Working with uncertainty

v0.65

Erik Štrumbelj

October 9, 2023

Contents

Preface

I would like to thank my students Jakob Božič, Benjamin Džubur, Greta Gašparac, Jan Hartman, Valter Hudovernik, Leon Hvastja, Martin Jurkovič, Maša Kljun, Uroš Kozole, Timur Kulenović, Miha Markež, Tomaž Martinčič, Andrej Miščič, Jurij Nastran, Samo Pahor, Boris Radovič, Žiga Rot, Jovana Videnović, and Luka Žontar, whose comments and suggestions helped improve this text.

Special thanks to Aljaž Zalar for reading an early version of the book and providing feedback.

Mathematical notation

$\mathbb{R}$	set of real numbers
$\mathbb{N}$	set of natural numbers
$\mathbb{Z}$	set of integers
$\mathbb{Q}$	set of rational numbers
$\triangleq$	defined as
Σ	sum
$\prod$	product
$\longrightarrow$	map
$\Longrightarrow$	implication
$\Longleftrightarrow$	equivalence
iff	if an only if
$\wedge$	and
$\vee$	or
$\neg$	not
$\emptyset$	empty set
$\{\omega\}$	a singleton set
$\in, \notin$	set membership
$\cap$	set intersection
$\cup$	set union
$\setminus$	set difference
$\subseteq$	subset
$\subset$	proper subset
$\times$	Cartesian product
A^c	set complement
$\mathcal{B}_A$	Borel $\sigma-$algebra on A
$\sigma(A)$	$\sigma-$algebra generated by A
$P(\cdot)$	Probability measure
$\lambda(\cdot)$	Lebesgue measure
$x \sim y$	object x is in relation with object y
$X \sim$	random variable X is distributed as
$\det A$	determinant of matrix A

Part I

The language of uncertainty

Chapter 1

Introduction to probability theory

Uncertainty - a lack of complete information about something - can be the result of many things. In Probability and statistics courses uncertainty is often synonymous with the randomness of the phenomena we study. However, that need not always be the case. Incomplete information could also be due to lack of understanding, errors in our knowledge or measurements, ignorance, or just plain laziness. In fact, in practice, randomness is one of the less common causes of uncertainty.

Whatever the reason might be for our uncertainty, quantitative reasoning with uncertainty requires a complete and precise description of the studied phenomena and the uncertainty. And whenever such precision is required it inevitably leads to mathematics. In our case the area of mathematics called probability theory.

Our view on probability theory will primarily be that it is a language for describing uncertainty. And our treatment of probability will be more abstract than what can typically be found in undergraduate probability courses. We will use measure theory of which probability theory is a special case. This will allow us a more general investigation of random variables and expectations and their limiting properties. However, we will also connect these more general results to special cases that we are already familiar with, such as discrete and continuous random variables.

1.1 Why do we need measure theory?

Before we can start talking about probability, we must introduce the minimal necessary structure: the set of all possible outcomes Ω (the sample space) and

a set of events $\mathcal{F}$. The set of events is in essence the set of sets that we allow ourselves to assign probabilities to. In introductory probability courses we explicitly or implicitly assume that the set of events is the power set of the set of outcomes. It turns out, however, that this can lead to probabilistic questions that we can't answer. More formally, there is sometimes no way to assign probability in a meaningful and coherent way to every possible subset of the sample space.

The following more intuitive example is due to Ross and Peköz (2007, p. 10) (for a more formal treatment see Theorem 2.1.1):

Example 1.1.1. *In this example we consider a circle of radius 1. We define a relation between points on this circle, such that two points are related if the distance between them (on the circle, in either direction) is a multiple of 1.*

We can check that this is an equivalence relation. It partitions the points on the circle into classes, such that any point in a class can be reached from any other point in steps of size 1. Additionally, every class is countably infinite - because the circumference of this circle is an irrational number, we can never return to the same point with steps of 1.

Now suppose that every class elects one of its points as its 'leader'. If we select a point X uniformly at random from the circle, what is the probability that X is the 'leader' of its family?

Define A as the event that X is the leader and define A_i and B_i as events that the point i steps clockwise and counter-clockwise, respectively, is the leader. Because every family certainly has a leader, we should have:

$$P(A) + \sum_{i=1}^{\infty} (P(A_i) + P(B_i)) = 1.$$

But since we selected X at random, $P(A)$, $P(A_i)$, and $P(B_i)$ should all have the same probability $p = P(A)$. Thus:

$$p + \sum_{i=1}^{\infty} 2p = 1.$$

However, there exists no such $0 \leq p \leq 1$ where the above holds. That is, there is no consistent way of computing $P(A)$.

We will never encounter such an example in practice. However, it is very important and somewhat surprising that not all subsets can be assigned probabilities. That is, in general, we have to give up on the assumption that all subsets can be assigned probabilities or we will not be able to construct even the most basic

uniform probability distribution. We will later use a more abstract formulation of this example to motivate the measure-theoretic construction of continuous probability spaces.

This example also illustrates how our intuition can sometimes fail us. We will encounter other similar examples when dealing with uncountable probability spaces.

1.2 Measure and probability spaces

The example from the previous section suggests that the power set might not always be the appropriate choice for the set of events. Instead, we will use a more general mathematical object called a sigma algebra (or σ-algebra):

Definition 1.2.1 (Sigma algebra)**.** A set $\mathcal{F}$ of subsets of Ω is a *σ-algebra* on Ω if it has the following three properties:

(i) $\emptyset \in \mathcal{F}$ *(contains the empty set).*

(ii) $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$*(closed under complementation).*

(iii) If $\{A_i\}$ is a countable sequence of elements of $\mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ *(closed under countable unions).*

In the context of truths and probabilistic questions, the σ-algebra starts with the following: We should always allow the question *what is the probability that nothing is true*. The remaining two requirements are implicit but intuitive - if we allow the question *is A true*, then we should also allow *is A false* (is anything other than A true). And, if we allow *is A true* and *is B true*, we should also allow *is A or B true*.

We can show that a σ-algebra is a strict generalization of the power set. That is, that every power set is a σ-algebra but not every σ-algebra is the power set - there exist *smaller* σ-algebras.

Proposition 1.2.1. *The following statements are true:*

(i) The power set 2^Ω is a σ-algebra on Ω.

(ii) There exists a set Ω and a set $\mathcal{F}$ of subsets of Ω, such that $\mathcal{F}$ is a σ-algebra on Ω and $\mathcal{F}$ is a strict subset of 2^Ω.

The proof of this proposition is left as an exercise.

Example 1.2.1. - *Consider $\Omega = \{0, 1\}$. Which of the following sets of subsets of Ω are σ-algebras on Ω?*

(a) $\mathcal{F}_1 = \{\emptyset, \{0\}, \{1\}, \Omega\}$.

(b) $\mathcal{F}_2 = \{\emptyset, \{1\}, \Omega\}$.

(c) $\mathcal{F}_3 = \{\{0\}, \{1\}, \Omega\}$.

(d) $\mathcal{F}_4 = \{\emptyset, \Omega\}$.

$\mathcal{F}_1$ *is the power set of* Ω *and thus a* σ*-algebra on* Ω.

$\mathcal{F}_2$ *is not a* σ*-algebra on* Ω*, because it does not contain the complement of* $\{1\}$. *Neither is* $\mathcal{F}_3$*, because it does not contain the empty set (or the complement of* Ω*).*

$\mathcal{F}_4$ *is a* σ*-algebra on* Ω *- it contains the empty set and it is closed for unions and intersections! It is the smallest possible* σ*-algebra, not just for this* Ω *but in general.*

We also define a generalization of the σ-algebra - an algebra - which requires only closure under finite unions. Algebras, while not of central interest, are useful, because it is often easier to check properties on an algebra and extend them to the sigma-algebra. As opposed to checking them directly on the sigma-algebra.

Definition 1.2.2 (Algebra)**.** A set $\mathcal{F}$ of subsets of Ω is an *algebra* on Ω if it has the following three properties:

(i) $\emptyset \in \mathcal{F}$ *(contains the empty set).*

(ii) $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$ *(closed under complementation).*

(iii) If $\{A_i\}$ is a finite sequence of elements of $\mathcal{F}$, then $\bigcup_{i=1}^{n} A_i \in \mathcal{F}$ *(closed under finite unions).*

Three other properties of σ-algebras follow from the above definitions and set theory:

Proposition 1.2.2. *If* $\mathcal{F}$ *is a* σ*-algebra on* Ω *then:*

(a) $\mathcal{F}$ *is an algebra on* Ω.

(b) $\Omega \in \mathcal{F}$.

(c) If $\{A_i\}$ *is a countable sequence of elements of* $\mathcal{F}$ *then* $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$.

The proof of Proposition 1.2.2 is left as an exercise.

A set of outcomes and a σ-algebra on that set together form a measurable space:

Definition 1.2.3 (Measurable space)**.** A *measurable space* is a 2-tuple $(\Omega, \mathcal{F})$ that contains a non-empty set Ω and a σ-algebra $\mathcal{F}$ on Ω.

Now that we have precisely defined the structure to assign probabilities to, we are ready to define probability itself. We start with the more general notion of measure:

Definition 1.2.4 (Measure)**.** Let $(\Omega, \mathcal{F})$ be a measurable space. A *measure* μ on $(\Omega, \mathcal{F})$ is a function $\mu : \mathcal{F} \to [0, \infty]$ with the following properties:

(i) $\mu(\emptyset) = 0$ *(null empty set).*

(ii) For every countable sequence $\{A_i\}$ of disjoint sets in $\mathcal{F}$ we have $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ *(countable additivity).*

In this general definition of a measure we allow for infinite measure. The $[0, \infty]$ represents the extended non-negative reals, extended by $\{\infty\}$, with $a + \infty = \infty$ and $a \cdot \infty = \infty$ for all $a \in [0, \infty]$.

Probability is a special case of a finite measure. In fact, many results that hold for probability are just special cases of more general results for finite measures. However, most measures in areas of practical importance are infinite, for example, measures associated with integration on $\mathbb{R}^n$.

Definition 1.2.5 (Finite measure)**.** A *finite measure* μ on $(\Omega, \mathcal{F})$ is a measure such that $\mu : \mathcal{F} \to [0, \infty)$.

Finally, the measure that will be of most interest to us - a probability measure - is a finite measure with total measure 1:

Definition 1.2.6 (Probability measure)**.** Let P be a finite measure on a measurable space $(\Omega, \mathcal{F})$. P is a *probability measure* if $P(\Omega) = 1$.

The set of outcomes, the set of events, and a probability measure form a complete and precise expression of probability:

Definition 1.2.7 (Measure space and probability space)**.** A *measure space* is 3-tuple $(\Omega, \mathcal{F}, \mu)$ that contains a measurable space and a measure μ on that space. If μ is a probability measure, the measure space is also defined as a *probability space.*

1.3 Properties of probability measures

Probability has several useful properties:

Proposition 1.3.1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. The following statements are true:*

(a) $\forall A \in \mathcal{F}$*:* $P(A) \leq 1$.

(b) $\forall A \in \mathcal{F}$*:* $P(A^c) = 1 - P(A)$.

(c) $\forall A, B \in \mathcal{F}$*: If* $A \subseteq B$*, then* $P(A) \leq P(B)$.

(d) $\forall A_1, A_2 \in \mathcal{F}$*:* $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$.

(e) $\forall A_1, ..., A_n \in \mathcal{F}$*:* $P(\bigcup_{i=1}^{n} A_i) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} P(A_1 \cap \cdots \cap A_n)$ *(inclusion-exclusion principle).*

(f) $\forall A_1, ..., A_n \in \mathcal{F}$: $P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$ *(Boole's inequality).*

Theorem 1.3.1 (Continuity of probability)**.** *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space. Let* $\{A_i\}$ *be a countable sequence of events from* $\mathcal{F}$*. Then*

$$P(\bigcup_{i=1}^{\infty} A_i) = \lim_{n\to\infty} P(\bigcup_{i=1}^{n} A_i).$$

Proof. We start by defining a new sequence $B_i = A_i \setminus \bigcup_{j<i} A_j$. We claim that the sequence has the following properties (proof left as an exercise):

(a) $\forall i \neq j$: $B_i \cap B_j = \emptyset$ (*the sets are disjoint*).

(b) $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$.

Therefore,

$$\begin{aligned}
P(\bigcup_{i=1}^{\infty} A_i) &= P(\bigcup_{i=1}^{\infty} B_i) && \text{(b)} \\
&= \sum_{i=1}^{\infty} P(B_i) && \text{(countable additivity and a)} \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} P(B_i) && \text{(def. of infinite series)} \\
&= \lim_{n\to\infty} P(\bigcup_{i=1}^{n} B_i) \\
&= \lim_{n\to\infty} P(\bigcup_{i=1}^{n} A_i).
\end{aligned}$$

■

These two corollaries follow from Theorem 1.3.1:

Corollary 1.3.1. *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space. If* $\{A_i\}$ *is a countable sequence of increasing nested events, that is,* $A_i \subseteq A_{i+1}$*, then*

$$P(\bigcup_{i=1}^{\infty} A_i) = \lim_{n\to\infty} P(A_n).$$

Corollary 1.3.2. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. If $\{A_i\}$ is a sequence of decreasing nested events, that is, $A_i \supseteq A_{i+1}$, then*

$$P(\bigcap_{i=1}^{\infty} A_i) = \lim_{n\to\infty} P(A_n).$$

The proof of these corollaries is left as an exercise.

1.4 Discrete probability spaces

Definition 1.4.1. A *discrete probability* space is a probability space with a countable set of outcomes.

For a countable set of outcomes and the power set as the set of events it is relatively straightforward to construct a valid probability space:

Proposition 1.4.1. *Let Ω be a countable and non-empty set. Let $\mathcal{F} = 2^{\Omega}$. Let $P_0 : \Omega \to [0, 1]$ be a function, such that $\sum_{\omega\in\Omega} P_0(\{\omega\}) = 1$.*

Then $(\Omega, \mathcal{F}, P)$, where $P(A) = \sum_{\omega\in A} P_0(\{\omega\})$, is a probability space.

Proof. $\mathcal{F}$ is clearly a σ-algebra of Ω, so what remains is to show that P is a probability measure on our measurable space $(\Omega, \mathcal{F})$.

By definition $P(\emptyset) = 0$. Furthermore, $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{\omega\in\bigcup_{i=1}^{\infty} A_i} P_0(\{\omega\}) = \sum_{i=1}^{\infty}\sum_{\omega\in A_i} P_0(\{\omega\}) = \sum_{i=1}^{\infty} P(A_i)$. Therefore, P is a measure.

Finally, $P(\Omega) = \sum_{\omega\in\Omega} P_0(\{\omega\}) = 1$, therefore, P is a probability measure. ■

Proposition 1.4.1 says that if we want to define a probability space on a countable set of outcomes, it suffices to use the power set as the set of events (our σ-algebra) and assign a probability to each outcome. The definition of the probability measure for each event then follows via countable additivity. Of course, our assignment of probabilities to singletons must obey the laws of probability measures - probabilities must be between 0 and 1 and they must sum up to 1.

We illustrate discrete probability spaces with two examples:

Example 1.4.1. *(Coin flip)*

A coin has two possible states - heads and tails. Without loss of generality, we can assign the number 1 to heads and number 0 to tails. Our set of outcomes is then $\Omega = \{0, 1\}$.

For the set of events, we use the power set $\mathcal{F} = \{\emptyset, \{0\}, \{1\}, \Omega\}$, although it is not the only possible choice (see below).

For the probability measure we must set $P(\emptyset) = 0$ (definition of measure) and $P(\Omega) = 1$ (definition of probability measure). We have the freedom to set $P(\{1\}) = \theta$, where $\theta \in [0, 1]$ (property of probability measures), but $P(\{0\}) = 1 - \theta$ (property of probability measures). This coin-with-probability-θ *measure is also called a Bernoulli measure or, when used to define a distribution of a random variable, a Bernoulli random variable.*

Note that $\Omega = \{0, 1\}$, $\mathcal{F} = \{\emptyset, \Omega\}$ with $P(\emptyset) = 0$ and $P(\Omega) = 1$ is also a probability space (as an exercise, you can verify it has all the defining properties). It is just a probability space where $P(heads)$ and $P(tails)$ do not exist, which arguably makes it a less useful probability space.

With Proposition 1.4.1 it is straightforward to construct a probability space over a countable set of outcomes even if the set of outcomes is infinite:

Example 1.4.2. *(Measures over natural numbers)*

Let $\Omega = \mathbb{N}$ and $\mathcal{F} = 2^{\mathbb{N}}$. Observe that the set of events is not countable. However, as long as the set of outcomes is, things remain simple.

The following are two very common probability measures:

(a) Geometric*: $P(\{k\}) = (1 - \theta)^k \theta$, $\theta \in (0, 1)$. Check that $\sum_{k \in N} P(\{k\}) = 1$.*

(b) Poisson*: $P(\{k\}) = \frac{\lambda^k e^{-\lambda}}{k!}$, where $\lambda > 0$. Check that $\sum_{k \in N} P(\{k\}) = 1$.*

Exercises

Exercise 1.1. Prove Proposition 1.2.1.

Exercise 1.2. Prove statements (a-c) from Proposition 1.2.2.

Exercise 1.3. Prove Proposition **??**.

Exercise 1.4. Prove statements (a-d) from Proposition 1.3.1. Which of them generalize to finite measures?

Exercise 1.5. Prove statement (e) from Proposition 1.3.1. Does it generalize to finite measures?

Exercise 1.6. Prove statement (f) from Proposition 1.3.1. Does it generalize to finite measures?

Exercise 1.7. Which statements from Proposition 1.3.1 apply to measures and not just probability measures? Provide a counter-example for those that do not.

Exercise 1.8. Prove statements (a) and (b) from proof of Theorem 1.3.1.

Exercise 1.9. Prove Corollary 1.3.1.

Exercise 1.10. Prove Corollary 1.3.2.

Chapter 2

Uncountable probability spaces

In the introduction we illustrated that for uncountable sets of outcomes it is not immediately clear how to choose the set of events and define coherent probabilities. In fact, we even hinted that it might not be possible to assign probabilities to every subset of an uncountable set. Now we formalize this notion and provide the means for the construction of uncountable probability spaces by introducing Borel sets and Lebesgue measure.

2.1 Existence of non-measurable sets

First, we focus on one of the most simple probabilistic statements - the uniform distribution on the unit interval. Attempting the approach we used for countable sets of outcomes fails immediately. Intuitively, for a uniform probability all singletons should have the same probability, however:

- If we give each singleton a positive probability, then by countable additivity, there will be subsets with infinite probability. For example, the subset of all rational numbers between 0 and 1 or the subset $\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots\}$.
- If we give each singleton zero probability, that is not enough to determine the probability of all other subsets of the unit interval, because countable additivity alone is not sufficient to define the probability of uncountable intervals, such as $[\frac{1}{3}, \frac{1}{2}]$.

While we might be perfectly comfortable with saying $X \sim U(0,1)$ and working with continuous probability distributions this shows that we might still lack a complete understanding of the underlying probability spaces.

Now we attempt a more formal construction of our uniform probability:

Definition 2.1.1 (Naive uniform probability measure)**.** Let P be a probability measure on measurable space $(\Omega = [0,1], \mathcal{F} = 2^{[0,1]})$. P is a *uniform probability measure* if it satisfies the following two properties:

(i) $P((a,b)) = P([a,b)) = P((a,b]) = P([a,b])$ for all $(a,b) \in \mathcal{F}$ *(uniformity).*

(ii) $P(A) = P(A \oplus \omega)$ for all $\omega \in \Omega$ and $A \in \mathcal{F}$ *(shift invariance).*

The shift operator $\oplus$ is defined as

$$A \oplus x \triangleq \{a + x | a \in A, a + x \leq 1\} \cup \{a + x - 1 | a \in A, a + x > 1\}.$$

Our goal is to construct a uniform probability measure over all subsets of the unit interval, so the choice of the outcome set and σ-algebra does not need further justification. The above two properties are properties that every uniform measure should have. We are not making the statement that these two properties are the only two properties a uniform measure should have, but they will suffice for our argument that such a measure does not exist (adding further requirements would make it at most more difficult to construct such a measure and does not contradict our argument):

Theorem 2.1.1 (Vitali set - a non-measurable set)**.** *A uniform probability measure as defined in Definition 2.1.1 does not exist.*

Proof. We will prove this by contradiction. Let us assume that such a probability measure exists.

We define an equivalence relation on Ω: $x \sim y$ iff $y - x \in \mathbb{Q}$. This relation partitions Ω into equivalence classes. Let $H \subset \Omega$ consist of precisely one element from each equivalence class (this requires the use of the Axiom of choice). Note that we assume, without loss of generality, that $0 \notin H$. If we were to allow for the case $0 \in H$, the union below would have to handle it as a special case, in order to be able to obtain a subset that contains 1.

Because H contains an element from each equivalence class, the union

$$\bigcup_{x \in [0,1), x \in \mathbb{Q}} H \oplus x$$

contains every point in $(0,1]$. Furthermore, sets $H \oplus x$ in the above union are all disjoint. By construction, each of them contains elements that are not exactly a rational number apart, so they can appear in another set only by looping around. However, this cannot happen, since we don't include 1.

Now we can use two properties of probability measures (total probability of 1 and countable additivity) and shift invariance to show that

$$P((0,1]) = 1 = \sum_{x\in[0,1),x\in\mathbb{Q}} P(H \oplus x) = \sum_{x\in[0,1),x\in\mathbb{Q}} P(H).$$

The rightmost sum is a countable sum of the same element and can only be 0 or infinite. This leads to a contradiction! ■

If our goal is to have a uniform probability measure, then we cannot relax the two properties or the defining properties of a probability measure. The only option that remains is to restrict the σ-algebra to something less than the power set. In other words, we must concede that certain probabilistic questions cannot be answered consistently.

2.2 Borel sets on $(0, 1]$

We have shown that it is impossible to construct a uniform probability measure on the unit interval $\Omega = [0, 1]$, if we set $\mathcal{F} = 2^\Omega$. Therefore, we must consider a smaller σ-algebra. Note that for convenience (some proofs are easier), we now focus on the set $\Omega = (0, 1]$.

We will construct such a σ-algebra implicitly by starting with a relatively small set of subsets of $(0, 1]$ that we definitely want to have in our set of events - the set of open intervals (why do we want at least these?) - and then extending this set the minimum required amount to make it a σ-algebra. Such an approach is justified by the following proposition:

Proposition 2.2.1 (Generated σ-algebras)**.** *For every set $\mathcal{C}$ of subsets of Ω there exists a smallest σ-algebra that contains all elements of $\mathcal{C}$.*

We denote such a σ-algebra by $\sigma(\mathcal{C})$ and we call it the σ-algebra generated by $\mathcal{C}$.

Proof. Let $\{\mathcal{F}_i\}$ be a set of all σ-algebras that contain $\mathcal{C}$. We know that this set is non-empty - it contains at least 2^Ω.

Now consider the intersection of all sets in our set $\mathfrak{F} = \bigcap_i \mathcal{F}_i$. Because every $\mathcal{F}_i$ contains $\mathcal{C}$ (by definition), we have $\mathcal{C} \in \mathfrak{F}$. Furthermore, the intersection of σ-algebras is a σ-algebra (left as an exercise). Therefore, $\mathfrak{F}$ is a σ-algebra that contains $\mathcal{C}$.

Finally, for every $\mathcal{F}_i$, we have $\mathfrak{F} \subseteq \mathcal{F}_i$. So, $\mathfrak{F}$ is at most as large as any σ-algebra that contains $\mathcal{C}$ and is therefore the smallest σ-algebra that contains $\mathcal{C}$. ■

So, if we define $\mathcal{C}$ to be the set of all open intervals $(a, b) \subset (0, 1]$, we know that there exists $\sigma(\mathcal{C})$ that is a σ-algebra with the fewest additional elements. If fact, such σ-algebras are so important that they have a name:

Definition 2.2.1 (Borel σ-algebra)**.** Let $\mathcal{C}$ be the set of all open intervals (a, b) in $(0, 1]$. The generated σ-algebra $\sigma(\mathcal{C})$ is called the **Borel σ-algebra** and is denoted by $\mathcal{B}_{(0,1]}$. Elements of Borel σ-algebras are called *Borel sets.*

Proposition 2.2.1 guarantees that $\mathcal{B}_{(0,1]}$ exists, but we at this point understand very little about this σ-algebra.

First, let us inquire about the cardinality of $\mathcal{B}_{(0,1]}$. We know that it contains more than just the open intervals (why?; see Exercise 2.2) and we have not excluded the possibility that the completion of the open subsets to a σ-algebra would lead to $\mathcal{B}_{(0,1]} = 2^{(0,1]}$. Luckily, that is not the case. In fact, it has been proven that the cardinality of the Borel σ-algebra is equal to the cardinality of $\mathbb{R}$. However, the proof of this statement is beyond the scope of this text.

Next, let us explore which sets are Borel sets.

Proposition 2.2.2. *Every singleton set $\{\omega\}$, $0 < \omega \leq 1$, is in $\mathcal{B}_{(0,1]}$.*

Proof. First, $\{1\}$ is in $\mathcal{B}_{(0,1]}$, because the complement $(0, 1)$ is by definition in $\mathcal{B}_{(0,1]}$. Also, $(0, b)$ and $(b, 1)$, for any $b \in (0, 1)$, are in $\mathcal{B}_{(0,1]}$ by definition. Then, by the properties of σ-algebras, the set $(0, b) \cup (b, 1) \cup \{1\}$ is also in $\mathcal{B}_{(0,1]}$. Its complement, which is also in $\mathcal{B}_{(0,1]}$, is $\{b\}$. ■

The following are an immediate consequence.

Corollary 2.2.1.

(a) $\mathcal{B}_{(0,1]}$ contains all half-open intervals in $(0, 1]$. That is, intervals of the form $(a, b]$ or $[a, b)$.

(b) $\mathcal{B}_{(0,1]}$ contains all closed intervals in $(0, 1]$. That is, intervals of the form $[a, b]$.

The proof of Corollary 2.2.1 is left as an exercise.

So, all intervals, singletons, countable unions, intersections, and complements thereof are Borel sets. In fact, all sets that will be of practical interest to most of us, are Borel sets.

While a set of subsets generates a unique σ-algebra, multiple different sets can generate the same σ-algebra:

Proposition 2.2.3. *Show for each of the following sets that $\sigma(\mathcal{C}) = \mathcal{B}_{(0,1]}$:*

(a) $\mathcal{C}$ is the set of all intervals in $(0, 1]$ of the form $(a, b]$.

(b) $\mathcal{C}$ is the set of all intervals in $(0, 1]$ of the form $[a, b]$.

(c) $\mathcal{C}$ is the set of all intervals in $(0, 1]$ of the form $(0, a]$.

Proof. We'll prove (a) and leave (b) and (c) as an exercise.

We can prove $\sigma(\mathcal{C}) = \mathcal{B}_{(0,1]}$ by showing that $\sigma(\mathcal{C}) \subseteq \mathcal{B}_{(0,1]}$ and $\mathcal{B}_{(0,1]} \subseteq \sigma(\mathcal{C})$. The former follows immediately from Corollary 2.2.1: $\mathcal{C} \subseteq \mathcal{B}_{(0,1]}$ (any set generated by a subset of a σ-algebra is already in the σ-algebra by definition).

To prove $\mathcal{B}_{(0,1]} \subseteq \sigma(\mathcal{C})$ it suffices to show that semi-open intervals generate all open intervals:

$$(a, b) = \bigcup_{i=1}^{\infty} (a, b - \frac{1}{i}].$$

Because the set of open intervals generates $\mathcal{B}_{(0,1]}$, the set of semi-open intervals generates at least $\mathcal{B}_{(0,1]}$. ■

2.3 Uniform measure on $(0, 1]$

Now we can return to the task of constructing a uniform measure λ on $\Omega = (0, 1]$. Instead of using the power set, our measurable space will be $((0, 1], \mathcal{B}_{(0,1]})$. Hopefully this removes the pathological sets that made it impossible (see Theorem 2.1.1).

Let us recall our notion of a uniform measure, as we defined it in Definition 2.1.1. We required such a method to have the following two properties:

(i) $P((a, b)) = P([a, b)) = P((a, b]) = P([a, b])$ for all $(a, b) \in \mathcal{F}$ *(uniformity).*

(ii) $P(A) = P(A \oplus \omega)$ for all $\omega \in \Omega$ and $A \in \mathcal{F}$ *(shift invariance).*

Property (i) states that the measure needs to be proportional to the length of the interval. So, for intervals of the form (a, b) (and for their half-open and closed counterparts), we could, without loss of generality, define our measure λ with the length of the interval: $\lambda((a, b)) \triangleq b - a$. As an exercise, verify that this measure is shift invariant as well.

The problem we face now is how to extend this measure to all Borel sets in $\mathcal{B}_{(0,1]}$, some of which can be very complicated. We will tackle this problem by starting with a more manageable set of sets and then invoking this powerful theorem from measure theory that will allow us to extend the probability measure to all Borel sets:

Theorem 2.3.1 (Caratheodory's extension theorem)**.** *Let $\mathcal{F}_0$ be an algebra of subsets of Ω. Let $\mu_0 : \mathcal{F}_0 \to [0, \infty)$, such that for every countable sequence $\{A_i\}$ of disjoint sets in $\mathcal{F}_0$ we have $\mu_0(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu_0(A_i)$ for any $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}_0$. Then there exists a unique measure μ on $(\Omega, \sigma(\mathcal{F}_0))$, such that $\mu(A) = \mu_0(A)$ for all $A \in \mathcal{F}_0$.*

The proof of Caratheodory's theorem is beyond the scope of this text. Note that the theorem holds in the more general case of σ-finite measures, which is a condition weaker than finite. The Lebesgue measure on $\mathbb{R}$ is an example of a measure that is not finite but is σ-finite.

The theorem states that in order to uniquely define a measure over some measurable space $(\Omega, \mathcal{F})$, it suffices to define a measure that is consistent on an algebra that generates $\mathcal{F}$. This is a very useful statement, because the required properties can be much easier to verify on an algebra. Note that measures defined only on an algebra are not really measures according to our definition, so we call them pre-measures.

In our case we have already determined that our uniform measure of an interval should be the length of an interval. However, the set of intervals is not an algebra on $(0, 1]$. For example, the complement of an interval or the union of two disjoint intervals is not always an interval. Instead, we start with half-open intervals and add all that is required to make this set an algebra.

Proposition 2.3.1. *Let $\mathcal{F}_0$ be the set of $\emptyset$ and all subsets of $(0, 1]$ which are finite unions of disjoint intervals of the form $(a, b]$. We have*

(a) $\sigma(\mathcal{F}_0) = \mathcal{B}_{(0,1]}$,

(b) $\mathcal{F}_0$ is an algebra, and

(c) $\mathcal{F}_0$ is not a σ-algebra.

Proof. The proof of (a) is straightforward. We have $\mathcal{F}_0 \subseteq \mathcal{B}_{(0,1]}$, so $\sigma(\mathcal{F}_0) \subseteq \mathcal{B}_{(0,1]}$. We also know from before that the set of all half-open intervals generates $\mathcal{B}_{(0,1]}$, so $\mathcal{B}_{(0,1]} \subset \sigma(\mathcal{F}_0)$.

In order for $\mathcal{F}_0$ to be an algebra, it must contain the empty set (it does, by definition) and must be closed under complementation and under finite unions. A union of two half-open intervals of the form $(a, b]$ is either another interval of the form $(a, b]$ or a union of two such intervals. Both cases are by definition in $\mathcal{F}_0$. Similarly, the complement of any finite union of such intervals is again a finite union of such intervals.

We can show (c) by observing the countable union $\cup_{i=1}^{\infty}(0, \frac{i}{i+1}]$. All of the terms in the union are intervals of the form $(a, b]$ and are therefore in $\mathcal{F}_0$. However, their union is $(0, 1)$, which is not in $\mathcal{F}_0$. ■

So far, we have introduced the Borel σ-algebra, which is the smallest $\sigma-$algebra that contains the sets we are interested in in practice. We have now introduced an algebra $\mathcal{F}_0$ that generates the Borel σ-algebra. Before we can invoke Caratheorory's theorem to show that our uniform probability can indeed be uniquely extended to the Borel σ-algebra, we must complete our uniform measure so that it is indeed a pre-measure on $\mathcal{F}_0$.

We have already determined how we are going to measure the intervals of the form $(a, b]$: $\lambda((a, b]) = b - a$. Because our measure will be finite, we have $\lambda(\emptyset) = 0$. All other sets in $\mathcal{F}_0$ are finite unions of disjoints half-open intervals, for example $(\frac{1}{3}, \frac{1}{2}] \cup (\frac{4}{5}, \frac{5}{6}]$, for which our measure λ is not yet defined. In general, sets

$$A = \bigcup_{i=1}^{n} (a_i, b_i],$$

where $0 \leq a_1$, $a_i < b_i$ and $b_i \leq a_{i+1}$. For such sets, we define their measure as the sum of the measures of individual intervals: $\lambda(A) = \sum_{i=1}^{n} \lambda((a_i, b_i])$. This should not be surprising, as it is necessary to define it like this if we are to respect countable (in this case only finite) additivity.

One step remains before we can invoke Theorem 2.3.1. We must show that our uniform measure is countably additive on $\mathcal{F}_0$. However, the proof is beyond the scope of this text.

This completes our argument that there exists a measure λ on the Borel $\sigma-$algebra on $(0, 1]$ that has the desired uniformity properties. We call this measure the Lebesgue measure. It is a generalization of the notion of length. And on $(0, 1]$ it is a probability measure (why?).

Our successful extension of this uniform measure to all Borel sets implies that the Vitali set is not a Borel set. Therefore, we have successfully avoided this and other pathological sets that are incompatible with the notion of uniform probability.

2.4 Lebesgue measure on $\mathbb{R}$

Borel sets and Lebesgue measure can be defined in a similar fashion for the real line (and $\mathbb{R}^n$, although this is too technical and out of the scope of this text). Note that there is also more general way of defining Borel sets on any topological space by starting with open sets. We can see how $\mathbb{R}$ and open intervals is just a special case.

Definition 2.4.1 (Borel σ-algebra on $\mathbb{R}$)**.** Let $\mathcal{C}$ be the set of all open intervals in $\mathbb{R}$. The generated σ-algebra $\sigma(\mathcal{C})$ is called the Borel σ-algebra and is denoted by $\mathcal{B}_{\mathbb{R}}$.

The following two propositions give two alternative but equivalent definitions of the Borel σ-algebra on the real line.

Proposition 2.4.1. *Let $\mathcal{C}$ be the set of all intervals of the form $(-\infty, a]$ in $\mathbb{R}$. Then $\sigma(\mathcal{C}) = \mathcal{B}_{\mathbb{R}}$.*

The proof of this proposition is left as an exercise.

Proposition 2.4.2. *Let $\mathcal{C}$ be the set of all sets $A \subseteq \mathbb{R}$, such that $A \cap (n, n+1]$ is a Borel set on $(n, n+1]$ for all $n \in \mathbb{Z}$. Then $\sigma(\mathcal{C}) = \mathcal{B}_{\mathbb{R}}$.*

The proof of this proposition is left as an exercise.

Now we are ready to extend the definition of Lebesgue measure to the real line.

Definition 2.4.2 (Lebesgue measure on the real line)**.** We define the Lebesgue measure of a set $A \in \mathcal{B}_{\mathbb{R}}$ as

$$\lambda(A) \triangleq \sum_{n=-\infty}^{\infty} \lambda^*(A \cap (n, n+1]),$$

where λ^* is the Lebesgue measure on the unit interval.

In essence, we partition the real line into unit intervals, measure the set's intersection with each interval and sum up the measures. However, we have yet to prove that it it a valid measure on the real line.

Proposition 2.4.3. $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \lambda)$ *is a measure space.*

Proof. $\mathcal{B}_{\mathbb{R}}$ is a σ-algebra on $\mathbb{R}$. It is also clear that λ is defined on all subsets of $\mathbb{R}$ and that it is non-negative (it is a sum of terms that are non-negative). Furthermore, the unit interval Lebesgue measure is a measure, so $\lambda^*(\emptyset) = 0$ and therefore $\lambda(\emptyset) = 0$. What remains to be shown to complete the proof is that λ is countably additive.

Let's take a sequence of pairwise disjoint sets $A_i \in \mathcal{B}_{\mathbb{R}}$:

$$\begin{aligned}
\lambda\left(\bigcup_{i=1}^{\infty} A_i\right) &= \sum_{n=-\infty}^{\infty} \lambda^*(\bigcup_{i=1}^{\infty} A_i \cap (n, n+1]) && \text{(by definition)} \\
&= \sum_{n=-\infty}^{\infty} \sum_{i=1}^{\infty} \lambda^*(A_i \cap (n, n+1]) && \text{(by countable additivity of } \lambda^*) \\
&= \sum_{i=1}^{\infty} \sum_{n=-\infty}^{\infty} \lambda^*(A_i \cap (n, n+1]) && \text{(Fubini's theorem)} \\
&= \sum_{i=1}^{\infty} \lambda(A_i) && \text{(by definition)}
\end{aligned}$$

■

Proposition 2.4.4. $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \lambda)$ *is not a probability space.*

The proof of this proposition is left as an exercise.

Exercises

Exercise 2.1. Prove that the intersection of two σ-algebras on Ω is a σ-algebra on Ω.

Exercise 2.2. Show that the set of open subsets of $(0, 1]$ is not a σ-algebra on $(0, 1]$.

Exercise 2.3. Prove Corollary 2.2.1.

Exercise 2.4. Prove (b) and (c) from Proposition 2.2.3.

Exercise 2.5. Prove that the Lebesgue measure of an interval (a, b) on $(0, 1]$ is shift-invariant.

Exercise 2.6. Let $\mathcal{F}_0$ be a set that contains $\emptyset$ and all subsets of $(0, 1]$ which are finite unions of disjoint intervals of the form $(a, b]$. Show that:

(a) $\mathcal{F}_0$ is an algebra.

(b) $\sigma(\mathcal{F}_0) = \mathcal{B}_{(0,1]}$.

Exercise 2.7. Prove that the Lebesgue measure of a singleton is 0. That is, $\lambda(\omega) = 0$, for all $\omega \in (0, 1]$.

Exercise 2.8. Prove Proposition 2.4.1.

Exercise 2.9. Prove Proposition 2.4.2.

Exercise 2.10. Prove Proposition 2.4.4.

Chapter 3

Conditional probability

Conditional probability defines how we should modify our uncertainty given some evidence or truth. As such it is the core mechanism of learning. Conditional probability also introduces an important special case of independent events - events where conditioning on one event has no effect on the uncertainty of the other.

Unless explicitly stated otherwise, all definitions, theorems and propositions in this chapter assume that we are working on a probability space $(\Omega, \mathcal{F}, P)$.

3.1 Conditional probability measure

Definition 3.1.1 (Conditional probability)**.** Let $B \in \mathcal{F}$, such that $P(B) > 0$. The conditional probability of event A conditional to event B is defined as

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

Conditional probability cannot be derived from the axioms of probability measures - it must be defined. If you want to learn more about why conditional probability is defined like this and why there are no reasonable alternative definitions, refer to Kadane (2011, Ch. 2).

A conditional probability is also a probability measure:

Theorem 3.1.1. *Let $B \in \mathcal{F}$ and $P(B) > 0$. The function $P(\cdot|B) : \mathcal{F} \to [0, 1]$ is a probability measure on $(\Omega, \mathcal{F})$.*

Proof. Being a ratio of probability measures, the conditional probability is non-negative. Furthermore,

$$P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

and

$$P(\emptyset|B) = \frac{P(\emptyset \cap B)}{P(B)} = \frac{P(\emptyset)}{P(B)} = 0.$$

What remains to be shown is that conditional probability is countably additive:

$$P(\bigcup_{i=1}^{\infty} A_i|B) = \frac{P((\cup_{i=1}^{\infty} A_i) \cap B)}{P(B)} = \frac{P(\cup_{i=1}^{\infty}(A_i \cap B))}{P(B)}$$
$$= \frac{\sum_{i=1}^{\infty} P(A_i \cap B)}{P(B)} = \sum_{i=1}^{\infty} P(A_i|B).$$

■

3.2 Properties of conditional probability

Definition 3.2.1 (Partition)**.** The *partition of a set* Ω is a countable set of disjoint events $\{A_i\}$, such that $\cup_{i=1}^{\infty} A_i = \Omega$.

Proposition 3.2.1 (Marginal probability)**.** *Let $A \in \mathcal{F}$ and let $\{B_i\}$ be a partition of Ω. Then,*

$$P(A) = \sum_{i=1}^{\infty} P(A \cap B_i).$$

Proof. Sets $A \cap B_i$ are disjoint, therefore

$$\sum_{i=1}^{\infty} P(A \cap B_i) = P(\bigcup_{i=1}^{\infty} A \cap B_i) = P(A \cap \bigcup_{i=1}^{\infty} B_i) = P(A \cap \Omega) = P(A).$$

Note that the statement $\bigcup_{i=1}^{\infty} A \cap B_i = A \cap \bigcup_{i=1}^{\infty} B_i$ that we used in the proof is not obvious. It is left as an exercise. ■

Proposition 3.2.2 (Law of total probability)**.** *Let $A \in \mathcal{F}$ and let $\{B_i\}$ be a partition of Ω. Then,*

$$P(A) = \sum_{i=1}^{\infty} P(A|B_i)P(B_i).$$

Proof. To prove this statement, we apply marginal probability and conditional probability to obtain

$$P(A) = \sum_{i=1}^{\infty} P(A \cap B_i) = \sum_{i=1}^{\infty} P(A|B_i)P(B_i).$$

■

The following theorem is the cornerstone of Bayesian statistics:

Theorem 3.2.1 (Bayes' rule)**.** *Let $B \in \mathcal{F}$, such that $P(B) > 0$. For any event A we have $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.*

Proof. We start with the definition of conditional probability and apply the definition again on the numerator. ■

Proposition 3.2.3 (Bayes' rule applied to partitions)**.** *Let $A \in \mathcal{F}$, $P(A) > 0$ and let $\{B_i\}$ be a partition of Ω. Then*

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{\infty} P(A|B_i)P(B_i)}.$$

Proof. We prove the statement by applying Bayes's rule and the Law of total probability:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{\infty} P(A|B_i)P(B_i)}.$$

■

Proposition 3.2.4 (Factorization of probability measures)**.** *For any countable set of events $\{A_i\}$ we have*

$$P(\bigcap_{i=1}^{\infty} A_i) = P(A_1) \prod_{i=2}^{\infty} P(A_i|A_1 \cap A_2 \cap \cdots \cap A_{i-1}).$$

This statement is conditional on all the conditional probabilities being well defined.

Proof. First, we prove the statement for a finite set of events. By applying the definition of conditional probability to all factors on the right-hand side, all but one of the terms cancel out:

$$P(A_1) \prod_{i=2}^{n} \frac{P(A_1 \cap A_2 \cap \cdots \cap A_i)}{P(A_1 \cap A_2 \cap \cdots \cap A_{i-1})} = P(\bigcap_{i=1}^{n} A_i).$$

Now

$$\lim_{n\to\infty} P(\bigcap_{i=1}^{n} A_i) = \lim_{n\to\infty} P(A_1)\prod_{i=2}^{n} P(A_i|A_1 \cap A_2 \cap \cdots \cap A_{i-1})$$

and, using continuity of probability on the left-hand side and the definition of an infinite sequence on the right-hand side, we get

$$P(\bigcap_{i=1}^{\infty} A_i) = P(A_1)\prod_{i=2}^{\infty} P(A_i|A_1 \cap A_2 \cap \cdots \cap A_{i-1}).$$

■

3.3 Independence

Definition 3.3.1 (Independence)**.** Events A and B are said to be independent if $P(A \cap B) = P(A)P(B)$.

Proposition 3.3.1 (Conditional probability of independent events)**.** *If A and B are independent and $P(B) > 0$, then $P(A|B) = P(A)$. Conversely, if $P(A|B) = P(A)$ then A and B are independent.*

Therefore, as long as $P(B) > 0$ the statement $P(A|B) = P(A)$ is equivalent to independence of A and B. The proof is left as an exercise.

Definition 3.3.2 (Conditional independence)**.** Events A and B are said to be conditionally independent given event C if $P(A \cap B|C) = P(A|C)P(B|C)$.

Definition 3.3.3 (Joint independence)**.** A countable set of events $\{A_i\}$, $i \in I$ are said to be jointly independent if for every non-empty finite subset $I_0 \subseteq I$ we have

$$P(\bigcap_{i\in I_0} A_i) = \prod_{i\in I_0} P(A_i).$$

The following proposition characterizes the (lack of) relationship between different types of independence.

Proposition 3.3.2.

(a) Conditional independence does not imply independence*: That is, for events A, B and C, such that $P(A|B \cap C) = P(A|C)$, that does not imply that $P(A|B) = P(A)$.*

(b) Independence does not imply conditional independence*: That is, there exist events A, B and C, such that $P(A|B) = P(A)$ and $P(A|B \cap C) \neq P(A|C)$.*

(c) Pairwise independence does not imply joint independence*: That is, for a countable set of events* $\{A_i\}$*, such that* $P(A_i|A_j) = P(A_i)$ *for all* $i \neq j$*, that does not imply that events* $\{A_i\}$ *are jointly independent.*

(d) Joint independence implies pairwise independence for all pairs..

The proof is left as an exercise.

Exercises

Exercise 3.1. Let A be an event and let $\{B_i\}$ be a partition. Show that $\bigcup_{i=1}^{\infty} A \cap B_i = A \cap \bigcup_{i=1}^{\infty} B_i$.

Exercise 3.2. Prove Proposition 3.3.1.

Exercise 3.3. Prove (d) in Proposition 3.3.2.

Exercise 3.4. Prove (a-c) in Proposition 3.3.2 by finding a counterexample.

Chapter 4

Abstract integration

In this chapter we will introduce a more general approach to integration - abstract integration of a function f with respect to a measure μ:

$$\int f d\mu.$$

While our treatment will be more general, we will primarily be interested in a few special cases: integration with respect to the Lebesgue measure λ (or the Lebesgue integral), integration with respect to a probability measure P, and integration with respect to the counting measure $\#$, which we define in this chapter.

The Lebesgue integral will allow us to integrate a more general class of functions (it is a strict generalization over the Riemann integral on bounded functions on bounded intervals). It also has other nice properties that make it more appropriate for rigorous probability theory. For example, limits of Lebesgue integrable functions tend to be Lebesgue integrable and the integral can be more easily extended to other, non-$\mathbb{R}^n$ spaces.

Integration with respect to a probability measure is particularly useful as it is directly related to the expectation of random variables and will allow us to treat this important quantity more generally, for all types of random variables.

4.1 A review of Riemann integration

First, we'll briefly review the Riemann integral of a bounded function f on an interval $[a, b]$ that has at most a countable number of discontinuities.

We define a partition P of $[a, b]$ as a finite set of points $x_0, x_1, ..., x_n$ such that $a = x_0 \leq x_1 \leq x_2 \leq ... \leq x_n = b$.

Additionally, we define the supremum and infimum of the function for each interval in the partition:

$$M_i \triangleq \sup\{f(x) : x_{i-1} \leq x \leq x_i\}$$

and

$$m_i \triangleq \inf\{f(x) : x_{i-1} \leq x \leq x_i\}.$$

Now we can define the upper and lower Riemann sums:

$$U(P, f) = \sum_{i=1}^{n} M_i(x_i - x_{i-1})$$

and

$$L(P, f) = \sum_{i=1}^{n} m_i(x_i - x_{i-1}).$$

Since f is bounded (say, by $m \leq f(x) \leq M$), the lower and upper Riemann sums are also bounded

$$m(b - a) \leq L(P, f) \leq U(P, f) \leq M(b - a).$$

Finally, we define the lower and upper Riemann integrals:

$$\overline{\int_a^b} f(x)dx = \inf_{\text{all partitions P of [a, b]}} U(P, f)$$

and

$$\underline{\int_a^b} f(x)dx = \sup_{\text{all partitions P of [a, b]}} L(P, f).$$

If the lower and upper Riemann integrals are the same, we say that f is Riemann integrable and the value of the integral equals the value of the lower and upper Riemann integral.

4.2 Integrating simple functions

Before we proceed with the definition of more abstract integration, we must first define what it means for a function to be measurable:

Definition 4.2.1 (Measurable function)**.** Let $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$ be measurable spaces. Function $f : \Omega \to S$ is said to be a *measurable function* if for every set $A \in \mathcal{S}$ we have $f^{-1}(A) \in \mathcal{F}$, where

$$f^{-1}(A) \triangleq \{\omega \in \Omega | f(\omega) \in A\}.$$

That is, the *pre-image* f^{-1} of every set in $\mathcal{S}$ is in $\mathcal{F}$.

Our interest lies in the measurable spaces where S is the real line and $\mathcal{S}$ is the Borel sigma algebra on the real line. Unless otherwise noted, $\mathcal{F}$-measurable implies that the preimage of every Borel set is in $\mathcal{F}$.

The following proposition, which we state without proof, will be useful when dealing with sums and products of measurable functions.

Proposition 4.2.1. *Let $(\Omega, \mathcal{F})$ be a measurable space and let $f : \Omega \to \mathbb{R}$ and $g : \Omega \to \mathbb{R}$ be $\mathcal{F}$-measurable functions. Then,*

- *the pointwise sum function $(f + g)(x) = f(x) + g(x), \forall x \in \Omega$ and*
- *the pointwise product function $(f \cdot g)(x) = f(x)g(x), \forall x \in \Omega$*

are $\mathcal{F}$-measurable.

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let $f : \Omega \to \mathbb{R}$ be an $\mathcal{F}$-measurable function.

Definition 4.2.2 (Simple function)**.** A function f is said to be *simple* if it can be written as

$$f(\omega) = \sum_{i=1}^{n} a_i I_{A_i}(\omega), \forall \omega \in \Omega,$$

where $a_i \in \mathbb{R}$, $A_i \in \mathcal{F}$, and $I_{A_i}(\omega)$ is the indicator function ($I_{A_i}(\omega) = 1$ if $\omega \in A_i$ and 0 otherwise).

That is, a function is simple if it attains a finite number of distinct values.

Whenever we'll be referring to a simple function, we will be referring to the unique canonical representation where a_i are distinct and A_i disjoint.

Definition 4.2.3. Let f be a non-negative simple function. The abstract integral of f with respect to a measure μ is defined as

$$\int f d\mu \triangleq \sum_{i=1}^{n} a_i \mu(A_i).$$

That is, the abstract integral of the non-negative simple function is a sum over all distinct values, each multiplied by the measure of the set of points that attain that value.

Now we can look at a function that is not Riemann integrable but is a simple function, so we can integrate it using our abstract integral.

Example 4.2.1 (A non-Riemann integrable function)**.** *Function f is defined as follows: $f(x) = 1$, if x is irrational and $0 \leq x \leq 1$, and $f(x) = 0$ everywhere else. We will try to compute the integral of this function over the unit interval $[0, 1]$.*

First, let's compute the Riemann integral. Observe that no matter how fine a partition we make, there will always be an irrational and a rational number in every segment. So, the supremum (M_i) and infimum (m_i) of the function for each segment will be 1 and 0, respectively. This implies that the upper and lower Riemann sums will be 1 and 0, respectively, regardless of the partition. Therefore, the upper and lower Riemann integrals will not be the same. This function is not Riemann integrable!

Now let's compute the abstract integral. So far, we've only learnt how to compute the abstract integral for simple functions. However, f is a simple function - it attains only two distinct values, 1 and 0. To compute the integral, we now need only the measures of the sets where it attains 1 and 0. That is, the measure of the irrationals and rationals on $[0, 1]$.

We'll compute the integral with respect to the Lebesgue measure. So, what is the Lebesgue measure of the set of rationals on $[0, 1]$? There are countably many rationals and the Lebesgue measure of a singleton is 0. So, by the countable additivity of measures, the measure of rationals on $[0, 1]$ is 0. Furthermore, the Lebesgue measure of the unit interval is 1. Because irrationals and rationals are disjoint and together constitute the entire unit interval, the measure of all irrationals on the unit interval must be 1. The Lebesgue integral of f is therefore $0 \times 0 + 1 \times 1 = 1$.

This function f is therefore an example of a function that is not Riemann integrable but is integrable with respect to the Lebesgue measure using the abstract integral we just defined. Further interesting results can be obtained if we integrate the function on the interval $(-\infty, x]$. If $x < 0$ then the Lebesgue integral is clearly 0 (the function is 0 everywhere) and if $x > 1$ the Lebesgue integral is 1. If $0 \leq x \leq 1$, then the Lebesgue integral is x.

In Chapter 5 we will see how this characterizes the uniform probability law - function f is the probability density function of the continuous uniform random variable! So, even though we set countably infinitely many values to 0 (all rationals) the density retains its properties. This illustrates how densities are unique only up to a set of measure 0 and emphasizes how the density is not a

direct expression of probability.

4.3 Arbitrary measurable functions

Before we fully generalize the abstract integral, we need one more intermediate step - the integral of a non-negative function:

Definition 4.3.1. Let $f : \Omega \to [0, \infty)$ be a non-negative $\mathcal{F}$-measurable function. Let $S(f)$ be the set of all non-negative simple functions $g : \Omega \to [0, \infty)$, such that $\forall \omega \in \Omega : g(\omega) \leq f(\omega)$.

The abstract integral of f with respect to a measure μ is defined as

$$\int f d\mu \triangleq \sup_{g \in S(f)} \int g d\mu.$$

Note that the above integral can be infinite. Now we are ready for the general definition:

Definition 4.3.2. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let $f : \Omega \to \mathbb{R}$ be an $\mathcal{F}$-measurable function. Let $f = f_+ - f_-$ be a decomposition of f into a non-negative component $f_+ \triangleq \max(f, 0) \geq 0$ and a non-positive component $f_- \triangleq -\min(f, 0) \geq 0$. Note without proof that both f_+ and f_- are measurable functions.

The abstract integral of f with respect to measure μ is defined as

$$\int f d\mu \triangleq \int f_+ d\mu - \int f_- d\mu,$$

where the integrals of f_+ and f_- refer to the previously defined integral of non-negative functions. If both of the integrals are infinite, the integral of f is left undefined (we also say that it does not exist).

Note that the above definition can be used to integrate over an arbitrary measurable set $A \in \mathcal{F}$: $\int_A f d\mu = \int f I_A d\mu$, where I_A is the indicator function. Function $g = f I_A$ is a product of two measurable functions and therefore measurable. Hence we can integrate g as stated above.

Integrability, however, is a slightly more strict term than the existence of the integral - it does not include infinite integrals:

Definition 4.3.3 (Integrability)**.** We say that function f is integrable if $\int |f| d\mu < \infty$. That is, if f is absolutely integrable - the integral of its absolute value is finite.

Proposition 4.3.1. *A function f is integrable iff both f_+ and f_- are integrable.*

The proof of this proposition is left as an exercise.

In the case where f is integrable wrt the Lebesgue measure, we say that function f is Lebesgue integrable.

Relationship between Riemann and Lebesgue integral

The abstract integral is well-defined but the definitions are not very practical. The following theorems, which we state without proof, can be very helpful:

Theorem 4.3.1. *If a function $f : [a, b] \to \mathbb{R}$ is Riemann integrable, then it is Lebesgue integrable, and the two integrals coincide.*

Theorem 4.3.2. *If a non-negative function f on $\mathbb{R}$ is improper Riemann integrable, then it is Lebesgue integrable, and the two integrals coincide.*

That is, in most practical scenarios the Lebesgue and Riemann integral are equivalent. In particular, for all Riemann integrable probability density functions (see Chapter 5). However, as we have already shown, not all probabilility densities are Riemann integrable. They are, of course, by definition, Lebesgue integrable. Also note that there exist functions that are improper Riemann integrable but not Lebesgue integrable, for example $\frac{\sin x}{x}$.

4.4 Properties of abstract integration

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let f and g be $\mathcal{F}$-measurable functions.

Definition 4.4.1. A property $\mathbb{P}$ is said to hold *almost everywhere* with respect to a measure μ (μ-a.e. for short) if there exists a set $N \in \mathcal{F}$, such that $\mu(N) = 0$ and all $\omega \in \Omega \setminus N$ have the property $\mathbb{P}$.

When dealing with probability measures, the expression *almost surely* (a.s. for short) is often used instead.

The following two theorems each provide an answer to the very important question - *When can we interchange the integral and the limit?*. These theorems are among the most important results of abstract integration and are used in the proofs of many other results.

Theorem 4.4.1 (Monotone convergence theorem (MCT))**.** *Let f_n be a non-decreasing sequence (that is, $f_n(\omega) \leq f_{n+1}(\omega)$ for all ω and all $n \geq 1$) of non-negative measurable functions with $\lim_{n\to\infty} f_n(\omega) = f(\omega)$ μ-a.e.. That is, f_n converges point-wise to f almost everywhere. Then,*

$$\lim_{n\to\infty} \int f_n d\mu = \int f d\mu.$$

Theorem 4.4.2 (Dominated convergence theorem (DCT))**.** *Let f_n be a sequence of measurable functions with $\lim_{n\to\infty} f_n(\omega) = f(\omega)$ μ-a.e.. If there exists an integrable function g, such that $|f_n(\omega)| \leq g(\omega)$ for all n and ω, then*

$$\lim_{n\to\infty} \int f_n d\mu = \int f d\mu.$$

The proof of these theorems is beyond the scope of this text.

Note that the MCT allows for infinite integrals - the function f is not necessarily integrable. The existence of an dominating function g in the DCT, however, implies that f and f_n are integrable.

Proposition 4.4.1. *Let f and g be integrable functions and A a measurable set. Some properties of abstract integration are:*

(a) $\int I_A d\mu = \mu(A)$.

(b) If $f \geq 0$, *then* $\int f d\mu \geq 0$.

(c) If $f = 0$ μ-*a.e., then* $\int f d\mu = 0$.

(d) For integrable functions f and g: $\int (f+g) d\mu = \int f d\mu + \int g d\mu$ (additivity).

(e) $\int af d\mu = a \int f d\mu$, *for* $a \in \mathbb{R}$.

Proof. We will prove (d) for the case of non-negative simple functions (the general case is beyond the scope of this text):

Let $f(\omega) = \sum_{i=1}^n a_i I_{A_i}(\omega)$ and $g(\omega) = \sum_{j=1}^m b_i I_{B_j}(\omega)$ be the canonical representations of f and g as simple functions. By definition of a canonical representation, sets A_i are disjoint as are sets B_j. So, the sets $A_i \cap B_j$ are also disjoint. Then

$$(f+g)(\omega) = \sum_{i=1}^n \sum_{j=1}^m (a_i + b_i) I_{A_i \cap B_j}(\omega).$$

$$\begin{aligned}
\int (f+g)(\omega) d\mu &= \sum_{i=1}^n \sum_{j=1}^m (a_i + b_i) \mu(A_i \cap B_j) \\
&= \sum_{i=1}^n a_i \sum_{j=1}^m \mu(A_i \cap B_j) + \sum_{j=1}^m b_i \sum_{i=1}^n \mu(A_i \cap B_j) \\
&= \sum_{i=1}^n a_i \mu(A_i) + \sum_{j=1}^m b_i \mu(B_j) \qquad \text{(marginalize)} \\
&= \int f(\omega) d\mu + \int g(\omega) d\mu
\end{aligned}$$

■

Statements (a-c) and (e) are left as an exercise.

(*) Summation is a special case of integration

Abstract integration also elegantly generalizes sums and infinite series. First, let's define a new measure:

Definition 4.4.2 (Counting function)**.** Let $(\Omega, \mathcal{F})$ be a measurable space. A function $\# : \mathcal{F} \longrightarrow [0, \infty)$ is defined as $\#(A) = |A|$ if $|A|$ is finite and ∞ otherwise.

Proposition 4.4.2. *The counting function $\#$ is a measure on $(\Omega, \mathcal{F})$.*

The proof of this proposition is left as an exercise.

With abstract integration at our disposal, we can interpret sums of finite and infinite sequences as an integral with respect to the counting measure:

Proposition 4.4.3. *Let $a_1, a_2, ...$ be a non-negative infinite sequence and define $f : \mathbb{N} \longrightarrow [0, \infty)$ as $f(i) = a_i$. Then*

$$\sum_{i=1}^{\infty} a_i = \int_{\mathbb{N}} f d\#.$$

Proof. First, we define a sequence of functions $f_n : \mathbb{N} \longrightarrow \mathbb{R}$, such that $f_n(k) = f(k)$ if $k \leq n$ and $f_n(k) = 0$, otherwise. That is, each f_n equals f up to the n-th number and 0 everywhere else beyond that number.

Clearly, f_n converges point-wise to f as n approaches ∞. Furthermore, f_n are non-decreasing, so we can apply the MCT to show

$$\lim_{n\to\infty} \int_{\mathbb{N}} f_n d\# = \int_{\mathbb{N}} \lim_{n\to\infty} f_n d\# = \int_{\mathbb{N}} f d\#.$$

Also

$$\begin{aligned} \int_{\mathbb{N}} f_n d\# &= \int_{\{1\}} f_n d\# + \cdots + \int_{\{n\}} f_n d\# + \int_{\{n+1,n+2,\dots\}} f_n d\# \\ &= f(1) + f(2) + \cdots + f(n) + 0, \end{aligned}$$

because all the individual terms are integrals of constant functions. From this and the above exchange of limit and integral, we have

$$\int_{\mathbb{N}} f d\# = \lim_{n\to\infty} \int_{\mathbb{N}} f_n d\# = \lim_{n\to\infty} \sum_{i=1}^{n} f(i) = \sum_{i=1}^{\infty} f(i).$$

■

Therefore, all the results for abstract integration apply to summation. For example, the exchange of limit and integral, which we already used in the proof above, and criteria for integrability (a series converges if it is absolutely convergent).

Exercises

Exercise 4.1. Prove Proposition 4.3.1.

Exercise 4.2. Prove Proposition 4.4.2. If the general proof for an arbitrary measurable space turns out to be too challenging, try to prove it for the special case of finite Ω or $\Omega = \mathbb{N}$ (both with the power set as the sigma-algebra).

Exercise 4.3. Prove statement (a) in Proposition 4.4.1.

Exercise 4.4. Prove statement (b) in Proposition 4.4.1.

Exercise 4.5. Prove statement (c) in Proposition 4.4.1.

Exercise 4.6. Prove statement (e) in Proposition 4.4.1.

Chapter 5

Random variables

A real-valued random variable (RV) is the fundamental building block of probabilistic expression. It can be thought of as a concise and precise statement about a probability space and the values we are interested in. That is, RVs are tools for precise and unambiguous expression of uncertainty.

In practice we mostly work with random variables and their distributions, because working with them is much easier than working directly with measurable spaces and probability measures.

Introductory probability courses typically focus on two families of RVs: discrete and continuous RVs. In this chapter we will precisely define RVs and show that there exist (infinitely) many RVs that are of practical interest but are neither discrete nor continuous.

5.1 Random variables are measurable functions

Random variable is in a way a very unfortunate name, because random variables are neither random nor variables! They are in fact functions! More precisely, they are measurable functions:

Definition 5.1.1. Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ a measurable space. A real-valued *random variable* X is a function $X : \Omega \to \mathbb{R}$ that is $\mathcal{F}$-measurable.

Recall that measurable (see Definition 4.2.1) in this context means that the preimage X^{-1} of every Borel set is in $\mathcal{F}$. Measurability is necessary - when we equip the measurable space $(\Omega, \mathcal{F})$ with a probability measure, we want to be able to measure the probability of sets of values of our random variable. We now proceed by doing just that. A random variable that maps from a probability space to a new measurable space induces a (new) probability measure on that new space:

Definition 5.1.2. Let $(\Omega, \mathcal{F}, P)$ be a probability space, $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ a measurable space, and X a random variable that maps from one to the other. The *probability law* of X is a function $P_X : \mathcal{B}_{\mathbb{R}} \to [0, 1]$, which is defined as

$$P_X(B) \triangleq P(\{\omega \in \Omega | X(\omega) \in B\}) = P \circ X^{-1}(B)$$

Proposition 5.1.1. *The probability law P_X is a probability measure on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$.*

Proof. $P_X(\emptyset) = 0$ follows from the fact that P_X is a finite measure. $P_X(\mathbb{R}) = P(\{\omega \in \Omega | f(\omega) \in \mathbb{R}\}) = P(\Omega) = 1$. What remains is to show countable additivity. Note that the pre-images under X of two disjoint Borel sets are disjoint, because X maps each element of Ω to a single element of $\mathbb{R}$. Therefore,

$$\begin{aligned} P_X(\bigcup_{i=1}^{\infty} A_i) &= P(\bigcup_{i=1}^{\infty} \{\omega \in \Omega | f(\omega) \in A_i\}) \\ &= \sum_{i=1}^{\infty} P(\{\omega \in \Omega | f(\omega) \in A_i\}) \\ &= \sum_{i=1}^{\infty} P_X(A_i). \end{aligned}$$

■

The probability law of a RV is an example of a *pushforward* measure - a measure that is obtained by *pushing forward* a measure from one measurable space to another using a measurable function.

5.2 Cumulative distribution function

In practical applications, we rarely work directly with probability spaces, σ-algebras, or even probability laws of RVs. Instead, we work with easier to understand and easier to use representations of RVs. The most important such representation is the cumulative distribution function (CDF):

Definition 5.2.1. The *cumulative distribution function* $F_X : \mathbb{R} \to [0, 1]$ of a random variable X on a probability space $(\Omega, \mathcal{F}, P)$ is defined as

$$F_X(x) \triangleq P_X((-\infty, x]) = P(\{\omega \in \Omega | X(\omega) \leq x\}).$$

Note that we often use compact notation $F_X(x) = P(X \leq x)$.

Proposition 5.2.1. *Let X and Y be random variables on $(\Omega, \mathcal{F}, P)$. Then,*

$$P_X(B) = P_Y(B), \forall B \in \mathcal{B}_{\mathbb{R}} \iff F_X(x) = F_Y(x), \forall x \in \mathbb{R}.$$

Proof. In one direction, the proof of the implication is straightforward: The CDF of a RV depends only on its probability law. If two RVs have identical probability laws, they have identical CDFs.

The proof in the other direction is more involved. One way of proving this is to invoke Dynkin's π-λ theorem. A corollary of that theorem is that if two measures agree on a π-system, they agree on a σ-algebra generated by that π-system. A π-system is an even more general set than an algebra. It is a non-empty set of subsets of Ω that is closed under finite intersection. The set of intervals $\{(-\infty, x] : x \in \mathbb{R}\}$ that a CDF is defined on is a π-system, so it follows that if two measures agree on a π-system (have the same CDF), they agree on the entire σ-algebra.

An alternative is to use Caratheodory's theorem in way very similar to our extension of the Lebesgue measure to all Borel sets on the unit interval. First, recall that the intervals $(a, b]$ generate the Borel σ-algebra on $\mathbb{R}$ and their algebra are finite unions of disjoint such intervals. Next, we introduce the (pre)measure $\mu((a, b]) = F(b) - F(a)$ and extend it to finite unions via addition (the measure is the sum of measures of disjoint intervals). This measure is finite. What remains is to show the final condition of Caratheodory's theorem - that the pre-measure is countably additive on the algebra (not trivial to prove!). So, the measure μ uniquely extends to all Borel sets. Because μ represents the probability law on intervals $(a, b]$ and depends only on the CDF, it follows that two RVs with the same CDF have the same probability law on all Borel sets. ■

The above statements say that if two RVs have identical probability laws, they have identical CDFs. And vice-versa, if they have identical CDFs, they have identical probability laws. That is, there is a one-to-one correspondence between the representations of RVs with their probability laws and their representations with CDFs. CDFs are, of course, much simpler and easier to understand representations. It might at first be surprising that the probability measure of all Borel sets can be represented by a single function from $\mathbb{R}$ to $\mathbb{R}$. However, recall that we have already noted that the cardinality of the Borel sets is $\mathbb{R}$.

Properties of CDFs

CDFs have the following properties.

Proposition 5.2.2. *Let X be a random variable with CDF $F_X(\cdot)$. Then $F_X(\cdot)$ has the following properties:*

(a) If $x \leq y$, then $F_X(x) \leq F_X(y)$ (non-decreasing function).

(b) $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.

(c) $\forall x \in \mathbb{R} : \lim_{\epsilon \downarrow 0} F_X(x + \epsilon) = F_X(x)$ (right continuity).

Proof. To prove (a), observe that if $x \leq y$ then $\{\omega \in \Omega | X(\omega) \leq x\} \subseteq \{\omega \in \Omega | X(\omega) \leq y\}$. Then use the property of probability measures in Proposition

1.3.1(c).

We will prove only the second part of (b) as the proof of the first part is symmetric.

$$\begin{aligned}
\lim_{x\to\infty} F_X(x) &= \lim_{x\to\infty} P(X \le x) && \text{(definition of CDF)} \\
&= \lim_{n\to\infty} P(X \le x_n) && \text{(sequence that goes to } \infty) \\
&= P(\bigcup_{n\in\mathbb{N}} \{\omega \in \Omega | X(\omega) \le x_n\}) && \text{(continuity of probability)} \\
&= P(\Omega) \\
&= 1
\end{aligned}$$

Claim (c) is proved in a similar way:

$$\begin{aligned}
\lim_{\epsilon\downarrow 0} F_X(x+\epsilon) &= \lim_{\epsilon\downarrow 0} P(X \le x+\epsilon) = && \text{(definition of CDF)} \\
&= \lim_{n\to\infty} P(X \le x+\epsilon_n) = && \text{(sequence that goes to 0)} \\
&= P(\bigcap_{n\in\mathbb{N}} \{\omega \in \Omega | X(\omega) \le x+\epsilon_n\}) && \text{(continuity of probability)} \\
&= P(X \le x) \\
&= F_X(x)
\end{aligned}$$

■

5.3 Quantile function

Definition 5.3.1 (Quantile function)**.** The generalized inverse $Q : (0,1) \to \mathbb{R}$ of the CDF F is defined as

$$Q(x) \triangleq \inf\{u : F(u) \ge x\}.$$

This generalization is necessary in order to define the inverse on the entire unit interval for CDFs with discontinuities. For continuous CDFs, $Q(x)$ and the ordinary inverse $F^{-1}(x)$ are the same (left as an exercise).

The quantile function plays an important role in practice, both as a means of summarizing the distribution of a random variable (for example, the median, $Q(\frac{1}{2})$) and for generating samples from the distribution via the inverse transform method (see Chapter 19.2). The latter is also interesting from a theoretical

perspective. Through the quantile function we can show that for every CDF F and the standard probability space on the unit interval there exists a RV whose CDF is F. That is, for every distribution there exists a RV with that distribution:

Theorem 5.3.1. *Let F be a function that satisfies the properties of CDFs from Proposition 5.2.2 and let Q be its corresponding quantile function. Consider the uniform probability space $((0,1], \mathcal{B}_{(0,1]}, \lambda)$, where λ is the Lebesgue measure. Let $X : (0,1] \to \mathbb{R}$ be $X(\omega) = Q(\omega)$ for all $\omega \in (0,1)$ and $X(1)$ is set to any real number. Then X is a RV and F is its CDF.*

Proof. The proof that X is a RV is left as an exercise.

$$F_X(x) = P_X\Big((-\infty, x]\Big) = \lambda\Big(\{\omega \in (0,1] | X(\omega) \leq x\}\Big)$$

$$= \lambda\Big(\{\omega \in (0,1] | Q(\omega) \leq x\}\Big) = \lambda\Big((0, F(x)]\Big) = F(x). \qquad \blacksquare$$

So, every function that has the properties of a CDF corresponds to a probability law of a RV.

5.4 Different RVs, same distribution

It is clear that two identical RVs will have identical probability laws and therefore identical CDFs. However, two different RVs can also have identical probability laws and CDFs!

Example 5.4.1. - *Let $((0,1], \mathcal{B}_{(0,1]}, \lambda)$ be our probability space. Let $X : (0,1] \to \mathbb{R}$ such that $X(\omega) = 1$, if $\omega < \frac{1}{2}$, and $X(\omega) = 0$, otherwise. Let $Y : (0,1] \to \mathbb{R}$ such that $Y(\omega) = 1 - X(\omega)$.*

Clearly, X and Y are RVs and they are not identical. However, both X and Y have the same CDF F: $F(x) = 0$, if $x < 0$, $F(x) = \frac{1}{2}$, if $0 \leq x < 1$, and $F(x) = 1$, if $x \geq 1$.

We can look at RV X as a fair coin and RV Y as a coin that always flips the opposite of X. While they both have the same distribution, they are not the same.

This distinction between a RV and its probability law (distribution, CDF) is so important that it deserves its own section.

5.5 Discrete random variables

Now we are ready to introduce discrete and continuous RVs as a special case of our more general treatment of RVs.

Definition 5.5.1. RV X is a *discrete random variable* if there exists a countable subset $B \in \mathcal{B}_{\mathbb{R}}$ such that $P_X(B) = 1$.

As a consequence the probability law of a discrete RV can be uniquely specified by assigning probabilities to at most a countable subset (and 0 everywhere else):

Definition 5.5.2 (Probability mass function (PMF))**.** Let X be a discrete random variable. The function $p_X : \mathbb{R} \to [0, 1]$, $p_X(x) \triangleq P_X(\{x\})$ is called the *probability mass function* of X.

Proposition 5.5.1 (Properties of discrete RVs)**.** *Let X be a discrete RV with PMF p_X and CDF F_X. Then*

(a) There exists a countable subset S of $\mathbb{R}$, such that $\sum_{x \in S} p_X(x) = 1$ and $\forall x \notin S : p_X(x) = 0$.

(b) The PMF completely characterizes the probability law: $P_X(B) = \sum_{x \in B \cap S} p_X(x)$, where S is as in (a).

(c) $F_X(x) = \sum_{x_i \in S : x_i \leq x} p_X(x_i)$, where S is as in (a).

(d) X is discrete $\iff$ $F_X(x)$ is piecewise constant.

Proof. The first part of (a) follows from the definition of a discrete RV. We prove the second part by using countable additivity of probability - a non-zero probability for an x outside of S would imply probability greater than 1. ■

Statements (b), (c), and (d) are left as an exercise.

5.6 Continuous random variables

Continuous RVs are most often defined through the existence of a probability density function (PDF) f_X:

Definition 5.6.1 (Continuous random variable)**.** RV X is a continuous RV if there exists a non-negative measurable function $f_X : \mathbb{R} \to [0, \infty)$, such that for any $B \in \mathcal{B}_{\mathbb{R}}$, we have

$$P_X(B) = \int_B f_X d\lambda,$$

where λ is the Lebesgue measure. Function f_X is the PDF of X.

By this definition, RV X is a continuous RV if there exists a non-negative function that characterizes the RVs probability law through its integral. That is, the probability of a set is the integral over that set.

In a typical first course in probability continuous RVs would be defined using the Riemann integral. The issue with that is that here exist PDFs that would characterize legitimate continuous random variables, but they are not integrable in the Riemann sense. Example 4.2.1 from Chapter 4 illustrates this point. First, it shows that there are functions that are not Riemann integrable. And second and more important, there are legitimate PDFs that are not Riemann integrable. The example is of a function that is almost everywhere the same as a uniform density, which is Riemann integrable. However, there also exist PDFs that are not Riemann integrable and are not almost everywhere the same as a Riemann integrable PDF. However, their construction is out of the scope of this text.

A common alternative but equivalent measure-theoretic definition of a continuous RV is that which is absolutely continuous with respect to the Lebesgue measure:

Definition 5.6.2 (Absolute continuity)**.** A measure μ is said to be absolutely continuous with respect to Lebesgue measure if $\lambda(A) = 0$ implies $\mu(A) = 0$ for every measurable subset A.

The well-known Radon-Nikodym theorem establishes that absolute continuity wrt Lebesgue measure implies the existence of a PDF (or the Radom-Nikodym derivative). There also exist RV whose CDF is continuous but they are not absolutely continuous - we introduce these in the next section.

We can show that the PDF has the following useful properties:

Proposition 5.6.1. *Let X be a continuous RV with PDF f_X. Then,*

(a) $F_X(x) = \int_{-\infty}^{x} f_X(y)d\lambda$.

(b) $\int_{-\infty}^{\infty} f_X(y)d\lambda = 1$.

(c) $f_X(x) \geq 0$, *for all* $x \in \mathbb{R}$.

(d) $F'_X(x) = f_X(x)$, *for almost every x with respect to the Lebesgue measure.*

Proof. The proof of (a-c) is left as an exercise. Statement (d) follows from (a) and the Fundamental theorem of calculus, which has a more general version that applies to Lebesgue-integrable functions (see Theorem 7.11 in Rudin (1987)). ■

The following properties are worth noting regarding continuous RV:

Proposition 5.6.2.

(a) The PDF of a RV does not have to be continuous.

(b) If the CDF of a RV is continuous that does not imply that the RV is continuous.

(c) The PDF of a RV is unique only up to a set of Lebesgue measure 0. This is unlike a CDF where two different CDFs characterize two different RVs.

Proof. A trivial counter-example that proves (a) is the uniform RV on the unit interval - it is discontinuous at 0 and 1. The function from Example 4.2.1 is another example - it is discontinuous everywhere.

Singular distributions that are covered later in this chapter are the counter-example that proves (b). For example, the Cantor distribution has a continuous CDF but is not a continuous RV (it does not admit a PDF).

Part (c) follows from the definition of a continuous RV (see Definition 5.6.1) - if some function f is the PDF then all functions that are almost everywhere the same as f (wrt the Lebesgue measure) will have the same values of the integral. For example, the function from Example 4.2.1 is a PDF of the uniform RV. ■

5.7 Singular random variables

There exists a third pure type of RVs (the other two being discrete and continuous random variables) - singular random variables. Before we define singular RVs, we first define continuous measures.

Definition 5.7.1 (Continuous of a measure)**.** A measure μ is said to be continuous if $\mu(\{\omega\}) = 0$ for every $\omega \in \Omega$.

Proposition 5.7.1. *If probability measure P is continuous then the corresponding CDF is a continuous function.*

Proof. If the measure of every singleton is 0, then the CDF has no jumps and is therefore continuous. ■

Proposition 5.7.2. *If measure μ is absolutely continuous with respect to the Lebesgue measure, it is continuous.*

Proof. The Lebesgue measure of a singleton is 0, so by absolute continuity, $\mu(\{\omega\}) = 0$ for every $\omega \in \Omega$. ■

So, absolute continuity of a measure implies its continuity. However, the converse is not true:

Definition 5.7.2. A random variable is said to be a *singular random variable* if its probability law P_X is a continuous measure and there $\exists A \in \mathcal{B}_{\mathbb{R}} : \lambda(A) = 0$ and $P_X(A) = 1$.

A singular random variable therefore concentrates all of its probability on a set of Lebesgue measure 0 where each element also has probability 0. It is implicit from the definition that this set must be uncountable, because a countable set of elements with probability 0 would not sum up to probability 1. Note that the requirement that all the probability is on a set of Lebesgue measure 0 is necessary in order for this to be a *pure* type. If only part of the probability would be on such a set then the variable would be a mixture of a singular and a continuous RV.

A singular RV is therefore continuous but not absolutely continuous, so it is not a continuous RV in the sense we are used to and it has no PDF (every integral wrt Lebesgue measure would be 0, because all the probability is assigned to a set of Lebesgue measure 0!) It also doesn't have a point mass, so it is not a discrete RV. It would be more precise to say that we have discrete and continuous RVs and that continuous RVs are further subdivided into continuous and absolutely continuous.

An example of a singular RV is the Cantor distribution.

5.8 Decomposition of probability measures

Theorem 5.8.1 (Decomposition theorem)**.** *Every CDF F can written as*

$$F = w_1 F_{continuous} + w_2 F_{discrete} + w_3 F_{singular},$$

where $w_i \geq 0$, $\sum w_i = 1$, and the CDFs on the right-hand side correspond to a continuous, discrete, and singular RV.

The theorem states that every RV is a combination of the three pure types. While singular random variables are only of theoretical interest, RVs that are a mix of a discrete and continuous RV are very common in practice.

Example 5.8.1. - *In one particular university course the scores of students that take the exam are uniform over $(0, 100\%]$. However, there is also a 0.2 probability that a student does not even attend the exam in which case he automatically receives a final score of 0%. Let the RV X be the final score received by a student. What is the CDF of X?*

This RV is not continuous, because its CDF is not absolutely continuous - it has a 0.2 point-mass on a the set $\{0\}$, which is a singleton and therefore has Lebesgue measure 0. It is also not discrete, because only 0.2 of the probability is concentrated on a countable subset of $\mathbb{R}$. And it is not singular, because it does not concentrate all of its probability on a set of Lebesgue measure 0 and the probability is not 0 for every singleton. So, X is not of a pure type - it is a mixture of a discrete and a continuous random variable.

5.9 Functions of random variables

In practice, we will often be interested not just in random variables, but also functions of random variables. The following proposition states that in most cases, but not always, the function of a random variable will again be a random variable:

Proposition 5.9.1. *Let $X : \Omega \to \mathbb{R}$ be a RV. Let $g : \mathbb{R} \to \mathbb{R}$ be a Borel measurable function. Then, $Y = g(X)$ is also a random variable.*

Proof. X is a RV and therefore, by definition, a measurable function. Therefore, we only need to show that the composition of two measurable functions is measurable. Because g is Borel measurable, the pre-image of a Borel set under g will be a Borel set. And, because X is measurable, the pre-image of a Borel set under X will be measurable with respect to our σ-algebra. Therefore, the pre-images of Borel sets under the composition $g \circ X(\cdot)$ will also be measurable with respect to our σ-algebra. ■

This proposition also suggests how to pick such a function g that $g(X)$ will not be a RV - we require a function that is not Borel measurable. That is, a function, such that the pre-image of a Borel set is not a Borel set. One possible choice would be a function that maps the elements of the Vitali set (see Theorem 2.1.1) to 5 and the rest to 0 - the probability of 5 would then not be measurable. This example is very theoretical. In practice, most of our functions will be Borel measurable and we will rarely be wrong in assuming that the function of a random variable is a random variable. In particular, all continuous functions are Borel measurable.

In general, the CDF of a transformed random variable can be computed as

$$F_Y(y) = P(g(X) \leq y) = P(\{\omega | g(X(\omega)) \leq y\}) = P_X(B_y),$$

where B_y is the set of all x, such that $g(x) \leq y$. This by itself is not very useful, but there are two special cases where it is easier to compute the transformed RV.

Proposition 5.9.2 (Transformation of a discrete RV)**.** *Let X be a RV, $g : \mathbb{R} \to \mathbb{R}$ a Borel measurable function, and $Y = g(X)$. If X is discrete, then:*

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

The proof is left as an exercise.

Example 5.9.1. *Let X be a discrete RV whose probability law represents a fair 6-sided die. That is $p_X(i) = \frac{1}{6}$, $i = 1..6$. Let $g(x) = (x-3)^2$. What is the probability law of RV $Y = g(X)$?*

X is a discrete RV so Y is also a discrete RV, because its support can't be more than that of X (a function maps a value to a single other value). The values produced by g from $1..6$ are $4, 1, 0, 1, 4$, and 9. So, Y has non-zero probabilities for 0, 1, 4, and 9: $p_Y(1) = p_Y(4) = \frac{2}{6}$ and $p_Y(0) = p_Y(9) = \frac{1}{6}$.

Proposition 5.9.3 (Transformation of a continuous RV)**.** *Let X be a continuous RV, $g : \mathbb{R} \to \mathbb{R}$ a Borel measurable, monotone increasing or monotone decreasing, and continuously differentiable function, and $Y = g(X)$. Then,*

$$f_Y(y) = f_X(g^{-1}(y))|\frac{d}{dy}g^{-1}(y)|,$$

almost everywhere with respect to the Lebesgue measure.

Proof. First note that the purpose of the conditions regarding g in the statement of the theorem are for g to have an inverse and for that inverse to be differentiable. To be more precise, non-differentiable at most on a set of Lebesgue measure 0, so that Y has a PDF. Another consequence of g^{-1} being differentiable almost everywhere is that the statement in the Proposition is also true almost everywhere.

Monotonicity also simplifies the relationship between the CDF of X and the CDF of Y. If g is a monotone increasing function, we have

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)).$$

Note that the above is useful on its own, because it applies to any RV and monotone increasing g.

Differentiating both sides with respect to y, we get

$$f_Y(y) = f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y).$$

Similarly, if g is a monotone decreasing function, we get

$$F_Y(y) = P(X \geq g^{-1}(y)) = 1 - P(X < g^{-1}(y)) = 1 - F_X(g^{-1}(y)).$$

Again, the above is useful on its own, because it applies to any RV and monotone decreasing g.

Differentiating both sides with respect to y, we get

$$f_Y(y) = -f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y).$$

However, because g is monotone decreasing, so is g^{-1}. The derivative is therefore negative and the increasing and decreasing cases can be summarized as

$$f_Y(y) = f_X(g^{-1}(y))|\frac{d}{dy}g^{-1}(y)|.$$

■

Example 5.9.2. *Let X have a continuous uniform distribution on the unit interval. Let $g(x) = e^x$. What is the PDF of RV $Y = g(X)$?*

For the uniform RV on the unit interval we have $f_X(x) = 1$. Function g is increasing, has an inverse $g^{-1}(x) = \log(x)$, and its inverse is differentiable. We have

$$f_Y(y) = f_X(g^{-1}(y))|\frac{d}{dy}g^{-1}(y)| = 1|\frac{1}{y}| = \frac{1}{y},$$

because we have $y \in [e^0, e^1] = [1, e]$.

We can check if f_Y is indeed a PDF. It is non-negative and the integral is

$$\int_1^e \frac{1}{x}dx = |\log(x)|_1^e = 1.$$

Exercises

Exercise 5.1. Prove the part of Theorem 5.3.1 that X is a RV.

Exercise 5.2. Prove statements (b-d) in Proposition 5.5.1.

Exercise 5.3. Prove statements (a-c) in Proposition 5.6.1.

Exercise 5.4. Prove that the inverse of the CDF and the generalized inverse from Definition 5.3.1 are equivalent for continuous RVs. Give an example that demonstrates that they are not equivalent for RVs with discontinuities in the CDF.

Exercise 5.5. Prove Proposition 5.9.2.

Chapter 6

Multiple random variables

6.1 Measure-theoretic background

Before we can talk about joint probability laws and CDF of two or more RVs, we must extend our understanding of probability spaces and integration to $\mathbb{R}^2$. Instead of assigning probabilities to subsets of the real line, we now have to assign probabilities to subsets of $\mathbb{R}^2$:

Definition 6.1.1 (Product σ-algebra)**.** Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be measurable spaces. The *product σ-algebra* is a σ-algebra for the corresponding product space $\Omega_1 \times \Omega_2$ and is defined as the σ-algebra generated by the rectangles $A_1 \times A_2$:

$$\mathcal{F}_1 \otimes \mathcal{F}_2 = \sigma(\{A_1 \times A_2 : A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}).$$

Now we will extend our favorite σ-algebra to $\mathbb{R}^n$ and show that it is a product σ-algebra.

Definition 6.1.2 (Borel σ-algebra)**.** The Borel σ-algebra on $\mathbb{R}^n$ is the σ-algebra generated by these sets of hyper-rectangles

$$\begin{aligned}\mathcal{B}_{\mathbb{R}^n} &= \sigma(\{(-\infty, b_1) \times (-\infty, b_2) \times \ldots \times (-\infty, b_n) : b_i \in \mathbb{R}\}) \\ &= \sigma(\{(a_1, b_1] \times (a_2, b_2] \times \ldots \times (a_n, b_n] : a_i, b_i \in \mathbb{R}\}) \\ &= \sigma(\{(a_1, b_1) \times (a_2, b_2) \times \ldots \times (a_n, b_n) : a_i, b_i \in \mathbb{R}\}).\end{aligned}$$

We will not prove the equivalence of these definitions, but the argument is similar to the arguments we used in the case of Borel sets on $(0, 1]$ and $\mathbb{R}$.

Now we can show that the Borel σ-algebra on $\mathbb{R}^n$ is a product algebra. And not only that, it is a product of n copies of Borel σ-algebras on $\mathbb{R}$!

Proposition 6.1.1.

$$\mathcal{B}_{\mathbb{R}^n} = \underbrace{\mathcal{B}_{\mathbb{R}} \otimes \mathcal{B}_{\mathbb{R}} \otimes \ldots \otimes \mathcal{B}_{\mathbb{R}}}_{n}.$$

Proof. For simplicity, will prove it for $n = 2$. The proof extends to n via induction.

First, we show that $\mathcal{B}_{\mathbb{R}^2} \subseteq \mathcal{B}_{\mathbb{R}} \otimes \mathcal{B}_{\mathbb{R}}$. By definition of a product σ-algebra, the rectangle $(a, b) \times (c, d)$ belongs to $\mathcal{B}_{\mathbb{R}} \otimes \mathcal{B}_{\mathbb{R}}$. Because the set of all such rectangles is a generating set of $\mathcal{B}_{\mathbb{R}^2}$ the product σ-algebra will contain at least all the sets in $\mathcal{B}_{\mathbb{R}^2}$.

To complete the proof, we show that $\mathcal{B}_{\mathbb{R}} \otimes \mathcal{B}_{\mathbb{R}} \subseteq \mathcal{B}_{\mathbb{R}^2}$. We introduce $\mathcal{F}$, a set of all subsets A of $\mathbb{R}$ such that $A \times \mathbb{R} \in \mathcal{B}_{\mathbb{R}^2}$. Note that $\mathcal{F}$ is a σ-algebra (why?). $\mathcal{F}$ also contains all intervals (a, b), because $(a, b) \times \mathbb{R}$ is in $\mathcal{B}_{\mathbb{R}^2}$, So, $\mathcal{B}_{\mathbb{R}} \subseteq \mathcal{F}$ and for every $A \in \mathcal{B}_{\mathbb{R}}$ we have $A \times \mathbb{R} \in \mathcal{B}_{\mathbb{R}^2}$.

Similarly, we can derive that for every $B \in \mathcal{B}_{\mathbb{R}}$ we have $\mathbb{R} \times B \in \mathcal{B}_{\mathbb{R}^2}$. It follows that

$$A \times B = (A \cap \mathbb{R}) \times (\mathbb{R} \cap B) = (A \times \mathbb{R}) \cap (\mathbb{R} \times B) \in \mathcal{B}_{\mathbb{R}^2}.$$

So, the σ-algebra generated by the set of rectangles $A \times B$, where $A, B \in \mathcal{B}_{\mathbb{R}}$ is a subset of $\mathcal{B}_{\mathbb{R}^2}$. ■

Definition 6.1.3 (Product measure)**.** Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be measure spaces and let $\mathcal{F}_1 \otimes \mathcal{F}_2$ be the product σ-algebra of their product space $\Omega_1 \times \Omega_2$. A *product measure* $\mu_1 \times \mu_2$ is a measure on the measurable space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ satisfying the property $\mu_1 \times \mu_2(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$, for all $A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2$.

Note that if the measure spaces in the definition of the product measure are σ-finite then the product measure exists, is unique, and is also σ-finite. We will not concern ourselves with the technical details and for us it will suffice that there exists a product Lebesgue measure. The product of two probability measures will also be of some interest, because of its connection with independence of RVs.

Example 6.1.1 (Lebesgue integration on $\mathbb{R}^2$)**.** *Let $(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2}, \lambda_1 \times \lambda_2)$ be our measurable space. So far in this chapter we have established that $\mathcal{B}_{\mathbb{R}^2}$ is a product σ-algebra (and a product of two Borel σ-algebras) and that the product measure $\lambda_1 \times \lambda_2$ exists and is unique (we know λ exists on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ and we have noted that it is σ-finite, although we did not go into the details of what that means, other than that it is a condition less strict than finite).*

By definition, we have $(\lambda_1 \times \lambda_2)(B_1, B_2) = \lambda_1(B_1)\lambda_2(B_2)$, so if the Lebesgue measure λ is a generalization of length then the product measure of two Lebesgue

measures can be seen as a generalization of area. For example, let's integrate function $f(x_1, x_2)$ which has value 1 on the unit square $[0,1]^2$ and 0 otherwise:

$$\int_{\mathbb{R}^2} f d(\lambda_1 \times \lambda_2).$$

Any product measure is by definition a measure, so we are already equipped to integrate wrt to a product measure. Function f is a simple function that takes only two possible values - 0 and 1. The measure of the set where it has value 1 is $\lambda_1 \times \lambda_2([0,1]^2) = \lambda_1([0,1])\lambda_2([0,1]) = 1$. The measure of the set where it has value 0 will be multiplied by 0, so it does not affect the value of the integral:

$$\int_{\mathbb{R}^2} f d(\lambda_1 \times \lambda_2) = 1.$$

It might be somewhat surprising that we do not have to introduce any additional theory for integration over 2 (or more) dimensions. However, note that dimensionality of the sample space does not play a role in the definition of abstract integration. We have measurable spaces and we measure (and integrate over) subsets of those measurable spaces. The product measure $\lambda \times \lambda$ does measure 2-dimensional sets, but it is just a measure. And $B = [0,1]^2$ is a 2-dimensional set, but it is still a set. The definition of the abstract integral does not depend on the dimensionality of the sample space (or any partition that would depend on its dimensionality), we only observe subsets with the same value.

In practice it is in most cases more convenient to integrate a function of 2 or more variables first wrt one variable, then wrt another, etc. A very important theorem from measure theory, which we state without proof, states that:

Theorem 6.1.1 (Tonelli/Fubini)**.** *Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be measure spaces, where μ_1 and μ_2 are σ-finite measures. Let $\mu = \mu_1 \times \mu_2$ be the product measure on the measurable space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$. Let $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$ be a measurable function. If $\int |f| d(\mu_1 \times \mu_2) < \infty$ then*

$$\int f d\mu = \int \left[\int f(\omega_1, \omega_2) d\mu_1(\omega_1) \right] d\mu_2(\omega_2) = \int \left[\int f(\omega_1, \omega_2) d\mu_2(\omega_2) \right] d\mu_1(\omega_1).$$

So, when integrating with respect to a product measure and if the function is integrable, we can integrate wrt one measure and then wrt the other. Note that all the definitions in this chapter so far generalize from 2 measures to n measures via induction. That is, a product measure of 3 measures is just a product of a product measure and a measure, etc.

6.2 Joint probability laws and CDFs

So far in this text we have only observed a single RV at a time. Now we are ready to extend this to multiple RVs (random vector, random matrix). The chapter focuses on joint distributions of two RVs. However, all the results readily generalize to three or more random variables.

First, we must ask the following question. If X and Y are RVs on the same probability space $(\Omega, \mathcal{F}, P)$, is $(X(\cdot), Y(\cdot)) : \Omega \to \mathbb{R}^2$ also a random vector on that probability space? The following theorem says that the answer is yes.

Theorem 6.2.1. *Let X and Y be random variables on probability space $(\Omega, \mathcal{F}, P)$. Then, $(X(\cdot), Y(\cdot)) : \Omega \to \mathbb{R}^2$ is $\mathcal{F}$-measurable.*

Proof. Let $\mathcal{F}_2 = \{S \subseteq \mathbb{R}^2 \colon (X^{-1}(S), Y^{-1}(S)) \in \mathcal{F}\}$. $\mathcal{F}_2$ contains all measurable rectangles $A \times B$, their unions, and their complements, because X and Y are measurable. So $\mathcal{F}_2$ is a σ-algebra that contains the Borel σ-algebra. ■

The probability law and CDF generalize to two or more RVs:

Definition 6.2.1. The joint probability law of RVs X and Y is defined as

$$P_{X,Y}(B) \triangleq P(\{\omega \in \Omega | (X(\omega), Y(\omega)) \in B\}), B \in \mathcal{B}_{\mathbb{R}^2}.$$

Definition 6.2.2. The joint CDF of two random variables is defined as

$$F_{X,Y}(x, y) \triangleq P_{X,Y}((-\infty, x] \times (-\infty, y]) = P(\{\omega \in \Omega | X(\omega) \leq x, Y(\omega) \leq y\}).$$

Typically, we use the more concise notation $F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$.

Proposition 6.2.1. *Properties of joint CDF:*

(a) $\lim_{x \to \infty, y \to \infty} F_{X,Y}(x, y) = 1.$

(b) $\lim_{x \to -\infty, y \to -\infty} F_{X,Y}(x, y) = 0.$

(c) For any $x_1 \leq x_2, y_1 \leq y_2$ we have $F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2)$ (nondecreasing).

(d) $\forall x, y \in \mathbb{R} : \lim_{u \downarrow 0, v \downarrow 0} F_{X,Y}(x + u, y + v) = F_{X,Y}(x, y)$ *(continuity from above).*

(e) $\lim_{y \to \infty} F_{X,Y}(x, y) = F_X(x)$ *and* $\lim_{x \to \infty} F_{X,Y}(x, y) = F_Y(y).$

The proof of this proposition is left as an exercise.

6.3 Independence of random variables

Recall that events A and B are said to be independent if $P(A \cap B) = P(A)P(B)$. Using this definition, we can extend the notion of independence to RVs.

Definition 6.3.1. RVs X and Y are said to be independent if for any two Borel sets $B_1, B_2 \in \mathcal{B}_{\mathbb{R}}$ the events $\{\omega : X(\omega) \in B_1\}$ and $\{\omega : Y(\omega) \in B_2\}$ are independent. That is $P(\{\omega : X(\omega) \in B_1, Y(\omega) \in B_2\}) = P(\{\omega : X(\omega) \in B_1\})P(\{\omega : Y(\omega) \in B_2\})$.

If we state this in terms of the probability laws of X and Y, we get a simpler but equivalent definition:

RVs X and Y are said to be independent if and only if $P_{X,Y}(B_1 \times B_2) = P_X(B_1)P_Y(B_2)$ for any two Borel sets $B_1, B_2 \in \mathcal{B}_{\mathbb{R}}$.

We can also view the above definition in terms of product measures. Two RVs are independent if and only if their joint probability law is a product measure of their individual probability laws!

Proposition 6.3.1. *RVs X and Y are independent if and only if $F_{X,Y}(x, y) = F_X(x)F_Y(y)$.*

Proof. First, let's show that independence implies that the joint CDF factors. $F_{X,Y}(x, y) = P_{X,Y}((-\infty, x] \times (-\infty, y]) = P_X((-\infty, x])P_Y((-\infty, y]) = F_X(x)F_Y(y)$.

Now let us show that the CDF factoring implies independence.

We have to prove that for every $A, B \in \mathbb{B}_{\mathbb{R}}$ the following holds:

$$P_{X,Y}(A \times B) = P_X(A)P_Y(B). \tag{6.1}$$

Let us fix B of the form $(-\infty, y]$, $y \in \mathbb{R}$. We define two measures $\mu_B, \nu_B : \mathbb{B}_{\mathbb{R}} \to \mathbb{R}$: $\mu_B(A) := P_{X,Y}(A \times B$ and $\nu_B(A) := P_X(A)P_Y(B)$.

Because of the assumption that the CDF factors, we have that for every A of the form $(-\infty, x]$, $x \in \mathbb{R}$: $\mu_B(A) = \nu_B(A)$. Since μ_B and ν_B are σ-finite, $\mu_B = \nu_B$ by the uniqueness from Caratheodory's theorem. So, (6.1) holds for every $A \in \mathbb{B}_{\mathbb{R}}$ and every B of the form $(-\infty, y]$.

Now we fix $A \in \mathbb{B}_{\mathbb{R}}$ and define two measures $\mu^A, \nu^A : \mathbb{B}_{\mathbb{R}} \to \mathbb{R}$: $\mu^A(B) := \mu_B(A)$ and $\nu^A(B) := \nu_B(A)$. Since by the previous paragraph, μ^A and ν^A coincide on $(-\infty, y]$, $y \in \mathbb{R}$ and are σ-finite, it follows that $\mu^A = \nu^A$ on $\mathbb{B}_{\mathbb{R}}$. Since $A \in \mathbb{B}_{\mathbb{R}}$ was arbitrary, it follows that (6.1) holds for every $A, B \in \mathbb{B}_{\mathbb{R}}$.

■

The definition of independence and all statements so far in this section can be generalized to countably many RVs. With 3 or more RVs we again, analogous to 3 or more events, have to distinguish between pairwise and joint independence

(joint implies pairwise, but the converse is not true). We only state the result that will be most useful to us:

Theorem 6.3.1. *Let $\{X_i\}$ be a countable set of RVs. RVs X_i are jointly independent if and only if*

$$F_{X_1,\dots}(x_1,\dots) = \prod_{i=1}^{\infty} F_{X_i}(x_i).$$

We state this theorem without proof, but the argument is similar to the one for 2 RVs.

6.4 Jointly discrete random variables

So far, our treatment has been general, at the level of probability laws and CDFs, which every RV has. In the case of jointly discrete or jointly continuous random variables, more specific and thus more useful results can be obtained.

It is a well-known result that the Cartesian product of two countable sets is countable. Therefore, the joint distribution of two discrete random variables is also discrete.

Definition 6.4.1 (Joint probability mass function)**.** Let X and Y be discrete random variables. The function $p_{X,Y} : \mathbb{R}^2 \to [0,1]$,

$$p_{X,Y}(x,y) \triangleq P(X = x, Y = y)$$

is called the *joint probability mass function* of X and Y.

The PMF of a jointly discrete RV completely characterizes its probability law $P_{X,Y}(B) = \sum_{(x,y)\in B; p_{X,Y}(x,y)>0} p_{X,Y}(x,y)$ and the marginal probability laws $p_X(x) = \sum_y p_{X,Y}(x,y)$ and $p_Y(y) = \sum_x p_{X,Y}(x,y)$.

Definition 6.4.2. Let X and Y be discrete random variables. The *conditional probability* of X given Y is defined as

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

where $p_Y(y) > 0$.

Theorem 6.4.1. *Discrete RVs X and Y are independent if and only if $\forall x, y \in \mathbb{R} : p_{X,Y}(x,y) = p_X(x)p_Y(y)$.*

Proof. If X and Y are independent, then $P(X \in B_1, Y \in B_2) = P(X \in B_1)P(Y \in B_2)$, for any B_1 and B_2, including $B_1 = \{x\}$, $B_2 = \{y\}$. Hence, $p_{X,Y}(x,y) = p_X(x)p_Y(y)$.

If $p_{X,Y}(x,y) = p_X(x)p_Y(y)$, we have

$$\begin{aligned} P(X \in B_1, Y \in B_2) &= \sum_{x \in B_1} \sum_{y \in B_2} p_{X,Y}(x,y) = \sum_{x \in B_1} \sum_{y \in B_2} p_X(x) p_Y(y) \\ &= \sum_{x \in B_1} p_X(x) \sum_{y \in B_2} p_Y(y) = P(X \in B_1) P(Y \in B_2). \end{aligned}$$

To simplify notation, we omit that we restrict ourselves to B_1 and B_2 such that $p_{X,Y}(x,y) > 0$.

What remains is to show that independence on these intervals is enough to imply independence on the entire σ-algebra.

■

6.5 Jointly continuous random variables

Similarly to a continuous RV the jointly continuous RVs are defined through the existence of a joint probability density function.

Definition 6.5.1 (Jointly continuous RVs)**.** X and Y are jointly continuous if there exists a non-negative measurable function $f_{X,Y} : \mathbb{R}^2 \to [0, \infty)$, such that for any $B \in \mathcal{B}_{\mathbb{R}^2}$, we have

$$P_{X,Y}(B) = \int_B f_{X,Y}(x,y) d(\lambda(x) \times \lambda(y)).$$

We call $f_{X,Y}$ the *joint probability density function.*

Note that unless we explicitly state otherwise, we will in the remainder of the book assume that we are integrating with respect to the Lebesgue measure. For example, dx, dy, du, dv will be shorthand for $d\lambda(x)$, $d\lambda(y)$, etc.

The joint PDF is a complete characterization of the joint distribution:

Proposition 6.5.1. *Let X and Y be jointly continuous. Then,*

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v) dv du.$$

The marginal PDFs can be derived from the joint PDF: $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$ and $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$.

Proof. By definition, $F_{X,Y}(x,y) = P_{X,Y}((-\infty, x] \times (-\infty, y])$. The result can be obtained from the definition of the joint probability density function by setting $B = (-\infty, x] \times (-\infty, y]$ and then applying Tonelli's theorem. ■

Theorem 6.5.1. *Jointly continuous RVs X and Y are independent if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ almost everywhere.*

Proof. By Proposition 6.3.1 we have $F_{X,Y}(x,y) = F_X(x)F_Y(y)$, $\forall x, y \in \mathbb{R}$. Inserting for continuous RVs, we get:

$$\begin{aligned}\int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)dvdu &= \left(\int_{-\infty}^{x} f_X(u)dv\right)\left(\int_{-\infty}^{y} f_Y(v)du\right) \\ &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_X(u)f_Y(v)dudv\end{aligned}$$

This implies that $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ almost everywhere. the set where this does not hold has Lebesgue measure 0 (see Chapter 4 where *almost everywhere* is defined). ■

Definition 6.5.2. The conditional PDF of continuous RV X given continuous RV Y is defined as

$$f_{X|Y}(x|y) \triangleq \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

for $f_Y(y) > 0$.

While the joint distribution of two discrete random variables is always discrete, the joint distribution of two continuous random variables is not always continuous. However, jointly continuous RVs are marginally continuous. The proof of these statements is left as an exercise.

6.6 Mixed joint density

Often we are interested in joint distributions of discrete and continuous RVs. For most practical purposes concerning marginals and conditionals, the PDF and PMF play an identical role. To avoid the technical details, we define the joint PDF-PMF and conditional for a discrete and a continuous RV.

Definition 6.6.1. The joint PDF-PMF of a continuous RV X and a discrete RV Y is defined as

$$f_{X,Y}(x,y) \triangleq f_{X|Y}(x|y)p_y(y) = p_{Y|X}(y|x)f_X(x),$$

where

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

and

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

Note that the marginal PDF of X can be obtained by summation over Y and the marginal PMF of Y by integration over X, analogous to jointly discrete and jointly continuous RVs.

Exercises

Exercise 6.1. Prove statements (a) and (b) from Proposition 6.2.1.

Exercise 6.2. Prove statement (c) from Proposition 6.2.1.

Exercise 6.3. Prove statement (d) from Proposition 6.2.1.

Exercise 6.4. Prove statement (e) from Proposition 6.2.1.

Exercise 6.5. Give an example where the joint distribution of two continuous RVs is not jointly continuous.

Exercise 6.6. Show that the marginals of a jointly continuous RV are continuous.

Chapter 7

Expected value

The expected value, expectation, or mean of a RVs is arguably the most important single-value representation of a probability distribution. Geometrically, it is the center of mass, and, unless the distribution is very skewed or multimodal, it will be a good summary of its location. Through the law of large numbers, the expected value is also related to the sample average. Expected value together with variance represent a complete representation of a normal distribution, which is most often parametrized with its mean and variance.

7.1 Definition of expectation

The expectation (expected value or mean) of a function of a random variable on a probability space is defined as the integral of the composition of that function and random variable with respect to the probability measure.

Definition 7.1.1. Let $(\Omega, \mathcal{F}, P)$ be a probability space, $X : \Omega \to \mathbb{R}$ a random variable and $g : \mathbb{R} \to \mathbb{R}$ a measurable function. The expected value of $g(X)$ is defined as

$$E[g(X)] \triangleq \int_{\Omega} g(X(\omega))dP(\omega),$$

if the integral exists.

If g is the identity, this simplifies to the expectation of the RV X:

$$E[X] \triangleq \int_{\Omega} X(\omega)dP(\omega).$$

Example 7.1.1 (Expected value of a Bernoulli RV)**.** *Later in this chapter we will derive the expected value of a Bernoulli distribution using the much more convenient definition for the special case of discrete RVs, which is derived from this more general definition.*

However, to illustrate the use of the general definition, let $((0,1], \mathbb{B}_{(0,1]}, \lambda)$ *be our probability space and* $X : (0,1] \to \{0,1\}$*, such that* $X((0,\theta]) = 1$ *and* $X((\theta,1]) = 0$*,* $\theta \in (0,1)$*. Clearly, the probability law of* X *is Bernoulli*(θ)*.*

Using the definition of expectation and observing that X *is a non-negative simple function, we have*

$$E[X] = \int_\Omega X(\omega)dP(\omega) = \int_{(0,1]} X(\omega)dP(\omega) = 1\lambda((0,\theta]) + 0\lambda((\theta,1]) = \theta + 0 = \theta.$$

Alternatively, we can derive the expected value of a RV X from its probability law $P_X(A) = P(X^{-1}(A))$:

Proposition 7.1.1. $E[X] = \int_{\mathbb{R}} x dP_X(x)$.

Proof. It is enough to prove that

$$E(X_+) = \int_{\mathbb{R}} x_+ dP_X.$$

Let $s_n(x) : \mathbb{R} \to [0,\infty)$ be a sequence of simple non-negative measurable functions such that $s_n \le s_{n+1} \le s_{n+2} \le \dots$ and $\lim_{n\to\infty} s_n(x) = x_+$.

Write $s_n = \sum_{i=1}^{m_n} s_{n,i} I_{A_i}$ and define $s'_n = \sum_{i=1}^{m_n} s_{n,i} I_{X^{-1}(A_i)}$.

Clearly $s'_n \le s'_{n+1} \le \dots$ and $\lim_{n\to\infty} s'_n(\omega) = X(\omega)$.

By MCT,

$$\lim_{n\to\infty} \int_\Omega s'_n dP = \int_\Omega X_+ dP,$$
$$\lim_{n\to\infty} \int_{\mathbb{R}} s_n dP_X = \int_{\mathbb{R}} x_+ dP_X.$$

From

$$\int_\Omega s'_n dP = \sum_i s_{n,i} P(X^{-1}(A_i)) = \sum_i s_{n,i} P_X(A_i) = \int_{\mathbb{R}} s_n dP_X,$$

it follows that the initial statement holds.

■

The two definitions above apply to all RVs. However, they are not very useful in practice. From it we can derive the more familiar and practically useful definitions for discrete and continuous RVs.

Proposition 7.1.2. *The expected value of a function of a discrete random variable X is*

$$E[g(X)] = \sum_{i=1}^{\infty} g(x_i) p_X(x_i),$$

where $x_1, x_2, \cdots$ are values that X attain.

Proof. For now, we'll assume that $g(X)$ is non-negative. Because X is a discrete RV, $g(X)$ has countably many values, so we can partition Ω into a countable number of parts A_i, such that all ω with the same value of $g(X(\omega)) = a_i$ are in the same partition. Then we can write:

$$g(X(\omega)) = \sum_{i=1}^{\infty} a_i I_{A_i}(\omega).$$

This is not a simple function representation due to the countable number of terms, but we will approximate it with the following sequence of functions.

$$g(X)_n(\omega) \triangleq \sum_{i=1}^{n} a_i I_{A_i}(\omega).$$

The sequence of functions $g(X)_n$ is non-decreasing and it is easy to check that $\lim_{n\to\infty} g(X)_n(\omega) = g(X(\omega))$. So, the MCT applies and the integral of X equals the limit of the integral of X_n:

$$\begin{aligned} E[g(X)] &= \lim_{n\to\infty} E[g(X)_n] = \lim_{n\to\infty} \sum_{i=1}^{n} a_i P(A_i) \\ &= \lim_{n\to\infty} \left[\sum_{i=1}^{n} g(x_i) \sum_{x_i : g(x_i) = a_i} P(X = x_i) \right] = \sum_{j=1}^{\infty} g(x_j) p_X(x_j). \end{aligned}$$

To generalize this to arbitrary $g(X)$, we can split the RV into a positive and negative part, do each separately (in absolute terms, to ensure non-negativity) and take the difference. The expectation will be defined if either part is not infinite. ∎

Example 7.1.2 (Expected value of a Bernoulli RV)**.** *Let $X \sim Bernoulli(\theta)$. The pmf of this discrete distribution, whose support is 0 and 1, is $p(1) = \theta$ and $p(0) = 1 - \theta$. Using the definition of the expected value, we have:*

$$E[X] = \sum_{i=0}^{1} ip(i) = 0 \cdot (1 - \theta) + 1 \cdot \theta = \theta.$$

Example 7.1.3 (Expected value of a Poisson RV)**.** *Let $X \sim Poisson(\lambda)$. The pmf of this discrete distribution, whose support is on non-negative integers, is*

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

Using the definition of the expected value, we have:

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

Note that in the 2nd equality we took out $\lambda e^{-\lambda}$, canceled the k, and ignored the $k = 0$ term of the series, which is 0. The 4th equality uses the fact that the series is a Taylor series expansion of the exponential.

Proposition 7.1.3. *The expected value of a function of a continuous random variable X is*

$$E[g(X)] = \int_{\mathbb{R}} g(x) f_X(x) d\lambda(x) = \int_{\mathbb{R}} g(x) f_X(x) dx,$$

where λ is the Lebesgue measure. Note that the second equality only applies when $g(x)f_X(x)$ is Riemann integrable.

Proof sketch without $g(\cdot)$. This follows from the definition of the density:

$$P_X(B) = \int_B f_X(x) d\lambda(x)$$

hence $dP_X(x) = f_X(x)d\lambda(x)$, which we substitute into

$$E[X] = \int_{\mathbb{R}} x dP_X(x)$$

from Proposition 7.1.1.

■

Example 7.1.4 (Expected value of Exponential RV)**.** *Let $X \sim Exp(\lambda)$. The pdf of this continuous distribution, whose support on the positive reals, is $f(x) = \lambda e^{-\lambda x}$. Using the definition of the expected value, we have:*

$$\begin{aligned}
E[X] &= \int_0^\infty x\lambda e^{-\lambda x} dx \\
&= \lambda \int_0^\infty x e^{-\lambda x} dx \\
&= \lambda \left(-\frac{x}{\lambda} e^{-\lambda x} \Big|_0^\infty - \int_0^\infty -\frac{1}{\lambda} e^{-\lambda x} dx \right) \\
&= -x e^{-\lambda x} \Big|_0^\infty + \int_0^\infty e^{-\lambda x} dx \\
&= -x e^{-\lambda x} - \frac{1}{\lambda} e^{-\lambda x} \Big|_0^\infty \\
&= \frac{1}{\lambda}.
\end{aligned}$$

In line 3 we used integration by parts with $u = x$ and $dv = e^{-\lambda x} dx$, so $du = dx$ and $v = -\frac{1}{\lambda} e^{-\lambda x}$.

Being an integral, expected value might not be integrable (for example, the Cauchy distribution) or might be integrable but infinite (for example, the Pareto distribution for some parameter values). That is, there exist RVs with an undefined expected value and RVs with infinite expectation.

Example 7.1.5 (RVs with infinite or undefined mean)**.** *Let X be a discrete RV that takes values 2^i, where i is a positive integer. Its pmf is $p(2^i) = \frac{1}{2^i}$. We have $\sum_{i=1}^\infty \frac{1}{2^i} = 1$, so X is indeed a RV.*

The expectation of X is

$$E[X] = \sum_{i=1}^\infty 2^i \frac{1}{2^i} = \sum_{i=1}^\infty 1 = \infty,$$

so X has an infinite expectation, even though it can only attain finite (but arbitrarily large) values.

We can use the same idea to define a RV with an undefined expected value. The key observation is that the expected value of X is an integral of the identity x and that an integral will be undefined if the positive and negative parts of x are both infinite.

Let Y be a discrete RV that takes values 2^i, where i is a non-zero integer. Its pmf is $p(2^i) = \frac{1}{2}\frac{1}{2^{|i|}} = \frac{1}{2^{|i|+1}}$. That is, Y is obtained by applying one half of X to the positive and one half to the negative integers.

The expectation of Y is

$$E[Y] = E[Y^+] - E[Y^-] = \frac{1}{2}\sum_{i=1}^{\infty} 2^i \frac{1}{2^{|i|}} - \frac{1}{2}\sum_{i=-\infty}^{-1} |2^i| \frac{1}{2^{|i|}} = \infty - \infty$$

and thus undefined.

We complete the definitions of the expected value with a more general definition in terms of the CDF of the RV:

Proposition 7.1.4.

$$\mathrm{E}[X] = \int_0^{\infty} (1 - F(x))dx - \int_{-\infty}^{0} F(x)dx.$$

Proof. We will prove this with the defining property of abstract integration that the integral is the difference between the positive and negative part of the function: $E[X] = E[X^+] - E[X^-]$.

First, we have for the absolute negative part of X and any $\omega \in \Omega$:

$$\begin{aligned} X^-(\omega) &= \int_{-X^-(\omega)}^{0} 1dx \\ &= \int_{-\infty}^{0} I_{X^-(\omega) \geq -x} dx \\ &= \int_{-\infty}^{0} I_{X(\omega) \leq x} dx. \end{aligned}$$

The first line is just the area of a rectangle with sides of length 1 and $-X^-(\omega)$. The second line is the same integral, but this time over the negative reals and an indicator indicating the interval $(-x, 0)$. The final line replaces the non-negative

part X^- with X. We may do this, because we are integrating only over the negative part of the real line.

It follows from the definition of expectation that

$$\begin{aligned} E[X^-] &= \int_\Omega X^-(\omega) dP(\omega) = \int_\Omega \left[\int_{-\infty}^0 I_{X(\omega) \leq x} dx \right] dP(\omega) \\ &= \int_{-\infty}^0 \left[\int_\Omega I_{X(\omega) \leq x} dP(\omega) \right] dx = \int_{-\infty}^0 P(X \leq x) dx = \int_{-\infty}^0 F(x) dx \end{aligned}$$

Similarly, we can show that

$$E[X^+] = \int_{-\infty}^0 P(X > y) dx = \int_{-\infty}^0 (1 - F(x)) dx,$$

which completes the proof. ■

7.2 Properties of expectation

Expected values then have all the properties of abstract integration from the previous chapter. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let X and Y be random variables.

Proposition 7.2.1. *Properties of expected values:*

(a) $E[I_A] = P(A)$.

(b) If $X \geq 0$ *then* $E[X] \geq 0$.

(c) If $X = 0$ *a.s. then* $E[X] = 0$.

(d) For finite $E[X]$ *and* $E[Y]$*:* $E[X + Y] = E[X] + E[Y]$ (additivity).

(e) $E[aX] = aE[X]$ (homogeneity).

(f) $E[aX + bY] = aE[X] + bE[Y]$ (linearity).

(g) If $X \geq 0$ *a.s. and* $E[X] = 0$ *then* $X = 0$ *a.s.*.

Proof. (a-e) are just special cases of Proposition 4.4.1. (f) combines (d) and (e). We state (g) without proof. ■

7.3 Variance and covariance

TODO: Move covariance/correlation (the 2 RV stuff) to Ch8 on multivariate. Maybe rename Ch8 to Random vectors and multivariate distributions.

Definition 7.3.1 (Variance)**.** Let X be a real-valued random variable, such that $E[X]$ is finite. The *variance* of X is defined as

$$Var[X] = \sigma_X^2 \triangleq E[(X - E[X])^2].$$

We refer to σ_X as the *standard deviation* of X.

Proposition 7.3.1. *Let X be a real-valued random variable. $Var[X] = 0$ if and only if X is constant a.s..*

Proof. Let c be a real constant. First, we show that $X = c$ *a.s.* is sufficient. From property (c) from Proposition 7.2.1 and linearity of expectation we have $E[X] = c$. Therefore, $X - E[X] = 0$ *a.s.* and $(X - E[X])^2 = 0$ *a.s.*. This is also a RV, so, using property (c) again, we get $E[(X - E[X])^2] = 0$.

Now we show that it is necessary. $Var[X] = E[(X - E[X])^2] = 0$ and $(X - E[X])^2 \geq 0$, so, by property (f) from Proposition 7.2.1, we have that $(X - E[X])^2 = 0$ a.s.. Therefore, $X = E[X]$ a.s. and X is a constant a.s.. ■

Proposition 7.3.2. *Let X be a real valued random variable. Then,*

$$Var[X] = E[X^2] - E[X]^2.$$

The proof is left as an exercise.

Proposition 7.3.3 (Jensen inequality)**.** *Let X be a random variable and g a convex function. Then,*

$$g(E[X]) \leq E[g(X)].$$

Similarly, if g is a concave function,

$$g(E[X]) \geq E[g(X)].$$

Proof. We will prove the first statement, the proof of the second is similar. Let $bx + a$ be a line tangent to g at the point $E[X]$. That is, $bE[X] + a = g(E[X])$. Because g is convex, it lies above any of it's tangents, so

$$E[g(X)] \geq E[bX + a] = bE[X] + a = g(E[X]).$$

■

Proposition 7.3.4. *Let X be a real-valued random variable. Then,*

$$E[X^2] \geq E[X]^2.$$

Proof. Observe that the square is a convex function. The proposition follows from Jensen's inequality. ■

Definition 7.3.2 (Covariance)**.** Let X and Y be real-valued RVs, such that $E[X]$ and $E[Y]$ are finite. The *covariance* of X and Y is defined as

$$Cov[X, Y] \triangleq E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y].$$

Definition 7.3.3 (Correlation)**.** Let X and Y be real-valued RVs, such that $0 < Var[X] < \infty$ and $0 < Var[Y] < \infty$. The *correlation* of X and Y is defined as

$$\rho[X, Y] \triangleq \frac{Cov[X, Y]}{\sqrt{Var[X]Var[Y]}}.$$

Definition 7.3.4 (Zero correlation)**.** Random variables X and Y are said to be *uncorrelated* if $\rho[X, Y] = 0$.

Equivalent conditions are if $Cov[X, Y] = 0$ or $E[XY] = E[X]E[Y]$.

Proposition 7.3.5. *Let X and Y be a random variables, such that $E[X]$ and $E[Y]$ are finite and X and Y are independent. Then,*

$$E[XY] = E[X]E[Y].$$

Proof. By definition of independence, $P_{X,Y}$ is a product measure of P_X and P_Y. Therefore,

$$\begin{aligned}
E[XY] &= \int_\Omega X(\cdot)Y(\cdot)dP \\
&= \int_{\mathbb{R}^2} xydP_{X,Y} \\
&= \int_{\mathbb{R}^2} xyd(P_X \otimes P_Y) \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} xydP_X \right) dP_Y \qquad \text{(Fubini's Theorem)} \\
&= \left(\int_{\mathbb{R}} ydP_Y \right) \left(\int_{\mathbb{R}} xdP_X \right) \\
&= E[X]E[Y].
\end{aligned}$$

■

Corollary 7.3.1. *If X and Y are independent, then they are uncorrelated.*

The corollary states that independence implies uncorrelatedness. It is important to note that the converse is not true.

Example 7.3.1 (Uncorrelated but dependent RVs)**.** *Let X be a standard normal RV. By definition, $E[X] = 0$ and, using Example 9.2.1, $E[X^3] = 0$.*

Let $Y = X^2$. Clearly, X and Y are dependent - knowing the value of X determines the value of Y. However,

$$Cov[X, Y] = E[XY] - E[X]E[Y] = E[X^3] - E[X]E[X^2] = 0.$$

As this example shows, there exist RVs that are uncorrelated but dependent. Covariance is just a special case of dependence (linear dependence).

For the special case of jointly multivariate normal RVs, uncorrelatedness does imply independence.

TODO: Proposition and proof.

However, the above holds only for jointly multivariate normal, but not for marginally normal that are not jointly multivariate normal.

TODO: Example. Note that this example also solves Exercise 8.5.

Proposition 7.3.6. *Let X and Y be real-valued RVs and $E[X^2] < \infty$, $E[Y^2] < \infty$. Then,*

$$Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y].$$

Proof.

$$\begin{aligned} Var[X+Y] &= E[(X+Y)^2] - (E[X]+E[Y])^2 \\ &= E[X^2+Y^2+2XY] - (E[X]^2+E[Y]^2+2E[X]E[Y]) \\ &= E[X^2] - E[X]^2 + E[Y^2] - E[Y]^2 + 2(E[XY] - E[X]E[Y]) \\ &= Var[X] + Var[Y] + 2Cov[X,Y]. \end{aligned}$$

■

Proposition 7.3.7. *Variance, covariance and correlation have some other useful properties. Assuming $E[X^2] < \infty$, $E[Y^2] < \infty$:*

(a) $Var[aX+b] = a^2 Var[X]$.

(b) $Cov[aX, bY] = abCov[X,Y]$.

(c) $\rho(aX, bY) = \frac{abCov[X,Y]}{|a||b|\sqrt{Var[X]Var[Y]}}$, *additionally assuming* $Var[X] > 0$, $Var[Y] > 0$.

The proof is left as an exercise.

Proposition 7.3.8. *Let X_i, $i = 1, ..., n$ be real valued RVs and $E[X_i^2] < \infty$. Then,*

$$Var[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} a_i^2 Var[X_i] + 2 \sum_{i,j:i<j} a_i a_j Cov[X_i, X_j].$$

Proof. The proof is similar to the proof for the special case of two variables.

$$\begin{aligned} Var[\sum_i a_i X_i] &= E[(\sum_i a_i X_i)^2] - (\sum_i E[a_i X_i])^2 \\ &= \sum_{i,j} E[a_i a_j X_i X_j] - \sum_{i,j} E[a_i X_i] E[a_j X_j] \\ &= \sum_{i,j} E[a_i a_j X_i X_j] - E[a_i X_i] E[a_j X_j] \end{aligned}$$

Now we split the terms into two groups:

$$\begin{aligned} &= \sum_{i,j:i=j} (E[a_i a_j X_i X_j] - E[a_i X_i] E[a_j X_j]) \\ &+ 2 \sum_{i,j:i<j} (E[a_i a_j X_i X_j] - E[a_i X_i] E[a_j X_j]) \\ &= \sum_{i=1}^{n} a_i^2 Var[X_i] + 2 \sum_{i,j:i<j} a_i a_j Cov[X_i, X_j]. \end{aligned}$$

■

Theorem 7.3.1. *Suppose $E[X^k]$ is finite for some k then $E[X^j]$ is finite for $j < k$.*

The proof of this theorem is left as an exercise. In words, the theorem says that if the X^k-th moment exists and is finite, then all lower-order moments also exist and are finite. In particular, if a RV has finite variance it has a finite mean.

Proposition 7.3.9 (Cauchy-Schwarz inequality on expectations)**.** *Let X and Y be random variables with finite variance. Then,*

$$\mathrm{E}[XY] \leq \sqrt{\mathrm{E}[X^2]\,\mathrm{E}[Y^2]}.$$

Proof. This is a direct application of the Cauchy-Schwarz inequality with the (semi) inner product defined with the expectation of the product of random variables. ■

Proposition 7.3.10 (Cauchy-Schwarz inequality on covariance)**.** *Let X and Y be random variables with finite variance. Then,*

$$\mathrm{Cov}[X, Y]^2 \leq \mathrm{Var}[X]\,\mathrm{Var}[Y].$$

Proof. This is a direct application of the Cauchy-Schwarz inequality with the (semi) inner product defined with $Cov[X, Y]$. ■

Corollary 7.3.2. *Correlation between random variables is between -1 and 1.*

7.4 Conditional expectation

Definition 7.4.1. The conditional expectation of a discrete RV X is defined as

$$E[X|Y = y] \triangleq \sum_x x p_{X|Y}(x|y).$$

The conditional expectation of a continuous RV X is defined as

$$E[X|Y = y] \triangleq \int x f_{X|Y}(x|y)dx$$

Note that if the value y is not known, the conditional expectation is a RV (a function of Y) and we write $E[X|Y]$.

Proposition 7.4.1 (Law of iterated expectation)**.** *Let X and Y be RVs. Suppose $E[X]$ is defined, then*

$$E[E[X|Y]] = E[X].$$

Partial proof assuming jointly discrete RVs.

$$\begin{aligned}
E[E[X|Y]] &= \sum_y \left(\sum_x x p_{x|y}(x|y) \right) p(y) = \\
&= \sum_y \left(\sum_x x \frac{p(x,y)}{p(y)} \right) p(y) = \\
&= \sum_y \sum_x x p(x,y) = \\
&= \sum_x \sum_y x p(x,y) = \\
&= \sum_x x \sum_y p(x,y) = \\
&= \sum_x x p(x) = \\
&= E[X]
\end{aligned}$$

■

Proposition 7.4.2 (A more general law of iterated expectation)**.** *Let X be a random variable, such that $E[X]$ is defined. For any RV Y on the same probability space and measurable function g we have*

$$E[E[X|Y]g(Y)] = E[Xg(Y)].$$

Proof. We use the Law of iterated expectation and the fact that constants can be moved outside expectations ($g(Y)$ is known conditional on Y):

$$E[Xg(Y)] = E[E[Xg(Y)|Y]] = E[E[X|Y]g(Y)].$$

■

Corollary 7.4.1.

$$0 = E[E[X|Y]g(Y)] - E[Xg(Y)] = E[(E[X|Y] - X)g(Y)].$$

Note that $g(Y) - E[X|Y]$ is also a function of X, so this Corollary suggests that the conditional expectation of X on Y takes out all 'linear' information from X that is in Y. The remainder is uncorrelated with any function of Y. The conditional expectation $E[X|Y]$ appears to be in some sense an optimal estimator of X.

Proposition 7.4.3. *If $E[X^2] < \infty$ then for any measurable function g,*

$$E[(X - E[X|Y])^2] \leq E[(X - g(Y))^2].$$

Proof.

$$\begin{aligned} E[(X - g(Y))^2] &= E[(X - E[X|Y] + E[X|Y] - g(Y))^2] \\ &= E[(X - E[X|Y])^2] + E[(E[X|Y] - g(Y))^2] \\ &\quad - 2E[(X - E[X|Y])(E[X|Y] - g(Y))] \\ &\geq E[(X - E[X|Y])^2] \end{aligned}$$

The final line is due to $E[(X - E[X|Y])(E[X|Y] - g(Y))] = 0$ (see Corollary 7.4.1) and $E[(X|Y] - g(Y))^2] \geq 0$. ■

That is, the conditional expectation is the optimal estimator with respect to mean squared error.

Exercises

Exercise 7.1. Give an example of two random variables that are uncorrelated but not independent.

Exercise 7.2. Prove Theorem 7.3.1.

Exercise 7.3. Prove Proposition 7.3.2.

Exercise 7.4. Prove Proposition 7.3.4 without using Jensen's inequality.

Exercise 7.5. Prove Proposition 7.3.7.

Exercise 7.6. Prove the law of iterated expectation for jointly continuous RVs.

Chapter 8

Multivariate distributions

8.1 Expectation, variance, and covariance

Definition 8.1.1. A *multivariate (k-variate) RV* or random vector X is a column vector $X = [X_1, X_2, ..., X_k]^T$, $k \geq 1$ whose components are random variables on the same probability space.

Definition 8.1.2. Let X and Y be k-variate and r-variate random variables, respectively.

The expectation of X is defined as

$$\mathrm{E}[X] = \mu \triangleq (\mathrm{E}[X_1], \mathrm{E}[X_2], ..., \mathrm{E}[X_k])^T.$$

$\mathrm{Cov}[X, Y]$, the cross-covariance matrix of X and Y is a $k \times r$ matrix with components

$$\mathrm{Cov}[X, Y]_{i,j} \triangleq \mathrm{Cov}[X_i, Y_j].$$

In matrix form,

$$\mathrm{Cov}[X, Y] = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])^T] = \mathrm{E}[XY^T] - E[X]E[Y]^T.$$

Note that the expected value of the matrix valued $E[XY^T]$ is defined analogously to the expectation of a random vector - with expectations of individual components.

Similarly, $\rho[X, Y]$, the cross-correlation matrix of X and Y, is a $k \times r$ matrix with components

$$\rho[X, Y]_{i,j} = \rho[X_i, Y_j].$$

When $k = r$, the cross-covariance and cross-correlation matrices are squared. If $X = Y$, we refer to $\text{Cov}[X, X] = \text{Cov}[X] = \text{Var}[X]$ as the covariance matrix and to $\rho[X, X] = \rho[X]$ as the correlation matrix. The covariance and correlation matrices are squared and symmetric.

Proposition 8.1.1. *The properties of univariate expectation and variance transfer to the multivariate random variables:*

$$\begin{aligned}
\text{E}[BX + a] &= B\,\text{E}[X] + a \\
\text{Cov}[BX + a] &= B\,\text{Cov}[X]B^T \\
\text{Var}[X + Y] &= \text{Var}[X] + \text{Var}[Y] + \text{Cov}[X, Y] + \text{Cov}[Y, X] \\
\rho[X] &= \left(\sqrt{diag(\text{Var}[X])}\right)^{-1} \text{Cov}[X] \left(\sqrt{diag(\text{Var}[X])}\right)^{-1}
\end{aligned}$$

For any $m \times k$ matrix B and $m-$vector a. Note that $diag(A)$ is a matrix with diagonal elements the same as the square matrixA and 0 everywhere else. So, $\sqrt{diag(\text{Var}[X])}$ is a matrix with standard deviations on the diagonal and zeroes everywhere else.

The proof is left as an exercise.

8.2 The multinomial distribution

The multinomial distribution is the generalization of the binomial distribution to more than one outcome.

Definition 8.2.1. The probability mass function of a k-variate multinomial distribution ($k \geq 1$) is

$$p(x_1, x_2, ..., x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

where the two parameters are a positive integer n (number of trials) and $p_i \geq 0, \sum_{i=1}^{k} p_i = 1$ (outcome probabilities).

Note that for $k = 2$ we get the binomial distribution.

The expectation, variance, and covariance of the multinomial distribution are as follows:

Proposition 8.2.1.

$$\begin{aligned}\mathrm{E}[X_i] &= np_i \\ \mathrm{Var}[X_i] &= np_i(1-p_i) \\ \mathrm{Cov}[X_i, X_j] &= -np_ip_j, i \neq j\end{aligned}$$

The proof of these properties is left as an exercise.

8.3 Transformations

Proposition 8.3.1. *Let $(X_1, X_2, ..., X_n)$ be RVs with joint pdf $f(x_1, x_2, ..., x_n)$. Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be an invertible and continuously differentiable function given by $Y_i = g_i(X_1, X_2, ..., X_n)$. Let $g^{-1}(y) = (h_1(y), h_2(y), ..., h_n(y))$ be the inverse of g. Then, the joint density of Y_i is*

$$f_{Y_1,Y_2,...,Y_n}(y_1, y_2, ..., y_n) = f_{X_1,X_2,...,X_n}(h_1(y), h_2(y), ..., h_n(y))|\det J_h|,$$

where J_h is the Jacobian matrix

$$J_h = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_1}{\partial y_n} & \frac{\partial x_2}{\partial y_n} & \cdots & \frac{\partial x_n}{\partial y_n} \end{bmatrix}$$

and $\frac{\partial x_i}{\partial y_j} = \frac{\partial h_i(y_1,y_2,...,y_n)}{\partial y_j}$.

It turns out that $\det J_h = 1/\det J_g$, where J_g has elements of the form $\frac{\partial y_i}{\partial x_j} = \frac{\partial g_i(x_1,x_2,...,x_n)}{\partial x_j}$.

We omit the proof of this proposition.

8.4 The multivariate normal distribution

Definition 8.4.1. The probability density function of a k-variate normal distribution ($k \geq 1$) is

$$f(x) = \frac{1}{(2\pi)^{\frac{k}{2}}(\det \Sigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right), x \in \mathbb{R}^k$$

where the two parameters are $\mu \in \mathbb{R}^k$ (mean vector) and $k \times k$ positive definite matrix Σ (covariance matrix).

Proposition 8.4.1. *Let $Z = [Z_1, Z_2, ..., Z_k]^T$ be a random vector, such that Z_i are independent univariate standard normal variables $Z_i \sim N(0,1)$. Let $\mu \in \mathbb{R}^k$ and A a $k \times k$ non-singular matrix. Then,*

$$AZ + \mu \sim N(\mu, \Sigma), \text{ with } \Sigma = AA^T.$$

Proof. The Z_i are independent, so their joint PDF is

$$f(z_1, ..., z_k) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi}} \exp(-\frac{z_i^2}{2}) = (2\pi)^{-\frac{k}{2}} \exp(-\frac{1}{2} z^T z).$$

We have an affine transformation $g(Z) = AZ + \mu$ with inverse $Z = g^{-1}(X) = A^{-1}(X - \mu)$. We can apply Proposition 8.3.1 with $\det J_h = \det(A^{-1}) = (\det A)^{-1}$, so

$$\begin{aligned} f(x) &= |(\det A)^{-1}| f(A^{-1}(x - \mu)) \\ &= (2\pi)^{-\frac{k}{2}} |(\det A)^{-1}| \exp\left(-\frac{1}{2}(x-\mu)^T (AA^T)^{-1}(x-\mu)\right) \\ &= (2\pi)^{-\frac{k}{2}} |(\det A)^{-1}| \exp\left(-\frac{1}{2}(x-\mu)^T (\Sigma)^{-1}(x-\mu)\right), x \in \mathbb{R}^k. \end{aligned}$$

■

Corollary 8.4.1. *Every $k-$variate normal distribution is a transformation $AZ + \mu$ of a k-variate random vector Z with standard normal components.*

Note: By definition, any positive definite matrix Σ can be written as a product $\Sigma = AA^T$, where A is non-singular.

Proposition 8.4.2. *Let X be a $k-$variate normal distribution $X \sim N(\mu, \Sigma)$. Let $b \in \mathbb{R}^k$ and B be a non-singular $k \times k$ matrix. Then,*

$$BX + b \sim N(B\mu + b, B\Sigma B^T).$$

Proof. By Corollary 8.4.1 we can write $X = AZ + \mu$ with $\Sigma = AA^T$. Then $BX + b = B(AZ + \mu) + b = BAZ + B\mu + b$, where B is non-singular and $BAA^T B^T = B\Sigma B^T$. ■

Proposition 8.4.3. *If $X \sim N(\mu, \Sigma)$ then $E[X] = \mu$ and $\mathrm{Cov}[X, X] = \Sigma$.*

Proof. By the properties of multivariate expectation and covariance and Corollary 8.4.1 we can write $X = AZ + \mu$ with $\Sigma = AA^T$. Then $E[X] = E[AZ + \mu] = AE[Z] + \mu = \mu$. $\text{Cov}[X] = \text{Cov}[AZ + \mu] = A\,\text{Cov}[Z]A^T = AA^T = \Sigma$. ■

Proposition 8.4.4 (Marginal and conditional distribution)**.** *Let X be a k-variate normal distribution, $k > 1$. If we partition the components of X into two random vectors X_A and X_B, each with at least one component, we can write*

$$X = \begin{bmatrix} X_A \\ X_B \end{bmatrix} \sim N\left(\mu = \begin{bmatrix} \text{E}[X_A] \\ \text{E}[X_B] \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{AA}, \Sigma_{AB} \\ \Sigma_{BA}, \Sigma_{BB} \end{bmatrix}\right),$$

where $\Sigma_{WZ} = \text{Cov}(X_W, X_Z)$. For any such partition, we have

(a) $X_A \sim N(\text{E}[X_A], \Sigma_{AA})$.

(b) $X_A | X_B = x_b \sim N(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \text{E}[X_B]), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})$.

The proof is left as an exercise.

Note that the univariate marginals of a multivariate normal are univariate normal. However, the converse is not always true. There exist multivariate distributions with univariate normal marginals that are not multivariate normal. Finding a counterexample is left as an exercise.

Another important property of the MVN is that because relationship between univariate normals is only linear, we have that uncorrelatedness implies independence. This is also left as an exercise.

Exercises

Exercise 8.1. Prove all the statements in Proposition 8.2.1.

Exercise 8.2. Prove all the statements in Proposition 8.1.1.

Exercise 8.3. Prove Corollary 8.4.1.

Exercise 8.4. Prove Proposition 8.4.4.

Exercise 8.5. Find a bivariate distribution where the two marginal distributions are univariate normal but the distribution is not bivariate normal.

Exercise 8.6. Show that unit diagonal correlation (uncorrelatedness) of a multivariate normal distribution implies independence of individual random variables (that is, that the joint density factors into individual marginals).

Chapter 9

Alternative representations of distributions

In this chapter we are going to present probability generating functions and moment generating functions. These are alternative representations of distributions (PMFs and PDFs), which are sometimes more convenient for deriving the expected value, the variance, and other properties of distributions.

9.1 Probability generating functions

Definition 9.1.1. The probability generating function (PGF) of an non-negative integer-valued random variable X is defined as

$$\alpha_X(t) \triangleq E[t^X] = \sum_{i=0}^{\infty} t^i p_X(i).$$

Note that PGFs are defined for non-negative random values. However, in this text we restrict ourselves to integer-valued (discrete) random variables.

Proposition 9.1.1.

(a) $\alpha_X(1) = 1$.

(b) If X and Y are independent random variables: $\alpha_{X+Y}(t) = \alpha_X(t)\alpha_Y(t)$.

(c) If random variables X and Y have identical probability generating functions, then they have the same distribution.

(d) If $\alpha_X(t)$ has a radius of convergence $\rho > 1$ then

$$E[X] = \frac{d}{dt}\alpha_X(t)|_{t=1}$$

and

$$Var[X] = \frac{d^2}{dt^2}\alpha_X(t)|_{t=1} + \frac{d}{dt}\alpha_X(t)|_{t=1} - (\frac{d}{dt}\alpha_X(t))^2|_{t=1}.$$

(e) $p_X(i) = \left(\frac{1}{i!}\right)\frac{d^i}{dt^i}\alpha_X(t)|_{t=0}$.

Proof. (a) Left as an exercise.

(b) Left as an exercise.

(c) Follows from uniqueness of power series representations of functions.

(d) The radius of convergence justifies the differentiation and evaluation at $t = 1$:

$$\frac{d}{dt}\alpha_X(t) = \sum_{i=0}^{\infty} i p_X(i) t^{i-1}$$

and

$$\frac{d}{dt}\alpha_X(t)|_{t=1} = \sum_{i=0}^{\infty} i p_X(i) = E[X].$$

Taking the second derivative and evaluating at $t = 1$:

$$\frac{d^2}{dt^2}\alpha_X(t) = \sum_{i=2}^{\infty} i(i-1) p_X(i) t^{i-2}$$

and

$$\frac{d^2}{dt^2}\alpha_X(t)|_{t=1} = E[X(X-1)] = E[X^2] - E[X].$$

Finally,

$$Var[X] = E[X^2] - E[X]^2 = \frac{d^2}{dt^2}\alpha_X(t)|_{t=1} + \frac{d}{dt}\alpha_X(t)|_{t=1} - (\frac{d}{dt}\alpha_X(t))^2|_{t=1}.$$

(e) Left as an exercise. ■

Example 9.1.1. *Expectation of a Geometric RV*

Let the PMF of X be $p(i) = \theta(1-\theta)^i$, for $i \geq 0$, and 0 otherwise. That is, X has a Geometric distribution.

By definition, the PGF of X is

$$\alpha(t) = \sum_{i=0}^{\infty} t^i \theta(1-\theta)^i = \theta \sum_{i=0}^{\infty} (t(1-\theta))^i = \theta \frac{1}{1-t(1-\theta)}.$$

The final step is based on the fact that the series is a Geometric series. The step is justified when the series converges: $|t(1-\theta)| < 1$. By rearranging and taking into account that for a Geometric distribution we have $0 < \theta < 1$, we get $|t| \leq \frac{1}{1-\theta}$.

The convergence radius is therefore greater than 1 for any $0 < \theta < 1$ between 0 and 1, so we can use

$$E[X] = \frac{d}{dt}\alpha(t)|_{t=1}$$

.

Taking the derivative, we get

$$= \frac{d}{dt}\frac{\theta}{1-t(1-\theta)}|_{t=1} = \frac{\theta(1-\theta)}{(1-t(1-\theta))^2}|_{t=1} = \frac{\theta(1-\theta)}{\theta^2} = \frac{1-\theta}{\theta}.$$

Example 9.1.2. *Sum of two independent Poisson RVs is a Poisson RV*

Let $X \sim Poisson(\lambda_X)$ and $Y \sim Poisson(\lambda_Y)$, $\lambda_X, \lambda_Y > 0$. What is the distribution of $X + Y$?

Recall that the PMF of a Poisson is $p(i) = \frac{\lambda^i}{i!}e^{-\lambda}$. Its PGF is then

$$\alpha(t) = \sum_{i=0}^{\infty} t^i p(i) = \sum_{i=0}^{\infty} t^i \frac{\lambda^i}{i!} e^{-\lambda} = e^{-\lambda} \sum_{i=0}^{\infty} \frac{(t\lambda)^i}{i!} = e^{\lambda(t-1)},$$

where the final step is based on recognizing the Taylor series expansion of $e^{t\lambda}$ at 0.

Using the property of PGFs that the sum of two independent RVs' PGF is the product of their PGFs, we have

$$\alpha_{X+Y}(t) = e^{\lambda_X(t-1)}e^{\lambda_Y(t-1)} = e^{(\lambda_X+\lambda_Y)(t-1)}.$$

Because the PGF uniquely determines the distribution, the sum of independent Poisson RVs is also Poisson with rate $\lambda_X + \lambda_Y$.

Example 9.1.3. *Expected value and variance of a Poisson RV*

Let $X \sim Poisson(X)$ We have shown in Example 9.1.2 that the PGF of a Poisson RV with rate λ is $\alpha(t) = e^{\lambda(t-1)}$. So

$$\mathrm{E}[X] = \frac{d}{dt}e^{\lambda(t-1)}|_{t=1} = \lambda e^{\lambda(t-1)}|_{t=1} = \lambda.$$

Differentiating one more time, we have

$$\frac{d^2}{dt^2}e^{\lambda(t-1)}|_{t=1} = \lambda^2 e^{\lambda(t-1)}|_{t=1} = \lambda^2$$

and

$$\mathrm{Var}[X] = \frac{d^2}{dt^2}\alpha(t)|_{t=1} + \frac{d}{dt}\alpha(t)|_{t=1} - (\frac{d}{dt}\alpha(t))^2|_{t=1} = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

9.2 Moment generating functions

Definition 9.2.1. The *moment generating function* (MGF) of a random variable X is defined as

$$M_X(t) \triangleq E[e^{tX}].$$

In the special case of a discrete random variable this reduces to

$$M_X(t) = \sum_x e^{tx} p_X(x)$$

and for a continuous random variable

$$M_X(t) = \int_x e^{tx} f_X(x)dx.$$

Although t can be complex, we restrict ourselves to $t \in \mathbb{R}$.

Note that there is a relationship between PGFs and MGFs. If a non-negative RV has a MGF and a PGF, we have $M(t) = \mathrm{E}[e^{tX}] = \mathrm{E}[(e^t)^X] = \alpha(e^t)$.

The following theorem, which we state without proof, comes from the properties of analytic functions:

Theorem 9.2.1. *Let X and Y be random variables, such that $M_X(t) = M_Y(t)$, $\forall t \in [-\epsilon, \epsilon]$ for some $\epsilon > 0$. Then, X and Y have the same CDF.*

Proposition 9.2.1.

(a) $M_X(0) = 1$.

(b) If X and Y are independent RVs: $M_{X+Y}(t) = M_X(t)M_Y(t)$.

(c) $M_{aX+b}(t) = e^{tb}M_X(at)$.

(d) Let $M_X(t)$ be finite for $t \in [-\epsilon, \epsilon]$ for some $\epsilon > 0$ then $\frac{d^k}{dt^k}M_X(t)|_{t=0} = E[X^k]$, for $k \geq 1$. (moment generating property).

Proof. Properties (a-c) are left as an exercise. The proof of property (d) is more involved. We have $\frac{d^k}{dt^k}M_X(t) = \frac{d^k}{dt^k}E[e^{tX}] = E[\frac{d^k}{dt^k}e^{tX}] = E[X^k e^{tX}]$. Evaluating at $t = 0$ we get $E[X^k]$. Note that the exchange of derivatives and expectation in this proof is not trivial - we have to invoke the dominated convergence theorem. It suffices to show $\mathrm{E}[\lim_{h\downarrow 0}\frac{e^{hX}-1}{h}] = \lim_{h\downarrow 0}\mathrm{E}[\frac{e^{hX}-1}{h}]$ (recall the definition of the derivative - this is its value at 0 for this function e^{hX}). We show it by first showing that $\frac{e^{hX}-1}{h} \leq Xe^{hX}$ and that Xe^{hX} converges. That is, $E[Xe^{hX}] \leq \infty$. ■

Example 9.2.1 (MGF of the standard normal RV)**.** *First, let's derive the MGF of a standard normal RV. Using the definition of a MGF and plugging in the PDF of the standard normal, we get*

$$\begin{aligned} M(t) = \mathrm{E}[e^{tX}] &= \int_{-\infty}^{\infty} e^{tx}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2-2tx)}dx \\ &= e^{\frac{t^2}{2}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-t)^2}dx \\ &= e^{\frac{t^2}{2}}. \end{aligned}$$

The key steps in the above derivation are completing the square in the exponent and then recognizing that we have an integral over a PDF of a normal distribution, which has to integrate to 1.

Next, we have

$$E[X] = \frac{d}{dt} e^{\frac{t^2}{2}}|_{t=0} = te^{\frac{t^2}{2}}|_{t=0} = 0,$$

$$E[X^2] = \frac{d^2}{dt^2} e^{\frac{t^2}{2}}|_{t=0} = e^{\frac{t^2}{2}} + t^2 e^{\frac{t^2}{2}}|_{t=0} = 1.$$

$$E[X^3] = \frac{d^3}{dt^3} e^{\frac{t^2}{2}}|_{t=0} = \frac{d}{dt}(1+t^2)e^{\frac{t^2}{2}}|_{t=0} = 2te^{\frac{t^2}{2}} + (1+t^2)te^{\frac{t^2}{2}}|_{t=0} = (t^3+3t)e^{\frac{t^2}{2}} = 0.$$

Example 9.2.2. *Distribution of the sum of two independent normal RVs*

From Example 9.2.1 we know that the MGF of the standard normal $Z \sim N(0,1)$ is $e^{\frac{t^2}{2}}$. Next, recall that every normal distribution can be obtained as an affine transformation $aZ + b$ of the standard normal.

Next, using the property of MGFs that $M_{aX+b}(t) = e^{tb} M_X(at)$, we can derive the MGF of $X \sim N(\mu, \sigma^2)$:

$$M_{\sigma Z+\mu}(t) = e^{t\mu} e^{\sigma^2 \frac{t^2}{2}}.$$

Let $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ be independent RVs. The MGF of their sum is

$$M_{X+Y}(t) = M_X(t)M_Y(t) = e^{t\mu_X} e^{\frac{\sigma_X^2 t^2}{2}} e^{t\mu_Y} e^{\frac{\sigma_Y^2 t^2}{2}} = e^{t(\mu_X+\mu_Y)} e^{(\sigma_X^2+\sigma_Y^2)\frac{t^2}{2}}.$$

This implies that the sum of two independent normal RVs is a normal RV with mean $\mu_X + \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$.

Exercises

Exercise 9.1. Prove statements (a), (b) and (e) in Proposition 9.1.1.

Exercise 9.2. Prove statements (a-c) in Proposition 9.2.1.

Chapter 10

Concentration inequalities

10.1 Markov inequality

Proposition 10.1.1 (Markov inequality)**.** *Let X be a non-negative random variable and let $E[X]$ exist. Then, for any $a > 0$,*

$$P(X \geq a) \leq \frac{E[X]}{a}.$$

Proof.

$$\begin{aligned}
E[X] &= E[XI_{\{X<a\}} + XI_{\{X\geq a\}}] \\
&= E[XI_{\{X<a\}}] + E[XI_{\{X\geq a\}}] && \text{(linearity of expectation)} \\
&\geq E[XI_{\{X\geq a\}}] && (X \text{ is non-negative}) \\
&\geq E[aI_{\{X\geq a\}}] \\
&= aE[I_{\{X\geq a\}}] \\
&= aP(X \geq a)
\end{aligned}$$

■

10.2 Chebyshev inequality

Proposition 10.2.1 (Chebyshev inequality)**.** *Let X be random variable with expectation μ and variance $\sigma^2 < \infty$. Then, for any $a > 0$,*

$$P(|X - \mu| \geq a\sigma) \leq \frac{1}{a^2}.$$

Or, equivalently, by setting $b = a\sigma$,

$$P(|X-\mu| \geq b) \leq \frac{\sigma^2}{b^2}.$$

Proof. We will use the Markov inequality on the non-negative RV $|X-\mu|^2$:

$$\begin{aligned} P(|X-\mu| \geq a\sigma) = P(|X-\mu|^2 \geq (a\sigma)^2) &\leq \frac{E[|X-\mu|^2]}{(a\sigma)^2} \\ &= \frac{\sigma^2}{(a\sigma)^2} \\ &= \frac{1}{a^2} \end{aligned}$$

■

10.3 Chernoff bound

Proposition 10.3.1 (Generic Chernoff bound)**.** *Let X be a random variable. Then, for any a and every $t > 0$,*

$$P(X \geq a) \leq \frac{E[e^{tX}]}{e^{ta}}.$$

Proof. We prove this by applying the Markov inequality to $P(e^{tX} \geq e^{ta})$. ■

Proposition 10.3.2 (Chernoff bound for Bernoulli variables)**.** *Let $X_1, ..., X_n \sim_{iid} Bernoulli(p)$, with $p > \frac{1}{2}$. Let $S_n = \sum_{i=1}^n X_i$. Then, for every $\delta > 0$,*

$$P(S_n \geq (1+\delta)np) \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{np}.$$

Proof. For a Bernoulli variable we have $E[e^{tX_i}] = pe^t + (1-p) = 1 + p(e^t - 1) \leq e^{p(e^t-1)}$. So, from independence of X_i, we have $E[e^{tS_n}] \leq e^{np(e^t-1)}$.

For any $\delta > 0$ and taking $t = \ln(1+\delta) > 0$ and $a = (1+\delta)np$, we have $E[e^{tS_n}] \leq e^{\delta np}$ and $e^{-ta} = (1+\delta)^{-(1+\delta)np}$. Substituting into the generic Chernoff bound, we have

$$P(S_n - np \geq \delta np) = P(S_n \geq (1+\delta)np) \leq \frac{e^{\delta np}}{(1+\delta)^{(1+\delta)np}} = \left(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right)^{np}.$$

■

Example 10.3.1. *Application of inequalities in the Bernoulli case*

Suppose a system has n independent components, each with a $p = \frac{1}{2}$ probability of failing. The system fails if more than $\frac{3}{4}$ components fail. Let's use the concentration inequalities to put an upper bound on the probability that the system fails.

Let X be the RV that represents the number of components that fail. Because it is a sum of iid Bernoulli, it has a binomial distribution $X \sim Binomial(n, \frac{1}{2})$.

First, let's use the Markov inequality $P(X \geq a) \leq \frac{E[X]}{a}$. In our case we have $a = \frac{3}{4}n$ and $\mathrm{E}[X] = \frac{n}{2}$, which gives us

$$P(X \geq \frac{3}{4}n) \leq \frac{4n}{6n} = \frac{2}{3}.$$

Before using the Markov inequality we should of course always check if the conditions are met - a should be positive and X should be a non-negative RVs, which are both true in our case.

The Markov inequality does not provide us with a very tight bound and what is even worse, it is constant for all n. That is, it does not get tighter with increasing n, although we know that the probability of system failure decreases with increasing n.

Next, let's apply the Chebyshev inequality, which does improve with increasing n. We start with the two sided $P(|X - \mu| \geq b) \leq \frac{\sigma^2}{b^2}$, which implies the one sided bound $P(X - \mu \geq b) \leq \frac{\sigma^2}{b^2}$ (if the bound is true for the absolute of the value, it is also true for the value, which is the same or less). In our case we have $\mu = E[X] = \frac{n}{2}$, $b = \frac{1}{4}n$, and $\sigma^2 = n\frac{1}{2}(1 - \frac{1}{2}) = \frac{1}{2}n$. Putting it all together, we get

$$P(X \geq \frac{3}{4}n) = P(X - \frac{1}{2}n \geq \frac{1}{4}n) \leq \frac{\frac{1}{4}n}{(\frac{1}{4}n)^2} = \frac{4}{n}.$$

Finally, let's apply the Chernoff bound for Bernoulli variables. We have $P(X \geq \frac{3}{4}n) = P(X \geq (1 + \frac{1}{2})\frac{1}{2}n)$, which identifies $\delta = \frac{1}{2} > 0$, so

$$P(X \geq \frac{3}{4}n) \leq \left(\frac{e^{\frac{1}{2}}}{(1 + \frac{1}{2})^{(1+\frac{1}{2})}}\right)^{\frac{1}{2}n} = \left(\frac{e^{\frac{1}{2}}}{\frac{3}{2}^{\frac{3}{2}}}\right)^{\frac{n}{2}} = \left(\frac{e}{\frac{27}{8}}\right)^{\frac{n}{4}}.$$

As we can see, each subsequent inequality achieves an asymptotically tighter bound by incorporating more information about the RV (variance for Chebyshev and moment generating function for Chernoff).

10.4 Hoeffding inequality

Proposition 10.4.1 (Hoeffding inequality)**.** *Let $X_1, X_2, ..., X_n$ be independent random variables, bounded by $a_i \leq X_i \leq b_i$. Let $S = \sum_{i=1}^n X_i$. Then, for every $t > 0$,*

(a) $P(S_n - E[S_n] \geq t) \leq e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$,

(b) $P(S_n - E[S_n] \leq -t) \leq e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$,

(c) $P(|S_n - E[S_n]| \geq t) \leq 2e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$.

Proof. By the generic Chernoff bound we have $P(X \geq t) \leq e^{-st} E[e^{sX}]$ for every $s > 0$. If we apply this to $S_n - E[S_n]$, we get

$$\begin{aligned} P(S_n - E[S_n] \geq t) &\leq e^{-st} E[e^{s(S_n - E[S_n])}] \\ &= e^{-st} E[e^{s(\sum_{i=1}^n X_i - E[S_n])}] \\ &= e^{-st} E[\prod_{i=1}^n e^{s(X_i - E[S_n])}] \\ &= e^{-st} \prod_{i=1}^n E[e^{s(X_i - E[S_n])}]. \quad \text{(from independence of } X_i) \end{aligned}$$

What remains is to find a good bound for $E[e^{s(X_i - E[S_n])}]$. To do this, we will rely on an intermediate result, which we state without proof:

Lemma 10.4.1. *Let X be a random variable, such that $E[X] = 0$ and $a \leq X \leq b$. Then, for all $s > 0$*

$$E[e^{sX}] \leq e^{\frac{s^2 (b-a)^2}{8}}.$$

$$\begin{aligned} P(S_n - E[S_n] \geq t) &\leq e^{-st} \prod_{i=1}^n E[e^{s(X_i - E[S_n])}] \\ &\leq e^{-st} \prod_{i=1}^n e^{\frac{s^2 (b_i - a_i)^2}{8}} \qquad \text{(by the above lemma)} \\ &= e^{-st} e^{s^2 \sum_{i=1}^n \frac{(b_i - a_i)^2}{8}} \\ &= e^{\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}} \qquad \text{(let } s = \frac{4t}{\sum_{i=1}^n (b_i - a_i)^2}) \end{aligned}$$

This concludes the proof of (a). Claim (b) can be proven by applying (a) to $-X_i$ and statement (c) can be proven by combining (a) and (b).

■

Exercises

Exercise 10.1. Show that the Chernoff bound for Bernoulli variables is a special case of the Hoeffding inequality.

Chapter 11

Convergence of random variables

11.1 Types of convergence

We start with the type of convergence that might already be familiar to us from sequences and limits of real numbers:

Definition 11.1.1 (Convergence of a sequence of real numbers)**.** A sequence of real numbers $\{x_n\}$ is said to converge to some $x \in \mathbb{R}$ if for any $\epsilon > 0$ there exists an n_0, such that

$$|x_n - x| < \epsilon, \forall n \geq n_0.$$

We write $\lim_{n\to\infty} x_n = x$ or just $x_n \longrightarrow x$.

We cannot apply point-wise convergence directly to RVs, because RVs are not real numbers but functions from Ω to $\mathbb{R}$. However, if we limit ourselves to some $\omega \in \Omega$, we get a sequence of values $X_n(\omega)$ and a value $X(\omega)$. If the RVs converge in the sequence sense above at every ω, we have point-wise convergence:

Definition 11.1.2 (Point-wise convergence)**.** A sequence of random variables X_i is said to converge point-wise to X if $X_n(\omega) \longrightarrow X(\omega)$, $\forall \omega \in \Omega$.

We write $X_n \xrightarrow{p.w.} X$.

Some parts of Ω might have, even when combined, zero probability. Those ω do not contribute to the behaviour of our RVs. In this sense point-wise convergence is unnecessarily restrictive. A much more commonly used type is almost sure convergence, which discards ω with zero probability.

Definition 11.1.3 (Almost sure convergence)**.** Let X_i be a sequence of random variables and X a RV, all of which are defined on the same probability space $(\Omega, \mathcal{F}, P)$. X_i is said to converge almost surely to X if

$$P(\{\omega : X_n(\omega) \longrightarrow X(\omega)\}) = 1.$$

We write $X_n \xrightarrow{a.s.} X$.

That is, we have almost sure convergence, if we do not have point-wise convergence at most on a set of ω with probability 0.

Another type of convergence that is less strict than almost sure convergence but in practice usually enough that things behave nicely as our sample size increases is convergence in probability.

Definition 11.1.4 (Convergence in probability)**.** Let X_i be a sequence of random variables and X a RV, all of which are defined on the same probability space $(\Omega, \mathcal{F}, P)$. X_i is said to converge in probability to X if

$$\lim_{n\to\infty} P(\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}) = 0, \forall \epsilon > 0.$$

We write $X_n \xrightarrow{P} X$.

See Example 11.2.1 for the proof and an explanation of the difference between the two types of convergence. Note that in probability and statistics textbooks the above definition of convergence in probability usually appears in a simplified but equivalent form:

Definition 11.1.5 (Convergence in probability)**.** Let X_i be a sequence of random variables and X a RV, all of which are defined on the same probability space $(\Omega, \mathcal{F}, P)$. X_i is said to converge in probability to X if

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0, \forall \epsilon > 0.$$

Definition 11.1.6 (Convergence in r-th mean)**.** Let X_i be a sequence of random variables and X a RV, all of which are defined on the same probability space $(\Omega, \mathcal{F}, P)$. X_i is said to converge in $r-$th mean ($r \geq 1$) to X if

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

We write $X_n \xrightarrow{r} X$ and in the special case of $r = 2$, we write $X_n \xrightarrow{q.m.} X$ (in quadratic mean).

Convergence in r-th mean is often easier to prove than convergence in probability. This is particularly useful combined with the fact that it implies convergence in probability (see Example 11.2.3).

Definition 11.1.7 (Convergence in distribution)**.** A sequence of random variables X_i is said to converge in distribution to X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \forall x \in R, \text{ where } F_X(\cdot) \text{ is continuous.}$$

We write $X_n \xrightarrow{D} X$.

Convergence in distribution is the weakest type of convergence of RVs in the sense that it is implied by all other types of convergence that we have introduced in this chapter. Unlike the other types of convergence it does not focus on the behaviour of the RVs at particular ω. Instead, it is interested only in the probability law or distribution of the RVs, represented here with the CDF. As such, it does not require the RVs to be defined on the same probability space. See Example 11.2.3 for an illustration of what it means for RVs to be similar in behaviour or only similar in distribution.

11.2 Relationships between types of convergence

Theorem 11.2.1. $X_n \xrightarrow{p.w.} X \implies X_n \xrightarrow{a.s.} X$.

Proof. This is immediately clear from the definitions. Point-wise convergence implies that convergence holds not only for a set of measure 1 but for all ω. ■

Theorem 11.2.2. $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{P} X$.

The proof of this is beyond the scope of this text.

Example 11.2.1 (Convergence in probability does not imply almost sure convergence)**.** *Consider the following sequence of random variables:*

$$X_n = \begin{cases} 1 & \textit{with probability } \frac{1}{n}, \\ 0 & \textit{with probability } 1 - \frac{1}{n}. \end{cases}$$

We have $\lim_{n\to\infty} P(|X_n| > \epsilon) = \lim_{n\to\infty} \frac{1}{n} = 0$*, so* $X_n \xrightarrow{P} 0$*. On the other hand, let* A_n *be the event that* $X_n = 1$*. These events are independent and* $\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$*. The Second Borel-Cantelli lemma (see Chapter 12.1) states that infinitely many* A_n *will occur, so* X_n *does not converge to 0 almost surely.*

This example illustrates the difference between convergence in probability and in the almost sure sense. Convergence in probability states that the probability of the unwanted deviation goes to 0 as n approaches infinity, but still allows for the deviation to happen an infinite number of times, albeit at less and less frequent

intervals. Almost sure convergence is much more strict, as the probability of deviation not only goes to 0, but does it in a way that at some point there is not enough probability left combined over all X_i beyond some n for the deviation to occur infinitely many times. That is, almost sure convergence is a statement about the probability of the entire tail from n onwards. As loose analogy would be that of the convergence of a sequence - there exist sequences that converge to 0, but the series sum can still diverge, because the convergence is not "fast" enough.

Theorem 11.2.3. $X_n \xrightarrow{r} X \implies X_n \xrightarrow{P} X$.

Proof. We will use the Markov inequality. For $Y \geq 0$ and $\epsilon > 0$ we have

$$P(Y > \epsilon) = P(Y^r > \epsilon^r) \leq \frac{E[Y^r]}{\epsilon^r}.$$

By applying this to $Y = |X_n - X|$, we get

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

Since $\lim_{n\to\infty} E[|X_n - X|^r] = 0$, $\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$ ■

Example 11.2.2 (Convergence in probability does not imply convergence in r-th mean)**.** *Consider the following sequence of random variables:*

$$X_n = \begin{cases} n^3 & \text{with probability } \frac{1}{n^2}, \\ 0 & \text{with probability } 1 - \frac{1}{n^2}. \end{cases}$$

We have $\lim_{n\to\infty} P(|X_n| > \epsilon) = \lim_{n\to\infty} \frac{1}{n^2} = 0$, *so* $X_n \xrightarrow{P} 0$. *However,* $E[X_n] = n$ *diverges.*

Theorem 11.2.4. $X_n \xrightarrow{r} X \implies X_n \xrightarrow{s} X$, *for* $r > s \geq 1$.

Proof. We can prove this using Lyapunov's inequality. This inequality states that for a random variable X and numbers $0 < s < r < \infty$ we have $E[|X|^r]^{\frac{1}{r}} \geq E[|X|^s]^{\frac{1}{s}}$. The result follows immediately. ■

Theorem 11.2.5. $X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X$.

Proof. For any $\epsilon > 0$ we have

$$\begin{aligned} F_{X_n}(x) &= P(X_n \leq x) \\ &= P(X_n \leq x \cap X \leq x + \epsilon) + P(X_n \leq x \cap X > x + \epsilon) \\ &\leq F_X(x + \epsilon) + P(|X_n - X| > \epsilon). \end{aligned}$$

Similarly, we have

$$F_X(x - \epsilon) \leq F_{X_n}(x) + P(|X_n - X| > \epsilon).$$

So,

$$F_X(x - \epsilon) - P(|X_n - X| > \epsilon) \leq F_{X_n}(x) \leq F_X(x + \epsilon) + P(|X_n - X| > \epsilon).$$

Since $X_n \xrightarrow{P} X$, $P(|X_n - X| > \epsilon) \to 0$ and we have

$$F_X(x - \epsilon) \leq \liminf_{n\to\infty} F_{X_n}(x) \leq \limsup_{n\to\infty} F_X(x) + P(|X_n - X| > \epsilon).$$

If F_X is continuous at x, then we take $\epsilon \to 0$, which proves the result. ■

Example 11.2.3 (Convergence in distribution does not imply convergence in probability)**.** *Consider $X \sim$ Bernoulli$(\frac{1}{2})$, the sequence $X_i = X$ and $Y = 1 - X$. Clearly, $X_n \xrightarrow{D} Y$, because they have the same distribution. However, $|X_i - Y| = 1$, for all i, so X_n does not converge to Y in probability.*

This example illustrates the difference between the specific behaviour of RVs and their general behaviour as represented by their distribution. The restriction that one coin flips the opposite of the other simplifies the proof but is not neccessary. The argument would hold even if the coins were independent fair coins. In words, two independent fair coins will have the same distribution, but one can still flip heads when the other flips tails or vice versa.

11.3 Useful theorems

Here we state, without proof, several useful theorems:

Proposition 11.3.1. *For any real-valued constants a and b we have*

(a) If $X_n \xrightarrow{a.s.} X$ and $Y_n \xrightarrow{a.s.} Y$, then $aX_n + bY_n \xrightarrow{a.s.} aX + bY$ and $X_nY_n \xrightarrow{a.s.} XY$.

(b) If $X_n \xrightarrow{P} X$ *and* $Y_n \xrightarrow{P} Y$*, then* $aX_n + bY_n \xrightarrow{P} aX + bY$ *and* $X_nY_n \xrightarrow{P} XY$*.*

(c) If $X_n \xrightarrow{r} X$ *and* $Y_n \xrightarrow{r} Y$*, then* $aX_n + bY_n \xrightarrow{r} aX + bY$*.*

Theorem 11.3.1 (Slutsky's theorem)**.** *For any real-valued constant* c *we have:*

(a) If $X_n \xrightarrow{D} X$ *and* $Y_n \xrightarrow{D} c$*, then* $X_n + Y_n \xrightarrow{D} X + c$*.*

(b) If $X_n \xrightarrow{D} X$ *and* $Y_n \xrightarrow{D} c$*, then* $X_nY_n \xrightarrow{D} cX$*.*

Theorem 11.3.2 (Continuous mapping theorem)**.** *Let* g *be a function that is discontinuous at most on a set of measure 0. Then,*

(a) If $X_n \xrightarrow{a.s.} X$*, then* $g(X_n) \xrightarrow{a.s.} g(X)$*.*

(b) If $X_n \xrightarrow{P} X$*, then* $g(X_n) \xrightarrow{P} g(X)$*.*

(c) If $X_n \xrightarrow{D} X$*, then* $g(X_n) \xrightarrow{D} g(X)$*.*

Chapter 12

Limit theorems

12.1 Borel-Cantelli lemmas

The Borel-Cantelli lemmas are several results that talk about finite or infinite occurrence of events. We state the two most common ones.

The first Borel-Cantelli lemma says that if the sum of probabilities of a sequence of events is finite, then the probability of infinitely many of them occurring is 0 (that is, finitely many of them will occur almost surely):

Theorem 12.1.1 (First Borel-Cantelli lemma)**.** *Let $\{A_i\}$ be a sequence of events, such that $\sum_{n=1}^{\infty} P(A_n) < \infty$. Then $P(\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k) = 0$.*

Proof.

$$\begin{aligned}
P(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m) &= P(\bigcap_{n=1}^{\infty} B_n) && (\text{notation } B_n = \bigcup_{m=n}^{\infty} A_m) \\
&= \lim_{n\to\infty} P(B_n) && (\text{continuity of probability}) \\
&= \lim_{n\to\infty} P(\bigcup_{m=n}^{\infty} A_m) \\
&\leq \lim_{n\to\infty} \sum_{m=n}^{\infty} P(A_m) && (\text{Boole's inequality}) \\
&= 0. && (\text{sum is finite by assumption})
\end{aligned}$$

■

The second Borel-Cantelli lemma says that if the sum of probabilities of a sequence of independent events is infinite, then the probability of only finitely

many of them occurring is 0 (that is, infinitely many of them will occur almost surely):

Theorem 12.1.2 (Second Borel-Cantelli lemma)**.** *Let $\{A_i\}$ be a sequence of independent events, such that $\sum_{n=1}^{\infty} P(A_n) = \infty$. Then $P(\cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k^c) = 0$.*

Proof.

$$\begin{aligned} P(\bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m^c) &\leq \sum_{n=1}^{\infty} P(\bigcap_{m=n}^{\infty} A_m^c) && \text{(Boole's inequality)} \\ &= \sum_{n=1}^{\infty} \lim_{k\to\infty} P(\bigcap_{m=n}^{k} A_m^c) && \text{(continuity of probability)} \\ &= \sum_{n=1}^{\infty} \lim_{k\to\infty} \prod_{m=n}^{k} P(A_m^c) && \text{(assumed independence)} \\ &= 0. && \text{(by Lemma 12.1.3)} \end{aligned}$$

■

The following Lemma is required in the above proof:

Lemma 12.1.3. *Let p_i be a sequence of numbers between 0 and 1. If $\sum_{i=1}^{\infty} p_i = \infty$ then $\lim_{n\to\infty} \prod_{i=1}^{n} (1 - p_i) = 0$.*

Proof. Since $\log(1 - p_i) \leq -p_i$, we have

$$\prod_{i=1}^{n} (1 - p_i) = \prod_{i=1}^{n} e^{\log(1-p_i)} \leq \prod_{i=1}^{n} e^{-p_i} = e^{-\sum_{i=1}^{n} p_i}.$$

By taking the limit of both sides, we get the desired result. ■

12.2 Weak Law of Large Numbers

Theorem 12.2.1 (WLLN)**.** *Let X_i be a sequence of identically distributed independent random variables with mean $E[X]$. Define $S_n = \sum_{i=1}^{n} X_i$. Then, $\frac{S_n}{n} \xrightarrow{P} E[X]$.*

Partial proof assuming X_i have finite variance σ^2. By the linearity of expectation and variance, we can show that $E[\frac{S_n}{n}] = E[X]$ and $Var[\frac{S_n}{n}] = \sigma_S^2 = \frac{\sigma^2}{n}$.

By Chebyshev's inequality we have $P(|\frac{S_n}{n} - E[X]| > a\sigma_S) \leq \frac{1}{a^2}$ for any $a > 0$. Substituting $a = \frac{\epsilon}{\sigma_S}$, where $\epsilon > 0$, we get

$$P(|\frac{S_n}{n} - E[X]| \geq \epsilon) \leq \frac{\sigma_S^2}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

By increasing n the right-hand side can be made arbitrarily close to 0 for any ϵ and σ, therefore, in the limit, the left-hand side tends to 0.

■

12.3 Strong Law of Large Numbers

Theorem 12.3.1 (SLLN)**.** *Let X_i be a sequence of identically distributed independent random variables with finite mean $E[X]$. Define $S_n = \sum_{i=1}^{n} X_i$. Then, $\frac{S_n}{n} \xrightarrow{a.s.} E[X]$.*

Partial proof assuming X_i have finite variance σ^2. We will assume that the $X_i \geq 0$ and generalize at the end of the proof. From the proof of the WLLN, we already have

$$P(|\frac{S_n}{n} - E[X]| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Now we consider a deterministic subsequence of squared indices:

$$\sum_{j=1}^{\infty} P(|\frac{S_{j^2}}{j^2} - E[X]| \geq \epsilon) \leq \sum_{j=1}^{\infty} \frac{\sigma^2}{j^2\epsilon^2} < \infty.$$

We took squared indices to obtain convergence of the right-hand side. Finiteness of the left-hand side implies, together with the First Borel-Cantelli lemma, that $\frac{S_{j^2}}{j^2} \xrightarrow{a.s.} E[X]$.

Now we show that this also holds for $j^2 \leq n \leq (j+1)^2$. Since $X_i \geq 0$, we have

$$\begin{aligned}
S_{j^2} \leq S_n &\leq S_{(j+1)^2} \\
\frac{S_{j^2}}{(j+1)^2} \leq \frac{S_n}{n} &\leq \frac{S_{(j+1)^2}}{j^2} \\
\frac{S_{j^2}}{(j+1)^2}\frac{i^2}{i^2} \leq \frac{S_n}{n} &\leq \frac{S_{(j+1)^2}}{j^2}\frac{(j+1)^2}{(j+1)^2} \\
\frac{S_{j^2}}{j^2}\frac{j^2}{(j+1)^2} \leq \frac{S_n}{n} &\leq \frac{S_{(j+1)^2}}{(j+1)^2}\frac{(j+1)^2}{j^2}.
\end{aligned}$$

As $j \to \infty$, we have

$$E[X] \leq \frac{S_n}{n} \leq E[X].$$

Therefore, $\frac{S_n}{n} \xrightarrow{a.s.} E[X]$.

To generalize to arbitrary RVs with finite variance, we write $X_n = X_n^+ - X_n^-$. Since both terms on the right-hand side are non-negative and have finite variance, the same arguments apply.

■

12.4 Central Limit Theorem

Theorem 12.4.1 (CLT)**.** *Let X_i be a sequence of identically distributed independent random variables with mean $E[X]$ and finite variance $Var[X]$. Define $S_n = \sum_{i=1}^n X_i$. Then, $\sqrt{n}(\frac{S_n}{n} - E[X]) \xrightarrow{D} N(0, Var[X])$. Or, equivalently $\frac{S_n}{n} \xrightarrow{D} N(E[X], \frac{Var[X]}{n})$.*

Partial proof assuming the existence of the MGF. Define a new sequence of random variables $Y_i = \frac{X_i - E[X]}{\sqrt{Var[X]}}$. Clearly, Y_i are also independent and identically distributed, with $E[Y] = 0$ and $Var[Y] = 1$.

The Taylor expansion of the MGF of Y_i around 0 is

$$M_Y(t) = E[e^{tY}] = M_{Y_i}(0) + tE[Y] + \frac{t^2}{2}Var[Y] + t^2 h(t) = 1 + \frac{t^2}{2} + t^2 h(t),$$

where $h(t)$ goes to 0 as t goes to 0.

Now we introduce $Z_n = \frac{\sum_{i=1}^n Y_i}{\sqrt{n}}$. Using the properties of MGF, the MGF of Z_n is

$$M_Z(t) = M_Y(\frac{t}{\sqrt{n}})^n = (1 + \frac{t^2}{2n} + \frac{t^2}{n}h(\frac{t}{\sqrt{n}}))^n.$$

As $n \to \infty$ $M_Z(t) \to e^{\frac{t^2}{2}}$. This is the MGF of the standard normal distribution, which completes the proof. ■

We state the following three theorems without proof.

Theorem 12.4.2 (Berry-Esseen inequality)**.** *Suppose that X_i also have a finite third moment. Then,*

$$\sup_s |P(\frac{\sqrt{n}(\frac{S_n}{n} - E[X])}{\sqrt{Var[X]}} \leq s) - \Phi(s)| \leq \frac{33}{4}\frac{E[|X - E[X]|^3]}{\sigma^3\sqrt{n}},$$

where Φ is the CDF of the standard normal distribution.

Theorem 12.4.3 (Multivariate CLT)**.** *Let X_i be a sequence of identically distributed independent k-dimensional random vectors with finite mean $E[X] = [\mu_1, ..., \mu_k]^T$ and covariance Σ. Define $S_n = \sum_{i=1}^{n} X_i$. Then, $\sqrt{n}(\frac{S_n}{n} - E[X]) \xrightarrow{D} N(0, \Sigma)$.*

Theorem 12.4.4 (The Delta method)**.** *Let X_i be a sequence of random variables, such that $X_n \xrightarrow{D} N\left(\mu, \frac{\sigma^2}{n}\right)$. Let g be a differentiable function, such that $g'(\mu) \neq 0$. Then,*

$$g(X_n) \xrightarrow{D} N\left(g(\mu), (g'(\mu))^2 \frac{\sigma^2}{n}\right).$$

Chapter 13

Markov chains

A Markov chain is a mathematical model that has numerous practical applications, including some that are particularly useful for data analysis methods and computation, such as simulation and sampling-based approximation to expectations. Note that our treatment of the subject is biased towards results that are essential for the introduction of Markov Chain Monte Carlo (MCMC).

Markov chains are a type of *stochastic process*. Take a probability space $(\Omega, \mathcal{F}, P)$ and a measurable space $(S, \mathcal{S})$. A stochastic process is a family of random variables $X : \Omega \to S$ indexed by t: $\{X(t), t \in T\}$, $X(t) \in S$. The set S is the *state space* of the stochastic process. The index set T can be thought of as time and can be uncountable. Continuous-time Markov chains have many applications and are interesting in their own right. However, for our purposes we can restrict ourselves to the index set of natural numbers. We can think of a *discrete-time* stochastic process as a sequence of random variables $X_0, X_1, X_2, \ldots$ that take values in set S.

13.1 Countable state space

Definition 13.1.1. A homogeneous discrete-time *Markov chain* with a countable state space is a discrete-time stochastic process with a countable state space, such that

$$P(X_{i+1} = x_{i+1} | X_i = x_i, X_{i-1} = x_{i-1}, \ldots, X_0 = x_0) = P(X_{i+1} = x_{i+1} | X_i = x_i).$$

In words, given the current state of the Markov chain, the transition probabilities to the next state are conditionally independent of the history of the process. This defining property is also known as the *Markov property*.

We will only be concerned concerned with *homogeneous* or *stationary* Markov chains - Markov chains where $P(X_{i+1} = x_{i+1} | X_i = x_i)$ does not depend on i

(that is, does not change over time).[1] We can compactly represent the transition probabilities of a homogeneous Markov chain:

Definition 13.1.2. The one time step *transition matrix* of a homogeneous countable state space Markov chain is the function $K(x, y) : S \times S \to [0, 1]$, such that $K(x, y) \triangleq P(X_{i+1} = y | X_i = x)$.

For finite S the transition matrix K can be represented with a matrix, where each row and column correspond to a state and the values in a row represent the probability vector for transitions from that state to all other states.

Proposition 13.1.1 (Chapman-Kolmogorov)**.** *For every $m, n \geq 0$ and $x, y \in S$, we have*

$$K^{m+n}(x, y) = \sum_{z \in S} K^m(x, z) K^n(z, y).$$

The proof is left as an exercise.

Theorem 13.1.1. $P(X_{i+m} = y | X_i = x) = K^m(x, y)$.

That is, the m-step transition probability is the $m-$th power of the transition matrix K. The proof is left as an exercise.

We now introduce the first concepts that will allow us to study Markov chains and identify those that are of particular interest.

Definition 13.1.3. We say that a state y is *reachable* from state x if there exists a $m \geq 0$ such that $K^m(x, y) > 0$.

Definition 13.1.4. We say that states x and y *communicate* and write $x \sim y$ if y is reachable from x and x is reachable from y.

Proposition 13.1.2. *The communicate relation is an equivalence relation (it is reflexive, symmetric, and transitive).*

The proof is left as an exercise.

The communicate relation partitions S into equivalence classes. We will be particularly interested in cases where there is only one class. That is, where every state communicates with every other state.

Definition 13.1.5. A Markov chain is *irreducible* if $x \sim y$ for every pair $x, y \in S$.

We also want to rule out another class of Markov chains - chains that exhibit periodic behavior.

Definition 13.1.6. The *period* of a Markov chain state x is the greatest common divisor of all n such that $K^n(x, x) > 0$. A state with period > 1 is called

[1] Non-homogeneous Markov chains are too general to allow for strong theoretical results.

periodic and a state with period 1 is *aperiodic*. Similarly, a Markov chain where all states have period 1 is called *aperiodic*.

Observe how a period $a > 1$ implies that the Markov chain can return to state x only at step counts that are multiples of a, hence, periodically.

Proposition 13.1.3. *If $K(x, x) > 0$, then state x is aperiodic.*

The proof is left as an exercise. Note that this property is relatively easy to satisfy in practice and therefore a convenient way of ensuring aperiodicity when constructing Markov chains for MCMC.

Proposition 13.1.4. *If states x and y communicate, then they have the same period.*

The proof is left as an exercise. A direct corollary of this proposition is that all states in an irreducible Markov chain have the same period.

Now we are ready to discuss the limiting behavior of a Markov chain. First, we define the stationary distribution of a Markov chain:

Definition 13.1.7. A distribution $\pi(x)$ on S is stationary for a Markov chain with transition matrix K if $\pi K = \pi$ or, equivalently,

$$\sum_{y \in S} \pi(y) K(y, x) = \pi(x).$$

In words, a stationary distribution is invariant - if we are at some point in time distributed with that distribution and make one step according to the transition matrix, we remain in that distribution.

A Markov chain can have more than one stationary distribution or it can be without a stationary distribution. However, limiting ourselves to irreducible and aperiodic Markov chains substantially simplifies asymptotic behavior, as shown by two very important theorems:

Theorem 13.1.2. *For a irreducible aperiodic Markov chain with a stationary distribution π we have $\lim_{n\to\infty} K^n(x, y) = \pi(y)$ for any initial distribution.*

The theorem basically says that an irreducible aperiodic countable state space Markov chaing will converge to its stationary distribution, if it has one, regardless of where we start. We state this theorem without proof.

Finally, we have the SLLN analogue for Markov chains:

Theorem 13.1.3. *For any irreducible Markov chain with stationary distribution π and function $f : S \to \mathbb{R}$ with $E_\pi[f(X)] = \sum_{x \in S} f(x)\pi(x) < \infty$ we have*

$$\frac{1}{n} \sum_{i=1}^{n} f(X_i) \xrightarrow{a.s.} E_\pi[f(X)]$$

for any initial distribution.

We state this theorem without proof. Note that aperiodicity is not necessary for the SLLN to apply.

Existence and uniqueness of a stationary distribution

The above limit theorems for countable-space Markov chains assume the existence of a stationary distribution. In this section we provide more tools for determining its existence, its uniqueness, and identifying whether a distribution is a stationary distribution of a Markov chain.

First, we need to add to our classification of states:

Definition 13.1.8. The *first positive return time* T_x of state x is $T_x \triangleq \min\{n \geq 1 : X_n = x | X_0 = x\}$.

That is T_x is the number of steps before the chain that started in x returns to x for the first time. Note that T_x is a random variable.

Definition 13.1.9. A state is called *transient* if $P(T_x < \infty) < 1$. If it is not transient it is *recurrent*.

In words, recurrent states have that we will return to them in a finite number of steps with probability 1 (almost surely).

This classification can be further refined as follows.

Definition 13.1.10. A state x is *positive-recurrent* if it is recurrent and $\mathrm{E}[T_x] < \infty$. If a state is recurrent and $\mathrm{E}[T_x] = \infty$, then it is *null-recurrent*. Otherwise it is transient.

This refinement is necessary, because for a countably infinite state space we can have cases where we will return in a finite number of steps with probability 1, but our expected return time would be infinite.[2] That is, despite the state being recurrent, the visits would not be frequent enough for the chain to have a stationary distribution (see Example 13.1.1).

Proposition 13.1.5. *Positive-recurrence, null-recurrence and transience are class properties - all states in a communicating class share them.*

Proof. Let x and y be states that communicate. That is, there exists a $m \geq 0$, such that $K^m(x, y) > 0$. Suppose x is recurrent. It follows that y must also be recurrent, because every time we re-visit x there is a non-zero probability that we will re-visit y in m steps. A similar argument can be used for null-recurrence and transience. ■

[2] While it might at first be counter-intuitive, there exists random variables that take only finite values but still have an infinite expectation.

Therefore, if one state in an irreducible chain is positive-recurrent, all states are positive recurrent. If all states in a chain are positive-recurrent, we say that the chain is positive-recurrent.

Theorem 13.1.4. *An irreducible Markov chain has a stationary distribution if and only if it is positive recurrent. If it does, then the stationary distribution π is unique and*

$$\pi(x) = \frac{1}{\mathrm{E}[T_x]} > 0.$$

Partial proof of uniqueness assuming $K(x, y) > 0$ for all pairs of states. Suppose an irreducible Markov chain has more than one stationary distribution and take two of those stationary distributions, π_1 and π_2. Let x be the state that maximizes $\frac{\pi_1(z)}{\pi_2(z)}$ over all $z \in S$ and let $a = \frac{\pi_1(x)}{\pi_2(x)}$. It follows that $a\pi_2(z) \geq \pi_1(z)$, for all $z \in S$. Because the chain is irreducible, we have

$$\pi_1(x) = \sum_{z \in S} \pi_1(z)K(z, x) \leq \sum_{z \in S} a\pi_2(z)K(z, x) = a\pi_2(x) = \pi_1(x).$$

The inequality must therefore never be strict and we have $\pi_1(z) = a\pi_2(z)$ for all $z \in S$. This is where we use the assumption $K(x, y) > 0$ for all $x, y \in S$ - if $K(z, x) = 0$ the equality would hold even if $\pi_1(z) < a\pi_2(z)$. Because the stationary distributions must sum to 1, it follows that $\pi_1(z) = \pi_2(z)$, for all $z \in S$. ■

Because all states in an irreducible finite-state Markov chain are recurrent and positive-recurrent, we have:

Corollary 13.1.1. *An irreducible finite-state Markov chain has a unique stationary distribution.*

Proof. Because positive-recurrence is a class property, it suffices to show that at least one state is positive-recurrent. Because the state space is finite, if we start in state i at least one state j must be visited an infinite number of times with positive probability. However, because the chain is irreducible, there is a positive probability of getting from j to i. Therefore, there is a positive probability that the chain starting in j will visit j and infinite number of times, which implies that j is positive-recurrent. ■

So, a Markov chain with a countable state space will have a unique stationary distribution if not only can we get from every state to every other state but also visit every state frequently enough. For finite state spaces this simplifies, because being able to get from every state to every other state implies that every state will be visited infinitely many times and with finite return time.

Example 13.1.1 (A null-recurrent chain)**.** *Take a Markov chain whose state space are positive integers. Let* $K(i, i+1) = \frac{i}{i+1}$ *and let* $K(i, 1) = \frac{1}{i+1}$*. All other transition probabilities are 0. That is, from state* i *the chain moves to the next integer with probability* $\frac{i}{i+1}$ *or moves back to state* 1 *with probability* $\frac{1}{i+1}$*.*

All states communicate, so we have an irreducible chain. To classify the states based on recurrence, it therefore suffices to classify one of the states. Let's focus on state 1. We have:

$P(T_1 < \infty) = \frac{1}{2} + \frac{1}{2}\frac{1}{3} + \frac{1}{2}\frac{2}{3}\frac{1}{4} + \frac{1}{2}\frac{2}{3}\frac{3}{4}\frac{1}{5} + ... = \frac{1}{1\times 2} + \frac{1}{2\times 3} + \frac{1}{3\times 4} + \frac{1}{4\times 5}... =$

$\sum_{i=1}^{\infty} \frac{1}{i(i+1)} = \sum_{i=1}^{\infty}(\frac{1}{i} - \frac{1}{i+1}) = 1 - \lim_{n\to\infty} \frac{1}{n} = 1.$

So, the chain is recurrent. However:

$E[T_1] = 1\frac{1}{2} + 2\frac{1}{2}\frac{1}{3} + 3\frac{1}{2}\frac{2}{3}\frac{1}{4} + 4\frac{1}{2}\frac{2}{3}\frac{3}{4}\frac{1}{5} + ... = \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5}... = \sum_{i=1}^{\infty} \frac{1}{i+1} = \infty.$

So, this irreducible chain is recurrent, but not positive recurrent. It is null-recurrent. Therefore, it does not have a stationary distribution! Even though we will return to each state almost surely, it does not happen often enough to result in a stationary distribution.

Finally, the following will be particularly useful for constructing Markov Chains with desirable stationary distributions:

Definition 13.1.11 (Detailed balance)**.** Consider a Markov chain with state space S and transition matrix K. A distribution π is said to satisfy *detailed balance* for this Markov chain if for every pair of states $x, y \in S$ we have

$$\pi(x)K(x, y) = \pi(y)K(y, x).$$

Note that Markov chains that satisfy detailed balance are also referred to as time reversible (or just *reversible*).

Theorem 13.1.5. *If distribution* π *satisfies detailed balance for a Markov chain, then it is a stationary distribution of the Markov chain.*

Proof. We must show that detailed balance of π implies that $\sum_{y\in S} \pi(y)K(y, x) = \pi(x)$ (the definition of a stationary distribution).

From detailed balance, we have

$$\sum_{y\in S} \pi(y)K(y, x) = \sum_{y\in S} \pi(x)K(x, y) = \pi(x) \sum_{y\in S} K(x, y) = \pi(x).$$

The second step follows from the fact that the probabilities of x transitioning to some state y must sum to 1. ■

Note that while it is a sufficient condition, it is not a necessary condition. There exist Markov chains that do not satisfy detailed balance but have a stationary distribution.

13.2 A note on general state space Markov chains

The results for countable state space Markov chains that are relevant to MCMC transfer to general state spaces, albeit with additional measure-theoretic considerations. In this section we briefly discuss this general setting. A detailed treatment can be found in Robert and Casella (2013) or Meyn and Tweedie (2012).

The main difference when moving to a general, possibly uncountably infinite state space S is similar to moving from discrete to continuous random variables. We can no longer talk just about transition probabilities between states (that is, $K(x, y)$, where $x, y \in S$), because all of these probabilities could be 0.

Instead, we introduce a measurable space and we specify the transition probabilities by defining a transition kernel (or just kernel).

Definition 13.2.1. Let $(S, \mathcal{S})$ be a measurable space. A transition kernel $K : S \to \mathcal{S}$ of a Markov Chain is a map

$$K(x, B) = P(X_i \in B | X_{i-1} = x),$$

where $x \in S$ and $B \subseteq \mathcal{S}$, that satisfies:

- For every x, $K(x, B)$ is a probability measure on $(S, \mathcal{S})$.
- For every B, $K(x, B)$ is a measurable function.

Although K now denotes a kernel, the transition matrix that we used for countable S uniquely determines a kernel. That is, for countable state spaces, it suffices to specify a transition matrix.

The Chapman-Kolmogorov equation and definition of a stationary distribution (measure) transfers to the general setting:

Proposition 13.2.1 (Chapman-Kolmogorov)**.** *For every $m, n \geq 0$, $x \in S$, and $A \in \mathcal{S}$, we have*

$$K^{m+n}(x, A) = \int_S K^m(x, dz) K^n(z, A).$$

Definition 13.2.2. A sigma-finite measure $\pi(x)$ on S is stationary for a Markov chain with transition kernel K if

$$\int_S K(y, A)d\pi = \pi(A).$$

The statement that being in detailed balance implies that the distribution is a stationary distribution of the Markov chain also transfers to the general setting. In particular, for continuous state space, we have the detailed balance condition $\pi(x)k(x, y) = \pi(y)k(y, x)$, where π is a density and $k(x, y)$ are densities (also known as transition functions), such that $K(x, A) = \int_A k(x, y)dy$.

The definition of irreducibility does not transfer to general state spaces. While we can talk about the probability of visiting a set $A \in \mathcal{S}$ from state $x \in S$, we can not talk, at least not in general, about the probability of visiting a state x. For example, when we are dealing with densities. Instead, an analogue to irreducibility is constructed via $\phi-$irreducibility, where ϕ is a measure.

Less formally, a chain is $\phi-$irreducible if there exists a measure ϕ on $\mathcal{S}$ such that, whenever $\phi(A) > 0$, the probability of reaching A from x is positive for all $x \in S$. In essence $\phi-$reducibility identifies sets that are always reached with some positive probability, regardless of the starting state.

A chain can be $\phi-$irreducible for many different ϕ. For example, suppose that a chain is $\phi-$irreducible, then it is irreducible for any non-trivial restriction of ϕ. A more useful definition is obtained via the so called *maximal irreducibility measure* ψ, which is an irreducibility measure such that $\psi(A) = 0 \Rightarrow \phi(A) = 0$, for every irreducibility measure ϕ and all sets A. In other words, the maximal irreducibility measure is the irreducibility measure that has the least null sets.

It can be shown, that the maximal irreducibility measure ϕ exists as long as at least one irreducibility measure ψ exists. It is also unique in the sense that all maximal irreducibility measures identify the same null sets. Furthermore, it can be shown that if the chain has a stationary distribution π and is $\phi-$irreducible (for any ϕ), then the chain is recurrent, the stationary distribution is unique, and the chain is $\pi-$irreducible.

Even though a chain is $\phi-$irreducible and has a stationary distribution (hence, a unique stationary distribution), problems with convergence to the stationary distribution might arise if the chain is started at an x that is in the null set. On the other hand, it can be shown that starting the chain at an x from set A where $\phi(A) > 0$ guarantees not only that every $\phi-$positive set will be visited with probability one, but that it will be visited infinitely many times with probability one. The final step is to remove these null sets from the state space. This comes at no harm, because the null set can not be visited from outside the null set. The formalization of this is called *Harris recurrence*. A $\phi-$irreducible chain is Harris-recurrent if the probability of visiting set A from state x is 1 for all x and all $\phi-$positive A. Any $\phi-$irreducible chain can trivially be made Harris recurrent by removing the null set.

Using the above measure-theoretic technicalities, the two Theorems that are key for MCMC also transfer to the general setting:

Theorem 13.2.1. *For a Harris-recurrent aperiodic Markov chain with a stationary measure π we have $\lim_{n\to\infty} K^n(x, A) = \pi(A)$ for any initial distribution.*[3]

Theorem 13.2.2. *For any Harris-recurrent Markov chain with stationary measure π and function $f : S \to \mathbb{R}$ with $E_\pi[f(X)] = \int_S f(x)d\pi < \infty$ we have*

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow{a.s.} E_\pi[f(X)]$$

for any initial distribution.

13.3 Central Limit Theorem for Markov Chains

As we have stated earlier in this chapter, Markov chains, which are a sequence of dependent random variables, admit a SLLN that is no different from the SLLN for sequences of independent random variables. This should not be that surprising, because we know that expectation is linear - the sum of expectations is the expectation of the sum of random variables, regardless of whether the random variables are independent or not.

The same, however, does not apply to variance - the variance of a sum of dependent random variables depends on the covariances. This suggests that if Markov chains do admit a CLT, it would not be the same as the one we stated for independent random variables.

Before we state the CLT, we introduce lag-k autocovariance of a discrete time stationary stochastic process:

$$\gamma_k = \mathrm{Cov}[f(X_i), f(X_{i+k})].$$

That is, the lag-k autocovariance is just the covariance of random variables in the stochastic process that are exactly k apart. Because we are assuming a stationary process, the γ_k is the same for all pairs exactly k apart.

Again, we are interested in estimating the integral $E_\pi[f(X)] = \int f(x)d\pi(x)$ with the average $\hat{f}_n = \frac{1}{n}f(X_i)$ of samples X_i from P. We already discussed the SLLN for Markov chains - for relatively well-behaved chains, we have $\hat{f}_n \xrightarrow{a.s.} E_\pi[f(X)]$. Now we are ready to state the CLT for Markov chains.

[3] A stronger statement is true, that the convergence is not only set-wise but in total variation. That is, uniformly over sets.

Theorem 13.3.1. *For an irreducible Markov chain with stationary distribution π and starting distribution π, we have*

$$n \operatorname{Var}[\hat{f}_n] \to \sigma^2 = \sum_{k=-\infty}^{+\infty} \gamma_k$$

and, if σ^2 is finite

$$\sqrt{n}(\hat{f}_n - E_\pi[f(X)]) \xrightarrow{D} N(0, \sigma^2).$$

The proof of this theorem is beyond the scope of this text. As we can see, the CLT is identical to the CLT for iid RVs, except for the computation of variance of the estimator, which must now take into account the dependencies between the RVs.

The estimation of the above variance is very important in practice, because it leads to Monte Carlo standard errors or some other quantification of the error of our Monte Carlo estimator. We will explore the topic further by showing how the lag-k covariances arise and how we can estimate them in practice.

From basic probability theory we know that for a mean of n possibly dependent RVs $X_1, ..., X_n$ we have

$$\sigma_n^2 = n \operatorname{Var}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} \operatorname{Var}[X_i] + \frac{1}{n} 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{Cov}[X_i, X_j].$$

If we assume that RVs k apart have the same covariance (in the context of stochastic processes, homogeneity) and use the lag-k autocovariance notation ($\gamma_0 = \operatorname{Var}[X]$, zero-lag autocovariance), we get

$$\sigma_n^2 = \gamma_0 + 2 \sum_{i=1}^{n-1} \frac{n-i}{n} \gamma_i,$$

which converges to

$$\sigma_n^2 = \gamma_0 + 2 \sum_{i=1}^{\infty} \gamma_i = \sum_{i=-\infty}^{\infty} \gamma_i$$

as $n \longrightarrow \infty$. Note that, by definition, $\gamma_k = \gamma_{-k}$.

Estimating this variance is an important practical problem. We discuss it further in Section 20.2.

Exercises

Exercise 13.1. Prove Proposition 13.2.1.

Exercise 13.2. Prove Theorem 13.1.1. Hint: Use Proposition 13.2.1.

Exercise 13.3. Prove Proposition 13.1.2.

Exercise 13.4. Prove Proposition 13.1.3.

Exercise 13.5. Prove Proposition 13.1.4.

Part II

Reasoning with uncertainty

Chapter 14

Introduction to statistical inference

14.1 Data, model, parameters

The essence of statistical analysis is inferring (learning) the properties of the underlying process that generated the data.

For example, observe data that were generated by flipping a (not necessarily fair) coin 10 times. Is it a fair coin?

H T H T H H T H H H.

We often refer to this a sample from our data generating process. The sample has certain properties, for example, the relative frequency of heads is 0.7.

While it might be tempting to use the properties of the sample as a substitute for the properties of the data generating process, they are not the same. To further illustrate this point, let's flip the same coin again using the exact same process:

T T H T T H H H T H.

The sequence is different than before as is the relative frequency of heads (0.5). The properties of the process are unchanged, but the properties of the data are different. Clearly, the two are not the same. And thinking that they are is one of the most common sources of misunderstanding, misinterpretation and flawed statistical analyses.

Without a doubt, the properties of the data depend on the properties of the data generating process. If they didn't, we couldn't learn anything about the process from the data it generates. Furthermore, the bigger our sample size, the more are the properties of the generating process reflected in the properties

of data. In fact, this is formally suggested by the limit theorems. We know, for example, that the more coin flips we have, the closer their relative frequency will be to the underlying expectation of flipping heads. However, in statistical analyses, there will always be at least some uncertainty associated with where the properties of the process lie and one of the main tasks of statistical inference is to quantify that uncertainty.

The properties of data are easy to compute and the properties of the generating process that interest us are typically determined by the problem we are trying to solve. In our case, for example, the expected value of the coin flipping heads. The challenge and art of data analysis lies in selecting a suitable relationship between the two - a hypothesis of how the data were generated or *a model*.

In our example, it is reasonable to interpret the data as if they were generated by drawing independent samples from a Bernoulli distribution with parameter $\theta \in [0, 1]$:

$$Y_1 = y_1, ..., Y_{20} = y_{20} | \theta \sim_{iid} \text{Bernoulli}(\theta).$$

Because the sum of independent and identically distributed Bernoulli variables is distributed Binomial, we could equivalently write

$$Y = \sum_{i=1}^{20} y_i | \theta \sim \text{Binomial}(20, \theta).$$

In fact, if we can assume that the coin tosses are independent and stationary (their expectation is constant over time), then this is one of the rare cases where there is only one choice of model. Also, our parameter θ has a straightforward interpretation $\theta = E[Y]$.

In this case our model is an explicit distributional assumption with a finite number of parameters. We will refer to these kind of models as *parametric*. Models that are not parametric will be called *non-parametric*. In non-parametric models the distributional assumption is implicit and/or the number of parameters is infinite. Note that this is a non-rigorous practical distinction. The theoretical differences between parametric and non-parametric models are more nuanced and out of the scope of this text.

Our choice of parametric model determines the distribution of the data given the parameters. The notation $Y = \sum_{i=1}^{20} y_i | \theta \sim \text{Binomial}(20, \theta)$ can be unrolled into the underlying explicit distributional assumption that

$$p(Y = y | \theta) = \binom{20}{y} \theta^y (1 - \theta)^{20-y},$$

where $y = \sum_{i=1}^{20} y_i$.

The function $p(y|\theta)$ plays a central role in parametric inference. It is called the *likelihood* function and we will often write $L(\theta; y) = p(y|\theta)$ and $\ell(\theta; y) = \log p(y|\theta)$. This naming convention and notation are deliberate to make the distinction between the likelihood and PMFs/PDFs. The function $p(y|\theta)$ when viewed as a function of θ with y known (as is typically the case with statistical analyses) is not a PMF or PDF.

What remains is to infer θ from the data. There are many different approaches to statistical inference to choose from, each with its advantages and disadvantages. The remainder of the chapter is dedicated to three examples that will briefly illustrate three of the most common: maximum likelihood, Bayesian inference and null-hypothesis significance testing. Each of these will be covered in more detail in the following chapters.

14.2 Approaches to statistical inference

Example 14.2.1 (Maximum likelihood estimate for a Binomial proportion)**.** *Maximum likelihood is, as the name suggests, concerned with finding the parameter values that maximize the likelihood. That is, the parameter value that is out of all parameter values the most likely to have generated the data.*

In our example we have a sample with $n = 20$ observations, 12 of which are heads and 8 are tails. Inserting this data into the chosen likelihood, we get

$$L(\theta; y) = \binom{20}{12} \theta^{12}(1-\theta)^8.$$

From this point maximum likelihood estimation becomes an optimization problem of finding the value of θ that maximizes the value of $L(\theta; y)$. In most cases it is more convenient to work with the log-likelihood $\ell(\theta; y)$. The log-likelihood is also numerically more stable, because products of probabilities/densities are turned into sums of their logarithms. Because the logarithm is a monotone increasing function the maximum of the likelihood is the same as the maximum of the log-likelihood. Taking the derivative

$$\frac{d}{d\theta}\ell(\theta; y) = 12\frac{1}{\theta} - 8\frac{1}{1-\theta},$$

we can see that it is 0 at $\theta = \frac{12}{20}$. The maximum likelihood estimate of θ is therefore 0.6. It should not be too surprising that the maximum likelihood estimate corresponds to the sample average.

This type of estimation falls into the category of point estimation, because the result is only a point in the parameter space. Point estimates are often good

enough, especially if we are interested only in prediction, but they lack the quantification of uncertainty that is necessary for making decisions. To illustrate this point, imagine a coin that flips heads 6000 times out of 10000. The maximum likelihood point estimate would be the same, 0.6, yet our intuition (and the limit theorems) tells us that the second estimate is more reliable. If we had to choose a coin that we believe is more likely to be unfair, we would choose the second coin.

In Chapter 16 we discuss maximum likelihood in more detail, including how to quantify uncertainty in maximum likelihood estimates. In this example we will only construct a crude confidence interval with the tools we already have. We know, by the CLT, that the sample relative frequency of 20 Bernoulli trials will be distributed approximately normally around the mean

$$y \approx N(\theta, \frac{\sigma_Y^2}{20}).$$

The variance of a Bernoulli RV is at most 0.25, so we state with at least 95% probability that the interval $[0.6 - 2 \times \frac{0.5}{\sqrt{20}}, 0.6 + 2 \times \frac{0.5}{\sqrt{20}}]$ *or* $[0.38, 0.82]$ *contains the true* θ*. This interval is wide and contains* 0.5*, so it's not strong evidence against the coin being fair. That is, if the true mean were* 0.5 *it would not be that surprising to get such a sample.*

Note that this construction is based on the assumption that our sample size is large enough for the CLT to apply. Better techniques exist for binary data and will be discussed later.

Example 14.2.2 (Bayesian inference for a Binomial proportion)**.** *The fundamental difference between Bayesian and so-called frequentist or classical statistics lies not in statistics but in how we view probability. If our view in Example 14.2.1 was that the data* y *are random and the parameter* θ *is an unknown constant, the Bayesian view is that* θ *is random variable. Not because it is random, it might very well be a constant, but because we choose to represent our uncertainty in what the value of* θ *might be with a random variable.*

The main objective of Bayesian statistics is to compute the posterior distribution $p(\theta|y)$ *- the distribution of the parameter after we see the data. This is done using Bayes' theorem which gave Bayesian statistics its name:*

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}.$$

To compute the posterior distribution, we need the likelihood and $p(\theta)$ *- the* prior distribution *or* prior*. The prior can be interpreted as our probabilistic opinion*

about the parameter before we see the data.

How do we select a prior? There are many different approaches, subjective and objective with respect to some criterion, which we will discuss in Chapter 18. For now, we will assume that we don't and we assume that a uniform distribution over all possible θ is an adequate representation of this:

$$\theta \sim \text{Unif}(0,1) \text{ or, equivalently } p(\theta) = 1.$$

Now we have the prior and the likelihood and we can compute the posterior:

$$\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\
&\propto p(y|\theta)p(\theta) && \text{(proportional to)} \\
&= p(y|\theta) && \text{(prior density is 1)} \\
&= \binom{20}{12}\theta^{12}(1-\theta)^8 \\
&\propto \theta^{12}(1-\theta)^8
\end{aligned}$$

Because PDFs (PMFs) integrate (sum) to 1, the scale of the PDF (PMF) is irrelevant, because it can be recovered from the shape of the distribution. That is, the shape of the PDF/PMF is enough to uniquely identify it, so we ignore multiplicative constants. The shape $\theta^{12}(1-\theta)^8$ is that of the Beta distribution with parameters $\alpha = 13$ and $\beta = 9$:

$$p(\theta|y) = \frac{1}{B(13,9)}\theta^{13-1}(1-\theta)^{9-1}.$$

The main advantage of Bayesian statistics is that we can answer probabilistic questions about θ - all the information we need is contained in the posterior distribution $p(\theta|y)$. For example, we can compute an interval where the true value of θ lies with 95% probability. That is, the interval

$$(Q_{\theta|y}(0.025), Q_{\theta|y}(0.975)) \approx (0.38, 0.78),$$

where Q is the quantile function of the posterior.

Similarly, we could, for example, compute the probability that this coin's θ is greater than 0.5:

$$P(\theta > 0.5|y) = \int_{0.5}^{1.0} p(\theta|y)d\theta \approx 0.81.$$

Example 14.2.3 (Exact Binomial test for proportion)**.** *Null-hypothesis significance testing, as the name suggests, focuses on testing a particular hypothesis. While there are many different tests, they all follow the same process:*

- *First, the hypothesis of interest that is to be tested is assumed to be true. Hence, the name* null hypothesis*.*
- *Second, a test statistic is chosen and we compute its distribution given the null hypothesis.*
- *Third, the value of the test statistic for the sample is compared to the distribution. If it is very unlikely to have been generated were the null hypothesis true, we choose to reject the null hypothesis.*

In our case we might be interested in testing if the coin is fair. Therefore, the null hypothesis would be that it is:

$$H_0 : \theta = \frac{1}{2}.$$

A reasonable test statistic in this case would be the deviation from the expected equi-distribution of tails and heads. The bigger the absolute difference between number of tails and number of heads, the less likely it is that the coin is fair. Our sample has an absolute difference of $12-8 = 4$*. The probability of obtaining a difference at least this large is*

$$P(|\#heads \text{ - } \#tails| \geq 4|H_0) = P(y \leq 8 \text{ or } y \geq 12|H_0) = 2F(8|H_0) \approx 0.50.$$

Null-hypothesis significance testing has a long tradition of both use and misuse. We will discuss this approach to inference in more detail in Chapter 17.

Chapter 15

Plug-in estimators and the bootstrap

15.1 Empirical CDF

Definition 15.1.1. Let $\{X_n\}$ be independent and identically distributed RVs. Their empirical cumulative distribution function (ECDF) is defined as

$$F_n(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leq x}.$$

Proposition 15.1.1 (Properties of the ECDF)**.** *For all x, we have*

(a) $E[F_n(x)] = F(x)$.

(b) $Var[F_n(x)] = \frac{F(x)(1-F(x))}{n}$.

(c) $F_n(x) \xrightarrow{a.s.} F(x)$.

(d) $\sup_x |F_n(x) - F(x)| \xrightarrow{a.s.} 0$ (Glivenko-Cantelli theorem).

Proof. The indicator random variable for a particular x takes value 1 with probability $F(x)$. Hence, we are dealing with a Bernoulli random variable, so (a) and (b) follow. Property (c) follows from the SLLN. Property (d) is stated without proof. ■

15.2 Statistical functionals and the plug-in principle

Definition 15.2.1. A *statistical functional* T is a map from the space of CDFs to $\mathbb{R}$.

Definition 15.2.2. The *plug-in estimator* of statistical functional $\theta = T(F)$ is defined as $\hat{\theta}_n \triangleq T(F_n)$.

15.3 Properties of point estimators

Before we discuss the properties of plug-in estimators, we define some general notions that are helpful in characterizing the usefulness of estimators of statistical functionals.

Definition 15.3.1. A point estimator $\hat{\theta}_n$ of θ is *unbiased* if $\mathrm{E}[\hat{\theta}_n] = T(F)$.

Definition 15.3.2. A point estimator $\hat{\theta}_n$ of θ is *consistent* if $\hat{\theta}_n \xrightarrow{P} \theta$.

Definition 15.3.3. The mean square error (MSE) of a point estimator is $\mathrm{MSE}(\hat{\theta}) \triangleq \mathrm{E}[(\hat{\theta}_n - \theta)^2]$.

Proposition 15.3.1 (Bias-variance decomposition)**.** *MSE*$(\hat{\theta}) = (\mathrm{E}[\hat{\theta}] - \theta)^2 + \mathrm{Var}[\hat{\theta}]$.

The proof is left as an exercise.

Proposition 15.3.2. *If MSE*$(\hat{\theta}_n) \to 0$ *then* $\hat{\theta}_n$ *is consistent.*

The proof is left as an exercise.

Definition 15.3.4. A point estimator $\hat{\theta}_n$ is *asymptotically normal* if

$$\frac{\hat{\theta}_n - \theta}{\sigma_{\hat{\theta}_n}} \xrightarrow{D} N(0, 1).$$

For most common functionals, their plug-in estimators have at least some of these desirable properties (see Exercises). Unfortunately, there are no general rules for when a plug-in estimator is a good estimator. Although the Glivenko-Cantelli theorem has very strong implications, it is not enough to imply $T(F_n) \to T(F)$, because a small change in F_n could still cause a big change in $T(F_n)$. In order for the implication to work, T has to be sufficiently smooth, where the notion of smoothness here is more general, because we are talking about functions.

15.4 TODO: Linear functionals, influence function

15.5 Bootstrapping the variance of an estimator

Plug-in estimators are very simple and convenient estimators of parameters. In most cases our plug-in estimator will have nice asymptotic properties, but in practice, we are interested in how good our estimate is on a finite sample. In this section we will introduce a very simple but powerful and general approach to estimating the uncertainty associated with a plug-in estimator: *the bootstrap*.

We start by describing the algorithm:

Algorithm 15.5.1. *Let the ECDF F_n be our sample from F and let T be a functional of interest. Then, the following algorithm returns an estimate of the variance of the plug-in estimator $\hat{\theta}_n$:*

1: ***procedure*** BOOTSTRAP-VARIANCE*(F_n, T, m)*
2: ***for*** $i \leftarrow 1 : m$ ***do*** ▷ *number of bootstrap samples*
3: *sample $X_1^*, ..., X_n^*$ iid from F_n* ▷ *sampling with replacement*
4: *let F_n^* represent $X_1^*, ..., X_n^*$*
5: $\hat{\theta}_{n,i}^* \leftarrow T(F_n^*)$
6: ***end for***
7: ***return*** $\frac{1}{m-1} \sum_{i=1}^{m} \left(\hat{\theta}_{n,i}^* - \bar{\theta}_n^*\right)^2$, *where* $\bar{\theta}_n^* = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_{n,i}^*$.
8: ***end procedure***

The key idea of the bootstrap is to simulate m replications of the original sample by sampling with replacement. For each replication we compute the value of the functional and we use these m values of the functional to estimate the variance.

Before we formally state the statement made by this algorithm, we formally define the objects from the algorithm.

Definition 15.5.1. The *bootstrap sample* from ECDF F_n is defined as a set of n independent samples from F_n. That is,

$$X_1^*, X_2^*, ..., X_n^* \sim F_n.$$

Definition 15.5.2. The bootstrap empirical cumulative distribution function is defined as

$$F_n^*(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} I_{X_i^* \leq x}.$$

Definition 15.5.3. The bootstrap plug-in estimator is defined as $\hat{\theta}_n^* \triangleq T(F_n^*)$.

Definition 15.5.4 (Estimated bootstrap variance)**.** $\widehat{\mathrm{Var}}_B[\hat{\theta}_n^*] = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\theta}_{n,i}^* - \bar{\theta}_n^*\right)^2$, where $\bar{\theta}_n^* = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_{n,i}^*$ and m is the number of bootstrap replications.

The statement we are making with this algorithm is that the estimated bootstrap variance $\widehat{\mathrm{Var}}_B[\hat{\theta}_n^*]$ is a good estimate for the variance of the plug-in estimator $\mathrm{Var}[\hat{\theta}_n]$. Note that $\widehat{\mathrm{Var}}_B[\hat{\theta}_n^*]$ itself is just an estimate for the bootstrap variance $\mathrm{Var}_B[\hat{\theta}_n^*|F_n]$, which is in practice very difficult to compute exactly due to the combinatorial explosion of possible combinations:

Definition 15.5.5 (Bootstrap variance)**.** $\mathrm{Var}[\hat{\theta}_n^*|F_n] = \frac{1}{n^n}\sum_{i}^{n^n}\left(\hat{\theta}_{n,i}^* - \bar{\theta}_n^*\right)^2$, where $\bar{\theta}_n^* = \frac{1}{n^n}\sum_{i}^{n^n}\hat{\theta}_{n,i}^*$.

So, in order for the bootstrap estimate of variance to work, we must have

$$\widehat{\mathrm{Var}}_B[\hat{\theta}_n^*] \approx \mathrm{Var}[\hat{\theta}_n^*|F_n] \approx \mathrm{Var}[\hat{\theta}_n].$$

The first approximation is obvious - estimated bootstrap variance is a consistent and unbiased estimator of bootstrap variance. Furthermore, we can make it arbitrarily accurate by increasing the number of bootstrap samples m!

The second approximation is more difficult to show in general. We provide some intuition: Imagine that we could draw not just one F_n but an arbitrary number of samples of size n from our population. We could use the values of the functional on these samples to estimate the variance of the estimate of the functional $\mathrm{Var}[\hat{\theta}_n]$. With the bootstrap, we are doing exactly this, but pretending that F_n is F. And, as n grows large, F_n is a better and better approximation to F. A bootstrap sample F_n^* is to the ECDF F_n what the ECDF F_n is to the underlying population F.

15.6 Bootstrapping confidence intervals

Confidence intervals are among the most common ways of summarizing the uncertainty associated with an estimator, second only to the variance/standard deviation of an estimator. Before we proceed with describing three different procedures for constructing confidence intervals, we will first formalize the concept of a confidence interval.

Definition 15.6.1 (Confidence interval)**.** A $1-\alpha$ confidence interval for a parameter θ is an interval $C_n = [a, b]$, such that

$$P(\theta \in C_n) = 1 - \alpha,$$

for all possible values of θ. The bounds a and b are functions of the data.

It is beneficial to think about confidence interval not as an interval but as a procedure to construct an interval for the given data. If an experiment would be repeated many times and we used a $1-\alpha$ confidence interval procedure each time, the $1-\alpha$ of the constructed confidence intervals would contain the true value of the parameter. Hence, the confidence interval is random, not the parameter, and it is in general incorrect to say that the true value of the parameter lies in the confidence interval with $1-\alpha$ probability.

We will use the term *confidence level* to refer to $1-\alpha$ and *coverage probability* to refer to the proportion of cases when the true parameter is covered by the confidence interval. Ideally, the two would be the same for our procedure for constructing confidence intervals. However, in practice, it is difficult to guarantee this in general, so confidence intervals in practice are often too wide or too narrow.

Definition 15.6.2. The $(1-\alpha)$ *Bootstrap standard confidence interval* is based on the bootstrap estimate of variance:

$$C_n = [\hat{\theta}_n - z_{1-\frac{\alpha}{2}}\hat{\sigma}, \hat{\theta}_n + z_{1-\frac{\alpha}{2}}\hat{\sigma}],$$

where z_x is the z-score (quantile function of the standard normal distribution) at x and $\hat{\sigma} = \sqrt{\widehat{\text{Var}}_B[\hat{\theta}^*_n]}$.

Note that the definition assumes a symmetric two-sided confidence interval. That needn't be the case - we can use the same process to construct one-sided or asymmetric intervals.

The bootstrap standard confidence interval is very similar to the classic standard confidence interval, the only difference is that we employ bootstrap to estimate the standard deviation, instead of estimating $\hat{\sigma}_{\hat{\theta}}$ directly. Other than simplicity of computation, there is no other advantage of the bootstrap standard confidence interval.

Standard confidence intervals serve as a decent quick quantification of uncertainty, but the assumption of normality leads to incorrect coverage, especially with skewed distributions. Furthermore, bootstrap standard confidence intervals systematically underestimate the coverage probability.

Now we introduce another intuitive approach to constructing CI that typically gives better coverage than standard normal intervals:

Definition 15.6.3. Let $F_{\hat{\theta}^*_{n,m}}$ be the ECDF based on m bootstrap replications $\hat{\theta}^*_{n,1}, \ldots, \hat{\theta}^*_{n,m}$. The $(1-\alpha)$ symmetric *percentile confidence interval* is based on the quantiles of these replications:

$$C_n = [Q(\frac{\alpha}{2}), Q(1-\frac{\alpha}{2})].$$

We can compute percentile confidence intervals using Algorithm 20.3.1.

15.7 Practical considerations

The bootstrap is an extremely powerful technique but it does have certain limitations. Theoretically, the most important question is for which statistical functionals will the bootstrap variance (or quantiles of the distribution of bootstrapped values) tend to the true variance (or quantiles). Answering this question is not trivial and in most cases requires advanced tools from functional analysis.

One of the few more general cases where the bootstrap is guaranteed to work are linear statistical functionals, as long as the 2-nd moment of the functional is finite. That is, functionals of the form $T(F) = \int r(x)dF(x)$ with the corresponding plug-in estimator

$$T(F_n) = \int r(x)dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} r(X_i).$$

Note that most common functionals such as the mean, median, and variance are linear functionals.

One clear example where the bootstrap certainly fails are extreme order statistics, such as the maximum.

Even when the bootstrap is guaranteed to work, the guarantees are typically asymptotic. That is, as n (sample size) and m (number of bootstrap replications) tend to infinity. In practice, however, we have to deal with finite n and m, albeit m can be arbitrarily large.

Dealing with m is straightforward - in essence, we are dealing with a standard case of Monte Carlo approximation. As a rule of thumb, $m = 100$ should be enough for means and medians, while $m = 10000$ might be required for functionals such as extreme quantiles. Of course, this also depends on the approximation error that is still acceptable. However, we needn't rely on such recipes, because we can in most cases estimate the approximation error from the variability of the bootstrapped values. And we can make m arbitrarily large, subject to constraints on time or computational resources.

Unfortunately, in practice we rarely have a choice regarding how large n is. Again, for functionals like the median and mean, even n in the 10s would be enough for reasonable approximations of variance and even reasonable coverage of typical confidence intervals. For more extreme functionals, 100s of observations might be required.

Exercises

Exercise 15.1. Prove Proposition 15.3.1.

Exercise 15.2. Prove Proposition 15.3.2.

Exercise 15.3. Show that the sample mean is an unbiased, consistent and asymptotically normal estimator of the mean.

Exercise 15.4. Show that sample variance is a consistent but biased estimator of variance.

Exercise 15.5. Show that sample correlation is a consistent estimator of correlation.

Chapter 16

Maximum likelihood estimation

Now we return to parametric methods. Like with non-parametric inference, the goal is still to infer the properties of a data generating process based on a sample independent observations from that process:

$$Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n \sim_{\text{iid}} \boxed{\begin{array}{l}\text{unknown data}\\ \text{generating process}\end{array}}.$$

However, unlike with the plug-in estimators and bootstrap inference, parametric approaches have to explicitly hypothesize what the unknown data generating process might be.

16.1 Parametric models and the likelihood

Definition 16.1.1. A *parametric model* $\mathcal{H}$ for data y is a parametrized set of distributions

$$\mathcal{H} = \left\{p(y|\theta) : \theta \in \Theta = \mathbb{R}^k\right\},$$

where p is the joint PDF/PMF of the data, Θ is our *parameter space* and $\theta = (\theta_1, ..., \theta_k)$ is our *parameter*.

Definition 16.1.2. The *likelihood* function is defined as

$$L(\theta; y) \triangleq p(y|\theta).$$

The *log-likelihood* function is defined as $\ell(\theta; y) \triangleq \log L(\theta; y)$.

If iid observations are assumed, the likelihood factorizes to $L(\theta; y) = \prod_{i=1}^{n} p(y_i|\theta)$. However, keep in mind, that this is an assumption. In general, the observations need not be conditionally independent.

Note that the likelihood is not a density or probability mass function - instead, it is the density/probability of the data viewed as a function of the parameters, not the data. That is, the likelihood does not necessarily integrate to 1.

16.2 The maximum likelihood estimator

Definition 16.2.1. The *maximum likelihood estimator* (MLE) of the parameter θ is defined as

$$\hat{\theta}_n \triangleq \arg\max_{\theta} L(\theta; y).$$

Proposition 16.2.1 (Invariance)**.** *Let $\hat{\theta}$ be the MLE of θ. Let g be a function. Then, $g(\hat{\theta})$ is the MLE of $g(\theta)$.*

Proof. The proposition does not state that g is bijective, so we must consider the possibility that $g(\theta) = \nu$ maps distinct θ to the same ν. Let $g^{-1}(\nu) = \{\theta : g(\theta) = \nu\}$ be the preimage of all θ that are mapped to a particular ν. The domain of g is the parameter space, so the MLE $\hat{\theta}$ must be in $g^{-1}(\nu)$ for exactly one ν, which we will call $\hat{\nu}$. As $\hat{\theta}$ is the maximum of $L(\theta)$, $\hat{\nu}$ must be the maximum $L(\nu)$. And we know $\hat{\nu} = g(\hat{\theta})$. ■

The MLE has several other nice properties. The MLE is consistent and asymptotically normal, two properties of estimators that we have already defined. The MLE is also asymptotically efficient, which we will show in Section 16.3. These properties only hold under one or more conditions. We list them here and invoke them as required by a particular theorem. Note that θ_0 here means the true value of the parameter, while θ is used to represent a value of the parameter:

R1 $\theta \neq \theta_0 \Leftrightarrow p(\cdot|\theta) \neq p(\cdot|\theta_0)$ *(identifiability).*

R2 The support of $p(y|\theta)$ is the same for all θ. That is, the same values of y have non-zero $p(y|\theta)$ for all θ.

R3 The point θ_0 is an interior point of the parameter space Θ.

R4 $p(y|\theta)$ is differentiable in θ on Θ.

R5 $p(y|\theta)$ is three times differentiable in θ on Θ and for all $\theta \in \Theta$ there exist a constant c and a function $M(y)$ such that

$$\left|\frac{\partial^3}{\partial\theta^3} \log p(y|\theta)\right| \leq M(y),$$

with $\mathrm{E}_{\theta_0}[M(y)] < \infty$, for all $\theta_0 - c < \theta < \theta_0 + c$ and all y in the support of Y.

R6 The integral $\int p(x|\theta)dx$ can be differentiated twice under the integral sign as a function of θ.

These conditions are often omitted or summarized as *under certain regularity conditions*.

Theorem 16.2.1 (Consistency of MLE)**.** *Assume regularity conditions R1-R4. The MLE $\hat{\theta}_n$ is a consistent estimator of θ.*

The proof of this theorem is out of the scope of this text. We refer the interested reader to Hogg et al. (2005, ch. 6.).

Example 16.2.1 (Linear regression)**.** *We'll demonstrate maximum likelihood estimation on the most popular parametric model - linear regression. The model assumes a linear relationship between the expectation of the target (dependent) variable y and the input (independent) variable(s) x:*

$$y_i = \beta^T x_i + \epsilon_i,$$

where ϵ_i (the residuals) are assumed to be identically and independently distributed with mean 0. Additionally, we will assume that their distribution is normal: $\epsilon_i \sim_{iid} N(0, \sigma^2)$.

Let's derive the maximum likelihood estimator for the coefficients β. First, we write the likelihood explicitly:

$$L(\beta, \sigma^2; y, x) = \prod_{i=1}^{n} p(y_i|x_i, \beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_i - \beta^T x_i)^2}{2\sigma^2}).$$

The log-likelihood is then $\ell(\beta, \sigma^2; y, x) =$

$$\sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_i - \beta^T x_i)^2}{2\sigma^2})\right) = n \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2.$$

Observe that the likelihood is maximized (in terms of β) where $(y_i - \beta^T x_i)^2$ is minimized. That is, the MLE for β is obtained where the sum of squared residuals is minimized and it does not depend on σ^2.

So, for this model, maximizing the likelihood corresponds to minimizing the mean squared error! Is that always the case? No, far from it - it is a consequence of using a likelihood based on the normal distribution. If we instead assumed, for example, Laplace-distributed residuals, the MLE would correspond to minimizing the sum of absolute errors.

Finally, we derive, in matrix form, the MLE estimate (or the ordinary least-squares estimate, if you prefer). We want to minimize

$$(y - X\beta)^T(y - X\beta) = (y^T - \beta^T X^T)(y - X\beta) = y^T y - 2\beta^T X^T y + \beta^T X^T X\beta.$$

Taking the derivative with respect to β, we get $-2X^T y + 2X^T X\beta$, which equals 0 where $X^T y = X^T X\beta$. Assuming $X^T X$ is invertible, we get

$$\beta_{MLE} = (X^T X)^{-1} X^T y.$$

16.3 Asymptotic normality and efficiency of MLE

First, we'll introduce a quantity that plays an important role in several areas of statistics - Fisher information. Fisher information deals with the question of how much the data are expected to constrain the parameter values. That is, how much information are the data expected to bring.

If the data bring a lot of information, then we expect the likelihood to form a sharp peak - the data could not have been generated by many different parameter values. On the other hand, if the data bring little information, the likelihood will be more flat - a wider range of parameter values is likely to have generated the data.

The goal is to quantify this difference. One way of doing this is by observing the derivative of the log-likelihood $\frac{d}{d\theta} \log f(X|\theta)$, also known as the *score function*. Sharper peaks will have higher slopes and flatter peaks lower slopes. The absolute or squared value of the score function would therefore be a good indicator.

Of course, we do not know what the data will be, so we can only compute the expected value. This leads to the following definition of Fisher information:

Definition 16.3.1. *(Unit) Fisher information* is defined as

$$I(\theta) \triangleq \mathrm{E}\left[\left(\frac{d}{d\theta} \log f(X|\theta)\right)^2 \middle| \theta\right] = \int \left(\frac{d}{d\theta} \log f(x|\theta)\right)^2 f(x|\theta) dx.$$

In the case of not just one but multiple n iid observations, Fisher information is defined as $I_n(\theta) \triangleq nI(\theta)$.

Proposition 16.3.1. *An alternative view is that Fisher information is the variance of the score function*

$$I(\theta) = \mathrm{Var}\left[\frac{d}{d\theta} \log f(X|\theta) \middle| \theta\right]$$

.

The proof is based on showing that the expected value of the score function is 0 and is left as an exercise.

We can also write Fisher information as:

Proposition 16.3.2. *If* $\log f(X|\theta)$ *is twice differentiable, then*

$$I(\theta) = -\operatorname{E}\left[\frac{d^2}{d\theta^2}\log f(X|\theta)\middle|\theta\right].$$

The proof is left as an exercise.

The first definition shows that Fisher information can be expressed as the variance of the score function. The motivation behind such a definition would be that the score function is expected to vary more if there is a peak and less if the likelihood is flat.

The alternative definition stated as a proposition shows that Fisher information can be expressed as the (negative) expectation of the second derivative of the log-likelihood - the curvature of the log-likelihood. Higher curvature of course implies higher peaks.

Theorem 16.3.1 (Asymptotic normality of MLE)**.** *Assume regularity conditions R1-R6. If the Fisher information* $I(\theta)$ *is positive and finite, then for any consistent sequence of MLE* $\hat{\theta}_n$ *we have*

$$\hat{\theta}_n - \theta \xrightarrow{D} N\left(0, I_n(\theta)^{-1}\right).$$

The proof of this theorem is out of the scope of this text. We refer the interested reader to Hogg et al. (2005, ch. 6.1).

Corollary 16.3.1. *When Theorem 16.3.1 applies, we also have*

$$\hat{\theta}_n - \theta \xrightarrow{D} N\left(0, I_n(\hat{\theta}_n)^{-1}\right).$$

That is, plugging in a consistent estimator does not affect asymptotic normality. The proof of this corollary is left as an exercise. The corollary provides an estimate for the asymptotic standard error of the MLE: $\hat{\sigma}_{\text{MLE}} \approx \sqrt{I_n(\theta_n)^{-1}}$.

We can utilize the asymptotic normality of MLE to construct confidence intervals:

Definition 16.3.2. The $(1-\alpha)$ *MLE standard confidence interval* is

$$C_n = [\hat{\theta}_n - z_{1-\frac{\alpha}{2}}\widehat{\sigma}_{\text{MLE}}, \hat{\theta}_n + z_{1-\frac{\alpha}{2}}\widehat{\sigma}_{\text{MLE}}],$$

where z_x is the z-score.

Example 16.3.1. *Let's revisit the Bernoulli example from Chapter 14. We already know that the MLE estimate of the proportion is the ratio of 1s ($\hat{\theta}_n = \frac{\sum y_i}{n}$) and the likelihood is Binomial*

$$L(\theta; y) = \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}.$$

To get to the confidence intervals we first derive the Fisher information, which is the (minus) expected value of the second derivative of the log-likelihood. For brevity, we write $k = \sum y_i$:

$$\frac{d}{d\theta}\ell(\theta; y) = \frac{d}{d\theta}\left(k\log\theta + (n-k)\log(1-\theta)\right) = \frac{k}{\theta} - \frac{n-k}{1-\theta}.$$

Differentiating for the second time:

$$\frac{d}{d\theta}\left(\frac{k}{\theta} - \frac{n-k}{1-\theta}\right) = -\frac{k}{\theta^2} - \frac{n-k}{(1-\theta)^2}.$$

Now we take the expectation over k, taking into account that $E[k|\theta] = n\theta$ (the expected value of a Binomial or a sum of Bernoulli):

$$E\left[-\frac{k}{\theta^2} - \frac{n-k}{(1-\theta)^2}\middle|\theta\right] = -\frac{n}{\theta} - \frac{n}{1-\theta} = -\frac{n}{\theta(1-\theta)}.$$

So, the Fisher information is

$$I_n(\theta) = \frac{n}{\theta(1-\theta)}$$

and unit Fisher information is

$$I(\theta) = \frac{1}{\theta(1-\theta)}.$$

This leads to the confidence interval

$$C_n = \left[\hat{\theta}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\theta(1-\theta)}{n}}, \hat{\theta}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\theta(1-\theta)}{n}}\right].$$

By plugging in $\hat{\theta}_n$, we arrive to the same interval as we did with the normal approximation argument in Chapter 14. While that works in some cases, Fisher information is a more general approach to constructing CIs for MLE estimators.

Asymptotic efficiency of MLE

Definition 16.3.3. Consider two estimators θ_a and θ_b, such that

$\sqrt{n}(\theta_a - \theta) \xrightarrow{D} N\left(0, \sigma_a^2\right)$ and $\sqrt{n}(\theta_b - \theta) \xrightarrow{D} N\left(0, \sigma_b^2\right)$.

We define asymptotic relative efficiency as $ARE(\theta_a, \theta_b) \triangleq \frac{\sigma_a^2}{\sigma_b^2}$

Theorem 16.3.2. *If θ_n is the MLE and $\hat{\theta}_n$ is an other estimator, then $ARE(\theta_n, \hat{\theta}_n) \leq 1$.*

The theorem states that out of all in some sense well-behaved estimators, MLE has, asymptotically, the smallest variance. This result is a combination of the asymptotic normality result from above and the Cramer-Rao lower bound theorem, which states that the inverse of the Fisher information is the lower bound on the variance of any unbiased estimator.

Multi-parameter case

The normality and efficiency arguments also extend to multiple parameters. Fisher information generalizes to the Fisher information matrix.

Definition 16.3.4. The *Fisher information matrix* is defined (component-wise) as

$$[I(\theta)]_{ij} \triangleq \mathrm{E}\left[\left(\frac{\partial}{\partial\theta_i}\log f(X|\theta)\right)\left(\frac{\partial}{\partial\theta_j}\log f(X|\theta)\right)\middle|\theta\right].$$

In the case of not just one but multiple n iid observations, Fisher information is defined as $I_n(\theta) \triangleq nI(\theta)$.

Analogous to the univariate case, we can also compute the Fisher information matrix

Proposition 16.3.3. *If the corresponding differentiation can be made, then*

$$[I(\theta)]_{ij} = -\mathrm{E}\left[\frac{\partial}{\partial\theta_i\theta_j}\log f(X|\theta)\middle|\theta\right].$$

The proof is left as an exercise.

Proposition 16.3.4. *The Fisher information matrix is symmetric and positive semi-definite.*

The proof is left as an exercise.

Theorem 16.3.1 and Corollary 16.3.1 also hold for the multivariate case (we state this without proof) with the inverse of the Fisher information matrix as the covariance matrix.

This can be utilized to construct confidence regions (a generalization of CIs), however, these are rarely used in practice. Instead, we still only focus on individual parameters. For the $i-th$ parameter we have

$$\hat{\theta}_{ni} - \theta_{0i} \xrightarrow{D} N\left(0, \left[I_n(\theta)^{-1}\right]_{ii}\right).$$

Note that this requires us to compute the inverse of the Fisher information matrix and take the $i-$th diagonal element. This is not the same as computing the univariate Fisher information for the $i-$th parameter and taking the reciprocal value (although it is in some cases).

Example 16.3.2 (The univariate normal distribution)**.** *The univariate normal distribution has 2 parameters, μ and σ^2. The likelihood and log-likelihood of a normal distribution model for n iid observations are*

$$L(\mu, \sigma^2; y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_i-\mu)^2}{2\sigma^2}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp(-\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2})$$

and

$$\ell(\mu, \sigma^2; y) = -\frac{n}{2}(\log 2\pi + \log \sigma^2) - \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}.$$

To get the Fisher information matrix, we first compute the 2nd order partial derivatives with respect to the parameters

$\frac{\partial}{\partial\mu}\ell(\mu, \sigma^2; y) = \frac{\sum_{i=1}^{n}(y_i-\mu)}{\sigma^2}$,

$\frac{\partial}{\partial\sigma^2}\ell(\mu, \sigma^2; y) = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^4}$,

$\frac{\partial^2}{\partial\mu^2}\ell(\mu, \sigma^2; y) = -\frac{n}{\sigma^2}$,

$\frac{\partial^2}{\partial\mu\sigma^2}\ell(\mu, \sigma^2; y) = -\frac{\sum_{i=1}^{n}(y_i-\mu)}{\sigma^4}$,

$\frac{\partial^2}{\partial(\sigma^2)^2}\ell(\mu, \sigma^2; y) = \frac{n}{2\sigma^4} - \frac{2\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^6}$,

Note that we don't have to compute $\frac{\partial^2}{\partial\sigma^2\mu}\ell(\mu, \sigma^2; y)$ - we can use the symmetry of the Fisher information matrix. Finally, we compute the expected values (with respect to y) of the partial derivatives

$\mathrm{E}[\frac{\partial^2}{\partial \mu^2}\ell(\mu, \sigma^2; y)] = \mathrm{E}[-\frac{n}{\sigma^2}] = -\frac{n}{\sigma^2}$,

$\mathrm{E}[\frac{\partial^2}{\partial \mu \sigma^2}\ell(\mu, \sigma^2; y)] = \mathrm{E}[-\frac{\sum_{i=1}^n (y_i - \mu)}{2\sigma^4}] = 0$,

$\mathrm{E}[\frac{\partial^2}{\partial (\sigma^2)^2}\ell(\mu, \sigma^2; y)] = \mathrm{E}[\frac{n}{2\sigma^4} - \frac{2\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^6}] = -\frac{n}{2\sigma^4}$,

so the (unit) Fisher information matrix is

$$I(\mu, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}.$$

So, asymptotically, the error (variance) of the MLE estimator for μ is $\frac{\sigma^2}{n}$ and for σ^2 it is $\frac{2\sigma^4}{n}$. The former is already familiar to us, but it is worth noting that the error for the mean does not depend on the value of the mean, just on the variance (this is not true for all distributions). Furthermore, the errors for the mean and the variance are not correlated.

Exercises

Exercise 16.1. Prove Proposition 16.3.1.

Exercise 16.2. Prove Proposition 16.3.2.

Exercise 16.3. Prove Proposition 16.3.3.

Exercise 16.4. Prove Corollary 16.3.1.

Exercise 16.5. Prove Proposition 16.3.4.

Chapter 17

Null-hypothesis significance testing

17.1 General framework

Hypothesis testing is a family of statistical inference methods that focus on testing the truth (or falsehood) of a well-defined hypothesis. In this chapter we will focus on the most popular framework of hypothesis testing - null-hypothesis significance testing (NHST).

Formally, suppose we have a parametric model parametrized with θ (may be a vector) and we want to test a well-defined hypothesis about where θ might lie against an alternative hypothesis. That is, we partition the parameter space into two disjoint sets $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$.

H_0 is called the *null-hypothesis* and H_1 is called the *alternative hypothesis*. If the null-hypothesis completely specifies the distribution (for example, $\Theta_0 = \{\theta_0\}$), we call it a *simple hypothesis*, otherwise it is a *composite hypothesis*.

We proceed by defining a random variable X and a *rejection region* R, which is a subset of the values of X. If $X \in R$, we reject the null-hypothesis, otherwise we do not reject it (we retain it). This is a very general formulation of the process. In most cases the random variable is a *test statistics* (a function of the data that describes how extreme the sample is if the null-hypothesis were true) and the rejection region is defined by a threshold - a *critical value*. If the sample is extreme beyond the critical value, we reject the null-hypothesis.

The hypothesis testing process has 2 possible decisions, reject the null-hypothesis or retain the null-hypothesis, and therefore 4 possible outcomes. The two desirable ones are that we reject a false null-hypothesis and retain a true null-hypothesis. The two errors are rejecting a true null-hypothesis (*Type I error*)

or retaining a false null-hypothesis (*Type II error*).

Of course, the goal is to define a hypothesis test, such that the probability of Type I or Type II error is minimal. These probabilities play a central role in hypothesis testing and we proceed with a more formal definition.

Definition 17.1.1. The *power function* of a test with rejection region R is defined as $\beta(\theta) \triangleq P(X \in R|\theta)$.

This is an abstract definition, but for any well-defined test this will be a function that maps parameter values to the probability of rejecting the null-hypothesis if that parameter value is the true parameter value.

Now we can define the significance level of a test:

Definition 17.1.2. A test has *significance level* α if its size is less or equal to $\alpha \geq \sup_{\theta \in \Theta_0} \beta(\theta)$.

The quantity $\sup_{\theta \in \Theta_0} \beta(\theta)$ is called the *size* of a test and represents the largest probability of rejecting the null-hypothesis that is true.

In other words, a test having significance level α means that the probability of Type I error is at most α.

In practice we typically determine the significance level that serves our purpose. Ideally, we would then like to use a test that has the lowest Type II error (highest power under the alternative-hypothesis) among all tests at that significance level. In some cases, such *most powerful tests* are known, for example, the likelihood-ratio test for simple hypotheses that we cover later in this chapter. However, in most cases they are not or they do not exist, so we use one of the widely used tests. Note that for a given test and data sample there is always a trade-off - lower significance level reduces probability of Type I error but increases the probability of Type II error (and vice-versa). The only way to reduce both is to gather more data.

Example 17.1.1 (Z-test)**.** *Suppose we have a sample $X_1, ..., X_n$ from a normal population with unknown mean μ and known variance σ^2. And suppose we want to test if the mean is at most a particular value $H_0 : \mu \leq \mu_0$ against the alternative that it is greater $H_1 : \mu > \mu_0$.*

A popular test statistic in such cases is

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}},$$

where $\overline{X}$ is the sample average. Standardized deviation from the hypothesized mean looks like a sensible test statistic - if the actual mean equals the true mean the test statistic of the sample should be close to 0. The further it deviates, the

less likely it is that the sample had been generated from a distribution with the hypothesized mean.

We will reject the null-hypothesis if $Z > z_\alpha$ (the standard score at level α), so the power function is $\beta(\mu) = P(Z > z_\alpha)$. Under the null-hypothesis Z has a standard normal distribution, but if μ is the true mean the distribution of Z shifts by $\mu - \mu_0$ (standardized), so:

$$\beta(\mu) = P(Y + \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha) = P(Y > z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}) = 1 - \Phi(z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}),$$

where $Y \sim N(0, 1)$ and Φ is the CDF of the standard normal.

Finally, we compute the size of the test. Observe that $\beta(\mu_0) = \alpha$ and that $\beta(\mu)$ is an increasing function. Therefore, for all $\mu \leq \mu_0$ we have $\beta(\mu) \leq \alpha$. So, the test has significance level α. Note that this does not mean that the probability of rejecting a true null-hypothesis is α! Because we have a composite hypothesis we can only claim that the probability of rejecting a true null-hypothesis is at most α (it could be less).

Rejecting the null-hypothesis is often referred to as finding a *statistically significant* result.

Statistical significance should not be confused with practical importance. The former talks about something being true with some level of certainty while the latter is concerned with the size of the effect and depends on the given context. For example, if we gathered a random sample of babies, we would be able to determine that there are statistically significantly more boys than girls (about 1.05 boys are born for every girl). This result is also an interesting fact about humans and therefore of practical significance to science. However, it is probably of no practical significance to manufacturers of newborn baby greeting cards - they would probably still make the same amount of *it's a girl* and *it's a boy* cards.

17.2 TODO: The Wald test

17.3 TODO: Testing with confidence intervals

17.4 TODO: The likelihood ratio test

17.5 TODO: Testing multiple hypotheses

Chapter 18

Bayesian inference

18.1 The Bayesian perspective

The differences between Bayesian statistics and the classical approaches to inference are rooted in a fundamental difference in how probability is viewed. In classical approaches we view probability as the property of random experiments. In Bayesian statistics, however, we view probability as a tool for expressing uncertainty.

Bayesian statisticians share the statistical modelling approach and use the same models (that is, the same likelihoods $p(y|\theta)$) but the above difference leads to a fundamentally different treatment of a model's parameters θ.

This difference can be illustrated by trying to answer this arguably very natural question: *Given this model $p(y|\theta)$ and some data y, what is the probability that $\theta > 0.5$?*

In maximum likelihood inference, for example, we treated θ as an unknown constant, the data on the other hand we treated as random variables, realizations from the idealized random experiment that we are using to interpret the process we are studying. Because θ is treated as a constant the above question is not even allowed in the maximum likelihood inference framework (NHST is the same)! MLE inference then proceeds in a different way - finding the parameter value that maximizes the probability of the data. Furthermore, uncertainty is quantified through the parameter estimator, which is a random variable, because it is a function of the data, which are random.

In Bayesian statistics, on the other hand, we treat the parameter as a random variable. Not necessarily because we would indeed think it is random, but because we don't know what its value is and *we choose to represent our uncertainty* with a random variable. The question $p(\theta > 0.5|y)$ now becomes a legitimate question! What remains, of course, is to provide a means for com-

puting the answer. For that we turn to the cornerstone of Bayesian statistics, Bayes' theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \propto p(y|\theta)p(\theta).$$

The theorem is already familiar to us as is one of the terms - the likelihood $p(y|\theta)$. The likelihood is determined by our choice of model.

The distribution of the parameter after observing the data $p(\theta|y)$ is referred to as the posterior distribution (density) or just the *posterior*. The posterior completely describes the uncertainty associated with the parameters after seeing the data and can be used to answer any probabilistic question regarding the parameters.

To compute the posterior we require two more terms. The term $p(y)$ only serves the purpose of normalizing the posterior and is, as illustrated above, an integral and difficult to compute. Because is is not a function of θ it does not affect the shape of the posterior and we can use that fact to avoid ever having to compute it.

Finally, $p(\theta)$ is known as the prior distribution (density) or just the *prior*. It is the distribution of the parameter before we see the data - it represents our prior uncertainty about where the parameter is. It is the quintessentially Bayesian concept and root of all the advantages and issues with Bayesian statistics. It's makes sense that if we are to be uncertain after seeing the data, we must be uncertain before seeing the data and that uncertainty has to be quantified in order to compute the posterior uncertainty.

Definition 18.1.1 (Conjugate prior)**.** If for some likelihood $p(y|\theta)$ the posterior distribution $p(\theta|y)$ and the prior distribution $p(\theta)$ are in the same family, we say that the prior is a *conjugate prior* for the likelihood.

Conjugate priors simplify computation and allow us to incrementally learn our models, using the posterior from one iteration as the prior for the next iteration of learning. Historically, conjugate priors were very important, because Bayesian computation would otherwise be infeasible. However, many models that are commonly used in practice do not have conjugate priors (in fact most), so we have to rely on numerical methods to do Bayesian inference (see Chapter 20[1])

Example 18.1.1 (Conjugate prior for the Bernoulli)**.** *We've already seen in Chapter 14 that using a $U(0,1)$ prior for the parameter of the Bernoulli model results in a Beta posterior. If we combine that with the fact that $U(0,1)$ is a special case of beta - $Beta(1,1)$ - we should consider the possibility that Beta might be conjugate for this likelihood.*

[1]The random number generation chapter is excluded for now.

Let's compute the posterior for the likelihood $y_i|\theta \sim_{iid} Bernoulli(\theta)$ with the prior $\theta \sim Beta(\alpha, \beta)$:

Now we have the prior and the likelihood and we can compute the posterior ($k = \sum y_i$, for brevity):

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ &\propto p(y|\theta)p(\theta) && \textit{(proportional to)} \\ &= \left(\binom{n}{k}\theta^k(1-\theta)^{n-k}\right)\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} && \textit{(insert densities)} \\ &\propto \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} && \textit{(remove constants)} \end{aligned}$$

We can see that the posterior is shaped like a Beta density and we can recognize what the parameters are:

$$\theta|y \sim Beta(\sum y_i + \alpha, n - \sum y_i + \beta).$$

Therefore, the Beta distribution is a conjugate prior for the Bernoulli likelihood! We can check that for $\alpha = 1$ and $\beta = 1$ we get the result for the $U(0,1)$ prior. The prior also has a straightforward interpretation - α and β represent the counts of 1s and 0s that our prior opinion is based on. And inference for this Bernoulli-Beta model reduces to adding the newly observed 1s and 0s to the prior counts.

In practice it might sometimes be infeasible or unnecessary to compute the posterior. In such cases we might compute only the peak of the posterior, which is also known as the MAP estimator:

Definition 18.1.2 (MAP estimator)**.** The *maximum a-posteriori* (MAP) estimator of the parameter θ is defined as

$$\hat{\theta}_n \triangleq \arg\max_\theta p(\theta|y).$$

Analogous to the MLE being the maximum of the likelihood the MAP is the maximum of the posterior distribution. Note that the two can, under certain conditions, be the same (left as an exercise). MAP estimation is also the first step of trying to estimate the Bayesian posterior with a normal distribution.

In large samples there is a strong relationship between frequentist and Bayesian inference:

Theorem 18.1.1 (Bernstein-von Mises (informal))**.** *Under certain regularity*

conditions the Bayesian posterior in large samples is approximately normal with mean approximately θ_{MLE} and covariance matrix approximately $I_n(\theta)^{-1}$.

In particular, this says that frequentist confidence intervals and Bayesian posterior intervals will be approximately the same.

Exercises

Exercise 18.1. Derive the condition under which the MLE and MAP estimator will be the same, assuming that both estimators exist.

Part III

Computational methods

Chapter 19

Monte Carlo method

19.1 Monte Carlo integration

Computing integrals of the form

$$I = \int_\Omega f(x)dx,$$

where f is a real-valued function on $\Omega \subseteq \mathbb{R}^k$ is a common computational problem in many areas, including statistics. In particular Bayesian statistics, where we rely heavily on being able to integrate the posterior distribution of parameters of our statistical model.

The main idea of Monte Carlo integration is to approximate the above integral using random sampling:

$$I = \int_\Omega f(x)dx = \int_\Omega \frac{f(x)}{p(x)}p(x)dx = \mathrm{E}[\frac{f(x)}{p(x)}],$$

where $p(x)$ is a PDF such that $p(x) > 0$ whenever $f(x) \neq 0$. That is, we have introduced a random variable X on Ω, such that the integral can be expressed as an expectation of the ratio $\frac{f(x)}{p(x)}$ over that random variable.

By the law of large numbers we have

$$x_n = \frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)}{p(X_i)} \xrightarrow{a.s.} \mathrm{E}[\frac{f(x)}{p(x)}] = I,$$

which leads to the conclusion that we can estimate I by generating independent samples from X. Additionally, if $\mathrm{Var}[\frac{f(x)}{p(x)}] < \infty$, then we know, by the CLT,

that, for n of reasonable size, the error of our Monte Carlo approximation will be approximately

$$\sigma_{MC} \approx \frac{\sqrt{\text{Var}[\frac{f(x)}{p(x)}]}}{\sqrt{n}} = \frac{SD[\frac{f(x)}{p(x)}]}{\sqrt{n}}.$$

Monte Carlo approximation error therefore decreases with the root of the number of samples we draw. While we can make it arbitrarily close to 0 by drawing enough samples, this rate of convergence $O(n^{-\frac{1}{2}})$ is very poor compared to even the most simple quadrature methods, which converge $O(n^{-3})$.

However, the main advantage of Monte Carlo integration is that this rate of convergence is independent of the dimension of the integral! None of the steps we took rely on f being a single variable function or X a univariate random variable. So, unlike quadrature methods, Monte Carlo integration scales to high-dimensional integration. In fact, when dealing with high-dimensional integration, Monte Carlo methods are in most cases the only option. Other advantages of Monte Carlo integration are its simplicity and wide applicability.

If f is defined on a bounded subset, for example, without loss of generality, $\Omega = [0, 1]^k$, we can always use the uniform distribution on Ω (the uniform PDF/PMF is positive everywhere and therefore satisfies the condition that $p(x) > 0$ whenever $f(x) \neq 0$). Then Monte Carlo integration simplifies to

$$I = \int_{[0,1]^k} f(x)dx = E[f(x)].$$

Note that Monte Carlo integration can also be used when Ω is countable or countable in some dimensions and uncountable in others, as long as we can define a suitable random variable.

19.2 Generating random numbers

Monte Carlo methods are an excellent tool but to apply them, we have to be able to sample from the desired distribution. Random number generation is a very broad and rich field. In this chapter we will only review the most basic approaches and refer the interested reader to other sources. Chapter 20, however, is dedicated to Markov Chain Monte Carlo, a family of Monte Carlo methods that are indispensable to modern Bayesian statistics and machine learning.

The key component of random number generators (RNGs) is the uniform RNG. The uniform RNGs that we find in modern programming languages and software are linear congruential RNGs. They generate deterministic (pseudo-random) sequences of numbers and are often referred to as pseudo-random to emphasize

this fact. For most practical tasks, however, sequences of numbers generated by pseudo-RNGs are statistically indistinguishable from true random sequences and their period is long enough so it is practically impossible to consume the entire sequence. Pseudo-RNGs also require a seed - a starting point in the deterministic sequence. Manually setting the seed aids in the repeatability and reproducibility of analyses and algorithms with random components.

Random variates from other distributions are then generated using a uniform RNG and applying transformation, rejection or weighting. In the remainder of the chapter we describe representatives of these approaches.

Inverse transformation method

The inverse transformation method is a very simple and effective approach to generating random variates from a target distribution for which we can evaluate the quantile function (inverse of the CDF).

Proposition 19.2.1 (Inverse transformation)**.** *Let* $U \sim \text{Unif}(0, 1)$ *and let* F *be a CDF. Then,* $Q(U)$ *has the CDF* F*.*

Proof. Let $X = Q(U)$. $F_X(x) = P(X \leq x) = P(Q(U) \leq x) = P(U \leq F(x)) = F(x)$. ■

This leads to the following algorithm.

Algorithm 19.2.1. *Let* Q *be the generalized inverse of our target distribution. Then, the following algorithm returns* m *independent samples from our target distribution:*

procedure INVERSE-SAMPLING(Q, m)
  for $i \leftarrow 1 : m$ ***do*** ▷ *for each sample*
    sample u *from* $\text{Unif}(0, 1)$
    $x_i \leftarrow Q(u)$
  end for
  return x
end procedure

Rejection sampling

Sometimes no closed-form transformation exists that would transform samples from the proposal distribution that we can easily sample from to the target distribution that we are interested in. One common approach in such situations is to sample from the proposal distribution and then reject samples that are less probable under the target distribution. Before we introduce the basic acceptance-rejection sampling algorithm, we motivate it with an example.

TODO !!!!

The example above illustrates a very important idea that sampling from a distribution is equivalent to sampling uniformly from the area or volume under the PMF/PDF of that distribution. If the area is complicated, we can instead envelop it with an area that is easier to sample from uniformly and reject samples that fall outside the area of the target distribution. While generating random samples from a distribution requires a complete understanding of its PMF/PDF, checking if a sample falls in the area typically only requires us to evaluate the PMF/PDF at a point. This is the main idea of rejection sampling methods.

Algorithm 19.2.2. *Let g be our proposal density. Let f be a function such that $f(x) = Cp(x)$ for some $C > 0$ and all x.*[1] *And let M be a positive constant such that $f(x) \leq Mg(x)$, for all x. Then, the following algorithm returns m independent samples from density p:*

```
1:  procedure REJECTION-SAMPLING(f, g, M, m)
2:      for i ← 1 : m do                          ▷ for each sample
3:          repeat                                ▷ repeat until accepted
4:              sample y from g
5:              sample u from Unif(0, 1)
6:          until u ≤ f(y) / (M g(y))
7:          x_i ← y
8:      end for
9:      return x
10: end procedure
```

Proof that rejection sampling works. The samples produced by the algorithm are independent and identically distributed. Let h be their density. We have

$h(y) = P(Y = y \cap \text{accept } Y) = g(y)P(\text{accept } Y|Y = y) = g(y)\frac{f(y)}{Mg(y)} \propto f(y) \propto p(y)$.

Furthermore, we can show the unconditional acceptance probability.

$$P(\text{accept } Y) = P\left(U \leq \frac{f(Y)}{Mg(Y)}\right) = \int P\left(U \leq \frac{f(Y)}{Mg(Y)}|Y = y\right) g(y)dy$$
$$= \int \frac{f(y)}{Mg(y)} g(y)dy = \frac{C}{M}.$$

Therefore, the number of samples required to accept one sample follows a Geometric distribution with $p = \frac{C}{M}$ and mean $\frac{M}{C}$. This shows that the efficiency of rejection sampling depends on how tightly $Mg(y)$ envelops $f(y)$. In the ideal case of $Mg(y) = f(y)$, we have $P(\text{accept } Y) = 1$, so we require just one sample to generate a sample from f (and therefore p). ■

[1] Note that we may have $f = p$ as a special case. However, with this more general formulation we can show that it it sufficient to know p only up to a normalization constant. This is very convenient when the normalization constant is difficult to compute, as is the case with most Bayesian posteriors.

The efficiency of the above rejection sampling algorithm is proportional to how well the envelope fits the target density. In higher dimensions it becomes difficult to find a tight-fitting envelope, so this algorithm is not suitable for multivariate distributions.

Importance sampling

Suppose we want to approximate the following integral via Monte Carlo integration

$$I = \int_\Omega f(x)p(x)dx = E_p[f(x)],$$

however, we are unable to efficiently sample from $p(x)$.

Rejection methods compensate for drawing samples from the proposal distribution instead of the target distribution by rejecting some of the samples. Weighting methods achieve the same by weighting the samples from the proposal distribution.

Importance sampling, the main representative of this idea, generalizes Monte Carlo integration by introducing a proposal distribution g which we do know how to efficiently sample from. Then,

$$I = \int_\Omega f(x)p(x)dx = \int_\Omega \frac{f(x)p(x)}{g(x)}g(x)dx = E_g[\frac{f(x)p(x)}{g(x)}].$$

The importance sampling estimator of I is then

$$x_n = \frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)p(X_i)}{g(X_i)}.$$

How do we choose g? Of course, g needs to be easy to sample from and $g(x) > 0$ whenever $f(x)p(x) \neq 0$ (from Monte Carlo integration). The following proposition sheds some light on what the shape of g should be.

Proposition 19.2.2. *The proposal distribution g that minimizes the variance of the importance sampling estimator $Var[x_n]$ is*

$$g^*(x) = \frac{|f(x)|p(x)}{\int |f(t)|p(t)dt}.$$

Proof. The variance of the estimator is

$$\mathrm{Var}[\frac{f(X)p(X)}{g(X)}] = \mathrm{E}_g[(\frac{f(X)p(X)}{g(X)})^2] - (\mathrm{E}_g[\frac{f(X)p(X)}{g(X)}])^2.$$

The second term in the expression above is the integral we are approximating squared and is therefore independent of choice of g. To minimize the variance, we need to chose a g that minimizes the first term. Using Jensen's inequality, we have

$$\mathrm{E}_g\left[(\frac{f(X)p(X)}{g(X)})^2\right] \geq \mathrm{E}_g\left[\frac{|f(X)|p(X)}{g(X)}\right]^2 = \left(\int |f(x)|p(x)dx\right)^2,$$

which gives us a lower bound on the variance that is independent of choice of g. If we plug the supposed optimal $g*$ into the above, we can see that it attains the lower bound. That is, no other choice of g can have lower variance, so $g*$ is indeed optimal.

■

This proposition is not directly useful. If we could sample from this optimal density, then we could probably sample from p as well. However, it does suggest that the shape of g should closely match the shape of $|f(x)|p(x)$.

Another thing we need to be careful is that some choices of g can lead to the estimator having infinite variance. The second moment is (see proof above) $\int \frac{f(x)^2p(x)^2}{g(x)}dx$. If f has thinner tails than g, this might be infinite. Typically, we want to choose g with thicker tails than p.

Finally, this corollary reveals an important fact about the efficiency of importance sampling:

Corollary 19.2.1. *Let g be the optimal proposal distribution. Then,*

$$\mathrm{Var}[X_n] \leq \mathrm{Var}_f[\frac{1}{n}\sum_{i=1}^{n} f(X_i)].$$

Proof.

$$\mathrm{Var}_f[f(X)] = E_f[f(X)^2] - I^2 \geq E_f[|f(X)|]^2 - I^2 = \left(\int |f(x)|p(x)dx\right)^2 - I^2$$

$$= \mathrm{Var}[\frac{f(x)p(x)}{g_{\mathrm{optim}}(x)}].$$

■

That is, importance sampling can be more efficient than sampling from the target distribution. Why? Areas that contribute the most to the integral are areas which are both probable and where f is large absolutely. The optimal proposal distribution puts more emphasis on those areas.

Chapter 20

Markov Chain Monte Carlo

20.1 Metropolis-Hastings

We will focus on the continuous state space $S = \mathbb{R}^n$ and target density $p(x)$ which we want to sample from. The main idea of the Metropolis-Hastings algorithm is to start with some Markov chain with transition function $k(x, y)$ and then modify it so that it will satisfy detailed balance with p thus making p the stationary distribution of the modified Markov chain.

We will assume $k(x, y)$ corresponds to an aperiodic and irreducible Markov chain on S, but no more. If p is not already the stationary distribution of the Markov chain defined by k, then there must be a pair of states x and y where detailed balance is not satisfied. Without loss of generality, let's assume that for those x and y we have

$$p(x)k(x, y) > p(y)k(y, x).$$

We want these to be in balance. The target density $p(x)$ can not be changed, because the goal is to sample from that density. What remains is to modify the transition function. There are two ways of looking at this - either we have too many transitions from x to y or too few transitions from y to y for the two sides to be equal.

It is much easier to reject transitions than to add transitions in a smart way, so we opt for the former. That is, we will reject transitions from x to y with probability $\alpha(x, y)$ such that

$$p(x)k(x, y)\alpha(x, y) = p(y)k(y, x).$$

We have not defined α, but both sides are non-negative, so there definitely exists a factor between 0 and 1, which balances the two sides:

$$\alpha(x, y) = \frac{p(y)k(y, x)}{p(x)k(x, y)}.$$

So, every time we will propose a transition from x to y, we will only accept the transition with probability $\alpha(x, y)$, which can be computed from the original transition function k and target density p. But to be completely general, we also have to consider the case where the two states are imbalanced so that we have too few transitions from x to y. In such cases we have $p(y)k(y, x) > p(x)k(x, y)$ and $\alpha(x, y) > 1$. This leads to the final form of the Metropolis-Hastings correction:

$$\alpha(x, y) = \min\left\{1, \frac{p(y)k(y, x)}{p(x)k(x, y)}\right\}.$$

If k is symmetric, this reduces to

$$\alpha(x, y) = \min\left\{1, \frac{p(y)}{p(x)}\right\},$$

which is the original Metropolis correction.

Observe that in order to compute the Metropolis-Hastings correction, we need only to evaluate the ratio $\frac{p(y)}{p(x)}$. That is, it is sufficient if we can evaluate p only up to a multiplicative constant.

Algorithm 20.1.1. *Let $f(x) \propto p(x)$ a function that is proportional to our target density (trivially, it can be the actual target density), k a transition function, $x_0 \in S$ a starting state, and m the number of samples that we want to draw. The following algorithm returns m (possibly dependent) samples from p:*

1: ***procedure*** METROPOLIS-HASTINGS*(f, k, m, x_0)*
2: ***for*** $i \leftarrow 1 : m$ ***do*** ▷ *number of samples*
3: *sample candidate state* $x^* \sim k(x_{i-1}, x^*)$
4: $\alpha \leftarrow \min\left\{1, \frac{p(x^*)k(x^*, x_{i-1})}{p(x_{i-1})k(x_{i-1}, x^*)}\right\}$
5: *sample* $u \sim U(0, 1)$
6: ***if*** $u \leq \alpha$ ***then***
7: $x_i \leftarrow x^*$ ▷ *accept transition*
8: ***else***
9: $x_i \leftarrow x_{i-1}$ ▷ *transitions to self do not spoil detailed balance*
10: ***end if***
11: ***end for***
12: ***return*** $x_1, ..., x_m$.
13: ***end procedure***

It is clear that the Metropolis-Hastings algorithm also results in a Markov-Chain - the distribution of the next state depends only on the current state.

The choice of the transition function k (also known as the proposal or candidate distribution) is not easy. At a minimum, it must be such that the Markov chain is aperiodic and irreducible. The former is trivial, as we are guaranteed aperiodicity if we have a non-zero probability of transitioning to the same state. Irreducibility is typically also not a practical problem with continuous state spaces.

In general, we would like our k to be such that the autocovariance of the chain is as low as possible, resulting in an efficient sampler. For example, $k(\cdot, x) \approx p(x)$. However, in practice, we do not understand p well enough - if we did, we would sample from it directly. Instead, we use some local proposal distribution, such as a normal distribution centered on the current state to propose the next state. This leads to a trade-off. If the proposal distribution is too broad, we will propose states far away which are likely to have low $p(x)$ and will be rejected most of the time, resulting in a very autocorrelated chain. On the other hand, if the proposal distribution is too narrow, we will propose states that are close (have similar $p(x)$) and will be accepted with high probability, however, the moves will be very short, again resulting in a very autocorrelated chain. The main challenge is to construct a MCMC algorithm that proposes states that are far away but still likely to be accepted - this is addressed by some of the more advanced MCMC algorithms, such as Hamiltonian Monte Carlo (HMC), which is the basis for modern inference software.

20.2 Practicalities of MCMC

Estimating the variance of MCMC estimates

Naively estimating lag-k autocovariances with empirical covariances is computationally intensive and will not lead to a consistent estimator - a re-weighting is required. For further details and a discussion of the most common approaches to estimating variance see (Geyer, 1992, Section 3).

How many MCMC samples to take?

A longer chain is always preferred to a shorter chain for several reasons. First, if we have chosen a poor starting value (a value that is not really typical in our target distribution; that is, a value that has low probability/density), a longer chain is more likely to move into the typical set and deflate the impact of the atypical starting values. Second, a longer chain is more likely to reveal problematic behavior (see Section 20.2). And third, every additional sample reduces our MCMC approximation error.

We can gather from the above that the chain should always we as long as our time and resource constraints permit. Or at least as long as necessary to reduce

the MCMC approximation error to less than the precision at which we want to interpret the quantities of interest. As a rule of thumb, ESS (see Section 20.2) of the order of 100 are good enough for means but ESS of the order of 10000 are required for more extreme quantities such as 95% intervals.

Two techniques are often used when dealing with MCMC samples - *thinning* and *burn-in*. Thinning is discarding some of the MCMC samples, typically keeping only samples at multiples of some integer, for example, every second or every fifth sample. In terms of the quality of our estimates this is strictly worse than keeping all the samples, however, it might sometimes be useful when we have memory constraints and a highly autocorrelated chain. Thinning will result in a shorter but less autororrelated chain where the information lost might be negligible relative to the gain in memory used.

Burn-in is the process of discarding some number of samples from the start of our chain. The motivation behind this is to deal with the effects of a poorly chosen starting value. If our starting value is far from the typical values in the distribution and our chain is not long enough, this value and possibly several subsequent samples, until we get to the typical set of the distribution, will skew our estimates. Therefore, we will benefit from discarding them. However, this is just an elaborate approach to choosing the starting value as effectively that is all we do at the expense of the number of samples that we discard. If our starting value is chosen sensibly, burn-in will not be necessary. Note that burn-in should not be confused with warmup phases that many MCMC samplers have to tune their proposal distributions and other MCMC parameters. Warmup samples have to be discarded because they are not from the same Markov Chain.

How many MCMC chains to run?

In terms of the quality of our MCMC samples a long chain of length nm is always at least as good if not better than n chains of shorter length m. That is, it is better to have one chain of length 1000 than 5 chains of length 200. For example, it is possible than none of the shorter chains even reached the typical set. There is a benefit to running multiple independent chains from different starting values - if everything is OK, all the chains should behave the same so any differences help us diagnose slow mixing, multiple modes, etc. (see Section 20.2).

Note that with the availability of multiple cores or processors it is now easy to run m independent chains of length n in approximately the same time we would need to run a single chain of length n. This is of course strictly better than having a single chain of length n as we get the MCMC diagnostics benefits of multiple chains and m times as many samples.

MCMC diagnostics

Before we proceed with interpreting any quantities that are the result of MCMC, we should also diagnose if our samples exhibit any problematic behavior that

could invalidate any results. Theory informs us that problematic behavior can arise due to reducibility, periodicity, and strong autocorrelation. In practice we can add to that a poor choice of starting value that requires us to take many steps before we reach the typical set.

Before we introduce the most common MCMC diagnostics techniques note that we rarely diagnose the MCMC samples as the multivariate samples they are. Instead, we focus on the univariate (marginal) distributions of individual dimensions or scalar functions of dimensions. In the context of statistical models, we focus on one parameter at a time.

If we have a reducible chain, we will have multiple modes in sampling, depending on where we start. Unless we know what values of the parameter we can expect, it is impossible to diagnose multiple modes with a single chain as the single mode of a multi-modal chain is indistinguishable from a uni-modal chain. However, running multiple independent chains from different starting values will identify this issue. Often, a simple inspection of a joint traceplot of several chains is enough to identify multiple modes. Note that a traceplot is just a line plot of the values of the parameter against the sampling iteration. That is, we observe how the sample values change over time.

Running multiple chains can also help us identify other issues. If everything is OK with our chains and they indeed sample from the same target distribution and indeed the autocorrelation is low enough so that with our samples we have *converged* to the target distribution, then the chains in terms of their global behavior should be indistinguishable from each other. A traceplot can therefore help us identify not only multiple modes but also when at least some of the chains have not (yet) exhibit correct limiting behavior. This notion can also be quantified. The most common such diagnostic is the $\hat{R}$ (R-hat) diagnostic, also referred to as the Gelman-Rubin diagnostic. It has many variants, but the basic principle is that we compare, for a parameter, the between-chain variability with the average of the within-chain variabilities. If the chains are indeed samples from the same target distribution, then each chain should be very similar to all chains combined and the ratio of the between-chain and within-chain variability will be close to 1. If they are not similar, then the between-chain variability will be greater than 1 and an indication that something is not OK. Of course, we need a large enough sample to get a good estimate of these variabilities.

Periodicity does not require much attention, because in practice our MCMC sampler will always have some nonzero probability of remaining in the current state and therefore cannot be periodic.

We have already briefly discussed poorly chosen starting values in Section 20.2. We can identify if a starting value is far from the typical set of values by inspecting the traceplots. If we did pick our starting values poorly and the total number of samples is not large enough to deflate the influence of the samples at beginning of the chain, we can consider discarding these *burn-in* samples to improve our estimates.

The CLT for Markov Chains informs us that approximation error depends not only on the variance of the samples but also on their covariance/correlation and that strong (positive) autocorrelation will result in high approximation error. Assuming that there are no serious issues with our chain, such as multiple modes or a very poorly chosen starting value, we can in practice estimate the covariances reasonably well (see Section 20.2). This allows us to estimate the MCMC approximation error for any quantity of interest and whenever we interpret any such value, we should always interpret it in the context of its MCMC approximation error.

An often used single-number summary of the quality of a MCMC sample for some parameter θ is the Effective Sample Size (ESS):

$$\text{ESS}_\theta = m\frac{\sigma^2_{\text{MC}}}{\sigma^2_{\text{MCMC}}},$$

where m is the number of samples and σ^2_{MC} and σ^2_{MCMC} are the MC variance and MCMC variance, respectively. Note that the MCMC variance is just the MC variance plus all the covariances. If we have little autocorrelation, then MC and MCMC variance will be similar and ESS will be similar to the actual number of samples. That is, the effectiveness of our chain is similar to the effectiveness of m independent samples from the target distribution. In practice, autocorrelation will be positive and ESS will be less than the number of samples.

Once we have determined that we have very strong autocorrelation, we need to adjust (increase) the number of samples we take, so that we get an acceptable approximation error. If computation is very expensive and/or autocorrelation is very strong, the number of samples required might be infeasible. In such cases we have to change our sampler or simplify the problem.ite

To summarize, all these diagnostics tools help us identify relatively obvious issues with our MCMC chains. However, the absence of any issues does not confirm that the chain is OK. That is, these diagnostics can reveal when something is wrong but are not proof that everything is OK.

20.3 Hamiltonian Monte Carlo

In this section we will provide a short introduction to Hamiltonian Monte Carlo (HMC). For a more detailed treatment of this topic, we refer the reader to the tutorials by Neal (2011) and Betancourt (2017).

HMC is currently the state-of-the-art MCMC method for general-purpose Bayesian inference and an essential part of statistics and machine learning frameworks such as Stan, PyMC3, Tensorflow, and Pyro.

HMC deals with the relatively inefficient exploration of the target distribution and poor scaling to higher dimensions of random walk Metropolis-Hastings and

its variants. This is achieved by a physics-inspired approach to proposing the next state and by utilizing the gradient of the target distribution for a better understanding of its geometry.

Hamiltonian dynamics

Before we introduce the basic ideas of HMC, we will briefly discuss Hamiltonian dynamics, which are fundamental to understanding HMC and give it its name.

In general, Hamiltonian dynamics consists of a $d-$dimensional position vector q and a $d-$dimensional momentum vector p. The evolution of the system is determined by the function $H(q,p)$ (the Hamiltonian) and the equations:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i},$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}.$$

For HMC the Hamiltonian H is typically chosen so that it is separable. That is, that it can be written as $H(q,p) = U(q) + K(p)$, where $U(q)$ is the potential energy and $K(p)$ the kinetic energy of the system.

Simulating Hamiltonian dynamics

To simulate Hamiltonian dynamics with a computer, we need to discretize time with some step size ϵ. We will introduce the most commonly used approach - the Leapfrog method. For a more detailed discussion of this topic, see Neal (2011).

The Leapfrog method involves doing a half-step update of momentum, a full step update of position, completed by another half-step update of momentum:

$$p_i(t + \frac{\epsilon}{2}) = p_i(t) - \frac{\epsilon}{2}\frac{\partial U}{\partial q_i}(q(t)),$$

$$q_i(t + \epsilon) = q_i(t) + \epsilon\frac{\partial K}{\partial p_i}(p(t + \frac{\epsilon}{2})),$$

$$p_i(t + \epsilon) = p_i(t + \frac{\epsilon}{2}) - \frac{\epsilon}{2}\frac{\partial U}{\partial q_i}(q(t + \epsilon)).$$

Most often the kinetic energy is taken to be of the form $K(p) = \frac{1}{2}p^T M^{-1} p$, where M (the mass matrix) is diagonal, with elements $m_1, ..., m_d$. In that case the kinetic energy simplifies to $K(p) = \sum_{i=1}^{d} \frac{p_i^2}{2m_i}$ and the second row of the Leapfrog method simplifies to

$$q_i(t+\epsilon) = q_i(t) + \epsilon \frac{p(t+\frac{\epsilon}{2})}{m_i}.$$

Note that when we simulate the dynamics for several steps, we can combine the last half-step of an iteration with the first half-step of the previous iteration. The result is a half-step update for momentum, followed by several pairs of full-step updates for position and momentum and finally a half-step update for momentum.

Properties of Hamiltonian dynamics

Hamiltonian dynamics have several properties, which are important for HMC to work (see Neal (2011) for details):

- They preserve the value of the Hamiltonian. That is, the total energy of the system remains constant. This is key for HMCs ability to propose states that are far away but with a high probability of being accepted.
- They are reversible. That is, running the dynamics from a state for some time s has an inverse. For the separable Hamiltonian and kinetic energy such that $K(p) = K(-p)$ (holds for the most typical choice $K(p) = \frac{1}{2}p^T M^{-1} p$), the inverse dynamics are obtained by negating the momentum, running the dynamics for the same number of steps and negating the momentum again.
- They are symplectic and, as a consequence, volume preserving. These, together with reversibility, are key for proving that HMC leaves the target distribution invariant.

The HMC algorithm

The main idea of HMC is to introduce the distribution we want to sample from as the potential energy of the Hamiltonian dynamics. First, let's introduce the joint density of position and momentum, which is determined by the value of the Hamiltonian (the second equality is assuming that the Hamiltonian is separable, as is most often the choice with HMC):

$$p(q,p) \propto e^{-H(q,p)} = e^{-U(q)} e^{-K(p)}$$

Now we take $U(q) = -\log f(q)$, where f is proportional to the distribution we want to sample from, and use the most typical kinetic energy:

$$p(q,p) \propto f(q) e^{-\frac{1}{2}p^T M^{-1} p}.$$

The resulting joint distribution $p(q,p)$ can be seen as the target distribution over the position vector q augmented by an independent multivariate Gaussian for the momentum vector p, with mean 0 and covariance M.

Hamiltonian dynamics conserves the Hamiltonian, so all states on a trajectory will have the same density $p(\cdot, \cdot)$. That makes Hamiltonian dynamics very suitable for proposing the next state in a MCMC algorithm, because a trajectory can propose a state very far away in position q from the current state, but still with acceptance probability 1.

Starting at some state (q, p), selecting the next state by running a trajectory for L steps and step size ϵ from the current state, and repeating that process, will only be able to produce states on a part of the density p. All (infinitely many) other states in (q, p) won't be visited (the chain is not irreducible), so the stationary distribution of such a Markov chain will not be the desired $p(q, p)$. To reach every possible state, we instead sample a new momentum from the multivariate Gaussian implied by our choice of mass matrix. Because the kinetic and potential energy parts of the joint density are independent and we are sampling from the actual distribution of momentum p, this sampling leaves the target distribution invariant. That is, $p(q, p)$ remains the stationary distribution of the Markov chain.

In practice, however, the Leapfrog method, while being a stable simulation of Hamiltonian dynamics, will not conserve the Hamiltonian exactly - there will be relatively small fluctuations. That is why we still have to apply a Metropolis correction when considering the proposed state.

Putting it all together, we get the basic HMC algorithm:

Algorithm 20.3.1. *Let $f(x)$ a function that is proportional to our target density, $q_0 \in \mathbb{R}^d$ the starting value, $\epsilon > 0$ a step size, L the number of steps, M a diagonal mass matrix with diagonal elements m_i, and m the number of samples that we want to draw. Note that $H(q, p) = -\log f(x) + \frac{1}{2} p^T M^{-1} p$.*

procedure HMC*(f, q_0, ϵ, L, m)*
  for $i \leftarrow 1 : m$ ***do*** ▷ *number of samples*
    $p \sim N(0, M)$ ▷ *sample new momentum*
    get (q^*, p^*) *by running* L *Leapfrog steps with step size* ϵ *from* (q_{i-1}, p)
    $\alpha \leftarrow \min\left\{1, \frac{e^{-H(q^*, p^*)}}{e^{-H(q_{i-1}, p)}} = e^{-H(q^*, p^*) + H(q_{i-1}, p)}\right\}$ ▷ *Metropolis corr.*
    sample $u \sim U(0, 1)$
    if $u \leq \alpha$ ***then***
      $q_i \leftarrow q^*$ ▷ *accept transition*
    else
      $q_i \leftarrow q_{i-1}$
    end if
  end for
  return $q_1, ..., q_m$.
end procedure

Practicalities of HMC

First, note that HMC only works for continuous (differentiable) distributions. There is currently no generalization to discrete parameters that is stable and efficient enough for general purpose use. The most common approach to dealing with discrete parameters is to marginalize over them or to use a more specific algorithm.

The basic HMC algorithm is relatively simple to implement. Most of the complexity of a flexible general-purpose practical implementation is not in HMC itself but in the tuning of HMC parameters and the computation of the gradients. The latter is typically done using auto-differentiation. Popular tools like Stan, Pyro, and Tensorflow are all equipped with a mathematics library which fully supports auto-differentiation.

In order to use HMC, we have to determine step size, number of steps, and the mass matrix M. HMC is very sensitive to the values of these parameters. If the step size is too small, the exploration will be too slow, if it is too large, the simulation of Hamiltonian dynamics will be inaccurate. If the number of steps is too small, the trajectories will be too short and HMC will resemble random walk Metropolis Hastings. But if it is too large, we will be doing a lot of unnecessary computation and potentially returning close to the origin of the trajectory. Finally, if the inverse of the mass matrix is a poor estimate of the scale/covariance in the target distribution, lower step sizes will be required to maintain stability. In general purpose tools HMC parameters are tuned during a warmup phase[1].

HMC also allows some additional diagnostics to complement the usual MCMC diagnostics like the traceplot, ESS, and observing the agreement of multiple independent chains. The most commonly used is the number of divergent transitions (trajectories) - more

[1]For an example, this is how tuning is done in Stan: `https://mc-stan.org/docs/2_29/reference-manual/hmc-algorithm-parameters.html`

Bibliography

Betancourt, M., 2017. A Conceptual Introduction to Hamiltonian Monte Carlo. URL: https://arxiv.org/abs/1701.02434.

Geyer, C.J., 1992. Practical Markov Chain Monte Carlo. Statistical science , 473–483.

Hogg, R.V., McKean, J., Craig, A.T., 2005. Introduction to mathematical statistics. 6th ed., Pearson Education.

Kadane, J.B., 2011. Principles of uncertainty. Chapman and Hall/CRC.

Meyn, S.P., Tweedie, R.L., 2012. Markov chains and stochastic stability. Springer Science & Business Media.

Neal, R.M., 2011. MCMC Using Hamiltonian Dynamics. CRC Press. chapter 5.

Robert, C., Casella, G., 2013. Monte Carlo statistical methods. Springer Science & Business Media.

Ross, S., Peköz, E., 2007. A Second Course in Probability. ProbabilityBookstore.com.

Rudin, W., 1987. Real and Complex Analysism 3rd Ed. Mathematics series, McGraw-Hill.