take the particular case where

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

Then,

$$S_n = \text{fraction of times } \omega \in A.$$

The weak law of large numbers states that the fraction of times the outcome is in a given set converges in probability to $E[X_1]$, which is the probability of this set, $P(A)$.

Proof: This is an immediate consequence of the Chebyshev inequality, for by the additivity of the expectation and the variance (for independent random variables),

$$\mathbb{E}[S_n] = \mu \qquad \text{and} \qquad \text{Var}[S_n] = \frac{\sigma^2}{n}.$$

Then,

$$P(|S_n - \mu| > \epsilon) \leq \frac{\text{Var}[S_n]}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0.$$

■

Comment: the first proof is due to Jacob Bernoulli (1713), who proved it for the particular case of binomial variables.

7.3 The central limit theorem

Theorem 7.2 (Central limit theorem) Let (X_n) be a sequence of i.i.d. random variables with $\mathbb{E}[X_i] = 0$ and $\text{Var}[X_i] = 1$. Then, the sequence of random variables

$$S_n = \frac{X_1 + \cdots + X_n}{\sqrt{n}}$$

converges in distribution to a random variables $X \sim \mathcal{N}(0,1)$. That is,

$$\lim_{n\to\infty} P(S_n \leq a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-y^2/2}\, dy.$$

Comments:

① If $\mathbb{E}[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2$ then the same applies for

$$S_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}.$$

② The central limit theorem (CLT) is about a running average rescaled by a factor of $\sqrt{n}$. If we denote by Y_n the running average,

$$Y_n = \frac{X_1 + \cdots + X_n}{n},$$

then the CLT states that

$$P\left(Y_n \le \frac{a}{\sqrt{n}}\right) \sim \Phi(a),$$

i.e., it provides an estimate of the distribution of Y_n at distances $O(n^{-1/2})$ from its mean. It is a theorem about ***small deviations*** from the mean. There exist more sophisticated theorems about the distribution of Y_n far from the mean, part of the so-called theory of ***large deviations***.

③ There are many variants of this theorem.

Proof: We will use the following fact, which we won't prove: if the sequence of moment generating functions $M_{X_n}(t)$ of a sequence of random variables (X_n) converges for every t to the moment generating function $M_X(t)$ of a random variable X, then X_n converges to X in distribution. In other words,

$$M_{X_n}(t) \to M_X(t) \text{ for all } t \text{ implies that } X_n \xrightarrow{D} X.$$

Thus, we need to show that the moment generating functions of the S_n's tends as $n \to \infty$ to $\exp(t^2/2)$, which is the moment generating function of a standard normal variable.

Recall that the PDF of a sum of two random variables is the convolution of their PDF, but the moment generating function of their sum is the product of the their moment generating function. Inductively,

$$M_{X_1+X_2+\ldots,+X_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) = [M_{X_1}(t)]^n,$$

where we have used the fact that they are i.i.d., Now, if a random variable Y has a moment generating function M_Y, then

$$M_{Y/a}(t) = \int_{\mathbb{R}} e^{ty} f_{Y/a}(y)\, dy,$$

but since $f_{Y/a}(y) = a\, f_Y(ay)$ we get that

$$M_{Y/a}(t) = a \int_{\mathbb{R}} e^{ty} f_Y(ay)\, dy = \int_{\mathbb{R}} e^{aty/a} f_Y(ay)\, d(ay) = M_Y(t/a),$$

from which we deduce that

$$M_{S_n}(t) = \left[M_{X_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n .$$

Take the logarithm of both sides, and write the left hand side explicitly,

$$\log M_{S_n}(t) = n \log \int_{\mathbb{R}} e^{tx/\sqrt{n}} f_{X_1}(x)\, dx.$$

Taylor expanding the exponential about $t = 0$ we have,

$$\begin{aligned} \log M_{S_n}(t) &= n \log \int_{\mathbb{R}} \left(1 + \frac{tx}{\sqrt{n}} + \frac{t^2x^2}{2n} + \frac{t^3x^3}{6n^{3/2}} e^{\xi x/\sqrt{n}}\right) f_{X_1}(x)\, dx \\ &= n \log\left(1 + 0 + \frac{t^2}{2n} + O(n^{-3/2})\right) \\ &= n\left(\frac{t^2}{2n} + O(n^{-3/2})\right) \to \frac{t^2}{2}. \end{aligned}$$

■

Example: Suppose that an experimentalist wants to measure some quantity. He knows that due to various sources of errors, the result of every single measurement is a random variable, whose mean μ is the correct answer, and the variance of his measurement is σ^2. He therefore performs independent measurements and averages the results. How many such measurements does he need to perform to be sure, within 95% certainty, that his estimate does not deviate from the true result by $\sigma/4$?

The question we're asking is how large should n be in order for the inequality

$$P\left(\mu - \frac{\sigma}{4} \le \frac{1}{n}\sum_{k=1}^{n} X_k \le \mu + \frac{\sigma}{4}\right) \ge 0.95$$

to hold. This is equivalent to asking what should n be for

$$P\left(-\frac{\sqrt{n}}{4} \le \frac{1}{\sqrt{n}}\sum_{k=1}^{n} \frac{X_k - \mu}{\sigma} \le \frac{\sqrt{n}}{4}\right) \ge 0.95.$$

By the central limit theorem the right hand side is, for large n, approximately

$$\frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{n}/4} e^{-y^2/2}\, dy,$$

which turns out to be larger than 0.95 for ≥ 62.

The problem with this argument that it uses the assumption that "n is large", but it is not clear what large is. Is $n = 62$ sufficiently large for this argument to hold? This problem could have been solved without this difficulty but resorting instead to the Chebyshev inequality:

$$P\left(-\frac{\sqrt{n}}{4} \le \frac{1}{\sqrt{n}}\sum_{k=1}^{n} \frac{X_k - \mu}{\sigma} \le \frac{\sqrt{n}}{4}\right) = 1 - P\left(\left|\frac{1}{\sqrt{n}}\sum_{k=1}^{n} \frac{X_k - \mu}{\sigma}\right| \ge \frac{\sqrt{n}}{4}\right)$$
$$\ge 1 - \frac{16}{n},$$

and the right hand side is larger than 0.95 if

$$n \ge \frac{16}{0.05} = 320.$$

▲▲▲

Example: The number of students X who are going to fail in the exam is a Poisson variable with mean 100, i.e, $X \sim \text{Poi}(100)$. I am going to admit that the exam was too hard if more than 120 student fail. What is the probability for it to happen?

We know the exact answer,

$$P(X \ge 120) = e^{-100} \sum_{k=120}^{\infty} \frac{100^k}{k!},$$

which is a quite useless expression. Let's base our estimate on the central limit theorem as follows: a Poisson variable with mean 100 can be expressed as the sum of one hundred independent variables $X_k \sim \text{Poi}(1)$ (the sum of independent Poisson variables is again a Poisson variable), that is $X = \sum_{k=1}^{100} X_k$. Now,

$$P(X \geq 120) = P\left(\frac{1}{\sqrt{100}} \sum_{k=1}^{100} \frac{X_k - 1}{1} \geq \frac{20}{10}\right),$$

which by the central limit theorem equals approximately,

$$P(X \geq 120) \approx \frac{1}{\sqrt{2\pi}} \int_2^\infty e^{-y^2/2}\, dy \approx 0.228.$$

▲▲▲

Example: Let us examine numerically a particular example. Let $X_i \sim \mathcal{Exp}(1)$ be independent exponential variable and set

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - 1).$$

A sum of n independent exponential variables has distribution $\mathcal{Gamma}(n, 1)$, i.e., its pdf is

$$\frac{x^{n-1} e^{-x}}{\Gamma(n)}.$$

The density for this sum shifted by n is

$$\frac{(x+n)^{n-1} e^{-(x+n)}}{\Gamma(n)},$$

with $x > -n$ and after dividing by $\sqrt{n}$,

$$f_{S_n}(x) = \sqrt{n}\, \frac{(\sqrt{n}x + n)^{n-1} e^{-(\sqrt{n}x+n)}}{\Gamma(n)},$$

with $x > -\sqrt{n}$. See Figure 7.1 for a visualization of the approach of the distribution of S_n toward the standard normal distribution.

▲▲▲

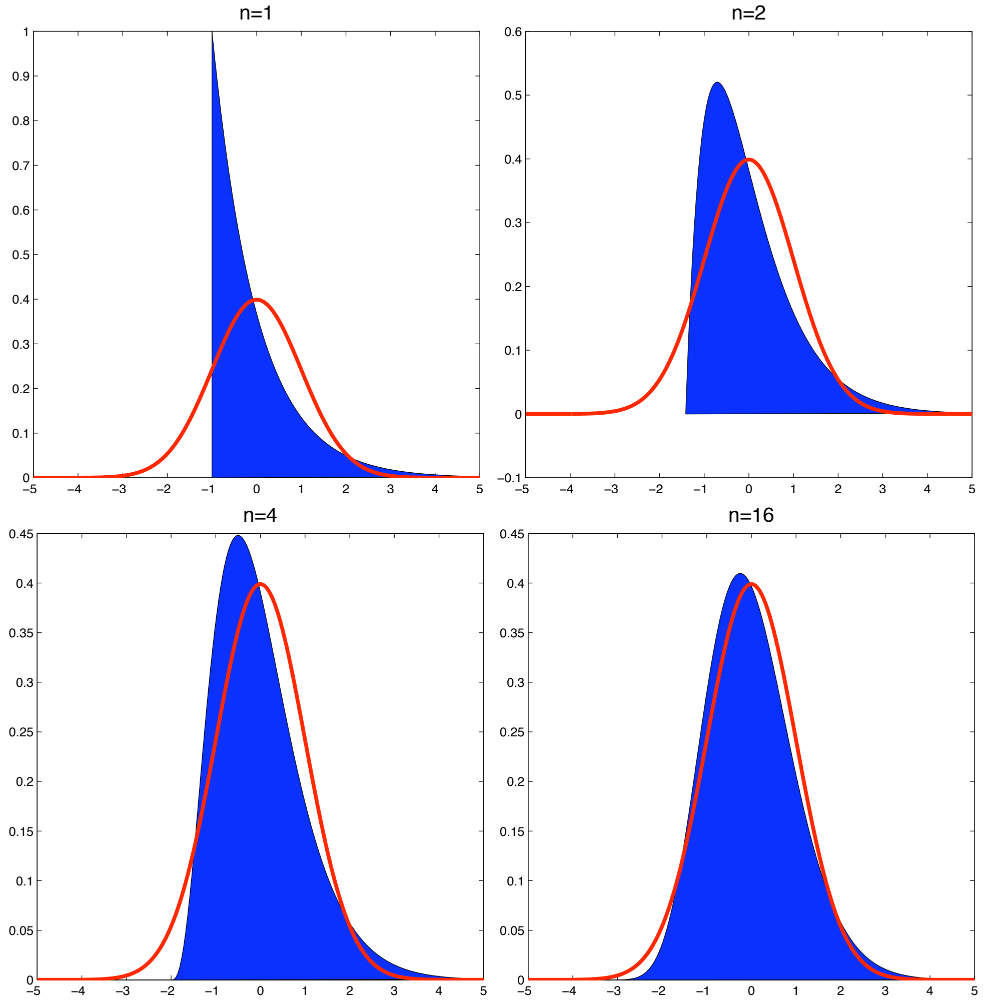


Figure 7.1: The approach of a normalized sum of $1, 2, 4$ and 16 exponential random variables to the normal distribution.

7.4 The strong law of large numbers

Our next limit theorem is the strong law of large number, which states that the running average of a sequence of i.i.d. variables converges to the mean almost-surely (thus strengthening the weak law of large number, which only provides convergence in probability).

Theorem 7.3 (Strong law of large numbers) Let (X_n) *be an i.i.d. sequence of random variables with* $\mathbb{E}[X_i] = 0$ *and* $\operatorname{Var}[X_i] = \sigma^2$*, then with probability one,*

$$\frac{X_1 + \cdots + X_n}{n} \to 0.$$

Proof: Set $\epsilon > 0$, and consider the following sequence of events:

$$A_n = \left\{ \max_{1 \le k \le n} \left| \sum_{i=1}^{k} \frac{X_i}{i} \right| > \epsilon \right\}.$$

From Komogorov's inequality,

$$P(A_n) \le \frac{1}{\epsilon^2} \sum_{k=1}^{n} \frac{\sigma^2}{k^2}.$$

The (A_n) are an increasing sequence hence, by the continuity of the probability,

$$\lim_{n\to\infty} P(A_n) = P\left(\lim_{n\to\infty} A_n \right) = P\left(\max_{1 \le k} \left| \sum_{i=1}^{k} \frac{X_i}{i} \right| > \epsilon \right) \le \frac{C\sigma^2}{\epsilon^2},$$

or equivalently,

$$P\left(\max_{1 \le k} \left| \sum_{i=1}^{k} \frac{X_i}{i} \right| \le \epsilon \right) \ge 1 - \frac{C\sigma^2}{\epsilon^2}$$

Since this holds for all $\epsilon > 0$, taking $\epsilon \to \infty$ results in

$$P\left(\max_{1 \le k} \left| \sum_{i=1}^{k} \frac{X_i}{i} \right| < \infty \right) = 1.$$

from which we infer that

$$P\left(\sum_{i=1}^{\infty} \frac{X_i}{i} < \infty\right) = 1,$$

It is then a consequence of a lemma due to Kronecker that

$$\sum_{i=1}^{\infty} \frac{X_i}{i} < \infty \qquad \text{implies} \qquad \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k = 0.$$

■

Lemma 7.1 *Let $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ be independent. Then,*

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Proof: Rather then working with the convolution formula, we use the moment generating functions. We have seen that

$$\begin{aligned} M_{X_1}(t) &= \exp\left(\frac{\sigma_1^2}{2}t^2 + \mu_1 t\right) \\ M_{X_2}(t) &= \exp\left(\frac{\sigma_2^2}{2}t^2 + \mu_2 t\right). \end{aligned}$$

Since X_1, X_2 are independent,

$$M_{X_1+X_2}(t) = \mathbb{E}\left[e^{X_1+X_2}\right] = \mathbb{E}\left[e^{X_1}\right]\mathbb{E}\left[e^{X_2}\right],$$

hence,

$$M_{X_1+X_2}(t) = \exp\left(\frac{\sigma_1^2 + \sigma_2^2}{2}t^2 + (\mu_1 + \mu_2)t\right),$$

from which we identify $X_1 + X_2$ as a normal variable with the desired mean and variance. ■

Corollary 7.1 *Let $Z_1, Z_2 \cdots \sim N(0, 1)$ be independent. Then for every $n \in \mathbb{N}$,*

$$\frac{Z_1 + \cdots + Z_n}{\sqrt{n}} \sim N(0, 1).$$

Theorem 7.4 *Let (X_n) be a sequence of i.i.d. RVs with $\mathbb{E}[X_i] = 0$, $\mathrm{Var}[X_i] = 1$ and $\mathbb{E}|X_i|^3 < \infty$; let S_n be defined as above. Let $Z_1, Z_2 \cdots \sim N(0, 1)$ be independent. Then, for every non-negative function $\varphi(x)$, uniformly three-times differentiable, and with compact support,*

$$\lim_{n\to\infty}\left|\mathbb{E}\,\varphi\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) - \mathbb{E}\,\varphi\left(\frac{Z_1 + \cdots + Z_n}{\sqrt{n}}\right)\right| = 0.$$

Comments: Suppose we could have taken $\varphi(x) = I_{[a,b]}(x)$. Then,

$$\mathbb{E}\varphi(X) = P\left(a \le X \le b\right),$$

and the theorem would imply that

$$\lim_{n\to\infty}\left|P\left(a \le \frac{X_1+\cdots+X_n}{\sqrt{n}} \le b\right) - \frac{1}{\sqrt{2\pi}}\int_a^b e^{-y^2/2}\,dy\right| = 0.$$

We are somewhat more restricted by the requirement that φ be three time differentiable, but we can approximate the indicator function by a sequence of smoother functions.

Proof: We start with

$$\begin{aligned}
&\mathbb{E}\,\varphi\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right) - \mathbb{E}\,\varphi\left(\frac{Z_1+\cdots+Z_n}{\sqrt{n}}\right)\\
&\quad= \mathbb{E}\,\varphi\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right) - \mathbb{E}\,\varphi\left(\frac{X_1+\cdots+X_{n-1}+Z_n}{\sqrt{n}}\right)\\
&\quad+ \mathbb{E}\,\varphi\left(\frac{X_1+\cdots+X_{n-1}+Z_n}{\sqrt{n}}\right) - \mathbb{E}\,\varphi\left(\frac{X_1+\cdots+X_{n-2}+Z_{n-1}+Z_n}{\sqrt{n}}\right)\\
&\quad+\ldots,
\end{aligned}$$

so that

$$\begin{aligned}
&\left|\mathbb{E}\,\varphi\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right) - \mathbb{E}\,\varphi\left(\frac{Z_1+\cdots+Z_n}{\sqrt{n}}\right)\right|\\
&\quad\le \sum_{i=1}^{n}\left|\mathbb{E}\,\varphi\left(\frac{X_1+\cdots+X_i+Z_{i+1}+\cdots+Z_n}{\sqrt{n}}\right) - \mathbb{E}\,\varphi\left(\frac{X_1+\cdots+X_{i-1}+Z_i+\cdots+Z_n}{\sqrt{n}}\right).\right|
\end{aligned}$$

Each of the summands can be estimated as follows:

$$\begin{aligned}
&\left|\mathbb{E}\,\varphi\left(\frac{X_1+\cdots+X_i+Z_{i+1}+\cdots+Z_n}{\sqrt{n}}\right) - \mathbb{E}\,\varphi\left(\frac{X_1+\cdots+X_{i-1}+Z_i+\cdots+Z_n}{\sqrt{n}}\right)\right|\\
&\quad= \left|\mathbb{E}\,\varphi\left(\frac{X_i}{\sqrt{n}}+u_n\right) - \mathbb{E}\,\varphi\left(\frac{Z_i}{\sqrt{n}}+u_n\right)\right|,
\end{aligned}$$

where the u_n represent all the other terms. We then Taylor expand up to the third term, and replace the expectation by

$$\mathbb{E}[\cdot] = \mathbb{E}[\mathbb{E}[\cdot|u_n]].$$

Using the fact that X_n and Z_n have the same first and second moments, and the uniform boundedness of the third derivative of φ, we get

$$\left|\mathbb{E}\,\varphi\left(\frac{X_i}{\sqrt{n}} + u_n\right) - \mathbb{E}\,\varphi\left(\frac{Z_i}{\sqrt{n}} + u_n\right)\right| \leq \frac{C}{n\sqrt{n}}.$$

Substituting above we conclude that

$$\left|\mathbb{E}\,\varphi\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) - \mathbb{E}\,\varphi\left(\frac{Z_1 + \cdots + Z_n}{\sqrt{n}}\right)\right| \leq \frac{C}{\sqrt{n}} \to 0.$$

■

Chapter 8

Markov chains

8.1 Definitions and basic properties

Let S be a set that could be either finite or countable. S is called the ***state space***; elements of S are denoted by indices $i, j, \dots, i_0, i_1, \dots$.

Definition 8.1 A **Markov chain** *over S is a sequence of random variables $X_1, X_2, \dots,$ satisfying*

$$P(X_n = i_n | X_0 = i_0, \dots, X_{n-1} = i_{n-1}) = P(X_n = i_n | X_{n-1} = i_{n-1}).$$

That is, "the present determines the distribution of the future independently of the past". The right hand side is called the **transition probability***. It depends on n, and on the two states i_{n-1} and i_n.*

Comments:

① A Markov chain is an instance of a ***stochastic process***.

② We often interpret the index n as ***time***. In a Markov chain time is a discrete parameter; there are also Markov processes in continuous time (continuous-time Markov processes) and Markov processes over uncountable state spaces (e.g., Brownian motion).

③ A Markov chain is called ***time homogeneous*** if the transition probability is time-independent, i.e., if

$$P(X_n = j | X_{n-1} = i) = P(X_m = j | X_{m-1} = i) \equiv p_{i,j}.$$

> The matrix P whose entries are $p_{i,j}$ is called the ***transition matrix*** (don't be bothered by infinite matrices). It is *the probability to transition from state i to state j in a single step.*

The transition matrix P is a ***stochastic matrix***. That is, $p_{i,j} \geq 0$ for all $i, j \in S$ and

$$\sum_{j \in S} p_{i,j} = 1$$

for all $i \in S$ (note that the sums over columns do not need to be one; a stochastic matrix having this additional property, i.e., that P^T is also stochastic, is called ***doubly stochastic***.

In addition to the transition matrix, a Markov chain is also characterized by its initial distribution, which can be represented by a vector $\boldsymbol{\mu}$ with entries

$$\mu_i = P(X_0 = i).$$

Proposition 8.1 The transition distribution $\boldsymbol{\mu}$ and the transition matrix P fully determine the distribution of the process (i.e., the distribution of the sequence of random variables).

Proof: A sequence of random variables is fully determined by all its finite-dimensional marginal distributions. Using the product formula,

$$P(A \cap B \cap C \cap \cdots \cap Z) = P(A|B \cap C \cap \cdots \cap Z)\, P(B|C \cap \cdots \cap Z) \cdots P(Z),$$

along with the Markov property, we have for every n,

$$\begin{aligned} P(X_0 = i_0, \dots, X_n = i_n) &= P(X_n = i_n | X_0 = i_0, \dots, X_{n-1} = i_{n-1}) \\ &\qquad \times P(X_{n-1} = i_{n-1} | X_0 = i_0, \dots, X_{n-2} = i_{n-2}) \cdots \\ &\qquad \times P(X_1 = i_1 | X_0 = i_0) P(X_0 = i_0) \\ &= \mu_{i_0} p_{i_0,i_1} \cdots p_{i_{n-2},i_{n-1}} p_{i_{n-1},i_n}. \end{aligned}$$

■

Summing the above identity over all values of $i_1, \dots, i_{n-1}$ we obtain,

$$P(X_n = i_n | X_0 = i_0) = (P^n)_{i_0,i_n} \equiv p^{(n)}_{i_0,i_n}.$$

Thus, the n-th power of the transition matrix is the n-step transition matrix; $p_{i,j}^{(n)}$ is the probability to transition from state i to state j in n steps. It follows at once that for all n, m,

$$p_{i,j}^{(n+m)} = \sum_{k\in S} p_{i,k}^{(n)} p_{k,j}^{(m)}.$$

It is also customary to define

$$p_{i,j}^{(0)} = \delta_{i,j}.$$

Finally, if $\boldsymbol{\mu}^{(n)}$ denotes the distribution at time n, i.e.,

$$\mu_j^{(n)} = P(X_n = j),$$

then

$$\mu_j^{(n)} = \sum_{i\in S} P(X_n = j | X_0 = i) P(X_0 = i) = \sum_{i\in S} \mu_i p_{i,j},$$

namely,

$$\boldsymbol{\mu}^{(n)} = \boldsymbol{\mu} P^n,$$

where we interpret $\boldsymbol{\mu}$ as a row vector.

Example: Consider a Markov chain with finite state space. We can represent the states in S as nodes of a graph. Every two nodes i, j are joined by a directed edge labeled by the probability to transition from i to j is a single step. If $p_{i,j} = 0$ then no edge is drawn. This picture is useful for imagining a ***simulation*** of the Markov chain. The initial state is drawn from the distribution $\boldsymbol{\mu}$ by "tossing a coin" (well, more precisely by sampling from $\mathcal{U}(0,1)$). Then, we transition from this state i_0 by drawing the next state from the distribution $p_{i_0,j}$, and so on. ▲▲▲

Example: Another example is ***random walk on*** $\mathbb{Z}^d$. For $d = 1$ we start, say, at the origin, $X_0 = 0$. Then

$$X_{n+1} = \begin{cases} X_n - 1 & \text{w.p. } 1/2 \\ X_n + 1 & \text{w.p. } 1/2. \end{cases}$$

In higher dimensions we will assume, for technical reasons, that the process moves at each step along all d axes. Thus, for $d = 2$, we have

$$X_{n+1} = X_n + (\pm 1, \pm 1),$$

each four transitions having equal probability. ▲▲▲

8.2 Transience and recurrence

We define now the following probability:

$$f_{i,j}^{(n)} = P(X_n = j, X_{n-1} \neq j, \dots, X_1 \neq j | X_0 = i).$$

It is the probability that the process arrives at state j at time n *for the first time*, given that it started in state i. Note that the events "the process arrived to state j for the first time at time n", with n varying, are mutually disjoint, and their countable union is the event that the process arrived to state j eventually. That is,

$$f_{i,j} = \sum_{n=1}^{\infty} f_{i,j}^{(n)}$$

is the probability that a process that started in state i will eventually get to state j. In particular, $f_{j,j}$ is the probability that a process starting in state j will eventually return to its stating point.

Definition 8.2 A state $j \in S$ is called ***persistent*** *if a process stating in this state has probability one to eventually return to it, i.e., if $f_{j,j} = 1$. Otherwise, it is called* ***transient***.

Suppose that the process started in state i. The probability that it visited state j for the first time at time n_1, for the second time at time n_2, and so on, until the k-th time at time n_k is

$$f_{i,j}^{(n_1)} f_{j,j}^{(n_2-n_1)} \cdots f_{j,j}^{(n_k-n_{k-1})}.$$

The probability that there were eventually k visits in j is obtained by summing over all possible values of the $n_1, \dots, n_k$, giving,

$$P(\text{at least } k \text{ visits in } j | X_0 = i) = f_{i,j} f_{j,j}^k.$$

Letting $k \to \infty$ we get that the probability of having infinitely many visits in state j is

$$P(X_n = j \text{ i.o.} | X_0 = i) = \begin{cases} f_{i,j} & f_{j,j} = 1 \\ 0 & f_{j,j} < 1. \end{cases}.$$

Taking $j = i$, we get that the process returns to its stating point infinitely many times with a probability that is either zero or one (yet another zero-one law),

$$P(X_n = i \text{ i.o.} | X_0 = i) = \begin{cases} 1 & f_{i,i} = 1 \\ 0 & f_{i,i} < 1. \end{cases}.$$

This means that

> One guaranteed return is equivalent to infinitely many guaranteed returns.

The question is how to identify whether a state is recurrent or transient given the Markov transition matrix. This is settled by the following theorem:

Theorem 8.1 The following characterization of persistence are equivalent:

$$f_{i,i} = 1 \quad \Leftrightarrow \quad P(X_n = i \ i.o.|X_0 = i) = 1 \quad \Leftrightarrow \quad \sum_{n=1}^{\infty} p_{i,i}^{(n)} = \infty.$$

Similarly for transience,

$$f_{i,i} < 1 \quad \Leftrightarrow \quad P(X_n = i \ i.o.|X_0 = i) = 0 \quad \Leftrightarrow \quad \sum_{n=1}^{\infty} p_{i,i}^{(n)} < \infty.$$

Proof: It is enough to show the correctness for transience. We have already seen that transience is equivalent to the probability of infinitely many returns being zero. By the first Borel-Cantelli lemma the finiteness of the series implies as well that the probability of infinitely many returns is zero. It remains to show that transience implies the finiteness of the series. We have

$$\begin{aligned}
p_{i,j}^{(n)} &= \sum_{s=0}^{n-1} P(X_n = j, X_{n-s} = j, X_{n-s-1} \neq j, \ldots, X_1 \neq j | X_0 = i) \\
&= \sum_{s=0}^{n-1} P(X_n = j | X_{n-s} = j) P(X_{n-s} = j, X_{n-s-1} \neq j, \ldots, X_1 \neq j | X_0 = i) \\
&= \sum_{s=0}^{n-1} p_{j,j}^{(s)} f_{i,j}^{(n-s)}.
\end{aligned}$$

This decomposition is called a ***first passage time decomposition***. We then set $i = j$ and sum over n,

$$\sum_{n=1}^{m} p_{i,i}^{(n)} = \sum_{n=1}^{m} \sum_{s=0}^{n-1} p_{i,i}^{(s)} f_{i,i}^{(n-s)} = \sum_{s=0}^{m-1} \sum_{n=s+1}^{m} p_{i,i}^{(s)} f_{i,i}^{(n-s)} = \sum_{s=0}^{m-1} p_{i,i}^{(s)} \sum_{n=1}^{m-s} f_{i,i}^{(n)} \leq f_{i,i} \sum_{s=0}^{m} p_{i,i}^{(s)}$$

This means that

$$\sum_{n=1}^{m} p_{i,i}^{(n)} \le f_{i,i} + f_{i,i} \sum_{s=1}^{m} p_{i,i}^{(s)},$$

where we have used the fact that $f_{i,i}(0) = 1$, or,

$$(1 - f_{i,i}) \sum_{n=1}^{m} p_{i,i}^{(n)} \le f_{i,i}.$$

If $f_{i,i} < 1$ then we have a uniform bound on the series. ■

Example: Random walks in $\mathbb{Z}^d$. Since

$$p_{0,0}^{(n)} \sim \frac{1}{n^{d/2}},$$

then the origin (and any other state by symmetry) is recurrent in dimensions one and two and transient in dimensions higher than two. ▲▲▲

Definition 8.3 A Markov chain is called **irreducible** *if, loosely speaking, it is possible to reach every state from every other state. More precisely, if for every* $i, j \in S$ *there exists an* n *for which*

$$p_{i,j}^{(n)} > 0.$$

Theorem 8.2 In an irreducible Markov chain one of the following holds:

① *All the states are transient, and for all* i, j,

$$P\left(\cup_{j\in S} \{X_n = j \; i.o.\} | X_0 = i\right) = 0,$$

and

$$\sum_{n=1}^{\infty} p_{i,j}^{(n)} < \infty.$$

② *All the states are persistent, and for all* i, j,

$$P\left(\cap_{j\in S} \{X_n = j \; i.o.\} | X_0 = i\right) = 1,$$

and

$$\sum_{n=1}^{\infty} p_{i,j}^{(n)} = \infty.$$

Comment: This implies that provided that one state is persistent, irrespectively of the initial state, the chain is guaranteed to visit every state infinitely many times. This is highly non-trivial.

Proof: Let $i, j \in S$ be given. Since the chain is irreducible, then there exist m, r such that $p_{i,j}^{(m)} > 0$ and $p_{j,i}^{(r)} > 0$. For all n,

$$p_{i,i}^{(m+n+r)} \le p_{i,j}^{(m)} p_{j,j}^{(n)} p_{j,i}^{(r)}.$$

Summing over n, it follows that if j is transient so is i. That is, transience (and hence persistence) is a property of all states.

We saw that

$$p(X_n = j \text{ i.o.} | X_0 = i) = \begin{cases} 0 & f_{j,j} < 1 \\ f_{i,j} & f_{j,j} = 1. \end{cases}$$

Thus, if the chain is transient, then this is zero for all i, j. By Boole's inequality a countable union of the event of zero probability has zero probability. Finally, using again the first passage trick,

$$p_{i,j}^{(n)} = \sum_{s=0}^{n} f_{i,j}^{(s)} p_{j,j}^{(n-s)},$$

hence,

$$\sum_{n=1}^{\infty} p_{i,j}^{(n)} = \sum_{n=1}^{\infty} \sum_{s=0}^{n} f_{i,j}^{(s)} p_{j,j}^{(n-s)} = \sum_{s=0}^{\infty} f_{i,j}^{(s)} \sum_{n=s}^{\infty} p_{j,j}^{(n-s)} = f_{i,j} \sum_{n=0}^{\infty} p_{j,j}^{(n)} < \infty.$$

Conversely, suppose the chain is persistent, i.e., $f_{j,j} = 1$ for all j. We then know that

$$p(X_n = j \text{ i.o.} | X_0 = i) = f_{i,j}.$$

For every m,

$$\begin{aligned} p_{j,i}^{(m)} &= P(\{X_m = i\} \cap \{X_n = j \text{ i.o.}\} | X_0 = j) \\ &\le \sum_{n>m} P(X_m = i, X_{m+1} \neq j, \dots, X_n = j | X_0 = j) \\ &= \sum_{n>m} p_{j,i}^{(m)} f_{i,j}^{(n-m)} = f_{i,j} p_{j,i}^{(m)}. \end{aligned}$$

Since there exists an m for which the left-hand side is positive, it follows that $f_{i,j} = 1$. Then, the countable intersection of certain events has probability one. Finally, if

$$\sum_{n=1}^{\infty} p_{i,j}^{(n)} < \infty,$$

then by the first Borel-Cantelli lemma

$$p(X_n = j \text{ i.o.} | X_0 = i) = 0,$$

which is a contradiction. ■

Corollary 8.1 *A finite irreducible Markov chain is persistent.*

Proof: Since for all i,

$$\sum_{j \in S} p_{i,j}^{(n)} = 1,$$

then

$$\frac{1}{|S|} \sum_{j \in S} \sum_{n=1}^{\infty} p_{i,j}^{(n)} = \infty.$$

It follows that $\sum_{n=1}^{\infty} p_{i,j}^{(n)} = \infty$ for all i, j. ■

8.3 Long-time behavior

We now turn to discuss the ***long time behavior*** of Markov chains. Recall that if $\boldsymbol{\mu}^{(n)}$ is the distribution at time n then

$$\mu_i^{(n)} = \sum_{j \in S} \mu_j^{(n-1)} p_{j,i},$$

and inductively,

$$\mu_i^{(n)} = \sum_{j \in S} \mu_j^{(0)} (P^n)_{j,i}.$$

The question is whether the distribution has a limit as $n \to \infty$ and what it is. Note first that if a limit $\boldsymbol{\mu} = \boldsymbol{\pi}$ exists, then

$$\boldsymbol{\pi} = \boldsymbol{\pi} P,$$

i.e., it is a ***stationary distribution*** of the Markov chain; it is also a left eigenvector of P with eigenvalue 1.

Example: To get some idea on what is going on, consider a two-state Markov chain with transition matrix

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix},$$

with $0 < p, q < 1$. For such a simple matrix we can easily calculate its n-th power. The eigenvalues satisfy

$$\begin{bmatrix} 1-p-\lambda & p \\ q & 1-q-\lambda \end{bmatrix} = 0,$$

i.e.,

$$\lambda^2 - (2-p-q)\lambda + (1-p-q) = 0.$$

Setting $1 - p - q = \alpha$ we have

$$\lambda_{1,2} = \frac{1}{2}\left((1+\alpha) \pm \sqrt{(1+\alpha)^2 - 4\alpha}\right) = 1, \alpha.$$

The eigenvector that corresponds to $\lambda = 1$ is $(1,1)^T$. The eigenvector that corresponds to $\lambda = \alpha$ satisfies

$$\begin{pmatrix} q & p \\ q & p \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = 0,$$

i.e., $(p, -q)^T$. Normalizing we have

$$S = \begin{pmatrix} 1/\sqrt{2} & p/\sqrt{p^2+q^2} \\ 1/\sqrt{2} & -q/\sqrt{p^2+q^2} \end{pmatrix} \qquad \text{and} \qquad S^{-1} = -\frac{\sqrt{2}\sqrt{p^2+q^2}}{p+q}\begin{pmatrix} -q/\sqrt{p^2+q^2} & -p/\sqrt{p^2+q^2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}.$$

Since

$$P = S\Lambda S^{-1},$$

it follows that $P^n = S\Lambda^n S^{-1}$ and as $n \to \infty$

$$\lim_{n\to\infty} P^n = S\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} S^{-1} = \frac{1}{p+q}\begin{pmatrix} q & p \\ q & p \end{pmatrix}.$$

For every $\boldsymbol{\mu}^{(0)}$ we get

$$\lim_{n\to\infty} \boldsymbol{\mu}^{(n)} = \frac{(q,p)}{p+q}.$$

Thus, the distribution converges to the (unique) stationary distribution irrespectively of the initial distribution. ▲▲▲