Probability and Stochastic Processes

David M. McClendon

Department of Mathematics
Ferris State University

2024 edition

Contents

Chapter 1

Probability spaces

1.1 The big picture

First question

What is *probability*?

Some history of probabiity

Pascal & Fermat (1654): correspondence regarding fair odds in games of chance

Bernoulli (1713), de Moivre (1718): basic laws of discrete probability

Boltzmann (1896), Gibbs (1902): statistical mechanics of gases expressed in terms of the random motion of large numbers of particles

Kolmogorov (1933): formal, mathematical foundation of the subject

Black-Scholes (1973): application of probability to pricing of derivatives

General setup of probability

1. You intend to perform an *experiment* which has different possible *outcomes*.

2. Use *mathematical language* to predict frequencies of these outcomes under repetitions of the experiment.

Motivating Examples

1. Roll a die repeatedly, and record the number you roll (the number is the outcome).

   In this setting, you might be interested in knowing things like:

   - What is the likelihood (a.k.a. probability) you will roll a 4 on the third roll?
   - What is the probability you will roll between nine and twelve 4s if you roll the die 60 times?
   - How many rolls on the average will it take you until you roll a 4 for the eighth time?
   - What is the probability you eventually roll nineteen 5s in a row?
   - What is the probability that the sum of the first 200 numbers you roll is less than 650?

2. A driver will be involved in a random number of accidents over the course of a year, and each of these accidents will cause a random amount of damage to his/her car.

   - How long will it take (on the average) for the driver to be involved in three accidents?
   - What is the probability the driver can be accident-free for at least six years?
   - What is the probability the driver will cause more than $\$3000$ worth of damage over the course of two years?
   - What is the smallest number A such that you can be 99% sure that the driver will cause less than $\$A$ worth of damage over the next three years?
   - What amount of damage should the driver expect (on the average) to cause over the course of a year?

Probability is the branch of mathematics which solves these types of questions. To solve them, and questions like them, we will

1. learn about a bunch of *common models* for probabilistic problems, and

2. learn the *general theory* of arbitrary probabilistic models.

Both the common models and the general theory involves mastery of three intertwining mathematical concepts: *probability spaces*, *random variables* and *stochastic processes*. Loosely speaking:

1. a *probability space* is a structure on which one can formulate a mathematically legal method of computing probability;
2. a *random variable* is a measured quantity arising randomly as the result of some experiment (like the number you roll or the amount of damage done in an accident);
3. a *stochastic process* is a collection of random variables indexed by time (like the running total of the numbers you roll or the running amount of total damage done by the driver or the price of a stock).

1.2 Probability spaces

RECALL

We seek mathematical language to describe probabilistic experiments.

Definition 1.1 (Outcomes, sample spaces and events)

1. *Any single possible result of a probabilistic experiment is called an* **outcome***.*
2. *The set of all possible outcomes is called the* **sample space***. This set is usually denoted* Ω*.*
3. *Any "observable" (more on what "observable" means later) subset of the sample space is called an* **event***. Events are usually denoted by capital letters like* A*,* B*,* E*,* F*, etc.*

Notice that definitions (2) and (3) above contain the words *set* and *subset*. So to understand these definitions, we need to review some material about sets.

Sets

Definition 1.2 (Basic language associated to sets)

1. *A* **set** *is any definable collection of objects. Sets are usually denoted by capital letters.*

2. *The members of a set are called* **elements** *of the set; if x is an element of set A then we write $x \in A$. If x is not an element of A, we write $x \notin A$.*

3. *If every element of set E is also an element of set F, we say E is a* **subset** *of F and write $E \subseteq F$ or $F \supseteq E$.*

4. *Two sets E and F are said to be* **equal** *if $E \subseteq F$ and $F \subseteq E$, in which case we write $E = F$.*

5. *The* **empty set**, *denoted $\emptyset$, is the set with no elements.*

Remarks:

1. the key word in part (1) of the above definition is "definable". This basically means that the set can be described without creating any kind of logical contradiction. For more on a collection which isn't definable, Google "Russell's paradox".

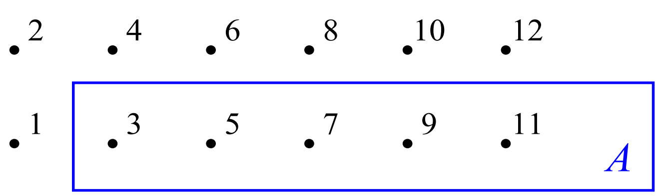
2. To say two sets are equal means that they contain exactly the same elements.

3. Note the difference between "$\in$" and "$\subseteq$": the first symbol should be preceded by an element; the second symbol should be preceded by a subset.

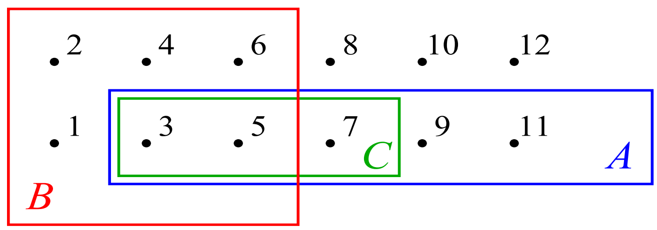
4. There is only one empty set, so we say "*the* empty set", not "*an* empty set".

Venn diagrams

A useful way to think about sets is to draw pictures called **Venn diagrams**. To draw a Venn diagram, traditionally you represent each set you're thinking about by a circle (or an oval, or a square, or a rectangle, or some other shape); think of an object as being an element of the set if and only if it is inside the shape corresponding to the set. For example, a Venn diagram for the set $A = \{3, 5, 7, 9, 11\}$ might be something like



because the box describing A contains exactly the elements of A (nothing more and nothing less). Similarly, a Venn diagram representing three sets A, B and C might be something like



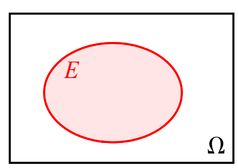
Here, this Venn diagram tells us that statements like these are all true:

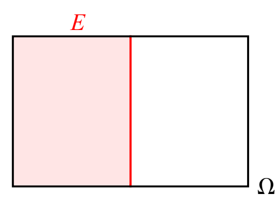
$$3 \in A \qquad 10 \notin B \qquad 2 \in B \qquad C \subseteq A \qquad B \not\subseteq A.$$

In probability, the sample space Ω is the "universal set" containing all possible outcomes of the experiment, so in any of our Venn diagrams, we can draw a box containing "everything" and label that box Ω.

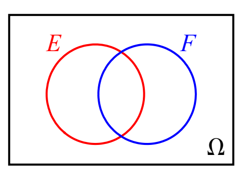
Also, in probability our Venn diagrams tend to be more abstract (since we don't actually have a list of elements of our sets), so they look more like these:

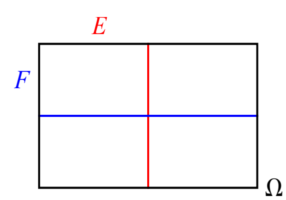
To show
one event E:



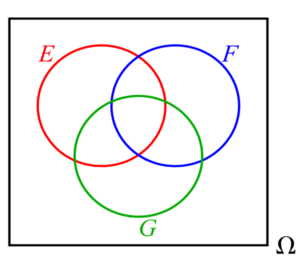


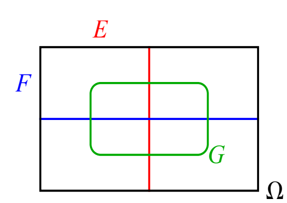
To show
two events
E and F:





To show
three events
E, F and G:





The more useful pictures in probability tend to be the ones drawn **on the right** above.

Set-builder notation

We often describe sets with "set-builder" notation. For instance, to say something like

$$E = \{x \in \mathbb{R} : 2 < x \leq 5\}$$

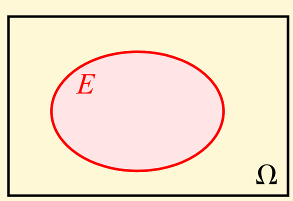
means (in English) that E is the set of real numbers x such that $2 < x \leq 5$ (in other words, E is the interval $(2, 5]$).

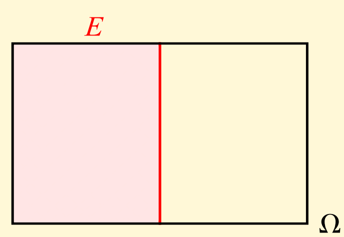
A picture of this E would look something like this:

Set operations

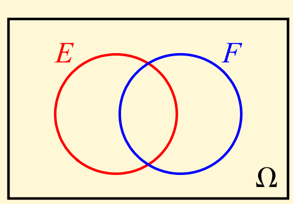
Next, we want to discuss some operations on sets which arise naturally when describing results of an experiment:

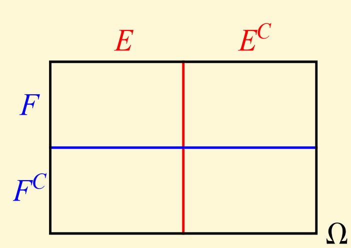
Definition 1.3 (Complements) *Given an event E, the event "E does not occur" is the* **complement** *of E and is denoted E^c, E^C, E', $\Omega - E$, $\widetilde{E}$, $\overline{E}$ (and other ways as well).*





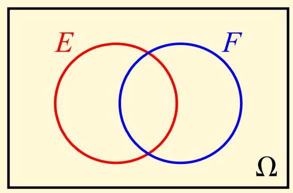
Definition 1.4 (Unions) *Given two events E and F, the event "E or F or both happen" is the* **union** *of E and F and is denoted $E \cup F$.*

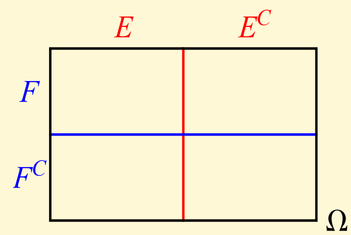




Given a bunch of events E_α indexed by α, the **union** *of these events, denoted $\bigcup_\alpha E_\alpha$, is the event that at least one of the E_α occur.*

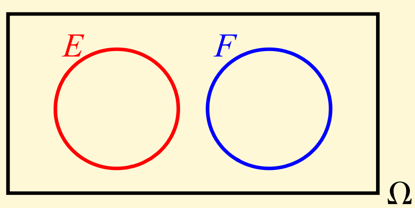
Definition 1.5 (Intersections) *Given two events E and F, the event "E and F both happen" is the* **intersection** *of E and F and is denoted $E \cap F$.*

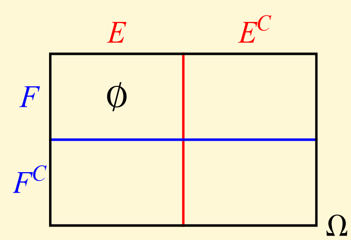




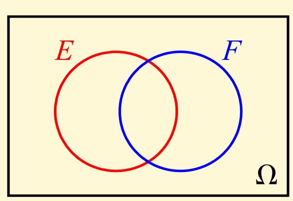
Given a bunch of events E_α indexed by α, the **intersection** *of these events, denoted $\bigcap_\alpha E_\alpha$, is the event that all of the E_α occur.*

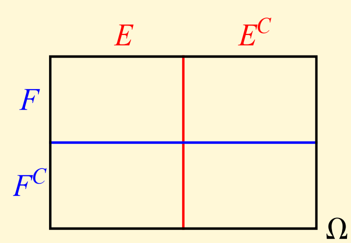
Definition 1.6 (Mutual exclusivity) *Two events E and F are called* **mutually exclusive** *or* **disjoint** *if they cannot both occur, i.e. if $E \cap F = \emptyset$.*





Definition 1.7 (Differences) *Given two events E and F, the event "E occurs, but not F" is the* **difference** *of E and F. This difference is denoted $E - F$ (also $E \backslash F$).*





Notice that $E - F = E \cap F^C$.

EXAMPLE 1

Let $E = [0, 3]$, let $F = (-\infty, 2)$, and let $G = [1, \infty)$. Describe each of these sets:

1. $E \cup F$

2. $E^C \cap G$

3. $F - G$

4. $E - F$

5. $(F \cup G)^C$

More examples of this vocabulary can be found on the next page:

		EXPERIMENT	SAMPLE SPACE	OUTCOMES	EXAMPLES OF EVENTS
DISCRETE PROBABILITY (countably many outcomes)	FINITELY MANY OUTCOMES	Toss a coin			
		Roll a die	$\Omega = \{1,2,3,4,5,6\}$	$1,2,3,\ldots,6$ $n \leftrightarrow$ rolling n face up	$\emptyset =$ rolling a 7 $E = \{1,3,5\} =$ rolling an odd number $F = \{2,3,5\} =$ rolling a prime number
	∞ MANY OUTCOMES	Flip a coin over and over until you flip a heads; record the # of flips			
CONTINUOUS PROBABILITY		Record the amount of time (starting now) until your phone rings	$\Omega = [0, \infty)$		$E = [0,5] =$ your phone rings within 5 minutes $F = \mathbb{Q} =$ your phone rings in a rational amount of time

Observability and the definition of a probability space

Start with a sample space Ω, which is just a mathematical set. We want to describe "observable" subsets of Ω, that is, subsets which we can distinguish.

Some philosophy: **I.**

II.

Given these philosophical constraints, nothing mathematical "forces" an event to be observable. We are allowed (in the most general sense) to choose a collection $\mathcal{F}$ of subsets of Ω which obey **I** and **II** above and decree the subsets belonging to $\mathcal{F}$ to be observable. The idea is that our choice of $\mathcal{F}$ should be a list of observable sets which appropriately models the problem at hand.

(It turns out that there are only two reasonable choices of $\mathcal{F}$ in MATH 414, but things get more interesting in MATH 416.)

Definition 1.8 *Let Ω be a set. A nonempty collection $\mathcal{F}$ of subsets of Ω is called a* $\sigma-$**algebra** *(a.k.a.* $\sigma-$**field***) if*

1. *$\mathcal{F}$ is "closed under complements", i.e. whenever $E \in \mathcal{F}$, $E^C \in \mathcal{F}$.*
2. *$\mathcal{F}$ is "closed under finite and countable unions and intersections", i.e. whenever $E_1, E_2, E_3, ... \in \mathcal{F}$, both $\bigcup_j E_j$ and $\bigcap_j E_j$ belong to $\mathcal{F}$ as well.*

A subset E of Ω is called $\mathcal{F}-$**measurable** *(or just* **measurable***) if $E \in \mathcal{F}$.*

The phrases "event", "measurable set" and "observable set" are synonyms.

Theorem 1.9 *Let $\mathcal{F}$ be a $\sigma-$algebra of subsets of Ω. Then $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$.*

PROOF By definition, $\mathcal{F}$ is nonempty.
Therefore, there is some set E which belongs to $\mathcal{F}$.
Since $\mathcal{F}$ is closed under complements, E^C is also in $\mathcal{F}$.
Now, since $\mathcal{F}$ is closed under finite intersections, $E \cap E^C = \emptyset \in \mathcal{F}$.
Also, since $\mathcal{F}$ is closed under finite unions, $E \cup E^C = \Omega \in \mathcal{F}$. $\square$

Examples of σ-algebras

Suppose you have a six-sided die where the sides are labeled with a red 1, a red 2, a red 3, a green 1, a green 2, and a green 3. Roll the die once and let Ω be the set of outcomes, i.e.

$$\Omega = \{⚀, ⚁, ⚂, ⚀, ⚁, ⚂\} = \{R1, R2, R3, G1, G2, G3\}.$$

Let's look at some $\sigma-$algebras on Ω.

1. Suppose a blind man rolls the die. He can tell whether the die has been rolled (by the sound), but has no idea what number is rolled. Thus the only sets he can observe are $\emptyset$ (the die hasn't been rolled) and Ω (the die has been rolled). He cannot observe the set $\{R1, R2\}$ or $\{R1, G3\}$, because to determine whether or not the outcome lies in that set, he would have to see the die.

   The $\sigma-$algebra representing the subsets a blind person can see is $\boxed{\mathcal{F} = \{\emptyset, \Omega\}}$.

   (Notice that this collection $\mathcal{F}$ of sets is a $\sigma-$algebra, meaning that it is closed under complements, countable unions and countable intersections.)

2. Suppose a red-green colorblind person rolls the die. She can observe sets like $\{R1, G1\}$, because to determine whether the outcome is in that set she only needs to see that the top face of the die has one spot. But she can't observe sets like $\{R1\}$, because she can't tell the background color of the face (so she can't distinguish between ⚀ and ⚀).

   The $\sigma-$algebra $\mathcal{F}$ representing the subsets a colorblind person can see can't be easily listed, but can be described as follows:

   $\mathcal{F}$ is the collection of sets E satisfying this property:
   $Rj \in E$ if and only if $Gj \in E$, for all $j \in \{1, 2, 3\}$.

   (Notice that this $\mathcal{F}$ is also a $\sigma-$algebra, since it is closed under complements, countable unions and countable intersections.)

3. Suppose a person with 20/20 vision rolls the die. She can distinguish any outcome. Thus the $\sigma-$algebra $\mathcal{F}$ representing the subsets she can see is the collection of $\boxed{\text{all subsets of } \Omega}$. ($\mathcal{F}$ is clearly closed under complements, countable unions and countable intersections).

Examples 1 and 3 above generalize:

Definition 1.10 *Let Ω be any set.*

- *$\mathcal{F} = \{\emptyset, \Omega\}$ is a $\sigma-$algebra called the* **trivial** *$\sigma-$algebra on Ω.*
- *The collection of all subsets of Ω, called the* **power set of** Ω *and denoted 2^Ω, is a $\sigma-$algebra on Ω.*

Fact If the sample space Ω is finite or countable (including all cases where $\Omega \subseteq \mathbb{Z}$), then we can decree $\mathcal{F}$ to be the power set of Ω and never have a problem. Thus **every subset of a finite or countable sample space can be thought of as measurable**.

Next, we want to calculate the probability of measurable sets:

More philosophy: Given a set Ω and a $\sigma-$algebra $\mathcal{F}$:

I.

II.

III.

Given these philosophical constraints, nothing else mathematical is forced on us. We are free to choose any assignment of probabilities to events that satisfies these rules. Our choice should appropriately model the context of the original problem.

Definition 1.11 *Given a set Ω and a $\sigma-$algebra $\mathcal{F}$ of subsets of Ω, a* **probability measure** *on $(\Omega, \mathcal{F})$ is a function $P : \mathcal{F} \to \mathbb{R}$ satisfying*

1. *P is* **normalized**, *meaning $P(\Omega) = 1$;*
2. *P is* **positive**, *meaning $P(E) \geq 0$ for all $E \in \mathcal{F}$;*
3. *P is* **countably additive on disjoint sets**, *meaning that if $E_1, E_2, ... \in \mathcal{F}$ are all mutually disjoint, then $P\left(\bigcup_j E_j\right) = \sum_j P(E_j)$.*

Note: Statement (3) above necessarily implies that if there are infinitely many j with $P(E_j) > 0$, then the infinite series $\sum_j P(E_j)$ must converge.

Definition 1.12 *A* **probability space** *is a triple* $(\Omega, \mathcal{F}, P)$ *where* Ω *is a set (called the* **sample space***),* $\mathcal{F}$ *is a* $\sigma-$*algebra on* Ω *(members of* $\mathcal{F}$ *are called* **events***) and* P *is a probability measure on* $(\Omega, \mathcal{F})$*.*

EXAMPLE 2

Describe a probability space which represents the result when a fair coin is tossed.

Important Remark

If the sample space Ω is finite or countable (including all situations where $\Omega \subseteq \mathbb{Z}$), then we can define $P : \mathcal{F} \to \mathbb{R}$ by writing down $P(\omega)$ for each $\omega \in \Omega$.

This is because for any event E, we can set $P(E) = \sum_{\omega \in E} P(\omega)$.

EXAMPLE 3

Suppose you roll a weighted die where 3 and 4 are three times as likely to appear as any of the other four numbers (3 and 4 are equally likely to occur). Describe a probability space which represents this experiment.

EXAMPLE 4

Flip a fair coin repeatedly until a tail lands for the first time. Describe a probability space which records the number of flips, and verify that you have constructed a probability space.

Observability in uncountable sample spaces

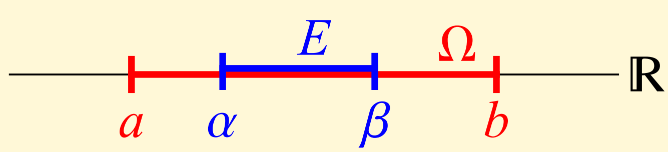
EXAMPLE 5

Choose a real number from the interval $[0,1]$ with all numbers "relatively equally likely". What is a probability space that models this problem?

Even more philosophy:

Definition 1.13 *Given any interval Ω of finite length (Ω does not have to be closed):*

1. *There is a $\sigma-$algebra $L(\Omega)$ of subsets of Ω called the* **Lebesgue $\sigma-$algebra** *which includes all intervals, all single points, and all countable unions of intervals, and*
2. *furthermore, there is a probability measure P on $(\Omega, L(\Omega))$ which assigns the probability of any interval to be its normalized length:*



$$P(E) = \frac{\text{length}(E)}{\text{length}(\Omega)} = \frac{\beta - \alpha}{b - a}.$$

This $(\Omega, L(\Omega), P)$ is a probability space called the **uniform distribution** *or* **normalized Lebesgue measure** *on Ω.*

Fact You cannot take $\mathcal{F}$ to be the power set of Ω and obtain a probability measure on $(\Omega, 2^{\Omega})$ which assigns the probability of any interval to be its normalized length.

Thus if the sample space of some experiment is represented by an interval of real numbers, and if we are going to compute probabilities in a reasonable way, we must assume that there are some sets which are not observable. It is beyond the scope of MATH 414 and 416 to actually characterize such a set; if you are interested, do a Google search for *Vitali set*. Fortunately, non-measurable sets do not arise in any real world applications of probability, so we will ignore this issue for the rest of the course.

EXAMPLE 6

Pick a number X from $[-2, 6)$ with the uniform distribution (i.e. pick the number "uniformly").

1. What is the probability that $X < 0$?

2. What is the probability that $X = 1$?

3. What is the probability that $X < 0$ or $X > 5$?

The idea of a uniform distribution generalizes to higher dimensions. The big difference is that we have to use a different notion of the "size" of a set:

dimension d	notion of "size"	how the "size" is computed
1 (i.e. $\mathbb{R}$)	length	by subtracting endpoints
2 (i.e. $\mathbb{R}^2$)		
3 (i.e. $\mathbb{R}^3$)		
> 3 (i.e. $\mathbb{R}^d$)		

Definition 1.14 *Given any set $\Omega \subseteq \mathbb{R}^d$ whose size is finite, there is a $\sigma-$algebra $L(\Omega)$ on Ω and a probability measure P on $(\Omega, L(\Omega))$ such that*

1. *$L(\Omega)$ contains all subsets of Ω whose volume is calculable using integrals;*
2. *$(\Omega, L(\Omega), P)$ is a probability space;*
3. *If $E \in L(\Omega)$, then $P(E) = \dfrac{\textit{size}(E)}{\textit{size}(\Omega)}$.*

This $(\Omega, L(\Omega), P)$ is called the **uniform distribution** *or* **normalized Lebesgue measure** *on Ω, and $L(\Omega)$ is called the* **Lebesgue $\sigma-$algebra** *on Ω .*

EXAMPLE 6

Pick a point (X, Y) from the square with vertices $(0, 0), (2, 0), (0, 2)$ and $(2, 2)$ uniformly.

1. What is the probability that $Y \geq X$?

2. What is the probability that $Y = 2X$?

3. What is the probability that $Y < X^2$?

Summary so far

- A **probability space** is a triple, consisting of
  - a **sample space** Ω (the set of all outcomes);
  - a $\sigma-$**algebra** $\mathcal{F}$ (the collection of measurable sets, closed under complements, countable unions and countable intersections);
  - and a **probability measure** P on $(\Omega, \mathcal{F})$ (P is a function which measures the probability of each measurable set; P must be *normalized*, *positive*, and *countably additive on disjoint sets*).
- If the sample space Ω is *finite* or *countable*, we can always decree every subset of Ω to be measurable (i.e. set $\mathcal{F} = 2^\Omega$) and can define P as a function on outcomes, rather than a function on events.
- If the sample space Ω is a subset of $\mathbb{R}^d$, we generally set $\mathcal{F} = L(\Omega)$, the Lebesgue $\sigma-$algebra on Ω. This $\sigma-$algebra contains all reasonable subsets of Ω, but not all subsets of Ω.
- To calculate probabilities associated to *uniform* choices of numbers or points, we compute lengths/areas/volumes as appropriate.

1.3 Elementary properties of probability spaces

RECALL

A probability space is a triple $(\Omega, \mathcal{F}, P)$ where $\mathcal{F}$ is a $\sigma-$algebra of subsets of Ω and P is a function from $\mathcal{F}$ to $\mathbb{R}$ so that P is

1.
2.
3.

We are now going to derive a long list of properties which hold in any probability space. They are called *elementary* properties of probability spaces, because they follow from the definition of a probability space without introducing other deep mathematical ideas.

Theorem 1.15 (Complement Rule) *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Then, for any event E, $P(E^C) = 1 - P(E)$.*

PROOF E and E^C are disjoint and $E \cup E^C = \Omega$, so by additivity of P, we have

$$1 = P(\Omega) = P(E \cup E^C) = P(E) + P(E^C).$$

Subtract $P(E)$ from both sides of this equation to get the result. □

Theorem 1.16 (Maximum/minimum probability) *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Then for any event E, $P(E) \in [0, 1]$.*

PROOF By definition, $P(E) \geq 0$.

By the complement rule, $P(E) = 1 - P(E^C)$.

Since $P(E^C) \geq 0$, that means $P(E) \leq 1$. □

Theorem 1.17 *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Then $P(\emptyset) = 0$.*

PROOF Apply the Complement Rule to $E = \Omega$. □

WARNING: $P(E) = 0$ does <u>**not**</u> imply $E = \emptyset$.

Theorem 1.18 (Monotonicity) *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let E and F be events. If $E \subseteq F$, then $P(E) \leq P(F)$.*

PROOF HW (as a hint, start by writing F as $F = E \cup (F \cap E^C)$.) □

Theorem 1.19 (De Morgan Law) *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let E_j be an event for all j. Then $P\left(\bigcup_j E_j\right) = 1 - P\left(\bigcap_j E_j^C\right)$.*

PROOF We will first show

$$\left(\bigcup_j E_j\right)^C = \bigcap_j E_j^C$$

and then apply the Complement Rule.

Recall from the previous page that we wanted to show

$$\left(\bigcup_j E_j\right)^C = \bigcap_j E_j^C.$$

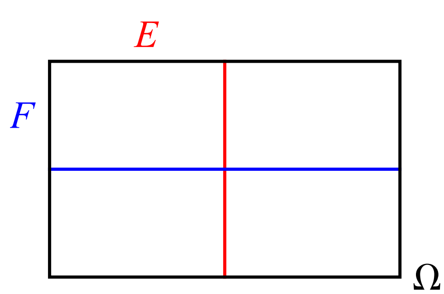
To do this, observe

$$\begin{aligned}
\omega \in \left(\bigcup_j E_j\right)^C &\iff \omega \text{ is not in } \bigcup_j E_j \\
&\iff \omega \text{ is not in at least one of the } E_j \\
&\iff \omega \text{ is in none of the } E_j \\
&\iff \omega \text{ is in all of the } E_j^C \\
&\iff \omega \in \bigcap_j \left(E_j^C\right). \square
\end{aligned}$$

Theorem 1.20 (Inclusion-Exclusion) *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let E and F be events. Then*

$$P(E \cup F) = P(E) + P(F) - P(E \cap F).$$

PROOF Start with a Venn diagram, and label each compartment of that Venn diagram with a lowercase letter representing the probability of that compartment:



Theorem 1.21 (Bonferonni Inequality) *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let E and F be events. Then $P(E \cap F) \geq P(E) + P(F) - 1$.*

PROOF HW (as a hint, use Theorems 1.16 and 1.20). □

Theorem 1.22 (General subadditivity) *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let E and F be events. Then $P(E \cup F) \leq P(E) + P(F)$.*

PROOF By Inclusion-Exclusion, $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.
Since $P(E \cap F) \geq 0$, $P(E \cup F) \geq P(E) - P(F) - 0 = P(E) + P(F)$ as wanted. □

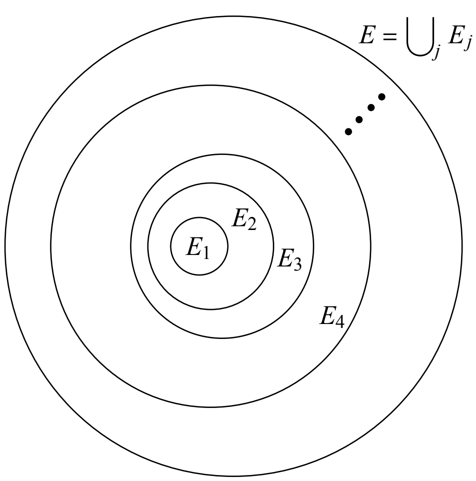
Theorem 1.23 (General subadditivity) *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $E_1, E_2, E_3, ...$ be events. Then $P\left(\bigcup_j E_j\right) \leq \sum_j P(E_j)$.*

PROOF Follows from Theorem 1.22 and induction on j. □

Theorem 1.24 (Continuity of probabliity measures I) *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $E_1, E_2, E_3, ...$ be events with $E_1 \subseteq E_2 \subseteq ....$ Let $E = \bigcup_j E_j$. Then $P(E) = \lim_{j\to\infty} P(E_j)$.*

PROOF The first step of this proof is to "disjointify" the E_j.
This means we will define a sequence of sets F_1, F_2, F_3, ... with two properties:

- The sets F_j are disjoint.
- $\bigcup_j F_j = \bigcup_j E_j$.



Theorem 1.25 (Continuity of probabliity measures II) *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $E_1, E_2, E_3, \ldots$ be events with $E_1 \supseteq E_2 \supseteq E_3 \supseteq \ldots$. Let $E = \bigcap_j E_j$. Then $P(E) = \lim_{j\to\infty} P(E_j)$.*

PROOF From the hypothesis, $E_1^C \subseteq E_2^C \subseteq E_3^C \subseteq \ldots$. Therefore

$$\begin{aligned}
P(E) &= 1 - P(E^C) \\
&= 1 - P\left[(\bigcap_j E_j)^C\right] \quad \text{(by definition of } E\text{)} \\
&= 1 - P\left(\bigcup_j (E_j^C)\right) \quad \text{(by De Morgan)} \\
&= 1 - \lim_{j\to\infty} P(E_j^C) \quad \text{(by Continuity I)} \\
&= 1 - \lim_{j\to\infty} [1 - P(E_j)] \quad \text{(by the Complement Rule)} \\
&= 1 - 1 + \lim_{j\to\infty} P(E_j) \\
&= \lim_{j\to\infty} P(E_j). \ \square
\end{aligned}$$

Applications

EXAMPLE 7

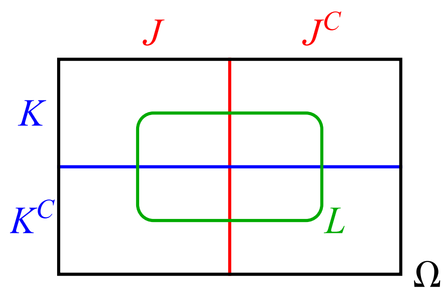
Assume $A \cup B = \Omega$, $P(A \cap B^C) = \dfrac{1}{4}$ and $P(A^C) = \dfrac{1}{3}$. Find $P(B)$.

EXAMPLE 8

Suppose events J, K and L in a probability space are such that

$$P(J) = .5,\ P(K) = .4,\ P(L) = .3 \text{ and } P(J \cup K \cup L) = .9.$$

If J and L are mutually exclusive and $P(K \cap L)$ is twice $P(K^C \cap L)$, what is $P(J - K)$?



EXAMPLE 9

The chance you lose your umbrella is at least 80%. The chance you lose your glasses is at least 75%. The chance you lose your keys is at least 60%. What is the minimum chance you lose all three items?

Generalized Inclusion-Exclusion

RECALL

Theorem 1.20 (Inclusion-Exclusion) above says:

$$P(E \cup F) =$$

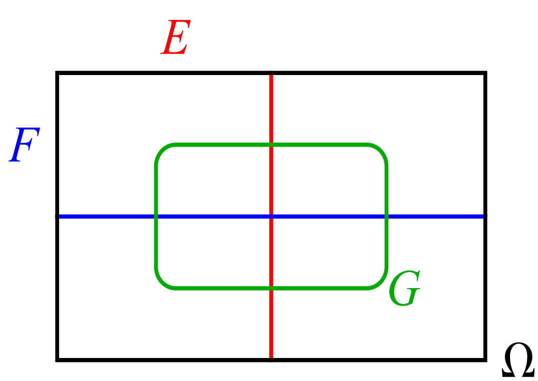
QUESTION

Can you say something similar about $P(E \cup F \cup G)$?

Theorem 1.26 (3-way Inclusion-Exclusion) *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let E, F and G be events. Then*

$$\begin{aligned} P(E \cup F \cup G) = {} & P(E) + P(F) + P(G) \\ & - P(E \cap F) - P(E \cap G) - P(F \cap G) \\ & + P(E \cap F \cap G). \end{aligned}$$

PROOF Start with a Venn diagram:



QUESTION

What about $P(E_1 \cup E_2 \cup E_3 \cup ... \cup E_n)$?

Theorem 1.27 (General Inclusion-Exclusion) *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $E_1, E_2, E_3, ..., E_n$ be events. Then*

$$P\left(\bigcup_{j=1}^{n} E_j\right) = S_1 - S_2 + S_3 - S_4 ... \pm S_n = \sum_{r=1}^{n} (-1)^r S_r$$

where

$$S_r = \sum_{1 \leq i_1 < i_2 < ... < i_r \leq n} P\left(E_{i_1} \cap E_{i_2} \cap ... \cap E_{i_r}\right).$$

EXAMPLE 10

Suppose that there are three risk factors which affect the chance one will contract a certain disease. Suppose that for any one risk factor, the probability that a randomly chosen person has any one particular risk factor is .45. Suppose that for any two risk factors, the probability that a randomly chosen person has those two risk factors is .2, and suppose that the probability that a person has all three risk factors is .07. What is the probability that a person has none of the three risk factors?

1.4 Conditional probability and independence

MOTIVATING EXAMPLE

Suppose you roll two fair dice. What is the probability that you roll two numbers that sum to 10?

	1	2	3	4	5	6
⚀ $\leftrightarrow$ 1						
2						
3						
4						
⚄ $\leftrightarrow$ 5						
6						

A CHANGE TO THE MOTIVATING EXAMPLE

Again, roll two fair dice. What is the probability that you roll two numbers that sum to 10, *given that at least one die roll is a 6*?

When you are asked to compute the probability of one event ***given*** that another one occurs, the quantity you compute is called a *conditional probability*:

Definition 1.28 *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let E and F be events with $P(F) > 0$. The* **conditional probability of** E **given** F*, denoted $P(E \mid F)$, is defined as*

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)}.$$

The definition of conditional probability can be rearranged by multiplying through the equation in Definition 1.28 by $P(F)$ to obtain

Theorem 1.29 (Multiplication Principle) *Let $(\Omega, \mathcal{F}, P)$ be a probability space, an let E and F be events with $P(F) > 0$. Then*

$$\begin{aligned} P(E \cap F) &= P(F) \cdot P(E \mid F) \\ &= P(E) \cdot P(F \mid E). \end{aligned}$$

This law is useful for computing probabilities like these, which come from experiments that have multiple stages or steps:

EXAMPLE 11

A jar contains 8 marbles, 3 of which are red. If you draw 2 marbles from the jar (one at a time, without replacement), what is the probability that both of the marbles you draw are red?

Independence

Definition 1.30 *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let E and F be events. E and F are said to be* **independent** *if $P(E \cap F) = P(E) \cdot P(F)$.*

If E and F are independent, we write $E \perp F$. Otherwise we write $E \not\perp F$.

Consequences of this definition

Lemma 1.31 *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let E and F be events. If $P(E) = 0$ or $P(E) = 1$, then $E \perp F$.*

(So in particular, $\emptyset \perp F$ and $\Omega \perp F$ for any event F.)

Furthermore, if $E \perp E$, then $P(E) = 0$ or $P(E) = 1$.

PROOF We'll prove here that if $P(E) = 0$, then $E \perp F$.

Towards that end, suppose $P(E) = 0$.

Therefore, since $E \cap F \subseteq E$, $P(E \cap F) = 0$.

Therefore $P(E \cap F) = 0 = 0P(F) = P(E)P(F)$, so $E \perp F$ by definition.

Proofs of the other statements in this lemma are HW. □

Lemma 1.32 *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let E and F be events with $P(E) > 0$ and $P(F) > 0$. Then, the following three statements are equivalent*[a]*:*

1. $E \perp F$
2. $P(E \mid F) = P(E)$
3. $P(F \mid E) = P(F)$

[a]To say statements are equivalent means that if any one of them are true, the others are true, and if any one of them is false, the others are false. We use the symbol $\iff$ in between statements that are equivalent.

PROOF This follows from basic algebra:

$$\begin{aligned}\boxed{P(E \mid F) = P(E)} &\iff \frac{P(E \cap F)}{P(F)} = P(E) \\ &\iff P(E \cap F) = P(E)P(F) \iff \boxed{E \perp F} \\ &\iff \frac{P(E \cap F)}{P(E)} = P(F) \\ &\iff \boxed{P(F \mid E) = P(F)}. \ \square\end{aligned}$$

Lemma 1.32 is interpreted like this: to say that two events are independent means heuristically **that the probability that either event occurs is not affected by knowing whether or not the other event occurs**.

Lemma 1.33 *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let E and F be events. Then, the following five statements are equivalent (HW):*

1. $E \perp F$
2. $F \perp E$
3. $E \perp F^C$
4. $E^C \perp F$
5. $E^C \perp F^C$

PROOF HW

EXAMPLE 12

Roll two fair dice. Let E be the event that you roll at least one 6, and let F be the event that you roll a total of at least 10. Are E and F independent? Give a heuristic justification of your answer, and then justify your answer algebraically.

EXAMPLE 13

Flip a fair coin six times consecutively. Compute the probability that out of the first, fourth and flips, at least one of those flips is heads.

Pairwise and mutual independence

Definition 1.34 *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space, and let* $E_1, ..., E_n$ *be events. The events* $E_1, ..., E_n$ *are called* **pairwise independent** *if* $E_i \perp E_j$ *for any* $i \neq j$.

Heuristic interpretation: To say events are *pairwise independent* means that knowing whether or not any **one** of the events occurring does not, by itself, affect the likelihood of any **one** other event.

Definition 1.35 *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space, and let* $E_1, ..., E_n$ *be events. The events* $E_1, ..., E_n$ *are called* **mutually independent** *(or just* **independent***) if for any subset* $J \subseteq \{1, ..., n\}$,

$$P\left(\bigcap_{j \in J} E_j\right) = \prod_{j \in J} P(E_j).$$

Heuristic interpretation: To say that a collection of events is *independent* means that knowing whether or not any **subcollection** of events occur does not affect the likelihood of any other **collection** of events (including any other single event) occurring.

(Mutual) independence implies pairwise independence, but not the other way around, as we see in this example:

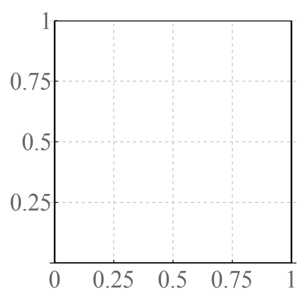
EXAMPLE 14

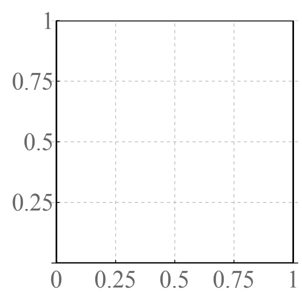
Let $\Omega = \{1, 2, 3, 4\}$ have the uniform distribution. Let $E = \{1, 2\}$, $F = \{1, 3\}$ and $G = \{2, 3\}$. Are E, F, G pairwise independent? Are E, F, G independent?

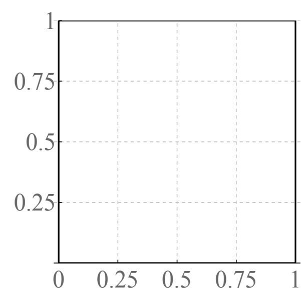
EXAMPLE 15

Let $\Omega = [0,1] \times [0,1]$ have the uniform distribution. Let $E = \left[0, \frac{1}{2}\right] \times [0,1]$, $F = [0,1] \times \left[0, \frac{1}{2}\right]$ and $G = \left([0,1] \times \left[0, \frac{1}{4}\right]\right) \cup \left([0,1] \times \left[\frac{1}{2}, \frac{3}{4}\right]\right)$.

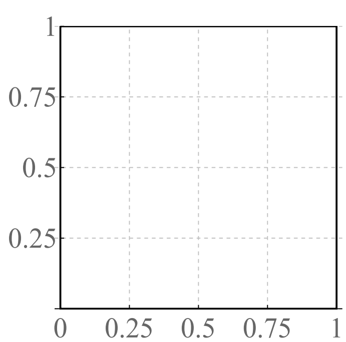
1. Are E, F, G pairwise independent?







2. Are E, F, G independent?



EXAMPLE 16 (THE MONTY HALL PROBLEM)

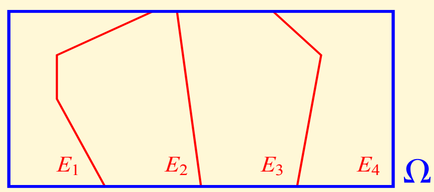
There are three doors on a game show "Let's Make a Deal". One door has a car behind it; two doors have piles of manure behind them. You pick a door. Then the game show host shows you that behind a door you did not pick, there is a pile of manure. Then he gives you the option of keeping your door, or switching to the other door you haven't seen yet. Should you switch?

1.5 The Law of Total Probability and Bayes' Law

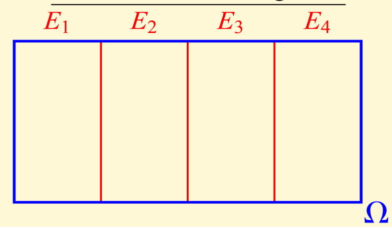
Definition 1.36 *A* **partition** *of a probability space* $(\Omega, \mathcal{F}, P)$ *is a collection of events* $E_1, ..., E_n$ *such that*

1. $P(E_i \cap E_j) = 0$ *for all* $i \neq j$ *(i.e. the* E_j *are essentially disjoint); and*
2. $P(E_1 \cup E_2 \cup ... \cup E_n) = 1$ *(i.e. the union of the* E_j*s is essentially* Ω*).*

generic picture



more useful diagram



EXAMPLE

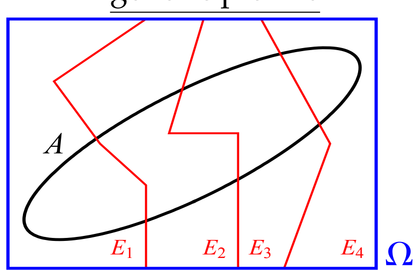
$E_1 = \left[0, \frac{1}{2}\right]$ and $E_2 = \left[\frac{1}{2}, 1\right)$ form a partition of $[0, 1]$ (with Lebesgue measure).

Theorem 1.37 (Law of Total Probability (LTP)) *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space, and let* $E_1, E_2, E_3, ..., E_n$ *be a partition. Then for any event* A*,*

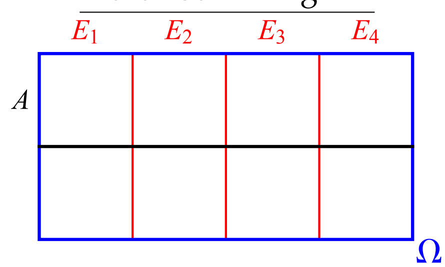
$$P(A) = \sum_{j=1}^{n} P(E_j)P(A \mid E_j).$$

PROOF Start by splitting A into its intersections with each of the E_j:

generic picture



more useful diagram



As a special case of the Law of Total Probability, note that for any event E, E and E^C form a partition of Ω. Thus for any two events A and E, the Law of Total Probability gives us

$$P(A) = P(E)P(A \mid E) + P(E^C)P(A \mid E^C).$$

EXAMPLE 17

A fair coin is flipped. If the coin lands heads, a fair die is rolled once. If the coin lands tails, a die is rolled twice independently. Find the probability that the number(s) rolled sum to 5.

EXAMPLE 18

A survey shows 54% of people age 40 or older believe in aliens, and 33% of people aged less than 40 believe in aliens. If 48% of people are age 40 or older, what percent of people believe in aliens?

Tree diagrams

Implementing the Law of Total Probability in more complicated situations often involves drawing a diagram called a *tree diagram*, rather than formally describing the events with capital letters:

EXAMPLE 19

A vase contains 3 red and 5 blue marbles. One marble is drawn from the jar and its color recorded, after which it is returned to the jar along with 2 marbles of the opposite color. Then another marble is drawn and its color recorded, after which it is returned to the jar with 2 marbles of the same color. Finally a third marble is drawn. What is the probability that of the three marbles drawn, two of them are blue?

EXAMPLE 20

(from Nate Silver's book *The Signal and the Noise*) Studies show that the chance that a woman in her forties will develop breast cancer is 1.4%. Studies also show that if a woman in her forties does not have cancer, a mammogram will incorrectly claim that she does 10% of the time, and if a woman in her forties does have breast cancer, a mammogram will detect it 75% of the time. Suppose a woman in her forties has a mammogram which indicates she has breast cancer. Given this, what is the probability she actually has breast cancer?

Without reading ahead, guess the answer to this question:

Theorem 1.38 (Bayes' Law) *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $E_1, ..., E_n$ be a partition. Then for any event A and any $k \in \{1, ..., n\}$,*

$$P(E_k \mid A) = \frac{P(E_k)P(A \mid E_k)}{\sum\limits_{j=1}^{n} P(E_j)P(A \mid E_j)}.$$

PROOF By direct calculation:

$$\begin{aligned} P(E_k \mid A) &= \frac{P(E_k \cap A)}{P(A)} && \text{(by def'n of conditional probability)} \\ &= \frac{P(E_k)P(A \mid E_k)}{P(A)} && \text{(by Multiplication Principle)} \\ &= \frac{P(E_k)P(A \mid E_k)}{\sum\limits_{j=1}^{n} P(E_j)P(A \mid E_j)} && \text{(by LTP)} \ \square \end{aligned}$$

Importance: Bayes' Law tells you how to get $P(E_k \mid A)$ given all the $P(A \mid E_j)$.

Application: Think of the E_j as hypotheses and think of the A as some bit of evidence. Theoretically, you should have an idea as to the likelihood that each hypothesis is true (i.e. you know the *prior probabilities* $P(E_k)$). Suppose you actually witness evidence A; what is the likelihood that hypothesis E_k is the correct hypothesis? This *posterior probability* $P(E_k \mid A)$ can be computed from the prior probability using Bayes' Law.

Again, note that for any event E, E and E^C form a partition of Ω. Thus for any two events A and E, Bayes' Law gives us

$$P(E \mid A) = \frac{P(E)P(A \mid E)}{P(E)P(A \mid E) + P(E^C)P(A \mid E^C)}.$$

EXAMPLE 20, REPEATED

Studies show that the chance that a woman in her forties will develop breast cancer is 1.4%. Studies also show that if a woman in her forties does not have cancer, a mammogram will incorrectly claim that she does 10% of the time, and if a woman in her forties does have breast cancer, a mammogram will detect it 75% of the time. Suppose a woman in her forties has a mammogram which indicates she has breast cancer. Given this, what is the probability she actually has breast cancer?

EXAMPLE 21

An insurance company offers three levels of insurance: A,B and C. Assume that in the next year:

- 35% of level A policyholders will file a claim;
- 12% of level B policyholders will file a claim;
- 16% of level C policyholders will file a claim.

If 80% of all policyholders are level A, and 15% are level B, what is the probability that a claim within the next year came from a level A policyholder?

Solution: Let E be the event that a claim is filed, and let A, B and C be the events that the policyholder has level A,B and C insurance, respectively. Here is the given information:

$$\begin{aligned} P(A) &= .8 & P(E \mid A) &= .35 \\ P(B) &= .15 & P(E \mid B) &= .12 \\ & & P(E \mid C) &= .16 \end{aligned}$$

Also, since $\{A, B, C\}$ forms a partition of Ω, we can figure out that

$$P(C) = 1 - P(A) - P(B) = 1 - .8 - .15 = .05.$$

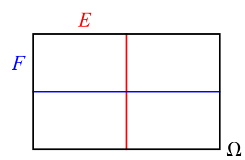
So by Bayes' Law, we have

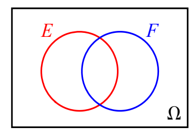
$$\begin{aligned} P(A \mid E) &= \frac{P(A)P(E \mid A)}{P(A)P(E \mid A) + P(B)P(E \mid B) + P(C)P(E \mid C)} \\ &= \frac{.8(.35)}{.8(.35) + .15(.12) + .05(.16)} \\ &= \boxed{.915033}. \end{aligned}$$

1.6 Chapter 1 Homework

Exercises from Section 1.2

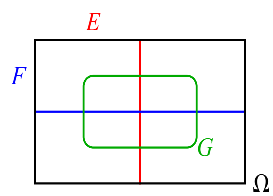
1. a) Shade the region corresponding to the set $E \cup F^C$ on the Venn diagram shown below at left.

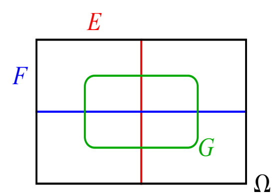




   b) Shade the region corresponding to the set $(E \cup F)^C \cap (E \cup F)^C$ on the Venn diagram shown above at right.

   c) Shade the region corresponding to the set $E \cup F \cup G^C$ on the Venn diagram shown below at left.





   d) Shade the region corresponding to the set $F \cap (E - G)$ on the Venn diagram shown above at right.

   e) Shade the region corresponding to the set $(E^C \cap F) \cup G^C$ on a Venn diagram similar to the ones given in parts (c) and (d).

2. In this problem, assume $E = \{0, 2, 4, 6, 8, 10, 12\}$, $F = \{0, 1, 2, ..., 8\}$, $G = \{4, 5, 6, ..., 12\}$ and $H = \{0, 3, 6, 9, 12\}$. (The universal set is $\Omega = \{0, 1, 2, ..., 12\}$.) For each given set, list the elements in the set (using proper notation, i.e. surrounding the list with braces):

   a) $E - H$
   b) F^C
   c) $E \cap F$
   d) $G^C \cup H$
   e) $(H - E) \cup (H - F^C)$
   f) $E \cap ((F \cup G) - E)$

3. Suppose you flip a fair coin three times, and record the outcomes with Hs and Ts. Describe the following events in words (your description should be as efficient as possible):

   a) $E = \{HHH, TTT\}$
   b) $E = \{HHT, HTH, THH\}$

c) $E = \{HHH, HHT, HTH, HTT\}$

d) $E = \{HHT, HTH, HTT, THH, THT, TTH, TTT\}$

4. A box contains 4 marbles: 2 red, 1 green, and 1 blue.

   a) Consider an experiment that consists of taking 1 marble from the box, putting it back and drawing a second marble from the box (recording both choices in order). Describe the sample space for this experiment (your sample space should be constructed so that all the outcomes are equally likely).

   b) Suppose you didn't put the first marble back before you drew the second marble. Describe the sample space in this context (again, your sample space should be constructed so that all the outcomes are equally likely).

5. In each part of this problem, you are given a set Ω and a description of some a collection $\mathcal{F}$ of subsets of Ω. Determine, with some justification, whether or not the collection $\mathcal{F}$ forms a σ-algebra:

   a) $\Omega = \{1, 2, 3, 4\}$; $\mathcal{F} = \{\emptyset, \{1, 2\}, \{3, 4\}, \Omega\}$

   b) $\Omega = \{1, 2, 3, 4\}$; $\mathcal{F} = \{\emptyset, \{1\}, \{2\}, \{1, 2\}, \Omega\}$

   c) $\Omega = \mathbb{R}$; $\mathcal{F}$ is the collection of sets E which have the property that either E is finite or E^C is finite.

6. Suppose you perform an experiment where there are eight possible outcomes. Assuming that every subset of outcomes constitutes an event, how many distinct events are there?

   Hint: You may want to try this problem in the situation where there are two, three and/or four outcomes, and look for a pattern.

7. Suppose you roll a fair die repeatedly until a 4 turns up. You record the number of rolls it takes to roll a 4. Describe a probability space for this experiment. Verify that you have constructed a probability space.

8. Suppose a point (x, y) is picked at random (with the uniform distribution) from the triangle in the xy-plane with vertices at $(0, 0)$, $(4, 0)$, and $(4, 4)$.

   a) What is the probability that $x \geq 2$?

   b) What is the probability that $x < y^2$?

9. Suppose a point (x, y) is chosen uniformly from the rectangle in the xy-plane whose vertices are $(0, 0)$, $(4, 0)$, $(0, 2)$ and $(4, 2)$. Let E be the event that $y > x$, and let F be the event that $x < 1$.

   a) Compute $P(E \cup F)$.

b) Compute $P(E^C \cap F)$.

10. Suppose a point (x, y, z) is chosen from the unit cube (this means a cube with opposite vertices at $(0, 0, 0)$ and $(1, 1, 1)$).

    a) Compute the probability that $x < \frac{1}{2}$ and $y > \frac{1}{3}$.

    b) Compute the probability that $x + y + z < 1$.

11. Four players, Al, Bal, Cal and Dal, take turns flipping a fair coin (Al goes first followed by Bal, then Cal, then Dal, then Al again, then Bal, etc.). The first player to flip a head wins. What is the probability of each player winning?

    Hint: Construct a probability space for this experiment where the outcomes correspond to the number of flips it takes for someone to win. Then, to find the probability that Al wins, add up the probabilities associated to the numbers of flips that would result in Al winning. Proceed from there.

12. (This is a famous problem in probability called *The Triangle Problem*.) Suppose you take a stick of length 1 and break it into three pieces, choosing the break points uniformly and independently. What is the probability that the three pieces can be used to form a triangle?

    Hint: In a triangle, the sum of the lengths of any two sides must be at least the length of the third side.

Exercises from Section 1.3

13. Prove that there is no such thing as a uniform distribution on $\mathbb{N} = \{1, 2, 3, ...\}$.

    Hint: Prove this by contradiction: suppose that there *is* a uniform distribution on $\mathbb{N}$. This means that $P(m) = P(n)$ for every $m, n \in \mathbb{N}$. There are two possibilities: either $P(1) = 0$ or $P(1) > 0$. Explain why both of these cases are impossible thinking about what the value of $P(\Omega)$ would end up being.

14. Let $(\Omega, \mathcal{A}, P)$ be a probability space. Prove (using the definition of probability space) the **Monotonicity** law, which says that if E and F are events with $E \subseteq F$, then $P(E) \leq P(F)$.

    Hint: Write F as the union of the two sets $E \cap F$ and $E' \cap F$.

15. Prove the **Bonferonni Inequality**, which says that given any two events E and F, $P(E \bigcap F) \geq P(E) + P(F) - 1$.

16. Suppose two fair dice are rolled and that the 36 possible outcomes are equally likely. Compute the probability that the sum of the numbers on the two faces is even.

17. (AE) (The "(AE)" means this is, or closely resembles, an old actuarial exam problem.) The probability that a small fire in a kitchen destroys a microwave oven is 70%. The probability that a small fire in a kitchen destroys a refrigerator is 50%. If the probability that a small fire destroys both is 45%, find the probability that the fire destroys neither the microwave nor the refrigerator.

18. A survey reveals that 20% of the population is afraid of ghosts, 35% of the population is afraid of vampires, and 40% of the population is afraid of zombies. 15% of the population fears ghosts and vampires; 12% of the population fears ghosts and zombies, and 20% of the population fears vampires and zombies. If 8% of the population fears ghosts, vampires and zombies, what percent of the population isn't afraid of any of the three mythical creatures discussed in the survey?

19. (AE) Suppose events A and B are such that $P(A) = \dfrac{2}{5}$ and $P(B) = \dfrac{2}{5}$. If you also know $P(A \cup B) = \dfrac{1}{2}$, compute $P(A \cap B)$.

Exercises from Section 1.4

20. Suppose a point is picked uniformly from the square whose vertices are $(0,0)$, $(1,0)$, $(0,1)$ and $(1,1)$. Let E be the event that the selected point is in the triangle bounded by the lines $y = 0$, $x = 1$ and $x = y$, and let F be the event that it is in the rectangle with vertices $(0,0), (1,0), (1, \frac{1}{2})$, and $(0, \frac{1}{2})$.

    a) Compute $P(E|F)$.
    b) Compute $P(F|E)$.
    c) Are E and F independent? Why or why not? (You need an algebraic proof.)

21. Suppose a number x is selected uniformly from the interval $[0, 100]$. Let J be the event that the number selected is in $[0, 50]$, let K be the event that the number selected is in $[30, 60]$, and let L be the event that the number selected is in $[20, 70]$.

    a) Compute $P(J \cup K \mid L)$.
    b) Compute $P(J \cap L^C \mid K \cup L^C)$.

    Note: Part of the point of this problem is to teach you order of operations with conditional probabilities. In particular, there are "invisible parentheses" that surround everything in front of any | and everything after any | in any conditional probability expression.

22. (AE) If $P(A) = .7$, $P(A \bigcap B^C) = .6$ and $A \perp B$, what is $P(B)$?

23. A coin is tossed three times. Consider the following events:
    - $A =$ flipping heads on the first toss
    - $B =$ flipping tails on the second toss
    - $C =$ flipping heads on the third toss
    - $D =$ flipping the same side of the coin all three times
    - $E =$ flipping heads exactly once in the three tosses

    a) Which one or ones of the following pairs of these events are independent? A and B, A and D, A and E, D and E (No proof is required here, if you want to use the heuristic idea of independence.)

    b) Which one or ones of the following triples of these events are independent? A, B and C; A, B and D; C, D and E (No proof is required here, if you want to use the heuristic idea of independence.)

24. a) Suppose that an event E has probability 1. Prove that E is independent of any other event F.

    b) Prove that if an event E is pairwise independent with itself, then $P(E) = 0$ or $P(E) = 1$.

25. a) Suppose E and F are independent. Prove that E^C and F^C are independent.

    b) Suppose E and F are independent. Prove that E and F^C are independent.

26. A fair die is rolled repeatedly until the first time a 5 is rolled. Given that it takes an even number of rolls to obtain that first 5, what is the probability that a 5 is rolled within the first 10 rolls?

27. A point is chosen uniformly from the unit square $[0, 1] \times [0, 1]$. Find a positive number c so that the events $E = \{(x, y) : y + cx \leq 1\}$ and $F = \{(x, y) : y \leq 2x/3\}$ are independent.

    Hint: There are two values of c which solve this problem; you need to find one or the other, not both.

Exercises from Section 1.5

28. There are three boxes, labeled I, II and III. Box I contains 2 white balls and 2 black balls; box II contains 2 white balls and 1 black ball; and box III contains 1 white ball and 3 black balls.

    a) One ball is selected from each box (the draws are independent of one another). Calculate the probability of drawing all white balls.

b) Suppose you have five slips of paper, two labeled “I”, two labeled “II” and one labeled “III”. One of these five slips is drawn uniformly and then a ball is drawn from the box indicated by the slip of paper chosen. Calculate the probability that the drawn ball is white.

29. An urn contains 3 red and 2 blue marbles. One marble is drawn from the jar and its color noted. That marble, along with 2 extra marbles of the same color, is then returned to the jar. A second marble is drawn from the jar.

    a) What is the probability that the two marbles drawn are of the same color?

    b) What is the probability that the second marble drawn is red?

30. Suppose a student takes a multiple choice exam where each question has 5 possible answers, exactly one of which is correct. If the student knows the answer to the question, she selects the correct answer. Otherwise, she guesses uniformly from the 5 possible answers. Assume that the student knows the answer to 70% of the questions.

    a) What is the probability that on any single given question, the student gets the correct answer?

    b) What is the probability that the student knows the answer to a question, given that she got the question correct?

31. (AE) Suppose a factory has two machines A and B which make 64% and 36% of the total production, respectively. Of their output, machine A produces 2% defective items and machine B produces 5% defective items. Find the probability that a given defective part was produced by machine B.

32. (AE) The probability that a randomly chosen male has a blood circulation problem is $.325$. Males who have a circulation problem are twice as likely to be smokers as those who do not have a blood circulation problem. What is the conditional probability that a male has a blood circulation problem, given that he is a smoker?

Calculus review

Later in the course, we’ll see that we need calculus to do lots of computations. To make sure you are up to speed, the first few chapters of these notes have some review problems incorporated into the homework. Here is the first batch:

33. a) Let $F(x) = 2x + 3 - 4x^2$. Compute $\frac{d}{dx}F$.

    b) Let $F(x) = (x^2 - 2)^5$. Compute $F'(x)$.

c) Let $F(x) = \dfrac{2}{\sqrt{x}} - \dfrac{1}{x} + \sqrt[4]{x}$. Compute $\frac{dF}{dx}$.

34. a) Let $F(x) = x^2 e^{4x}$. Compute $F'(x)$.
    b) Let $F(x) = 2e^{-x/4} - 6e^{x/3}$. Compute $F'(x)$.
    c) Let $M(t) = (.3 + .7e^t)^8$. Compute $M''(t)$.

35. a) Let $M(t) = \dfrac{4}{4-t}$. Compute $M'''(0)$.
    b) Let $G(t) = e^{3(e^t-1)}$. Compute $G''(1) - [G'(1)]^2$.

36. Compute each integral, and then use a calculator to get a decimal approximation of your answer:
    a) $\displaystyle\int_3^6 \frac{2}{x^{3.5}}\,dx$
    b) $\displaystyle\int_{1/2}^1 \frac{4}{x^8}\,dx$
    c) $\displaystyle\int_1^4 x^2 \frac{3.25}{x^{4.25}}\,dx$

37. a) Determine a value of b so that $\displaystyle\int_0^b \frac{x^2}{12}\,dx = \frac{1}{2}$.
    b) Determine a value of c so that $\displaystyle\int_2^4 c\sqrt{x}\,dx = 1$.
    c) Suppose $\displaystyle\int_0^b ax^3\,dx = 1$. Solve for b in terms of a.

Chapter 2

Discrete random variables

2.1 Introducing random variables

Definition 2.1 *A* **random variable (r.v.)** X *is a (measurable) function* $X : \Omega \to \mathbb{R}^d$*, where* $(\Omega, \mathcal{F}, P)$ *is a probability space. The* **range** *of* X *is the set of values taken by* X*.*

Definition 2.2 *A r.v. is called* **real-valued** *if its range is a subset of* $\mathbb{R}$*. It is called* **vector-valued** *(or* $d-$**dimensional** *or a* **joint distribution***) if its range is a subset of* $\mathbb{R}^d$ *for* $d > 1$*.*

Technical remark: In MATH 414, the adjective "measurable" can be ignored without a problem. To be technically precise, a function $X : \Omega \to \mathbb{R}^d$ is measurable if given any subset S of the codomain $\mathbb{R}^d$ whose volume you can compute with calculus, the inverse image of S under X is an event. We'll get into this more in MATH 416, but you would never need to worry about this technicality too much unless you go to graduate school in mathematics.

EXAMPLES OF RANDOM VARIABLES

Example A: Roll a fair die and let X be the number rolled.

Example B: Flip a fair coin 3 times and let X be the number of times you flip heads.

Example C: Roll a die repeatedly; let X be the number of rolls it takes for the running total of your rolls to be even.

Example D: Let X be the smallest amount of time between successive text messages you receive in the next 48 hours.

Example E: You and your friend plan to meet at The Rock between 6 and 7 PM. Let X record both your arrival time and your friend's arrival time, in terms of the number of minutes after 6 that you each arrive.

Classifying random variables

On the face of things, it seems (based on the definition) that you need a lot of information to describe a random variable: the Ω, the $\mathcal{F}$, the P and the rule for X. **In practice, you don't actually use any of this to characterize a random variable.**

First concept: Random variables can be partitioned into three types:

1.

2.

3.

The way you think about a r.v. (and the way you perform calculations related to the r.v.) depends heavily on which type of r.v. you are dealing with. **So the first thing you must do when dealing with any r.v. is to determine which of these three types it is.**

2.2 Density functions of discrete random variables

For now, we study discrete r.v.s; we'll deal with the others in Chapter 3.

Definition 2.3 *A subset S of $\mathbb{R}^d$ is called* **discrete** *if given any $x \in S$, you can draw a circle (or sphere) of positive radius around x such that the only point inside that circle belonging to S is x itself.*

EXAMPLES

$\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Z}^d$ are discrete; any finite set is discrete; any subset of a discrete set is discrete.

NONEXAMPLES

$\mathbb{Q}$, $\mathbb{R}$, $\mathbb{Q}^d$ are not discrete; any set containing an interval or a curve is not discrete.

Remark: Knowing the examples and nonexamples listed above is sufficient for MATH 414 and MATH 416.

Some enrichment: Discreteness is not really a concept of probability theory. It comes from a branch of mathematics called *topology*. In fact, a better definition of discreteness comes from topology - a subset of a metric space is discrete if and only if it has no cluster points (if and only if all its points are isolated).

Definition 2.4 *A random variable X is called* **discrete** *if its range is a discrete set.*

QUESTION

Which one or ones of Examples A,B,C,D,E given above are discrete r.v.s?

RETURN TO EXAMPLE B

(Flip a fair coin 3 times and record the number of heads)

Definition 2.5 *Let $X : \Omega \to \mathbb{R}^d$ be a discrete random variable. A* **density function** *(a.k.a.* **PDF** *a.k.a.* **pdf** *a.k.a.* **mass function** *a.k.a.* **pmf***) for X is a function*

$$f_X : Range(X) \to \mathbb{R}$$

which satisfies

$$f_X(x) = P(X = x)$$

for all $x \in \mathbb{R}^d$.

We express density functions either by giving a formula for them, or by giving a chart:

EXAMPLE 1

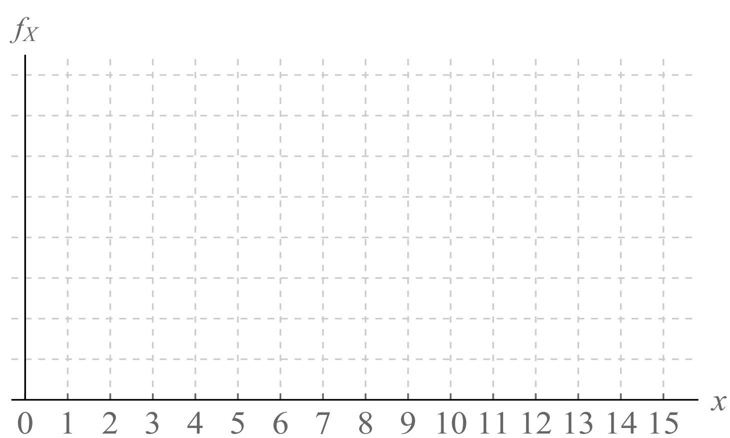
Find density functions for the r.v.s described in Examples A and B above.

EXAMPLE 2

There are two dice which are rolled; one which is normal and one whose sides are numbered $1, 1, 2, 4, 4, 6$. Let X represent the sum of the two numbers rolled. Find a density function of X, and sketch its graph. Finally, explain how you can compute $P(X \geq 10)$ from the density function.

$W \backslash N$	1	2	3	4	5	6
1						
1						
2						
4						
4						
6						

x	$f_X(x)$



Key idea: If you want to do any probabilistic calculations related to a discrete real-valued r.v., all you need to be given (or all you need to figure out) is the density function of that r.v. This is because if you are given any set $E \subseteq \mathbb{R}^d$,

$$P(X \in E) = \sum_{x \in E} P(X = x) = \sum_{x \in E} f_X(x)$$

so long as X is discrete.

Properties of density functions of discrete r.v.s

Theorem 2.6 (Properties of density functions) *A function f is the density function of a discrete r.v. if and only if:*

1. $f(x) \geq 0$ *for all* x;
2. $\{x : f(x) > 0\}$ *is a discrete set; and*
3. $\displaystyle\sum_{x \in \{x : f(x) > 0\}} f(x) = 1.$

EXAMPLE 3

Suppose a r.v. X takes only the values $2, 3$ and 4 and has a density function that is proportional to $\dfrac{1}{x^2}$. What is the probability that $X = 2$?

2.3 Counting principles

The first situation we want to model using random variables is when we select a number (or vector or some other kind of object) from a finite set, with all numbers (vectors/objects) equally likely. The random variable that describes this is called a discrete uniform r.v.:

Definition 2.7 *Let $\Omega \subseteq \mathbb{R}^d$ be a finite set. A* **(discrete) uniform** *random variable on Ω is a r.v. X whose density function is*

$$f_X(x) = \begin{cases} \dfrac{1}{\#(\Omega)} & \text{if } x \in \Omega \\ 0 & \text{else} \end{cases}.$$

If X is uniform on Ω, we write $X \sim Unif(\Omega)$.

EXAMPLE 4

Let X be the number rolled if you roll one fair die. Describe X, by giving its density function and characterizing X with appropriate language using the $\sim$ symbol.

Theorem 2.8 *Suppose $X \sim Unif(\Omega)$. Then given, any subset E of Ω, we can compute the probability that $X \in E$ by* **counting***:*

$$P(E) = P(X \in E) = \frac{\#(E)}{\#(\Omega)}.$$

EXAMPLE 5

Deal 2 cards from a 52 card deck. What is the probability that you get two aces?

To solve problems like this, it behooves us to learn how to count certain sets of objects quickly. The study of counting complicated sets of objects is called **combinatorics**.

(In what follows, $\#(E)$ refers to the number of elements in set E; all sets in this section should be assumed finite.)

Basic counting principles

The first principle of counting is very simple: if you can divide the things you are counting into two disjoint groups, you can count the groups separately and add the answers. For example, if you have 5 red apples and 3 green apples, how many apples do you have?

Theorem 2.9 (Addition Principle of Counting) *Let E and F be finite sets. If $E \cap F = \emptyset$, then*

$$\#(E \cup F) = \#(E) + \#(F).$$

If you divide the things you are counting into two groups which overlap, you can use Inclusion-Exclusion to count them. The proof of this principle is virtually identical to the probabilistic version given in the previous chapter:

Theorem 2.10 (Inclusion-Exclusion Principle (Counting Version)) *Let E and F be finite sets. Then:*

$$\#(E \cup F) = \#(E) + \#(F) - \#(E \cap F).$$

EXAMPLE 6

Suppose that 17 students surveyed like pepperoni on their pizza, 13 students surveyed like mushroom on their pizza and 20 students like pepperoni or mushroom on their pizza. How many students like pepperoni and mushroom on their pizza?

Theorem 2.11 (Multiplication Principle of Counting) *If E is a finite set of objects, each of which can be described as the result of a sequence of independent "choices", where:*

- *there are m_1 options for the first choice;*
- *each of the first choices allows m_2 options for the second choice;*
- *each of the first two choices allows m_3 options for the third choice; etc.*

then

$$\#(E) = m_1 m_2 m_3 \cdots m_n.$$

EXAMPLE 7

How many different license plates can a state make if each plate has 4 letters followed by 3 nonzero digits?

Orderings and factorials

EXAMPLE 8

How many different orderings of the letters in the English alphabet are there?

The result of the previous example generalizes:

Definition 2.12 *Let $n \in \mathbb{N}$. Then $n!$, read n* **factorial**, *is*

$$n! = n(n-1)(n-2)(n-3)\cdots 3 \cdot 2 \cdot 1.$$

As a special definition, we let $0! = 1$.

Notice: For any $n \in \mathbb{N}$, $n \cdot (n-1)! = n!$ (this explains why $0!$ should be 1).

The significance of factorials is as follows:

Theorem 2.13 (Orderings) *The number of distinct ways to order n different objects is $n!$.*

QUESTION

Is there such a thing as $(3.5)!$ or $\pi!$? If so, what might that be?

Permutations

EXAMPLE 9

There are 10 people in a club. How many different sets of officers (president, VP, secretary and treasurer) can be selected from this club?

In Example 9, we are selecting an **ordered** subset of 4 from a set of 10. These ordered subsets have names:

Definition 2.14 *An ordered subset taken from a larger finite set is called a* **permutation**.

Theorem 2.15 (Permutations) *The number of ordered sets of size k, taken from a set of size n is*

$$\frac{n!}{(n-k)!} = n(n-1)(n-2)\cdots(n-k+1).$$

Combinations

EXAMPLE 10

If there are 10 people in a club, how many different $4-$person committees can be formed? (In other words, how many **unordered** groups of 4 from the group of 10 are there?)

Definition 2.16 *An unordered subset (equivalently, just a subset) taken from a larger finite set is called a* **combination**.

Theorem 2.17 (Combinations) *The number of unordered sets of size k taken from a set of size n is denoted $\binom{n}{k}$ (read "n choose k") or $C(n,k)$ or ${}_nC_k$ and is given by the formula*

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

EXAMPLE 11

$$\binom{7}{3} = \frac{7!}{3!(7-3)!} = \frac{7\cdot 6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1}{3\cdot 2\cdot 1\cdot 4\cdot 3\cdot 2\cdot 1} =$$

The numbers $\begin{pmatrix} n \\ k \end{pmatrix}$ are called *binomial coefficients*:

Theorem 2.18 (Properties of binomial coefficients) *Let $n, k \in \mathbb{N}$. Then:*

Binomial symmetry: $\binom{n}{k} = \binom{n}{n-k}$.

Anything choose zero (or itself) is 1: $\binom{n}{0} = \binom{n}{n} = 1$.

Anything choose 1 is itself: $\binom{n}{1} = \binom{n}{n-1} = n$.

Binomial addition formula: $\binom{n}{k-1} + \binom{n}{k} = \binom{n+1}{k}$.

PROOF The first three statements follow from Theorem 2.17 directly.

We will prove the binomial addition formula with some algebra:

$$\binom{n}{k-1} + \binom{n}{k} = \frac{n!}{(k-1)!(n-k+1)!} + \frac{n!}{k!(n-k)!}$$

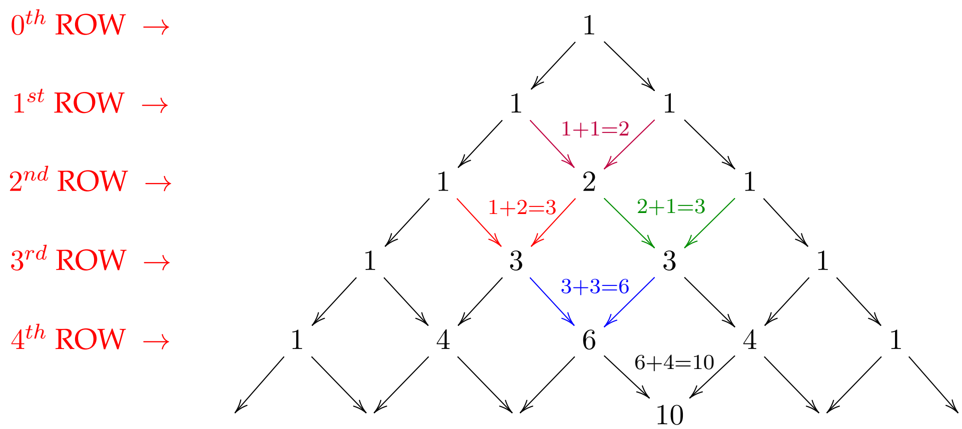
Now add these fractions by finding a common denominator:

$$\begin{aligned}
\frac{n!}{(k-1)!(n-k+1)!} + \frac{n!}{k!(n-k)!} &= \frac{n!{\color{red}k}}{k!(n-k+1)!} + \frac{n!{\color{blue}(n-k+1)}}{k!(n-k+1)!} \\
&= \frac{n![{\color{red}k} + {\color{blue}(n-k+1)}]}{k!(n-k+1)!} \\
&= \frac{(n+1)!}{k!(n+1-k)!} \\
&= \binom{n+1}{k}. \ \square
\end{aligned}$$

Definition 2.19 *Let $n, k \in \mathbb{N}$. If $n < k$, we set* $\binom{n}{k} = 0$.

This definition makes sense because if $k > n$, there are no subsets of size k that can be taken from a set of size n.

Pascal's Triangle



From statements (2) and (4) of Theorem 2.18 above, the entries of Pascal's Triangle must be the binomial coefficients (because they have the same entries down the sides and they satisfy the same addition law). So Pascal's Triangle is really an array of the binomial coefficients:

$$\binom{0}{0} = 1$$

$$\binom{1}{0} = 1 \qquad \binom{1}{1} = 1$$

$$\binom{2}{0} = 1 \qquad \binom{2}{1} = 2 \qquad \binom{2}{2} = 1$$

$$\binom{3}{0} = 1 \qquad \binom{3}{1} = 3 \qquad \binom{3}{2} = 3 \qquad \binom{3}{3} = 1$$

$$\binom{4}{0} = 1 \qquad \binom{4}{1} = 4 \qquad \binom{4}{2} = 6 \qquad \binom{4}{3} = 4 \qquad \binom{4}{4} = 1$$

EXAMPLE 12

A restaurant has 12 appetizers, 20 entrees and 5 desserts. If your table splits 3 appetizers, 5 entrees and 2 desserts, how many different meals are possible (assuming no doubling up of the same appetizer, entree or dessert)?

EXAMPLE 13

Deal 5 cards from a standard deck. What is the probability of being dealt a full house?

EXAMPLE 14

Deal 5 cards from a standard deck. What is the probability of being dealt two pair (but not a full house and not 4-of-a-kind)?

Binomial coefficients are often used to expand expressions, for example

$$\begin{aligned}
(x+y)^4 &= (x+y)^2(x+y)^2 \\
&= (x^2+2xy+y^2)(x^2+2xy+y^2) \\
&= x^4+2x^3y+x^2y^2+2x^3y+2x^2y^2+2xy^3+x^2y^2+2xy^3+y^4 \\
&= x^4+4x^3y+6x^2y^2+4xy^3+y^4 \\
&\qquad \text{(now, write in reverse order)} \\
&= y^4+4xy^3+6x^2y^2+4x^3y+x^4 \\
&= 1x^0y^{4-0}+4x^1y^{4-1}+6x^2y^{4-2}+4x^3y^{4-3}+1x^4y^{4-4}
\end{aligned}$$

More generally, we have:

Theorem 2.20 (Binomial Theorem) *Let $x, y \in \mathbb{R}$ and let $n \in \mathbb{N}$. Then*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

PROOF Expand out $(x+y)^n$:

$$(x+y)^n = (x+y)(x+y)(x+y)(x+y)\cdots(x+y)(x+y)$$

When you expand this, each term of your answer will be the product of n numbers, all of which are x or y. So each term is of the form $x^k y^{n-k}$.
Next, fix k. The number of $x^k y^{n-k}$ terms in the expansion is the number of different ways to choose which k of the n $(x+y)$s being multiplied together contribute an x to the term.
There are $\binom{n}{k}$ such ways to do this, so the coefficient on $x^k y^{n-k}$ in the expansion is $\binom{n}{k}$. The theorem follows by adding these terms over the k. □

In MATH 414 & 416, the Binomial Theorem is most often used to simplify sums of series obtained in some probability computation:

EXAMPLE 15

$$\sum_{x=0}^{14} \binom{14}{x} 2^x y^{18-x} =$$

Corollary 2.21 *Let $n \in N$. Then*

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n.$$

PROOF

$$\sum_{k=0}^{n} \binom{n}{k} =$$

Distinguishable arrangements

EXAMPLE 16

How many different arrangements of the letters in the word MISSISSIPPI are there?

Theorem 2.22 (Distinguishable arrangements, a.k.a. MISSISSIPPI rule)
Suppose you have $n = n_1 + n_2 + ... + n_r$ objects of r different types:

- n_1 *objects of type* 1;
- n_2 *objects of type* 2;

  $\vdots$
- n_r *objects of type* r.

Then the number of distinguishable ways to order these objects is

$$\binom{n}{n_1, n_2, n_3, \cdots n_r} = \frac{n!}{n_1! n_2! n_3! \cdots n_r!}.$$

Note: Distinguishable arrangements can be thought of extending the idea of a combination. Suppose you have n objects of two types; where k objects are of the first type and $n-k$ objects are of the second type. The number of distinguishable arrangements of these objects is therefore

the same as the number of k combinations from a set of n. This is because arranging the objects is the same as choosing an unordered set of k "slots" in which to place the objects of the first type.

Sampling without replacement

EXAMPLE 17

A box contains 30 red marbles and 20 blue marbles. If you draw 9 marbles from the box all at once, what is the probability that of those 9 marbles, 7 are red?

Theorem 2.23 (Sampling without replacement) *Suppose you have $n = n_1 + n_2 + \ldots + n_r$ total objects of r different types:*

- *n_1 objects of type 1;*
- *n_2 objects of type 2;*

  $\vdots$

- *n_r objects of type r.*

Suppose you draw $k = k_1 + k_2 + \ldots + k_r$ objects simultaneously. Then, the probability that you draw k_j objects of type j (for each j) is

$$\frac{\binom{n_1}{k_1}\binom{n_2}{k_2}\cdots\binom{n_r}{k_r}}{\binom{n}{k}}.$$

Note: in this setting, drawing objects simultaneously is the same (mathematically) as drawing the objects one at a time without replacement (i.e. without putting back each object you draw before drawing the next object).

QUESTION

What if you draw the objects with replacement (i.e. put each draw back before drawing the next one)? We'll discuss that later.

Hypergeometric random variables

Suppose that there were only two types of objects: r of type 1 and $n-r$ of type 2. Then, if you draw k objects all at once, you can let X be the number of objects of type 1 you draw.

We summarize this in the following definition:

Definition 2.24 *Let $n > 0$, $k \leq n$ and $r \leq n$ be whole numbers. A* **hypergeometric** *random variable with parameters n, r and k is a discrete r.v. X with range $\{0, 1, 2, ..., \min(r, k)\}$ whose density function is*

$$f_X(x) = \frac{\binom{r}{x}\binom{n-r}{k-x}}{\binom{n}{k}}.$$

If X is hypergeometric with parameters n, r and k, we write $X \sim Hyp(n, r, k)$.

A $Hyp(n, r, k)$ r.v. **counts the number of special objects drawn when k objects are drawn at once from a set of n objects, r of which are special**.

Just to make sure the notation is clear, to say

$$\text{“}X \text{ is } Hyp(8, 5, 4)\text{” or “}X \sim Hyp(8, 5, 4)\text{”}$$

means X is a hypergeometric r.v. whose density function is

$$f_X(x) = \frac{\binom{5}{x}\binom{3}{4-x}}{\binom{8}{4}}.$$

Theorem 2.25 (Vandermonde's Identity) *Let $r, n, k \in \mathbb{N}$. Then*

$$\sum_{x=0}^{k} \binom{r}{x}\binom{n-r}{k-x} = \binom{n}{k}.$$

PROOF By the Binomial Theorem,

$$(1+t)^n = \sum_{k=0}^{n} \binom{n}{k} t^k. \tag{2.1}$$

At the same time, (also by the Binomial Theorem),

$$\begin{aligned} (1+t)^n = (1+t)^r(1+t)^{n-r} &= \left[\sum_{x=0}^{r} \binom{r}{x} t^x\right] \cdot \left[\sum_{y=0}^{n-r} \binom{n-r}{y} t^y\right] \\ &= \sum_{x=0}^{r} \sum_{y=0}^{n-r} \binom{r}{x}\binom{n-r}{y} t^{x+y} \end{aligned}$$

Next, we do an *index change* in the second sum: let $k = x + y$, i.e. $y = k - x$. That makes the new index $k = x + y$ go from $x + 0 = x$ to $r + (n - r) = n$. So the double sum above becomes (after the index change)

$$(1+t)^n = \sum_{x=0}^{r} \sum_{k=x}^{n} \binom{r}{x}\binom{n-r}{k-x} t^k = \sum_{k=0}^{n} \left[\sum_{x=0}^{k} \binom{r}{x}\binom{n-r}{k-x}\right] t^k.$$

To match equation (2.1) above, the term inside the bracket must equal $\binom{n}{k}$.

This is Vandermonde's identity. □

Corollary 2.26 *The density function of a hypergeometric r.v. is in fact a density function (its values sum to 1).*

PROOF Take Vandermonde's identity and divide through by $\binom{n}{k}$. □

More examples with combinatorics

EXAMPLE 18

Pick a random number with 5 digits (ex: 00312, 15923, etc.) Assuming every 5 digit string is equally likely,

1. What is the probability that any two digits are the same?
2. What is the probability that exactly two digits are the same?

EXAMPLE 19

Roll seven fair dice. What is the probability you roll 4 sixes, 2 threes and a one?

EXAMPLE 20 (THE COAT CHECK PROBLEM)

Suppose N people leave their coat at a coat check. The coats get jumbled randomly, so when the people leave, they each get a coat at random (that said, no two people get the same coat–each coat goes to one person).

1. What is the probability a specified person gets their coat back?
2. What is the probability n specified people get their coat back?
3. What is the probability at least one person gets their coat back?
4. Suppose there are an infinite number of people (i.e. let $N \to \infty$). What is the probability that no one gets their coat back?

Solution:

1.

2.

3. We apply Generalized Inclusion-Exclusion: let S_n be the event that *some* group of n people get their coat back.

   $$P(S_n) =$$

   Therefore, by Generalized Inclusion-Exclusion,

   $$\begin{aligned} P(\geq 1 \text{ person gets their coat back}) &= P(S_1) - P(S_2) + P(S_3) - P(S_4) + \ldots \\ &= \frac{1}{1!} - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \ldots \\ &= \boxed{\sum_{n=1}^{N} \frac{(-1)^{n+1}}{n!}}. \end{aligned}$$

4. Take the limit on the answer to # 3 as $N \to \infty$ to get $P(\geq 1$ person gets their coat back), which is

   $$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n!} = -\sum_{n=1}^{\infty} \frac{(-1)^n}{n!} = -\left[\sum_{n=0}^{\infty} \frac{(-1)^n}{n!} - 1\right] =$$

   Finally, by the complement rule, P(no one gets their coat back) is

2.4 Bernoulli processes

Definition 2.27 *A* **stochastic process** $\{X_t : t \in \mathcal{I}\}$ *is a collection of random variables indexed by* t*. The set* $\mathcal{I}$ *of values of* t *is called the* **index set** *of the stochastic process.*

Almost always, the index set is $\{0, 1, 2, 3, ...\}$ or $\mathbb{Z}$ (in which case we call the stochastic process a **discrete-time** process and often use n instead of t for the index), or the index set is $[0, \infty)$ or $\mathbb{R}$ (in which case we call the stochastic process a **continuous-time** process).

In MATH 414, we will focus on three stochastic processes which are of fundamental importance (we will learn a lot more about stochastic processes in MATH 416). The first one, called the Bernoulli process, is discussed in this section.

Definition 2.28 *Let* $p \in [0, 1]$*. A* **Bernoulli experiment** *is a probabilistic experiment consisting of a "subexperiment" called a* **trial** *which is repeated over and over again, where the trials have the following properties:*

1. *Each trial has two outcomes,* **success** *and* **failure***.*
2. *On any one trial, the probability of success is* p *(so the probability of failure is* $1 - p$*).*
3. *The result of any one trial is independent of the results of any other trials.*

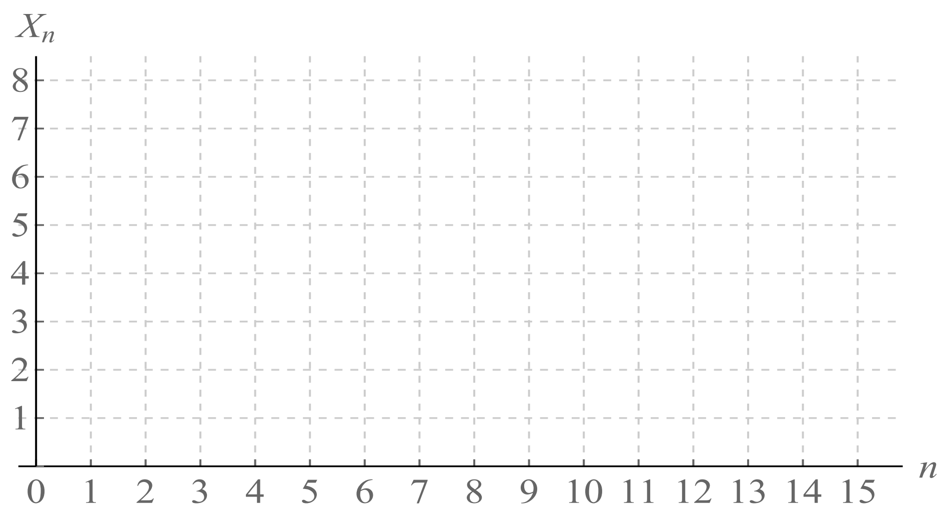
If we let, for $n \in \{0, 1, 2, 3, ...\}$*,* X_n *be the number of successes in the first* n *trials,* $\{X_n : n \in \{0, 1, 2, ...\}\}$ *is a stochastic process called a* **Bernoulli process** *and* p *is called the* **success probability***.*

To picture a Bernoulli process in your mind, think of flipping a coin repeatedly (which flips heads with probability p) and writing down the sequence of heads and tails you get. X_n is the number of heads you flip in the first n flips.

Suppose you flip this coin repeatedly and get the following results:

T H T H T T T H H T T H T T H ...

You can represent the result of this process by the following picture:



Each sequence of dots we get from a sequence of coin flips is called a **sample function** for the process.

Observations about any Bernoulli process $\{X_t\}$

1. $X_0 = 0$ (you can't flip a positive number of heads in zero flips);
2. every time you flip heads, the value of X_n goes up by 1;
3. every time you flip tails, the value of X_n stays the same;
4. X_n never decreases nor jumps by more than 1 unit at a time.

The definition of a Bernoulli process alone is enough to figure out some basic conditional probability questions:

EXAMPLE 21

Let $\{X_n\}$ be a Bernoulli process with success probability p.

1. Compute the probability that $X_8 = 5$, given that $X_6 = 3$.
2. Compute the probability that $X_7 = 3$, given that $X_3 = 3$.

3. Let X_n be a Bernoulli process with success probability p. Find the probability that $X_8 = 2$, given that $X_5 = 1$.

4. In Question 3 of this example, what is really relevant? For example, if I asked you to find the probability that $X_t = b$ given that $X_s = a$, what matters about s, t, a and b?

Binomial random variables

At this point, we want to define a random variable which counts the number of successes in n trials coming from a Bernoulli experiment:

Definition 2.29 *A* **binomial** *random variable with parameters $n \in \mathbb{N}$ and $p \in [0, 1]$ is a discrete r.v. taking values in $\{0, 1, 2, ...n\}$ whose density function is*

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

If X is binomial with parameters n and p, we write $X \sim b(n,p)$ or $X \sim bin(n,p)$.

A binomial r.v. with parameters n and p **counts the number of successes in n trials of a Bernoulli process with success probability p.**

The numbers which occur as values of the density function of binomial r.v.s are commonly encountered in probability. We denote by $b(n, p, k)$ the number

$$b(n,p,k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Theorem 2.30 *The density function of a binomial(n, p) r.v. is a density function (i.e. its values sum to* 1*).*

PROOF Use the binomial theorem:

$$\sum_{x=0}^{n} f_X(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} =$$

How binomial r.v.s relate to Bernoulli processes

Let $\{X_n\}$ be a Bernoulli process with success probability p. Then:

1. For any fixed m and n with $m < n$, $X_n - X_m \sim b(n-m, p)$;
2. For any fixed n, $X_n \sim b(n, p)$;

   NOTE: X_n is a r.v.; $\{X_n\}$ is a process.
3. If $m < n$, $P(X_n = y \mid X_m = x)$ equals the number $b(n-m, p, y-x)$.

Back to sampling with/without replacement

QUESTION

Suppose you have a bag containing 40 marbles, of which 8 are orange. If you draw 20 marbles from the bag, what is the probability that you draw exactly 5 orange marbles?

Solution: It depends on whether you draw the marbles **without replacement** (including if they are all drawn at once) or **with replacement** (i.e. you put each marble back before you draw again).

If the sampling is **without replacement**:

If the sampling is **with replacement**:

EXAMPLE 22

Suppose you guess at every question on a 10-question multiple choice test (four choices per question). What is the probability you get exactly 7 questions correct?

EXAMPLE 23 (CHALLENGE)

Suppose you know 75% of the questions that might be asked on a 10-question exam. If you guess at the other 25% of the questions, what is the probability you get all ten questions correct?

Remark: From *Mathematica*, this sum is $\dfrac{137858491849}{1099511627776} \approx .1253$.

EXAMPLE 24

A fair coin is tossed 11 times (equivalently, 11 fair coins are tossed at once).

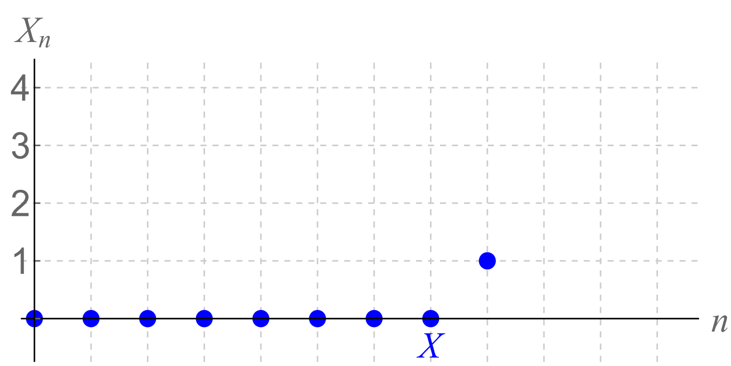
1. What is the probability of flipping exactly 7 heads?
2. What is the probability of at least 8 heads?
3. What is the probability of at least one head?

EXAMPLE 25

A machine produces parts which are defective 1% of the time. Out of 2000 parts produced, what is the probability that exactly 30 parts are defective?

Geometric and negative binomial random variables

We earlier discussed binomial random variables, which describe the height of the graph coming from a Bernoulli process at time n. Now we introduce random variables which describe horizontal measurements on the graph. For example, suppose $\{X_n\}$ is a Bernoulli process with success probability p. Let X be a r.v. which measures the amount of time that passes before the first time the graph of $\{X_n\}$ hits height 1. X is called a **geometric** random variable.



QUESTION

What is the density function of X?

Definition 2.31 *A* **geometric** *random variable with parameter $p \in (0, 1]$ is a discrete r.v. taking values in $\{0, 1, 2, 3, ...\}$ whose density function is*

$$f_X(x) = p(1-p)^x.$$

If X is geometric with parameter p, we write $X \sim Geom(p)$.

A $Geom(p)$ r.v. **counts the number of failures before the first success in a Bernoulli process with success probability** p.

Theorem 2.32 *The density function of a $Geom(p)$ r.v. is a density function (i.e. its values sum to 1).*

PROOF

$$\sum_{x=0}^{\infty} f_X(x) = \sum_{x=0}^{\infty} p(1-p)^x = \qquad \square$$

Theorem 2.33 (Hazard law for geometric r.v.s) *Let $X \sim Geom(p)$. Then for any $n \in \mathbb{N}$,*

$$P(X \geq n) =$$

PROOF

$$P(X \geq n) = \sum_{x=n}^{\infty} f_X(x) = \sum_{x=n}^{\infty} p(1-p)^x = \qquad \square$$

Geometric random variables are exactly the discrete random variables which have an important property called *memorylessness*:

Definition 2.34 *A random variable X is called* **memoryless** *if for all $m, n \geq 0$,*

$$P(X \geq m+n \,|\, X \geq m) = P(X \geq n).$$

To say that a r.v. is memoryless means that if you think of the r.v. as the time it takes for something to happen, if you know you have been waiting for m units, the probability you will wait at least another n units is the same as the probability you would wait at least n units from the get go (in other words, you "forget" that you have already waited m units).

Theorem 2.35 *A random variable X taking values in $\{0, 1, 2, ...\}$ is memoryless if and only if X is geometric.*

PROOF ($\Leftarrow$) Assume $X \sim Geom(p)$.

We will show X is memorylessness by verifying that

$$P(X \geq m + n \,|\, X \geq m) = P(X \geq n).$$

We do this by direct computation:

$$\begin{aligned} P(X \geq m + n \,|\, X \geq m) &= \frac{P(X \geq m + n \bigcap X \geq n)}{P(X \geq m)} \quad \text{(by def'n of cond'l probability)} \\ &= \frac{P(X \geq m + n)}{P(X \geq m)} \\ &= \frac{(1-p)^{m+n}}{(1-p)^m} \quad \text{(by the hazard law)} \\ &= (1-p)^n \\ &= P(X \geq n) \quad \text{(by the hazard law in reverse).} \end{aligned}$$

($\Rightarrow$) Assume X is memoryless and let $p = P(X = 0)$.

By the definition of memorylessness, for all m,

$$P(X \geq m + 1 \,|\, X \geq m) = P(X \geq 1) = 1 - P(X = 0) = 1 - p.$$

Therefore for all $m \geq 0$, we have

$$P(X \geq m + 1) = (1 - p)P(X \geq m). \tag{2.2}$$

Since $p = P(X = 0)$, we know

$$P(X \geq 1) = 1 - P(X = 0) = 1 - p$$

and therefore, by repeatedly applying (2.2), we see

$$\begin{aligned} P(X \geq 2) &= \\ P(X \geq 3) &= \\ &\vdots \\ P(X \geq m) &= (1-p)^m. \end{aligned}$$

Last,

$$\begin{aligned} f_X(x) = P(X = x) &= P(X \geq x) - P(X \geq x+1) \\ &= (1-p)^x - (1-p)^{x+1} \\ &= [1-(1-p)](1-p)^x \\ &= p(1-p)^x \end{aligned}$$

meaning $X \sim Geom(p)$ as wanted. □

Let's now generalize the idea of a geometric random variable. Suppose we wanted to count the number of failures before the r^{th} success in a Bernoulli process, where $r \in \mathbb{N}$. Let X be such a r.v.; what is the density function of X?

Definition 2.36 *A* **negative binomial** *random variable with parameters $r \in \mathbb{N}$ and $p \in [0,1]$ is a discrete r.v. taking values in $\{0, 1, 2, 3, ...\}$ whose density function is*

$$f_X(x) = \binom{x+r-1}{r-1} p^r (1-p)^x =$$

If X is negative binomial with parameters r and p, we write $X \sim NB(r,p)$.

That this function is in fact a density function will not be proven here. It uses the Taylor series expansion of the function $(1-p)^{-x}$. (The "$-$" sign here is why we call this the "negative" binomial r.v.)

A $NB(r,p)$ r.v. **counts the number of failures before the r^{th} success in a Bernoulli process with success probability** p.

Note that a negative binomial r.v. with parameters 1 and p is the same thing as a geometric r.v. with parameter p. (We shorthand this fact by writing "$NB(1,p) \sim Geom(p)$".)

Examples with geometric and negative binomial r.v.s

EXAMPLE 26

Let X be a geometric r.v. so that $P(X \geq 5) = .3$. What is $P(X = 1)$?

EXAMPLE 27

The number of hurricanes that hit Florida in a given year is assumed to be geometric with parameter $.85$. What is the probability that either 3 or 4 hurricanes will hit Florida this year?

EXAMPLE 28

An urn contains 30 red, 20 green and 50 blue marbles. Marbles are drawn from the urn, one at a time with replacement. What is the probability that the fifth time a green marble is drawn is on the 18th draw?

2.5 Summary of Chapter 2

- A discrete random variable is a function $X : \Omega \to \mathbb{R}^d$ taking values in a discrete set (like $\mathbb{N}$ or $\mathbb{Z}$ or $\mathbb{Z}^d$).
- We can completely describe a discrete r.v. X by giving its density function $f_X : Range(X) \to [0, 1]$, which is defined by

$$f_X(x) = P(X = x).$$

  Such a function must take only values between 0 and 1, and its values must sum to 1. The density function of a discrete r.v. is used to compute probabilities by adding its values: if E is any subset of the range of X,

$$P(X \in E) = \sum_{x \in E} f_X(x).$$

- Classes of commonly encountered discrete random variables include the following:
  1. **uniform** r.v.s, which assign equal likelihood to all values in the range of X;
  2. **hypergeometric** r.v.s, which count the number of special objects drawn when a sample is drawn without replacement;
  3. **binomial** r.v.s, which count the number of successes in n trials of a Bernoulli process (and also describe sampling with replacement);
  4. **geometric** r.v.s, which count the number of failures before the first success in a Bernoulli process (and are the only memoryless discrete r.v.s);
  5. **negative binomial** r.v.s, which count the number of failures before the r^{th} success in a Bernoulli process.

  You must know (or be able to refer to on your cheat sheet) the range, density function and other relevant facts about each of these common r.v.s.

- We solve probability questions associated to uniform r.v.s by counting. Techniques used to count sets include inclusion-exclusion, the multiplication principle, permutations, combinations, distinguishable arrangements, and partition formulas.

2.6 Chapter 2 Homework

Exercises from Section 2.2

1. Suppose X is a discrete r.v. with density function f given by

x	-3	-1	0	1	2	3	5	8
$f_X(x)$	$.1$	$.2$	$.15$	$.2$	$.1$	$.15$	$.05$	$.05$

   a) Compute the probability that X is negative.
   b) Compute the probability that X is not positive.
   c) Compute the probability that X is even.
   d) Compute $P(X \in [1, 8])$.
   e) Compute $P(X = -3 \,|\, X \leq 0)$.
   f) Compute $P(X \geq 3 \,|\, X > 0)$.

2. Choose two of (a), (b), (c):

   a) Suppose a box has 12 balls numbered 1 to 12. Two balls are selected from the box independently, with replacement. Let X denote the larger of the two numbers on the selected balls. Compute the density of X.
   b) Suppose you choose a zip code (i.e. a five-digit sequence of numbers) uniformly from all possible zip codes and let X be the number of nonzero digits in the zip code. Calculate the density function of X.
   c) Suppose you uniformly and independently choose three whole numbers from 0 to 9. Let X be the first digit of the number you get when you add these whole numbers together. Calculate the density function of X.

Exercises from Section 2.3

3. (AE) Among a group of 20000 people, 7200 are below age 40, 8200 are childless and 12300 are male. In the same group, there are 5400 males below age 40, 4700 childless persons below age 40 and 6000 childless males. Finally, there are 3100 childless males below age 40. How many people are females above 40 who have children?

4. A 7-person committee, consisting of 3 Democrats, 3 Republicans and 1 Independent, is to be chosen from a group of 20 Democrats, 15 Republicans and 10 Independents. How many different committees are possible?

5. A bus starts with 6 people and stops at 10 different stops. Assuming that each passenger is equally likely to depart at any stop, calculate the probability that the 6 people get off at 6 different stops.

6. My niece's iPhone has 100 songs on it, of which 10 are performed by Taylor Swift. If she sets her iPod to shuffle mode, which will play all 100 songs in a random order (without repeating any songs until they are all played once), what is the probability that the first Taylor Swift song my niece hears is the eighth song played?

7. A **domino** is a rectangular block divided into two equal subrectangles as below, where each subrectangle has a number on it:

x	y

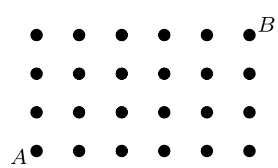
(The numbers x and y might be the same or different.) Since dominos are symmetric, the domino (x, y) is the same as (y, x). How many different domino blocks can be made if the x and y are to be chosen from n different numbers?

Hint: Count the dominos where $x = y$ separately from the dominos where $x \neq y$. Then add these two separate counts.

8. How many distinct arrangements of the letters in each of the following words are possible?

a) COFFEE   b) ASSESS   c) BOOKKEEPER

9. a) Consider the grid of points shown below. Suppose that starting at the point A you move from point to point, moving only one unit to the right or one unit up at a time, ending at the point B. How many different paths from A to B are possible?



b) The above picture gives a 6×4 grid of dots. Answer the same question that was posed in part (a), if the grid is $m \times n$ (i.e. it has n horizontal rows, each containing m dots).

10. How many distinct, non-negative integer-valued vectors $(x_1, x_2, ..., x_5)$ satisfy $x_1 + x_2 + x_3 + x_4 + x_5 = 12$?

Hint: This has something to do with distinguishable arrangements, and might have something to do with Problem 9, depending on how you think about it.

In Problems 11-15, you are to give both a formula for the answer in terms of standard combinatorial notation, and a decimal approximation of your answer.

11. Suppose you deal a five-card hand from a standard deck of cards. Compute the probability of being dealt each of the following hands:

    a) A royal flush (the A,K,Q,J and 10 of the same suit)

    b) A flush (any five cards of the same suit)

    c) Three-of-a-kind, but not a full house or four-of-a-kind

    d) A straight (five cards in a sequence, regardless of suit)
    Note: An ace may be the highest card (10-J-Q-K-A) or lowest card (A-2-3-4-5) in a straight, but a sequence like K-A-2-3-4 is not a straight because the ace is in the middle.

    e) A hand which contains no pair (nor three- nor four-of-a-kind)

12. In Texas Hold'Em, each player is dealt 2 cards from a standard deck.

    a) What is the probability that a Texas Hold'Em player is dealt a pair?

    b) What is the probability that a Texas Hold'Em player's hand is a "Broadway" hand (i.e. both cards are 10 or higher)?

    c) What is the probability that a Texas Hold'Em player is dealt "suited connectors", meaning that the cards are of the same suit and adjoining rank (like (A-2) or (8-9) or (10-J) or (K-A))?

13. In the card game Bridge, each player is dealt 13 cards from a standard deck.

    a) A **Yarborough** is a (terrible) Bridge hand that contains no card higher than a 10 (i.e. no jacks through aces). Compute the probability that a Bridge hand is a Yarborough.

    b) A Bridge hand is said to have a **void** if there is at least one suit for which the hand has no cards in that suit. Compute the probability that a Bridge hand has exactly one void.

14. In the card game Shanghai Rummy, two 54-card decks (each including the standard 52 cards and 2 jokers) are shuffled together. Then, each player is dealt a 12-card hand. What is the probability that a Shanghai Rummy hand contains at least one joker?

15. **Set** is a card game played with a deck of 81 different cards. Unlike normal playing cards, which have two attributes (a suit and a rank), each card in a Set deck has four attributes: a color (one of red, green, or purple), a shape (one of diamonds, ovals or waves), a number (1, 2 or 3), and a pattern (solid, striped, or open).

    a) If you choose five cards randomly from a Set deck, what is the probability that your hand is a color flush (meaning all five cards are of the same color)?

b) If you choose five cards randomly from a Set deck, what is the probability that your hand is a color and shape flush (meaning all five cards are of the same color and shape)?

c) If you choose five cards randomly from a Set deck, what is the probability that your hand is a color and shape and pattern flush (meaning all five cards are of the same color, shape and pattern)?

d) If you choose five cards randomly from a Set deck, what is the probability that your hand is a flush with respect to any two attributes?

e) If you choose five cards randomly from a Set deck, what is the probability that your hand is a flush with respect to at least one attribute?

Hint: Use Inclusion-Exclusion, together with previous parts of this problem.

Exercises from Section 2.4

16. A fair die is rolled 12 times (independently). Compute the probability of rolling exactly 2 sixes, and the probability of rolling at most 2 sixes.

17. (AE) Experience shows that 20% of the people reserving tables at a certain restaurant never show up. If the restaurant has 50 tables and takes 52 reservations, what is the probability that it will be able to accommodate everyone who shows up?

18. A circular target of radius 1 is divided into four annular zones (an "annular" shape is like a ring) of outer radii $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$ and 1, respectively. Suppose 10 shots are fired at the target independently, and that each shot hits a random point in the target chosen uniformly.

    a) Compute the probability that exactly four shots land in the region of radius $1/4$.

    b) What is the probability that at most three shots land in the zone bounded on the inside by the circle of radius $1/2$ and on the outside by the circle of radius $3/4$?

    c) If exactly 5 shots land inside the circle of radius $1/2$, determine the probability that at least one shot lands inside the circle of radius $1/4$.

19. (AE) You own a business that gets bolts from two bolt manufacturers: A and B (you get 70% of your bolts from A and 30% from B). Suppose that 5% of all bolts from manufacturer A are defective, and that 20% of all bolts from manufacturer B are defective. You get a shipment of 12 bolts from one of the two manufacturers. If exactly 3 of the 12 bolts are defective, what is the probability that the shipment came from manufacturer B?

20. There are 40 gumballs in a bag, of which 20 are red, 10 are orange, 8 are green, and 2 are purple.

    a) If you draw 10 gumballs from the bag without replacement, what is the probability that you draw 5 red, 3 orange, and 2 purple gumballs?

    b) If you draw 7 gumballs from the bag without replacement, what is the probability that you draw exactly 4 green gumballs?

    c) If you draw 7 gumballs from the bag with replacement, what is the probability that you draw exactly 4 green gumballs?

    d) If you draw 6 gumballs from the bag without replacement, what is the probability you draw at least 5 orange gumballs?

    e) If you draw 10 gumballs from the bag with replacement, what is the probability that you draw 3 orange gumballs?

21. Continuing with the same bag of gumballs as in the previous problem:

    a) If you draw 15 gumballs from the bag without replacement and take a bite out of them, then put them back in the bag, and if you subsequently draw 5 gumballs from the bag with replacement, what is the probability that you drew 3 gumballs that you bit?

    b) Suppose you draw gumballs from the bag repeatedly, with replacement. What is the probability that the first time you draw a purple gumball is on the 9th draw?

    c) Suppose you draw gumballs from the bag repeatedly, with replacement. What is the probability that the fifth time you draw a red gumball is on the 14th draw?

    d) Suppose you draw gumballs from the bag two at a time, putting each group back after you draw it. What is the probability that the first time you draw 2 red gumballs (on a single draw) is the 4th time you draw 2 gumballs from the bag?

    e) Divide the 40 gumballs randomly into four disjoint groups of 10. What is the probability that the first and second groups have the same number of green gumballs?

22. Suppose $X \sim Geom(.8)$. Compute the following:

    a) $P(X > 3)$

    b) $P(4 \leq X \leq 7 \text{ or } X > 9)$

    c) $P(X \leq 2 \,|\, X \leq 3)$

    d) $P(X \geq 85 \,|\, X \geq 80)$

Calculus review

23. Evaluate each integral:

a) $\int e^x \, dx$

b) $\int e^{-x} \, dx$

c) $\int e^{2x} \, dx$

d) $\int e^{x/5} \, dx$

e) $\int e^{-(3/8)x} \, dx$

f) $\int e^{(5x-3)/4} \, dx$

24. Based on your answers to Exercise 23, what is $\int e^{rx} \, dx$? (There are two cases, depending on whether or not $r = 0$.)

25. Based on your answer to Exercise 24, evaluate the following integrals, simplifying your answer as much as possible. Try to do them quickly, i.e. without writing a u-substitution.

a) $\int 3e^{-4x} \, dx$

b) $\int_0^1 \frac{1}{2} e^{-(2/3)x} \, dx$

c) $\int_2^5 2e^{4x} \, dx$

d) $\int_4^{12} \frac{e^{-x/4}}{8} \, dx$

e) $\int_5^7 ae^{-bx} \, dx$
(in (e), assume $b \neq 0$)

26. Evaluate each improper integral:

a) $\int_1^\infty 3e^{-x} \, dx$.

b) $\int_5^\infty 2e^{-x/4} \, dx$

c) $\int_a^\infty re^{-sx} \, dx$ (assume in this problem that $s \neq 0$)

27. Determine the value of c so that $\int_0^\infty 4e^{-cx} \, dx = 1$.

Chapter 3

Continuous random variables

3.1 Density functions of continuous random variables

RECALL

A r.v. X is a function $X : \Omega \to \mathbb{R}^d$, where $(\Omega, \mathcal{A}, P)$ is a probability space.

In Chapter 2, we studied discrete r.v.s, meaning those whose range is finite or countable. Now, we will study non-discrete r.v.s. First, a definition:

Definition 3.1 *A r.v. $X : \Omega \to \mathbb{R}^d$ is called* **continuous (cts)** *if, for every $x \in \mathbb{R}^d$, we have*

$$P(X = x) = 0.$$

Definition 3.2 *A r.v. $X : \Omega \to \mathbb{R}^d$ is called* **mixed** *if if it neither discrete nor continuous.*

EXAMPLE 1

Pick a number uniformly from $[0, 3]$ and let X be the result.

Recall

To describe a discrete real-valued r.v., we write down a ______________________ for that r.v. This object tells us two things:

1.

2.

Question

What is the analogue of this for a cts r.v.?

Bad news: Unfortunately, we can't accomplish both (1) and (2) above when X is cts:

Definition 3.3 *Let $X : \Omega \to \mathbb{R}$ be a cts r.v. We say X has a* **density function** f_X *(equivalently, f_X is a* **density function** *for X) if $f_X : \mathbb{R} \to [0, \infty)$ satisfies, for any real numbers $a \leq b$,*

$$P(X \in [a, b]) = \int_a^b f_X(x)\,dx.$$

Example 1, continued

What is a density function for the uniform r.v. on $[0, 3]$?

Theorem 3.4 *A function $f : \mathbb{R} \to \mathbb{R}$ is the density function of a cts r.v. $X : \Omega \to \mathbb{R}$ if and only if all of the following hold:*

1. *f is measurable (meaning you can compute $\int_a^b f(x)\,dx$ for every a and b);*
2. *$f(x) \geq 0$ for all x;*
3. *$\int_{-\infty}^{\infty} f(x)\,dx = 1$.*

EXAMPLE 2

Suppose X is a continuous r.v. whose density function is

$$f_X(x) = \begin{cases} cx & \text{if } 0 \leq x \leq 3 \\ 0 & \text{else} \end{cases}$$

for some constant c.

1. What is the range of X?
2. What is the value of c?
3. Find $P(X \leq 1)$.
4. Find $P(X \geq 2)$.
5. Find $P(X > 2)$.
6. Which is more likely, that $X = 1$ or $X = 2$?
7. Which is more likely, that X is close to 1 or X is close to 2?

Key idea: If you want to do any probabilistic calculations related to a continuous r.v., all you need to be given (or all you need to figure out) is the density function of that r.v. This is because if you are given any set $E \subseteq \mathbb{R}^d$,

$$P(X \in E) = \int_E f_X(x)\, dx,$$

so long as X is continuous.

Contrast this with how you compute probabilities for discrete r.v.s:

	DISCRETE R.V.S	CONTINUOUS R.V.S
How the density function is defined		
How probabilities are computed using the density		

Bad news: There are continuous r.v.s that do not have a density function.
Good news: You would not encounter these r.v.s in any normal situation.

Uniform continuous r.v.s

The most common type of continuous r.v. is where you choose from a set where all subsets of the same size have equal probability This is called a *uniform* r.v.:

Definition 3.5 *Let $\Omega \subseteq \mathbb{R}$ be a union of intervals whose total length is finite. A* **(continuous) uniform** *random variable on Ω is the cts r.v. X with density function*

$$f_X(x) = \begin{cases} \dfrac{1}{\textit{total length}(\Omega)} & \textit{if } x \in \Omega \\ 0 & \textit{else} \end{cases}$$

If X is uniform on a single interval $[a, b] \subseteq \mathbb{R}$, we write $X \sim Unif([a, b])$.

Example 1 describes a uniform r.v. on $[0, 3]$, for instance.

EXAMPLE 3

Describe a density function for the uniform r.v. on $[0,4] \cup [10,11) \cup \{13\}$.

Remark: The density function of a cts r.v. is never unique – it can be altered on any finite or countable set without affecting any probability computations.

EXAMPLE 4

Find a density function for X, if $X \sim Unif(\left[0, \frac{1}{2}\right])$.

Remark: Unlike density functions for discrete r.v.s, density functions for cts r.v.s can take values greater than 1.

3.2 Distribution functions

In this section, we address two questions:

1. How do we represent a r.v. which is mixed (neither discrete nor cts)?
2. Is there an object which describes r.v.s, which unifies the theory of discrete, cts and mixed r.v.s?

The answer to these questions is given in the following definition. For now, we'll stick to real-valued r.v.s (and discuss vector-valued r.v.s later).

Definition 3.6 *Let $X : \Omega \to \mathbb{R}$ be a r.v. The* **cumulative distribution function** *(a.k.a.* **distribution function** *a.k.a.* **cdf***) of X is the function $F_X : \mathbb{R} \to \mathbb{R}$ defined by*

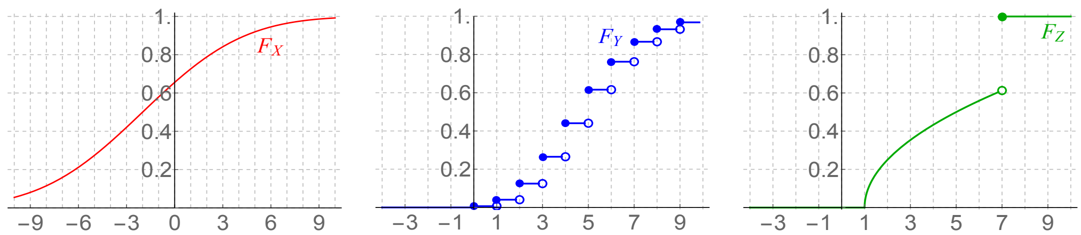
$$F_X(x) = P(X \leq x).$$

EXAMPLE 5

What is the cdf for the uniform r.v. on $[0, 4]$?

EXAMPLE 6

Shown below are graphs of the cdfs for three r.v.s X, Y and Z. What can you tell about X, Y and Z from these graphs? What are the commonalities across these three graphs?

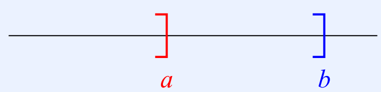


Theorem 3.7 (Properties of distribution functions) *Let $X : \Omega \to \mathbb{R}$ be a r.v. whose cdf is F_X. Then:*

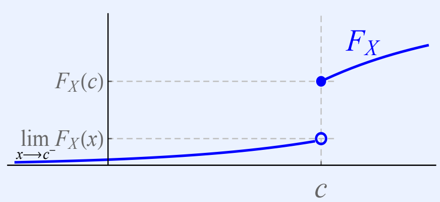
1. *F_X is the only cdf of X;*
2. *F_X is nondecreasing;*
3. $\lim_{x\to-\infty} F_X(x) = 0$;
4. $\lim_{x\to\infty} F_X(x) = 1$;
5. *If $Range(X) \subseteq (a, b)$, then $F_X(x) = 0$ for all $x \le a$;*
6. *If $Range(X) \subseteq (a, b)$, then $F_X(x) = 1$ for all $x \ge b$;*
7. *F_X is right-continuous everywhere*
   (meaning $\lim_{x\to c^+} F_X(x) = F_X(c)$ for all c).

Theorem 3.8 (Calculating probabilities from distribution functions) *Let $X : \Omega \to \mathbb{R}$ be a r.v. whose cdf is F_X. Then:*

1. *$P(X \in (a, b]) = F_X(b) - F_X(a)$ for all $a < b$.*



2. $P(X = c) = F_X(c) - \lim_{x\to c^-} F_X(x)$
   (this is the size of the jump in F_X at c).



3. *$P(X = c) = 0$ if and only if F_X is continuous at c.*

The next theorem generalizes what we observed in Example 6:

Theorem 3.9 *Let $X : \Omega \to \mathbb{R}$ be a r.v. whose cdf is F_X. Then:*

1. *X is cts if and only if F_X is a continuous function;*
2. *X is discrete if and only if F_X is piecewise constant.*

EXAMPLE 7

Suppose X is a real-valued r.v. that has distribution function

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{1}{4}\sqrt{x} & \text{if } x \in (0,1) \\ \frac{1}{2}x & \text{if } x \in [1,2) \\ 1 & \text{if } x \geq 2 \end{cases}$$

Compute each probability:

1. $P(X = x)$ (do this for every real number x)
2. $P(X \leq 1)$
3. $P(X < 1)$
4. $P(X \geq 1)$
5. $P(X > 1)$
6. $P(\frac{1}{2} \leq X < \frac{3}{2})$

Theorem 3.10 (Relationship between density and dist. functions) *Suppose that $X : \Omega \to \mathbb{R}$ is a cts r.v. with density function f_X. Then:*

1. $\dfrac{d}{dx}\left(F_X(x)\right) = f_X(x)$*; and*
2. $\displaystyle\int_{-\infty}^{x} f_X(t)\,dt = F_X(x)$.

PROOF Statement (2) follows from definitions of density and distribution functions:

$$\int_{-\infty}^{x} f_X(t)\,dt = P(X \in (-\infty, x]) = F_X(x).$$

Statement (1) follows from (2) and the Fundamental Theorem of Calculus:

$$\frac{d}{dx}F_X(x) = \frac{d}{dx}\left[\int_{-\infty}^{x} f_X(t)\,dt\right] = f_X(x). \ \Box$$

EXAMPLE 8

Suppose X is a cts. r.v. whose distribution function is

$$F_X(x) = \begin{cases} 0 & x \le 0 \\ \sin x & 0 < x \le \frac{\pi}{2} \\ 1 & x > \frac{\pi}{2} \end{cases}.$$

1. Find a density function of X.
2. Compute $P\left(X < \frac{\pi}{6}\right)$ using the cdf of X.
3. Compute $P\left(X < \frac{\pi}{6}\right)$ using a density function of X.

Survival functions

Definition 3.11 *Let X be a real-valued r.v. The* **survival function** *of X is the function $S_X(x) = P(X > x) = 1 - F_X(x)$.*

Note: if X is cts, then $S_X(x) = P(X \ge x)$ as well.

EXAMPLE 9

Compute the survival function of X, if $X \sim Unif([0, 8])$.

3.3 Transformations of random variables

Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a function and let X be a real-valued r.v. (By the way, φ is "phi".) Then $Y = \varphi(X)$ is a r.v. which is called a **transformation** of X. The object of this section is to compute the density function of a transformation when you are given a density function of the original r.v.

When X is discrete

In this situation, $Y = \varphi(X)$ must also be discrete. To compute the density function of Y, first determine the range of Y. Then, for y belonging to the range of Y, start with the definitions as follows:

$$f_Y(y) = P(Y = y) = P(\varphi(X) = y)$$

and then solve the equation inside the parentheses for X. Then use a density function of X to compute probabilities.

EXAMPLE 10

Suppose $X \sim Unif(\{-2, -1, 0, 1, 2\})$. Let $Y = X^4$. Find a density function of Y.

Remark: Once you have a density function of Y, you can compute any probability associated to Y by adding values of f_Y (or integrating f_Y, if Y is continuous).

When X is continuous

In this situation, $Y = \varphi(X)$ could be discrete, continuous or neither. Since you don't even know that Y has a density function, the best way to proceed is to **find the distribution function of Y first**. First, determine the range of Y. If this range is $[a, b]$ or (a, b), you know that

and

Next, let y be in the range of Y. By the definition of cdf, we get

$$F_Y(y) = P(Y \leq y) = P(\varphi(X) \leq y).$$

Solve the inequality $\varphi(X) \leq y$ for X (this may involve multiple cases) and use either the density or distribution function of X to obtain the cdf of Y. Finally, differentiate F_Y to obtain f_Y.

EXAMPLE 11

Let X be uniform on $[0, 2]$ and let $Y = X^3$. Compute a density function of Y.

EXAMPLE 12

Suppose that an insurance company has to make two kinds of annual payments, "direct" and "indirect". If X is the size of the direct payment and Y is the size of the indirect payment the company has to make, assume that (X, Y) is modeled by a uniform r.v. on the unit square (this is the square whose vertices are $(0,0), (1,0), (0,1)$ and $(1,1)$). Determine a density function of the total annual payment the insurance company has to make.

EXAMPLE 13

Choose a point (X, Y) uniformly from the rectangle whose vertices are the four points $(1, 0), (1, 1), (4, 0)$ and $(4, 1)$. Let $Z = Y/X$; compute f_Z.

EXAMPLE 14

You and your friend decide to meet at the library to study math. Each of you choose a random time (uniformly and independently) to arrive at the library between 6 and 7 PM. What is the density function of the length of time the first person to arrive has to wait for the second person to arrive?

Definition 3.12 *A continuous, real-valued r.v.* Y *is called* **Cauchy** *if* $Y = \tan X$, *for* X *uniform on* $\left[\frac{-\pi}{2}, \frac{\pi}{2}\right]$.

The Cauchy r.v. **measures the slope of an angle which is uniformly chosen**, because $\tan\theta$ is the slope of a line at angle θ to the horizontal.

EXAMPLE 15

Compute a density function of the Cauchy r.v.

Solution: First, notice that since $X \sim Unif([\frac{-\pi}{2}, \frac{\pi}{2}])$,

$$f_X(x) = \frac{1}{\frac{\pi}{2} - \frac{-\pi}{2}} = \frac{1}{\pi}$$

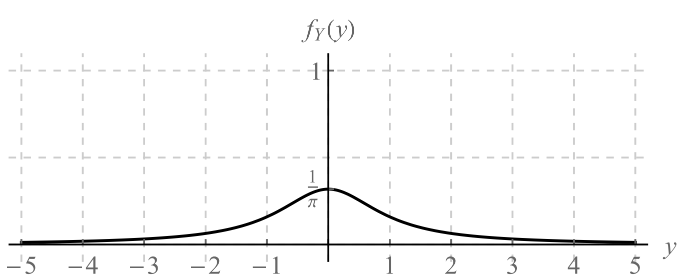
for $x \in [\frac{-\pi}{2}, \frac{\pi}{2}]$ (and $f_X(x) = 0$ otherwise).

Now, let $Y = \tan X$; the range of Y is $\mathbb{R}$. For any $y \in \mathbb{R}$,

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(\tan X \le y) \\ &= P(X \le \arctan y) \\ &= \int_{\pi/2}^{\arctan y} f_X(x)\,dx \\ &= \int_{\pi/2}^{\arctan y} \frac{1}{\pi}\,dx \\ &= \frac{1}{\pi}\left(\arctan y - \frac{\pi}{2}\right) \\ &= \frac{1}{\pi}\arctan y - \frac{1}{2}. \end{aligned}$$

Therefore a density function for Y is

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}\left[\frac{1}{\pi}\arctan y - \frac{1}{2}\right] =$$



3.4 Poisson processes

In the last chapter, we discussed Bernoulli processes, which count the number of *successes* occurring when time is kept track of discretely (i.e. in terms of the number of trials that have been performed). In this section we describe a second important type of process, which can be thought of as keeping track of the number of "successes" called *births* occurring when time is kept track of continuously (i.e. in terms of elapsed physical time). Such a process is called a *Poisson process*:

Definition 3.13 *Suppose "births" are occurring at random times in $[0, \infty)$ according to the following three rules:*

No simultaneous births: *the probability of two births happening at the same time is zero;*

Time homogeneity: *the number of births happening in any interval of time depends only on the length of that interval (and not on the starting point or endpoint of that interval); and*

Independent increments: *the number of births occurring on any collection of disjoint intervals are mutually independent of one another.*

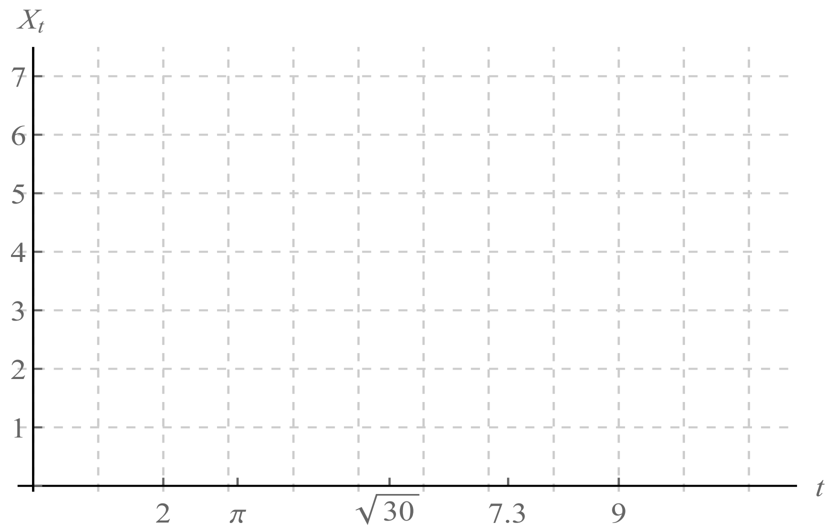
In this setting, if we define X_t to be the number of births in time interval $[0, t]$, we obtain a continuous-time stochastic process $\{X_t : t \in [0, \infty)\}$ called a **Poisson process**.

Things from the real world modeled by Poisson processes include:

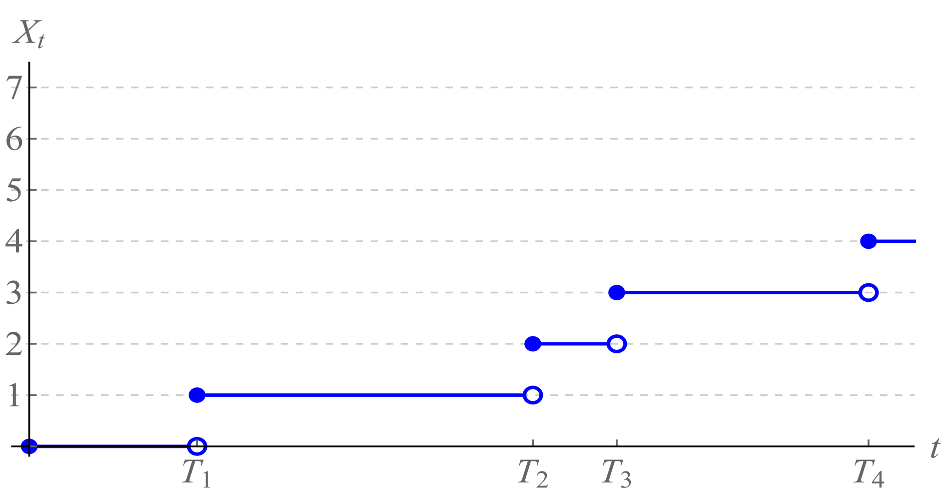
- births of new individuals in a population;
- arrivals of customers to a service center;
- times of radioactive emissions;
- times when a cell phone receives a text message;
- times when an earthquake hits the San Andreas Fault;
- times at which insurance companies acquire new customers;
- times at which insurance companies' customers file claims;
- times when an error occurs during a transmission.

In all these situations, each time when one of these things occurs is the time of a "birth".

To get a picture of what a Poisson process "looks like", suppose births happen at times 2, π, $\sqrt{30}$, 7.3, 9, ... If we graph X_t against t, we get this sample function:



More generally, suppose the times of births are (in increasing order) $T_1, T_2, T_3, \ldots$. This produces the following picture of a sample function, from which we can define random variables associated to the Poisson process:



Definition 3.14 *Let $\{X_t\}$ be a Poisson process. For $j = 1, 2, 3, ...$, define the following r.v.s:*

$$\begin{aligned} T_j &= \min\{t : X_t = j\} \\ &= \textit{the } j^{th} \textit{ smallest time at which a birth occurs (set } T_0 = 0) \\ W_j = T_j - T_{j-1} &= \textit{the } j^{th} \textbf{ waiting time} \\ &\quad \textit{(the time between the } (j-1)^{st} \textit{ and } j^{th} \textit{ births)} \end{aligned}$$

Note the parallels between these r.v.s and the r.v.s arising from a Bernoulli process:

	Bernoulli process	Poisson process
time measurement	discrete ($t \in \mathbb{N}$)	continuous ($t \in [0, \infty)$)
parameter	success probability p	
distribution of X_t	binomial(t, p)	
$W \sim$ time to first success/birth	$Geom(p)$ (memoryless)	
$T_r \sim$ time to r^{th} success/birth	$NB(r, p)$	

Exponential random variables

Our goal is to determine the density function for each of the r.v.s associated to a Poisson process. We start with the distribution of the waiting times W_j:

Quick observations about waiting times:

1. $T_j = W_1 + W_2 + W_3 + ... + W_j$.
2. If $i \neq j$, the values of $W_i \perp W_j$ (follows from independent increments).
3. For any j, the density function of W_j is the same as the density function of any other W_i, hence the same as the density function of W_1 (follows from time homogeneity). So we can call each of the waiting times W.
4. W is continuous (follows from time homogeneity).
5. W is memoryless (see next page).

Lemma 3.15 *If W is the waiting time between births in a Poisson process, then W is memoryless, meaning that for all $m, n \geq 0$,*

$$P(W \geq m+n \,|\, W \geq m) = P(W \geq n).$$

PROOF The important observation to prove this is that in a Poisson process, $W \geq w$ means there are no births in a time interval of length w. The rest of this proof is a calculation based on this observation:

$$\begin{aligned}
P(W \geq n) &= P(\text{no births take place in the time interval } [0,n)) \\
&\qquad \text{(since waiting time to first birth is at least } n) \\
&= P(\text{no births take place in the time interval } [m, m+n)) \\
&\qquad \text{(by time homogeneity)} \\
&= P\left(\begin{array}{c|c} \text{no births take place} & \text{no births take place} \\ \text{in the time interval} & \text{in the time interval} \\ [m, m+n) & [0, m) \end{array} \right) \\
&\qquad \text{(by the independent increment property)} \\
&= \frac{P(\text{no births in } [0,m) \bigcap \text{no births in } [m, m+n))}{P(\text{no births in } [0,m))} \\
&\qquad \text{(by definition of conditional probability)} \\
&= \frac{P(\text{no births in } [0, m+n))}{P(\text{no births in } [0,m))} \\
&= \frac{P(W \geq m+n)}{P(W \geq m)} \\
&= \frac{P(W \geq m+n \bigcap W \geq m)}{P(W \geq m)} \\
&= P(W \geq m+n \,|\, W \geq m). \;\square
\end{aligned}$$

RECALL

If X is discrete, we showed that any memoryless r.v. X must be ____________________.
The waiting time W in a Poisson process is memoryless, but is continuous. To classify it, we use the following theorem:

Theorem 3.16 *Let X be a continuous r.v. taking values in $[0, \infty)$ which is memoryless. Then X has density function*

$$f_X(x) =$$

PROOF First, let F_X be the cdf of X and consider the survival function

$$S_X(x) = 1 - F_X(x) = P(X > x) = P(X \geq x).$$

Note $S_X(x) \in (0, 1)$ so $-\ln S_X(1) > 0$.
We can then let $\lambda = -\ln S_X(1)$, which means $S_X(1) = e^{-\lambda}$.

Since X is memoryless,

$$\frac{P(X \geq m+n)}{P(X \geq m)} = P(X \geq n) \quad \Rightarrow \quad P(X \geq m+n) = P(X \geq m)P(X \geq n)$$
$$\Rightarrow$$

So for any positive integer m,

$$S_X(m) = S_X(1+1+...+1) = S_X(1)S_X(1)\cdots S_X(1) = [S_X(1)]^m = e^{-\lambda m}.$$

Now for any positive rational number $\frac{m}{n}$,

$$S_X(m) = S_X\left(\frac{m}{n} + ... + \frac{m}{n}\right) = S_X\left(\frac{m}{n}\right) S_X\left(\frac{m}{n}\right) \cdots S_X\left(\frac{m}{n}\right) = \left[S_X\left(\frac{m}{n}\right)\right]^n$$

so by taking n^{th} roots of both sides of the above equation we get

$$S_X\left(\frac{m}{n}\right) = \sqrt[n]{S_X(m)} = \sqrt[n]{S_X(1)^m} = [S_X(1)]^{m/n} = e^{-\lambda(m/n)}.$$

Since $S_X(\frac{m}{n}) = e^{-\lambda(m/n)}$ for all rational numbers m/n, and since S_X is continuous (because X is cts by hypothesis), it must be that for all **real** numbers x, $S_X(x) = e^{-\lambda x}$. Thus

$$F_X(x) = 1 - S_X(x) = 1 - e^{-\lambda x}$$

and

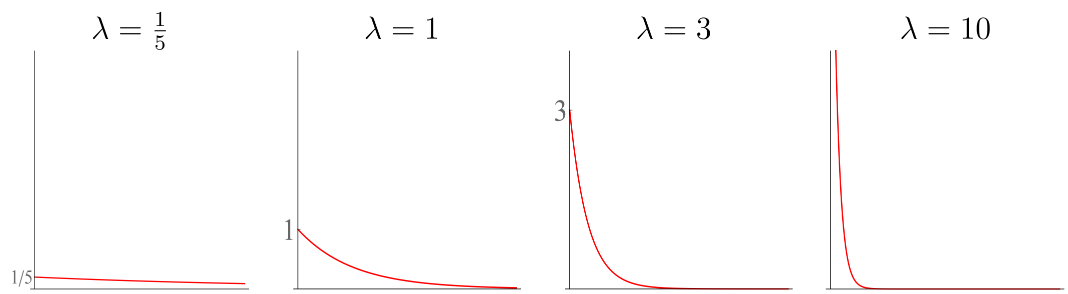
$$f_X(x) =$$

Definition 3.17 *An* **exponential** *r.v. X with parameter $\lambda \in (0, \infty)$ is a continuous r.v. whose density function is*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \textit{else} \end{cases}.$$

If X is exponential with parameter λ, we write $X \sim Exp(\lambda)$.

Here are some plots of density functions of $Exp(\lambda)$ r.v.s for various λ:



Thus if X is exponential, we are more likely to get smaller values for X if λ is large, and more likely to get larger values for X if λ is small.

Theorem 3.18 (Properties of exponential r.v.s) *Let X be a real-valued r.v. The following statements are equivalent:*

1. *$X \sim Exp(\lambda)$.*
2. *X is memoryless and continuous.*
3. *The cdf of X is $F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$.*
4. *The survival function of X is $S_X(x) = e^{-\lambda x}$.*
5. *X models the time between births in a Poisson process.*

Corollary 3.19 (Waiting times are exponential) *Let $\{X_t\}$ be a Poisson process. Then there is a number $\lambda > 0$, called the* **rate** *or* **birth rate** *of the process, such that the waiting times between each births are $Exp(\lambda)$.*

An exponential r.v. with parameter λ **gives the waiting time between births in a Poisson process with rate** λ.

Poisson random variables

Now, we turn our attention to figuring out the density of X_t (for a fixed t). Notice first that each X_t is discrete because it **counts** the number of births in $[0,t]$. Furthermore,

$$P(X_t = x) = P(\text{exactly } x \text{ births in the time interval } [0,t]).$$

Take the time interval $[0,t]$ and divide it into n equal-length subintervals:

The length of each subinterval is

and therefore the probability of no birth in each subinterval is

so the probability of at least one birth in each subinterval is

Now, if n is large enough, then these subintervals will be very, very small, so by the property of no simultaneous births, we will not have more than one birth any any of these subintervals. So for large enough n, each subinterval will have

one birth (with probability $1 - e^{-\lambda t/n}$)
or
zero births (with probability $e^{-\lambda t/n}$).

That means we can think of each subinterval as being a trial of a Bernoulli experiment (where a "success" means that the subinterval contains a birth), and therefore

$$\begin{aligned} P(X_t = x) &= P(\text{exactly } x \text{ births in the time interval } [0,t]) \\ &= P(x \text{ successes in } n \text{ trials}) \\ &= b(n, 1 - e^{-\lambda t/n}, x) \end{aligned}$$

Of course, this only works **if n is large enough**. How large is large enough? Well, ∞ is definitely large enough, so we conclude

$$P(X_t = x) = \lim_{n \to \infty} b(n, 1 - e^{-\lambda t/n}, x).$$

and we work out this limit on the next page.

From the previous page,

$$
\begin{aligned}
P(X_t = x) &= \lim_{n\to\infty} b\left(n, 1-e^{-\lambda t/n}, x\right) \\
&= \lim_{n\to\infty} \binom{n}{x} \left(1-e^{-\lambda t/n}\right)^x \left(e^{-\lambda t/n}\right)^{n-x} \\
&= \lim_{n\to\infty} \frac{n!}{x!(n-x)!} \left(1-e^{-\lambda t/n}\right)^x e^{-\lambda t(n-x)/n} \\
&= \frac{1}{x!} \lim_{n\to\infty} \frac{n!}{(n-x)!} \left(1-e^{-\lambda t/n}\right)^x \exp\left[-\lambda t\left(\frac{n-x}{n}\right)\right] \\
&= \frac{1}{x!} \lim_{n\to\infty} \frac{n!}{(n-x)!} \cdot \frac{n^x}{n^x} \left(1-e^{-\lambda t/n}\right)^x \exp\left[-\lambda t\left(1-\frac{x}{n}\right)\right] \\
&= \frac{1}{x!} \exp\left[-\lambda t(1)\right] \lim_{n\to\infty} \frac{n(n-1)(n-2)\cdots(n-x+1)}{n^x} \cdot n^x \left(1-e^{-\lambda t/n}\right)^x \\
&= \frac{1}{x!} \exp\left[-\lambda t(1)\right] \lim_{n\to\infty} \frac{n^x + \text{ smaller powers of } n}{n^x} \cdot n^x \left(1-e^{-\lambda t/n}\right)^x \\
&= \frac{e^{-\lambda t}}{x!} \lim_{n\to\infty} \left(1+n^{\text{negative powers}}\right) \left[n\left(1-e^{-\lambda t/n}\right)\right]^x \\
&= \frac{e^{-\lambda t}}{x!} (1+0) \lim_{n\to\infty} \left[n\left(1-e^{-\lambda t/n}\right)\right]^x \\
&= \frac{e^{-\lambda t}}{x!} \lim_{n\to\infty} \left[n\left(1-e^{-\lambda t/n}\right)\right]^x \\
&= \frac{e^{-\lambda t}}{x!} \left[\lim_{n\to\infty} \frac{1-e^{-\lambda t/n}}{\frac{1}{n}}\right]^x \\
&\overset{L}{=} \frac{e^{-\lambda t}}{x!} \left[\lim_{n\to\infty} \frac{e^{-\lambda t/n}\left(\frac{\lambda t}{n^2}\right)}{\frac{-1}{n^2}}\right]^x \\
&= \frac{e^{-\lambda t}}{x!} \left[\lim_{n\to\infty} \lambda t\, e^{-\lambda t/n}\right]^x \\
&= \frac{e^{-\lambda t}}{x!} \left[\lambda t\left(e^0\right)\right]^x \\
&= \boxed{\frac{(\lambda t)^x e^{-\lambda t}}{x!}}.
\end{aligned}
$$

Definition 3.20 *Let $\lambda \in (0, \infty)$. A* **Poisson** *r.v., denoted $Pois(\lambda)$, is a discrete r.v. taking values $\{0, 1, 2, 3, ...\}$ whose density function is*

$$f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

λ is called the **parameter** *of the Poisson r.v.*

Theorem 3.21 *The density function of a Poisson r.v. is in fact a density function (its values sum to 1).*

PROOF Apply the formula for the Taylor series of e^λ:

$$\sum_{x=0}^{\infty} f_X(x) = \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1. \ \square$$

Theorem 3.22 *Let $\{X_t : t \in [0, \infty)\}$ be a Poisson process with rate λ. Then for each t, $X_t \sim Pois(\lambda t)$.*

A Poisson r.v. with parameter λ **counts the number of events taking place in a Poisson process with rate λ over any one unit of time**.
A Poisson r.v. with parameter λt **counts the number of events taking place in a Poisson process with rate λ over any time period of length** t.

There is a relationship between binomial and Poisson r.v.s:

Theorem 3.23 (Law of Small Numbers (LSN)) $\lim\limits_{n\to\infty} b(n, \frac{\lambda}{n}) = Pois(\lambda)$. *Restated, this means that for any $x \in \{0, 1, 2, 3, ...\}$,*

$$\lim_{n\to\infty} b\left(n, \frac{\lambda}{n}, x\right) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

PROOF HW

The LSN says that if you perform more and more trials in a Bernoulli experiment, but simultaneously lower the probability of success on each trial so that the expected number of successes is kept equal to the constant λ, you achieve a Poisson r.v. in the limit. So a Poisson r.v. is kind of like a binomial r.v. with infinitely many trials and an infinitely small success probability.

Gamma random variables

The last r.v. associated to a Poisson process whose density we need to find is the time T_r to the r^{th} success in a Poisson process. We start by noting that the range of $T = T_r$ is $[0, \infty)$; next we compute its distribution function. Let $t \in [0, \infty)$. Then

$$\begin{aligned}
F_T(t) = P(T_r \leq t) &= P(X_t \geq r) \\
&\quad \text{(both these inequalities describe the} \\
&\quad \text{event of at least } r \text{ births in } [0,t]) \\
&= 1 - P(X_t < r) \\
&= 1 - P(Pois(\lambda t) < r) \\
&= 1 - \sum_{x=0}^{r-1} \frac{e^{-\lambda t}(\lambda t)^x}{x!}.
\end{aligned}$$

$$\begin{aligned}
\Rightarrow f_T(t) = \frac{d}{dt} F_T(t) &= \frac{d}{dt}\left[1 - \sum_{x=0}^{r-1} \frac{e^{-\lambda t}(\lambda t)^x}{x!}\right] \\
&= -\sum_{x=0}^{r-1} \frac{\lambda^x}{x!} \frac{d}{dt}\left[e^{-\lambda t} t^x\right] \\
&= -\sum_{x=0}^{r-1} \frac{\lambda^x}{x!} \frac{d}{dt}\left[e^{-\lambda t} e^{x \ln t}\right] \\
&= -\sum_{x=0}^{r-1} \frac{\lambda^x}{x!} \frac{d}{dt}\left[e^{x \ln t - \lambda t}\right] \\
&= -\sum_{x=0}^{r-1} \frac{\lambda^x}{x!} \left[e^{x \ln t - \lambda t}\right] \left(\frac{x}{t} - \lambda\right) \\
&= \sum_{x=0}^{r-1} \frac{\lambda^x}{x!} \left[t^x e^{-\lambda t}\right] \left(\lambda - \frac{x}{t}\right) \\
&= \sum_{x=0}^{r-1} \frac{\lambda^{x+1}}{x!} t^x e^{-\lambda t} - \sum_{x=0}^{r-1} \frac{x \lambda^x}{x!} t^{x-1} e^{-\lambda t} \\
&= \sum_{x=0}^{r-1} \frac{\lambda^{x+1}}{x!} t^x e^{-\lambda t} - \sum_{x=1}^{r-1} \frac{x \lambda^x}{x!} t^{x-1} e^{-\lambda t} \\
&= \sum_{x=1}^{r} \frac{\lambda^x}{(x-1)!} t^{x-1} e^{-\lambda t} - \sum_{x=1}^{r-1} \frac{\lambda^x}{(x-1)!} t^{x-1} e^{-\lambda t} \\
&= \boxed{\frac{\lambda^r}{(r-1)!} t^{r-1} e^{-\lambda t}}.
\end{aligned}$$

Definition 3.24 *Let $\lambda \in (0,\infty)$ and let $r \in \{1,2,3,...\}$. A* **gamma** *r.v., denoted $\Gamma(r,\lambda)$, is a cts r.v. X taking values in $[0,\infty)$ whose density function is*

$$f_X(x) = \begin{cases} \dfrac{\lambda^r}{(r-1)!} x^{r-1} e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases}.$$

r and λ are called the **parameters** *of the gamma r.v.*

We will prove that f_X is actually a density function later.

Theorem 3.25 *Let $\{X_t : t \in [0,\infty)\}$ be a Poisson process with rate λ. Then for each $r \in \{1,2,3,...\}$, T_r (the time to the r^{th} birth) is $\Gamma(r,\lambda)$.*

In particular, a $\Gamma(1,\lambda)$ r.v. is the same thing as an $Exp(\lambda)$ r.v. (i.e. $\boxed{\Gamma(1,\lambda) \sim Exp(\lambda)}$).

PROOF The first part of this was derived on the previous page.

For the second part, let $X \sim \Gamma(1,\lambda)$:

$$f_X(x) =$$

This is the same density as an $Exp(\lambda)$ r.v., so $X \sim Exp(\lambda)$. $\square$

A $\Gamma(r,\lambda)$ r.v. **measures the time until the r^{th} birth in a Poisson process with rate λ.**

Problems with r.v.s related to Poisson processes

EXAMPLE 17

The number of people in a community who live to 100 years of age is a Poisson r.v. with parameter 6.

1. Compute the probability that exactly 4 people live to 100.
2. Compute the probability that at least 2 people live to 100.

EXAMPLE 18

The time (in hours) it takes to repair a machine is an exponential r.v. with parameter $\frac{1}{2}$. Find the probability that the repair time is at least 2 hours.

EXAMPLE 19

Suppose X is exponential with parameter 4. Let $Y = X^2$; find a density function of Y.

EXAMPLE 20

Suppose that hits to a certain website follows a Poisson process with rate 200.

1. What is the probability there are (exactly) 630 hits in the first 3 units of time?
2. Suppose there is a hit at time 10. What is the probability that there are no hits between times 10 and 11?
3. Write a density function of the r.v. measuring the time to the fifth hit.

3.5 More on gamma random variables

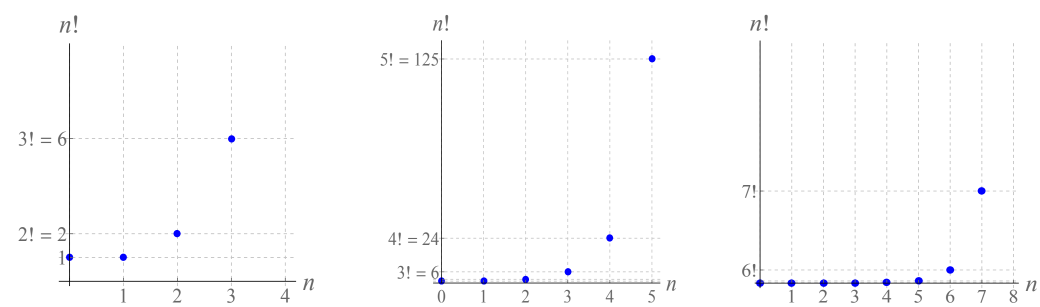
The gamma function

We begin this question by trying to determine the value of $n!$ when n is not a whole number. For example, what is $\frac{1}{2}!$? What is $\pi!$?

More precisely, we seek a function $f : \mathbb{R} \to \mathbb{R}$ (or at least $f : [0, \infty) \to \mathbb{R}$) with the following properties:

1. $f(n) = n!$ for all $n \in \{0, 1, 2, \ldots\}$;
2. f is continuous;
3. $xf(x-1) = f(x)$ for all x.

Such an f would be a "continuous version of factorial":

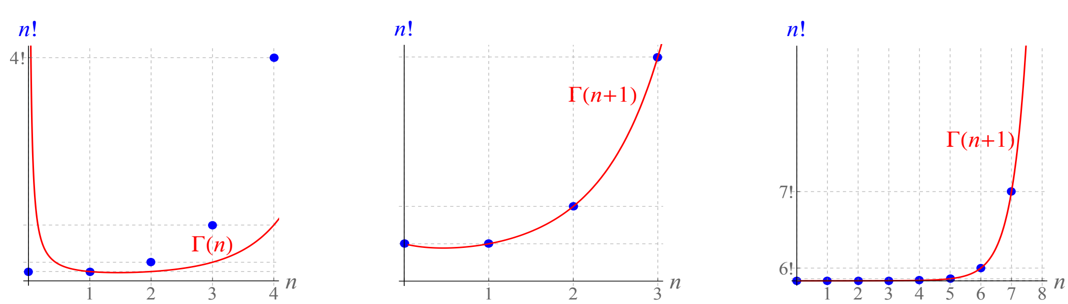


To do this, we will start by trying to incorporate property (3) above through some creative integration by parts. Our attempt will be slightly off, but "close enough".

Definition 3.26 *The* **gamma function** *is the function* $\Gamma : (0, \infty) \to \mathbb{R}$ *defined by*

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x}\, dx.$$

It turns out that $\Gamma(r) = \dfrac{1}{r} \displaystyle\prod_{n=1}^{\infty} \frac{\left(1 + \frac{1}{n}\right)^r}{1 + \frac{r}{n}}$ (this isn't relevant to MATH 414 or 416).



Theorem 3.27 (Properties of the gamma function) *Let Γ be the gamma function. Then:*

1. $\Gamma : (0, \infty) \to \mathbb{R}$ *is continuous.*
2. $\Gamma(1) = 1$.
3. *For every* $r > 0$, $\Gamma(r + 1) = r\Gamma(r)$.
4. *For every* $r > 1$, $\Gamma(r) = (r - 1)\Gamma(r - 1)$.
5. *For* $n \in \{1, 2, 3, ...\}$, $\Gamma(n) = (n - 1)!$.
6. *For every* $n \in \mathbb{N}$, $n! = \Gamma(n + 1)$.

PROOF (1) All functions defined as integrals are cts by the Fund. Thm. of Calculus.

(2) $\Gamma(1) = \int_0^\infty e^{-x}\,dx = -e^{-x}|_0^\infty = 0 - (-1) = 1$.

(4) follows from (3) by replacing all the rs with $r - 1$.

(5) follows from (2) and repeated application of (4).

(6) follows from (5) by replacing each n with $n + 1$. That leaves (3).

To establish (3), use integration by parts with $u = x^r$ and $dv = e^{-x}\,dx$:

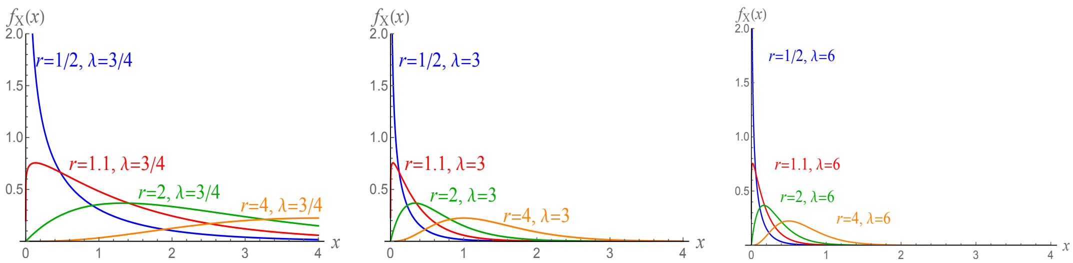
We can now extend the definition of gamma random variables to the situation where r is not necessarily a whole number:

Definition 3.28 *Let* $\lambda \in (0, \infty)$ *and let* $r \in (0, \infty)$. *A* **gamma** *r.v., denoted* $\Gamma(r, \lambda)$, *is a cts r.v.* X *taking values in* $[0, \infty)$ *whose density function is*

$$f_X(x) = \begin{cases} \dfrac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases}.$$

r *and* λ *are called the* **parameters** *of the gamma r.v.*

Here are some graphs of density functions of $\Gamma(r, \lambda)$ r.v.s:



Theorem 3.29 *The density function of a $\Gamma(r, \lambda)$ r.v. is in fact a density function.*

PROOF Perform the u-substitution $u = \lambda x$; $du = \lambda\, dx$ inside the integral:

$$\int_0^\infty f_X(x)\, dx = \int_0^\infty \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}\, dx =$$

Corollary 3.30 (Gamma Integral Formula) *Let $r, \lambda > 0$. Then:*

$$\int_0^\infty x^{r-1} e^{-\lambda x}\, dx = \frac{\Gamma(r)}{\lambda^r}$$

Application: $\displaystyle\int_0^\infty 4z^6 e^{-2z}\, dz =$

3.6 Summary of Chapter 3

- A continuous random variable is a function $X : \Omega \to \mathbb{R}^d$ such that the probability of any individual value of X is zero.
- We usually describe continuous r.v.s by specifying a density function $f_X : \mathbb{R}^d \to [0, \infty)$, which satisfies

$$P(X \in E) = \int_E f_X(x)\,dx$$

  for any set E. Such a function must be everywhere nonnegative and must integrate to 1. We compute probabilities associated to a cts r.v. by integrating the density function as above.
- All real-valued r.v.s can be described by giving a distribution function $F_X : \mathbb{R} \to [0,1]$ defined by

$$F_X(x) = P(X \le x).$$

  Distribution functions have many properties; notably X is cts if and only if F_X is cts; and if X is cts with density f_X,

$$f_X(x) = \frac{d}{dx}F_X(x) \quad \text{and} \quad F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt.$$

- To find the density function of a continuous transformation Y of a continuous r.v X., first find the range of Y, then compute F_Y by back-substitution. Last, differentiate F_Y to get f_Y.
- Classes of commonly encountered continuous random variables include the following:
  1. **uniform** r.v.s, which assign relatively equal likelihood to all values in the range of X;
  2. **exponential** r.v.s, which measure the amount of time until a birth happens in a Poisson process (and are the only memoryless cts r.v.s);
  3. **gamma** r.v.s, which measure the amount of time until the r^{th} birth in a Poisson process;
  4. the **Cauchy** r.v., which gives the tangent of a uniformly chosen angle.

  You should know the range, distribution and density function of each of these common r.v.s, and additional facts relevant to each class.
- One additional class of discrete r.v.s not previously encountered are **Poisson** r.v.s, which count the number of births over a fixed length of time in a Poisson process.

- The gamma function Γ, defined by
$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x}\, dx,$$
extends the idea of factorial to positive real numbers (for $n \in \mathbb{N}$, $n! = \Gamma(n+1)$).

3.7 Chapter 3 Homework

Exercises from Section 3.1

1. Suppose you choose a real number X from the interval $[2, 10]$ with a density function of the form $f_X(x) = Cx$, where C is some constant.
   a) What is the value of C?
   b) Compute $P(X > 5)$.
   c) Compute $P(X \leq 7)$.
2. a) (AE) The loss due to a fire in a commercial building is modeled by a continuous r.v. X with density function $f(x) = k(20 - x)$ for $0 < x < 20$ ($f(x) = 0$ otherwise). Given that a fire loss exceeds 8, what is the probability that it exceeds 16?

b) In part (a) of this problem, did you have to determine the value of k to find the answer? Why or why not?

c) In general, for what types of probability computations would one not need to find the value of an unknown multiplicative constant (like the k in part (a)) in a density function?

Exercises from Section 3.2

3. Let X be a r.v. whose distribution function is

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \dfrac{x}{4} & \text{if } 0 \leq x < 1 \\ \dfrac{x}{2} & \text{if } 1 \leq x < 2 \\ 1 & \text{if } x \geq 2 \end{cases}$$

Compute each quantity:

a) $P\left(\frac{1}{4} \leq X \leq \frac{5}{4}\right)$

b) $P\left(\frac{1}{4} \leq X \leq 1\right)$

c) $P\left(\frac{1}{4} \leq X < 1\right)$

d) $P\left(1 \leq X \leq \frac{7}{4}\right)$

e) $P\left(1 < X < 2\right)$.

f) $P\left(X \text{ is an integer}\right)$

4. Suppose X is a r.v. whose cdf is

$$F_X(x) = \begin{cases} 0 & \text{if } x < 1 \\ \dfrac{1}{10} & \text{if } 1 \leq x < 3 \\ \dfrac{x}{10} & \text{if } 3 \leq x < 4 \\ K - \dfrac{2}{x} & \text{if } x \geq 4 \end{cases}$$

where K is a constant. Compute each quantity:

a) K

b) $P(X = 3)$

c) $P(2 < X < 3)$

d) $P(3 < X < 4)$

e) $P\left(X > 1\right)$

f) $P\left(X > 4 \,|\, X \geq 4\right)$

g) $P\left(X < 3.5 \,|\, X \leq 4\right)$

h) $P\left(X > 2 \,|\, X > 3\right)$

5. Suppose X is a continuous r.v. with survival function

$$S(x) = \begin{cases} x^{-3/2} & \text{if } x \geq 0 \\ 1 & \text{if } x < 0. \end{cases}$$

   a) Compute $P(X \geq 7)$.
   b) Compute $P(X < 5)$.
   c) Compute $P(3 < X < 10)$.
   d) Compute the cdf of X.
   e) Compute a density function of X.

Exercises from Section 3.3

6. Let X be a discrete r.v. with density function f_X defined as follows:

$$f_X(-1) = \frac{1}{5},\ f_X(0) = \frac{1}{5},\ f_X(1) = \frac{2}{5},\ f_X(2) = \frac{1}{5}.$$

   a) Compute a density function of $Y = 2X + 1$.
   b) Compute a density function of $Z = X^2$.

7. Suppose $X \sim Unif([1, 10])$. Compute a density of $Y = \ln X$.

8. Suppose X has the density

$$f_X(x) = \begin{cases} \frac{3}{8}x - \frac{3}{32}x^2 & \text{if } 0 \leq x \leq 4 \\ 0 & \text{else} \end{cases}$$

   Compute a density function of $Y = \sqrt{X}$.

9. Let X be a continuous, real-valued r.v. with some unknown distribution function F_X and density function f_X.

   a) Compute (in terms of F_X) the distribution function of $Y = e^X$.
   b) Compute (in terms of f_X) a density function of $Y = e^X$.

10. Suppose a point (X, Y) is chosen uniformly from the triangle whose vertices are $(0, 0)$, $(4, 0)$, and $(4, 4)$. Compute a density function of $W = Y - X$.

11. Suppose a point (X, Y) is chosen uniformly from the rectangle whose vertices are $(1, 0)$, $(5, 0)$, $(1, 2)$ and $(5, 2)$. Compute a density function of $V = X + Y$.

12. Compute the density function of $Z = XY$, where X and Y are chosen as in the previous problem.

13. A point is chosen uniformly from the interval $(-10, 10)$. Let X be the r.v. defined so that X denotes the coordinate of the chosen point if the point is in $[-5, 5]$, $X = -5$ if the point is in the interval $(-10, -5)$, and $X = 5$ if the point is in the interval $(5, 10)$.

    a) Compute the distribution function F_X of the r.v. X.
    b) Sketch the graph of F_X.
    c) Classify X as discrete, continuous or mixed, with appropriate justification.
    d) Does X have a density function? Why or why not?

Exercises from Section 3.4

14. a) The number of bad checks that a bank receives during a 5-hour business day is a Poisson r.v. with $\lambda = 2$. What is the probability that the bank will receive no more than 2 bad checks in its business day?
    b) The mileage (in thousands of miles) that car owners get with a certain kind of radial tire is a r.v. whose distribution is exponential with parameter $\frac{1}{40}$. Compute the probability that one of these tires will last at least 20,000 miles.

15. (AE) You are given the following information about N, the annual number of claims for a randomly selected insured person:

$$P(N = 0) = \frac{1}{2}; \quad P(N = 1) = \frac{3}{8}; \quad P(N = 2) = \frac{1}{8}.$$

    Let S denote the total annual claim amount for an insured. When $N = 1$, S is exponentially distributed with parameter $\frac{1}{6}$. When $N > 1$, S is exponentially distributed with parameter $\frac{1}{10}$. Compute $P(4 < S < 8)$.

    Hint: Use the Law of Total Probability.

16. Suppose that births occur according to a Poisson process with hourly rate $\lambda = 3$, where $t = 0$ corresponds to midnight. Let p be the probability that no births occur between 8 AM and 10 AM.

    a) Compute p, using the density function of an appropriate discrete r.v.
    b) Compute p, using the density function of an appropriate continuous r.v.

17. Suppose that births occur according to a Poisson process with hourly rate $\lambda = 3$, where $t = 0$ corresponds to midnight.

a) What is the probability that exactly 5 births occur by 2 AM?

b) What is the probability that at least 3 births have occurred by 7 AM?

c) What is the probability that exactly one birth occurs between 8 and 9 AM and exactly two births occur between 2 and 4 PM?

d) What is the conditional probability that at least one birth takes place between 8 AM and noon, given that no births take place between 8 AM and 10 AM?

e) What is the probability that exactly one birth occurs between 8 and 10 AM and exactly one birth occurs between 9 and 11 AM?

Hint: Split this situation into two disjoint events; compute the probability of each event, and add.

18. Suppose that births occur according to a Poisson process with rate λ.

a) Suppose you are given that v births occur between times 0 and t. Let $s < t$; compute the probability that exactly x of the v births occur between times 0 and s.

b) Suppose you are given that v births occur between times 0 and t. Let $s < t$. If X records the number of births occurring between times 0 and s, what kind of r.v. is X? Include its parameters.

Hint: You computed the density function of X in part (a). Simplify this density function and identify it as the density of a common r.v.

c) Suppose nine births occur between times 15 and 27. What is the probability that (exactly) three of those births occurred after time 22?

NOTE: You should remember the result you derived in part (b) of the preceding problem.

17. Suppose X is exponential with parameter λ, where λ is such that $P(X \geq .02) = .35$. Determine the number t such that $P(X \geq t) = .85$.

18. (AE) Suppose the number of claims filed by an insurance policyholder is a Poisson r.v. If the filing of (exactly) one claim is four times as likely as the filing of (exactly) two claims, find the probability the policyholder files exactly five claims.

19. Choose (a) or (b):

a) Let X have an exponential density with parameter λ. Compute the density of $Y = cX$, where $c > 0$ is a positive constant.

b) Let X have the Cauchy density. Compute the density of $Y = a + bX$, where a and b are constants such that $b > 0$.

20. Prove the **Law of Small Numbers**, which says that for any constant $\lambda > 0$,
$$\lim_{n\to\infty} b\left(n, \frac{\lambda}{n}, x\right) = \frac{e^{-\lambda}\lambda^x}{x!}.$$
Hint: As a model, follow the (long) computation done on the page before Definition 3.20.

21. (AE) The damage done to a house by a natural disaster is an exponential r.v. with $P(X \geq 30) = .3$. If a natural disaster strikes 15 houses, and the damages to each house are independent, what is the probability that of the 15 houses, at least 2 of them suffer damage at least 20?

Exercises from Section 3.5

22. a) Evaluate $\Gamma(7)$.
    b) Simplify $\dfrac{\Gamma(3.2)}{\Gamma(5.2)}$.
    c) Suppose x is some number so that $\Gamma(x) = 100000$. Compute $\Gamma(x-2)$, in terms of x.

    A useful and amazing fact to know about the gamma function is the following:
$$\Gamma(r)\Gamma(1-r) = \frac{\pi}{\sin(\pi r)}.$$
    Use this fact to evaluate each given expression:

    d) $\Gamma\left(\frac{1}{3}\right)\Gamma\left(\frac{2}{3}\right)$
    e) $\Gamma\left(\frac{7}{6}\right)\Gamma\left(\frac{5}{6}\right)$

23. Evaluate each integral:

    a) $\displaystyle\int_0^\infty x^7 e^{-x/3}\,dx$
    b) $\displaystyle\int_0^\infty 4x^2 e^{-4x}\,dx$
    c) $\displaystyle\int_0^\infty 8x^3 e^{3-2x}\,dx$
    d) $\displaystyle\int_0^\infty (3x)^t e^{yx}\,dx$

Calculus review

24. For each given function f, compute $\dfrac{\partial f}{\partial x}$, $\dfrac{\partial f}{\partial y}$, $\dfrac{\partial^2 f}{\partial x^2}$, $\dfrac{\partial^2 f}{\partial y^2}$ and $\dfrac{\partial^2 f}{\partial x \partial y}$.

a) $f(x, y) = x^2 + 2xy + 3y^2$

b) $f(x, y) = e^{2x-3y}$

25. For each given function f, compute $\dfrac{\partial f}{\partial x}$ and $\dfrac{\partial f}{\partial y}$.

a) $f(x, y) = e^{4x} + e^{xy} - e^{3y}$

b) $f(x, y) = e^{-x^2+4xy-2y^2}$

26. In each part of this exercise, you are given an iterated integral. Some of them represent valid mathematics, and some of them are nonsense. Determine whether each expression is valid, or nonsense:

a) $\displaystyle\int_0^1 \int_0^1 f(x, y)\, dx\, dy$

b) $\displaystyle\int_0^1 \int_0^x f(x, y)\, dy\, dx$

c) $\displaystyle\int_0^1 \int_0^x f(x, y)\, dy\, dx$

d) $\displaystyle\int_0^1 \int_0^y f(x, y)\, dy\, dx$

e) $\displaystyle\int_0^w \int_0^y f(x, y)\, dx\, dy$

f) $\displaystyle\int_0^w \int_0^x f(x, y)\, dx\, dy$

g) $\displaystyle\int_0^y \int_0^x f(x, y)\, dy\, dx$

h) $\displaystyle\int_0^y \int_0^y f(x, y)\, dx\, dy$

i) $\displaystyle\int_0^x \int_0^x f(x, y)\, dy\, dx$

j) $\displaystyle\int_0^1 \int_0^w f(x, y)\, dx\, dy$

27. Compute each iterated integral:

a) $\displaystyle\int_0^1 \int_0^y 6x^2y^3\, dx\, dy$

b) $\displaystyle\int_0^\infty \int_y^\infty e^{-x-y}\, dx\, dy$

c) $\displaystyle\int_0^1 \int_x^4 xy\, dy\, dx$

d) $\displaystyle\int_0^1 \int_y^{2-y} dx\, dy$

Note:
I didn't forget anything in (d).

Chapter 4

Joint distributions

4.1 Introducing joint distributions

Suppose that in a probabilistic experiment you are taking more than one measurement, say d distinct (real-valued) random variables. Often, the right way to think of these d quantities is as a single random variable which takes values in $\mathbb{R}^d$.

EXAMPLE 1

Pick a sample of 6 marbles (simultaneously) from an urn with 10 red, 12 blue, 18 black and 20 green marbles in it. Let

$$\begin{aligned} X_1 &= \# \text{ of red marbles drawn} \\ X_2 &= \# \text{ of blue marbles drawn} \\ X_3 &= \# \text{ of black marbles drawn} \\ X_4 &= \# \text{ of green marbles drawn} \end{aligned}$$

Obtain $\mathbf{X} = \overrightarrow{X} = X = (X_1, X_2, X_3, X_4) : \Omega \to \mathbb{R}^4$ (discrete, $4-$diml r.v.)

EXAMPLE 2

Pick a point uniformly from the unit square. Let

$$\begin{aligned} X &= x - \text{coordinate of the chosen point} \\ Y &= y - \text{coordinate of the chosen point} \end{aligned}$$

Obtain $\mathbf{X} = \overrightarrow{X} = (X, Y) : \Omega \to \mathbb{R}^2$ (cts, $2-$diml r.v.)

EXAMPLE 3

Pick a point uniformly (or not) from the triangle whose vertices are $(0,0)$, $(0,2)$ and $(4,0)$. Let

$$X = x - \text{coordinate of the chosen point}$$
$$Y = y - \text{coordinate of the chosen point}$$

Obtain $\mathbf{X} = \overrightarrow{X} = (X, Y) : \Omega \to \mathbb{R}^2$ (cts, $2-$diml r.v.)

Notice: In Example 2, you obtain no information about either X or Y when you are told the value of the other coordinate. This is not the case in Examples 1 and 3; as you learn information about one or more coordinates, your belief about the values of the remaining coordinates changes.

Definition 4.1 *A $d-$**dimensional random variable** (a.k.a. $d-$**dimensional random vector** is a random variable whose range is a subset of $\mathbb{R}^d$. We denote such a r.v. by $\mathbf{X}$ or X or $\overrightarrow{X}$.*

Definition 4.2 *Let $\mathbf{X} : \Omega \to \mathbb{R}^d$ be a discrete $d-$dim'l r.v. with density function $f_{\mathbf{X}}$. The coordinates $X_1, X_2, ..., X_d$ of X are called its **marginals**, and any such $\mathbf{X}$ is called a **joint distribution** of its marginals. $f_{\mathbf{X}}$ is called the **joint density (function)** of $\mathbf{X}$.*

Note: Given a bunch of marginals $X_1, ..., X_d$, one can construct lots of different joint distributions $\mathbf{X}$ of those marginals (see Examples 4 and 5 coming up).

4.2 Discrete joint distributions

Similar to real-valued discrete r.v.s, a discrete $d-$dim'l r.v. is determined by a density function

$$f_{\mathbf{X}} = f_X = f_{\vec{X}} : \mathbb{R}^d \to [0, \infty)$$

satisfying

$$f_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}) \text{ for all } \mathbf{x} \in \mathbb{R}^d$$

and

$$P(\mathbf{X} \in E) = \sum_{\mathbf{x} \in E} f_{\mathbf{X}}(\mathbf{x})$$

for any event E.

Theorem 4.3 (Density function of marginals, discrete case) *Let $\mathbf{X} : \Omega \to \mathbb{R}^d$ be a discrete $d-$dim'l r.v. with density function $f_{\mathbf{X}}$. Then the density function of the j^{th} marginal X_j is*

$$f_{X_j}(x) = P(X_j = x) = \sum_{\{\mathbf{x} \in \mathbb{R}^d : \mathbf{x}_j = x\}} f_{\mathbf{X}}(\mathbf{x}).$$

In other words, this theorem says that to find the density function of a marginal, you **add up the values of the joint density over all the coordinates other than the marginal you want**. As a special case, given a two-dimensional joint density $f_{X,Y}$,

$$f_X(x) = \sum_y f_{X,Y}(x, y) \quad \text{and} \quad f_Y(y) = \sum_x f_{X,Y}(x, y).$$

EXAMPLE 4

Independently roll a fair die and flip a fair coin. Let X record the number on the die and let Y record 0 for tails and 1 for heads. Describe the joint density of X and Y, and the marginals.

EXAMPLE 5

Roll a fair die and flip a coin, with the assumption that the coin "knows" what number is rolled, i.e. if you roll an even number then the coin flips heads with probability $2/3$ and if you roll an odd number then the coin flips heads with probability $1/3$. Let X record the number on the die and let Y record 0 for tails and 1 for heads. Describe the joint density of X and Y, and the marginals.

$Y \backslash X$	1	2	3	4	5	6
0						
1						

EXAMPLE 6

Draw 4 balls without replacement from an urn with 15 green and 5 black balls in it. Let X and Y be the number of green and black balls drawn, respectively. Describe the joint density of X and Y, and the marginals.

$Y \backslash X$	0	1	2	3	4
0					
1					
2					
3					
4					

EXAMPLE 7

1000 people are surveyed, and the results are summarized in the following table:

	SMOKERS	NON-SMOKERS
UNDER AGE 30	10%	38%
AGE 30+	18%	34%

For each question, give the correct notation for what the question is asking, and answer the question.

1. What % of those surveyed are under age 30?
2. What is the probability that a surveyed person aged 30+ smokes?
3. What is the probability that a given non-smoker is under 30?

EXAMPLE 8

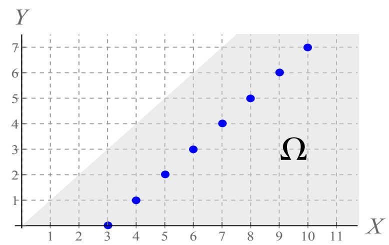
Suppose X and Y are integer-valued r.v.s with joint density

$$f_{X,Y}(x,y) = \begin{cases} \frac{c}{4^y} & \text{if } 0 \leq x \leq y \\ 0 & \text{else} \end{cases}$$

where c is a constant.

1. Determine the value of c.
2. Compute $P(X = 5, Y = 8)$.
3. Compute the density function of the marginal X.
4. Compute $P(X - Y = 3)$.
5. Write an expression involving sums and/or integrals that could be evaluated to give $P(X + Y \leq 12)$. (You do not need to evaluate this expression.)

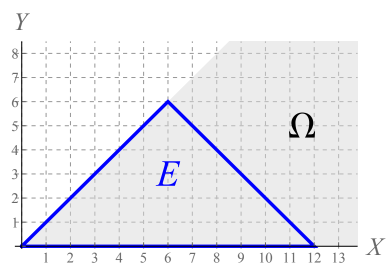
4. First, sketch a picture of the points (X, Y) so that $X - Y = 3$:



Now compute the probability by adding up values of the density function over all the (x, y) marked in the picture:

$$P(X - Y = 3) = P(Y = X - 3) = \sum_{x=3}^{\infty} f_{X,Y}(x, x-3) = \sum_{x=3}^{\infty} \frac{9}{16 \cdot 4^{x-3}}$$
$$= \frac{9}{16} \sum_{x=0}^{\infty} \left(\frac{1}{4}\right)^x$$
$$= \frac{9}{16}\left(\frac{1}{1-\frac{1}{4}}\right) = \boxed{\frac{3}{4}}.$$

5. First, sketch a picture of the set E of $(X, Y) \in \Omega$ so that $X + Y \leq 12$:



Now compute the probability by adding up values of the density function over all $(x, y) \in E$:

EXAMPLE 9

Suppose X and Y are discrete r.v.s with joint distribution

$$f_{X,Y}(x,y) = \begin{cases} pq(1-p)^x(1-q)^y & \text{for } x \geq 0,\, y \geq 0 \\ 0 & \text{else} \end{cases}$$

where p and q are constants.

1. Compute the density of the marginal Y.
2. Compute the density of $Z = \min(X, Y)$.

4.3 Multinomial and hypergeometric distributions

MOTIVATING PROBLEM: SAMPLING

A jar contains 100 marbles of various colors: 30 red, 25 white, 15 green, and 40 black. You draw a sample of 8 marbles from the jar. Let

$$\mathbf{X} = (R, W, G, B)$$

record the number of marbles of each color you draw. This is a 4-dimensional r.v.

Question: What is the joint density function of $\mathbf{X}$?

Answer:

Sampling without replacement

In this setting, the joint density comes from the partition problem formula we described in Chapter 2:

Definition 4.4 *Let $n \in \mathbb{N}$ and let $n_1, ..., n_d \in \mathbb{N}$ be such that $\sum_j n_j = n$. Let $k \leq n$. A discrete joint distribution $\mathbf{X} : \Omega \to \mathbb{R}^d$ is called* **hypergeometric** *(or $d-$***dim'l hypergeometric** *if it has density function*

$$f_{\mathbf{X}}(x_1, ..., x_d) = \begin{cases} \dfrac{\binom{n_1}{x_1}\binom{n_2}{x_2}\cdots\binom{n_d}{x_d}}{\binom{n}{k}} & \text{for } (x_1, ..., x_d) \in \mathbb{N}^d \text{ satisfying } \displaystyle\sum_{j=1}^{d} x_j = k \\ 0 & \text{else} \end{cases}$$

In this case, we write $\mathbf{X} \sim Hyp(n, (n_1, n_2, ..., n_d), k)$ or $\mathbf{X} \sim Hyp(\mathbf{n}, k)$ where $\mathbf{n} = (n_1, ..., n_d)$.

d-dimensional hypergeometric r.v.s model the situation where you have n_j objects of type j in a jar (for a total of n objects) and you draw k objects **without replacement**. If you let X_j be the number of objects of type j you draw, then $\mathbf{X} = (X_1, ..., X_d) \sim Hyp(n, (n_1, ..., n_d), k)$.

In our motivating example above, if the sampling is without replacement the joint density of X would be

and the probability of drawing 3 red, 1 white, 2 green and 2 black marbles is

Theorem 4.5 *Suppose* $\mathbf{X} \sim Hyp(n, (n_1, n_2, ..., n_d), k)$*. Then* $X_j \sim Hyp(n, n_j, k)$*, where* X_j *is the* j^{th} *marginal of* $\mathbf{X}$*.*

Sampling with replacement

In this setting, we can think of each draw from the jar as an independent repetition of a "Bernoulli-like" trial, except that the trial has d different outcomes ($d = 4$ in our example; this is the number of different colors). Now, the probability of getting the j^{th} outcome x_j times in n trials is

Definition 4.6 *Let $n \in \mathbb{N}$ and let $p_1, ..., p_d \geq 0$ be such that $\sum_j p_j = 1$. A discrete joint distribution $\mathbf{X} : \Omega \to \mathbb{R}^d$ is said to be* **multinomial** *with parameters n and $\mathbf{p} = (p_1, ..., p_d)$ if it has joint density*

$$f_{\mathbf{X}}(x_1, ..., x_d) = \left(\begin{array}{c} n \\ x_1, x_2, ..., x_d \end{array} \right) \prod_{j=1}^{x_d} p^{x_j} = \frac{n!}{x_1! x_2! \cdots x_d!} p_1^{x_1} p_2^{x_2} \cdots p_d^{x_d}$$

(for nonnegative integers $x_1, ..., x_d$ satisfying $\sum_{j=1}^{d} x_j = n$; the joint density is 0 otherwise). In this setting, we write $\mathbf{X} \sim multi(n, (p_1, p_2, ..., p_d))$ or $\mathbf{X} \sim multi(n, \mathbf{p})$.

Multinomial r.v.s describe **sampling with replacement**.

In our motivating example above, if the sampling is with replacement then the joint density of X would be

and the probability of drawing 3 red, 1 white, 2 green and 2 black marbles is

Theorem 4.7 *Suppose $\mathbf{X} \sim multi(n, \mathbf{p})$. Then $X_j \sim b(n, p_j)$, where X_j is the j^{th} marginal of $\mathbf{X}$.*

4.4 Continuous joint distributions

In this section, we take the usual language associated to non-discrete, real-valued r.v.s and extend it to joint distributions.

As usual, given 2 real-valued r.v.s X and Y, we think of $\mathbf{X} = (X, Y) : \Omega \to \mathbb{R}^2$.

(Similarly, write $\mathbf{X} = (X_1, ..., X_d) : \Omega \to \mathbb{R}^d$.)

Joint distribution functions

DIMENSION	DEFINITION OF DIST. FUNCTION	APPLICATION TO PROBABILITIES
$d = 1$ $(X : \Omega \to \mathbb{R})$	$F_X : \mathbb{R} \to [0, 1]$ $F_X(x) = P(X \leq x)$	$P(a < X \leq b) =$ $F_X(b) - F_X(a)$
$d = 2$ $(\mathbf{X} : \Omega \to \mathbb{R}^2)$ $(\mathbf{X} = (X, Y))$		
general d $\mathbf{X} : \Omega \to \mathbb{R}^d$		

Moral Distribution functions are not as useful for joint distributions as they are for real-valued r.v.s.

Theorem 4.8 (Properties of joint distribution functions) *Let* $\mathbf{X} : \Omega \to \mathbb{R}^d$ *be a joint distribution with joint cdf* $F_{\mathbf{X}} : \mathbb{R}^d \to [0, 1]$. *Then:*

1. $\lim_{x_j \to \infty \, \forall j} F_{\mathbf{X}}(\mathbf{x}) = 1.$

2. $\lim_{x_j \to -\infty \, \forall j} F_{\mathbf{X}}(\mathbf{x}) = 0.$

3. *If all but one coordinate is fixed,* $F_{\mathbf{X}}$ *is increasing with respect to that coordinate.*

Marginal distribution functions

As with the discrete case, the coordinates of a joint non-discrete r.v. are called its **marginals**. We can compute the cdf of a marginal from a joint cdf by taking limits:

Theorem 4.9 (Distribution functions of marginals) *Let* $\mathbf{X} : \Omega \to \mathbb{R}^d$ *be a joint distribution with joint cdf* $F_{\mathbf{X}} : \mathbb{R}^d \to [0, 1]$. *Then the cdf* F_{X_j} *of the* j^{th} *marginal* X_j *is*

$$F_{X_j}(x_j) = P(X_j \leq x_j) = \lim_{x_i \to \infty \, \forall i \neq j} F_{\mathbf{X}}(x_1, ..., x_d).$$

PROOF

$$\begin{aligned}
F_{X_j}(x_j) &= P(X_j \leq x_j) \\
&= P(X_1 < \infty, X_2 < \infty, ..., X_{j-1} < \infty, X_j \leq x_j, X_{j+1} < \infty, ..., X_d < \infty) \\
&= \text{“}F_{\mathbf{X}}(\infty, \infty, ..., \infty, x_j, \infty, ..., \infty)\text{”} \\
&= \lim_{x_i \to \infty \forall i \neq j} F_{\mathbf{X}}(x_1, x_2, ..., x_d). \; \square
\end{aligned}$$

As a special case, given joint distribution (X, Y) with joint cdf $F_{X,Y}$, we have

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

Density functions for continuous joint distributions

RECALL

A r.v. $\mathbf{X} : \Omega \to \mathbb{R}^d$ is called **continuous** if $P(\mathbf{X} = \mathbf{x}) = 0$ for every $\mathbf{x} \in \mathbb{R}^d$.

When $d = 1$, most cts r.v.s have a density function which is used to compute probabilities: if $X : \Omega \to \mathbb{R}$ is continuous with density function f_X, then

Definition 4.10 *Let* $\mathbf{X} : \Omega \to \mathbb{R}^d$ *be a r.v. We say that a function* $f_{\mathbf{X}} : \mathbb{R}^d \to [0, \infty)$ *is a* **(joint) density function** *for* $\mathbf{X}$ *if for every subset* $E \subseteq \mathbb{R}^d$ *whose size (i.e. length/area/volume/etc.) can be computed using calculus,*

$$P(\mathbf{X} \in E) = \int_E f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}.$$

Note: The integral in the above definition is really a multiple integral:

$$d = 1: \quad \int_E f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \text{ means } \int_a^b f_X(x)\, dx$$

$$d = 2: \quad \int_E f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \text{ means } \iint_E f_{X,Y}(x, y)\, dA$$

$$d = 3: \quad \int_E f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \text{ means } \iiint_E f_{X,Y,Z}(x, y, z)\, dV$$

etc.

Note: Density functions for a specific cts joint distribution X are not unique (they can be changed at single points, etc. without affecting probability computations).

Theorem 4.11 (Properties of joint density functions) *Let* $\mathbf{X} : \Omega \to \mathbb{R}^d$ *be a d-dimensional r.v.*

1. *If* $\mathbf{X}$ *is mixed, then it has no density function.*
2. *A (measurable) function* $f : \mathbb{R}^d \to \mathbb{R}$ *is the density function of a cts joint distribution* $\mathbf{X}$ *if and only if*
   (i) $f(\mathbf{x}) \geq 0$ *for all* $\mathbf{x} \in \mathbb{R}^d$*, and*
   (ii) $\displaystyle\int_{\mathbb{R}^d} f(\mathbf{x})\, d\mathbf{x} = 1.$
3. *Suppose continuous* $\mathbf{X} : \Omega \to \mathbb{R}^d$ *has joint distribution function* $F_{\mathbf{X}}$ *and joint density function* $f_{\mathbf{X}}$*. Then for all* $\mathbf{x} \in \mathbb{R}^d$*,*

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^d}{\partial x_1 \partial x_2 \cdots \partial x_d} F_{\mathbf{X}}(\mathbf{x}).$$

Remark: There are continuous joint distributions that do not have a density function, but we don't have to worry about those in MATH 414 or 416.

As a special case of (3), we see that if (X, Y) is a cts joint distribution with joint cdf $F_{X,Y}(x, y)$ and joint density $f_{X,Y}(x, y)$, then

$$\boxed{f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y)}.$$

PROOF (WHEN $d = 2$) Let $x, y \in \mathbb{R}$. Then

$$\int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s, t)\, dt\, ds = P(X \leq x, Y \leq y) = F_{X,Y}(x, y)$$

Differentiate both sides of this equation with respect to x:

$$\frac{\partial}{\partial x} \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s, t)\, dt\, ds = \frac{\partial}{\partial x} F_{X,Y}(x, y)$$

Now differentiate both sides with respect to y:

$$\frac{\partial}{\partial y} \int_{-\infty}^{y} f_{X,Y}(x, t)\, dt = \frac{\partial}{\partial y} \left[\frac{\partial}{\partial x} F_{X,Y}(x, y) \right]$$

EXAMPLE 10

Suppose X and Y are cts r.v.s with joint density

$$f_{X,Y}(x,y) = \begin{cases} 6xy^2 & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{else} \end{cases}.$$

Compute $P\left(X + Y \leq \frac{1}{2}\right)$.

Density functions of marginals (continuous case)

Theorem 4.12 (Density functions of marginals, continuous case) *Let* $\mathbf{X} : \Omega \to \mathbb{R}^d$ *be a cts joint distribution with joint density function* $f_{\mathbf{X}} : \mathbb{R}^d \to [0, \infty)$*. Then:*

1. *Each marginal* X_j *is continuous and has a density function;*
2. *For each* j*,*

$$f_{X_j}(x_j) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) \, dx_1 dx_2 \cdots dx_{j-1} dx_{j+1} \cdots dx_d.$$

This theorem tells us that to find the density function of the marginal of a continuous joint distribution, you **integrate the joint density with respect to all the other coordinates**.

As a special case, if X and Y are cts r.v.s with joint density function $f_{X,Y}(x, y)$, then

$$\boxed{f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy} \quad \text{and} \quad \boxed{f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dx}.$$

PROOF (WHEN $d = 2$) Let $x \in \mathbb{R}$. Then:

$$P(X \le x) = P(X \in (-\infty, x]) = \int_{-\infty}^{x} f_X(s) \, ds.$$

At the same time,

$$\begin{aligned} P(X \le x) = P(X \in (-\infty, x]) &= P(X \in (-\infty, x], Y \in (-\infty, \infty)) \\ &= \iint_{(-\infty,x]\times(-\infty,\infty)} f_{X,Y}(s, y) \, dA \\ &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(s, y) \, dy \, ds. \end{aligned}$$

By equating the two expressions above we found for $P(X \le x)$, we get

$$\int_{-\infty}^{x} f_X(s) \, ds = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(s, y) \, dy \, ds.$$

Differentiate both sides of this with respect to x; by the FTC we get

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy. \ \square$$

4.5 Independence of random variables

RECALL

Earlier in the course we talked about what it means for two *events* to be independent:

Now, we want to extend the notion of independence to *random variables*.

> **Definition 4.13** *Let $X_1, ..., X_d$ be real-valued r.v.s with joint distribution $\mathbf{X}$. The r.v.s (just as well, the distribution) are (is) called* **(mutually) independent** *if*
>
> $$F_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^{d} F_{X_j}(x_j)$$
>
> *for all $\mathbf{x} = (x_1, ..., x_d) \in \mathbb{R}^d$, where $F_{\mathbf{X}}$ is the joint cdf and the F_{X_j} are the cdfs of the marginals.*

Notation: If two r.v.s X and Y are independent, we write $X \perp Y$; otherwise we write $X \not\perp Y$.

Idea: To say two r.v.s are independent means that given any information about one of them does not affect your assessment of any probability associated to the other one.

> **IMPORTANT:** Whether r.v.s are independent depends on the joint distribution, and not just on the marginals. Look back at Examples 4 and 5 from earlier in this chapter, which have the same marginals X and Y.
>
> - In Example 4, $X \perp Y$.
> - In Example 5, $X \not\perp Y$.

Theorem 4.14 *Let $X_1, ..., X_d$ be continuous real-valued r.v.s with joint density $f_{\mathbf{X}} : \mathbb{R}^d \to [0, \infty)$. Then the X_j are independent if and only if*

$$f_{\mathbf{X}}(x_1, ..., x_d) = \prod_{j=1}^{d} f_{X_j}(x_j) \text{ for all } (x_1, ..., x_d) \in \mathbb{R}^d.$$

As a special case, we see that $\boxed{X \perp Y \text{ iff } f_{X,Y}(x,y) = f_X(x) f_Y(y)}$ for all x, y.

Proof (when $d = 2$)
($\Rightarrow$) Suppose $X \perp Y$.
Then $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ by definition of $\perp$.
Take mixed second-order partials of both sides of this to get

$$\begin{aligned}
\frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) &= \frac{\partial}{\partial x}\frac{\partial}{\partial y} F_X(x)F_Y(y) \\
f_{X,Y}(x,y) &= \frac{\partial}{\partial x} F_X(x) \, \frac{\partial}{\partial y} F_Y(y) \\
f_{X,Y}(x,y) &= f_X(x) f_Y(y).
\end{aligned}$$

($\Leftarrow$) Suppose $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. Then

$$\begin{aligned}
F_{X,Y}(x,y) = P(X \le x, Y \le y) &= \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s,t)\, dt\, ds \\
&= \int_{-\infty}^{x} \int_{-\infty}^{y} f_X(s) f_Y(t)\, dt\, ds \\
&= \int_{-\infty}^{x} f_X(s)\, ds \cdot \int_{-\infty}^{y} f_Y(t)\, dt \\
&= F_X(x)F_Y(y).
\end{aligned}$$

so $X \perp Y$ by definition. $\square$

A similar result holds for density functions of discrete r.v.s:

Theorem 4.15 *Let $X_1, ..., X_d$ be discrete, real-valued r.v.s with joint distribution $\mathbf{X}$. The r.v.s are independent if and only if*

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^{d} f_{X_j}(x_j)$$

for all $\mathbf{x} = (x_1, ..., x_d) \in \mathbb{R}^d$.

4.6 Example computations with joint distributions

EXAMPLE 11

Pick a point (X, Y) uniformly from the region $\left\{(x, y) : 0 \leq x \leq 6, y \leq \frac{1}{2}x\right\}$.

1. Determine the joint density of X and Y.
2. Determine the density functions of the marginals.
3. Determine whether X and Y are independent.

EXAMPLE 12

Suppose $X \sim Geom(p)$. Find the density of $X + X$.

EXAMPLE 13

Suppose X and Y are continuous r.v.s whose joint density

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{C}{(x+y)^4} & \text{if } x \geq 1, y \geq 1 \\ 0 & \text{else} \end{cases}$$

1. Determine the value of C.
2. Compute $P(Y \leq 2X)$.
3. Compute the densities of the marginals.
4. Determine if $X \perp Y$.

3. We compute the density of X first, by integrating with respect to ___:

$$\begin{aligned} f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \int_1^{\infty} \frac{24}{(x+y)^4}\,dy \\ &= -8(x+y)^{-3}\Big|_1^{\infty} \\ &= 0 - (-8(x+1)^{-3}) \\ &= 8(x+1)^{-3}. \end{aligned}$$

This holds when $x \geq 1$; otherwise $f_X(x) = 0$. So formally, the density is

$$\boxed{f_X(x) = \begin{cases} 8(x+1)^{-3} & \text{if } x \geq 1 \\ 0 & \text{else} \end{cases}}$$

Next, we compute the density of Y by integrating with respect to x:

$$\begin{aligned} f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \int_1^{\infty} \frac{24}{(x+y)^4}\,dx \\ &= -8(x+y)^{-3}\Big|_1^{\infty} \\ &= 0 - (-8(1+y)^{-3}) \\ &= 8(1+y)^{-3}. \end{aligned}$$

Formally, the answer is

$$\boxed{f_Y(y) = \begin{cases} 8(1+y)^{-3} & \text{if } y \geq 1 \\ 0 & \text{else} \end{cases}}$$

4. To determine whether or not $X \perp Y$, we test as follows:

4.7 Conditional density

RECALL

Given a probability space $(\Omega, \mathcal{A}, P)$ and an *event* E with $P(E) > 0$, we defined the **conditional probability of F given E** by

$$P(F \mid E) = \frac{P(E \cap F)}{P(E)}.$$

Our goal is to create something similar on the level of *random variables*:

QUESTION

Let X, Y be real-valued r.v.s. (either cts or discrete). What is the "probability" of X given a particular value of Y? e.g.

$$\text{“}P(X = x \mid Y = y)\text{''} =$$

Definition 4.16 *Let X and Y be real-valued r.v.s with joint density function $f_{X,Y}$. The* **conditional density of X given** *Y is the function $f_{X|Y} : \mathbb{R}^2 \to [0, \infty)$ defined by*

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

where f_Y is the density of the marginal Y (if $f_Y(y) = 0$, we say $f_{X|Y}(x|y) = 0$).

Theorem 4.17 (Properties of conditional densities) *Let X and Y be real-valued r.v.s. Then:*

Conditional densities are densities: *For every y such that $f_Y(y) > 0$, $f_{X|Y}(x|y)$ is a density function for a random variable $X|Y$ (whose value is x), i.e.*

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y)\,dx = 1.$$

Multiplicative property: *We can compute the joint density of X and Y by multiplying the density of one marginal times the conditional density of the other one, given the first:*

$$f_{X|Y}(x|y) \cdot f_Y(y) = f_{X,Y}(x,y).$$

Conditional probability calculations: *We compute conditional probabilities associated to one r.v. given the value of the other as follows:*

$$P(X \in E \,|\, Y = y) = \begin{cases} \int_E f_{X|Y}(x|y)\,dx & \text{if } X \text{ is cts} \\ \sum_x f_{X|Y}(x|y) & \text{if } X \text{ is discrete} \end{cases}$$

EXAMPLE 14

Suppose X and Y have joint density

$$f_{X,Y}(x,y) = \begin{cases} cy(2-x-y) & \text{if } (x,y) \in [0,1]^2 \\ 0 & \text{else} \end{cases}.$$

1. Find the conditional density of Y given X.
2. Find the conditional density of Y given $X = \frac{1}{3}$.
3. Find the probability that $Y \in \left[\frac{1}{4}, \frac{3}{4}\right]$ given that $X = \frac{1}{3}$.

Solution: 1.

2. Having computed $f_{Y|X}$ in # 1, we compute this simply by plugging in $x = \frac{1}{3}$:

$$f_{Y|X}\left(y|\frac{1}{3}\right) = \frac{6y(2-\frac{1}{3}-y)}{4-3(\frac{1}{3})} = \boxed{2y\left(\frac{5}{3}-y\right) \text{ for } y \in [0,1]}.$$

3. Integrate the conditional density found in # 2:

$$\begin{aligned} P\left(Y \in \left[\frac{1}{4}, \frac{3}{4}\right] \mid X = \frac{1}{3}\right) &= \int_{1/4}^{3/4} f_{Y|X}(y|\frac{1}{3})\, dy \\ &= \int_{1/4}^{3/4} 2y\left(\frac{5}{3}-y\right) dy \\ &= \int_{1/4}^{3/4} \left(\frac{10}{3}y - 2y^2\right) dy \\ &= \left[\frac{5}{3}y^2 - \frac{2}{3}y^3\right]_{1/4}^{3/4} = \boxed{\frac{9}{16}}. \end{aligned}$$

REMARK: If X is cts, there is a big difference between

$$P\left(Y \in \left[\frac{1}{4}, \frac{3}{4}\right] \,\middle|\, X = \frac{1}{3}\right) \qquad \text{and} \qquad P\left(Y \in \left[\frac{1}{4}, \frac{3}{4}\right] \,\middle|\, X \leq \frac{1}{3}\right):$$

EXAMPLE 15

Suppose that $X \sim Exp(\lambda)$, and that $Y|X \sim Exp(x)$.

1. Determine the joint density of X and Y.
2. Compute a density function of Y.

4.8 Transformations of continuous joint distributions

We want to consider two types of transformation problems:

Class 1: Compute the density of a real-valued r.v. U obtained as a function of several r.v.s $X_1, ..., X_d$ which have some given joint distribution.

Example: Given a joint density of X and Y, find a density of $Z = X + 2Y$.

Class 2: Compute the joint density of some r.v.s $U_1, ..., U_d$ obtained as functions of several r.v.s $X_1, ..., X_d$ which have some given joint distribution.

Example: Given a joint density of X and Y, find a joint density of U and V, where $U = X + Y$ and $V = \dfrac{X}{X+Y}$.

We handle problems in each of these two classes separately.

Class 1 Examples

Setup: $\varphi : \mathbb{R}^d \to \mathbb{R}$ is some function; $U = \varphi(X_1, ..., X_d) = \varphi(\overrightarrow{X})$ is real-valued.

Method of solution:

1. Classify U as discrete or continuous.
2. Find the range of U.
3. If U is discrete, compute the density by back-substitution:
$$f_U(u) = P(U = u) = P(\varphi(\mathbf{X}) = u) = P(\mathbf{X} \in \varphi^{-1}(u))$$
$$= \begin{cases} \displaystyle\int_{\varphi^{-1}(u)} f_{\mathbf{X}}(\mathbf{x})\,d\mathbf{x} & \text{if } \mathbf{X} \text{ is cts} \\ \displaystyle\sum_{\mathbf{x} \in \varphi^{-1}(u)} f_{\mathbf{X}}(\mathbf{x}) & \text{if } \mathbf{X} \text{ is discrete} \end{cases}.$$
4. If U is continuous, first compute the cdf of Y by back-substitution:
$$F_U(u) = P(U \le u) = P(\varphi(\mathbf{X}) \le u) = P(\mathbf{X} \in \varphi^{-1}(-\infty, u])$$
$$= \begin{cases} \displaystyle\int_{\varphi^{-1}(-\infty,u]} f_{\mathbf{X}}(\mathbf{x})\,d\mathbf{x} & \text{if } \mathbf{X} \text{ is cts} \\ \displaystyle\sum_{\mathbf{x} \in \varphi^{-1}(-\infty,u]} f_{\mathbf{X}}(\mathbf{x}) & \text{if } \mathbf{X} \text{ is discrete} \end{cases}.$$
Then differentiate F_U with respect to u to obtain f_U.

EXAMPLE 16

Let (X, Y) be independent, exponential r.v.s, both with parameter λ. Determine a density of $X + Y$.

EXAMPLE 17

Suppose that the amount X an insurance company pays in claims and the amount Y it collects in premiums are modeled by a joint density

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{3}{500}x & \text{if } 0 \leq x \leq y \leq 10 \\ 0 & \text{else} \end{cases}$$

Let R be the ratio of premiums to claims; find the distribution function of R.

Class 2 Examples

Setup: $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ is some function (we will assume that φ is invertible, otherwise the problem is much harder); $\mathbf{U} = (U_1, .., U_d) = \varphi(X_1, ..., X_d) = \varphi(\mathbf{X})$ is a joint distribution. $f_{\mathbf{U}}(\mathbf{u})$ = ?

Let's write $\varphi(x_1, ..., x_d) = (u_1, ..., u_d)$ for convenience.

This problem has a theoretical solution: suppose for now that $d = 2$. Then, the joint density of $\mathbf{U}$ should satisfy, for every (measurable) set $E \subseteq \mathbb{R}^2$,

Motivation from Calculus 1: u-substitutions

$$\int_a^b f(\varphi(x))\,\varphi'(x)\,dx \underset{\begin{pmatrix} u = \varphi(x), \\ du = \varphi'(x)\,dx \end{pmatrix}}{=} \int_{\varphi(a)}^{\varphi(b)} f(u)\,du$$

Since this holds for every $E \subseteq \mathbb{R}^2$, we have

$$f_{\mathbf{U}}(u_1, u_2) \cdot |J(\varphi)| = f_{\mathbf{X}}(x_1, x_2) \quad \Rightarrow \quad \boxed{f_{\mathbf{U}}(u_1, u_2) = \frac{1}{|J(\varphi)|} f_{\mathbf{X}}(x_1, x_2)}$$

where $J(\varphi)$ is the **Jacobian** of φ:

$$J(\varphi) = \det \begin{pmatrix} \dfrac{\partial u_1}{\partial x_1} & \dfrac{\partial u_1}{\partial x_2} \\ \\ \dfrac{\partial u_2}{\partial x_1} & \dfrac{\partial u_2}{\partial x_2} \end{pmatrix} = \det \begin{pmatrix} (u_1)_{x_1} & (u_1)_{x_2} \\ \\ (u_2)_{x_1} & (u_2)_{x_2} \end{pmatrix} = \det D\varphi.$$

This generalizes:

Theorem 4.18 (Transformation theorem, higher-dimensions)
Suppose $\mathbf{X} = (X_1, ..., X_d)$ has joint density $f_{\mathbf{X}} : \mathbb{R}^d \to [0, \infty)$.
Suppose that $\mathbf{U} = (U_1, ..., U_d) = \varphi(X_1, ..., X_d) = \varphi(\mathbf{X})$, where $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ is a C^1 function[a].
If the Jacobian determinant

$$J(\varphi) = \det \begin{pmatrix} \frac{\partial u_1}{\partial x_1} & \frac{\partial u_1}{\partial x_2} & \cdots & \frac{\partial u_1}{\partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial u_d}{\partial x_1} & \frac{\partial u_d}{\partial x_2} & \cdots & \frac{\partial u_d}{\partial x_d} \end{pmatrix}_{d \times d}$$

is everywhere nonzero, then the U_j are all continuous and have joint density given by

$$f_{\mathbf{U}}(u_1, ..., u_d) = \frac{1}{|J(\varphi)|} f_{\mathbf{X}}(x_1, ..., x_d),$$

i.e.

$$f_{\mathbf{U}}(\mathbf{u}) = \frac{1}{|J(\varphi)|} f_{\mathbf{X}}(\varphi^{-1}(\mathbf{u})).$$

[a] A function is called C^1 if all its partial derivatives exist everywhere and are continuous.

EXAMPLE 18

Let (X_1, X_2) be uniform on $[0, 1]^2$. Compute a joint density of $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$.

EXAMPLE 19

Suppose $X_1 \sim \Gamma(\alpha, \lambda)$, $X_2 \sim \Gamma(\beta, \lambda)$ and $X_1 \perp X_2$. Find the joint density of $Y_1 = X_1 + X_2$ and $Y_2 = \dfrac{X_1}{X_1 + X_2}$.

4.9 Chapter 4 Homework

Exercises from Section 4.2

1. Suppose X and Y are discrete, integer-valued r.v.s with joint density function

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{2}{9} \cdot \dfrac{2^x}{3^{x+y}} & \text{if } x \geq 0, y \geq 0 \\ 0 & \text{if } x < 0 \text{ or } y < 0 \end{cases}.$$

   a) Verify that this $f_{X,Y}$ is in fact a density function (by showing that its values sum to 1).
   b) Compute the probability that $X = 3$ and $Y = 4$.
      Note: This is one question, asking for the probability that ($X = 3$ and $Y = 4$).
   c) Compute the probability that $X = 2$.
   d) Calculate a density function of the marginal Y.
   e) Based on the computation you did in part (d), how would you describe Y as a common r.v.? (Include any appropriate parameters.)

2. Suppose you have two dice numbered 1 to 6 that you can load however you want (i.e. you can assign whatever probabilities you want to each number on each die). Is it possible to load the dice in such a manner that makes every sum from 2 to 12 equally likely when the dice are rolled independently? If so, explain how. If not, explain why not.

   Hint: Call the two dice X and Y. Let $p_1 = f_X(1) = P(X = 1)$, $p_6 = f_X(6) = P(X = 6)$, $q_1 = f_Y(1) = P(Y = 1)$ and $q_6 = f_Y(6) = P(Y = 6)$. Now, consider the probability that the sum of the numbers rolled is 2 and the probability that the sum of the numbers rolled is 11. What must each of these probabilities be, in terms of p_1, p_6, q_1 and q_6? What must these equal, since every sum is supposed to be equally likely? This gives you two equations involving p_1, p_6, q_1 and q_6. Finally, consider the probability that the sum of the numbers rolled is 7. This will lead you to an inequality involving p_1, p_6, q_1 and q_6 from which you can derive something useful.

3. Let X and Y be r.v.s having joint density function given by the following table:

$Y \backslash X$	-1	0	2	6
-2	$\frac{1}{27}$	$\frac{1}{9}$	$\frac{1}{27}$	$\frac{1}{9}$
1	$\frac{1}{9}$	0	$\frac{1}{9}$	$\frac{2}{9}$
3	0	$\frac{2}{27}$	$\frac{1}{9}$	$\frac{2}{27}$

a) Compute the probability that X is even.

b) Compute the probability that XY is odd.

4. Let X and Y have the joint density given in Exercise 3.

a) Compute the probability that $X > 0$ and $Y \geq 0$.

b) Compute the probability that $X > 0$ or $Y \geq 0$.

5. Let X and Y have the joint density given in Exercise 3.

a) Compute the density function of X.

b) Compute the density function of Y.

6. Let $X \sim Unif(\{0,1\})$ and $Y \sim Unif(\{0,1\})$. Characterize all possible joint distributions of X and Y. For each of these joint distributions, compute the density of $X+Y$.

Hint: The idea here is to think about the most general way in which you could make a chart similar to the ones we made for Examples 4, 5 and 6. For instance, if you put a number a in one of the boxes in that chart, what would have to go in the other boxes?

7. Suppose X and Y are discrete r.v.s, **each taking values on the nonnegative integers**, with joint density function $f_{X,Y}$. For each given probability, write an expression, involving one or more sums, which gives the probability.

As an example, if asked to compute $P(0 \leq X \leq 5, 2 \leq Y \leq 4)$, one possible correct answer is

$$P(0 \leq X \leq 5, 2 \leq Y \leq 4) = \sum_{x=0}^{5} \sum_{y=2}^{4} f_{X,Y}(x,y).$$

a) $P(5 \leq X < 10, 0 < Y < 4)$

Note: in this type of statement, the comma always means "and".

b) $P(X = 6, 9 \leq Y)$

c) $P(X = 5 \text{ or } Y \geq 4)$

d) $P(X = 1)$

8. Same directions as Exercise 7:

a) $P(3 \leq X, 12 \leq Y \leq 20)$

b) $P(X + Y = 11)$

c) $P(X - Y = 9)$

d) $P(0 \leq X \leq Y)$

9. Same directions as Exercise 7:

   a) $P(0 \leq X \leq Y \leq 10)$
   b) $P(X + Y \leq 15)$
   c) $P(X + Y = z)$, where z is a constant
   d) $P(Y - X = z)$, where z is a nonnegative constant

10. Suppose X and Y are as described in Exercise 1.

    a) Compute the probability that $X + Y = 8$.
    Hint: I want an answer with no "Σ"s in it. To evaluate your sum, you will need the formula for a finite geometric sum given on the pink sheet.
    b) Compute the probability that $X + Y \geq 12$ (again, no "Σ"s in your answer are allowed).

Exercises from Section 4.3

11. There are 40 gumballs in a bag, of which 20 are red, 10 are orange, 8 are green, and 2 are purple.

    a) Suppose you randomly draw 15 gumballs from the bag, one at a time, with replacement. What is the probability you draw 5 red, 5 orange, and 5 green gumballs?
    b) Suppose you randomly draw 15 gumballs from the bag simultaneously. What is the probability you draw 5 red, 5 orange, and 5 green gumballs?

Exercises from Sections 4.4 to 4.6

12. Suppose X and Y are continuous r.v.s such that $X \geq 0$ and $Y \geq 0$, with joint density function $f_{X,Y}$. For each given probability, write an expression involving integrals which gives the probability. As an example, if asked to compute $P(0 \leq X \leq 5, 2 \leq Y \leq 4)$, one possible correct answer is

$$P\left(0 \leq X \leq 5, 2 \leq Y \leq 4\right) = \int_0^5 \int_2^4 f_{X,Y}(x, y)\, dy\, dx.$$

    a) $P(3 \leq X < 8, 0 < Y < 5)$
    b) $P(X \geq 4)$
    c) $P(X + Y \leq 8)$
    d) $P(\min(X, Y) \leq 6)$

13. Same directions as Exercise 12:

a) $P(\max(X,Y) \leq 6)$
b) $P(X \leq Y)$
c) $P(Y/X < 5)$
d) $P(X - 2Y > 5)$

14. Repeat Exercise 12, but under the extra assumptions that X and Y take values only in the square whose vertices are $(0,0)$, $(10,0)$, $(0,10)$ and $(10,10)$.

15. Suppose X and Y are continuous r.v.s such that $0 < Y < X$, with joint density function $f_{X,Y}$. For each given probability, write an expression involving integrals which gives the probability.

a) $P(X \geq 14)$
b) $P(Y \leq 2)$
c) $P(X + Y \leq 8)$
d) $P(3 \leq X \leq 10, 5 \leq Y \leq 8)$

16. Suppose X and Y are two continuous real-valued r.v.s with joint density function

$$f_{X,Y}(x,y) = \begin{cases} C\left(x^2 + \dfrac{xy}{2}\right) & \text{if } 0 < x < 1, 0 < y < 2 \\ 0 & \text{else} \end{cases}$$

where C is some constant. Compute each quantity:

a) C
b) $f_X(x)$
c) $P(X > Y)$
d) $P\left(Y > \frac{1}{2} \,\middle|\, X < \frac{1}{2}\right)$

17. Let Ω be the triangle in the $xy-$plane whose vertices are $(0,0)$, $(2,0)$ and $(0,2)$. Suppose X and Y are r.v.s with joint density

$$f_{X,Y}(x,y) = \begin{cases} c\,x^2 y & \text{if } (x,y) \in \Omega \\ 0 & \text{else} \end{cases}$$

where c is some constant.

a) Compute c.
b) Calculate the probability that $X \geq 1$.
c) Calculate the probability that both X and Y are greater than $\frac{1}{2}$.
d) Determine a density function of the marginal Y.
e) Are X and Y independent? Why or why not?

18. (AE) A device runs until either of two components fails, at which point the device stops running. The joint density function of the lifetimes of the two components, both measured in hours, is

$$f(x,y) = \begin{cases} \frac{1}{8}(x+y) & 0 \le x, y \le 2 \\ 0 & \text{else} \end{cases}$$

What is the probability that the device fails during its first hour of operation?

19. (AE) An insurance company insures a large number of drivers. Let X be the r.v. representing the company's losses under collision insurance, and let Y represent the company's losses under liability insurance. X and Y have joint density function

$$f(x,y) = \begin{cases} \frac{1}{4}(2x+2-y) & x \in (0,1), y \in (0,2) \\ 0 & \text{else} \end{cases}$$

What is the probability that the total loss is at least 1?

20. Suppose X and Y are real-valued r.v.s with joint density

$$f_{X,Y}(x,y) = \begin{cases} \lambda^2 e^{-\lambda y} & 0 \le x \le y \\ 0 & \text{else} \end{cases}.$$

a) Compute the marginal densities of X and Y.
b) Compute the probability that $Y \le 4$.

21. Let X and Y denote the coordinates of a point chosen uniformly from the unit square. Let $Z_1 = X^2$, let $Z_2 = Y^2$ and let $Z_3 = X + Y$.

a) Are Z_1 and Z_2 independent? Why or why not? (Give a heuristic argument only.)
b) Are Z_1 and Z_3 independent? Why or why not? (Give a heuristic argument only.)

22. Let X and Y be independent r.v.s, where $X \sim Geom(p)$ and $Y \sim Geom(q)$ (do not assume any relationship between p and q in this problem).

a) Compute $P(X = Y)$.
b) Compute $P(X \ge Y)$.

Exercises from Section 4.7

23. Suppose X and Y are continuous r.v.s with joint density

$$f_{X,Y}(x,y) = \begin{cases} xe^{-x(y+1)} & \text{if } x > 0, y > 0 \\ 0 & \text{else} \end{cases}.$$

Compute the conditional density of X given Y.

24. Suppose X and Y are discrete r.v.s, taking values in the integers, whose joint density is

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{x!\,y!}\lambda^x e^{-\lambda-x-1}(x+1)^y & \text{if } 0 \le x, 0 \le y \\ 0 & \text{else} \end{cases}.$$

Compute the conditional density of Y given $X = 3$.

25. (AE) An insurance company supposes that each person has an accident parameter a and that the yearly number of accidents of someone who has accident parameter a is a Poisson r.v. X with parameter a. The company also supposes that the parameter of a newly insured person is itself a $\Gamma(r, \lambda)$ r.v. If a newly insured person has n accidents in his first year,

   a) Compute the conditional density of his accident parameter.
   b) Identify the conditional density you found in part (a) as the density of a common r.v. (including appropriate parameters).

26. Let $Y \sim Exp(\lambda)$, where λ is itself a r.v. $\Lambda \sim \Gamma(r, \beta)$.

   a) Compute a density of Y.
   b) Compute the conditional density of Λ given $Y = y$.

27. The distribution of Y, given X, is uniform on $[0, X]$. The marginal density of X is $f_X(x) = 2x$ for $0 < x < 1$ ($f_X(x) = 0$ otherwise). Find the conditional density of X given $Y = y$ (where this conditional density is positive).

28. Compute the conditional density $f_{Y|X}$, for the joint density given in Exercise 20.

29. (AE) An auto insurance policy will pay for damage to both the policyholder's car and the other driver's car in the event that the policyholder is responsible for an accident. Assume that the size X of the payment for damage to the policyholder's car is uniform on $(0, 1)$, and that given $X = x$, the size Y of the payment to the other driver's car is uniform on $(x, x + 1)$. If the policyholder is responsible for an accident, what is the probability that the payment for damage to the other driver's car is greater than $\frac{1}{2}$?

30. Suppose X and Y are discrete r.v.s whose joint density is given in the chart in Exercise 3.

    a) Calculate $P(X < 4 \,|\, Y = 1)$.
    b) Calculate $P(Y < 3 \,|\, X = 6)$.

Exercises from Section 4.8

31. Let X and Y be independent r.v.s, where $X \sim Pois(\lambda_1)$ and $Y \sim Pois(\lambda_2)$. Prove that $X+Y$ is Poisson; what is its parameter? (The way you do this for now is to explicitly compute the density function of $X+Y$.)

    NOTE: The fact you just proved in Exercise 31 should be memorized (and will be generalized later).

32. Suppose (X, Y) have joint density

$$f_{X,Y}(x,y) = \begin{cases} 4xy & \text{if } 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{else} \end{cases}$$

    Compute the density of $W = X + Y$.

    Hint: The computation requires separate cases, depending on whether $W \ge 1$ or $W < 1$.

33. (AE) A company offers earthquake insurance. Annual premiums are modeled by an exponential random variable with parameter 1. Annual claims are modeled by an exponential random variable with parameter 2. Assume that the annual premiums and claims are independent; let X denote the ratio of claims to premiums. What is the density function of X?

34. If $X \sim \Gamma(r, \lambda)$, what is the density of $Y = \sqrt{X}$?

35. (AE) The time T that a computer is not working is a random variable whose cumulative distribution function is $F(t) = 1 - \frac{1}{4}t^{-2}$ for $t > 2$. The resulting cost X to the business as a result of the computer malfunctioning is $X = T^2$. Find the density function of X (when $X > 4$).

36. Let X and Y be independent exponential r.v.s, with respective parameters λ and μ. Compute the joint density of X and $Z = X + Y$.

37. Let X and Y be continuous r.v.s with joint density function

$$f_{X,Y}(x,y) = \begin{cases} e^{-y} & \text{if } 0 \le x \le y \\ 0 & \text{else.} \end{cases}.$$

    Compute the joint density of W and Z, where $W = Y/X$ and $Z = X + Y$.

38. Let X and Y be continuous r.v.s with $0 \leq X$ and $0 \leq Y$ that have some unknown density function f. Compute, in terms of f, the joint density of $T = X^2 Y$ and $U = XY$.

39. Let X and Y be independent Poisson r.v.s, with respective parameters λ and μ. Let $Z = X + Y$.

    a) Compute the joint density of X and Z.

       Hint: In terms of X and Z, the joint density of X and Z is $f_{X,Z}(x, z) = P(X = x, Z = z)$. Back-substitute to see what this is in terms of X and Y.

    b) Compute the conditional density of X given Z.

       Hint: You should know what the density of Z is without computing its marginal again (since you studied this situation in Exercise 31).

40. Suppose $X_1, ..., X_d$ are independent, continuous r.v.s.

    a) Let $MAX = \max(X_1, ..., X_d)$ be the maximum of the X_js. Derive a formula for F_{MAX} in terms of the F_{X_j}.

    b) (AE) A company decides to accept the highest of five sealed bids on a property. The sealed bids are regarded as five independent r.v.s, each with common cumulative distribution function

       $$F(x) = \frac{(x-3)^2}{4} \text{ for } 3 \leq x \leq 5.$$

       Find the density function of the accepted bid.

41. Suppose $X_1, ..., X_d$ are independent, continuous r.v.s.

    a) Let $MIN = \min(X_1, ..., X_d)$. Derive a formula for the survival S_{MIN} of the minimum, in terms of the survival functions S_{X_j} of the marginals.

    b) Prove that if $X_1, ..., X_d$ are independent exponential r.v.s with respective parameters $\lambda_1, ..., \lambda_d$, then $\min(X_1, ..., X_d)$ is exponential with parameter $\lambda_1 + ... + \lambda_d$.

NOTE: The facts you prove in Exercises 40 (a) and 41 (a) and (b) are good to memorize for the actuarial exam, and for MATH 416. The maximum and minimum of the r.v.s $X_1, ..., X_d$ are part of what are called the **order statistics** of the X_j.

Chapter 5

Expected value

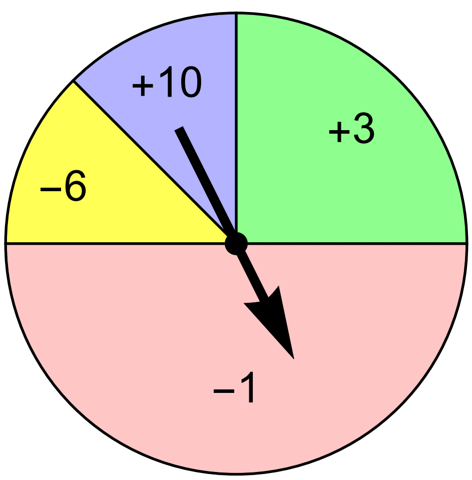
5.1 Definition of expected value

MOTIVATING QUESTION

What is the "average" value of a random variable?

EXAMPLE 1

You and your friend play a game with a spinner. You spin the spinner and then exchange money depending on where the spinner lands:



Flawed definition: The **expected value** of a discrete, real-valued r.v. X, denoted EX, is

$$EX = \sum_{x \in Range(X)} x\, f_X(x).$$

Technical point: The range of X might be an infinite set (i.e. it might be $\mathbb{Z}$). Then there are potential issues with the convergence of the infinite series

$$\sum_{x \in Range(X)} x\, f_X(x)$$

if we try to rearrange terms. To get around any problems, we require that this series *converge absolutely*.

RECALL FROM CALCULUS 2

A series $\sum a_n$ is said to **converge absolutely** if $\sum |a_n|$ converges. Absolutely convergent sequences can be rearranged and/or regrouped without changing the sum of the series.

In our setting, to say $\displaystyle\sum_{x \in Range(X)} x f_X(x)$ converges absolutely means

With this in mind, we make the following definition:

Definition 5.1 *Let $X : \Omega \to \mathbb{R}$ be a discrete r.v., with density f_X. We say X* **has finite expectation** *(and write $EX < \infty$) if*

$$\sum_{x \in Range(X)} |x|\, f_X(x) < \infty;$$

in which case we say the **expected value** *(a.k.a.* **mean** *a.k.a.* **expectation***) of X is the real number*

$$EX = \sum_{x \in Range(X)} x\, f_X(x).$$

If $\displaystyle\sum_{x \in Range(X)} |x|\, f_X(x) = \infty$, we say X **does not have finite expectation** *and we write $EX = \infty$.*

A similar definition works for continuous, real-valued r.v.s:

Definition 5.2 *Let $X : \Omega \to \mathbb{R}$ be a continuous r.v., with density f_X. We say X* **has finite expectation** *(and write $EX < \infty$) if*

$$\int_{-\infty}^{\infty} |x| f_X(x)\, dx < \infty$$

in which case we say the **expected value** *(a.k.a.* **mean** *a.k.a.* **expectation***) of X is the real number*

$$EX = \int_{-\infty}^{\infty} x\, f_X(x)\, dx.$$

If $\int_{-\infty}^{\infty} |x| f_X(x)$ diverges, we say X **does not have finite expectation** *and we write $EX = \infty$.*

Notation: EX is also denoted μ, μ_X, $E[X]$, $E(X)$ and $\mathbb{E}(X)$.

Note: If $X : \Omega \to \mathbb{R}$ is neither discrete nor cts, then it makes no sense to talk about EX.

Also, if X isn't real-valued (such as when $\mathbf{X} : \Omega \to \mathbb{R}^d$ is a joint distribution), it makes no sense to talk about EX.

EXAMPLE 2

Suppose X has density function $f_X(x) = \frac{3}{28}x^2$ for $-1 \leq x \leq 3$ (and $f_X(x) = 0$ otherwise). Compute the expected value of X.

Note: If the range of X is bounded above **and** bounded below (like in Example 2), then $EX < \infty$ is automatic.

If the range of X is either bounded above **or** bounded below, then you can simultaneously check that X has finite expectation and compute EX by computing

$$\sum_x x f_X(x) \text{ (if } X \text{ is discrete) or } \int_{-\infty}^{\infty} x f_X(x)\, dx \text{ (if } X \text{ is continuous).}$$

So in practice, you never actually have to mess with computing $\sum_x |x| f_X(x)\, dx$ or $\int_{-\infty}^{\infty} |x| f_X(x)\, dx$.

LOTUS (Expected values of transformations)

QUESTION

Suppose you know the density of r.v. X. To get the expected value of X, you compute

$$EX = \sum_x x f_x(x) \qquad \text{or} \qquad EX = \int x f_X(x)\, dx.$$

How would you compute the expected value of a transformation of X, i.e. what is EY if $Y = \varphi(X)$?

Long way:

Seemingly dumb way:

Actually, this seemingly dumb way works! It's called "LOTUS", which is an acronym for the Law of the Unconscious Statistician:

Theorem 5.3 (LOTUS) *Suppose* $\mathbf{X} : \Omega \to \mathbb{R}^d$ *is a r.v. with density* $f_{\mathbf{X}}$. *Let* $U = \varphi(\mathbf{X})$ *where* $\varphi : \mathbb{R}^d \to \mathbb{R}$ *is a real-valued function of* d *variables. Then:*

(a) U *has finite expectation if and only if*

$$\begin{cases} \sum_{\mathbf{x}} |\varphi(\mathbf{x})| f_{\mathbf{X}}(\mathbf{x}) < \infty & \text{if } \mathbf{X} \text{ is discrete} \\ \int_{-\infty}^{\infty} |\varphi(x)| f_X(x)\, dx < \infty & \text{if } \mathbf{X} = X \text{ is cts and real-valued} \\ \int_{\mathbb{R}^d} |\varphi(\mathbf{x})| f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} < \infty & \text{if } \mathbf{X} \text{ is cts and vector-valued} \end{cases}$$

(b) if $EU < \infty$, *then*

$$EU = \begin{cases} \sum_{\mathbf{x}} \varphi(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) & \text{if } \mathbf{X} \text{ is discrete} \\ \int_{-\infty}^{\infty} \varphi(x) f_X(x)\, dx & \text{if } \mathbf{X} = X \text{ cts and real-valued} \\ \int_{\mathbb{R}^d} \varphi(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} & \text{if } \mathbf{X} \text{ is cts and vector-valued} \end{cases}$$

Remark 1: In practice, we'll never have to worry about part (a) of this theorem, because we will deal with r.v.s that have finite expectation.

Remark 2: If $\mathbf{X}$ is a joint distribution, then the integrals here are actually multiple integrals. For instance, if $U = \varphi(X, Y)$, then

$$EU = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \varphi(x, y) f_{X,Y}(x, y)\, dA.$$

PROOF (WHEN $\mathbf{X}$ IS DISCRETE) In this case, U is also discrete, so we can denote the values in the range of U $u_1, u_2, \ldots$.

For each j, let $A_j = \varphi^{-1}(u_j) = \{\mathbf{x} \in Range(\mathbf{X}) : \varphi(\mathbf{x}) = u_j\}$.

The A_j form a partition of the range of X.

Now, since $\mathbf{X} \in A_j$ if and only if $U = u_j$, we see that

$$f_U(u_j) = P(U = u_j) = P(\mathbf{X} \in A_j) = \sum_{\mathbf{x} \in A_j} f_{\mathbf{X}}(\mathbf{x}).$$

Therefore

$$E|U| = \sum_j |u_j| f_U(u_j) = \sum_j |u_j| \sum_{x \in A_j} f_{\mathbf{X}}(\mathbf{x}) \quad \text{(from the previous page)}$$
$$= \sum_j \sum_{\mathbf{x} \in A_j} |u_j| f_{\mathbf{X}}(\mathbf{x})$$
$$= \sum_{\mathbf{x} \in Range(\mathbf{X})} |\varphi(x)| f_{\mathbf{X}}(\mathbf{x}).$$

Therefore $EU < \infty$ if and only if $\sum_{\mathbf{x}} |\varphi(\mathbf{x})| f_{\mathbf{X}}(\mathbf{x}) < \infty$, proving statement (a).

For statement (b), repeat the argument that proved part (a), but with no absolute values around the u_j.

The proof of LOTUS when X is continuous is beyond the scope of this course, as it uses a branch of mathematics called *measure theory*. $\square$

EXAMPLE 3

Suppose X has density function $f_X(x) = x + \frac{1}{2}$ for $0 < x < 1$ (and $f_X(x) = 0$ otherwise). Let $Y = 3X^2 + 6X + 7$. Find EY.

EXAMPLE 4

Let $X \sim Pois(\lambda)$. Find $E[e^X]$.

Expected values and survival functions

A useful, alternate method to compute expected values is by means of the survival function. Recall that for a real-valued r.v. X, $S_X(x) = P(X > x) = 1 - F_X(x)$.

Theorem 5.4 (Expected value from survival function) *Suppose X is a random variable taking values in $[0, \infty)$. Then:*

1. *if X is discrete, then $EX = \sum_{x=0}^{\infty} S_X(x)$.*
2. *if X is continuous, then $EX = \int_0^{\infty} S_X(x)\, dx$.*

PROOF If X is discrete, then

$$\begin{aligned}
EX &= \sum_{x=0}^{\infty} x f_X(x) \\
&= 1f_X(1) + 2f_X(2) + 3f_X(3) + 4f_X(4) + ... \\
&= [f_X(1) + f_X(2) + f_X(3) + ...] + [f_X(2) + f_X(3) + ...] + [f_X(3) + ...] \\
&= P(X > 0) + P(X > 1) + P(X > 2) + ... \\
&= \sum_{x=0}^{\infty} P(X > x) = \sum_{x=0}^{\infty} S_X(x).
\end{aligned}$$

If X is continuous, then

$$EX = \lim_{b\to\infty} \int_0^b x f_X(x)\, dx = \lim_{b\to\infty} \left[x F_X(x)|_0^b - \int_0^b F_X(x)\, dx \right]$$

$$\begin{aligned}
&= \lim_{b\to\infty} \left[bF_X(b) - \int_0^b F_X(x)\, dx \right] \\
&= \lim_{b\to\infty} \left[\left(\int_0^b 1\, dx \right) F_X(b) - 1 \int_0^b F_X(x)\, dx \right] \\
&= \lim_{b\to\infty} \left[F_X(b) \int_0^b 1\, dx \right] - \left[\lim_{b\to\infty} F_X(b) \right] \int_0^b F_X(x)\, dx \\
&= \lim_{b\to\infty} \left[F_X(b) \int_0^b 1\, dx \right] - \lim_{b\to\infty} \left[F_X(b) \int_0^b F_X(x)\, dx \right] \\
&= \lim_{b\to\infty} \left[F_X(b) \int_0^b \left[1 - F_X(x)\right]\, dx \right] \\
&= 1 \cdot \lim_{b\to\infty} \int_0^b \left[1 - F_X(x)\right]\, dx \\
&= \int_0^\infty S_X(x)\, dx. \ \square
\end{aligned}$$

EXAMPLE 5

Suppose X is a continuous, real-valued r.v. with cdf $F_X(x) = 1 - \dfrac{x^2}{(x^3+1)^2}$ for $x \geq 0$ ($F_X(x) = 0$ for $x < 0$.) Compute the expected value of X.

Example 6

A dishwasher manufacturer offers a warranty program, under which they agree to cover the full cost of repair of a broken dishwasher within the first five years after purchase and agree to cover one-fourth of the cost of a repair after five years have elapsed from the purchase. If the cost of a repair is $\$160$, and the time until the dishwasher breaks has density $f_T(t) = \dfrac{3}{8}t^{-4}$ for $t > \dfrac{1}{2}$ (and $f_T(t) = 0$ otherwise), compute the expected amount the manufacturer pays to cover repairs.

5.2 Properties of expected value

RECALL (FROM LINEAR ALGEBRA)

Let V and W be vector spaces. A transformation $T : V \to W$ is called **linear** if

The first thing to know about expected value is that it is a linear transformation from the vector space of random variables to the vector space $\mathbb{R}$:

Theorem 5.5 (Linearity of Expected Value) *Suppose X and Y are real-valued r.v.s with $EX < \infty$ and $EY < \infty$. Then:*

1. *$X + Y$ has finite expectation and $E[X + Y] = EX + EY$.*
2. *For any constant c, cX has finite expectation and $E[cX] = c\,EX$.*

PROOF Suppose X and Y have finite expectation.

For the first statement, let $Z = X + Y = \varphi(X, Y)$. Then if X and Y are discrete,

$$\begin{aligned}\sum_{x,y} |x + y| f_{X,Y}(x, y) &\leq \sum_{x,y} (|x| + |y|) f_{X,Y}(x, y) \\ &= \sum_{x,y} |x| f_{X,Y}(x, y) + \sum_{x,y} |y| f_{X,Y}(x, y).\end{aligned}$$

and if X and Y are continuous,

$$\begin{aligned}\int_{\mathbb{R}^2} |x + y| f_{X,Y}(x, y)\, dA &\leq \int_{\mathbb{R}^2} (|x| + |y|) f_{X,Y}(x, y)\, dA \\ &= \int_{\mathbb{R}^2} |x| f_{X,Y}(x, y)\, dA + \int_{\mathbb{R}^2} |y| f_{X,Y}(x, y)\, dA.\end{aligned}$$

Since $EX < \infty$, the red sum/integral is finite, and since $EY < \infty$, the blue sum/integral is finite.
So the entire (red + blue) expression is finite, so by LOTUS, $E[X + Y] < \infty$.
Now, again using LOTUS,

$$E[X + Y] = \sum_{x,y} (x + y) f_{X,Y}(x, y) = \sum_{x,y} x f_X(x, y) + \sum_{x,y} y f_{X,Y}(x, y) = EX + EY.$$

Now for the second statement: if X is discrete, then so is cX and

$$\sum_x |cx| f_X(x) = |c| \sum_x |x| f_X(x) < \infty$$

so by LOTUS cX has finite expectation. Then, again using LOTUS,

$$E[cX] = \sum_x cx f_X(x) = c \sum_x x f_X(x) = cEX.$$

If X is continuous, the same proof works using integrals instead of sums. □

Theorem 5.6 (Expectation preserves constants) *Let X be a real-valued r.v. If $P(X = c) = 1$, then $EX = c$.*

PROOF If $P(X = c) = 1$, then X is discrete and $f_X(c) = 1$. So

$$EX = \sum_x x f_X(x) = c \cdot 1 = c. \ \square$$

EXAMPLE 7

Suppose $EX = 8$ and $EY = -3$. Compute the expected value of $2X + 5Y + 3$.

Inequality properties

Theorem 5.7 (Inequality Properties of Expected Value) *Suppose X and Y are real-valued r.v.s with $EX < \infty$ and $EY < \infty$. Then:*

Positivity: *If $P(X \geq 0) = 1$, then $EX \geq 0$.*

Monotonicity: *If $P(X \geq Y) = 1$, then $EX \geq EY$.*

Triangle inequality: $|EX| \leq E|X|$.

Preservation of bounds: *If $P(|X| \leq M) = 1$, then $|EX| \leq M$.*

Definiteness: *If $P(X \geq Y) = 1$ and $EX = EY$ then $P(X = Y) = 1$.*

PROOF We begin by proving positivity. If $P(X \geq 0) = 1$ and X is discrete, then

$$EX = \sum_x x f_X(x) \geq 0$$

since all the numbers in the sum are nonnegative.
If $P(X \geq 0) = 1$ and X is continuous, then $Range(X) \subseteq [0, \infty)$ so

$$EX = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_0^{\infty} x f_X(x)\,dx \geq 0$$

since the integrand is positive. This proves positivity.

Next, we prove monotonicity: let $Z = X - Y$; by linearity, $EZ = EX - EY$.
If $P(X \geq Y) = 1$, then $P(Z \geq 0) = 1$.
So by positivity, $EZ \geq 0$.
Thus $EX - EY \geq 0$ so $EX \geq EY$, proving monotonicity.

To establish the triangle inequality, suppose $-|X| \leq X \leq |X|$.
This implies $-E|X| \leq EX \leq E|X|$ by monotonicity. Thus $|EX| \leq E|X|$.

Preservation of bounds follows from the triangle inequality and monotonicity.

For definiteness, again let $Z = X - Y$, since $EX = EY$ we have $EZ = 0$.
Assuming Z is discrete, repeating the argument we made for positivity, we have (since $P(Z \geq 0) = 1$)

$$EZ = \sum_{z=0}^{\infty} z f_Z(z) = 0$$

and since all the zs in the sum are ≥ 0 and all the $f_Z(z)$s are ≥ 0, the only way this can be consistent with $\sum_z f_Z(z) = 1$ is if $f_Z(0) = 1$ (otherwise there would be a positive term without any negative term that could cancel it).
Thus $P(Z = 0) = 1$ so $P(X - Y = 0) = 1$ so $P(X = Y) = 1$.
If Z is continuous, the proof of definiteness is harder (take MATH 430). $\square$

Theorem 5.8 (Independence Properties of Expected Value) *Suppose X and Y are real-valued r.v.s with $EX < \infty$ and $EY < \infty$. If $X \perp Y$, then for any functions $\varphi, \psi : \mathbb{R} \to \mathbb{R}$, if $\varphi(X)$ and $\psi(Y)$ both have finite expectation, then so does $\varphi(X)\psi(Y)$, and*

$$E[\varphi(X)\psi(Y)] = E[\varphi(X)] \cdot E[\psi(Y)].$$

In particular, $E[XY] = EX \cdot EY$.

WARNING: The converse of this is false, i.e. $E[XY] = EX \cdot EY$ does not imply $X \perp Y$.

PROOF Note that $X \perp Y$ implies $f_{X,Y}(x,y) = f_X(x) f_Y(y)$.
Then, if X and Y are discrete,

$$\begin{aligned}\sum_{x,y} |\varphi(x)\psi(y)| f_{X,Y}(x,y) &= \sum_x \sum_y |\varphi(x)||\psi(y)| f_X(x) f_Y(y) \\ &= \left(\sum_x |\varphi(x)| f_X(x)\right) \left(\sum_y |\psi(y)| f_Y(y)\right),\end{aligned}$$

and if X and Y are continuous,

$$\begin{aligned}\int_{\mathbb{R}^2} |\varphi(x)\psi(y)| f_{X,Y}(x,y)\, dA &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |\varphi(x)||\psi(y)| f_X(x) f_Y(y)\, dx\, dy \\ &= \left(\int_{-\infty}^{\infty} |\varphi(x)| f_X(x)\, dx\right) \left(\int_{-\infty}^{\infty} |\psi(y)| f_Y(y)\, dy\right).\end{aligned}$$

Since $E[\varphi(X)] < \infty$, the red sum/integral is finite.
Since $E[\psi(Y)] < \infty$, the blue sum/integral is finite.
Thus the entire expression is finite so by LOTUS, $E[\varphi(X)\psi(Y)] < \infty$.
So if $\varphi(X)\psi(Y)$ is discrete,

$$\begin{aligned}E[\varphi(X)\psi(Y)] = \sum_{x,y} \varphi(x)\psi(y) f_{X,Y}(x,y) &= \sum_x \sum_y \varphi(x)\psi(y) f_X(x) f_Y(y) \\ &= \left(\sum_x \varphi(x) f_X(x)\right) \left(\sum_y \psi(y) f_Y(y)\right) \\ &= E[\varphi(X)] \cdot E[\psi(Y)].\end{aligned}$$

(and if $\varphi(X)\psi(Y)$ is cts, the same type of computation works with integrals). □

5.3 Variance

MOTIVATION

Here are some random variables, all of which have mean 10:

$$X = 10 \qquad X \sim Unif(\{9, 11\}) \qquad X \sim Unif([0, 20])$$

To distinguish these r.v.s, we can think of how much the values of the r.v. are spread out. To do this, we use a quantity called *variance*:

Definition 5.9 *Let $X : \Omega \to \mathbb{R}$ be a r.v. such that $EX < \infty$ and $E[(X - EX)^2] < \infty$. The* **variance** *of X, denoted $Var(X)$ (or $V(X)$ or σ^2 or σ_X^2), is*

$$Var(X) = E[(X - EX)^2].$$

The **standard deviation** *of X, denoted σ or σ_X, is $\sigma = \sqrt{Var(X)}$.*

Observations:

1. $Var(X) \geq 0$.
2. The more spread out X is, the further from zero $X - EX$ is, so the greater $Var(X)$ is. Thus variance is a measure of spread of a random variable.

Theorem 5.10 (Variance of a constant) *Let X be a real-valued r.v. $Var(X) = 0$ if and only if X is constant (i.e. $\exists\, c$ s.t. $P(X = c) = 1$).*

PROOF ($\Rightarrow$) Suppose $Var(X) = 0$. Then $E[(X - EX)^2] = 0$.
Since $(X - EX)^2 \geq 0$, by definiteness, that means $P((X - EX)^2 = 0) = 1$.
This is equivalent to $P(X = EX) = 1$, i.e. X is constant with probability 1.

($\Leftarrow$) Suppose X is constant, say $X = c$.
Then $EX = c$ so $(X - EX)^2 = (c - c)^2 = 0$, and therefore $Var(X) = E[0] = 0$. □

Theorem 5.11 (Variance Formula) *Let X be a real-valued r.v. so that $Var(X)$ exists. Then*

$$\begin{aligned} Var(X) &= EX^2 - (EX)^2 \\ &= \text{“second moment”} - \text{“mean squared”}. \end{aligned}$$

PROOF This is just algebra, together with properties of expected value:

$$\begin{aligned} Var(X) &= E[(X - EX)^2] \\ &= E[(X - EX)(X - EX)] \\ &= E[X^2 - 2(EX)X + (EX)^2] \end{aligned}$$

EXAMPLE 8

Suppose X is a continuous r.v. with density $f_X(x) = \frac{1}{4}x^3$ for $0 \leq x \leq 2$. Compute the variance of X.

Theorem 5.12 (Properties of Variance) *Let X be a r.v. with finite variance. Then:*

1. *For any constant b, $Var(X + b) = Var(X)$;*
2. *For any constant a, $Var(aX) = a^2 Var(X)$.*

PROOF HW (as a hint, these follow from either the definition of variance or the variance formula)

EXAMPLE 9

Suppose X is a r.v. with variance 12. Compute the variance of $5X + 8$.

5.4 Expected values and variances of common random variables

Theorem 5.13 *Expected values and variances of common r.v.s are as follows:*

X	EX	$Var(X)$
$Unif(\{1,2,...,n\})$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
$Geom(p)$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
$NB(r,p)$	$r\left(\frac{1-p}{p}\right)$	$r\left(\frac{1-p}{p^2}\right)$
binomial(n,p)	np	$np(1-p)$
$Pois(\lambda)$	λ	λ
$Hyp(n,r,k)$	$\frac{kr}{n}$	$\frac{kr}{n}\left(\frac{n-r}{n}\right)\frac{n-k}{n-1}$
$Unif([a,b])$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Exp(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$\Gamma(r,\lambda)$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$
Cauchy	∞	*DNE*
std. normal $n(0,1)$	0	1
normal $n(\mu,\sigma^2)$	μ	σ^2

Remark: "Standard normal" and "normal" random variables will be introduced in Chapter 6.

PROOF (OF SOME OF THESE) In the homework, you will prove the expected value formulas when X is hypergeometric, exponential, and gamma, and the variance formulas when X is continuous uniform, exponential and gamma.

$X \sim Unif(\{1, 2, ..., n\})$:

$$EX = \sum_x x f_X(x) = \sum_{x=1}^{n} x \frac{1}{n} =$$

$$EX^2 = \sum_x x^2 f_X(x) = \sum_{x=1}^{n} x^2 \frac{1}{n} = \frac{1}{n} \sum_{x=1}^{n} x^2 = \frac{1}{n}\left[\frac{n(n+1)(2n+1)}{6}\right] = \frac{(n+1)(2n+1)}{6};$$

$$Var(X) = EX^2 - (EX)^2 = \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \cdots = \boxed{\frac{n^2-1}{12}}.$$

$X \sim Pois(\lambda)$:

$$EX = \sum_x x f_X(x) = \sum_{x=0}^{\infty} x e^{-\lambda} \frac{\lambda^x}{x!} =$$

$$EX^2 = \sum_x x^2 f_X(x) = \sum_{x=0}^{\infty} x^2 e^{-\lambda} \frac{\lambda^x}{x!} =$$

$$Var(X) = EX^2 - (EX)^2 = [\lambda^2 + \lambda] - \lambda^2 = \boxed{\lambda}.$$

$X \sim b(n, p)$:

$$\begin{aligned}
EX = \sum_x x f_X(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} &= \sum_{x=1}^{n} x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} \\
&= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x} \\
&= np \sum_{x=0}^{n-1} \frac{(n-1)!}{x!(n-x-1)!} p^x (1-p)^{n-x-1} \\
&= np \sum_{x=0}^{n-1} \binom{n-1}{x} p^x (1-p)^{n-1-x} \\
&= np[p + (1-p)]^{n-1} \\
&= np(1) = \boxed{np}.
\end{aligned}$$

$\underline{X \sim Unif([a,b])}$:

$\underline{X \sim Cauchy}$:

5.5 Covariance and correlation

MOTIVATING QUESTION

We have seen variance is **not** linear, because $Var(aX) = a^2Var(X)$, not $a\,Var(X)$.

But we haven't looked at whether or not variance respects addition. In particular, does $Var(X+Y) = Var(X) + Var(Y)$? If not, what is a formula for $Var(X+Y)$ in terms of $Var(X)$ and $Var(Y)$?

Answer:

Definition 5.14 *Given two r.v.s X and Y, each having finite variance, the* **covariance** *between X and Y, denoted $Cov(X,Y)$ (or $C(X,Y)$ or $\sigma_{X,Y}$ or σ_{XY}) is*

$$Cov(X,Y) = E\left[(X - EX)(Y - EY)\right].$$

The covariance between two random variables measures the "tendency of the r.v.s to change together". In other words:

- If $Cov(X,Y) > 0$, then as X increases, we expect Y to increase and as X decreases, we expect Y to decrease.
- If $Cov(X,Y) < 0$, then as X increases, we expect Y to decrease and as X decreases, we expect Y to increase.
- If $Cov(X,Y) = 0$, then changes in X should not lead to any expected change in Y.

Properties of covariance

Theorem 5.15 (Bilinearity of covariance) *Suppose that all the r.v.s mentioned in these equations are real-valued, and have finite mean and variance. Then:*

1. $Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$;
2. $Cov(X, Y_1 + Y_2) = Cov(X, Y_1) + Cov(X, Y_2)$;
3. *For any constant* a, $Cov(aX, Y) = a\,Cov(X, Y) = Cov(X, aY)$.

PROOF HW

Theorem 5.16 (Properties of covariance) *Let* X *and* Y *be real-valued r.v.s having finite variance. Then:*

Covariance formula: $Cov(X, Y) = E[XY] - EX \cdot EY$.

Symmetry: $Cov(X, Y) = Cov(Y, X)$.

Self-covariance is variance: $Cov(X, X) = Var(X)$.

Variance sum formula: $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$.

PROOF The variance sum formula was established earlier.

For the covariance formula, notice

$$\begin{aligned}
Cov(X, Y) &= E[(X - EX)(Y - EY)] \\
&= E[XY - EX \cdot Y - EY \cdot X + EX \cdot EY] \\
&= E[XY] - E[EX \cdot Y] - E[EY \cdot X] + E[EX \cdot EY] \\
&= E[XY] - EX \cdot EY - EY \cdot EX + EX \cdot EY \\
&= E[XY] - EX \cdot EY.
\end{aligned}$$

Symmetry of covariance is obvious from the definition.

To prove that self-covariance is variance, observe

$$Cov(X, X) = E[XX] - EX \cdot EX = EX^2 - (EX)^2 = Var(X). \ \square$$

Theorem 5.17 (Independent r.v.s have zero covariance) *Suppose that X and Y are real-valued r.v.s with finite mean and variance. If $X \perp Y$, then $Cov(X,Y) = 0$ and $Var(X+Y) = Var(X) + Var(Y)$.*

WARNING: The converse of this is false. There are r.v.s X and Y with covariance 0 that are not independent.

PROOF If $X \perp Y$, then $E[XY] = EX \cdot EY$ by a previous theorem.
Therefore $Cov(X,Y) = E[XY] - EX \cdot EY = 0$. □

A PROBLEM WITH COVARIANCE

Suppose X and Y, both measured in hours, have covariance 2. Then if we let X_M and Y_M be the same quantities as X and Y, *but measured in minutes rather than hours*, we have

$$Cov(X_M, Y_M) =$$

Thus the covariance between two quantities *depends greatly on the units the quantities are measured in*. We don't really want this, because the covariance is "supposed" to measure how correlated the random variables are. To fix this, we invent a new quantity called "correlation":

Definition 5.18 *Given two r.v.s X and Y, each having finite variance, the* **correlation** *between X and Y, denoted $\rho(X,Y)$ (or ρ_{XY} or $\rho_{X,Y}$) is*

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}.$$

X and Y are **uncorrelated** *if $\rho(X,Y) = 0$ (equivalently, if $Cov(X,Y) = 0$).*

From Theorem 5.17, independent r.v.s are uncorrelated, but heed the warning after Theorem 5.17: uncorrelated r.v.s may not be independent (we'll see a specific example in the HW).

Theorem 5.19 (Schwarz Inequality) *Let X and Y be real-valued r.v.s with finite variances. Then*
$$(E[XY])^2 \leq EX^2 \cdot EY^2.$$

PROOF The proof of the Schwarz inequality has two cases:

Case 1: If $P(Y=0)=1$, then
$$E([XY])^2 = 0 \leq 0 = EX^2 \cdot 0 = EX^2 \cdot EY^2$$
as desired.

Case 2: Suppose $P(Y=0)<1$.

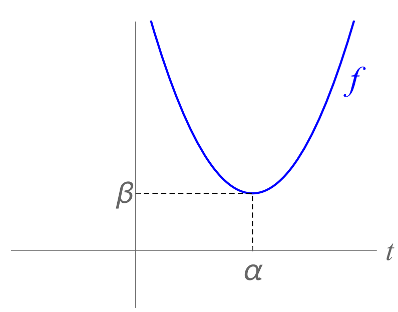
This implies $P(Y^2=0)<1$ so $E[Y^2]>0$; this will allow us to divide through by EY^2 later on (which we couldn't do in Case 1).

Now, define a function $f:\mathbb{R}\to\mathbb{R}$ by
$$f(t) = E[(X-tY)^2].$$

Note that $f(t)\geq 0$ for all t since f is the expected value of a nonnegative r.v.

Expanding f, we get
$$\begin{aligned} f(t) = E[(X-tY)(X-tY)] &= E[X^2 - 2tXY + t^2Y^2] \\ &= EX^2 - 2tE[XY] + t^2EY^2. \end{aligned}$$

Thus f is a quadratic function of t whose graph is a parabola that opens upward. Since $f(t)\geq 0$ for all t, the vertex (α,β) of this parabola must lie above the t-axis:



Now, let's find the coordinates of this vertex using some calculus:
$$f'(t) = 2tEY^2 - 2E[XY]$$

Set $f'(t)=0$ and solve for t (a.k.a. α) to get $\alpha = \dfrac{E[XY]}{EY^2}$.

The y-coordinate of the vertex is therefore

$$\begin{aligned}\beta = f(\alpha) = f\left(\frac{E[XY]}{EY^2}\right) &= EX^2 - 2\left(\frac{E[XY]}{EY^2}\right)E[XY] + \left(\frac{E[XY]}{EY^2}\right)^2 EY^2 \\ &= EX^2 - 2\frac{(E[XY])^2}{EY^2} + \frac{(E[XY])^2}{EY^2} \\ &= EX^2 - \frac{(E[XY])^2}{EY^2}.\end{aligned}$$

Putting this all together, we have

$$\begin{aligned} & 0 \le \beta \\ \Rightarrow \quad & 0 \le EX^2 - \frac{(E[XY])^2}{EY^2} \\ \Rightarrow \quad & \frac{(E[XY])^2}{EY^2} \le EX^2 \\ \Rightarrow \quad & (E[XY])^2 \le EX^2 \cdot EY^2 \end{aligned}$$

which is the Schwarz inequality. □

Some context

You may recall from linear algebra another inequality called the Cauchy-Schwarz Inequality (important in the context of computing projections of one vector onto another, angles between vectors, etc.). That inequality is basically the same as this one; it says that for two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

$$|\mathbf{x} \cdot \mathbf{y}| \le ||\mathbf{x}||\,||\mathbf{y}||$$

where $||\;||$ denotes the norm or length of a vector (recall that $||\mathbf{x}|| = \sqrt{\mathbf{x} \cdot \mathbf{x}}$). Denoting the “dot product” of two random variables as “$X \cdot Y$” $= E[XY]$, the Schwarz inequality here is exactly the same thing as the C-S inequality from linear algebra... if you square both sides of the C-S inequality you get

$$\begin{aligned}(\mathbf{x} \cdot \mathbf{y})^2 &\le ||\mathbf{x}||^2\,||\mathbf{y}||^2 \\ (\mathbf{x} \cdot \mathbf{y})^2 &\le (\mathbf{x} \cdot \mathbf{x})\,(\mathbf{y} \cdot \mathbf{y}).\end{aligned}$$

Properties of correlation

Theorem 5.20 (Properties of Correlation) *Let X and Y be r.v.s with finite variance. Then:*

Correlation is symmetric: $\rho(X,Y) = \rho(Y,X)$.

Self-correlation is 1: $\rho(X,X) = 1$.

Correlation is between -1 and 1: $|\rho(X,Y)| \leq 1$.

$\perp$ r.v.s are uncorrelated: *If $X \perp Y$, then $\rho(X,Y) = 0$ (the converse of this is false).*

Correlation is unchanged under linear transformations: *For any positive constants a and b, and any constants c and d,*

$$\rho(aX + c, bY + d) = \rho(X,Y).$$

Correlation of ± 1 implies linear relationship: $\rho(X,Y) = \pm 1$ *if and only if there are constants a and b (with $a \neq 0$) such that $Y = aX + b$.*

PROOF The first statement is clear, since $Cov(X,Y) = Cov(Y,X)$.

The second is a direct calculation:

$$\rho(X,X) = \frac{Cov(X,X)}{\sqrt{Var(X) \cdot Var(X)}} = \frac{Var(X)}{\sqrt{(Var(X))^2}} = \frac{Var(X)}{Var(X)} = 1.$$

For the bounds on ρ, apply the Schwarz Inequality to $X - EX$ and $Y - EY$:

$$E[(X - EX)(Y - EY)]^2 \leq E[(X - EX)^2] \cdot E[(Y - EY)^2]$$

i.e. $Cov(X,Y)^2 \leq Var(X) \cdot Var(Y)$.
Take the square root of both sides to get

$$|Cov(X,Y)| \leq \sqrt{Var(X) \cdot Var(Y)}$$

i.e.

$$|\rho(X,Y)| = \frac{|Cov(X,Y)|}{\sqrt{Var(X) \cdot Var(Y)}} \leq 1.$$

The fact that independent r.v.s are uncorrelated follows from the fact that $X \perp Y$ implies $Cov(X,Y) = 0$.

The last two statements are HW problems. □

EXAMPLE 10

Suppose X and Y are chosen from $[0,1]^2$ with joint density $f_{X,Y}(x,y) = x+y$. Compute the correlation between X and Y.

Solution: Compute a lot of expected values using LOTUS:

$$\begin{aligned}
EX &= \int_0^1 \int_0^1 x(x+y)\,dy\,dx = \cdots = \frac{7}{12} \\
EY &= \int_0^1 \int_0^1 y(x+y)\,dy\,dx = \cdots = \frac{7}{12} \\
EX^2 &= \int_0^1 \int_0^1 x^2(x+y)\,dy\,dx = \cdots = \frac{5}{12} \\
EY^2 &= \int_0^1 \int_0^1 y^2(x+y)\,dy\,dx = \cdots = \frac{5}{12} \\
EXY &= \int_0^1 \int_0^1 xy(x+y)\,dy\,dx = \cdots = \frac{1}{3}
\end{aligned}$$

Then

$$\begin{aligned}
Var(X) &= EX^2 - (EX)^2 = \frac{5}{12} - \left(\frac{7}{12}\right)^2 = \frac{11}{144} \\
Var(Y) &= EY^2 - (EY)^2 = \frac{5}{12} - \left(\frac{7}{12}\right)^2 = \frac{11}{144} \\
Cov(X,Y) &= EXY - EX \cdot EY = \frac{-1}{144}
\end{aligned}$$

and finally,

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\frac{-1}{144}}{\sqrt{\left(\frac{11}{144}\right)\left(\frac{11}{144}\right)}} = \frac{\frac{-1}{144}}{\frac{11}{144}} = \boxed{-\frac{1}{11}}.$$

5.6 Conditional expectation and conditional variance

Definition 5.21 *Let X and Y be real-valued r.v.s. The* **conditional expectation of Y given X**, *also called the* **regression of Y on X**, *is the function*

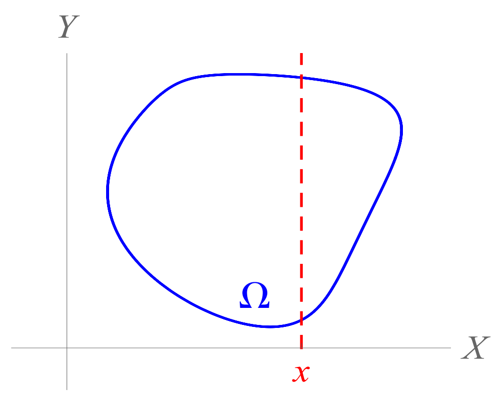
$$E(Y|X) = \begin{cases} \sum_y y f_{Y|X}(y|x) & \text{if } Y|X \text{ is discrete} \\ \int_{-\infty}^{\infty} y f_{Y|X}(y|x)\, dy & \text{if } Y|X \text{ is cts} \end{cases}.$$

In this setting, there is also a conditional expectation of X given Y, defined by

$$E(X|Y) = \begin{cases} \sum_x x f_{X|Y}(x|y) & \text{if } X|Y \text{ is discrete} \\ \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx & \text{if } X|Y \text{ is cts} \end{cases}.$$

Important: $E(Y|X)$ is a function of x, not a number.
(Similarly, $E(X|Y)$ is a function of y.)
That said, we can think of $E(Y|X)$ as a r.v. by thinking of it as a function of X: as an example, if $E(Y|X)(x) = x^2 - 3x$, we can also write $E(Y|X) = X^2 - 3X$.

What does conditional expectation mean? As an example, suppose X and Y are chosen from this set Ω with some joint density function:



Theorem 5.22 (Properties of conditional expectation) *Suppose that any r.v.s mentioned in this theorem have finite expectation, and let c be an arbitrary constant. Then:*

Law of Total Expectation: $E[E(Y|X)] = EY$. *This means:*

$$\int_{-\infty}^{\infty} E(Y|X)(x) f_X(x)\,dx = EY \text{ if } X \text{ is cts}$$
$$\text{or } \sum_x E(Y|X)(x) f_X(x) = EY \text{ if } X \text{ is disrete}$$

Linearity: $E(Y_1 + Y_2 \,|\, X) = E(Y_1|X) + E(Y_2|X)$ *and* $E(cY|X) = c\,E(Y|X)$

Independence: *The following are equivalent:*

- $X \perp Y$
- $E(Y|X)$ *is a constant function.*
- $E(Y|X) = EY$ *for all* x.

Preservation of constants: $E[c|X] = c$ *for any constant* c.

Stability/"pulling out what's given": *For any function* φ,

$$E[\varphi(X)Y|X] = \varphi(X)\,E[Y|X].$$

In particular, $E(X|X) = X$.

Useful integral formulas when computing conditional expectations

Gamma integral formula: $\displaystyle\int_0^{\infty} x^{r-1}e^{-\lambda x}\,dx = \frac{\Gamma(r)}{\lambda^r}$

Beta integral formula: $\displaystyle\int_0^{1} x^{\alpha-1}(1-x)^{\beta-1}\,dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

Gaussian integral formula: (coming later) $\displaystyle\int_{-\infty}^{\infty} e^{\frac{-(x-\mu)^2}{2\sigma^2}}\,dx = \sigma\sqrt{2\pi}$

Note: If you recognize the conditional density $f_{Y|X}$ as a common density, then you can immediately conclude the value of $E(Y|X)$ from the facts known about expected values of common r.v.s.

EXAMPLE 11

Suppose the conditional density of Y given X is (for $x, y > 0$)

$$f_{Y|X}(y|x) = xe^{-xy}.$$

Then we know

EXAMPLE 12

Let X and Y have joint density

$$f_{X,Y}(x,y) = \begin{cases} \frac{12}{5}y(2-x-y) & \text{if } (x,y) \in [0,1]^2 \\ 0 & \text{else} \end{cases}.$$

Find the conditional expectation of Y given X and the conditional expectation of Y given $X = \frac{1}{3}$.

EXAMPLE 13

Three contestants on a game show are given the same question, and each person answers the question correctly with probability $1 - x$ (their answers are independent). The difficulty x of the question is itself a r.v. chosen from $(0, 1)$ with density function $6x(1 - x)$. Find the expected difficulty level of the question, given that all three contestants answer incorrectly.

Conditional variance

Definition 5.23 *Let X and Y be real-valued r.v.s. The* **conditional variance of Y given X**, *is the function*

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

That the two formulas given in the box above are the same is a HW problem.

As with conditional expectation, the conditional variance is a function of x (and can be thought of as a random variable).

Theorem 5.24 (Law of Total Variance) *Let X and Y be real-valued r.v.s. Then*

$$Var(Y) = E[Var(Y|X)] + Var[E(Y|X)].$$

PROOF HW (use the definitions and crunch the symbols appropriately)

This theorem is extremely useful for computing the variance of Y, when X and $Y|X$ are given as common random variables:

EXAMPLE 14

The number of accidents on a stretch of highway is uniform on $\{1, 2, 3, ..., 9\}$. Given N accidents on the stretch of highway, the total amount of damage caused by the accidents is exponential with mean $2N$. Find the variance of the total amount of damage caused by accidents on this stretch of highway.

5.7 Probability generating functions

What is a generating function?

Take a sequence of numbers $a_0, a_1, a_2, a_3, \ldots$. To record this sequence, you can write down the entire sequence, or take the numbers and put them in as terms in a power series

$$f(t) = \sum_{n=0}^{\infty} a_n t^n = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + \ldots$$

This gives you a function of t, called the **generating function** of the sequence $\{a_n\}$. There are a couple of reasons why we would want to do this:

- the formula for $f(t)$ may be easier/shorter to write than the formula for a_n;
- properties of the generating function may give you useful information about the sequence.

In our setting, we start with a discrete r.v. X taking values in $\{0, 1, 2, 3, \ldots\}$. This naturally gives you a sequence coming from the probabilities of each value of X:

$$f_X(0), f_X(1), f_X(2), f_X(3), f_X(4), \ldots$$

The generating function associated to this sequence is therefore

This is called the *probability generating function* of X, and it turns out that this function has many useful properties.

Definition 5.25 *Let $X : \Omega \to \mathbb{N}$ be a discrete r.v., taking values only in $\{0, 1, 2, 3, \ldots\}$. The* **probability generating function** *of X (a.k.a.* **pgf** *or* **generating function***), denoted G_X or Φ_X, is the function $G_X : [-1, 1] \to \mathbb{R}$ defined by*

$$G_X(t) = E[t^X] = \sum_{x=0}^{\infty} f_X(x) t^x.$$

Note: The t in this definition is just a dummy variable. It doesn't really have any meaning.

Properties of probability generating functions

Theorem 5.26 (Properties of PGFs) *Let X be a discrete r.v. taking values in $\mathbb{N}$. Then:*

1. *G_X is a continuous and differentiable function of t on $[-1, 1]$.*
2. *$G_X(1) = 1$.*
3. *$G_X(0) = f_X(0) = P(X = 0)$ (the constant term on G_X).*
4. *$|G_X(t)| \leq 1$ for all t.*

PROOF Statement (1) follows from the fact that G_X is a power series (in Calculus 2, we learn that all power series are cts and diffble).
For statement (2), observe $G_X(1) = E[1^X] = E[1] = 1$.
For (3), notice $G_X(0) = f_X(0) = P(X = 0)$, the constant term on G_X.
For the last statement, note $|G_X(t)| = |E[t^X]| \leq E|t^X| \leq E[1] = 1$. □

Theorem 5.27 (PGFs and expectations) *Let X be a discrete r.v. taking values in $\mathbb{N}$. Then:*

1. *$G'_X(1) = EX$.*
2. *$G''_X(1) = E[X(X-1)] = EX^2 - EX$.*
3. *$G^{(r)}_X(1) = E[X(X-1)(X-2)(X-3)\cdots(X-r)]$.*
   (This quantity is called the r^{th} **factorial moment** *of X.)*
4. *$Var(X) = G''_X(1) + G'_X(1) - [G'_X(1)]^2$.*
5. *The equation $G_X(t) = t$ has a solution in $(0, 1)$ if and only if $EX > 1$.*

PROOF For statement (1), notice $G_X(t) = \sum_{x=0}^{\infty} t^x f_X(x)$ so $G'_X(t) = \sum_{x=1}^{\infty} x t^{x-1} f_X(x)$.
Therefore $G'_X(1) = \sum_{x=1}^{\infty} x 1^{x-1} f_X(x) = \sum_{x=1}^{\infty} x f_X(x) = \sum_{x=0}^{\infty} x f_X(x) = EX$.
To prove statement (2), differentiate $G'_X(t)$ to get $G''_X(t) = \sum_{x=2}^{\infty} x(x-1)t^{x-2} f_X(x)$.
This means

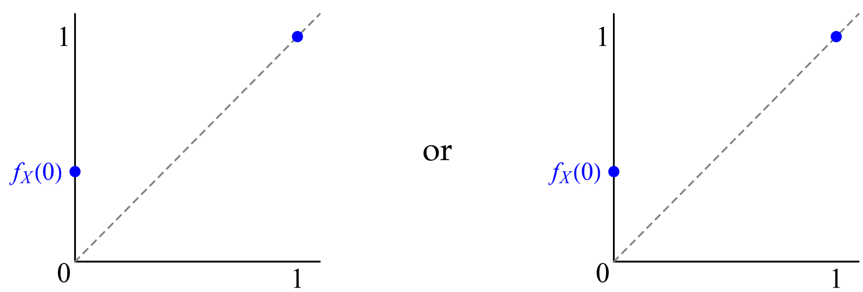
$$\begin{aligned} G''_X(1) &= \sum_{x=2}^{\infty} x(x-1)1^{x-2} f_X(x) = \sum_{x=2}^{\infty} x(x-1) f_X(x) \\ &= \sum_{x=0}^{\infty} x(x-1) f_X(x) = E[X(X-1)]. \end{aligned}$$

Statement (3) has a similar proof as (2), but uses induction.

Statement (4) follows from (1), (2) and the variance formula.

For the last statement, notice that the graph of G_X:
- is continuous (part (1) of Theorem 5.26),
- passes through $(1, 1)$ (part (2) of Theorem 5.26) with slope EX (statement (1) of this theorem),
- and passes through $(0, f_X(0))$ (part (1) of this theorem).

So the graph of G_X looks like



Theorem 5.28 (Independence property of PGFs) *Let $X : \Omega \to \mathbb{N}$ and $Y : \Omega \to \mathbb{N}$ be independent r.v.s with respective PGFs G_X and G_Y. Then*

$$G_{X+Y}(t) = G_X(t)\, G_Y(t).$$

PROOF This is a direct calculation:

$$\begin{aligned} G_{X+Y}(t) = E[t^{X+Y}] = E[t^X t^Y] &= E[t^X]E[t^Y] \text{ (since } X \perp Y\text{)} \\ &= G_X(t)G_Y(t). \ \square \end{aligned}$$

Theorem 5.29 (Uniqueness of PGFs) *Let X and Y be discrete r.v.s taking values in $\mathbb{N}$. Then:*

Inversion formula for PGFs: $f_X(n) = \dfrac{G_X^{(n)}(0)}{n!}$ *for all* $n \in \{0, 1, 2, 3, ...\}$

Uniqueness of PGFs: *If $G_X(t) = G_Y(t)$, then $X \sim Y$.*

PROOF The first part of this is the uniqueness of power series from Calculus 2. That means we can determine a r.v.'s density from its PGF. Thus if $G_X = G_Y$, $f_X = f_Y$, i.e. $X \sim Y$. $\square$

Theorem 5.30 (PGFs of common r.v.s) *For the discrete r.v.s encountered in Chapter 2, their probability generating functions are as follows:*

X	$G_X(t)$
$Unif(\{1,2,...,n\})$	$\dfrac{t(t^n-1)}{n(t-1)}$
$Geom(p)$	$\dfrac{p}{1-t(1-p)}$
$NB(r,p)$	$\left[\dfrac{p}{1-t(1-p)}\right]^r$
binomial(n,p)	$(pt+1-p)^n$
$Pois(\lambda)$	$e^{\lambda(t-1)}$

PROOFS (OF SOME OF THESE) The uniform discrete r.v. is left as HW.

A main application of probability generating functions is to derive facts about the sums of independent random variables. These arguments combine the PGFs of common r.v.s with the independence property of PGFs and the uniqueness of PGFs:

Theorem 5.31 *Suppose $X_1, ..., X_d$ are independent r.v.s, and let $S = X_1 + ... + X_d$. Then:*

1. *If each $X_j \sim Pois(\lambda_j)$, then $S \sim Pois(\lambda_1 + ... + \lambda_d)$.*
2. *If each $X_j \sim b(n_j, p)$ (same p), then $S \sim b(n_1 + ... + n_d, p)$.*
3. *If each $X_j \sim Geom(p)$ (same p), then $S \sim NB(d, p)$.*
4. *If each $X_j \sim NB(r_j, p)$ (same p), then $S \sim NB(r_1 + ... + r_d, p)$.*

PROOF First, we prove statement (1).

For statement (2), suppose $X_j \sim b(n_j, p)$. Then $G_{X_j}(t) = (pt + 1 - p)^{n_j}$ for each j, so

$$G_S(t) = \prod_{j=1}^{d} G_{X_j}(t) = \prod_{j=1}^{d} (pt + 1 - p)^{n_j} = (pt + 1 - p)^{\sum_j n_j} = G_{b\left(\sum_j n_j, p\right)}(t).$$

By uniqueness of PGFs, $S \sim b\left(\sum_j n_j, p\right)$.

Statement (3) is HW; the proof of statement (4) is similar and omitted. $\square$

5.8 Moments and moment generating functions

A DRAWBACK OF PGFS

Definition 5.32 *Let $X : \Omega \to \mathbb{R}$ and let $r \in \{0, 1, 2, 3, ...\}$. If X^r has finite expectation, then we define the r^{th}* **moment** *of X, denoted μ_r, to be $E[X^r]$. Otherwise, we say X* **does not have a moment of order** *r.*

Heuristic analogy:

r	interpretation of $f^{(r)}(0)$	interpretation of EX^r
0	$f(0) =$ height of f at $x = 0$	$EX^0 = 1$
1	$f'(0) =$ slope of f at $x = 0$	$EX^1 = EX =$ mean
2	$f''(0) =$ concavity of f at $x = 0$	$EX^2 =$ variance (sort of)
3	$f'''(0) =$ jerk	$EX^3 =$ skewness (sort of)

Let's take the moments of r.v. X and put them in a sequence:

$$1, EX, EX^2, EX^3, EX^4, ...$$

We could directly construct a generating function from this sequence, but since these moments are supposed to be like derivatives, we'll take some inspiration from Calculus 2 and divide the r^{th} moment by $r!$ (kind of like how you divide $f^{(r)}(0)$ by $r!$ to get the coefficient on x^r in the Taylor series of f). This gives us a sequence

$$1,\ EX,\ \frac{1}{2}EX^2,\ \frac{1}{3!}EX^3,\ \frac{1}{4!}EX^4,\ , ...$$

which has generating function

$$
\begin{aligned}
&1 + EXt + \frac{1}{2}EX^2t^2 + \frac{1}{3!}EX^3t^3 + \frac{1}{4!}EX^4t^4 + \ldots \\
&= E[1] + E[tX] + E\left[\frac{(tX)^2}{2}\right] + E\left[\frac{(tX)^3}{3!}\right] + E\left[\frac{(tX)^4}{4!}\right] + \ldots \\
&= E\left[1 + tX + \frac{(tX)^2}{2} + \frac{(tX)^3}{3!} + \frac{(tX)^4}{4!} + \ldots\right]
\end{aligned}
$$

This leads to the following definition:

Definition 5.33 *Given real-valued $r.v.$ X (X can be cts or discrete), the* **moment generating function (MGF)** *of X, denoted M_X or g_X, is defined by*

$$M_X(t) = E[e^{tX}].$$

The domain of M_X is the set of all $t \in \mathbb{R}$ such that e^{tX} has finite expectation.

EXAMPLE 15

Suppose X is a continuous r.v. taking values in $[0, 1]$ with density $f_X(x) = \dfrac{1}{e-1}e^x$. Compute the moment generating function of X.

Properties of moment generating functions

Many properties of MGFs are similar to those of PGFs:

Theorem 5.34 (Properties of MGFs) *Let $X : \Omega \to \mathbb{R}$ be a r.v. with mgf $M_X(t)$. Then:*

1. $M_X(0) = 1$.
2. **Expected value from MGF:** $M'_X(0) = EX = \mu_1$.
3. $M''_X(0) = EX^2 = \mu_2$.
4. **Moment formula:** *For all $r \in \{1, 2, 3, ...\}$, $M_X^{(r)}(0) = \mu_r = E[X^r]$.*
5. **Variance from MGF:** $Var(X) = M''_X(0) - [M'_X(0)]^2$.
6. **Linear translation formula:** *For any a and b, $M_{aX+b}(t) = e^{bt} M_X(at)$.*

PROOF The first five statements come from equating coefficients on two different ways of writing M_X as a power series:

$$M_X(t) = 1 + EXt + \frac{1}{2}EX^2t^2 + \frac{1}{3!}EX^3t^3 + \frac{1}{4!}EX^4t^4 + ...$$

$$= M_X(0) + M'_X(0)t + \frac{1}{2}M''_X(0)t^2 + \frac{1}{3!}M'''_X(0)t^3 + \frac{1}{4!}M_X^{(4)}(0)t^4 + ...$$

The last statement is a direct computation:

$$M_{aX+b}(t) = E\left[e^{(aX+b)t}\right] = E\left[e^{bt}e^{X(at)}\right] = e^{bt}E\left[e^{X(at)}\right] = e^{bt}M_X(at). \square$$

Theorem 5.35 (Independence property of MGFs) *Let $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ be independent r.v.s with respective mgfs M_X and M_Y. Then*

$$M_{X+Y}(t) = M_X(t)\,M_Y(t).$$

Similarly, if $X_1, ..., X_d$ are independent r.v.s with respective mgfs $M_{X_1}, M_{X_2}, ..., M_{X_d}$, then:

$$M_{\sum_{j=1}^{d} X_j}(t) = \prod_{j=1}^{d} M_{X_j}(t).$$

PROOF HW (similar to proof for PGFs)

Moment generating functions of common random variables

Theorem 5.36 (MGFs of common r.v.s) *For the common classes of random variables encountered in Chapters 2 and 3, their moment generating functions are as follows:*

X	$M_X(t)$
$Unif(\{1,2,...,n\})$	$\dfrac{e^t(e^{nt}-1)}{n(e^t-1)}$
$Geom(p)$	$\dfrac{p}{1-(1-p)e^t}$
$NB(r,p)$	$\left[\dfrac{p}{1-(1-p)e^t}\right]^r$
binomial(n,p)	$\left(1-p+pe^t\right)^n$
$Pois(\lambda)$	$e^{\lambda(e^t-1)}$
$Unif([a,b])$	$\dfrac{e^{tb}-e^{ta}}{t(b-a)}$
$Exp(\lambda)$	$\dfrac{\lambda}{\lambda-t}$ *for* $t<\lambda$
$\Gamma(r,\lambda)$	$\left(\dfrac{\lambda}{\lambda-t}\right)^r$ *for* $t<\lambda$
std. normal $n(0,1)$	$e^{t^2/2}$
normal $n(\mu,\sigma^2)$	$\exp\left(\mu t+\dfrac{\sigma^2t^2}{2}\right)$

PROOFS (OF SOME OF THESE) First, whenever X is discrete, then

$$M_X(t)=E[e^{tX}]=E[(e^t)^X]=G_X(e^t)$$

so the MGFs of all the discrete r.v.s come from replacing any ts in the PGF with e^t. Exponential and gamma r.v.s are left as HW; let's do the uniform cts r.v. here:

5.9 Uniqueness of MGFs

It turns out that you can explicitly recover the density function of a real-valued r.v. from its moment generating function:

Theorem 5.37 (Inversion formula) *Let $X : \Omega \to \mathbb{R}$ have mgf M_X. Then:*

1. *If X is discrete and integer-valued, then for every $x \in \mathbb{Z}$,*
$$f_X(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ixt} M_X(it)\, dt.$$
2. *If X is continuous, then X has density*
$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ixt} M_X(it)\, dt.$$

WARNING: If you are ever using these formulas to do a MATH 414 or 416 problem, you are doing the problem wrong.

We'll use formula (2) once, to discover one important fact later in the course.

Gaussian integral formula

To prove the inversion formulas, we first need the following important integral formulas (which will also be used for other purposes later):

Lemma 5.38 (Basic Gaussian Integral Formula)
$$\int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \sqrt{2\pi}.$$

PROOF Let $A = \displaystyle\int_{-\infty}^{\infty} e^{-x^2/2}\, dx$. ($A > 0$ since the integrand is positive.) Then

$$\begin{aligned}
A^2 = (A)(A) &= \left(\int_{-\infty}^{\infty} e^{-x^2/2}\, dx\right)\left(\int_{-\infty}^{\infty} e^{-x^2/2}\, dx\right) \\
&= \left(\int_{-\infty}^{\infty} e^{-x^2/2}\, dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2}\, dy\right) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-x^2/2} e^{-y^2/2}\, dx\, dy \\
&= \iint_{\mathbb{R}^2} e^{-(x^2+y^2)/2}\, dA
\end{aligned}$$

Continuing from the previous page, we next perform the u-sub

$$u = \frac{r^2}{2}, du = r\, dr$$

on the inside integral to get

$$\begin{aligned} A^2 = \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} r\, dr\, d\theta &= \int_0^{2\pi} \int_0^{\infty} e^{-u}\, du\, d\theta \\ &= \int_0^{2\pi} \left[-e^{-u} \right]_0^{\infty} d\theta \\ &= \int_0^{2\pi} d\theta \\ &= 2\pi. \end{aligned}$$

Since $A^2 = \pi$ and $A > 0$, $A = \sqrt{2\pi}$ as wanted. □

Theorem 5.39 (Gaussian Integral Formula) *Let μ, σ be constants with $\sigma > 0$. Then*

$$\int_{-\infty}^{\infty} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2} \right] dx = \sigma\sqrt{2\pi}.$$

PROOF Perform the u-sub $u = \dfrac{x-\mu}{\sigma}$, $du = \dfrac{1}{\sigma} dx$ in the integral to obtain

$$\begin{aligned} \int_{-\infty}^{\infty} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2} \right] dx &= \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] dx \\ &= \int_{-\infty}^{\infty} \exp\left[-\frac{u^2}{2} \right] \sigma\, du \\ &= \sigma \int_{-\infty}^{\infty} e^{-u^2/2}\, du \\ &= \sigma\sqrt{2\pi}. \ \square \end{aligned}$$

Observe: The value of this integral does not depend on μ (only on σ).

The Gaussian Integral Formula can be combined with an algebraic technique called *completing the square* to compute lots of integrals:

EXAMPLE 16

Compute this integral:

$$\int_{-\infty}^{\infty} e^{-2x^2+20x-39}\,dx$$

Solution: The goal is to rewrite the integral so that it matches the Gaussian Integral Formula given on the previous page:

$$\int_{-\infty}^{\infty} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]\,dx = \sigma\sqrt{\pi}$$

$$\int_{-\infty}^{\infty} \exp\left[-2x^2+20x-39\right]\,dx$$

At this point, our integral becomes

$$\begin{aligned}\int_{-\infty}^{\infty} \exp\left[-2(x-5)^2\right] e^{11}\,dx &= e^{11}\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}(4)(x-5)^2\right]\,dx \\ &= e^{11}\int_{-\infty}^{\infty} \exp\left[\frac{-(x-5)^2}{2\left(\frac{1}{4}\right)}\right]\,dx \\ &= e^{11}\int_{-\infty}^{\infty} \exp\left[\frac{-(x-5)^2}{2\left(\frac{1}{2}\right)^2}\right]\,dx\end{aligned}$$

This matches the Gaussian Integral Formula with $\mu = 5$, $\sigma = \frac{1}{2}$ so the integral evaluates to $\boxed{e^{11}\frac{1}{2}\sqrt{\pi}}$.

Proof of the inversion formula

Here is a proof of the inversion formula when X is continuous (the proof when X is discrete is similar, but omitted from these notes). Recall that our goal is to show

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixt} M_X(it)\,dt.$$

The proof is just a long calculation. The first step is to start with the right-hand side and insert an additional term in the inversion formula needed to make it look more like a Gaussian integral:

$$\begin{aligned}
\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixt} M_X(it)\,dt &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixt}(1) M_X(it)\,dt \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixt}\left(\lim_{\epsilon\to 0^+} e^{-\epsilon t^2}\right) M_X(it)\,dt \\
&= \lim_{\epsilon\to 0^+}\frac{1}{2\pi}\int_{-\infty}^{\infty} M_X(it) e^{-\epsilon t^2} e^{-ixt}\,dt
\end{aligned}$$

The second step is to expand the $M_X(it)$ term using the definition of MGF and LOTUS. This gives

$$\begin{aligned}
\lim_{\epsilon\to 0^+}\frac{1}{2\pi}\int_{-\infty}^{\infty} M_X(it) e^{-\epsilon t^2} e^{-ixt}\,dt &= \lim_{\epsilon\to 0^+}\frac{1}{2\pi}\int_{-\infty}^{\infty} E\left[e^{itX}\right] e^{-\epsilon t^2} e^{-ixt}\,dt \\
&= \lim_{\epsilon\to 0^+}\frac{1}{2\pi}\int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f_X(y) e^{ity}\,dy\right] e^{-\epsilon t^2} e^{-ixt}\,dt \\
&= \lim_{\epsilon\to 0^+}\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_X(y) e^{it(y-x)} e^{-\epsilon t^2}\,dy\,dt \\
&= \lim_{\epsilon\to 0^+}\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_X(y) e^{it(y-x)} e^{-\epsilon t^2}\,dt\,dy
\end{aligned}$$

Now, pull the $f_X(y)$ out of the dt integral, and evaluate the inside integral by completing the square and using the Gaussian Integral Formula:

$$\begin{aligned}
&\lim_{\epsilon\to 0^+}\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_X(y) e^{it(y-x)} e^{-\epsilon t^2}\,dt\,dy \\
&= \lim_{\epsilon\to 0^+}\frac{1}{2\pi}\int_{-\infty}^{\infty} f_X(y)\int_{-\infty}^{\infty} \exp\left[-\epsilon\left(t^2 - \frac{i(y-x)}{\epsilon}t\right)\right]\,dt\,dy \\
&= \lim_{\epsilon\to 0^+}\frac{1}{2\pi}\int_{-\infty}^{\infty} f_X(y)\int_{-\infty}^{\infty} \exp\left[-\epsilon\left(t^2 - \frac{i(y-x)}{\epsilon}t + \frac{-(y-x)^2}{4\epsilon^2} + \frac{(y-x)^2}{4\epsilon^2}\right)\right]\,dt\,dy \\
&= \lim_{\epsilon\to 0^+}\frac{1}{2\pi}\int_{-\infty}^{\infty} f_X(y)\int_{-\infty}^{\infty} \exp\left[\frac{-\left(t - \frac{i(y-x)}{2\epsilon}\right)^2}{2\cdot\left(\frac{1}{\sqrt{2\epsilon}}\right)^2}\right] \exp\left[\frac{-(y-x)^2}{4\epsilon}\right]\,dt\,dy
\end{aligned}$$

$$= \lim_{\epsilon \to 0^+} \frac{1}{2\pi} \int_{-\infty}^{\infty} f_X(y) \sqrt{\frac{1}{2\epsilon}} \sqrt{2\pi} \, \exp\left[\frac{-(y-x)^2}{4\epsilon}\right] dy$$

(Gaussian Integral Formula with $\mu = \frac{i(y-x)}{2\epsilon}$, $\sigma = \frac{1}{\sqrt{2\epsilon}}$)

$$= \lim_{\epsilon \to 0^+} \frac{1}{2\sqrt{\pi\epsilon}} \int_{-\infty}^{\infty} f_X(y) \, \exp\left[\frac{-(y-x)^2}{4\epsilon}\right] dy.$$

Next, use the u-sub $u = \frac{y-x}{2\sqrt{\epsilon}}$, $du = \frac{1}{2\sqrt{\epsilon}} dy$ to write the integral as

$$\begin{aligned}
&\lim_{\epsilon \to 0^+} \frac{1}{2\sqrt{\pi\epsilon}} \int_{-\infty}^{\infty} f_X(y) \, \exp\left[\frac{-(y-x)^2}{4\epsilon}\right] dy \\
&= \lim_{\epsilon \to 0^+} \frac{1}{2\sqrt{\pi\epsilon}} \int_{-\infty}^{\infty} f_X(x + 2\sqrt{\epsilon}\, u) \exp\left[-u^2\right] \, 2\sqrt{\epsilon}\, du \\
&= \lim_{\epsilon \to 0^+} \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} f_X(x + 2\sqrt{\epsilon}\, u) \, e^{-u^2} \, du.
\end{aligned}$$

Finally, move the limit back inside the integral and use the Gaussian Integral Formula one more time:

$$\begin{aligned}
&\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \left[\lim_{\epsilon \to 0^+} f_X(x + 2\sqrt{\epsilon}\, u)\right] \, e^{-u^2} \, du \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} f_X(x) e^{-u^2} \, du \\
&= \frac{1}{\sqrt{\pi}} f_X(x) \int_{-\infty}^{\infty} \exp\left[\frac{-u^2}{2 \cdot \left(\frac{1}{\sqrt{2}}\right)^2}\right] \, du \\
&= \frac{1}{\sqrt{\pi}} f_X(x) \frac{1}{\sqrt{2}} \sqrt{2\pi}
\end{aligned}$$

(Gaussian Integral Formula with $\mu = 0$, $\sigma = \frac{1}{\sqrt{2}}$)

$$= f_X(x).$$

This proves the inversion formula (when X is continuous). □

The significance of the inversion formulas is that they explain the following principle:

Corollary 5.40 (Uniqueness of MGFs) *Let $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ be any two real-valued r.v.s so that $M_X(t) = M_Y(t)$. Then $X \sim Y$.*

Sums of independent common r.v.s

As with PGFs, an important application of MGFs is to establish results about the sum of independent random variables:

Theorem 5.41 (Sums of $\perp$ r.v.s) *Suppose $X_1, ..., X_d$ are independent r.v.s, and let $S = X_1 + ... + X_d$. Then:*

1. *If each $X_j \sim Pois(\lambda_j)$, then $S \sim Pois(\lambda_1 + ... + \lambda_d)$.*
2. *If each $X_j \sim b(n_j, p)$ (same p), then $S \sim b(n_1 + ... + n_d, p)$.*
3. *If each $X_j \sim Geom(p)$ (same p), then $S \sim NB(d, p)$.*
4. *If each $X_j \sim NB(r_j, p)$ (same p), then $S \sim NB(r_1 + ... + r_d, p)$.*
5. *If each $X_j \sim Exp(\lambda)$ (same λ), then $S \sim \Gamma(d, \lambda)$.*
6. *If each $X_j \sim \Gamma(r_j, \lambda)$ (same λ), then $S \sim \Gamma(r_1 + ... + r_d, \lambda)$.*

PROOF (OF SOME OF THESE) Statement (6) is left as HW.

5.10 Joint moment generating functions

Definition 5.42 *Let $X_1, ..., X_d$ be real-valued r.v.s with some joint distribution $\mathbf{X}$. The* **joint moment generating function** *of $\mathbf{X}$, denoted $M_{\mathbf{X}}$ or $g_{\mathbf{X}}$, is the function $M_{\mathbf{X}} : \mathbb{R}^d \to \mathbb{R}$ defined by*

$$M_{\mathbf{X}}(\mathbf{t}) = E[e^{\mathbf{t}\cdot\mathbf{X}}].$$

The domain of $M_{\mathbf{X}}$ is the set of all $\mathbf{t} \in \mathbb{R}^d$ such that $e^{\mathbf{t}\cdot\mathbf{X}}$ has finite expectation.

Many of the same properties of MGFs carry over to the joint case:

Theorem 5.43 (Properties of joint MGFs) *Let $X_1, ..., X_d$ be real-valued r.v.s with joint MGF $M = M_{\mathbf{X}}$. Then:*

$M(\mathbf{0}) = 1$.

MGF of marginals: *For each $j \in \{1, ..., d\}$,*

$$M_{X_j}(t) = M_{\mathbf{X}}(0, 0, ..., 0, t, 0, ..., 0) \text{ (the } t \text{ is in the } j^{th} \text{ position).}$$

MGF of linear combination of marginals:

For any constants $a_1, ..., a_d$, $M_{a_1X_1+...+a_dX_d}(t) = M_{\mathbf{X}}(a_1t, ..., a_dt)$.

In vector language, this says that for any $\mathbf{a} = (a_1, ..., a_d)$, $M_{\mathbf{a}\cdot\mathbf{X}}(t) = M_{\mathbf{X}}(t\mathbf{a})$.

Moment formulas: *For each $j \in \{1, ..., d\}$,*

$$E[X_j] = \left.\frac{\partial M_{\mathbf{X}}}{\partial t_j}\right|_{\mathbf{t}=\mathbf{0}} \text{ and } E[X_j^r] = \left.\frac{\partial^r M_X}{\partial t_j^r}\right|_{\mathbf{t}=\mathbf{0}}.$$

Product moment formulas: *For any nonnegative integers $r_1, ..., r_d$,*

$$E[X_1^{r_1} X_2^{r_2} \cdots X_d^{r_d}] = \left.\frac{\partial^{r_1+...+r_d} M_{\mathbf{X}}}{\partial t_1^{r_1} \partial t_2^{r_2} \cdots \partial t_d^{r_d}}\right|_{\mathbf{t}=\mathbf{0}}.$$

Linear translation formula: *For any $a \in \mathbb{R}$ and $\mathbf{b} \in \mathbb{R}^d$,*

$$M_{a\mathbf{X}+\mathbf{b}}(\mathbf{t}) = e^{\mathbf{b}\cdot\mathbf{t}} M_{\mathbf{X}}(a\,\mathbf{t}).$$

Theorem 5.44 *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be two joint distributions of the same dimension.*

Inversion formula for joint MGFs: *If* $\mathbf{X}$ *is continuous, then*

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\mathbf{x}\cdot\mathbf{t}} M_{\mathbf{X}}(\mathbf{t})\, d\mathbf{t}.$$

Uniqueness of joint MGFs: *If* $M_{\mathbf{X}} = M_{\mathbf{Y}}$*, then* $\mathbf{X} \sim \mathbf{Y}$.

Theorem 5.45 (Independence test using joint MGF) *Let* $X_1, ..., X_d$ *be real-valued r.v.s. Then* $X_1, ..., X_d$ *are independent if and only if*

$$M_{\mathbf{X}}(\mathbf{t}) = \prod_{j=1}^{d} M_{X_j}(t_j)$$

for all $\mathbf{t} = (t_1, ..., t_d) \in \mathbb{R}^d$.

PROOF ($\Rightarrow$) Suppose the X_j are independent. Then

$$M_{\mathbf{X}}(\mathbf{t}) = E\left[e^{\mathbf{t}\cdot\mathbf{X}}\right] =$$

($\Leftarrow$) Suppose $M_{\mathbf{X}}(\mathbf{t}) = \prod_{j=1}^{d} M_{X_j}(t_j)$.

By uniqueness of joint MGFs, it must be that the r.v.s are independent (for if they weren't, their joint MGF would have to be something other than what we computed in the ($\Rightarrow$) direction). □

EXAMPLE 17

Suppose X and Y are real-valued r.v.s with joint MGF

$$M_{X,Y}(s,t) = \exp\left(-s^2 - 3s - 6st - 2t^2\right).$$

1. Compute the moment generating function of X.
2. Compute the expected value of $3X - 2Y$.
3. Compute $Cov(X,Y)$.

5.11 Markov and Chebyshev inequalities

In this section we discuss inequalities which give us quick bounds on certain probabilities related to the mean and variance of a random variable.

Theorem 5.46 (Markov inequality) *Let $X : \Omega \to [0, \infty)$ be a nonnegative r.v. with finite expected value. Then for all $a > 0$,*

$$P(X \geq a) \leq \frac{EX}{a}.$$

PROOF Let $I : \Omega \to \{0, a\}$ be defined by

$$I(\omega) = \begin{cases} a & \text{if } X \geq a \\ 0 & \text{else} \end{cases}$$

Notice that $X \geq I$, so

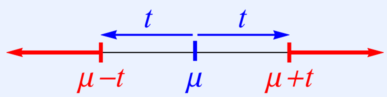
$$\begin{aligned} EX \geq EI &= a \cdot P(I = a) + 0 \cdot P(I = 0) \\ &= a\, P(X \geq a). \end{aligned}$$

Divide both sides by a to get the result. □

EXAMPLE 18

Suppose the time it takes for a radioactive element to decay is a random variable whose mean is 23. Use the Markov inequality to find an upper bound on the probability that it will take at least 230 units of time for the element to decay.

Theorem 5.47 (Chebyshev inequality) *Let $X : \Omega \to \mathbb{R}$ be a r.v. with finite expected value μ and finite variance σ^2. Then for all $t > 0$,*



$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} = \frac{Var(X)}{t^2}.$$

PROOF Apply the Markov inequality to the r.v. $(X - \mu)^2$ with $a = t^2$ to get

$$P((X - \mu)^2 \geq t^2) \leq \frac{E[(X - \mu)^2]}{t^2} = \frac{Var(X)}{t^2}.$$

But $P(|X - \mu| \geq t) = P((X - \mu)^2 \geq t^2)$. This proves the result. □

EXAMPLE 19

Suppose the number of items produced in a factory is a random variable with mean 100 and variance 40. Use the Chebyshev inequality to find a lower bound on the probability that between 90 and 110 items will be produced by the factory.

5.12 Chapter 5 Homework

Exercises from Section 5.1

1. Compute the expected value of each given r.v. X:

   a) X is cts and has density function $f(x)$ defined by $f(x) = \dfrac{3}{4}(1 - x^2)$ for $x \in (-1, 1)$ and $f(x) = 0$ otherwise.

   b) X has cdf $F_X(x)$ defined by $F_X(x) = 1 - \dfrac{5}{x}$ if $x \geq 5$ and $F_X(x) = 0$ otherwise.

   c) X is the marginal of the joint distribution obtained when one selects a point (X, Y) uniformly from the triangle with vertices $(0, 0)$, $(4, 0)$ and $(0, 4)$.

   d) X takes values in $\{0, 1, 2, ...\}$ and has survival function $S_X(x) = \dfrac{1}{x!}$.

NOTE: in all HW exercises from this point forward, you may assume without proof that all r.v.s under consideration have finite expectation.

2. a) Suppose $W \sim binomial(4, \frac{1}{3})$. Compute $E\left[\sin\left(\frac{\pi W}{2}\right)\right]$, evaluating all the trig expressions and simplifying your answer.

   b) Suppose $X \sim Pois(5)$. Calculate the mean of $(1 + X)^{-1}$.

   c) Let Y be the sine of an angle chosen uniformly from $(-\pi/2, \pi/3)$. Compute the expected value of Y.

3. Suppose you play a carnival game that works like this: there are two bags, each with discs numbered 1 to 5 in them. You draw one disc uniformly from each bag. Whatever disc is the smaller number you draw, you win that amount of money (for example, if you draw a 2 and a 4, you would win 2).

   a) How much would you expect to win if you played this game 100 times?

   b) How much should the person running the game charge you if she expects to make a profit of .30 per game?

   c) Suppose that there were n discs in each bag, numbered 1 to n. How much would you now expect to win if you played the same game 100 times?

Hint: The summation formulas

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2} \text{ and/or } \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

may be useful.

4. (AE) A new plasma TV costs $\$650$. The lifetime of the TV is exponentially distributed with parameter $\lambda = \frac{1}{4}$. Best Buy sells a warranty where they give a full refund to a buyer if the TV fails within the first two years, they give a half refund to a buyer if the TV fails during the third or fourth year, and they give no refund otherwise. How much should Best Buy expect to pay in refunds, if they sell 1000 plasma TVs?

5. (AE) Let T_1 be the time between a car accident and the reporting of a claim to an insurance company; let T_2 be the time between the reporting of this claim and the payment of this claim. Assume that (T_1, T_2) is uniform on the region of points (t_1, t_2) satisfying $0 < t_1 < 16$; $0 < t_2 < 16$; $0 < t_1 + t_2 < 20$. Find the expected amount of time between the accident and the payment of the claim.

6. Suppose that the density function f_X of X is:

$$f_X(x) = \begin{cases} a + bx^2 & \text{if } 0 \le x \le 1 \\ 0 & \text{else} \end{cases}.$$

If $EX = \frac{3}{5}$, determine the values of a and b.

Exercises from Section 5.2

7. Suppose X has expected value 3 and Y has expected value -1.

   a) What is the expected value of $3X - 5Y$?
   b) What is the expected value of $2X + 4$?
   c) What is the range of possible values of $E|Y|$?
   d) If $P(Z \le X) = 1$, what is the range of possible values of EZ?
   e) If $X \perp Y$, what is $E[3XY]$?

Exercises from Section 5.3

8. Suppose X is a cts r.v. with density f given by $f(x) = cx^3$ for $0 < x < 4$ and $f(x) = 0$ otherwise. Calculate the variance of X.

9. Let X be a r.v. with finite expectation and finite variance. Prove:

a) For any constant a, $Var(aX) = a^2 Var(X)$.

b) For any constant b, $Var(X + b) = Var(X)$.

10. a) Suppose X and Y are two independent r.v.s such that $EX^4 = 2$, $EY^2 = 1$, $EX^2 = 1$ and $EY = 0$. Compute the variance of X^2Y.

    b) Let S and T be two independent r.v.s with $ES = 5$, $ET = -3$, $Var(S) = 8$ and $Var(T) = 7$. Let $W = 2S + 3T - 4$; compute the mean and variance of W.

Exercises from Section 5.4

11. a) Prove that the expected value of an $Exp(\lambda)$ r.v. is $\frac{1}{\lambda}$.

    b) Prove that the expected value of an $\Gamma(r, \lambda)$ r.v. is $\frac{r}{\lambda}$.

    c) Verify that the expected value of a $Hyp(n, r, k)$ r.v. is $\frac{kr}{n}$.
    Hint: You will have to do an index change in your summation, and then apply Vandermonde's identity.

12. a) Let $X \sim Exp(\lambda)$. Compute $E(X^2)$ directly (using the change of variables formula together with the Gamma integral formula) and use your answer to verify that the variance of X is $\frac{1}{\lambda^2}$.

    b) Prove that the variance of the uniform distribution on the interval (a, b) is $\frac{(b-a)^2}{12}$.

    c) Prove that the variance of a $\Gamma(r, \lambda)$ r.v. is $\frac{r}{\lambda^2}$.

13. A pond contains equal numbers of four different types of fish. You go fishing, and each time you cast, you catch one of the four types of fish (each type is equally likely). What is the expected number of casts it will take you to have caught at least one of all four types of fish?

14. Choose two of (a),(b),(c):

    a) (AE) An actuary has discovered that policyholders are six times as likely to file three claims as they are to file four claims. If the number of claims filed has a Poisson distribution, what is the variance of the number of claims filed?

    b) (AE) A company has two electric generators. The time until failure for each generator is exponential with mean 13. The company will begin using the second generator immediately after the first one fails. What is the variance of the total time the generators produce electricity?

c) (AE) The profit for a new product is given by $Z = 5X - 4Y + 8$, where $X \perp Y$, $Var(X) = 3$ and $Var(Y) = 2$. What is the variance of the profit for the new product?

15. (AE) Let X represent the number of customers arriving during the morning hours, and let Y be the number of customers arriving during the evening hours to a restaurant. Assuming that X and Y are both Poisson, and that the first moment of X is 8 less than the first moment of Y, and that the second moment of X is 60% of the second moment of Y, what is the variance of Y?

16. Suppose that the departure of a tour is delayed by an amount of time that is modeled by an exponential r.v. with variance 9 hours. If the departure of the tour is delayed by less than 2 hours, the tour company pays no refund, but if the tour is delayed 2 to 4 hours, then the tour company pays a refund of $20t$, where t is the number of hours the tour is delayed. If the tour is delayed by more than 4 hours, the tour company pays a flat refund of 80. Compute the variance of the refund paid by the tour company.

Exercises from Section 5.5

17. Compute the covariance of X and Y, if they have joint density

$$f_{X,Y}(x,y) = \begin{cases} 2 & \text{if } x > 0, y > 0, \text{ and } x + y < 1 \\ 0 & \text{else} \end{cases}.$$

18. Suppose a box contains three balls numbered 1 to 3. Two balls are selected without replacement from the box. Let U be the number on the first ball selected, and let V be the number on the second ball selected. Compute $Cov(U,V)$ and $\rho(U,V)$.

    Hint: Start by making a chart which describes the joint density of U and V.

19. (AE) Let X and Y denote the price of two stocks at the end of a five-year period. Suppose X is uniform on $[0,6]$ and that given $X = x$, Y is uniform on $[0,x]$. Determine $Cov(X,Y)$.

20. Let (X,Y) be a point chosen uniformly from the finite set of four points

$$\{(0,1),(1,0),(0,-1),(-1,0)\}.$$

    Prove that X and Y are uncorrelated, but not independent.

21. (AE) Let X denote the size of a surgical claim, and let Y denote the size of the associated hospital claim. An actuary is using a model in which $EX = 6$, $EX^2 = 47.4$, $EY = 3$, $EY^2 = 21.4$ and $Var(X+Y) = 13.5$. Let $C_1 = X + Y$ be

the size of the combined claims before the application of a 20% surcharge on the hospital portion of the claim, and let C_2 denote the size of the combined claims after the surcharge. Calculate $Cov(C_1, C_2)$.

22. Prove any two of the following three statements:

    a) $Cov(X, Y) = Cov(Y, X)$
    b) $Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$
    c) $Cov(aX, Y) = a\,Cov(X, Y)$

    Note: These three statements generalize to the following important property of covariance called *bilinearity*:

$$Cov\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j Cov(X_i, Y_j)$$

23. Prove that **correlation is unchanged under linear transformations**, meaning that $\rho(aX + c, bY + d) = \rho(X, Y)$ for any constants a, b, c, d with $a > 0$ and $b > 0$.

24. a) Prove that if $Y = aX + b$ for constants a and b (with $a \neq 0$), then $\rho(X, Y) = \pm 1$.
    b) In this setting, under what conditions is $\rho(X, Y) = 1$ (as opposed to -1)?

In the next two exercises, we will prove that **a correlation of ± 1 implies a linear relationship between the r.v.s**, i.e. that if $\rho(X, Y) = \pm 1$, then $Y = aX + b$ where a and b are constants.

25. Define

$$\widehat{X} = \frac{1}{\sqrt{Var(X)}}(X - EX) \text{ and } \widehat{Y} = \frac{1}{\sqrt{Var(Y)}}(Y - EY).$$

    a) Compute $E\left[\widehat{X}\right]$ and $E\left[\widehat{Y}\right]$.
    b) Compute $E\left[\widehat{X}^2\right]$ and $E\left[\widehat{Y}^2\right]$.
    c) Prove that $\rho(X, Y) = Cov\left(\widehat{X}, \widehat{Y}\right)$.
    d) Prove that $Cov\left(\widehat{X}, \widehat{Y}\right) = E\left[\widehat{X}\widehat{Y}\right]$.

26. a) Use the results of Exercise 25 to prove that

$$E\left[\left(\widehat{Y} - \rho(X, Y)\widehat{X}\right)^2\right] = 1 - \rho(X, Y)^2. \qquad (\bigstar)$$

b) Use part (a) to prove that if $\rho(X,Y) = \pm 1$, then $\widehat{Y} = \rho(X,Y)\widehat{X}$ where a and b are constants.

Hint: If $\rho(X,Y) = \pm 1$, what must the right-hand side of (★) be? What does that imply about $\left(\widehat{Y} - \rho(X,Y)\widehat{X}\right)^2$?

c) Use part (b) to deduce that $Y = aX + b$ for suitable constants a and b.

Hint: Start with what you proved in (b), and back-substitute for X and Y. Rearrange what you get to show $Y =$(constant)$X+$constant, as wanted.

Exercises from Section 5.6

27. Let X and Y be r.v.s having joint density function given by the following table:

$Y \backslash X$	-1	0	2	6
-2	$\frac{1}{27}$	$\frac{1}{9}$	$\frac{1}{27}$	$\frac{1}{9}$
1	$\frac{1}{9}$	0	$\frac{1}{9}$	$\frac{2}{9}$
3	0	$\frac{2}{27}$	$\frac{1}{9}$	$\frac{2}{27}$

a) Calculate $E(Y \mid X)$.

b) Calculate $E(X^3 \mid Y = 1)$.

c) Calculate $Var(X \mid Y = 3)$.

28. Let (X,Y) be chosen uniformly from the triangle whose vertices are $(0,0)$, $(2,0)$ and $(1,2)$. Compute the conditional expectation of Y given X.

29. (AE) A fair die is rolled repeatedly. Let X be the number of rolls needed to obtain a 5 and let Y be the number of rolls needed to obtain a 6. Calculate $E[X \mid Y = 2]$.

30. Let X and Y be independent, where X is $\Gamma(r,\lambda)$ and Y is $\Gamma(s,\lambda)$. Compute $E[X \mid X + Y]$.

Hint: First calculate the joint density of X and $X + Y$.

31. Suppose $E[Y|X] = 2x + 1$ and $E[Z|X] = 3x$.

a) Compute $E[3Y - 4Z + 7|X]$.

b) Compute $E[2X^2Y|X = 2]$.

32. (AE) Let N_1 and N_2 represent the numbers of claims submitted to a life insurance company in January and February, respectively. The joint density function of N_1 and N_2 is

$$f_{N_1,N_2}(n_1,n_2) = \begin{cases} \frac{2}{3}\left(\frac{1}{3}\right)^{n_1} e^{-n_1-1}\left(1-e^{-n_1-1}\right)^{n_2} & \text{for } n_1, n_2 \in \{0,1,2,3,...\} \\ 0 & \text{else} \end{cases}$$

Calculate the expected number of claims that will be submitted to the company in February, if exactly 2 claims were submitted in January.

33. (AE) A driver and a passenger are in a car accident. Each of them independently has a probability $.3$ of being hospitalized. If they are hospitalized, the loss is uniform on $[0,1]$. When two hospitalizations occur, the losses are independent. Calculate the expected number of people who are hospitalized, given that the total loss due to hospitalizations from the accident is less than 1.

34. The time it takes an insurance company to process a claim of size S is uniform on $[S, S+1]$. If S is itself exponentially distributed with parameter $\frac{1}{2}$, what is the expected time to process a claim?

35. a) Prove that the two formulas given in the notes as definitions of conditional variance are the same.
    b) Prove the **Law of Total Variance**, which says:

$$E[Var(X|Y)] + Var[E(X|Y)] = Var(X).$$

36. a) (AE) The number of workplace injuries, N, occuring in a factory on any given day is Poisson with mean λ. The parameter λ is itself a r.v. depending on the level of activity in the factory, and is assumed to be uniformly distributed on the interval $[0,6]$. Compute $Var(N)$.
    b) (AE) The stock prices of two companies at the end of any given year are modeled with r.v.s X and Y whose joint density function is

$$f(x,y) = \begin{cases} 2x & \text{for } 0 < x < 1, x < y < x+1 \\ 0 & \text{otherwise} \end{cases}.$$

    What is the conditional variance of Y given $X = x$?

Exercises from Section 5.7

37. Suppose $X \sim Unif(\{1,2,3,...,n\})$. Compute the probability generating function of X.

38. Let $X \sim binomial(n, p)$. Use the probability generating function of X to compute the expected value and variance of X.

39. Let X be a discrete r.v. taking values in $\{0, 1, 2, 3, ...\}$ with pgf $G_X(t) = \exp(2t + 3t^{-1} - 5)$. Compute the variance of X.

40. Suppose X is geometric with mean 2. Compute $E[X(X-1)(X-2)(X-3)]$.

41. Let X be a discrete r.v. taking values in $\{0, 1, 2, 3, ...\}$ with pgf $G_X(t) = e^{t-1}$.

    a) What is $P(X = 0)$?

    b) What is $P(X = 4)$?

    Hint: Since $G_X(t) = \sum\limits_{x=0}^{\infty} f_X(x)t^x$, $f_X(4) = P(X = 4)$ is the coefficient on t^4 in the Taylor series expansion of $G_X(t)$. So start by writing the Taylor series of $G_X(t)$.

42. Prove that if $X_1, ..., X_d$ are independent geometric r.v.s, each with parameter p, then their sum $S = X_1 + ... + X_d$ is negative binomial with parameters d and p.

43. (AE) The number N of babies born in a hospital during any one week is a r.v. satisfying $P(N = n) = \frac{1}{2^{n+1}}$, for $n \in \{0, 1, 2, ...\}$. Suppose that the number of babies born in any one week is independent of the number of babies born in any other week. Determine the probability that exactly seventeen babies are born in a given four-week period.

Exercises from Section 5.8

44. a) Prove that the moment generating function of an $Exp(\lambda)$ r.v. is $\dfrac{\lambda}{\lambda - t}$.

    b) Prove that the moment generating function of a $\Gamma(r, \lambda)$ r.v. is $\left(\dfrac{\lambda}{\lambda - t}\right)^r$.

45. a) Compute the first and second moments of X, if its moment generating function is $M_X(t) = \dfrac{1}{\sqrt{1-4t}}$ for $t < \dfrac{1}{4}$.

    b) Suppose X and Y are exponential r.v.s with respective means 3 and 7. If $X \perp Y$, what is the moment generating function of $4X + Y$?

    c) (AE) Assume that the number of claims related to traffic accidents on a certain road is a r.v. X whose moment generating function is $M_X(t) = (1 - 2500t)^{-4}$. Find the standard deviation of the claim size for this class of accidents.

46. (AE) Let X, Y and Z be i.i.d. r.v.s, each taking the value 0 with probability p and the value 1 with probability $(1-p)$. Compute the moment generating function of $W = XYZ$.

47. Let X be a continuous r.v. having the density $f_X(x) = \frac{1}{2}e^{-|x|}$ for all x. Compute the moment generating function of X.

48. Explain why each of the following functions *cannot* be the moment generating function of a real-valued r.v. X:

    a) $h(t) = \dfrac{e^{-t}}{2-t}$ for $t < 2$;

    b) $j(t) = \dfrac{1+t}{1-t}$ for $t < 1$;

    c) $k(t) = \exp\left(\dfrac{-t^2}{2}\right)$ for $-\infty < t < \infty$.

49. Prove the independence property of MGFs, which says that if $X \perp Y$ then $M_{X+Y}(t) = M_X(t)M_Y(t)$.

50. In this problem we will derive the **Beta integral formula** (which can be useful to solve certain expected value and conditional expectation problems):

$$\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

    a) Let X be $\Gamma(\alpha, \lambda)$ and let Y be $\Gamma(\beta, \lambda)$. Suppose that $X \perp Y$. Determine the density function of $Z = X + Y$ using moment generating functions.

    b) Given X and Y as above, compute the joint density function of X and $Z = X + Y$ by the transformation method of Chapter 4.

    c) Use your answer to part (b) to compute the marginal density of Z (write your answer as an integral with respect to x).

    d) Derive the Beta integral formula by equating the answers to part (a) and (c) of this problem, and solving the resulting equation for the Beta integral above.

    Hint: in the integral you obtain from part (c), use the $u-$substitution $u = \dfrac{x}{z}$.

51. A **Beta random variable** with parameters $r_1 > 0$ and $r_2 > 0$ (denoted $B(r_1, r_2)$) is a continuous r.v. whose density is

$$f(x) = \begin{cases} \dfrac{\Gamma(r_1+r_2)}{\Gamma(r_1)\Gamma(r_2)} x^{r_1-1}(1-x)^{r_2-1} & \text{if } 0 < x < 1 \\ 0 & \text{else} \end{cases}.$$

a) Prove that the function above is in fact a density function.

b) Determine the expected value of a Beta $B(r_1, r_2)$ r.v.

Exercises from Section 5.9

52. Suppose Y is a discrete r.v. taking the values $0, 1, 4$ and 10 with respective probabilities $\frac{3}{8}, \frac{1}{8}, \frac{1}{3}$ and $\frac{1}{6}$. Compute $M_Y(t)$.

53. Suppose X is a r.v. with $EX = \frac{1}{2}$ whose moment generating function is

$$M_X(t) = \frac{1}{7} + \frac{2}{7}e^t + Ce^{-t} + De^{2t},$$

where C and D are constants.

a) Find C and D.

b) Find a density function of X.
Hint: Look at the moment generating function you computed in Exercise 52, and use that to make an educated guess as to the density of X. (Uniqueness of MGFs can be used to show that your guess is correct.)

c) Find $P(X \geq 0)$.

d) Find the variance of X.

54. (AE) Let X and Y be i.i.d. r.v.s such that the moment generating function of $X + Y$ is

$$M_{X+Y}(t) = .09e^{-2t} + .24e^{-t} + .34 + .24e^t + .09e^{2t}$$

for all t. Calculate $P(X \leq 0)$.

55. Evaluate each integral:

a) $\displaystyle\int_{-\infty}^{\infty} \exp\left[\frac{-(x+3)^2}{18}\right] dx$ b) $\displaystyle\int_{-\infty}^{\infty} \sqrt{\pi}e^{-t^2+12t}\, dt$

56. Prove that if $X_1, ..., X_d$ are independent r.v.s with $X_j \sim \Gamma(r_j, \lambda)$, then $S = X_1 + ... + X_j \sim \Gamma(r_1 + ... + r_d, \lambda)$.

Exercises from Section 5.10

57. (AE) Suppose X and Y are independent r.v.s which have the same moment generating function: $M_X(t) = M_Y(t) = e^{t^2}$. Determine the joint moment generating function of $W = X + Y$ and $Z = Y - X$.

58. Suppose X and Y are real-valued r.v.s whose joint moment generating function is
$$M_{X,Y}(s,t) = \frac{64}{(s-4)^2(s+t-2)^2}.$$
   a) Compute EY.
   b) Compute $Cov(X,Y)$.
   c) Compute the moment generating function of $Y - X$.
   d) Based on your answer to part (c), what common r.v. is $Y - X$?

Exercises from Section 5.11

59. Suppose X is a gamma r.v. with mean 2 and variance 4. Use the Markov inequality to find the largest possible value of $P(X \geq 6)$.

60. Use the Markov inequality to prove that for every $t \geq 0$, $e^{-t} \leq \frac{1}{t}$.

    Hint: Consider an exponential r.v. X with a particular value of λ, and use the Markov inequality with a particular value of a. At least one of the λ and/or a should have a t in it.

61. Let X be a discrete r.v. whose density is

x	1	2	3
$f_X(x)$	$\frac{1}{18}$	$\frac{16}{18}$	$\frac{1}{18}$

    Show that when $\delta = 1$, $P(|X-\mu| \geq \delta) = \dfrac{Var(X)}{\delta^2}$ (the point of this problem is to show that in general, the $\leq$ sign in Chebyshev's inequality cannot be replaced by a $<$).

62. A bolt manufacturer knows that 5% of his production is defective. He gives a guarantee on his shipment of 10000 parts by promising that no more than a bolts are defective. Use Chebyshev's inequality to find the smallest number a can be, so that the manufacturer is assured of not paying a refund more than 1% of the time.

63. Let X be Poisson with mean λ.
   a) Use Chebyshev's inequality to verify that $P\left(X \leq \dfrac{\lambda}{2}\right) \leq \dfrac{4}{\lambda}$.
   b) Use Chebyshev's inequality to verify that $P(X \geq 2\lambda) \leq \dfrac{1}{\lambda}$.

64. Suppose X is a r.v. with mean and variance both equal to 20. From Chebyshev's inequality, what can be said about $P(0 < X < 40)$? (In particular, what is the maximum or minimum value of this expression?)

65. Suppose X and Y are two real-valued r.v.s with

$$EX = 75,\ EY = 75,\ Var(X) = 10,\ Var(Y) = 12,\ Cov(X, Y) = -3.$$

Based on Chebyshev's inequality, what can be said about $P(|X - Y| \geq 15)$?

Chapter 6

I.i.d. processes and normal random variables

6.1 I.i.d. processes

We are interested in studying the results of an experiment which is repeated over and over. Examples include:

- attributes of items that are produced by a manufacturing process;
- measurement errors in experiments;
- claim sizes filed by a series of insurance policyholders;
- heights, weights, lifespans, etc. taken from a sample of organisms;
- daily medical readings of a patient (blood pressure, heart rate, blood sugar, etc.); etc.

Definition 6.1 *A discrete-time stochastic process $\{X_t : t \in \mathbb{N}\}$ is called an* **i.i.d. process** *if the process is "independent and identically distributed", i.e.*

- $X_j \perp X_k$ *for all* $j \neq k$*, and*
- *the* X_j *have the same distribution for all* j*.*

In this setting, we denote the mean of each X_j by μ and the variance of each X_j by σ^2.

The prototype example of an i.i.d. process is coin flipping: if you flip the same coin over and over again and let

$$X_j = \begin{cases} 1 & \text{if } j^{th} \text{ flip is heads} \\ 0 & \text{if } j^{th} \text{ flip is tails} \end{cases}$$

What we are most interested in is either the *sum* or the *average* behavior of such a process.

Definition 6.2 *Given an i.i.d. process* $\{X_t\}$*, define the following processes:*

1. $\{S_n\}_{n\in\mathbb{N}}$*, the* **sequence of sums***, is* $S_n = X_1 + X_2 + ... + X_n$*;*

2. $\{A_n\}_{n\in\mathbb{N}}$ *the* **sequence of averages***, is*

$$A_n = \frac{1}{n}S_n = \frac{X_1 + ... + X_n}{n};$$

3. $\{A_n^*\}_{n\in\mathbb{N}}$*, the* **sequence of normalized averages***, is*

$$A_n^* = \frac{A_n - E[A_n]}{\sqrt{Var(A_n)}} = \frac{A_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{S_n - \mu n}{\sigma\sqrt{n}}.$$

Notice that if each X_t has mean μ and variance σ^2, then

$$E[A_n] =$$

$$Var(A_n) =$$

$$E[A_n^*] =$$

$$Var(A_n^*) =$$

EXAMPLE 1

Suppose $\{X_t\}$ is an i.i.d. process where each $X_t \sim Unif([0,1])$. Compute the values of the r.v.s S_n, A_n and A_n^* for $n \in \{1,2,3\}$, if the values of the X_t are

$$X_1 = \frac{2}{3} \qquad X_2 = \frac{1}{4} \qquad X_3 = 0 \qquad X_4 = \frac{2}{5}$$

EXAMPLE 2

Suppose we have an i.i.d. process $\{X_t\}$ where each X_t is uniform on $\{1, 2, 3, 4, 5, 6\}$. (This process models repeated rolling of a fair die.) If the sequence of die rolls is

$$3,\ 5,\ 1,\ 3,\ 2,\ 6,\ 4,\ 1,\ 1,\ \ldots$$

compute the first six values of the corresponding sequence of averages.

6.2 Laws of Large Numbers

In this section, we investigate some results which give us information about the averages coming from an i.i.d. process. Recall that if $\{X_t\}$ is i.i.d., we denote each EX_j by μ and each $Var(X_j)$ by σ^2.

Quantitative Weak Law of Large Numbers

Theorem 6.3 (Quantitative Weak Law of Large Numbers (QWLLN)) *Let $\{X_t\}$ be an i.i.d. process, and for each $n \in \mathbb{N}$, set $A_n = \frac{1}{n}(X_1 + \ldots + X_n)$. Then for all $\delta > 0$,*

$$P(|A_n - \mu| \geq \delta) \leq \frac{\sigma^2}{n\delta^2}.$$

PROOF From the previous section, $E[A_n] = \mu$ and $Var(A_n) = \frac{\sigma^2}{n}$. Apply Chebyshev's inequality to A_n to get the QWLLN. $\square$

Idea: The QWLLN says that if you fix an "error tolerance" δ, if you take enough measurements (say n measurements), then the probability that the average of your measurements A_n is within δ of the theoretical average μ is high.

EXAMPLE 3

Marbles are drawn from a jar containing 3 red and 5 marbles, one at a time with replacement. What is the smallest number n such that you can be 99% assured that between 37% and 38% of the first n marbles drawn are red?

Weak Law of Large Numbers

Theorem 6.4 (Weak Law of Large Numbers (WLLN)) *Let $\{X_t\}$ be an i.i.d. process, where each X_j is a r.v. with finite expected value μ and finite variance σ^2. For each $n \in \mathbb{N}$, set $A_n = \dfrac{X_1 + ... + X_n}{n}$. Then for all $\delta > 0$,*

$$\lim_{n\to\infty} P(|A_n - \mu| \geq \delta) = 0.$$

PROOF Take the limit of each side of the inequality in the QWLLN as $n \to \infty$. $\square$

REMARK

One can derive the WLLN without the assumption that the X_j have finite variance.

Interpreting the WLLN

Loosely speaking, the WLLN says that if you take a large sample, the probability that the average of your sample is within δ of the "theoretical average" of each measurement (i.e. the expected value μ) is large, and that as the size of the sample increases, this probability goes to 1. Let's think about what this means in terms of flipping a fair coin repeatedly:

Let $\{X_t\}$ be an i.i.d sequence of r.v.s, each uniform on $\{0,1\}$ (think of $X_j = 1$ as corresponding to the j^{th} flip being heads and $X_j = 0$ meaning the j^{th} flip being tails). In this setting, $\mu = EX_j = \frac{1}{2}$.

Under these assumptions, what is A_n?

Let $\delta = \dfrac{1}{10}$. Let's say that a sequence of flips is "n-good" (or "n, δ-good") if $|A_n - \mu| < \delta$, i.e. the proportion of heads in the first n flips is between $\dfrac{4}{10}$ and $\dfrac{6}{10}$.

Example: H,H,H,H,T,T,T,T,... is not $4, \dfrac{1}{10}$-good, but is $8, \dfrac{1}{10}$-good.

The WLLN says: if you **fix δ**, and then choose a **large enough n**, **most sequences are n, δ-good**.

HOWEVER: what the WLLN doesn't tell you (and why it is called the "Weak" LLN) is any relationship between sequences that are good at different values of n.
For example, the WLLN does not guarantee that most sequences are "eventually good", i.e. are n-good for all sufficiently large n.
In particular, it might be the case that typical sequences of heads and tails are n-bad for infinitely many, very sparsely spaced n).

This weakness is fixed with the following stronger result, which says (among other things) that with probability 1, a randomly chosen sequence of heads and tails from a fair coin is eventually good (i.e. the proportion of heads in the sequence becomes close to $\frac{1}{2}$ *and stays close to $\frac{1}{2}$ forever*:

Strong Law of Large Numbers

Theorem 6.5 (Strong Law of Large Numbers (SLLN)) *Let $\{X_t\}$ be an i.i.d. process, where each X_j is a r.v. with finite expected value μ. Then*

$$P\left(\lim_{n\to\infty} A_n = \mu\right) = 1$$

PROOF (with the extra assumption that $EX^4 < \infty$) Suppose first that $EX_j = 0$. That means $Var(X_j) = EX_j^2 - 0^2 = EX_j^2$. Now,

$$\begin{aligned} E[S_n^4] &= E[(X_1 + ... + X_n)^4] \\ &= E[(X_1 + ... + X_n)(X_1 + ... + X_n)(X_1 + ... + X_n)(X_1 + ... + X_n)] \end{aligned}$$

$$\begin{aligned} \Rightarrow E[S_n^4] &= nEX_j^4 + \binom{n}{2}\binom{4}{2}[Var(X_j)]^2 \\ &= nEX_j^4 + 3n(n-1)[Var(X_j)]^2 \end{aligned}$$

$$\leq nEX_j^4 + 3n(n-1)EX_j^4.$$

Therefore

$$E\left[\frac{S_n^4}{n^4}\right] \leq \frac{nEX_j^4 + 3n(n-1)EX_j^4}{n^4} \leq \frac{1}{n^3}EX_j^4 + \frac{3}{n^2}EX_j^4$$

so

$$\lim_{n\to\infty} E[A_n^4] = \lim_{n\to\infty} E\left[\left(\frac{S_n}{n}\right)^4\right] = \lim_{n\to\infty} E\left[\frac{S_n^4}{n^4}\right] = 0.$$

By definiteness, $\lim_{n\to\infty} A_n^4 = 0$ with probability 1, so $P(\lim_{n\to\infty} A_n = 0) = 1$ as wanted.

If $EX_j = \mu \neq 0$, then apply the above to $X_j - \mu$ to see that $\lim_{n\to\infty} (A_n - \mu) = 0$ with probability 1, i.e. $P(\lim_{n\to\infty} A_n = \mu) = 1$ as wanted. □

6.3 Limits of normalized averages

QUESTION

Let $\{X_t\}$ be an i.i.d. process with normalized averages A_n^*. What happens to the distribution of A_n^* as $n \to \infty$?

Reminder: $A_n = \dfrac{X_1 + \ldots + X_n}{n} = \dfrac{S_n}{n}$ and $A_n^* = \dfrac{A_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \dfrac{S_n - n\mu}{\sigma\sqrt{n}}$.

To study this question, we use MGFs. Suppose that each X_j has MGF $M_X(t)$. Then

$$M_{S_n}(t) = M_{X_1+\ldots+X_n}(t) = \prod_{j=1}^{n} M_{X_j}(t) = [M_X(t)]^n$$

and therefore

$$\begin{aligned}
M_{A_n^*}(t) = E[e^{tA_n^*}] &= E\left[\exp\left(t\,\frac{S_n - n\mu}{\sigma\sqrt{n}}\right)\right] \\
&= E\left[\exp\left(\frac{t}{\sigma\sqrt{n}}S_n - \frac{n\mu t}{\sigma\sqrt{n}}\right)\right] \\
&= E\left[\exp\left(\frac{t}{\sigma\sqrt{n}}S_n\right)\right] \cdot \exp\left(\frac{-\mu t n}{\sigma\sqrt{n}}\right) \\
&= M_{S_n}\left(\frac{t}{\sigma\sqrt{n}}\right) \cdot \exp\left(\frac{-\mu t n}{\sigma\sqrt{n}}\right) \\
&= \left[M_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n \cdot \exp\left(\frac{-\mu t n}{\sigma\sqrt{n}}\right) \\
&= \exp\left[n\left(\ln M_X\left(\frac{t}{\sigma\sqrt{n}}\right) - \frac{\mu t}{\sigma\sqrt{n}}\right)\right] \\
&= \exp\left[\triangle\right].
\end{aligned}$$

Thus

$$\lim_{n\to\infty} M_{A_n^*}(t) = \lim_{n\to\infty} \exp[\triangle] = \exp\left[\lim_{n\to\infty} \triangle\right].$$

We are going to work out $\lim\limits_{n\to\infty} \triangle$. First, a special situation: if $t = 0$,

$$\lim_{n\to\infty} \triangle = \lim_{n\to\infty} n(\ln M_X(0) - 0) = \lim_{n\to\infty} n(\ln 1 - 0) = n \cdot 0 = 0 = \frac{0^2}{2}$$

The more general situation is when $t \neq 0$:

$$\begin{aligned}
\lim_{n\to\infty} \triangle &= \lim_{n\to\infty} n\left(\ln M_X\left(\frac{t}{\sigma\sqrt{n}}\right) - \frac{\mu t}{\sigma\sqrt{n}}\right) \\
&= \lim_{n\to\infty} \frac{\frac{t^2}{\sigma^2}\cdot\left(\ln M_X\left(\frac{t}{\sigma\sqrt{n}}\right) - \frac{\mu t}{\sigma\sqrt{n}}\right)}{\frac{t^2}{\sigma^2}\cdot\frac{1}{n}} \\
&= \frac{t^2}{\sigma^2}\lim_{n\to\infty} \frac{\ln M_X\left(\frac{t}{\sigma\sqrt{n}}\right) - \mu\left(\frac{t}{\sigma\sqrt{n}}\right)}{\left(\frac{t}{\sigma\sqrt{n}}\right)^2}.
\end{aligned}$$

Now, let $s = \dfrac{t}{\sigma\sqrt{n}}$ so that as $n \to \infty$, $s \to 0$. This makes

$$\begin{aligned}
\lim_{n\to\infty} \triangle &= \frac{t^2}{\sigma^2}\lim_{s\to 0}\frac{\ln M_X(s) - \mu s}{s^2} = \frac{\ln 1 - 0}{0} = \frac{0}{0} \\
&\stackrel{L}{=} \frac{t^2}{\sigma^2}\lim_{s\to 0}\frac{\frac{M_X'(s)}{M_X(s)} - \mu}{2s} = \frac{\frac{\mu}{1} - \mu}{0} = \frac{0}{0} \\
&\stackrel{L}{=} \frac{t^2}{\sigma^2}\lim_{s\to 0}\frac{\frac{M_X''(s)M_X(s) - (M_X'(s))^2}{[M_X(s)]^2}}{2} \\
&= \frac{t^2}{\sigma^2}\cdot\frac{\frac{[EX^2\cdot 1 - (EX)^2]}{1^2}}{2} \\
&= \frac{t^2}{\sigma^2}\cdot\frac{Var(X)}{2} = \frac{1}{2}t^2.
\end{aligned}$$

We have proven that for any t,

$$\boxed{\lim_{n\to\infty} M_{A_n^*}(t) = \exp\left(\frac{t^2}{2}\right).}$$

THIS IS AMAZING! We have proven, that **no matter what the original density of X was**, the MGF of A_n^* must approach this "magic" MGF $e^{t^2/2}$.

That means, by uniqueness of MGFs, that A_n^* must approach some "magic" r.v. whose MGF is $e^{t^2/2}$. **But what r.v. X has $M_X(t) = e^{t^2/2}$?**

To get from $M_X(t) = e^{t^2/2}$ back to the density $f_X(x)$, we use the ________________

________________ :

$$
\begin{aligned}
f_X(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixt} M_X(it)\,dt \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixt} e^{(it)^2/2}\,dt \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixt} e^{-t^2/2}\,dt \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(t^2 + 2ixt\right)\right]\,dt \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left[-\frac{-(t^2 - 2ixt - x^2 + x^2)}{2}\right]\,dt \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left[\frac{-(t-ix)^2}{2}\right] e^{-x^2/2}\,dt \\
&= \frac{1}{2\pi}e^{-x^2/2}\int_{-\infty}^{\infty} \exp\left[\frac{-(t-ix)^2}{2}\right]\,dt \\
&= \frac{1}{2\pi}e^{-x^2/2}\sqrt{2\pi} \\
&\qquad \text{(Gaussian Integral Formula with } \mu = ix,\ \sigma = 1\text{)} \\
&= \boxed{\frac{1}{\sqrt{2\pi}}e^{-x^2/2}}.
\end{aligned}
$$

Big picture: These computations lead us to our last class of common random variables, called *normal r.v.s*.

Based on what we have just done, these normal r.v.s will *approximate the normalized average of any i.i.d. sequence of r.v.s*.

We will see that this class of normal r.v.s also approximate the *averages and sums of any i.i.d. sequence of r.v.s*.

6.4 Normal random variables

The standard normal random variable

Definition 6.6 *The* **standard normal** *r.v., abbreviated* $n(0,1)$ *or* $\mathcal{N}(0,1)$ *or* Z*, is the continuous r.v. whose density function is*

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right).$$

The cumulative distribution function of the standard normal r.v. is denoted Φ*:*

$$\Phi(x) = P(n(0,1) \leq x) = \int_{-\infty}^{x} \phi(t)\, dt.$$

There is no better formula for Φ; values of Φ are estimated using a calculator or tables (one such table can be found in Appendix A.3 of these notes). On exams in MATH 414 and 416, we often will leave answers to questions in terms of Φ.

The standard normal r.v. **approximates the normalized average A_n^* of an i.i.d. sequence $\{X_t\}$ of r.v.s, no matter what distribution the individual X_t have**.

Theorem 6.7 (Properties of the standard normal density) *Let* ϕ *be the density of the standard normal r.v.. Then*

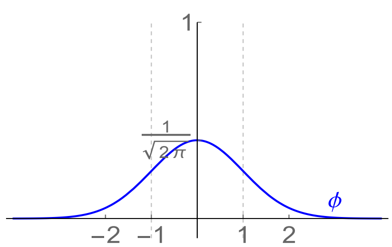
1. $\phi(-x) = \phi(x)$.
2. $\phi(0) = \dfrac{1}{\sqrt{2\pi}}$.
3. $\displaystyle\int_{-\infty}^{\infty} \phi(x)\, dx = 1$.
4. ϕ *is increasing on* $(-\infty, 0)$ *and decreasing on* $(0, \infty)$.
5. ϕ *is concave down on* $(-1, 1)$ *and concave up on* $(-\infty, -1) \cup (1, \infty)$.
6. $\displaystyle\lim_{x\to\infty} \phi(x) = \lim_{x\to-\infty} \phi(x) = 0$.

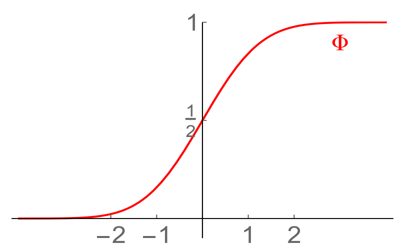
PROOF (1) and (2) are obvious; (3) follows from the Gaussian Integral Formula with $\mu = 0$, $\sigma = 1$; (4) comes from differentiating ϕ and analyzing the sign of ϕ'; (5) comes from analyzing the sign of ϕ''; (6) is a basic Calculus 1 limit. □

Theorem 6.8 (Properties of Φ) *Let Φ be the cdf of the standard normal. Then:*

1. *Φ is a cdf (so it has properties common to all cdfs);*
2. *Φ is continuous;*
3. *Φ is differentiable and $\Phi' = \phi$;*
4. *$\Phi(0) = \frac{1}{2}$;*
5. *For all x, $\Phi(-x) = 1 - \Phi(x)$.*
6. *For all x, $P(|n(0,1)| \leq x) = \Phi(x) - \Phi(-x) = 2\Phi(x) - 1$.*

These properties imply that the graph of ϕ looks like a "bell curve" (shown below at left), and that the graph of Φ looks like the picture at right:





Theorem 6.9 (Mean, variance and MGF of std. normal) *Let $X \sim n(0,1)$. Then:*

$$EX = 0 \qquad Var(X) = 1 \qquad M_X(t) = e^{t^2/2}.$$

PROOF We already know $M_X(t) = e^{t^2/2}$, because that was the MGF we plugged into the inversion formula to come up with the density of the standard normal r.v. in the first place. Therefore:

$$EX = M_X'(0) = \frac{d}{dt}\left[e^{t^2/2}\right]_{t=0} =$$

$$EX^2 = M_X''(0) = \frac{d}{dt}\left[te^{t^2/2}\right]_{t=0} = \left[e^{t^2/2} + t^2 e^{t^2/2}\right]_{t=0} = 1$$

$$Var(X) = EX^2 - (EX)^2 = 1 - 0^2 = 1. \square$$

Linear transformations of the standard normal

MOTIVATION

Suppose $\{X_t\}$ is an i.i.d. process with mean μ and variance σ^2. So far, we've learned that the normalized averages A_n^* must be approximately standard normal when n is large, i.e.

$$\boxed{A_n^* \approx n(0,1) \text{ for } n \text{ large.}}$$

As before, the sequences of sums $S_n = \sum_{j=1}^{n} X_j$ and averages $A_n = \frac{1}{n}\sum_{j=1}^{n} X_j$ are related to the normalized averages A_n^* by

$$A_n^* = \frac{A_n - \mu}{\frac{\sigma}{\sqrt{n}}} \Rightarrow \tag{6.1}$$

$$A_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}} \Rightarrow \tag{6.2}$$

We see that both A_n and S_n are a linear transformation of A_n^*, i.e.

$$A_n = (\text{constant})A_n^* + \text{constant} \qquad\qquad S_n = (\text{constant})A_n^* + \text{constant}$$

That means

$$A_n \approx (\text{constant}) \qquad + \text{constant} \tag{6.3}$$

$$S_n \approx (\text{constant}) \qquad + \text{constant} \tag{6.4}$$

so A_n **and** S_n **are both approximately linear transformations of the standard normal r.v.** This motivates the following definition:

Definition 6.10 *Let $\mu \in \mathbb{R}$ and $\sigma \in (0,\infty)$. A random variable X is called* **normal** *with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ if*

$$X = \mu + \sigma Z$$

where $Z \sim n(0,1)$. In this case, we write $X \sim n(\mu,\sigma^2)$ or $X \sim \mathcal{N}(\mu,\sigma^2)$.

In this context, μ is called the **mean parameter** *of X and σ^2 is called the* **variance parameter** *of X.*

Normal r.v.s **approximate averages and sums of a bunch of i.i.d. r.v.s, no matter what distribution the individual r.v.s have**.

More precisely, the content of equations (6.1)-(6.4) can be rephrased into what is commonly called the Central Limit Theorem:

Theorem 6.11 (Central Limit Theorem (CLT)) *Let $\{X_n\}$ be an i.i.d. process such that each of the X_j have finite mean μ and finite variance σ^2. Then:*

1. *if A_n is the (non-normalized) average of the first n X_j, A_n is approximated by a normal r.v. with parameters μ and $\frac{\sigma^2}{n}$.*
2. *if S_n is the sum of the first n X_j, S_n is approximated by a normal r.v. with parameters μn and $n\sigma^2$.*

The CLT is usually shorthanded as follows:

"If $\{X_t\}$ is i.i.d., then $\boxed{A_n \approx n\left(\mu, \frac{\sigma^2}{n}\right)}$ and $\boxed{S_n \approx n\left(n\mu, n\sigma^2\right)}$."

We'll return to applications of the CLT later.

Properties of normal r.v.s

Theorem 6.12 (Properties of normal r.v.s) *Let $X \sim n(\mu, \sigma^2)$. Then:*

Every normal r.v. is linear transformation of std. normal: $X = \mu + \sigma Z$ *where* $Z \sim n(0,1)$.

Linear transformations of normal r.v.s are normal: *If $X \sim n(\mu, \sigma^2)$, then $aX + b \sim n(a\mu + b, a^2\sigma^2)$.*

CDF: $F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$.

Density function: $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$.

Mean: $EX = \mu$.

Variance: $Var(X) = \sigma^2$.

MGF: $M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$.

Sums of $\perp$ normal r.v.s are normal: *If $X_j \sim n(\mu_j, \sigma_j^2)$ are independent normal r.v.s for $j \in \{1, ..., d\}$, then $X_1 + ... + X_d \sim n(\sum \mu_j, \sum \sigma_j^2)$.*

PROOF Throughout this proof, $Z \sim n(0,1)$.

The first statement is the definition of a normal r.v.

For the second statement, suppose $X \sim n(\mu, \sigma^2)$. Then $X = \mu + \sigma Z$.
Therefore $Y = aX + b = a(\mu + \sigma Z) + b = (a\mu + b) + (a\sigma)Z$.
So by the definition of a normal r.v., $Y \sim n(a\mu + b, a^2\sigma^2)$.

The CDF and PDF are direct calculations using transformation methods:

$$F_X(x) = P(X \leq x) = P(\mu + \sigma Z \leq x) =$$

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx}\left[\Phi\left(\frac{x-\mu}{\sigma}\right)\right] =$$

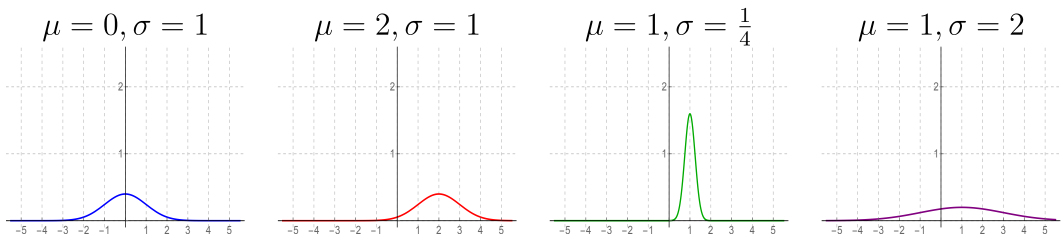
The mean, variance and MGF are HW problems (as a hint, use the fact $X = \mu + \sigma Z$; you already know the mean, variance and MGF of Z).

The last statement is also a HW problem, which is similar to some arguments we made in Chapter 4 about sums of other common r.v.s. □

EXAMPLE 20

Suppose X is normal with mean 20 and variance 36. Find, in terms of Φ, the probability that $12 < X \leq 20$.

Here are some plots of density functions for various values of μ and σ:



In general, the graph of the density function of any normal r.v. is a "bell curve" which has its peak at μ and inflection points at $\mu \pm \sigma$ (HW). This means that if σ is small, then the function has a tall, skinny peak (meaning that X takes values close to μ with very high probability) and if σ is large, the function has a short, wide peak (meaning that the values of X are more spread out).

Normal random variables arise naturally as averages of i.i.d. processes; examples of data which can be assumed to be normally distributed include:

1. Heights of people;
2. Exam grades;
3. Velocities of gas particles (Maxwell's Law);
4. Measurement errors in lab experiments;
5. The change in the price of a stock over a fixed period of time.

A connection between normal and gamma r.v.s

Theorem 6.13 *Let $X \sim n(0, \sigma^2)$ and let $Y = X^2$. Then $Y \sim \Gamma\left(\frac{1}{2}, \frac{1}{2\sigma^2}\right)$.*

PROOF Y has range $[0, \infty)$; let $y \geq 0$. Then

$$\begin{aligned} F_Y(y) = P(Y \leq y) = P(X^2 \leq y) &= P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \\ &= \Phi\left(\frac{\sqrt{y}}{\sigma}\right) - \Phi\left(\frac{-\sqrt{y}}{\sigma}\right) \\ &= \Phi\left(\frac{\sqrt{y}}{\sigma}\right) - \left[1 - \Phi\left(\frac{\sqrt{y}}{\sigma}\right)\right] \\ &= 2\Phi\left(\frac{\sqrt{y}}{\sigma}\right) - 1. \end{aligned}$$

Therefore

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}\left[2\Phi\left(\frac{\sqrt{y}}{\sigma}\right) - 1\right] = 2\phi\left(\frac{\sqrt{y}}{\sigma}\right)\cdot\frac{1}{\sigma\,2\sqrt{y}}$$
$$= \frac{1}{\sigma\sqrt{y}}\cdot\frac{1}{\sqrt{2\pi}}\exp\left[\frac{-y}{2\sigma^2}\right]$$
$$= \frac{\left(\frac{1}{2\sigma^2}\right)^{1/2}}{\sqrt{\pi}}y^{\frac{1}{2}-1}e^{-(1/2\sigma^2)y}$$

At the same time, the density of a $\Gamma\left(\frac{1}{2}, \frac{1}{2\sigma^2}\right)$ r.v. is

$$f_{\Gamma(\frac{1}{2},\frac{1}{2\sigma^2})}(y) = \frac{\left(\frac{1}{2\sigma^2}\right)^{1/2}}{\Gamma\left(\frac{1}{2}\right)}y^{\frac{1}{2}-1}e^{-(1/2\sigma^2)y}$$

Corollary 6.14 $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

6.5 Applications of the Central Limit Theorem

RECALL

The effective content of the CLT is as follows:

$$\text{“If } \{X_t\} \text{ is i.i.d., then } \boxed{A_n \approx n\left(\mu, \frac{\sigma^2}{n}\right)} \text{ and } \boxed{S_n \approx n\left(\mu n, n\sigma^2\right)}.\text{”}$$

These ideas can be used to approach many applied problems:

EXAMPLE 4

The weight of an adult male follows a probability distribution with mean 165 lb and standard deviation 30 lb. Compute the probability that 25 adult males collectively weigh at most 4400 lbs.

The continuity correction

If the quantity being studied in the CLT is *continuous* (like a length of time, or a physical measurement like temperature, mass, force, velocity, etc.), then most problems estimating the sum of i.i.d. copies of that quantity work like Example 4 above. However, if the quantity being studied in the CLT is *discrete* (like coin flips, poll results, dice rolls, etc.), then we have to tweak our procedure by applying what is called the *continuity correction*:

EXAMPLE 5

A basketball player expects to make 80% of his free throws (assume the result of each free throw is independent of any of the others). Use the Central Limit Theorem to estimate the probability that he makes at least 252 of 300 attempts.

EXAMPLE 6

The amount of gasoline purchased each week at a gas station follows a normal distribution with mean 60000 gal and standard deviation 10000 gal. If the gas station currently has a supply of 85000 gal and takes a weekly delivery of 57000 gal:

1. Compute the probability that, after 11 weeks, the gas station has a supply of at least 20000 gal.
2. What would the weekly delivery have to be, to ensure that after 11 weeks the gas station is 99.5% likely to have a supply of at least 20000 gal?

6.6 Stirling's formula

Theorem 6.15 (Stirling's Formula)

$$\lim_{n\to\infty} \frac{n!}{n^n e^{-n}\sqrt{2\pi n}} = 1 \qquad \textit{(more generally, } \lim_{n\to\infty} \frac{\Gamma(n+1)}{n^n e^{-n}\sqrt{2\pi n}} = 1\textit{).}$$

CONSEQUENCE

For large n, $n! = \Gamma(n+1)$ is approximately equal to $n^n e^{-n}\sqrt{2\pi n}$.
So in many instances (i.e. proofs), $n!$ can be replaced with $n^n e^{-n}\sqrt{2\pi n}$ without a problem.

PROOF (OF STIRLING'S FORMULA) Define $\psi : \mathbb{R} \to \mathbb{R}$ by

$$\psi(x) = \begin{cases} \dfrac{2}{x^2}(e^x - 1 - x) & \text{if } x \neq 0 \\ 1 & \text{if } x = 0 \end{cases}$$

By L'Hôpital's Rule, $\lim\limits_{x\to 0} \psi(x) = 1$. Therefore ψ is everywhere continuous.

Next, define $f : [0,1] \to \mathbb{R}$ by $f(t) = \dfrac{1}{\sqrt{\pi}} \displaystyle\int_{-\infty}^{\infty} \exp\left[-x^2\psi(xt)\right]\, dx$.
This improper integral does indeed converge.
Furthermore, f is continuous of t by the FTC, since it is defined as an integral of a continuous integrand.

We will prove Stirling's formula by computing $f(0)$ in two different ways.

First, we compute $f(0)$ directly:

$$\begin{aligned} f(0) &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left[-x^2\psi(0)\right]\, dx \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left[-x^2(1)\right]\, dx \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left[\frac{-x^2}{2\left(\sqrt{\frac{1}{2}}\right)^2}\right] dx \\ &= \frac{1}{\sqrt{\pi}} \sqrt{\frac{1}{2}}\sqrt{2\pi} \qquad \text{(Gaussian Integral Formula with } \mu = 0, \sigma = \sqrt{\tfrac{1}{2}}\text{)} \\ &= 1. \end{aligned}$$

Now, we compute $f(0)$ a different way, with a lot of symbol-crunching:

$$\begin{aligned} f(t) &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left[-x^2 \psi(xt)\right] dx \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left[-x^2 \left(\frac{2}{x^2 t^2}\right)\left(e^{xt} - 1 - xt\right)\right] dx \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left(\frac{-2}{t^2} e^{xt}\right) \exp\left(\frac{2}{t^2}\right) \exp\left(\frac{2x}{t}\right) dx \end{aligned}$$

Next, perform the u-sub $u = \frac{2}{t^2} e^{xt}$; $du = \frac{2}{t^2} e^{xt} t\, dx = ut\, dx$.
When doing this substitution,

$$e^{xt} = \frac{t^2 u}{2} \Rightarrow \exp\left(\frac{2x}{t}\right) = \exp\left(xt \cdot \frac{2}{t^2}\right) = [\exp(xt)]^{2/t^2} = \left(\frac{t^2 u}{2}\right)^{2/t^2}.$$

So after the substitution, we get

$$\begin{aligned} f(t) &= \frac{1}{\sqrt{\pi}} \exp\left(\frac{2}{t^2}\right) \int_0^{\infty} e^{-u} \left(\frac{t^2 u}{2}\right)^{2/t^2} \frac{1}{tu}\, du \\ &= \frac{1}{t\sqrt{\pi}} \exp\left(\frac{2}{t^2}\right) \left(\frac{t^2}{2}\right)^{2/t^2} \int_0^{\infty} e^{-u} u^{\left(\frac{2}{t^2} - 1\right)}\, du \\ &= \frac{1}{t\sqrt{\pi}} \exp\left(\frac{2}{t^2}\right) \left(\frac{t^2}{2}\right)^{2/t^2} \Gamma\left(\frac{2}{t^2}\right). \end{aligned}$$

(Gamma Integral Formula with $r = \frac{2}{t^2}$, $\lambda = 1$)

Now let $n = \frac{2}{t^2}$ so that $t = \sqrt{\frac{2}{n}}$. This makes

$$f(t) = f\left(\sqrt{\frac{2}{n}}\right) = \frac{1}{\sqrt{\frac{2\pi}{n}}} e^n \left(\frac{1}{n}\right)^n \Gamma(n) = \frac{1}{\sqrt{\frac{2\pi}{n}}} e^n n^{-n} \frac{\Gamma(n+1)}{n} = \frac{\Gamma(n+1)}{n^n e^{-n} \sqrt{2\pi n}}.$$

6.7 Bivariate normal densities

In the next two sections, we generalize the idea of a normal random variable to higher dimensions. These distributions arise in statistics, econometrics, signal processing and other fields. We will start by studying the situation in dimension 2.

To get started, we need some notation associated to any joint distribution:

Definition 6.16 *Let $\mathbf{X} = (X, Y)$ be a joint distribution of two real-valued r.v.s X and Y. The* **mean vector** *of $\mathbf{X}$ is the 2×1 vector*

$$\mu = \begin{pmatrix} EX \\ EY \end{pmatrix}_{2\times 1} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}_{2\times 1}.$$

The **covariance matrix** *of $\mathbf{X}$ is the 2×2 matrix*

$$\Sigma = \begin{pmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{pmatrix}_{2\times 2} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}_{2\times 2}.$$

Keep in mind that we think of vectors in $\mathbb{R}^2$ as 2×1 matrices, so their transposes are row matrices. For example,

$$\mu^T = \begin{pmatrix} \mu_1 & \mu_2 \end{pmatrix}_{1\times 2} \qquad \mathbf{b}^T = \begin{pmatrix} b_1 & b_2 \end{pmatrix}_{1\times 2} \qquad \text{etc.}$$

Theorem 6.17 (Properties of mean vectors and covariance matrices) *Let $\mathbf{X} = (X, Y)$ have mean vector μ and covariance matrix Σ. Then:*

1. *Σ is a symmetric matrix (meaning $\Sigma^T = \Sigma$).*
2. *The diagonal entries of Σ are the variances of X and Y.*
3. *For any vector $\mathbf{b} = (b_1, b_2) = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}_{2\times 1} \in \mathbb{R}^2$, $\mathbf{b} \cdot \mathbf{X} = b_1 X + b_2 Y$ is a real-valued r.v. with*

$$E[\mathbf{b} \cdot \mathbf{X}] = \mathbf{b} \cdot \mu \quad \textit{and} \quad Var(\mathbf{b} \cdot \mathbf{X}) = \mathbf{b}^T \Sigma \mathbf{b}.$$

4. *Σ is nonnegative definite (meaning that for any vector $\mathbf{b} \in \mathbb{R}^2$, $\mathbf{b}^T \Sigma \mathbf{b} \geq 0$).*
5. *$\det \Sigma = \sigma_X^2 \sigma_Y^2 - \sigma_{XY}^2 \geq 0$.*

PROOF Statements (1) and (2) are obvious.

For statement (3), we have

$$E[\mathbf{b}\cdot\mathbf{X}] = E[b_1X + b_2Y] = b_1EX + b_2EY = (b_1, b_2)\cdot(EX, EY) = \mathbf{b}\cdot\mu$$

and

$$\begin{aligned} Var(\mathbf{b}\cdot\mathbf{X}) = Var[b_1X + b_2Y] &= Var[b_1X] + Var[b_2Y] + 2Cov[b_1X, b_2Y] \\ &= b_1^2Var(X) + b_2^2Var(Y) + 2b_1b_2Cov(X,Y). \end{aligned}$$

At the same time,

$$\begin{aligned} \mathbf{b}^T\Sigma\mathbf{b} &= \begin{pmatrix} b_1 & b_2 \end{pmatrix}\begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \\ &= \begin{pmatrix} b_1 & b_2 \end{pmatrix}\begin{pmatrix} b_1Var(X) + b_2Cov(X,Y) \\ b_1Cov(X,Y) + b_2Var(Y) \end{pmatrix} \\ &= b_1^2Var(X) + b_2^2Var(Y) + 2b_1b_2Cov(X,Y), \end{aligned}$$

the same as what we computed for $Var(\mathbf{b}\cdot\mathbf{X})$.

For statement (4), notice $\mathbf{b}^T\Sigma\mathbf{b} = Var(\mathbf{b}\cdot\mathbf{X}) \geq 0$ (since all variances are nonnegative).

For the last statement, any nonnegative definite matrix has nonnegative determinant. □

Now for a definition of the class of joint distributions we want to study:

Definition 6.18 *Let $\mathbf{X} = (X, Y)$ be a joint distribution of two real-valued r.v.s X and Y. $\mathbf{X}$ is called* **bivariate normal** *(a.k.a.* **joint(ly) normal** *a.k.a.* **Gaussian** *if every finite linear combination of X and Y is normal, meaning that for any $\mathbf{b} = (b_1, b_2) \in \mathbb{R}^2$, $\mathbf{b}\cdot\mathbf{X} = b_1X + b_2Y$ is a normal r.v.*

Lemma 6.19 *If $\mathbf{X} = (X, Y)$ is bivariate normal, then X and Y are normal.*

PROOF

NOTE: The converse of this lemma is false: just because the X_j are normal does not mean they have a joint normal distribution (example in HW).

Lemma 6.20 *Suppose* $\mathbf{X} = (X, Y)$ *has a joint normal density with mean vector* μ *and covariance matrix* Σ. *Let* $V = b_1 X + b_2 Y = \mathbf{b} \cdot \mathbf{X}$. *Then*

$$V \sim n\left(\mathbf{b} \cdot \mu, \mathbf{b}^T \Sigma \mathbf{b}\right).$$

PROOF The definition of bivariate normal tells us V is normal; the mean and variance of V follow from (3) of Theorem 6.17. □

Main concept of this section

Everything you need to know about a bivariate normal distribution $\mathbf{X}$ can be determined from its mean vector μ and its covariance matrix Σ.

The joint MGF of a bivariate normal distribution

Theorem 6.21 *Suppose* $\mathbf{X} = (X, Y)$ *is bivariate normal with mean vector* μ *and covariance matrix* Σ. *Then the joint MGF of* $\mathbf{X}$ *is*

$$M_{\mathbf{X}}(\mathbf{t}) = \exp\left(\mathbf{t} \cdot \mu + \frac{1}{2}\mathbf{t}^T \Sigma \mathbf{t}\right).$$

PROOF Let $V = \mathbf{t} \cdot \mathbf{X} = t_1 X + t_2 Y$.

By Lemma 6.20, $V \sim n(\mathbf{t} \cdot \mu, \mathbf{t}^T \Sigma \mathbf{t})$ so

$$M_{\mathbf{X}}(\mathbf{t}) = E[e^{\mathbf{t} \cdot \mathbf{X}}] = E[e^V] = E[e^{1V}] = M_V(1) = \exp\left(\mathbf{t} \cdot \mu + \frac{1}{2}\mathbf{t}^T \Sigma \mathbf{t}\right)$$

as wanted. □

What this theorem tells us: If $\mathbf{X}$ is bivariate normal, we can compute the MGF $M_{\mathbf{X}}(\mathbf{t})$ from μ and Σ.

Since we have inversion formulas allowing us to compute the density of $\mathbf{X}$ from its MGF, the density of $\mathbf{X}$ will also depend only on μ and Σ.

What this theorem also will tell us: In the next corollary, we will see that for bivariate normal densities, uncorrelated marginals must be independent. (Recall that this isn't true in general.)

Corollary 6.22 *If $\mathbf{X} = (X, Y)$ is bivariate normal and $Cov(X, Y) = 0$, then $X \perp Y$.*

PROOF We have $X \sim n(\mu_X, \sigma_X^2)$ and $Y \sim n(\mu_Y, \sigma_Y^2)$. If $Cov(X, Y) = 0$, then

$$\begin{aligned}
M_{\mathbf{X}}(\mathbf{t}) &= \exp\left[\mathbf{t} \cdot \mu + \frac{1}{2}\mathbf{t}^T \Sigma \mathbf{t}\right] \\
&= \exp\left[(t_1, t_2) \cdot (\mu_X, \mu_Y) + \frac{1}{2}\begin{pmatrix} t_1 & t_2 \end{pmatrix}\begin{pmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{pmatrix}\begin{pmatrix} t_1 \\ t_2 \end{pmatrix}\right] \\
&= \exp\left[t_1\mu_X + t_2\mu_Y + \frac{1}{2}\sigma_X^2 t_1^2 + \frac{1}{2}\sigma_Y^2 t_2^2\right] \\
&= \exp\left[\mu_X t_1 + \frac{1}{2}\sigma_X^2 t_1^2\right] \exp\left[\mu_Y t_2 + \frac{1}{2}\sigma_Y^2 t_2^2\right] \\
&= M_X(t_1) M_Y(t_2).
\end{aligned}$$

Since $M_{\mathbf{X}}(\mathbf{t}) = M_X(t_1)M_Y(t_2)$, $X \perp Y$ by Theorem 5.45. □

Theorem 6.23 (Decomposition of bivariate normal) *Suppose $\mathbf{X} = (X, Y)$ is bivariate normal. Then there exist two normal r.v.s $\widehat{Y}$ and $\overline{Y}$ so that*

- $Y = \widehat{Y} + \overline{Y}$;
- $\widehat{Y}$ *is a multiple of* X;
- $\overline{Y} \perp X$.

PROOF Let $\widehat{Y} = \dfrac{\sigma_{XY}}{\sigma_X^2} X$ and $\overline{Y} = Y - \widehat{Y}$. The first two bullet points of the theorem are satisfied, so it remains to verify the third.

Observe next that for any $\mathbf{b} = (b_1, b_2)$,

$$\begin{aligned}
b_1 X + b_2 \overline{Y} = b_1 X + b_2\left(Y - \widehat{Y}\right) &= b_1 X + b_2\left(Y - \frac{\sigma_{XY}}{\sigma_X^2} X\right) \\
&= \left[b_1 - b_2\frac{\sigma_{XY}}{\sigma_X^2}\right] X + b_2 Y.
\end{aligned}$$

Since (X, Y) is normal, this linear combination is normal.

This verifies that any linear combination of X and $\overline{Y}$ is normal, meaning $(X, \overline{Y})$ is bivariate normal by definition.

Next,

$$\begin{aligned}
Cov(X,\overline{Y}) &= Cov\left(X, Y - \frac{\sigma_{XY}}{\sigma_X^2}X\right) \\
&= Cov(X,Y) - \frac{\sigma_{XY}}{\sigma_X^2}Cov(X,X) \\
&= \sigma_{XY} - \frac{\sigma_{XY}}{\sigma_X^2}\sigma_X^2 = 0
\end{aligned}$$

so by Corollary 6.22, $X \perp \overline{Y}$. $\square$

Remarks: This may remind you of something you learn in linear algebra: *projections*. To project a vector $\mathbf{y}$ onto another nonzero vector $\mathbf{x}$, we think of a picture like this:

The formula to compute this projection $\hat{\mathbf{y}}$ is $\hat{\mathbf{y}} = \dfrac{\mathbf{x}\cdot\mathbf{y}}{\mathbf{x}\cdot\mathbf{x}}\mathbf{x}$. (Compare this with the formula for $\widehat{Y}$ above; it's the same except that the dot products are replaced with covariances.)

Definition 6.24 *Let $\mathbf{X} = (X, Y)$ be bivariate normal. The normal r.v. $\widehat{Y} = \dfrac{\sigma_{XY}}{\sigma_X^2}X$ is called the* **projection** *of Y onto X.*

The joint density of a bivariate normal

Lemma 6.25 *Suppose $\mathbf{X} = (X, Y)$ is bivariate normal with mean vector μ and covariance matrix Σ. Then there are independent, standard normal r.v.s W and Z so that*

$$\begin{aligned}
X &= \sigma_X W + \mu_X \\
Y &= \frac{\sigma_{XY}}{\sigma_X}W + \frac{\sqrt{\det\Sigma}}{\sigma_X}Z + \mu_Y
\end{aligned}$$

.

Proof First, set $W = \dfrac{X-\mu_X}{\sigma_X}$; W is the $z-$score of X so $W \sim n(0,1)$. We have

$$\sigma_X W + \mu_X = \sigma_X W + \mu_X = \sigma_X\left(\frac{X-\mu_X}{\sigma_X}\right) + \mu_X = X.$$

That leaves figuring out what Z is. Toward that end, let $\widehat{Y}$ be the projection of Y onto X and let $\overline{Y} = Y - \widehat{Y}$.
Note $E[\overline{Y}] = EY - E\widehat{Y} = \mu_Y - \frac{\sigma_{XY}}{\sigma_X^2}\mu_X$ and

$$\begin{aligned}
Var(\overline{Y}) &= Var(Y - \widehat{Y}) \\
&= Var\left(Y - \frac{\sigma_{XY}}{\sigma_X^2}X\right) \\
&= Var(Y) + \frac{\sigma_{XY}^2}{\sigma_X^4}Var(X) - 2\frac{\sigma_{XY}}{\sigma_X^2}Cov(X,Y) \\
&= \sigma_Y^2 + \frac{\sigma_{XY}^2}{\sigma_X^2} - 2\frac{\sigma_{XY}^2}{\sigma_X^2} \\
&= \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2} = \frac{\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2}{\sigma_X^2} = \frac{\det \Sigma}{\sigma_X^2}.
\end{aligned}$$

Next, let

$$Z = \frac{\overline{Y} - E[\overline{Y}]}{\sqrt{Var(\overline{Y})}} = \frac{\overline{Y} - \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}\mu_X}{\frac{\sqrt{\det \Sigma}}{\sigma_X}}.$$

Z is the $z-$score of $\overline{Y}$, so $Z \sim n(0,1)$.
Also, since $X \perp \overline{Y}$, W depends only on X, and Z depends only on $\overline{Y}$, $W \perp Z$.
Last, we have

$$\begin{aligned}
&\frac{\sigma_{XY}}{\sigma_X}W + \frac{\sqrt{\det \Sigma}}{\sigma_X}Z + \mu_Y \\
&= \frac{\sigma_{XY}}{\sigma_X} \cdot \frac{1}{\sigma_X}(X - \mu_X) + \frac{\sqrt{\det \Sigma}}{\sigma_X}\left[\frac{\overline{Y} - \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}\mu_X}{\frac{\sqrt{\det \Sigma}}{\sigma_X}}\right] + \mu_Y \\
&= \frac{\sigma_{XY}}{\sigma_X^2}(X - \mu_X) + (Y - \widehat{Y}) - \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}\mu_X + \mu_Y \\
&= \frac{\sigma_{XY}}{\sigma_X^2}(X - \mu_X) + Y - \frac{\sigma_{XY}}{\sigma_X^2}X + \frac{\sigma_{XY}}{\sigma_X^2}\mu_X \\
&= \left[\frac{\sigma_{XY}}{\sigma_X^2} - \frac{\sigma_{XY}}{\sigma_X^2}\right]X + Y - \frac{\mu_X\sigma_{XY}}{\sigma_X^2} + \frac{\mu_X\sigma_{XY}}{\sigma_X^2} \\
&= Y
\end{aligned}$$

as wanted. □

Theorem 6.26 *Suppose* $\mathbf{X} = (X, Y)$ *is bivariate normal with mean vector* μ *and covariance matrix* Σ. *Then*

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{2\pi\sqrt{\det\Sigma}} \exp\left[\frac{-1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right].$$

PROOF From the preceding lemma, we can write

$$\begin{aligned} X &= \sigma_X W + \mu_X \\ Y &= \frac{\sigma_{XY}}{\sigma_X} W + \frac{\sqrt{\det\Sigma}}{\sigma_X} Z + \mu_Y \end{aligned}$$

where (W, Z) are $\perp$ $n(0,1)$ r.v.s.

We compute the joint density of X and Y with Jacobians. Let

$$\varphi(w, z) = \left(\sigma_X w + \mu_X, \frac{\sigma_{XY}}{\sigma_X} w + \frac{\sqrt{\det\Sigma}}{\sigma_X} z + \mu_Y\right)$$

so that $\varphi(W, Z) = (X, Y)$. Notice $J(\varphi) = \det\begin{pmatrix} \sigma_X & 0 \\ \frac{\sigma_{XY}}{\sigma_X} & \frac{\sqrt{\det\Sigma}}{\sigma_X} \end{pmatrix} = \sqrt{\det\Sigma}$.

This means

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{1}{|J(\varphi)|} f_{W,Z}(w,z) \\ &= \frac{1}{\sqrt{\det\Sigma}} f_W(w)\, f_Z(z) \qquad \text{(since } W \perp Z\text{)} \\ &= \frac{1}{\sqrt{\det\Sigma}} \left(\frac{1}{\sqrt{2\pi}} e^{-w^2/2}\right)\left(\frac{1}{\sqrt{2\pi}} e^{-z^2/2}\right) \qquad \text{(since } W, Z \sim n(0,1)\text{)} \\ &= \frac{1}{2\pi\sqrt{\det\Sigma}} \exp\left[-\frac{1}{2}(w^2+z^2)\right] \\ &= \frac{1}{2\pi\sqrt{\det\Sigma}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right] \end{aligned}$$

as wanted (the red parts above are equal because of a horrible algebra calculation shown below).

Why are the red parts equal?

First, by back-solving for w and z in terms of x and y, we get

$$w = \frac{x-\mu_X}{\sigma_X} \quad \text{and} \quad z = \frac{\sigma_X}{\sqrt{\det\Sigma}}\left(y - \mu_Y - \frac{\sigma_{XY}(x-\mu_X)}{\sigma_X^2}\right).$$

Then the red parts are

$$
\begin{aligned}
&w^2+z^2\\
&= \frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{\sigma_X^2}{\det\Sigma}\left(y-\mu_Y - \frac{\sigma_{XY}(x-\mu_X)}{\sigma_X^2}\right)^2\\
&= \frac{(x-\mu_X)^2\det\Sigma}{\sigma_X^2\det\Sigma} + \frac{\sigma_X^2}{\det\Sigma}\left[(y-\mu_Y)^2 - \frac{2\sigma_{XY}(x-\mu_X)(y-\mu_Y)}{\sigma_X^2} + \frac{\sigma_{XY}^2(x-\mu_X)^2}{\sigma_X^4}\right]\\
&= \frac{1}{\det\Sigma}\frac{\det\Sigma}{\sigma_X^2}(x-\mu_X)^2 + \frac{1}{\det\Sigma}\left[\sigma_X^2(y-\mu_Y)^2 - 2\sigma_{XY}(x-\mu_X)(y-\mu_Y) + \frac{\sigma_{XY}^2}{\sigma_X^2}(x-\mu_X)^2\right]\\
&= \frac{1}{\det\Sigma}\left[\left(\frac{\det\Sigma}{\sigma_X^2} + \frac{\sigma_{XY}^2}{\sigma_X^4}\right)(x-\mu_X)^2 - 2\sigma_{XY}(x-\mu_X)(y-\mu_Y) + \sigma_X^2(y-\mu_Y)^2\right]\\
&= \frac{1}{\det\Sigma}\left[\left(\frac{\det\Sigma}{\sigma_X^2} + \frac{\sigma_{XY}^2}{\sigma_X^4}\right)(x-\mu_X)^2 - 2\sigma_{XY}(x-\mu_X)(y-\mu_Y) + \sigma_X^2(y-\mu_Y)^2\right]\\
&= \frac{1}{\det\Sigma}\left[\left(\frac{\sigma_X^2\sigma_Y^2-\sigma_{XY}^2}{\sigma_X^2} + \frac{\sigma_{XY}^2}{\sigma_X^2}\right)(x-\mu_X)^2 - 2\sigma_{XY}(x-\mu_X)(y-\mu_Y) + \sigma_X^2(y-\mu_Y)^2\right]\\
&= \frac{1}{\det\Sigma}\left[\sigma_Y^2(x-\mu_X)^2 - 2\sigma_{XY}(x-\mu_X)(y-\mu_Y) + \sigma_X^2(y-\mu_Y)^2\right]\\
&= \frac{1}{\det\Sigma}\left[(x-\mu_X)(\sigma_Y^2(x-\mu_X) - \sigma_{XY}(y-\mu_Y)) + (y-\mu_Y)(-\sigma_{XY}(x-\mu_X) + \sigma_X^2(y-\mu_Y))\right]\\
&= \frac{1}{\det\Sigma}\begin{pmatrix} x-\mu_X & y-\mu_Y\end{pmatrix}\begin{pmatrix}\sigma_Y^2(x-\mu_X)-\sigma_{XY}(y-\mu_Y)\\ -\sigma_{XY}(x-\mu_X)+\sigma_X^2(y-\mu_Y)\end{pmatrix}\\
&= \begin{pmatrix} x-\mu_X & y-\mu_Y\end{pmatrix}\frac{1}{\det\Sigma}\begin{pmatrix}\sigma_Y^2 & -\sigma_{XY}\\ -\sigma_{XY} & \sigma_X^2\end{pmatrix}\begin{pmatrix}x-\mu_X\\ y-\mu_Y\end{pmatrix}\\
&= \begin{pmatrix} x-\mu_X & y-\mu_Y\end{pmatrix}\begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{XY} & \sigma_Y^2\end{pmatrix}^{-1}\begin{pmatrix}x-\mu_X\\ y-\mu_Y\end{pmatrix}\\
&= (\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu).
\end{aligned}
$$

This completes the proof. □

Practical computational stuff

Let X and Y have a bivariate normal distribution. Write

$$\Sigma^{-1} = \begin{pmatrix} a & b\\ b & d\end{pmatrix}.$$

Now

$$
\begin{aligned}
f_{X,Y}(x,y) &= \frac{1}{2\pi\sqrt{\det\Sigma}}\exp\left[\frac{-1}{2}(\mathbf{x}-\mu)^T\,\Sigma^{-1}\,(\mathbf{x}-\mu)\right]\\
&= \frac{1}{2\pi\sqrt{\det\Sigma}}\exp\left[\heartsuit\right]
\end{aligned}
$$

where

$$\begin{aligned}
\heartsuit &= \frac{-1}{2}\begin{pmatrix} x-\mu_X & y-\mu_Y \end{pmatrix}\begin{pmatrix} a & b \\ b & d \end{pmatrix}\begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix} \\
&= \frac{-1}{2}\begin{pmatrix} x-\mu_X & y-\mu_Y \end{pmatrix}\begin{pmatrix} ax-a\mu_X & by-b\mu_Y \\ bx-b\mu_X & dy-d\mu_Y \end{pmatrix} \\
&= \frac{-1}{2}\Big[ax^2 - a\mu_X x + bxy - b\mu_Y x - a\mu_X x + a\mu_X^2 - b\mu_X y + b\mu_X\mu_Y \\
&\qquad\qquad +bxy - b\mu_X y + dy^2 d\mu_Y y - b\mu_Y x + b\mu_X\mu_Y - d\mu_X\mu_Y + d\mu_Y^2\Big]
\end{aligned}$$

$$\begin{aligned}
\heartsuit &= \frac{-1}{2}\Big[ax^2 + 2bxy + dy^2 + (-2a\mu_X - 2b\mu_Y)x + (-2b\mu_X - 2d\mu_Y)y \\
&\qquad\qquad +(a\mu_X^2 + 2b\mu_X\mu_Y + d\mu_Y^2)\Big] \\
&= \frac{-1}{2}ax^2 - bxy - \frac{1}{2}dy^2 + (a\mu_X + b\mu_Y)x + (b\mu_X + d\mu_Y)y + (\text{constant}).
\end{aligned}$$

Punchline: Given $f_{X,Y}$ for a bivariate normal distribution (X, Y):

1. You can find a, b and d by looking at the coefficients on the x^2, xy and y^2 terms inside the exponential part of $f_{X,Y}$, respectively:

$$\begin{aligned}
a &= -2(\text{coefficient on } x^2 \text{ in } \heartsuit) \\
b &= -(\text{coefficient on } xy \text{ in } \heartsuit) \\
d &= -2(\text{coefficient on } y^2 \text{ in } \heartsuit)
\end{aligned}$$

   This tells you $\Sigma^{-1} = \begin{pmatrix} a & b \\ b & d \end{pmatrix}$.

2. You can then find Σ by inverting Σ^{-1}. This will tell you the variances of X and Y, and the covariance between them.

3. You can recover the expected values of X and Y by solving a system of equations coming from the coefficients on the x and y terms inside the exponential part of $f_{X,Y}$.

EXAMPLE 15

Suppose X and Y have bivariate normal density

$$f_{X,Y}(x,y) = K \exp\left[\frac{-1}{2}(5x^2 - 6xy + 2y^2 - 40x + 24y + 80)\right].$$

1. Write down the covariance matrix Σ.
2. Compute K.
3. Compute the covariance between X and Y.
4. Compute the variances of X and Y.
5. Compute the expected values of X and Y.
6. Determine the density functions of the marginals.
7. Let $V = 7X - 3Y$.
   a) Compute the mean and variance of V.
   b) Write down a density function of V.
   c) Compute the probability that $V \leq 30$.

6. Both X and Y are normal, so

$$X \sim n(4,2) \Rightarrow \boxed{f_X(x) = \frac{1}{\sqrt{2}\sqrt{2\pi}} \exp\left[\frac{-(x-4)^2}{2 \cdot 2}\right]}$$

$$Y \sim n(0,5) \Rightarrow \boxed{f_Y(y) = \frac{1}{\sqrt{5}\sqrt{2\pi}} \exp\left[\frac{-y^2}{2 \cdot 5}\right]}$$

7. $V = 7X - 3Y = (7,-3) \cdot (X,Y) = \mathbf{b} \cdot \mathbf{X}$ where $\mathbf{b} = (7,-3)$.
   Therefore $V \sim n(\mathbf{b} \cdot \mu, \mathbf{b}^T \Sigma \mathbf{b})$.
   We have $EV = \mathbf{b} \cdot \mu = (7,-3) \cdot (4,0) = \boxed{28}$
   and $VarV = \mathbf{b}^T \Sigma \mathbf{b} = \begin{pmatrix} 7 & -3 \end{pmatrix} \begin{pmatrix} 2 & 3 \\ 3 & 5 \end{pmatrix} \begin{pmatrix} 7 \\ -3 \end{pmatrix} = \begin{pmatrix} 7 & -3 \end{pmatrix} \begin{pmatrix} 5 \\ 6 \end{pmatrix} = \boxed{17}$.
   Therefore $V \sim n(28,17)$, so $\boxed{f_V(v) = \frac{1}{\sqrt{17}\sqrt{2\pi}} \exp\left[\frac{-(v-28)^2}{2 \cdot 17}\right]}$.
   Last, $V \sim n(28,17)$ means $V = 28 + \sqrt{17}Z$ where $Z \sim n(0,1)$.
   Finally,
   $$P(V \leq 30) = P(28 + \sqrt{17}Z \leq 30) = P\left(Z \leq \frac{2}{\sqrt{17}}\right) = \boxed{\Phi\left(\frac{2}{\sqrt{17}}\right)}$$
   $$= \Phi(.485) = \boxed{.686162}.$$

Conditional density, expectation and variance

Theorem 6.27 *Suppose* $\mathbf{X} = (X, Y)$ *is bivariate normal. Then* $Y|X$ *is normal with parameters*

$$E[Y|X] = \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(x - \mu_X); \qquad Var[Y|X] = \sigma_Y^2(1 - \rho^2).$$

(Here ρ *is the correlation between* X *and* Y*.)*

PROOF To compute $E[Y|X]$ and $Var[Y|X]$, let $\widehat{Y} = \frac{\sigma_{XY}}{\sigma_X^2} X$ be the projection of Y onto X, and let $\overline{Y} = Y - \widehat{Y}$. Recall from Lemma 6.25:

$$X \perp \overline{Y}; \qquad E\overline{Y} = \mu_Y - \frac{\sigma_{XY}}{\sigma_X^2}\mu_X; \qquad Var(\overline{Y}) = \frac{\det \Sigma}{\sigma_X^2} = \frac{\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2}{\sigma_X^2}.$$

Therefore

$$\begin{aligned}
E[Y|X] = E[\widehat{Y} + \overline{Y}|X] &= E[\widehat{Y}|X] + E[\overline{Y}|X] \\
&= E\left[\frac{\sigma_{XY}}{\sigma_X^2}X|X\right] + E[\overline{Y}] \qquad \text{(since } \overline{Y} \perp X\text{)} \\
&= \frac{\sigma_{XY}}{\sigma_X^2}E\left[X|X\right] + \mu_Y - \frac{\sigma_{XY}}{\sigma_X^2}\mu_X \\
&= \frac{\sigma_{XY}}{\sigma_X^2}x + \mu_Y - \frac{\sigma_{XY}}{\sigma_X^2}\mu_X \qquad \text{(since } E[X|X=x] = x\text{)} \\
&= \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(x - \mu_X)
\end{aligned}$$

and

$$\begin{aligned}
Var[Y|X] &= Var[\widehat{Y} + \overline{Y}|X] \\
&= Var[\widehat{Y}|X] + Var[\overline{Y}|X] + 2Cov[\widehat{Y}, \overline{Y}|X] \\
&= Var\left[\frac{\sigma_{XY}}{\sigma_X^2}X|X\right] + Var[\overline{Y}|X] + 2(0) \qquad \text{(since } \overline{Y} \perp \widehat{Y}\text{)} \\
&= \frac{\sigma_{XY}^2}{\sigma_X^4}Var\left[X|X\right] + Var(\overline{Y}) \qquad \text{(since } \overline{Y} \perp X\text{)} \\
&= \frac{\sigma_{XY}}{\sigma_X^2}0 + \frac{\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2}{\sigma_X^2} \qquad \text{(since } Var[X|X=x] = 0\text{)} \\
&= \sigma_Y^2\left(\frac{\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2}{\sigma_X^2\sigma_Y^2}\right) \\
&= \sigma_Y^2\left(1 - \left[\frac{\sigma_{XY}}{\sigma_X\sigma_Y}\right]^2\right) = \sigma_Y^2(1 - \rho^2).
\end{aligned}$$

Last, to see why $Y|X$ is normal, observe X is normal, so

$$\begin{aligned}
f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\
&= \frac{(*)\exp\left[(*)x^2 + (*)xy + (*)y^2 + (*)x + (*)y + (*)\right]}{(*)\exp\left[(*)x^2 + (*)x + (*)\right]} \\
&= (*)\exp\left[(*)x^2 + (*)xy + (*)y^2 + (*)x + (*)y + (*)\right]
\end{aligned}$$

which is the density of a normal r.v. □

EXAMPLE 16

Suppose (X, Y) have a bivariate normal density where

$$E[X|Y] = 3.7 - .15y; \qquad E[Y|X] = .4 - .6x; \qquad Var(Y|X) = 3.64.$$

1. Compute $\rho(X, Y)$.
2. Compute $Var(X)$.
3. Compute $Var(Y)$
4. Compute $Cov(X, Y)$.
5. Compute EX.
6. Compute EY.

6.8 Joint normal densities in higher dimensions

In this section, we briefly discuss how the machinery of the previous section can be adapted in dimension greater than 2.

Definition 6.28 *Let* $\mathbf{X}$ *be the joint distribution of real-valued r.v.s* $X_1, ..., X_d$. *The* **mean vector** μ *of* $\mathbf{X}$ *is*

$$\mu = \begin{pmatrix} EX_1 \\ EX_2 \\ \vdots \\ EX_d \end{pmatrix}_{d\times 1} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{pmatrix}_{d\times 1}.$$

The **covariance matrix** Σ *of* $\mathbf{X}$ *is*

$$\Sigma = \begin{pmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_d) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & \cdots & Cov(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_d, X_1) & Cov(X_d, X_3) & \cdots & Cov(X_d, X_d) \end{pmatrix}_{d\times d}.$$

Theorem 6.29 (Properties of mean vectors and covariance matrices) *Let* μ *and* Σ *be the mean vector and covariance matrix of any joint distribution* $\mathbf{X} = (X_1, ..., X_d)$. *Then:*

1. Σ *is* $d \times d$.
2. Σ *is symmetric (i.e.* $\Sigma^T = \Sigma$*).*
3. *The diagonal entries of* Σ *are the variances of the* X_j.
4. *For any vector* $\mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_d \end{pmatrix}$,

$$E[\mathbf{b} \cdot \mathbf{X}] = \mathbf{b} \cdot \mu \;\; \text{and} \;\; Var(\mathbf{b} \cdot \mathbf{X}) = Var\left(\sum_{j=1}^{d} b_j X_j\right) = \mathbf{b}^T \Sigma \mathbf{b}.$$

5. Σ *is nonnegative definite (for any vector* $\mathbf{b} \in \mathbb{R}^d$, $\mathbf{b}^T \Sigma \mathbf{b} \geq 0$*).*
6. $\det \Sigma \geq 0$.

Definition 6.30 *A collection* $\mathbf{X} = (X_1, ..., X_d)$ *of real-valued r.v.s with joint density* $f_{\mathbf{X}} : \mathbb{R}^d \to \mathbb{R}$ *is called* **joint(ly) normal** *or* **joint(ly) Gaussian** *if every finite linear combination of the marginals*

$$\mathbf{b} \cdot \mathbf{X} = \sum_{j=1}^{d} b_j X_j$$

(where $b_1, ..., b_d \in \mathbb{R}$*) is normal.*

Corollary 6.31 *If* $\mathbf{X}$ *is joint normal, then for any* $\mathbf{b} \in \mathbb{R}^d$*,*

$$\mathbf{b} \cdot \mathbf{x} \sim n\left(\mathbf{b} \cdot \mu, \mathbf{b}^T \Sigma \mathbf{b}\right).$$

Theorem 6.32 (Characterization of bivariate normal densities) *Let* $\mathbf{X}$ *be a joint distribution with mean vector* μ *and covariance matrix* Σ*. Then, the following are equivalent:*

1. $\mathbf{X}$ *is joint normal.*
2. $M_{\mathbf{X}}(\mathbf{t}) = \exp\left(\mathbf{t} \cdot \mu + \frac{1}{2}\mathbf{t}^T \Sigma \mathbf{t}\right)$.
3. $f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma}} \exp\left[\frac{-1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)\right]$.

Corollary 6.33 (Uniqueness of joint normal densities) *If two jointly normal distributions have the same means and same covariances between the marginals, then they are the same distribution.*

Corollary 6.34 *If* $\mathbf{X}$ *is a joint normal density such that* $Cov(X_i, X_j) = 0$ *for all* $i \neq j$*, then* $X_i \perp X_j$ *for all* $i \neq j$*.*

6.9 Chapter 6 Homework

Exercises from Section 6.1

1. Suppose $\{X_t\}$ is an i.i.d. sequence of exponential r.v.s, each having mean 3. If $X_1 = 2$, $X_2 = 4$, $X_3 = 1$ and $X_4 = 5$, compute the values of S_n, A_n and A_n^* for $n \in \{1, 2, 3, 4\}$.

Exercises from Section 6.2

2. Suppose that i.i.d. blood samples taken from a patient will show a hemoglobin level that averages 15 g/dl with a standard deviation of 3 g/dl. According to the QWLLN, what is the smallest number of samples that would need to be taken from this patient, so that the average hemoglobin level of the samples taken is 98% likely to lie between 14.9 g/dl and 15.1 g/dl?

3. Suppose $X_1, X_2, ...$ are i.i.d. r.v.s (taking only positive real values), each having finite mean μ. Show that with probability 1, the geometric averages of the X_j converge, where the *geometric average* of $X_1, ..., X_n$ is

$$G_n = \sqrt[n]{\prod_{j=1}^{n} X_j}.$$

Determine $\lim\limits_{n\to\infty} G_n$.

Hint: Apply the SLLN to $\log G_n$ (here $\log$ means natural logarithm).

Exercises from Section 6.3

4. For each $\lambda > 0$, let X_λ be Poisson with parameter λ and let $Y_\lambda = \dfrac{X_\lambda - \lambda}{\sqrt{\lambda}}$. Show that for all t, $\lim\limits_{\lambda\to\infty} M_{Y_\lambda}(t) = \exp\left(\dfrac{t^2}{2}\right)$.

5. Fix $\lambda > 0$ and for each $r > 0$ let X_r be $\Gamma(r, \lambda)$ and define $Y_r = \dfrac{X_\lambda - \left(\frac{r}{\lambda}\right)}{\left(\frac{\sqrt{r}}{\lambda}\right)}$. Show that for all t, $\lim\limits_{r\to\infty} M_{Y_r}(t) = \exp\left(\dfrac{t^2}{2}\right)$.

Exercises from Section 6.4

6. Suppose Z has the standard normal distribution. Compute decimal approximations to the following probabilities (trust me, there are no typos in these inequalities):

a) $P(Z < 1.33)$

b) $P(Z < -.425)$

c) $P(Z \geq .79)$

d) $P(.55 < Z < 1.22)$

e) $P(-1.90 \geq Z \geq .44)$

f) $P(Z > -.2)$

g) $P(-.63 \leq Z < .3)$

h) $P(Z < -1.3$ or $Z > .58)$

7. a) Prove that the mean of a $n(\mu, \sigma^2)$ r.v. is μ.
   b) Prove that the variance of a $n(\mu, \sigma^2)$ r.v. is σ^2.
   c) Verify that the MGF of a $n(\mu, \sigma^2)$ r.v. is $\exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$.

8. Prove that if $X_j \sim n(\mu_j, \sigma_j^2)$ are independent normal r.v.s, then $X_1 + ... + X_j \sim n\left(\sum_j \mu_j, \sum_j \sigma_j^2\right)$.

9. Let f be the density function of the normal r.v. with parameters μ and σ^2.
   a) Show that f has its maximum when $x = \mu$.
   b) Show that the $x-$coordinates of the inflection points of f are $x = \mu \pm \sigma$.

10. Suppose X is normal with parameters $\mu = 20$ and $\sigma^2 = 100$. Compute decimal approximations to the following probabilities:
    a) $P(X \geq 17)$
    b) $P(X < 24.5)$
    c) $P(X = 18)$
    d) $P(X < 17 \,|\, X \leq 23)$
    e) $P(X \geq 21 \,|\, X < 24)$
    f) $P(X < 15$ or $X \geq 24)$
    g) $P(W \leq 18)$, assuming $W = \frac{1}{2}X - 4$
    h) $P(Y \geq 54)$, assuming $Y = 3X + 9$

11. Let X and Y be independent standard normal r.v.s.
    a) Compute the joint density of X and Y/X.
       Hint: This is a transformation problem involving Jacobians.
    b) Compute the density of Y/X.
       Hint: Integrate the joint density you found in part (a) with respect to x.
    c) Simplify your answer to part (b), and identify Y/X as a common r.v.

12. Suppose that during periods of transcendental meditation the reduction of a person's oxygen consumption is a normal $n(37.6$ cc/min, 4.6^2 cc/min).

a) Calculate (a decimal approximation to) the probability that during a period of transcendental meditation a person's oxygen consumption will be reduced by at least 42.5 cc/min.

b) Calculate (a decimal approximation to) the probability that during a period of transcendental meditation a person's oxygen consumption will be reduced by anywhere from 30 to 40 cc/min.

13. A study shows that an experimental drug causes patient's blood pressure to lower by an amount that is normally distributed with parameters $\mu = 32$ mmHg and $\sigma^2 = 60$ mmHg. What is the probability that by taking this drug, a patient's blood pressure will be lowered by at least 25 mmHg but by no more than 35 mmHg? (Give both an answer in terms of Φ, and a decimal approximation.)

14. Let $X_\lambda \sim Pois(\lambda)$ and fix $c > 0$. Use the result of Exercise 4 to estimate, for large λ, the value of $P(X_\lambda \leq c\lambda)$. Your answer should be in terms of Φ, the cumulative distribution function of the standard normal r.v.

    Hint: Define Y_λ as in Exercise 4. Since, for large λ, $M_{Y_\lambda}(t) = \exp\left(\frac{t^2}{2}\right) = M_{n(0,1)}(t)$, that means by uniqueness of MGFs that $Y_\lambda \approx n(0,1)$. That means X_λ is approximately some other r.v.

15. Let Q be a Poisson r.v. with mean 5000. Use your answers to the previous parts of this question to estimate the following, in terms of Φ:

    a) $P(Q \leq 5100)$
    b) $P(Q < 4920)$
    c) $P(Q \geq 5050)$
    d) Let $X_r \sim \Gamma(r, \lambda)$. Use the result of Exercise 5 to estimate, for large r, the value of $\lim_{r\to\infty} P(X_r \leq r/\lambda)$. (Your answer should be in terms of Φ.)
    e) Suppose Q is a gamma r.v. with parameters $r = 2000$ and $\lambda = 2$. Estimate the following probabilities in terms of Φ:
        i. $P(Q \leq 1800)$
        ii. $P(Q > 1850)$
        iii. $P(Q < 2150)$

Exercises from Section 6.5

16. The amount of liquid a student puts in their drink at the Rock is a r.v. with mean 350 mL and variance 1500 mL. Use the Central Limit Theorem to estimate the probability that a randomly selected group of 12 students put an average of 320 mL or more in their drinks. Give both the exact answer in terms of Φ and a decimal approximation to the answer.

17. 1000 fair dice are rolled independently. Use the Central Limit Theorem to estimate the probability that the sum of these 1000 rolls is at least 3450 and no greater than 3650. Give both the exact answer in terms of Φ and a decimal approximation to the answer.

18. You play a game where you lose $\$1$ with probability .7, you lose $\$2$ with probability .2, and win $\$10$ with probability .1. If you play this game 10000 times, what is the probability that you will be ahead (that is, you have won more money than you have lost)? (You are to approximate this answer using the Central Limit Theorem; give both the exact answer in terms of Φ and a decimal approximation to the answer.)

19. Let $X \sim Pois(40)$. Let $p = P(X \geq 48)$. Approximate p using the Central Limit Theorem, by approximating X as the sum of 40 i.i.d. r.v.s. Give both the exact answer in terms of Φ and a decimal approximation to the answer.

20. A tobacco company claims that the amount of nicotine in one of its cigarettes is a r.v. with mean 2.2 mg and standard deviation $.8$ mg. Use the Central Limit Theorem to estimate the probability that 100 randomly chosen cigarettes would have an average nicotine content of at most 2.09 mg. Give both the exact answer in terms of Φ and a decimal approximation to the answer.

21. (AE) In an analysis of healthcare data, ages are rounded to the nearest multiple of 5 years. The difference between the true age and the rounded age is assumed to be uniformly distributed on the interval from -2.5 years to 2.5 years. The healthcare data is based on a random sample of 80 people. What is the approximate probability (as estimated using the CLT) that the mean of the rounded ages is within 0.125 years of the mean of the true ages?

22. (AE) A charity receives 3100 contributions, each of which are assumed to be independent and identically distributed with mean 150 and standard deviation 40. Use the CLT to approximate the number b so that it is 90% likely that the total contributions to the charity are less than or equal to b.

23. (AE) The total claim amount for a property insurance policy follows a distribution that has density function

$$f(x) = \frac{1}{2000} e^{-x/2000} \text{ for } x > 0.$$

The premium for the policy is set at 250 over the expected total claim amount. If the insurance company sells 300 policies, what is the approximate probability (as estimated using the CLT) that the insurance company will have claims exceeding the premiums collected?

Exercises from Section 6.6

24. Use Stirling's formula to show that for large n, $\binom{2n}{n} \approx \frac{4^n}{\sqrt{\pi n}}$.

    Remark: We will use this fact in MATH 416.

Exercises from Section 6.7

25. In this problem we show that just because X and Y are normal does not mean that they have a joint normal distribution. Let $X \sim n(0,1)$ and let R be uniform on the two numbers $\{-1,1\}$; suppose $R \perp X$. Let $Y = RX$.

    a) Prove (using transformation methods) that $Y \sim n(0,1)$.

    b) Prove that (X,Y) are <u>not</u> bivariate normal, by showing that the linear combination $X+Y$ is not normally distributed.

26. Suppose X and Y have the following bivariate normal density:

$$f_{X,Y}(x,y) = C \exp\left[\frac{-1}{54}\left(x^2 + 4y^2 + 2xy + 2x + 8y + 4\right)\right]$$

    Compute each quantity:

    a) EX

    b) EY

    c) $Var(X)$

    d) $Var(Y)$

    e) $Cov(X,Y)$

    f) $\rho(X,Y)$

    g) C

    h) The covariance matrix Σ

27. Let X and Y have the density given in Exercise 26.

    a) Compute the conditional density of Y given $X = x$.

    b) Compute the conditional density of X given $Y = -1$.

28. Let X and Y have the density given in Exercise 26.

    a) Compute the conditional variance of Y given $X = x$.

    b) Compute the density of $W = 8X + 5Y$.

Chapter 7

Applications to insurance

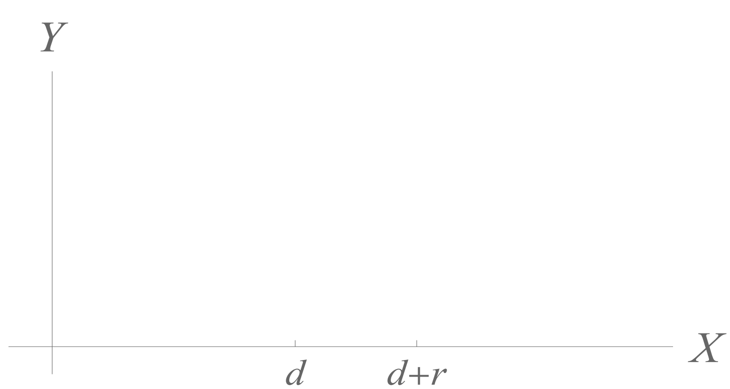
7.1 Deductibles

A common application of LOTUS occurs in the context of insurance. Most of the time, when you buy an insurance policy, the policy includes a **deductible** of some fixed amount d. That means that when you incur a loss covered by the insurance policy, you must pay the first d of the loss; the insurance company only covers anything that is left after that.

Why do insurance companies like deductibles?

1. Companies can pay out less in claims to policyholders
2. Reduces overhead costs (no processing of small claims)
3. Creates some risk for policyholder, which provides an incentive for the policyholder to be more risk-averse

Suppose a policyholder holding a policy with a deductible d incurs loss X (where X is some r.v.). If Y is the claim payment, what is Y as a function of X?



On the previous page we saw that

$$Y = \varphi(X) = \begin{cases} 0 & \text{if } X \leq d \\ X - d & \text{if } X > d \end{cases}$$

This means the cdf of Y can be computed as follows:

Therefore, Y has no density, because F_Y isn't continuous.

Nonetheless, we can use LOTUS to compute EY:

We have proven:

Theorem 7.1 (Expected value for insurance policy with deductible) *Suppose the loss incurred by a policyholder with deductible d is a cts r.v. X with finite expectation. If Y is the claim payment associated to this loss, then*

$$EY = \int_d^\infty (x - d) f_X(x)\, dx.$$

EXAMPLE 1

Suppose that a loss random variable is uniform on $[0, 10]$.

1. Compute the expected amount paid by the insurer, if the policy has a deductible of 3.

2. Compute the variance of the amount paid by the insurer, if the policy has a deductible of 3.

3. If a deductible is applied before any insurance payment, and the expected payment of the insurer is 1.5, determine the size of the deductible.

EXAMPLE 2

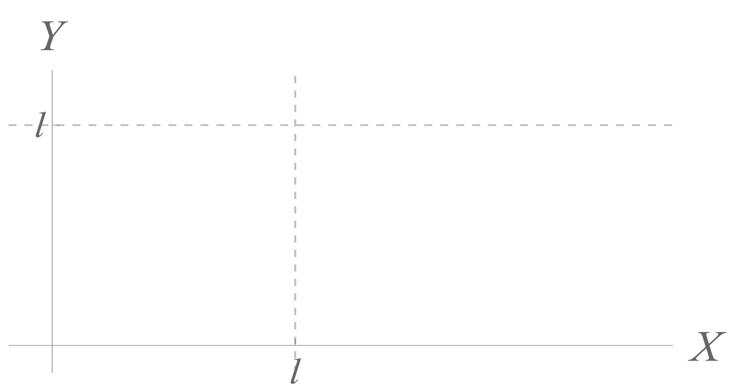
Suppose that a loss random variable is exponential with mean 10. If a deductible of size 5 is applied, find the expected payment of the insurer.

Solution: X exp. w/ mean 10 means $X \sim Exp(\quad)$, i.e. $f_X(x) =$

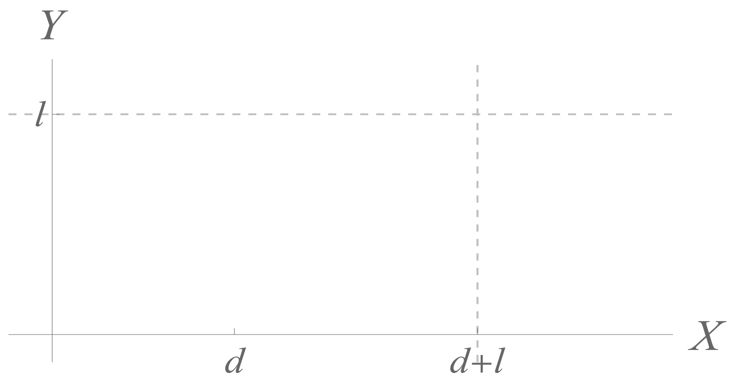
$$\begin{aligned} EY &= \int_d^\infty (x-d) f_X(x)\,dx \\ &= \int_5^\infty (x-5)\frac{1}{10}e^{-(1/10)x}\,dx \end{aligned}$$

7.2 Benefit limits

A second way that insurance companies mitigate their risk is by selling policies that have **benefit limits** (a.k.a. **coverage limits**). Suppose a policy has a benefit limit of l (and no deductible). This means that the maximum amount the insurance company will pay its policyholder for a loss is l. Then if loss X is incurred by the policyholder, the corresponding claim payment Y is



If there is both a benefit limit l and a deductible d, then if loss X is incurred by the policyholder, the corresponding claim payment Y is



Using LOTUS, we can derive these formulas:

Theorem 7.2 (Expected value for insurance policy with benefit limit) *Suppose the loss X incurred by a policyholder with a benefit limit l is a continuous r.v. with finite expectation. Then if Y is the claim payment associated to this loss,*

$$EY = \int_0^l x f_X(x)\, dx + l \cdot P(X \geq l).$$

PROOF This will follow from the next theorem by setting $d = 0$. □

Theorem 7.3 (Exp. value for policy with deductible and benefit limit) *Suppose the loss X incurred by a policyholder with deductible d and benefit limit l is a continuous r.v. with finite expectation. Then if Y is the claim payment associated to this loss,*

$$EY = \int_d^{d+l} (x-d) f_X(x)\, dx + l \cdot P(X \geq d+l).$$

PROOF From the previous page, we know

$$Y = \varphi(X) = \begin{cases} 0 & \text{if } X \leq d \\ X - d & \text{if } d < X < d+l \\ l & \text{if } X \geq d+l \end{cases}$$

So by LOTUS, we have

$$\begin{aligned}
EY &= \int_0^\infty \varphi(x) f_X(x)\,dx \\
&= \int_0^d 0 f_X(x)\,dx + \int_d^{d+l} (x-d) f_X(x)\,dx + \int_{d+l}^\infty l f_X(x)\,dx \\
&= 0 + \int_d^{d+l} (x-d) f_X(x)\,dx + l \int_{d+l}^\infty f_X(x)\,dx \\
&= \int_d^{d+l} (x-d) f_X(x)\,dx + l \cdot P(X \geq d+l). \ \square
\end{aligned}$$

EXAMPLE 3

Suppose that a loss random variable is uniform on $[0, 10]$. Find the expected amount paid by the insurer, if the policy has a deductible of 1 and a coverage limit of 6.

EXAMPLE 4

Suppose the loss from an accident is a continuous random variable with density $f(x) = \frac{24}{7}x^{-4}$ when $1 < x < 2$. Suppose that the insurance policy has a coverage limit of 1.5. What is the standard deviation of the loss to the insurance company?

Last, $Var(Y) = EY^2 - (EY)^2 = \frac{19}{12} - \left(\frac{157}{126}\right)^2 = \frac{122}{3969}$ and so

$$\sigma_Y = \sqrt{Var(Y)} = \sqrt{\frac{122}{3969}} = \boxed{\frac{\sqrt{122}}{63}} \approx \boxed{.1753}.$$

7.3 Proportional coverage

A third way insurance companies limit their exposure is by offering **proportional coverage**. This means that they only cover a fraction of the loss, as opposed to the entire loss. To compute quantities associated to proportional coverage, think of the original loss as X and the claim payment as Y. Write Y as a piecewise-defined function φ of X and answer the question asked (if the question asks for an expected value, variance or standard deviation, you have to do several integrals separately according to each piece of the function φ).

EXAMPLE 5

Suppose that the damage (in thousands of dollars) caused when a piece of equipment breaks is given by a continuous random variable with density $f(x) = \frac{2}{x^3}$ when $x > 1$. Suppose that the piece of equipment breaks 25% of the time. If an insurance company agrees to cover 100% of the first $\$3000$ in damage and 50% of the next $\$3000$ in damage, what is the expected value of the amount the insurance company will have to pay?

EXAMPLE 6

The cumulative distribution function for health care costs experienced by a policyholder is

$$F(x) = \begin{cases} 1 - e^{-x/100} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The policy has a deductible of 20. An insurer reimburses the policyholder for 100% of health care costs between 20 and 120 (less the deductible); health care costs from 120 to 420 are reimbursed at 50%; health care costs above 420 are not reimbursed. Find the cdf of the reimbursements.

7.4 Chapter 7 Homework

Exercises from Sections 7.1 and 7.2

1. Suppose that a loss random variable is uniform on $[0, 1000]$. Determine the expected amount paid by the insurer in each of the following cases:

   a) The policy has a deductible of 300.

   b) The policy has a coverage limit of 500.

   c) The policy has a deductible of 100 and a coverage limit of 600.

2. Calculate the variance of the amount paid by the insurer in situation (a) of the previous problem.

3. (AE) Suppose that a loss random variable is uniform on $[0, 1000]$. A deductible of size d is applied before any insurance payment. If the expected payment of the insurer is 150, find d.

4. (AE) Suppose that a loss random variable is exponential with mean 10. If a deductible of size 5 is applied, find the expected payment of the insurer.

5. (AE) An insurance policy pays for a random loss X subject to a deductible C, where $0 < C < 1$. The loss amount is modeled as a continuous r.v. whose density is $f(x) = 4x^3$ on $[0, 1]$. If the probability that the insurance payment is less than $.5$ is $.2401$, what is C?

6. Suppose that the loss from a hurricane is uniform on $[0, 8]$. A policyholder holds a hurricane insurance policy with coverage limit that is twice its deductible.

   a) If the deductible is 1, compute the expected insurance payment for a loss due to a hurricane.

   b) If the expected insurance payment for a loss due to a hurricane is $\frac{10}{7}$, compute the smallest possible value of the deductible.

   c) What deductible corresponds to the greatest expected insurance payment?

7. An insurance policy reimburses a loss up to a benefit limit of 15. The policyholder's loss follows a distribution with density function $(x) = \frac{2}{x^3}$ for $x > 1$.

   a) What is the probability that the benefit paid is less than 10?

   b) What is the expected value of the benefit paid under this policy?

Exercises from Section 7.3

8. Suppose the loss to a business from a fire, measured in thousands of dollars, is a r.v. X with density

$$f(x) = \begin{cases} \frac{1}{8}(4-x) & \text{for } 0 \le x \le 4 \\ 0 & \text{else} \end{cases}.$$

The business is covered by an insurance policy with no deductible that provides 50% coverage for the first $\$2000$ of damage and 25% coverage for the remaining damage.

a) Find the standard deviation of the amount the insurance policy will pay out, in the event of a loss.

b) Suppose that there is a 30% chance of one fire, and a 70% chance of no fire. Find the expected amount the insurance company will have to pay.

c) Suppose that the number of fires the business will suffer in one year is a Poisson r.v. with variance $\dfrac{1}{8}$. Assuming that the losses from each fire are independent, find the variance of the amount the insurance company will pay to the business.

9. A loss r.v. is uniform on $[0, 2]$. The loss is covered by a policy which has a deductible of 1. For losses exceeding the deductible, the policy provides proportional coverage covering $p\%$ of the loss. If the expected claim payment is $.2$, find the value of p.

10. A homeowner takes out a policy on his house. The policy has a deductible of 4 and reimburses the homeowner for 100% of the damage to the house for damage between 4 and 10. The policy only reimburses the homeowner for 60% of damage between 10 and 20, and reimburses the policyholder for 20% of damage above 20. If the damage to the house is a r.v. with density function

$$f(x) = \begin{cases} \dfrac{3}{125000}(x-50)^2 & \text{for } 0 \le x \le 50 \\ 0 & \text{else} \end{cases},$$

find the expected reimbursement.

11. (AE) The lifetime of a printer costing 100 is exponential with mean 2 years. The manufacturer agrees to pay a full refund to a buyer if the printer fails during the first year following its purchase, and a half refund if it fails during its second year. If the manufacturer sells 100 printers, how much should it expect to pay in refunds?

Chapter 8

Markov chains

8.1 What is a Markov chain?

In MATH 416, our primary goal is to describe probabilistic models which simulate real-world phenomena. As with all modeling problems, there is a "Goldilocks" issue:

- If the model is too simple,

- if the model is too complex,

In applied probability, we want to model phenomena which evolve randomly. The mathematical object which describes such a situation is a *stochastic process*:

Definition 8.1 *A* **stochastic process** $\{X_t : t \in \mathcal{I}\}$ *is a collection of random variables indexed by* t*. The set* $\mathcal{I}$ *of values of* t *is called the* **index set** *of the stochastic process, and members of* $\mathcal{I}$ *are called* **times***. We assume that each* X_t *has the same range, and we denote this common range by* $\mathcal{S}$*.* $\mathcal{S}$ *is called the* **state space** *of the process, and elements of* $\mathcal{S}$ *are called* **states***.*

Note: $\{X_t\}$ refers to the entire process (i.e. at all times t), whereas X_t is a single random variable (i.e. refers to the state of the process at a fixed time t).

Concept: Think of X_t as recording your "position" or "state" at time t. As t changes, you think of "moving" or "changing states". This process of "moving" will be random, and modeled using the language and theory of probability we learned in MATH 414.

Almost always, the index set is $\{0, 1, 2, 3, ...\}$ or $\mathbb{Z}$ (in which case we call the stochastic process a **discrete-time** process), or the index set is $[0, \infty)$ or $\mathbb{R}$ (in which case we call the stochastic process a **continuous-time** process). Chapter 8 of these notes focuses on discrete-time processes; Chapters 9 and 11 study continuous-time processes, and Chapter 10 contains ideas useful in both settings.

In MATH 414, we encountered three standard classes of stochastic processes:

1. The **Bernoulli process**, a discrete-time process $\{X_t\}$ with state space $\mathbb{N}$ where X_t is the number of successes in the first t trials of a Bernoulli experiment. Probabilities associated to a Bernoulli process are completely determined by a number $p \in (0, 1)$ which gives the probability of success on any one trial.

2. The **Poisson process**, a continuous-time process $\{X_t\}$ with state space $\mathbb{N}$ where X_t is the number of births in the first t units of time. Probabilities associated to a Poisson process are completely determined by a number $\lambda > 0$ called the **rate** of the process.

3. **i.i.d. processes** are discrete-time processes $\{X_t\}$ where each X_t has the same density and all the X_t are mutually independent. Sums and averages of random variables from these processes are approximately normal by the Central Limit Theorem.

We now define a class of processes which encompasses the three examples above and much more:

Definition 8.2 *Let $\{X_t\}$ be a stochastic process with state space $\mathcal{S}$. $\{X_t\}$ is said to have the* **Markov property** *if for any times $t_1 < t_2 < ... < t_n$ and any states $x_1, ..., x_n \in \mathcal{S}$,*

$$P(X_{t_n} = x_n \,|\, X_{t_1} = x_1, X_{t_2} = x_2, ..., X_{t_{n-1}} = x_{n-1}) = P(X_{t_n} = x_n \,|\, X_{t_{n-1}} = x_{n-1}).$$

A **Markov chain** *is a discrete-time stochastic process with finite or countable state space that has the Markov property.*

To understand this definition, think of time t_n as the "present" and the times $t_1 < ... < t_{n-1}$ as being times in the "past". If a process has the Markov property, then given some values of the process in the past, the probability of the present value of the process **depends only on the most recent given information**, i.e. only on $X_{t_{n-1}}$.

Note: Bernoulli processes, Poisson processes and i.i.d. processes all have the Markov property.

The three ingredients of a Markov chain

QUESTION

What are the "ingredients" of a Markov chain? In other words, what makes one Markov chain different from another one?

Answer:

1. The **state space** $\mathcal{S}$ of the Markov chain

   (Usually $\mathcal{S}$ is labelled $\{1, ..., d\}$ or $\{0, 1\}$ or $\{0, 1, 2, ...\}$ or $\mathbb{N}$ or $\mathbb{Z}$, etc.)

2. The **initial distribution** of the r.v. X_0, denoted π_0:

   $$\pi_0(x) = P(X_0 = x) \text{ for all } x \in \mathcal{S}$$

   $\pi_0(x)$ is the probability the chain starts in state x.

3. **Transition probabilities**, denoted $P(x, y)$ or $P_{x,y}$ or P_{xy}:

   $$P(x, y) = P_{xy} = P_{x,y} = P(X_t = y \,|\, X_{t-1} = x)$$

   $P(x, y)$ is the probability, given that you are in state x at a certain time $t - 1$, that you are in state y at the next time (which is time t).

   Technically, transition probabilities depend not only on x and y but on t, but throughout our study of Markov chains we will assume (often without stating it) that the transition probabilities do not depend on t; that is, that they have the following property:

Definition 8.3 *Let $\{X_t\}$ be a Markov chain. We say the transition probabilities of $\{X_t\}$ are* **time homogeneous** *if for all $s, t \in \mathcal{S}$,*

$$P(X_t = y \,|\, X_{t-1} = x) = P(X_s = y \,| X_{s-1} = x),$$

i.e. that the transition probabilities depend only on x and y (and not on t).

The reason the transition probabilities are sufficient to describe a Markov chain is that by the Markov property,

$$P(X_t = x_t \,|\, X_0 = x_0, ..., X_{t-1} = x_{t-1}) = P(X_t = x_t \,|\, X_{t-1} = x_{t-1}) = P(x_{t-1}, x_t).$$

In other words, conditional probabilities of this type **depend only on the most recent transition** and ignore any past behavior in the chain.

Simulating a Markov chain

To get used to how Markov chains work, let's simulate one using a computer. Let's suppose:

- the state space is $\mathcal{S} = \{1, 2, 3\}$;
- the initial distribution π_0 satisfies $\pi_0(1) = \frac{1}{2}$, $\pi_0(2) = \frac{1}{6}$ and $\pi_0(3) = \frac{1}{3}$. We shorthand this by writing π_0 as

$$\pi_0 = \left(\frac{1}{2}, \frac{1}{6}, \frac{1}{3}\right).$$

- the transition probabilities are $P(1,1) = \frac{1}{3}$, $P(1,2) = \frac{1}{2}$, $P(1,3) = \frac{1}{6}$, $P(2,1) = \frac{2}{3}$, $P(2,2) = 0$, $P(2,3) = \frac{1}{3}$, $P(3,1) = \frac{1}{6}$, $P(3,2) = \frac{1}{3}$, $P(3,3) = \frac{1}{2}$. A shorthand way of writing all these is by treating them as entries of a matrix:

$$P = \begin{pmatrix} \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \\ \frac{2}{3} & 0 & \frac{1}{3} \\ \frac{1}{6} & \frac{1}{3} & \frac{1}{2} \end{pmatrix}$$

We can capture the state space and the transition probabilities with the following picture:

To simulate this Markov chain, we first have to select the state X_0 in which we start. This is done using π_0: we pick state 1 with probability $\frac{1}{2}$, state 2 with probability $\frac{1}{6}$, and state 3 with probability $\frac{1}{3}$.

One way to perform this random choice on a computer is to have the computer generate a "uniformly random" real number in $[0, 1]$ (in *Mathematica*, you use the `RandomReal[ ]` command to do this). If the number is less than $\frac{1}{2}$, let $X_0 = 1$; if the number is between $\frac{1}{2}$ and $\frac{1}{2} + \frac{1}{6}$, let $X_0 = 2$; otherwise $X_0 = 3$:

Suppose we picked $X_0 = 3$. Now, since $X_0 = 3$, we pick the next state X_1 using Row 3 of P. By this, I mean $X_1 = 1$ with probability $\frac{1}{6}$, $X_1 = 2$ with probability $\frac{1}{3}$, and $X_1 = 3$ with probability $\frac{1}{2}$ (if you did this on a computer by selecting a random real number in $[0, 1]$, then X_1 would be determined as follows:

Let's suppose our random choice led to $X_1 = 1$. The next thing to do is to pick the state X_2, which is done by using Row 1 of P (so $X_2 = 1$ with probability $\frac{1}{3}$, etc.). The idea expressed in the Markov property is that so long as we know X_1, the fact X_0 was 3 is no longer relevant to the computation of X_2, i.e. that $X_0 = 3$ is "old news".

Similarly, once you've figured X_2, the fact that $X_1 = 1$ doesn't influence how X_3 is generated, etc.

To get the rest of the chain $\{X_t\}$, you pick each state X_t from the previous one X_{t-1}: if $X_{t-1} = j$, X_t is chosen using Row j of P as described above.

8.2 Basic examples of Markov chains

EXAMPLE 1: I.I.D PROCESS $\{X_t\}$ (OF DISCRETE R.V.S)

State space: $\mathcal{S} =$

Initial distribution:

Transition probabilities: $P(x, y) = P(X_t = y \,|\, X_{t-1} = x) =$

EXAMPLE 2: BERNOULLI PROCESS $\{X_t\}$

State space: $\mathcal{S} = \mathbb{N} = \{0, 1, 2, 3, ...\}$

Initial distribution:

Transition probabilities:

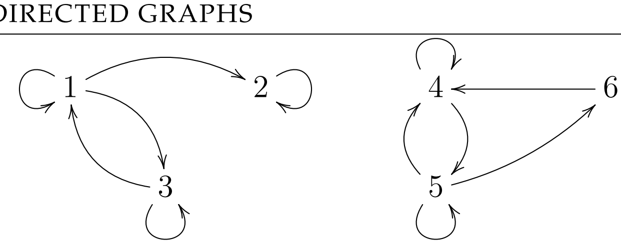
$$P(x, y) = P(X_t = y\,|X_{t-1} = x) = \left\{ \vphantom{\begin{array}{c} a \\ b \\ c \\ d \end{array}} \right.$$

We represent these transition probabilities with the following picture:

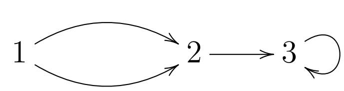
The above picture generalizes: Every Markov chain can be thought of as a random walk on a graph as follows:

Definition 8.4 *A* **directed graph** *is a finite or countable set of points called* **nodes**, *usually labelled by integers, together with "arrows" from one point to another, such that given two nodes x and y, there is either zero or one arrow going directly from x to y.*

EXAMPLES OF DIRECTED GRAPHS



NOT A DIRECTED GRAPH:



If one labels the arrow from x to y with a number $P(x, y)$ such that for each node x, $\sum\limits_{y} P(x, y) = 1$, then the directed graph represents the transition probabilities of a Markov chain, where the nodes are the states and the arrows represent the transitions. If you are in state x at time $t-1$ (i.e. if $X_{t-1} = x$), then to determine your state X_t at time t, you follow one of the arrows starting at x (with probabilities as indicated on the arrows which start at x).

EXAMPLE 3: BASIC URN MODEL

An urn initially holds 2 red and 2 green marbles. Every minute, you choose a marble uniformly from the urn. If you draw a red marble, you put the red marble back in the urn, and add two green marbles from the urn. If you draw a green marble, you leave it out of the urn. Let X_t be the number of green marbles in the jar after t draws.

EXAMPLE 4: SIMPLE UNBIASED RANDOM WALK

$\mathcal{S} = \mathbb{Z}$; $P(x, x+1) = P(x, x-1) = \frac{1}{2}$ for all $x \in \mathcal{S}$.

$$\begin{tikzcd}
\cdots -4 \arrow[r, bend left, "\frac{1}{2}"] & -3 \arrow[l, bend left, "\frac{1}{2}"] \arrow[r, bend left, "\frac{1}{2}"] & -2 \arrow[l, bend left, "\frac{1}{2}"] \arrow[r, bend left, "\frac{1}{2}"] & -1 \arrow[l, bend left, "\frac{1}{2}"] \arrow[r, bend left, "\frac{1}{2}"] & 0 \arrow[l, bend left, "\frac{1}{2}"] \arrow[r, bend left, "\frac{1}{2}"] & 1 \arrow[l, bend left, "\frac{1}{2}"] \arrow[r, bend left, "\frac{1}{2}"] & 2 \arrow[l, bend left, "\frac{1}{2}"] \arrow[r, bend left, "\frac{1}{2}"] & 3 \arrow[l, bend left, "\frac{1}{2}"] \arrow[r, bend left, "\frac{1}{2}"] & 4 \cdots \arrow[l, bend left, "\frac{1}{2}"]
\end{tikzcd}$$

EXAMPLE 5: GAMBLER'S RUIN

Make a series of $\$1$ bets in a casino, where you are 60% likely to win and 40% likely to lose each game. Let X_t be your bankroll after the t^{th} bet.

8.3 Matrix theory applied to Markov chains

Suppose $\{X_t\}$ is a Markov chain with **finite** state space $\mathcal{S} = \{1, ..., d\}$. Let $\pi_0 : \mathcal{S} \to [0, 1]$ give the initial distribution (i.e. $\pi_0(x) = P(X_0 = x)$) and let the transition probabilities be $P_{x,y}$ ($P_{x,y}$ is the same thing as $P(x, y)$).

So long as the state space is finite, the most convenient representation of the chain's transition probabilities is in a matrix:

Definition 8.5 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S} = \{1, ..., d\}$. The $d \times d$ matrix of transition probabilities*

$$P = \begin{pmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,d} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ P_{d,1} & P_{d,2} & \cdots & P_{d,d} \end{pmatrix}_{d \times d}$$

is called the **transition matrix** *of the Markov chain.*

WHY USE MATRICES?

We will see that we can answer almost any question about a finite state space Markov chain by performing some simple matrix algebra associated to the transition matrix P.

Stochastic matrices

A natural question to ask is exactly which matrices can be transition matrices of some Markov chain. Notice that all the entries of P must be nonnegative, and that the rows of P must sum to 1, since they represent the probabilities associated to all the places x can go in 1 unit of time.

Definition 8.6 *A $d \times d$ matrix of real numbers P is called a* **stochastic matrix** *if*

1. *P has only nonnegative entries, i.e. $P_{x,y} \geq 0$ for all $x, y \in \{1, ..., d\}$; and*

2. *each row of P sums to 1, i.e. for every $x \in \{1, ..., d\}$, $\sum_{y=1}^{d} P_{x,y} = 1$.*

Theorem 8.7 (Transition matrices are stochastic) *A $d \times d$ matrix of real numbers P is the transition matrix of a Markov chain if and only if it is a stochastic matrix.*

n-step transition probabilities

Definition 8.8 *Let $\{X_t\}$ be a Markov chain and let $x, y \in \mathcal{S}$. Define the n-***step transition probability** *from x to y by*

$$P^n(x,y) = P(X_{t+n} = y \,|\, X_t = x).$$

(Since we are assuming the transition probabilities are time homogeneous, these numbers will not depend on t.)

So $P^n(x,y)$ measures the probability, given that you are in state x, that you are in state y exactly n units of time from now.

Theorem 8.9 *Let $\{X_t\}$ be a Markov chain with finite state space $\mathcal{S} = \{1, ..., d\}$. If P is the transition matrix of $\{X_t\}$, then for every $x, y \in \mathcal{S}$ and every $n \in \{0, 1, 2, 3, ...\}$, we have*

$$P^n(x,y) = (P^n)_{x,y},$$

the $(x,y)-$entry of the matrix P^n.

PROOF I'm going to prove this only when $n = 2$ (the proof for general n uses a proof technique called "induction", for which $n = 2$ constitutes the base case). By time homogeneity,

$$P^2(x,y) = P(X_2 = y \,|\, X_0 = x)$$

Now, recall how matrix multiplication works:

Time n distributions

Definition 8.10 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. A* **distribution** *on $\mathcal{S}$ is a probability measure π on $(\mathcal{S}, 2^{\mathcal{S}})$, i.e. a function $\pi : \mathcal{S} \to [0,1]$ such that $\sum_{x \in \mathcal{S}} \pi(x) = 1$.*

The coordinates of any distribution must be nonnegative and sum to 1.

We denote distributions as *row* vectors, i.e. if $\mathcal{S} = \{1, 2, ..., d\}$ then

$$\pi = (\pi(1), \pi(2), ..., \pi(d)) = \begin{pmatrix} \pi(1) & \pi(2) & \cdots & \pi(d) \end{pmatrix}_{1 \times d}.$$

This is unusual, as normally one would represent a vector in $\mathbb{R}^d$ as a column matrix, but this convention makes upcoming formulas easier.

Definition 8.11 *Let $\{X_t\}$ be a Markov chain. The* **time n distribution** *of the Markov chain, denoted π_n, is the distribution π_n defined by*

$$\pi_n(x) = P(X_n = x).$$

So $\pi_n(x)$ gives the probability that at time n, you are in state x.

Theorem 8.12 *Let $\{X_t\}$ be a Markov chain with finite state space $\mathcal{S} = \{1, ..., d\}$. If*

$$\pi_0 = (\pi_0(1), \pi_0(2), ..., \pi_0(d))_{1 \times d}$$

is the initial distribution of $\{X_t\}$ (written as a $1 \times d$ row vector), and if P is the transition matrix of $\{X_t\}$, then for every $x, y \in \mathcal{S}$ and every $n \in \mathcal{I}$, we have

$$\pi_n(y) = (\pi_0 P^n)_y,$$

the y^{th}–entry of the $(1 \times d)$ row vector $\pi_0 P^n$.

PROOF This is a direct calculation:

$$\begin{aligned} \pi_n(y) = P(X_n = y) &= \sum_{x \in \mathcal{S}} P(X_n = y \,|\, X_0 = x) P(X_0 = x) \quad \text{(LTP)} \\ &= \sum_{x \in \mathcal{S}} (P^n)_{x,y} \, \pi_0(x) \quad \text{(Theorem 8.9)} \\ &= \sum_{x \in \mathcal{S}} \pi_0(x) \, (P^n)_{x,y} \\ &= [\pi_0 P^n]_y \quad \text{(def'n of matrix multiplication)} \quad \square \end{aligned}$$

EXAMPLE 6

Consider the Markov chain with state space $\{0, 1\}$ whose transition matrix is

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{pmatrix}$$

and whose initial distribution is uniform.

1. Sketch the directed graph representing this Markov chain.
2. Find the distribution of X_2.
3. Find $P(X_3 = 0)$.
4. Find $P(X_8 = 1 \mid X_7 = 0)$.
5. Find $P(X_7 = 0 \mid X_4 = 0)$.

Markov chains with infinite state space

Although the formulas for $n-$step transitions and time n distributions are motivated by those obtained earlier in this section, the big difference if $\mathcal{S}$ is infinite is that the transitions $P(x, y)$ **cannot be expressed in a matrix** (since the matrix would have to have infinitely many rows and columns). The proper notation is to use functions:

Definition 8.13 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$.*

1. *The* **transition function** *of the Markov chain is the function*

$$P : \mathcal{S} \times \mathcal{S} \to [0, 1] \text{ defined by } P(x, y) = P(X_t = y \mid X_{t-1} = x).$$

2. *The* **initial distribution** *of the Markov chain is the function*

$$\pi_0 : \mathcal{S} \to [0, 1] \text{ defined by } \pi_0(x) = P(X_0 = x).$$

3. *The* $n-$**step transition function** *of the Markov chain is the function P^n : $\mathcal{S} \times \mathcal{S} \to [0, 1]$ defined by*

$$P^n(x, y) = P(X_{t+n} = y \mid X_t = x).$$

4. *The* **time** n **distribution** *of the Markov chain is the function*

$$\pi_n : \mathcal{S} \to [0, 1] \text{ defined by } \pi_n(x) = P(X_n = x).$$

As with finite state spaces, the transition functions must be "stochastic":

Lemma 8.14 *$P : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is the transition function of a Markov chain with state space $\mathcal{S}$ if and only if*

1. *for every $x, y \in \mathcal{S}$, $P(x, y) \geq 0$, and*
2. *for every $x \in \mathcal{S}$, $\displaystyle\sum_{y \in \mathcal{S}} P(x, y) = 1$.*

Lemma 8.15 *If π_n is the time n distribution of a Markov chain with state space $\mathcal{S}$, then $\displaystyle\sum_{x \in \mathcal{S}} \pi_n(x) = 1$.*

Theorem 8.16 *Let $\{X_t\}$ be a Markov chain with transition function P and initial distribution π_0. Then:*

1. *For all $x_0, x_1, ..., x_n \in \mathcal{S}$,*

$$P(X_0 = x_0, X_1 = x_1, ..., X_n = x_n) = \pi_0(x_0) \prod_{j=1}^{n} P(x_{j-1}, x_j)$$

2. *For all $x, y \in \mathcal{S}$,*

$$P^n(x, y) = \sum_{z_1,...,z_{n-1} \in \mathcal{S}} P(x, z_1)P(z_1, z_2) \cdots P(z_{n-2}, z_{n-1})P(z_{n-1}, y)$$

3. *The time n distribution π_n satisfies, for all $y \in \mathcal{S}$,*

$$\pi_n(y) = \sum_{x \in \mathcal{S}} \pi_0(x)P^n(x, y).$$

8.4 The Fundamental Theorem of Markov chains

Many areas of mathematics have a central result which is key to understanding the ideas of the subject. These central results are called "Fundamental Theorems":

Fundamental Theorem of Arithmetic: every integer greater than 1 can be factored uniquely into a product of prime numbers.

Fundamental Theorem of Algebra: every polynomial whose coefficients are in $\mathbb{C}$ has a root in $\mathbb{C}$.

Fundamental Theorem of Calculus: if $f : \mathbb{R} \to \mathbb{R}$ is cts and $F(x) = \int_a^x f(t)\, dt$, then $F'(x) = f(x)$. (Also, if $f : \mathbb{R} \to \mathbb{R}$ is cts with antiderivative F, then $\int_a^b f(x)\, dx = F(b) - F(a)$.)

Fundamental Theorem of Line Integrals: If $\mathbf{f} = \nabla f$ is a conservative vector field on $\mathbb{R}^n$, then for any piecewise C^1 curve γ with initial point $\mathbf{a}$ and terminal point $\mathbf{b}$, $\int_\gamma \mathbf{f} \cdot d\mathbf{s} = f(\mathbf{b}) - f(\mathbf{a})$.

Fundamental Theorem of Linear Algebra: If $A \in M_{mn}(\mathbb{R})$, then $\dim C(A) = \dim R(A)$, $[R(A)]^\perp = N(A)$ and $[C(A)]^\perp = N(A^T)$.

This section is about the Fundamental Theorem of Markov Chains (FTMC). To get an idea of what this theorem is about, we'll do some experimentation.

What we (almost assuredly) saw in our experiment is that the Markov chain we invented had a special distribution π, so that as $n \to \infty$, the time n distributions π_n approached this distribution π, no matter what the initial distribution was. The FTMC says that for most (not all) Markov chains, this phenomenon holds:

Theorem 8.17 (Fundamental Theorem of Markov Chains (FTMC)) *Let $\{X_t\}$ be an irreducible, aperiodic, positive recurrent Markov chain. Then $\{X_t\}$ has a unique stationary distribution π, such that π is steady-state, meaning*

$$\lim_{n\to\infty} \pi_n(x) = \pi(x)$$

for all $x \in \mathcal{S}$, no matter the initial distribution π_0.

To understand this theorem, we need to learn the meaning of its vocabulary: *irreducible*, *aperiodic*, *positive recurrent*, *stationary*, *steady-state*. Learning this vocabulary is the goal of the next four sections.

8.5 Stationary and steady-state distributions

RECALL

A Markov chain is determined by three things:

-
-
-

From this, you get time n distributions π_n which give the probability of each state at time n:

$$\pi_n(y) = P(X_n = y) = \sum_{x\in\mathcal{S}} \pi_{n-1}(x)P(x,y) = \sum_{x\in\mathcal{S}} \pi_0(x)P^n(x,y)$$

(i.e. $\pi_n = \pi_0 P^n$ if $\mathcal{S}$ is finite and P is the transition matrix)

We are investigating the FTMC, which says that if $\{X_t\}$ is "irreducible", "aperiodic" and "positive recurrent", then there is a "stationary, steady-state" distribution π such that $\pi_n(x)$ approaches $\pi(x)$ for all $x \in \mathcal{S}$. This upshot of the FTMC is that for large n, $\pi_n(x)$ can be approximated by $\pi(x)$.

QUESTION

What do “stationary” and “steady-state” mean?

Stationary distributions

Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. Suppose π is a distribution on $\mathcal{S}$ so that, if the initial distribution π_0 is π, the time 1 distribution π_1 is also π. Then π is called “stationary” (because it didn’t change as time passed). More precisely:

Definition 8.18 *Let $\{X_t\}$ be a Markov chain. A distribution π on $\mathcal{S}$ is called* **stationary** *(with respect to $\{X_t\}$) if for all $y \in \mathcal{S}$,*

$$\sum_{x \in \mathcal{S}} \pi(x)P(x,y) = \pi(y).$$

If $\mathcal{S}$ is finite (say $\mathcal{S} = \{1, 2, 3, ..., d\}$, to say π is stationary means (in matrix multiplication terminology)

$$\pi P = \pi$$

if we write $\pi = \begin{pmatrix} \pi(1) & \pi(2) & \cdots & \pi(d) \end{pmatrix}_{1 \times d}$.

Lemma 8.19 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. If π is a stationary distribution, then for all $n > 0$ and all $y \in \mathcal{S}$, we have*

$$\pi(y) = \sum_{x \in \mathcal{S}} \pi(x)P^n(x,y).$$

(So if $\mathcal{S}$ is finite, this means $\pi = \pi P^n$ for all n.)

PROOF Definition of “stationary” + induction on n.

Lemma 8.20 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. An initial distribution π_0 is stationary if and only if the time n distributions are the same for every n.*

PROOF ($\Rightarrow$) Assume π_0 is stationary.

Then, applying the previous lemma to any $y \in \mathcal{S}$, we see

$$\pi_n(y) = \sum_{x \in \mathcal{S}} \pi_0(x)P^n(x,y) = \pi_0(y)$$

so $\pi_n = \pi_0$ as wanted.

($\Leftarrow$) Assume the time n distributions are the same for every n.

In particular, this means $\pi_1 = \pi_0$, meaning

$$\pi_0(y) = \pi_1(y) = \sum_{x \in \mathcal{S}} \pi_0(x) P(x, y).$$

By definition, π_0 is stationary. $\square$

Put another way, this lemma says that ***stationary distributions are those which do not change as time passes***.

Steady-state distributions

A steady-state distribution for a Markov chain is like the special one in our experiment: if π is steady-state for $\{X_t\}$, then no matter the initial distribution π_0, $\pi_n(x) \to \pi(x)$ as $n \to \infty$, so for large n, $\pi_n(x) \approx \pi(x)$. More precisely:

Definition 8.21 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. A distribution π on $\mathcal{S}$ is called* **steady-state** *(with respect to $\{X_t\}$) if*

$$\lim_{n \to \infty} P^n(x, y) = \pi(y) \text{ for all } x, y \in \mathcal{S}.$$

Theorem 8.22 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. Suppose π is a steady-state distribution for $\{X_t\}$. Then for any initial distribution π_0,*

$$\lim_{n \to \infty} \pi_n(y) = \lim_{n \to \infty} P(X_n = y) = \pi(y)\ \forall\, y \in \mathcal{S}.$$

PROOF By Theorem 8.16 (3), we get the top equation below. Then take the limit of both sides as $n \to \infty$:

$$\begin{tikzcd}
\pi_n(y) \arrow[d, "n\to\infty"'] & = P(X_n = y) & = \sum_{x \in \mathcal{S}} \pi_0(x) P^n(x, y) \arrow[d, "n\to\infty"] \\
\lim_{n\to\infty} \pi_n(y) & & \sum_{x \in \mathcal{S}} \pi_0(x) \pi(y)
\end{tikzcd}$$

So ***steady-state distributions are those which "attract" the time n distribution as n increases, no matter the initial distribution***.

EXAMPLE 7

Let $p, q \in (0,1)$ (there is no relationship between p and q). Consider a Markov chain with $\mathcal{S} = \{0, 1\}$ whose transition matrix is

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

Find all stationary distributions of this Markov chain (there might not be any).

IN GENERAL

You find stationary distributions for finite state-space Markov chains by solving a system of linear equations corresponding to $\pi P = \pi$ as in Example 6.

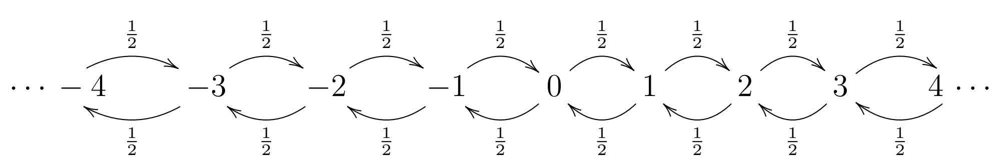
EXAMPLE 8

Find all stationary distributions of $\{X_t\}$, if $\{X_t\}$ has transition matrix

$$\begin{pmatrix} \frac{1}{7} & \frac{4}{7} & \frac{2}{7} \\ 0 & \frac{5}{7} & \frac{2}{7} \\ \frac{3}{7} & \frac{1}{7} & \frac{3}{7} \end{pmatrix}.$$

EXAMPLE 9

Let $\{X_t\}$ be simple, unbiased random walk on $\mathbb{Z}$ (this means $\mathcal{S} = \mathbb{Z}$, and for every $x \in \mathcal{S}$, $P(x, x+1) = P(x, x-1) = \frac{1}{2}$). Find all stationary distributions of $\{X_t\}$.



Number of stationary and steady-state distributions

BIG PICTURE QUESTIONS

Let $\{X_t\}$ be a Markov chain.

1. Does $\{X_t\}$ have a stationary distribution?
2. If so, how many stationary distributions does $\{X_t\}$ have?
3. Does $\{X_t\}$ have a steady-state distribution?
4. If so, how many steady-state distributions does $\{X_t\}$ have?

In the rest of this section, we are going to run through some theorems addressing these questions. We'll start with ideas related to Question 2 above.

Definition 8.23 *Suppose $\pi_1, \pi_2, \pi_3, ...$ are all distributions on a set $\mathcal{S}$ (there could be finitely or countably many distributions). A* **convex combination** *of these distributions is another distribution of the form*

$$\sum_j \alpha_j \pi_j$$

where the α_j are nonnegative numbers satisfying $\sum_j \alpha_j = 1$.

EXAMPLE

Let $\pi_1 = (.1, .5, .4)$, $\pi_2 = (0, 1, 0)$ and $\pi_3 = (.7, .2, .1)$. The distribution

$$\begin{aligned} .5\pi_1 + .3\pi_2 + .2\pi_3 &= .5(.1, .5, .4) + .3(0, 1, 0) + .2(.7, .2.1) \\ &= (.05, .25, .2) + (0, .3, 0) + (.14, .04, .02) \\ &= (.19, .59, .22) \end{aligned}$$

is a convex combination of π_1, π_2 and π_3 with $\alpha_1 = .5, \alpha_2 = .3$ and $\alpha_3 = .2$.

Lemma 8.24 *Any convex combination of distributions is a distribution.*

PROOF If

$$\pi = \sum_j \alpha_j \pi_j,$$

then

$$\sum_{x \in \mathcal{S}} \pi(x) = \sum_{x \in \mathcal{S}} \sum_j \alpha_j \pi_j(x) = \sum_j \alpha_j \sum_{x \in \mathcal{S}} \pi_j(x) = \sum_j \alpha_j \cdot 1 = 1.$$

Since all the α_j are nonnegative, we see $\pi(x) \geq 0$ for all x as well. Therefore π is a distribution. □

SPECIAL CASE

A convex combination of two distributions π_1 and π_2 is any distribution

$$\alpha\pi_1 + (1 - \alpha)\pi_2$$

where $\alpha \in [0, 1]$.

Theorem 8.25 (Any convex comb. of stat. distributions is stationary) *Suppose $\pi_1, \pi_2, \pi_3, \ldots$ are all stationary distributions for a Markov chain $\{X_t\}$. Then any convex combination of the π_j is also a stationary distribution for $\{X_t\}$.*

PROOF HW (you have to check that the stationarity equation $\pi(y) = \sum_{x \in \mathcal{S}} \pi(x) P(x, y)$ holds for the convex combination)

Corollary 8.26 (Number of stationary distributions) *A Markov chain must have either zero, one, or infinitely many stationary distributions.*

PROOF Suppose the Markov chain has two different stationary distributions, say π_1 and π_2. Then for any $\alpha \in [0, 1]$,

$$\alpha\pi_1 + (1 - \alpha)\pi_2$$

is also a stationary distribution. Since there are infinitely many choices for $\alpha \in [0, 1]$, the Markov chain will have infinitely many stationary distributions. □

Now we turn to Big Picture Question 4 from earlier (how many steady-state distributions can a Markov chain have?)

Theorem 8.27 (Uniqueness of steady-state distributions) *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. If the Markov chain has a steady-state distribution π, then*

1. *π is stationary for $\{X_t\}$; and*
2. *π is the only stationary distribution for $\{X_t\}$.*

PROOF OF STATEMENT (2) We prove this by contradiction.

Suppose π_{dif} is a stationary distribution, different from π.

That means there is $y \in \mathcal{S}$ such that $\pi_{dif}(y) \neq \pi(y)$.

Now, use π_{dif} as the initial distribution of the chain; by stationarity the time n distribution of state y is $[\pi_{dif}]_n(y) = \pi_{dif}(y)$. Thus

$$\lim_{n\to\infty}[\pi_{dif}]_n(y) = \lim_{n\to\infty}\pi_{dif}(y) = \pi_{dif}(y) \neq \pi(y).$$

This contradicts π being steady-state, completing the proof of (2). $\square$

PROOF OF STATEMENT (1) WHEN $\mathcal{S}$ IS FINITE

Suppose π is steady-state and let $y \in \mathcal{S}$. Then

$$\begin{aligned}
\pi(y) &= \lim_{n\to\infty} P^n(x,y) \quad \text{(by definition of steady-state)} \\
&= \lim_{n\to\infty} P^{n+1}(x,y) \\
&= \lim_{n\to\infty} \sum_{z\in\mathcal{S}} P^n(x,z)P(z,y) \quad \text{(LTP)} \\
&= \sum_{z\in\mathcal{S}} \lim_{n\to\infty} P^n(x,z)P(z,y) \\
&= \sum_{z\in\mathcal{S}} \pi(z)P(z,y) \quad \text{(by definition of steady-state).}
\end{aligned}$$

Since $\pi(y) = \sum_{z\in\mathcal{S}} \pi(z)P(z,y)$, π is stationary by definition.

QUESTION

Why isn't this proof valid if $\mathcal{S}$ is infinite?

The perils of interchanging limits and infinite sums

In the argument on the previous page, we had the expression

$$\lim_{n\to\infty} \sum_{z\in\mathcal{S}} P^n(x,z)P(z,y)$$

which is a specific case of a more general expression of the form

$$\lim_{n\to\infty} \sum_{z} a(z)b_n(z).$$

We'd like to interchange the limit and sum operations here (i.e. move the limit after the sum), and we can always do this if the sum is a finite sum, but if the sum is infinite, this may not be legal:

AN EXAMPLE WHERE INTERCHANGE OF LIMIT AND INFINITE SERIES FAILS

Suppose $z \in \{1,2,3,...\}$, $b_n(z) = \frac{z^2}{n^2}$ and $a(z) = \frac{1}{z^2}$. Then

$$\lim_{n\to\infty} \sum_{z=1}^{\infty} a(z)b_n(z) = \lim_{n\to\infty} \sum_{z=1}^{\infty} \frac{1}{z^2}\left(\frac{z^2}{n^2}\right) = \lim_{n\to\infty} \sum_{z=1}^{\infty} \frac{1}{n^2} = \lim_{n\to\infty} \infty = \boxed{\infty}$$

but if we interchange the limit and the sum, we get

$$\sum_{z=1}^{\infty} a(z) \lim_{n\to\infty} b_n(z) = \sum_{z=1}^{\infty} \frac{1}{z^2} \lim_{n\to\infty} \frac{z^2}{n^2} = \sum_{z=1}^{\infty} \frac{1}{z^2}(0) = \sum_{z=1}^{\infty} 0 = \boxed{0}.$$

In proofs like this, there are two ways to get around this problem. The first is to use what is called an *argument by exhaustion*, where we consider a finite subset of the zs being added and write down an inequality like this (assuming all the $b_n(z) \geq 0$):

$$\lim_{n\to\infty} \sum_{z=1}^{\infty} a(z)b_n(z) \geq \lim_{n\to\infty} \sum_{z=1}^{N} a(z)b_n(z) = \sum_{z=1}^{N} a(z) \lim_{n\to\infty} b_n(z).$$

Since this inequality holds for all N, we can then take a limit of both sides of this as $N \to \infty$ to get

$$\lim_{n\to\infty} \sum_{z=1}^{\infty} a(z)b_n(z) \geq \sum_{z=1}^{\infty} a(z) \lim_{n\to\infty} b_n(z).$$

In other words, $\boxed{\textit{if } b_n(z) \geq 0\textit{, the limit of an infinite sum} \geq \textit{the infinite sum of the limits}}$.

To complete an argument by exhaustion, you argue separately (somehow) that

$$\lim_{n\to\infty} \sum_{z=1}^{\infty} a(z)b_n(z) \leq \sum_{z=1}^{\infty} a(z) \lim_{n\to\infty} b_n(z);$$

this argument depends on the particular $a(z)$ and $b_n(z)$ under consideration.

PROOF OF STATEMENT (1) WHEN $\mathcal{S}$ IS INFINITE

Suppose π is steady-state. We'll prove π is stationary by establishing two claims:

Claim 1: For every $y \in \mathcal{S}$, $\pi(y) \geq \sum_{z\in\mathcal{S}} \pi(z)P(z,y)$.

To prove this, observe

$$\begin{aligned}
\pi(y) &= \lim_{n\to\infty} P^{n+1}(x,y) && \text{(since } \pi \text{ is steady-state)}\\
&= \lim_{n\to\infty} \sum_{z\in\mathcal{S}} P^n(x,z)P(z,y)\\
&\geq \sum_{z\in\mathcal{S}} \lim_{n\to\infty} P^n(x,z)P(z,y) && \text{(limit of inf. series} \geq \text{inf. series of limit)}\\
&= \sum_{z\in\mathcal{S}} \pi(z)P(z,y) && \text{(since } \pi \text{ is steady-state).}
\end{aligned}$$

Claim 2: For every $y \in \mathcal{S}$, $\pi(y) \leq \sum_{z\in\mathcal{S}} \pi(z)P(z,y)$.

To prove this, suppose not, i.e. $\exists\, y \in \mathcal{S}$ where $\pi(y) > \sum_{z\in\mathcal{S}} \pi(z)P(z,y)$.

This would mean that

$$\begin{aligned}
1 = \sum_{y\in\mathcal{S}} \pi(y) > \sum_{y\in\mathcal{S}}\sum_{z\in\mathcal{S}} \pi(z)P(z,y) &= \sum_{z\in\mathcal{S}}\sum_{y\in\mathcal{S}} \pi(z)P(z,y)\\
&= \sum_{z\in\mathcal{S}} \pi(z) \sum_{y\in\mathcal{S}} P(z,y)\\
&= \sum_{z\in\mathcal{S}} \pi(z)\cdot 1\\
&= 1
\end{aligned}$$

This is a contradiction ($1 > 1$ is false), so Claim 2 is true.

Claims 1 and 2 tell us that for every $y \in \mathcal{S}$, $\pi(y) = \sum_{z\in\mathcal{S}} \pi(z)P(z,y)$, i.e. that π is stationary, as wanted. This completes the proof of Theorem 8.27. $\square$

EXAMPLE 7, REVISITED

For the Markov chain whose state space is $\mathcal{S} = \{0,1\}$ and whose transition matrix is $P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$, we saw that the stationary distribution was

$$\pi = \left(\frac{q}{p+q}, \frac{p}{p+q}\right).$$

Is this distribution steady-state?

Solution: π is steady-state if $\lim_{n\to\infty} P^n(x,y) = \pi(y)$ for all $x, y \in \mathcal{S}$, i.e.

$$\lim_{n\to\infty} P^n = \begin{pmatrix} \frac{q}{p+q} & \frac{p}{p+q} \\ \frac{q}{p+q} & \frac{p}{p+q} \end{pmatrix}.$$

Q: How might we compute powers P^n of the matrix P?

A:

If you did all that for this matrix P, you'd find

$$\lambda = 1 \leftrightarrow (1,1) \qquad \lambda = 1-p-q \leftrightarrow (-p,q)$$

so

$$\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & 1-p-q \end{pmatrix} \qquad S = \begin{pmatrix} 1 & -p \\ 1 & q \end{pmatrix}$$

and therefore (after some calculation)

$$P^n = S\Lambda^n S^{-1} = \begin{pmatrix} \frac{q}{p+q} + \frac{p}{p+q}(1-p-q)^n & \frac{p}{p+q} - \frac{q}{p+q}(1-p-q)^n \\ \frac{q}{p+q} - \frac{p}{p+q}(1-p-q)^n & \frac{p}{p+q} + \frac{q}{p+q}(1-p-q)^n \end{pmatrix}.$$

Since $-1 < 1-p-q < 1$, $\lim_{n\to\infty} P^n = \begin{pmatrix} \frac{q}{p+q} & \frac{p}{p+q} \\ \frac{q}{p+q} & \frac{p}{p+q} \end{pmatrix}$, so $\lim_{n\to\infty} P^n(x,y) = \pi(y)$ and $\pi = \left(\frac{q}{p+q}, \frac{p}{p+q}\right)$ is indeed steady-state.

REMARK

There will be a better (less computationally intense) way of concluding π is steady-state, based on theory we will develop in this chapter.

8.6 Class structure and periodicity

What this section is about: The FTMC says that if $\{X_t\}$ is irreducible, aperiodic and positive recurrent, then it has a steady-state distribution.

In this section, we discuss what is meant by "irreducible" and "aperiodic".

To get started, we need to establish some notation that we'll use frequently.

Definition 8.28 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$.*

1. *Given an event E, define $P_x(E) = P(E \,|\, X_0 = x)$. This is the probability of event E, given that you start at x.*
2. *Given a r.v. Z, define $E_x(Z) = E(Z \,|\, X_0 = x)$. This is the expected value of Z, given that you start at x.*

Definition 8.29 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$.*

1. *Given a set $A \subseteq \mathcal{S}$, let T_A be the r.v. defined by*

$$T_A = \min\{t \geq 1 : X_t \in A\}.$$

   (We set $T_A = \infty$ if $X_t \notin A$ for all t.) T_A is called the **hitting time** *or* **first passage time** *to A.*

2. *Given a state $a \in \mathcal{S}$, denote by T_a the r.v. $T_{\{a\}}$.*

Note: $T_A : \Omega \to \{1, 2, 3, ...\} \bigcup \{\infty\}$, so $\sum_{n=1}^{\infty} P(T_A = n) = 1 - P(T_A = \infty) \leq 1$.

Class structure

Definition 8.30 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$.*

1. *For each $x, y \in \mathcal{S}$, define $f_{x,y} = P_x(T_y < \infty)$. This is the probability you get from x to y in some finite (positive) time.*
2. *We say x* **leads to** *y (and write $x \to y$) if $f_{x,y} > 0$. This means that if you start at x, there is some positive probability that you will eventually hit y.*
3. *We say x and y* **communicate** *(and write $x \leftrightarrow y$) if $x \to y$ and $y \to x$.*

Definition 8.31 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$, and let $C \subseteq \mathcal{S}$.*

1. *C is called* **closed** *if for every $x \in C$, if $x \to y$, then y must also be in C.*
2. *C is called a* **communicating class** *if C is closed and all members of C communicate.*
3. *$\{X_t\}$ is called* **irreducible** *if $\mathcal{S}$ is a communicating class.*

What these definitions mean:

- *closed* sets are those which are like the Hotel California: "you can never leave".
- A set of states is a *communicating class* if you never leave, and you can get from anywhere to anywhere within the class.
- A Markov chain is *irreducible* if you can get from any state to any other state.

Remark: whether or not a Markov chain is irreducible depends only on its transition probabilities, and not on its initial distribution.

EXAMPLE 10

Let $\{X_t\}$ be a Markov chain with state space $\{1, 2, 3, 4, 5, 6\}$ whose transition matrix is

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & 0 & \frac{1}{8} \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}$$

Find all closed sets and all communicating classes of $\{X_t\}$.

OBSERVATION

To solve Example 9, the only thing relevant is whether the entries of P are zero or nonzero. So long as an entry is nonzero, whether it is $\frac{1}{2}$ or $\frac{1}{4}$ or whatever doesn't affect the closed sets and communicating classes of $\{X_t\}$.

EXAMPLE 11

Each matrix below is the transition matrix of a Markov chain with state space $\{1, 2, 3, 4\}$. The "$+$" in the matrices represent arbitrary positive numbers. For each Markov chain, find all its communicating classes and determine if the chain is irreducible.

$$\begin{pmatrix} + & 0 & + & 0 \\ 0 & + & + & 0 \\ 0 & + & + & + \\ + & 0 & 0 & + \end{pmatrix} \qquad \begin{pmatrix} + & 0 & + & 0 \\ 0 & + & + & 0 \\ + & 0 & + & 0 \\ + & + & 0 & + \end{pmatrix}$$

Lemma 8.32 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. Then*

$$x \to y \iff P^n(x, y) > 0 \text{ for some } n \geq 1.$$

PROOF ($\Rightarrow$) Assume $x \to y$, i.e. $f_{x,y} = P_x(T_y < \infty) > 0$.

Notice

$$P_x(T_y < \infty) = \sum_{n=1}^{\infty} P_x(T_y = n),$$

so if this sum is > 0, at least one of its terms must be > 0, meaning there must be at least one N such that $P_x(T_y = N) > 0$.

Since $P^N(x, y) \geq P_x(T_y = N)$, we can conclude $P^N(x, y) > 0$ as wanted.

($\Leftarrow$) Suppose $P^n(x, y) > 0$ for one or more n. Take the smallest such n; for this n, we have

$$P_x(T_y = n) = P^n(x, y) > 0.$$

Therefore

$$f_{x,y} = P_x(T_y < \infty) \geq P_x(T_y = n) > 0,$$

so $x \to y$ as wanted. $\square$

Lemma 8.33 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. Then*

$$(x \to y \text{ and } y \to z) \Rightarrow x \to z.$$

PROOF Apply Lemma 8.32 twice:

$$x \to y \Rightarrow \text{ there exists } n_1 \text{ such that } P^{n_1}(x,y) > 0.$$

$$y \to z \Rightarrow \text{ there exists } n_2 \text{ such that } P^{n_2}(y,z) > 0.$$

Thus

$$P^{n_1+n_2}(x,z) \geq P^{n_1}(x,y)P^{n_2}(y,z) > 0,$$

so by Lemma 8.32 again, $x \to z$. □

RECALL

One of the necessary ingredients in the FTMC is that the chain is irreducible. In the next example, we see why irreducibility is important to ensuring the existence of a steady-state distribution.

EXAMPLE 12

Suppose $\{X_t\}$ is a Markov chain with state space $\{0, 1\}$ whose transition matrix is the 2×2 identity matrix ($P = I$).

1. Sketch the directed graph of this Markov chain, and find its communicating classes. Is $\{X_t\}$ irreducible?
2. Find all stationary distributions of this Markov chain.
3. Does $\{X_t\}$ have a steady-state distribution? Explain.

Periodicity

To explain the concept of periodicity, let's start with this simple example, which illustrates why "aperiodicity" is important in the FTMC:

EXAMPLE 13

Suppose $\{X_t\}$ is a Markov chain with state space $\{0,1\}$ whose transition matrix is

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

1. Sketch the directed graph of this Markov chain, and find its communicating classes. Is $\{X_t\}$ irreducible?
2. Find all stationary distributions of this Markov chain.
3. Suppose $\pi_0 = (1,0)$. Compute π_n for every n. Does $\lim\limits_{n\to\infty} \pi_n$ exist?
4. Does $\{X_t\}$ have a steady-state distribution? Explain.

The problem with the Markov chain in Example 13 (i.e. what causes its stationary distribution to not be steady-state) is that it is "periodic"... if you start in a certain state, you can only return to that state at times that are a multiple of 2. This means the chain has period 2. More generally:

Definition 8.34 *Let a and b be integers. We say a* **divides** *b (and write $a|b$) if b is a multiple of a. The* **greatest common divisor** *of a set E of integers, denoted $\gcd E$, is the largest integer dividing every number in that set.*

EXAMPLES

$$6 \,|\, 42 \qquad 5 \not| \, 42 \qquad 3 \,|\, 180$$

$$\gcd\{12,36\} = 12 \qquad \gcd\{18,27,15\} = 3 \qquad \gcd\{8,17\} = 1$$

Definition 8.35 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. Let $x \in \mathcal{S}$ be such that $f_x > 0$ (equivalently, $P^n(x,x) > 0$ for some $n \geq 1$; equivalently, $x \to x$). The* **period** *of x, denoted by d_x, is the largest integer which divides every n for which $P^n(x,x) > 0$. More formally,*

$$d_x = \gcd\{n : P^n(x,x) > 0\}.$$

Note: If $P(x,x) > 0$, then $d_x | 1$, so $d_x = 1$.

EXAMPLE 14

Let $\{X_t\}$ be simple, unbiased random walk on $\mathbb{Z}$:

$$\begin{tikzcd}
\cdots -4 \arrow[r, bend left, "\frac{1}{2}"] & -3 \arrow[l, bend left, "\frac{1}{2}"] \arrow[r, bend left, "\frac{1}{2}"] & -2 \arrow[l, bend left, "\frac{1}{2}"] \arrow[r, bend left, "\frac{1}{2}"] & -1 \arrow[l, bend left, "\frac{1}{2}"] \arrow[r, bend left, "\frac{1}{2}"] & 0 \arrow[l, bend left, "\frac{1}{2}"] \arrow[r, bend left, "\frac{1}{2}"] & 1 \arrow[l, bend left, "\frac{1}{2}"] \arrow[r, bend left, "\frac{1}{2}"] & 2 \arrow[l, bend left, "\frac{1}{2}"] \arrow[r, bend left, "\frac{1}{2}"] & 3 \arrow[l, bend left, "\frac{1}{2}"] \arrow[r, bend left, "\frac{1}{2}"] & 4 \cdots \arrow[l, bend left, "\frac{1}{2}"]
\end{tikzcd}$$

Determine the period of each state.

Theorem 8.36 (Communicating states have the same period) *Suppose $\{X_t\}$ is a Markov chain with state space $\mathcal{S}$. Let $x, y \in \mathcal{S}$ be such that $x \leftrightarrow y$. Then $d_x = d_y$.*

PROOF Suppose states x and y lead to one another. Then:

$$\begin{aligned} x \to y \ &\Rightarrow\ \exists\, n_1 \text{ s.t. } P^{n_1}(x,y) > 0 \\ y \to x \ &\Rightarrow\ \exists\, n_2 \text{ s.t. } P^{n_2}(y,x) > 0. \end{aligned}$$

(P.S. $\exists$ is short for "there exists".)

Therefore

$$P^{n_1+n_2}(x,x) \geq P^{n_1}(x,y)P^{n_2}(y,x) > 0 \ \Rightarrow\ d_x \,|\, (n_1 + n_2).$$

Now, take any n such that $P^n(y,y) > 0$. Then

$$P^{n_1+n+n_2}(x,x) \geq P^{n_1}(x,y)P^{n}(y,y)P^{n_2}(y,x) > 0 \Rightarrow d_x \,|\, (n_1 + n + n_2).$$

Notice that if d_x divides both $n_1 + n_2$ and $n_1 + n + n_2$, then d_x divides the difference, so $d_x \,|\, n$.

We have shown d_x divides any n such that $P^n(y,y) > 0$, meaning d_x is a common divisor of these n.

A symmetric argument (reversing the roles of x and y) shows $d_y \leq d_x$.
So $d_x = d_y$, as wanted. □

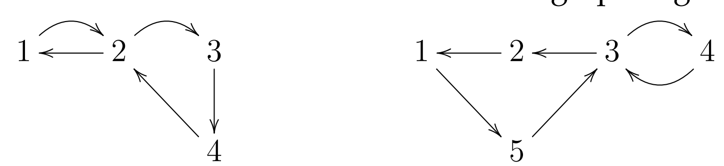
Theorem 8.36 shows that period is a **class property**, meaning that it is a property shared by all members of a communicating class. This implies:

Corollary 8.37 *If $\{X_t\}$ is an irreducible Markov chain, all states have the same period.*

Definition 8.38 *An irreducible Markov chain with state space $\mathcal{S}$ is called* **aperiodic** *if $d_x = 1$ for all $x \in \mathcal{S}$ and is called* **periodic with period** *d if $d_x = d > 1$ for all $x \in \mathcal{S}$.*

EXAMPLE 15

Find the period of each Markov chain whose directed graph is given below.



One important consequence of aperiodicity is that in an irreducible, aperiodic Markov chain, for every pair of states you can get from one to the other in any sufficiently large amount of time. This is made precise in Theorem 8.39:

Theorem 8.39 *Suppose $\{X_t\}$ is an irreducible, aperiodic Markov chain. Then, for every $x, y \in \mathcal{S}$, there is a number N such that $P^n(x, y) > 0$ for all $n \geq N$.*

PROOF For each $z \in \mathcal{S}$, let $I_z \subseteq \mathbb{N}$ be defined by $I_z = \{n : P^n(z, z) > 0\}$.
This means that I_z is the set of times that you can get from state z back to itself.
Notice that I_z is closed under addition: if $t, u \in I_z$, then $t + u \in I_z$.
Therefore if $t \in I_z$, then for any $k \geq 1$, $kt = t + t + t + \cdots + t \in I_z$ as well.
We know $1 = d = \gcd I_z$.

Claim 1: There is a number n_1 such that $n_1 \in I$ and $n_1 + 1 \in I_z$.
To prove this, suppose not. That means there is an integer $g \geq 2$ which is the smallest gap between two consecutive numbers in I_z.
Since g is a gap between numbers in I, we can select t so that $t \in I_z$ and $t + g \in I_z$.
But, since $\{X_t\}$ is aperiodic and $g \geq 2$, g is <u>not</u> the period of $\{X_t\}$, so there must be a time $u \in I_z$ which g does not divide.

Divide this u by g to get $u = gq + r$ where the remainder $r \in \{1, 2, ..., g-1\}$.
At this point, we know that since I_z is closed under addition and multiplication by scalars,

$$(q+1)(t+g) \in I_z \text{ and } u + (q+1)t \in I_z.$$

But the gap between these two numbers is

$$\begin{aligned}(q+1)(t+g) - [u + (q+1)t] &= (q+1)g - u = qg + g - gq - r \\ &= g - r < g,\end{aligned}$$

contradicting g being the smallest gap between numbers in I_z.

Claim 2: There is a number $n_2 = n_2(z)$ so that if $n \geq n_2$, $n \in I_z$.
To prove this, let $n_2 = n_1^2$, where n_1 is as in Claim 1 ($n_1 \in I_z$, $n_1 + 1 \in I_z$).
Now suppose $n \geq n_2$.
Divide $n - n_2$ by n_1 to get

$$n - n_2 = n_1 q + r \qquad (*)$$

where the remainder $r \in \{0, 1, 2, ..., n_1 - 1\}$.
Rewriting $(*)$, we get

$$\begin{aligned} n - n_2 &= n_1 q + r \\ n &= n_1 q + r + n_2 \\ n &= n_1 q + r + n_1^2 \\ n &= n_1 q + r + r n_1 - r n_1 + n_1^2 \\ n &= r(n_1 + 1) - (n_1 - r + q) n_1 \end{aligned}$$

and therefore $n \in I_z$ since $n_1 \in I_z$ and $n_1 + 1 \in I_z$.

Now, fix $x, y \in \mathcal{S}$. By irreducibility, $x \to y$, so $\exists\, n_3$ so that $P^{n_3}(x, y) > 0$.
Finally, let $N = n_3 + n_2(y)$. For any $n \geq N$,

$$P^n(x, y) \geq P^{n_3}(x, y) P^{n-n_3}(y, y) > 0,$$

using Claim 2 since $n - n_3 \geq N - n_3 = n_2(y)$. This proves the theorem. □

8.7 Recurrence and transience

What this section is about: We are going to divide the states of a Markov chain into different "types". There will be general laws which govern the behavior of each "type" of state, and the types of states of the chain gives you information about whether the chain has stationary distributions and/or a steady-state distribution.

Definition 8.40 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$.*

1. *For each $x \in \mathcal{S}$, set $f_x = f_{x,x} = P_x(T_x < \infty)$.*
2. *A state $x \in \mathcal{S}$ is called* **recurrent** *if $f_x = 1$.*

   The set of recurrent states of the Markov chain is denoted $\mathcal{S}_R$.

   The Markov chain $\{X_t\}$ is called **recurrent** *if $\mathcal{S}_R = \mathcal{S}$, i.e. all of its states are recurrent.*
3. *A state $x \in \mathcal{S}$ is called* **transient** *if $f_x < 1$.*

   The set of transient states of the Markov chain is denoted $\mathcal{S}_T$.

   The Markov chain $\{X_t\}$ is called **transient** *if all its states are transient.*

Recurrent and transient states are two of the "types" of states referred to earlier:

- a recurrent state (by definition) is "a state to which you ***must*** return" (with probability 1)
- a transient state is (by definition) "a state to which you ***might not*** return".

Expected number of visits to a state

The rest of this section is devoted to developing properties of recurrent and transient states. The key to deriving these properties is to connect recurrence/transience with n-step transition probabilities $P^n(x, y)$:

behavior of the n-step transitions $P^n(x,y)$	$\overset{\text{Thm 8.42}}{\longleftrightarrow}$		$\overset{\text{Thm 8.43}}{\longleftrightarrow}$	recurrence/ transience of x and/or y

To connect these concepts, we will need a new idea that involves the expected value of a new random variable that counts the number of times a Markov chain "visits" each state.

Definition 8.41 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$.*

- *For each $y \in \mathcal{S}$, define the r.v. V_y, the* **number of visits** *to y, by*

$$V_y = \#\text{ of times } t \geq 1 \text{ such that } X_t = y.$$

- *For each $y \in \mathcal{S}$ and $N \in \{1, 2, 3, ...\}$, define the r.v. $V_{y,N}$, the* **number of visits** *to y* **up to time** *N, by*

$$V_{y,N} = \#\text{ of times } t \in \{1, 2, ..., N\} \text{ such that } X_t = y.$$

Observe: $V_y : \Omega \to \{0, 1, 2, 3, ...\} \bigcup \{\infty\}$, but $V_{y,N} : \Omega \to \{0, 1, 2, 3, ..., N\}$.

Now, for the first major theorem of this section. This result connects n-step transition probabilities to a state with the expected number of visits to that state:

Theorem 8.42 (Formula for expected number of visits) *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. Then, for any $x, y \in \mathcal{S}$, we have*

$$E_x(V_y) = \sum_{n=1}^{\infty} P^n(x, y) \quad \text{and} \quad E_x(V_{y,N}) = \sum_{n=1}^{N} P^n(x, y).$$

The proof of this theorem is the first of several places where it will be convenient to use something called an **indicator function** or **characteristic function**. Suppose E is a set and X is a r.v. Denote by $\mathbb{1}_E$ the r.v. defined by

$$\mathbb{1}_E(X) = \begin{cases} 1 & \text{if } X \in E \\ 0 & \text{if } X \notin E \end{cases}.$$

Observe that if $\{X_t\}$ is a sequence of r.v.s, we can *count* the number of the X_t that are in set E *by adding* the $\mathbb{1}_E(X_t)$. In particular, PROOF The first equation follows as a direct calculation:

$$\begin{aligned}
E_x(V_y) = E_x\left[\sum_{n=1}^{\infty} \mathbb{1}_{\{y\}}(X_n)\right] &= \sum_{n=1}^{\infty} E_x(\mathbb{1}_{\{y\}}(X_n)) \\
&= \sum_{n=1}^{\infty} [1 \cdot P_x(X_n = y) + 0 \cdot P_x(X_n \neq y)] \\
&= \sum_{n=1}^{\infty} P_x(X_n = y) \\
&= \sum_{n=1}^{\infty} P^n(x, y).
\end{aligned}$$

The second equation has the same proof, with N instead of ∞ as the upper limit of the sum. □

Now for a theorem that connects expected number of visits with recurrence and transience:

Theorem 8.43 (Properties of recurrent/transient states) *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. Then:*

1. *If $y \in \mathcal{S}_T$, then for all $x \in \mathcal{S}$,*
$$P_x(V_y < \infty) = 1 \text{ and } E_x(V_y) = \frac{f_{x,y}}{1 - f_y} < \infty.$$

2. *If $y \in \mathcal{S}_R$, then*
$$P_x(V_y = \infty) = P_x(T_y < \infty) = f_{x,y}$$
(in particular $P_y(V_y = \infty) = 1$) and
   - *(a) if $f_{x,y} = 0$, then $E_x(V_y) = 0$;*
   - *(b) if $f_{x,y} > 0$, then $E_x(V_y) = \infty$.*

What this theorem says in English:

1. If y is transient, then no matter where you start, you only visit y a finite number of times (and the expected number of times you visit is $\frac{f_{x,y}}{1 - f_y}$).
2. If y is recurrent, then
   - it may be possible that you never hit y, but
   - if you hit y, then you must visit y infinitely many times.

PROOF First, observe that $V_y \geq 1 \iff T_y < \infty$, because both statements correspond to hitting y in a finite amount of time.

Therefore $P_x(V_y \geq 1) = P_x(T_y < \infty) = f_{x,y}$.

Now $P_x(V_y \geq 2) =$

Similarly $P_x(V_y \geq n) =$

That means that for all $n \geq 1$ we have $P_x(V_y = n) =$

Case 1: y is transient (i.e. $f_y = f_{y,y} < 1$).

Then $P_x(V_y = \infty) = \lim_{n\to\infty} P_x(V_y \geq n) = \lim_{n\to\infty} f_{x,y} f_y^{n-1} = 0$ so $P_x(V_y < \infty) = 1$ as wanted.

Also,

$$\begin{aligned} E_x(V_y) &= \sum_{n=0}^{\infty} n \cdot P_x(V_y = n) \\ &= \sum_{n=1}^{\infty} n \cdot P_x(V_y = n) \\ &= \sum_{n=1}^{\infty} n f_{x,y} f_y^{n-1}(1 - f_y) && \text{(from above)} \\ &= f_{x,y}(1 - f_y) \sum_{n=1}^{\infty} n f_y^{n-1} \\ &= f_{x,y}(1 - f_y) \frac{1}{(1 - f_y)^2} && \text{(pink sheet)} \\ &= \frac{f_{x,y}}{1 - f_y}. \end{aligned}$$

Case 2: y is recurrent (i.e. $f_y = f_{y,y} = 1$).

Then $P_x(V_y = \infty) = \lim_{n\to\infty} P_x(V_y \geq n) = \lim_{n\to\infty} f_{x,y} f_y^{n-1} = \lim_{n\to\infty} f_{x,y} 1^{n-1} = f_{x,y}$.

So if $f_{x,y} > 0$, then $E_x(V_y) = \infty$, since $P_x(V_y = \infty) = f_{x,y} > 0$.

If $f_{x,y} = 0$, then $P^n(x, y) = 0$ for all $n \geq 1$, so by Lemma 8.42,

$$E_x(V_y) = \sum_{n=1}^{\infty} P^n(x, y) = \sum_{n=1}^{\infty} 0 = 0. \square$$

Recurrence criteria

Putting the previous two results together yields these criteria, which can be useful to determine if a state is recurrent or transient:

Corollary 8.44 (Recurrence criteria) *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. Let $y \in \mathcal{S}$.*

1. $y \in \mathcal{S}_R \iff E_y(V_y) = \infty \iff \sum_{n=1}^{\infty} P^n(y,y)$ *diverges.*

2. *If there exists $x \in \mathcal{S}$ so that $\lim_{n\to\infty} P^n(x,y) \neq 0$, then $y \in \mathcal{S}_R$.*

   (Restated, if $y \in \mathcal{S}_T$, then $\lim_{n\to\infty} P^n(x,y) = 0$ for all $x \in \mathcal{S}$.)

PROOF Statement (1) follows from 2(b) of Theorem 8.43 and Theorem 8.42.

For statement (2), y being transient implies $E_x(V_y) < \infty$ by (1) of Theorem 8.43, and that implies $\sum_{n=1}^{\infty} P^n(x,y) < \infty$ by Theorem 8.42.

By the n^{th}-term Test for infinite series (Calculus 2), $\lim_{n\to\infty} P^n(x,y) = 0$. $\square$

EXAMPLE 16

Consider a Markov chain with state space $\{1,2,3\}$ and transition matrix

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1-p & p \end{pmatrix}$$

where $p \in (0,1)$.

1. Which states are recurrent? Which states are transient?
2. Find $f_{x,y}$ for all $x, y \in \mathcal{S}$.
3. Find the expected number of visits to each state, given that you start in any of the states.

Recurrence/transience and class structure

Theorem 8.45 (Recurrent states lead only to recurrent states) *Suppose $\{X_t\}$ is a Markov chain. If $x \in \mathcal{S}_R$ and $x \to y$, then:*

1. $f_{y,x} = f_{x,y} = 1$*; and*
2. $y \in \mathcal{S}_R$.

PROOF If $y = x$, this follows from the definition of "recurrent", so assume $y \neq x$. We are given $x \to y$, so $P^n(x, y) > 0$ for some $n \geq 1$. Let N be the smallest $n \geq 1$ such that $P^n(x, y) > 0$. Then we have a picture like this:

Proof that $f_{y,x} = 1$:

Suppose not, i.e. that $f_{y,x} < 1$. Then

$$1 - f_x \geq P(x, y_1)P(y_1, y_2)P(y_2, y_3) \cdots P(y_{N-1}, y_N)\,[1 - f_{y,x}] > 0$$

so $1 - f_x > 0$, so $f_x < 1$. This contradicts $x \in \mathcal{S}_R$. Therefore $f_{y,x} = 1$.

Proof that $y \in S_R$:

Since $f_{y,x} = 1$, $y \to x$, so $\exists N'$ such that $P^{N'}(y, x) > 0$.

This means for every $n \geq 0$, $P^{N'+n+N}(y, y) \geq P^{N'}(y, x)P^n(x, x)P^N(x, y)$.

We'll prove y is recurrent by showing $E_y(V_y) = \infty$:

$$\begin{aligned} E_y(V_y) &= \sum_{n=1}^{\infty} P^n(y, y) \\ &\geq \sum_{n=N'+N+1}^{\infty} P^n(y, y) \\ &\geq \sum_{n=1}^{\infty} P^{N'}(y, x)P^n(x, x)P^N(x, y) \end{aligned}$$

$$\begin{aligned}
E_y(V_y) &= P^{N'}(y,x)P^N(x,y)\sum_{n=1}^{\infty} P^n(x,x)...\\
&= P^{N'}(y,x)P^N(x,y)E_x(V_x)\\
&\qquad \text{(Formula for expected number of visits)}\\
&= \infty\\
&\qquad \text{(Recurrence criterion 1 of Corollary 8.44).}
\end{aligned}$$

By a recurrence criteria, since $E_y(V_y) = \infty$, $y \in \mathcal{S}_R$.

Proof that $f_{x,y} = 1$:

Since $y \in \mathcal{S}_R$ and $y \to x$, $f_{x,y} = 1$ by the first statement we proved. □

Corollary 8.46 (Finite state space Markov chains are not transient) *Let $\{X_t\}$ be a Markov chain with* **finite** *state space $\mathcal{S}$. Then the Markov chain is not transient (i.e. there is at least one recurrent state).*

PROOF Suppose not, i.e. all states are transient. Then by a recurrence criterion,

$$\begin{aligned}
0 &= \lim_{n\to\infty} P^n(x,y) \quad \forall\, x,y \in \mathcal{S}\\
\Rightarrow 0 &= \sum_{y\in\mathcal{S}} \lim_{n\to\infty} P^n(x,y)\\
\Rightarrow 0 &= \lim_{n\to\infty} \sum_{y\in\mathcal{S}} P^n(x,y)\\
\Rightarrow 0 &= \lim_{n\to\infty} 1.
\end{aligned}$$

This is a contradiction! Therefore there must be at least one recurrent state. □

Theorem 8.47 (Decomposition Theorem) *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. If $\mathcal{S}_R \neq \emptyset$, then we can write*

$$\mathcal{S}_R = \bigcup_j C_j$$

where the C_j are disjoint communicating classes (the union is either finite or countable).

PROOF Since $\mathcal{S}_R \neq \emptyset$, we can choose some $x \in \mathcal{S}_R$.

Define $C(x) = \{y \in \mathcal{S} : x \to y\}$, meaning that $C(x)$ is the set of states you can get to (eventually) from x.

Since x is recurrent, $x \in C(x)$. Thus $C(x) \neq \emptyset$, and $\mathcal{S}_R = \bigcup_{x\in\mathcal{S}_R} C(x)$.

Claim 1: $C(x)$ is closed.

Claim 2: $C(x)$ is a communicating class.

Claim 3: For $x, y \in \mathcal{S}_R$, the sets $C(x)$ and $C(y)$ are either disjoint or equal.

To verify this, suppose $z \in C(x) \cap C(y)$.

Summary of the theory developed so far

Theorem 8.48 (Main Recurrence/Transience Theorem) *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$.*

1. *If $C \subseteq \mathcal{S}$ is a communicating class, then either every state in C is recurrent (i.e. $C \subseteq \mathcal{S}_R$), or every state in C is transient (i.e. $C \subseteq \mathcal{S}_T$).*
2. *If $C \subseteq \mathcal{S}$ is a communicating class of recurrent states, then $f_{x,y} = 1$ for all $x, y \in C$.*
3. *If $C \subseteq \mathcal{S}$ is a finite communicating class, then $C \subseteq \mathcal{S}_R$.*
4. *If $\{X_t\}$ is irreducible, then $\{X_t\}$ is either recurrent or transient.*
5. *If $\{X_t\}$ is irreducible and $\mathcal{S}$ is finite, then $\{X_t\}$ is recurrent.*

Theorem 8.49 (State space decomposition) *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. We can write $\mathcal{S}$ as a disjoint union*

$$\mathcal{S} = \mathcal{S}_R \bigcup \mathcal{S}_T = \left(\underset{j}{\cup} C_j\right) \bigcup \mathcal{S}_T$$

where the C_j are recurrent communicating classes (there might be communicating classes in $\mathcal{S}_T$, but we don't care so much about those). Then:

1. *If you start in one of the C_j,*
   - *you will[a] stay in that C_j forever, and*
   - *you will visit every state in that C_j infinitely often.*
2. *If you start in $\mathcal{S}_T$, you either*
   a) *stay in $\mathcal{S}_T$ forever (but hit each state in $\mathcal{S}_T$ only finitely many times), or*
   b) *you will eventually enter a C_j, in which case you subsequently stay in that C_j forever and visit every state in that C_j infinitely often.*

   Situation 2 (a) above is only possible if $\mathcal{S}_T$ is infinite.

[a] Technicality: in this theorem, the phrase "you will" actually means "the probability that you will is 1".

Absorption probabilities

QUESTION

Suppose you have a Markov chain with state space decomposition as described above. Suppose you start at $x \in \mathcal{S}_T$. What is the probability that you eventually enter recurrent communicating class C_j?

Definition 8.50 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. Let $x \in \mathcal{S}_T$ and let C_j be a communicating class of recurrent states. The* **probability x is absorbed by C_j**, *denoted f_{x,C_j}, is*

$$f_{x,C_j} = P_x(T_{C_j} < \infty).$$

Lemma 8.51 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. Let $x \in \mathcal{S}_T$ and let C be a communicating class of recurrent states. Then for any $y \in C$, $f_{x,C_j} = f_{x,y}$.*

When $\mathcal{S}_T$ is finite, we can solve for these probabilities by solving a system of linear equations. Here is the method:

Suppose $\mathcal{S}_T = \{x_1, ..., x_n\}$.

Since $\mathcal{S}_T$ is finite, each x_j must eventually be absorbed by a C_j, so we have

$$\sum_i f_{x_j,C_i} = 1 \text{ for all } j.$$

Fix one of the C_i; then

$$f_{x_j,C_i} = P_{x_j}(T_{C_i} = 1) + P_{x_j}(T_{C_i} > 1)$$

If you write this equation for each $x_j \in \mathcal{S}_T$, you get a system of n equations in the n unknowns $f_{x_1,C_i}, f_{x_2,C_i}, f_{x_3,C_i}, ..., f_{x_n,C_i}$. This can be solved for the absorption probabilities for C_i; repeating this procedure for each i yields all the absorption probabilities of the Markov chain.

EXAMPLE 17

Consider a Markov chain with transition matrix

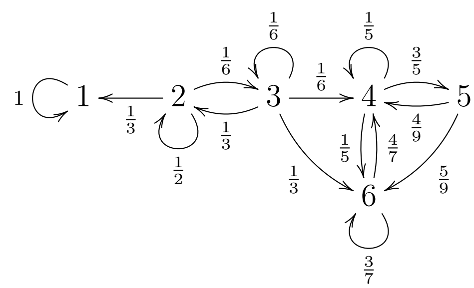
$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Determine which states of the chain are recurrent and which states are transient. For every $x \in \mathcal{S}_T$, compute $f_{x,1}$.

EXAMPLE 18

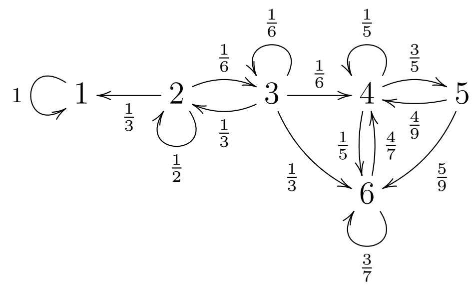
Let $\{X_t\}$ be a Markov chain with state space $\{1,2,3,4,5,6\}$ whose transition matrix and associated directed graph are

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{6} & 0 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} & 0 & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{5} & \frac{3}{5} & \frac{1}{5} \\ 0 & 0 & 0 & \frac{4}{9} & 0 & \frac{5}{9} \\ 0 & 0 & 0 & \frac{4}{7} & 0 & \frac{3}{7} \end{pmatrix}$$



Determine which states of the chain are recurrent and which states are transient. For each $x, y \in \mathcal{S}$, compute $f_{x,y}$.

(repeated for convenience)



8.8 Positive and null recurrence

What this section is about: We want to address big picture question (1) from earlier: when does a Markov chain have a stationary distribution?

We start with a couple of results telling us when there is *no* stationary distribution:

Theorem 8.52 (Stat. dists. give 0 probability to transient states) *Let π be a stationary distribution of Markov chain $\{X_t\}$. If $y \in \mathcal{S}_T$, then $\pi(y) = 0$.*

PROOF By stationarity, for all $n \geq 1$,

$$\sum_{x \in \mathcal{S}} \pi(x) P^n(x, y) = \pi(y).$$

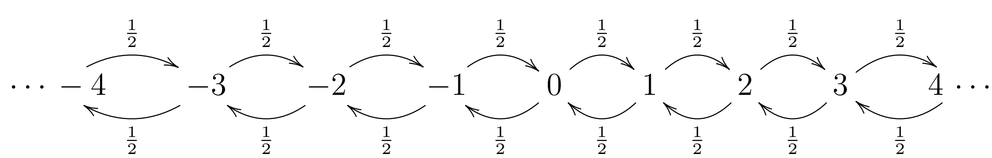
Take limits on both sides as $n \to \infty$. By the third recurrence criterion, since $y \in \mathcal{S}_T$, $\lim\limits_{n \to \infty} P^n(x, y) = 0$, so the equation above becomes $0 = \pi(y)$. □

Corollary 8.53 *If an irreducible Markov chain has a stationary (or steady-state) distribution, then the chain is recurrent.*

We'd like the converse of this corollary to be true (it would be great if every irreducible, recurrent Markov chain had a stationary distribution). **Unfortunately, it isn't.** To see, why, consider this example, which we've seen before:

EXAMPLE OF RECURRENT CHAIN WITH NO STATIONARY DISTRIBUTION

Simple, unbiased random walk on $\mathbb{Z}$:



Earlier, we saw that $\{X_t\}$ has no stationary distribution (because by symmetry, such a distribution would have to be uniform on $\mathbb{Z}$, and no such distribution exists).

Now, let's show that simple unbiased random walk is recurrent. Since $\{X_t\}$ is irreducible, it is sufficient to show that state 0 is recurrent. To do this, we'll use a recurrence criterion, and show that $\sum\limits_{n=1}^{\infty} P^n(0, 0)$ diverges.

To show $\sum_{n=1}^{\infty} P^n(0,0)$ diverges, notice first that

$$P^n(0,0) = \begin{cases} & \text{if } n \text{ is odd} \\ & \text{if } n = 2k \text{ is even} \end{cases}$$

By a HW problem from MATH 414 (that used Stirling's Formula), $\binom{2k}{k} \approx \frac{4^k}{\sqrt{\pi k}}$ for large k. So

$$\begin{aligned} \sum_{n=1}^{\infty} P^n(0,0) &= \sum_{k=1}^{\infty} P^{2k}(0,0) \\ &= \sum_{k=1}^{\infty} \binom{2k}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{2k-k} \\ &= \sum_{k=1}^{\infty} \binom{2k}{k} \frac{1}{4^k} \\ &\approx \sum_{k=1}^{\infty} \frac{4^k}{\sqrt{\pi k}} \frac{1}{4^k} \\ &= \frac{1}{\sqrt{\pi}} \sum_{k=1}^{\infty} \frac{1}{\sqrt{k}}. \end{aligned}$$

This series diverges by the __________________________________, so by a recurrence criterion $0 \in \mathcal{S}_R$, and by irreducibility the entire chain is recurrent.

Punchline: Simple unbiased random walk is an example of a Markov chain which is recurrent, but has no stationary distribution.

What's "wrong" in this example? Simple, unbiased random walk is recurrent, meaning that every state eventually returns to itself with probability 1. But it's only "barely" recurrent, because the *expected amount of time it takes to return to your initial value is infinite*. The technical term for this kind of recurrence is *null recurrence*.

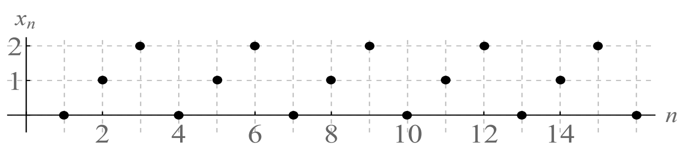
To have a stationary distribution, not only does an irreducible Markov chain need to be recurrent (meaning every state returns to itself with probability 1), but the *expected amount of time it takes to return to each state must be finite*. The term for this is *positive recurrence*, and this is the last ingredient in the FTMC.

Cesáro convergence

A sequence $\{a_n\}$ is said to **converge** to limit L if $\lim_{n\to\infty} a_n = L$, in which case we write $a_n \to L$.

EXAMPLES (AND A NON-EXAMPLE)

- $\dfrac{1}{n} \to 0.$
- $\dfrac{n+1}{n-1} \to 1.$
- The sequence $\{x_n\} = \{0, 1, 2, 0, 1, 2, 0, 1, 2, 0, 1, 2, ...\}$ does not converge:



However, this sequence does have some regular behavior:

Definition 8.54 *Let $\{x_n\}$ be a sequence of objects that can be added (like numbers, functions, vectors, random variables, etc.) The* **sequence of Cesàro averages** *of $\{x_n\}$ is the sequence $\{av(x)_n\}$ defined by setting*

$$av(x)_n = \frac{1}{n}\sum_{k=1}^{n} x_k$$

for all n. We say $\{x_n\}$ **converges in the Cesàro sense** *to L if the Cesàro averages converge to L, i.e. if*

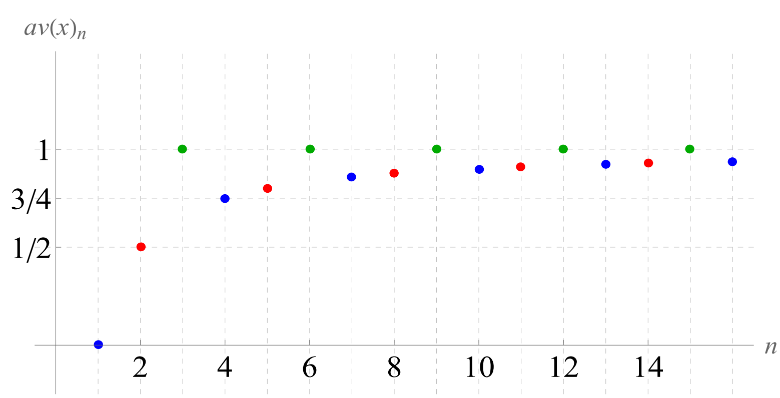
$$\lim_{n\to\infty} av(x)_n = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} x_k = L.$$

We write $x_n \xrightarrow{Ces} L$ to represent this.

EXAMPLE 19

Consider the sequence $\{x_n\} = \{0, 1, 2, 0, 1, 2, ...\}$.

1. Compute the first six terms of the sequence of Cesáro averages of $\{x_n\}$.
2. Determine what $av(x)_n$ is, in terms of n.
3. Show that $x_n \xrightarrow{Ces} 1$.



Facts about Cesàro convergence:

$$a_n \to L \text{ in the usual sense} \Rightarrow a_n \xrightarrow{Ces} L$$

$$a_n \xrightarrow{Ces} L \text{ and } \{a_n\} \text{ converges } \Rightarrow a_n \to L$$

"Cesàro convergence is weaker than usual convergence"

Why do we care about Cesàro convergence?

APPLICATION 1: SLLN

The Strong Law of Large Numbers (Chapter 6 / MATH 414) says

APPLICATION 2: MARKOV CHAINS

For any Markov chain, we will see that although $\lim_{n\to\infty} P^n(x,y)$ may not exist, the sequence $P^n(x,y)$ converges in the Cesàro sense for any $x, y \in \mathcal{S}$ (and the value to which the Cesáro averages converge has a lot to do with stationary and steady-state distributions, and with positive and null recurrence).

In particular, recall that $\sum_{k=1}^{n} P^k(x,y) =$

Therefore, the Cesàro averages of the sequence $\{P^n(x,y)\}$ are actually

$$\frac{1}{n}\sum_{k=1}^{n} P^k(x,y) =$$

Positive and null recurrence

Definition 8.55 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$ and transition function P.*

1. *Given $y \in \mathcal{S}_R$, define $m_y = E_y(T_y)$. m_y is a number (possibly ∞) called the* **mean return time** *to y.*

2. *A recurrent state y is called* **null recurrent** *if $m_y = \infty$. The set of null recurrent states of $\{X_t\}$ is denoted $\mathcal{S}_N$. If all the states of $\{X_t\}$ are null recurrent, $\{X_t\}$ is called* **null recurrent**.

3. *A recurrent state y is called* **positive recurrent** *if $m_y < \infty$. The set of positive recurrent states of $\{X_t\}$ is denoted $\mathcal{S}_P$. If all the states of $\{X_t\}$ are positive recurrent, $\{X_t\}$ is called* **positive recurrent**.

Note: The mean return time of any transient state is trivially ∞:

$$y \in \mathcal{S}_T \;\Rightarrow\; P_y(T_y = \infty) > 0 \;\Rightarrow\; E_y(T_y) = \infty \text{ automatically.}$$

Theorem 8.56 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. Let $y \in \mathcal{S}$.*

1. *If $T_y < \infty$ (i.e. if the chain hits y), $\displaystyle\lim_{n\to\infty} \frac{V_{y,n}}{n} = \frac{1}{m_y}$.*

2. *If $T_y = \infty$ (i.e. the chain never hits y), then $\displaystyle\lim_{n\to\infty} \frac{V_{y,n}}{n} = 0$.*

(Technically, these limits hold with probability 1.)

PROOF Let's start with statement (2). If the chain never hits y, then $V_{y,n} = 0$ for all n, so (2) follows.

To prove statement (1), assume WLOG[1] that you start in state y (since by hypothesis you must hit y at some point). Define these r.v.s:

- $T_y^r = \min\{n \geq 1 : V_{y,n} = r\} =$ time of r^{th} return to y
- $W_y^1 = T_y^1$
- $W_y^j = T_y^j - T_y^{j-1}$ for all $j \geq 2$

[1]"WLOG" means "without loss of generality"

The first key observation is that the W_y^j are i.i.d., each with mean m_y. So by the SLLN, $P\left(W_y^j \xrightarrow{Ces} m_y\right) = 1$. Restating this, we get

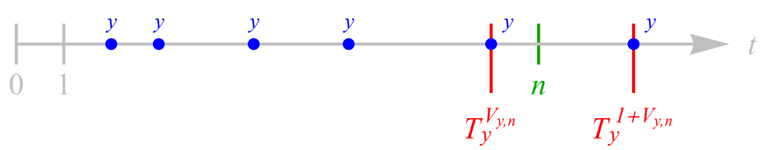
$$\begin{aligned}
P\left(W_y^j \xrightarrow{Ces} m_y\right) &= 1 \\
P\left(av(W_y^j)_n \to m_y\right) &= 1 \\
P\left(\lim_{n\to\infty} av(W_y^j)_n = m_y\right) &= 1 \\
P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} W_y^j = m_y\right) &= 1 \\
P\left(\lim_{n\to\infty} \frac{T_y^n}{n} = m_y\right) &= 1
\end{aligned}$$

Since $V_{y,n} \to \infty$ as $n \to \infty$, we can substitute $V_{y,n}$ for n to get

$$P\left(\lim_{n\to\infty} \frac{T_y^{V_{y,n}}}{V_{y,n}} = m_y\right) = 1. \qquad (\#)$$

The second key observation is that

which is explained by this picture:



Therefore

$$\frac{T_y^{V_{y,n}}}{V_{y,n}} \le \frac{n}{V_{y,n}} \le \frac{T_y^{1+V_{y,n}}}{} \cdot \frac{}{V_{y,n}}$$

Take reciprocals to get the result $\lim_{n\to\infty} \frac{V_{y,n}}{n} = \frac{1}{m_y}$. $\square$

Theorem 8.57 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. Let $x, y \in \mathcal{S}$.*

1. $\displaystyle\lim_{n\to\infty} \frac{E_x(V_{y,n})}{n} = \frac{f_{x,y}}{m_y}$.

2. $P^n(x,y) \xrightarrow{Ces} \dfrac{f_{x,y}}{m_y}$.

(Technically, these limits hold with probability 1.)

PROOF From the previous discussion, (1) implies (2), so it is sufficient to prove (1). To do this, note

$$\lim_{n\to\infty} \frac{E_x(V_{y,n})}{n} = \lim_{n\to\infty} E_x\left[\frac{V_{y,n}}{n}\right]$$

Corollary 8.58 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$.*

1. *Let $C \subseteq \mathcal{S}$ be a communicating class of recurrent states. Then for all $x, y \in C$,*

$$\lim_{n\to\infty} \frac{E_x(V_{y,n})}{n} = \frac{1}{m_y} \begin{cases} = 0 & \text{if } m_y = \infty \iff y \in \mathcal{S}_N \cup \mathcal{S}_T \\ > 0 & \text{if } m_y < \infty \iff y \in \mathcal{S}_P \end{cases}.$$

   Furthermore, if $P(X_0 \in C) = 1$, then $\displaystyle\lim_{n\to\infty} \frac{V_{y,n}}{n} = \frac{1}{m_y}\ \forall\, y \in C$.

2. *If $y \in \mathcal{S}_T \cup \mathcal{S}_N$, then for all $x \in \mathcal{S}$, $P^n(x,y) \xrightarrow{Ces} 0$.*

3. *If $y \in \mathcal{S}_P$, then $P^n(y,y) \xrightarrow{Ces} \dfrac{1}{m_y}$.*

PROOF Statement (1) follows immediately from Theorem 8.57.

For statement (2), notice that if $y \in \mathcal{S}_T \cup \mathcal{S}_N$, $m_y = \infty$ so

$$P^n(x,y) \xrightarrow{Ces} \frac{f_{x,y}}{m_y} = \frac{f_{x,y}}{\infty} = 0.$$

For statement (3), since y is recurrent, $f_y = f_{y,y} = 1$ so

$$P^n(y,y) \xrightarrow{Ces} \frac{f_{y,y}}{m_y} = \frac{1}{m_y}. \square$$

Note: Corollary 8.58 provides a new distinction between positive recurrent and null recurrent states:

$$\begin{aligned} y \in \mathcal{S}_N \cup \mathcal{S}_T &\iff P^n(y,y) \xrightarrow{Ces} 0 \\ y \in \mathcal{S}_P &\iff P^n(y,y) \xrightarrow{Ces} \frac{1}{m_y} > 0. \end{aligned}$$

Theorem 8.59 (Pos. recurrent states lead only to pos. recurrent states) *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. If $x \in \mathcal{S}_P$ and $x \to y$, then $y \in \mathcal{S}_P$.*

PROOF x is recurrent, so by a previous theorem $y \to x$. Thus $\exists\, n_1, n_2$ such that $P^{n_1}(x,y) > 0$ and $P^{n_2}(y,x) > 0$. Therefore, for all $m \geq 0$,

$$\begin{aligned} P^{n_1+m+n_2}(y,y) &\geq P^{n_1}(x,y)P^m(x,x)P^{n_2}(y,x) \\ \frac{1}{n}\sum_{m=1}^{n} P^{n_1+m+n_2}(y,y) &\geq \frac{1}{n}P^{n_1}(x,y)P^{n_2}(y,x)\sum_{m=1}^{n} P^m(x,x) \\ \lim_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n} P^{n_1+m+n_2}(y,y) &\geq \lim_{n\to\infty}\frac{1}{n}P^{n_1}(x,y)P^{n_2}(y,x)\sum_{m=1}^{n} P^m(x,x) \\ \lim_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n} P^{n_1+m+n_2}(y,y) &\geq P^{n_1}(x,y)P^{n_2}(y,x)\lim_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n} P^m(x,x) \end{aligned}$$

Corollary 8.60 (Null rec. states lead only to null rec. states) *Let $\{X_t\}$ be a Markov chain. If $x \in \mathcal{S}_N$ and $x \to y$, then $y \in \mathcal{S}_N$.*

PROOF HW (this is a short argument putting together facts facts from Theorems 8.45 and 8.59) □

Corollary 8.61 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. If $C \subseteq \mathcal{S}$ is a communicating class, then (every $x \in C$ is transient) or (every $x \in C$ is null recurrent) or (every $x \in C$ is positive recurrent).*

Theorem 8.62 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. If $C \subseteq \mathcal{S}$ is a finite communicating class, then every $x \in C$ is positive recurrent.*

PROOF For every $x \in C$ and $k \in \{1, 2, 3, ...\}$, we have $\sum\limits_{y \in C} P^k(x, y) = 1$. So

$$1 = \frac{1}{n} \cdot n = \frac{1}{n}(1 + 1 + 1 + ... + 1) = \frac{1}{n}\sum_{m=1}^{n} 1 = \frac{1}{n}\sum_{m=1}^{n}\sum_{y \in C} P^m(x, y)$$

Therefore there must be some $y \in C$ such that $m_y < \infty$, i.e. $y \in \mathcal{S}_P$. Since positive recurrence is a class property, every $x \in C$ is positive recurrent. □

Corollary 8.63 *Any irreducible Markov chain with a finite state space is positive recurrent.*

8.9 Existence and uniqueness of stationary distributions

What this section is about: We are going to prove that for an irreducible, positive recurrent Markov chain, then the chain has exactly one stationary distribution.

We begin by showing that for an irreducible Markov chain, values of any of its stationary distributions are determined by mean return times. To do this, we need to return to a previously encountered dilemma:

More about interchanging infinite series and limits

Recall that in general, the limit of an infinite series $\geq$ the series of the limits (but these aren't always equal):

$$\lim_{n\to\infty} \sum_{z=1}^{\infty} a(z) b_n(z) \geq \sum_{z=1}^{\infty} a(z) \lim_{n\to\infty} b_n(z). \tag{8.1}$$

If you need to argue that the two sides of (8.1) are equal, you need to work hard. One way to do this is an argument by exhaustion (discussed earlier); another way is to appeal to the following theorem that comes from a branch of mathematics called *real analysis*:

Theorem 8.64 (Bounded Convergence Theorem (BCT)) *Let $a(z)$ be nonnegative numbers such that $\sum_z a(z) < \infty$. Fix $B > 0$ and let $b_n(z)$ be numbers such that $\boxed{|b_n(z)| \leq B}$ for all z and n and*

$$\lim_{n\to\infty} b_n(z) = b(z) \text{ for all } z.$$

Then

$$\sum_z a(z) b_n(z) \overset{n\to\infty}{\longrightarrow} \sum_z a(z) b(z),$$

in other words

$$\lim_{n\to\infty} \sum_z a(z) b_n(x) = \sum_z \lim_{n\to\infty} a(z) b_n(z) = \sum_z a(z) \lim_{n\to\infty} b_n(z).$$

The condition $\boxed{|b_n(z)| \leq B}$ means that the $b_n(z)$ are **bounded**, giving this theorem its name.

Theorem 8.65 *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. If π is a stationary distribution for $\{X_t\}$, then for every $x \in \mathcal{S}$,*

$$\pi(x) = \frac{1}{m_x}.$$

PROOF Suppose π is stationary. Then, for all $k \in \{1, 2, 3, ...\}$ and all $z \in \mathcal{S}$, the stationarity equation gives

$$\sum_{z \in \mathcal{S}} \pi(z) P^k(z, x) \qquad = \qquad \pi(x).$$

Corollary 8.66 (Nonexistence of stationary distributions) .

1. *A transient Markov chain has no stationary distributions.*
2. *A null recurrent Markov chain has no stationary distributions.*

PROOF In either (1) or (2), $m_x = \infty$ for all $x \in \mathcal{S}$. By the preceding theorem, any stationary distribution π would have to satisfy $\pi(x) = \dfrac{1}{m_x} = 0$ for all $x \in \mathcal{S}$. But then $\sum_{x \in \mathcal{S}} \pi(x) = 0 \neq 1$, so π can't be a distribution. □

Theorem 8.67 (Existence/uniqueness of stationary distributions) *Let $\{X_t\}$ be an irreducible Markov chain.*

$\{X_t\}$ has a stationary distribution if and only if $\{X_t\}$ is positive recurrent, in which case the Markov chain has a unique stationary distribution π defined theoretically by

$$\pi(x) = \frac{1}{m_x} \quad \textit{for all } x \in \mathcal{S}.$$

PROOF What's left to prove is that for an irreducible, positive recurrent Markov chain $\{X_t\}$, the formula $\pi(x) = \dfrac{1}{m_x}$ actually defines a stationary distribution.

Case 1: $\mathcal{S}$ is finite:

In this situation, first notice that for all $m > 0$ and all $z \in \mathcal{S}$,

$$\begin{aligned}
\sum_{x \in \mathcal{S}} P^m(z,x) &= 1 \\
\Rightarrow \frac{1}{n} \sum_{m=1}^{n} \sum_{x \in \mathcal{S}} P^m(z,x) &= \frac{1}{n} \sum_{m=1}^{n} 1 = 1 \\
\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} \sum_{x \in \mathcal{S}} P^m(z,x) &= 1 \\
\lim_{n\to\infty} \sum_{x \in \mathcal{S}} \frac{1}{n} \sum_{m=1}^{n} P^m(z,x) &= 1 \\
\sum_{x \in \mathcal{S}} \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} P^m(z,x) &= 1 \qquad \text{(since } \mathcal{S} \text{ is finite)} \\
\sum_{x \in \mathcal{S}} \frac{f_{zx}}{m_x} &= 1 \qquad \text{(since } P^n(z,x) \xrightarrow{Ces} \frac{f_{zx}}{m_x}\text{)} \\
\sum_{x \in \mathcal{S}} \frac{1}{m_x} &= 1 \qquad (f_{zx} = 1 \text{ since } \{X_t\} \text{ pos. rec.}).
\end{aligned}$$

Therefore $\pi(x) = \dfrac{1}{m_x}$ defines a distribution.

Continuing with Case 1 (where $\mathcal{S}$ is finite), what's left is to show that the distribution defined by $\pi(x) = \dfrac{1}{m_x}$ is in fact stationary (we have to verify the stationarity equation). To do this, observe

$$P^{m+1}(z,y) = \sum_{x\in\mathcal{S}} P^m(z,x)P(x,y) \qquad (\#)$$

Now, take Cesáro limits of both sides of (#). Start with the left-hand side:

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} P^{m+1}(z,y) \\
&= \lim_{n\to\infty} \frac{1}{n}\left[\sum_{m=1}^{n+1} P^m(z,y) - P^1(z,y)\right] \\
&= \lim_{n\to\infty} \frac{n+1}{n}\cdot\left[\frac{1}{n+1}\sum_{m=1}^{n+1} P^m(z,y)\right] - \lim_{n\to\infty}\frac{1}{n}P(z,y)
\end{aligned}$$

Now for the right-hand side of (#):

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n}\sum_{x\in\mathcal{S}} P^m(z,x)P(x,y) \\
&= \lim_{n\to\infty}\sum_{x\in\mathcal{S}} \frac{1}{n}\sum_{m=1}^{n} P^m(z,x)P(x,y) \\
&= \sum_{x\in\mathcal{S}}\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} P^m(z,x)P(x,y) \qquad \text{(since } \mathcal{S} \text{ is finite)} \\
&= \sum_{x\in\mathcal{S}} P(x,y)\left[\lim_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n} P^m(z,x)\right]
\end{aligned}$$

By equating the Cesáro limits of the two sides of (#), we get the stationarity equation

$$\pi(y) = \frac{1}{m_y} = \sum_{x\in\mathcal{S}}\frac{1}{m_x}P(x,y) = \sum_{x\in\mathcal{S}}\pi(x)P(x,y),$$

proving the theorem in the situation where $\mathcal{S}$ is finite.

Notice that the argument above doesn't work when $\mathcal{S}$ is infinite because you can't interchange the infinite sums with the limits. To prove the theorem when $\mathcal{S}$ is infinite, we need an argument by exhaustion, given on the next page:

Case 2: $\mathcal{S}$ is infinite.

In this situation, let $\mathcal{S}_{\text{finite}} \subseteq \mathcal{S}$ be an arbitrary finite subset of $\mathcal{S}$.

Now, instead of (#) above, we know the *inequality*

$$P^{m+1}(z,y) = \sum_{x\in\mathcal{S}} P^m(z,x)P(x,y) \geq \sum_{x\in\mathcal{S}_{\text{finite}}} P^m(z,x)P(x,y) \qquad (\heartsuit)$$

Since $\mathcal{S}_{\text{finite}}$ is finite, we can repeat everything we did on the previous page to get

$$\pi(y) \geq \sum_{x\in\mathcal{S}_{\text{finite}}} \pi(x)P(x,y).$$

Because $\mathcal{S}_{\text{finite}}$ is arbitrary, by taking limits as $\mathcal{S}_{\text{finite}} \nearrow \mathcal{S}$ we get

$$\boxed{\pi(y) \geq \sum_{x\in\mathcal{S}} \pi(x)P(x,y)}.$$

Now suppose $\pi(y) > \sum_{x\in\mathcal{S}} \pi(x)P(x,y)$.

Then, by summing over all $y \in \mathcal{S}$, we get

$$\begin{aligned}\sum_{y\in\mathcal{S}} \pi(y) > \sum_{y\in\mathcal{S}}\sum_{x\in\mathcal{S}} \pi(x)P(x,y) &= \sum_{x\in\mathcal{S}}\sum_{y\in\mathcal{S}} \pi(x)P(x,y)\\ &= \sum_{x\in\mathcal{S}} \pi(x) \sum_{y\in\mathcal{S}} P(x,y)\\ &= \sum_{x\in\mathcal{S}} \pi(x)\cdot 1\\ &= \sum_{x\in\mathcal{S}} \pi(x)\end{aligned}$$

which is a contradiction. Therefore $\sum_{x\in\mathcal{S}} \pi(x)P(x,y) = \pi(y)$.

This doesn't mean π is stationary (because we don't know the values of π sum to 1, but it does mean that a multiple of π, say $M\pi$, is stationary.

However, by Theorem 8.65 the only value of M that is possible is $M = 1$.

Therefore $\pi(x) = \dfrac{1}{m_x}$ defines a stationary distribution. $\square$

Corollary 8.68 *Any irreducible Markov chain on a finite state space has a unique stationary distribution.*

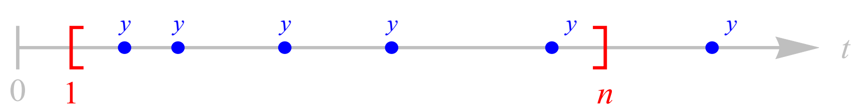
The ergodic theorem

Theorem 8.69 (Ergodic Theorem for Markov chains) *Let $\{X_t\}$ be an irreducible, positive recurrent Markov chain with state space $\mathcal{S}$ and let π be its unique stationary distribution. Then for all $y \in \mathcal{S}$,*

$$P\left(\lim_{n\to\infty} \frac{V_{y,n}}{n} = \pi(y)\right) = 1.$$

PROOF We've seen that $\pi(y) = \frac{1}{m_y}$; the result follows from Theorem 8.57. □

A picture to explain the ergodic theorem:



EXAMPLE 20

Suppose $\{X_t\}$ is a Markov chain with $\mathcal{S} = \{1, 2, 3, 4\}$ whose stationary distribution is $\left(\frac{1}{9}, \frac{2}{9}, \frac{4}{9}, \frac{2}{9}\right)$. Suppose $X_0 = 1$.

1. Estimate the number of times t in the interval $[1, 900]$ such that $X_t = 2$.

2. What is the mean return time to state 2?

Stationary distributions for non-irreducible Markov chains

Definition 8.70 *A distribution π on $\mathcal{S}$ is* **supported** *or* **concentrated** *on a subset $C \subseteq \mathcal{S}$ if $\pi(x) = 0$ for all $x \notin C$.*

EXAMPLE 21

If $\mathcal{S} = \{1, 2, 3, 4\}$ and $\pi = (\frac{1}{2}, 0, \frac{1}{2}, 0)$, we say π is supported on $\{1, 3\}$.

Given a non-irreducible Markov chain $\{X_t\}$, we can identify the various positive recurrent communicating classes of $\{X_t\}$ and treat each of those classes as their own Markov chain. That leads to this theorem, which sums up the content of this section:

Theorem 8.71 (Existence/uniqueness of stationary distributions) *Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$.*

1. *We can write $\mathcal{S}$ as the disjoint unioni*

$$\mathcal{S} = \mathcal{S}_T \bigcup \mathcal{S}_R = \mathcal{S}_T \bigcup \left(\mathcal{S}_N \bigcup \mathcal{S}_P \right).$$

2. *If $\mathcal{S}_P = \emptyset$, then $\{X_t\}$ has no stationary distribution.*

3. *If $\mathcal{S}_P \neq \emptyset$ consists of one communicating class, then $\{X_t\}$ has a unique stationary distribution π defined by*

$$\pi(x) = \begin{cases} \dfrac{1}{m_x} & \text{if } x \in \mathcal{S}_P \\ 0 & \text{else} \end{cases}$$

4. *If $\mathcal{S}_P \neq \emptyset$ consists of more than one communicating class, then for each communicating class $C \subseteq \mathcal{S}_P$ there is a unique stationary distribution supported on that class (call it π_C) defined by*

$$\pi_C(x) = \begin{cases} \dfrac{1}{m_x} & \text{if } x \in C \\ 0 & \text{else} \end{cases}$$

Convex combinations of these π_C are also stationary, so $\{X_t\}$ has infinitely many stationary distributions. (The stationary distributions are exactly the convex combinations of these π_C.)

EXAMPLE 22

Find all stationary distributions of the Markov chain with transition matrix

$$\begin{pmatrix} \frac{1}{4} & 0 & 0 & 0 & \frac{3}{4} & 0 \\ \frac{1}{8} & \frac{1}{2} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & 0 \\ 0 & 0 & \frac{3}{4} & 0 & 0 & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{4} & 0 & 0 & \frac{3}{4} \end{pmatrix}$$

EXAMPLE 23

Let $\{X_t\}$ be the Markov chain with state space $\mathcal{S} = \{0, 1, 2, 3, ...\}$ and transition function P defined by

$$P(x,y) = \begin{cases} \dfrac{1}{2} & \text{if } y = 0 \\ \dfrac{1}{4} & \text{if } y = x+1 \\ \dfrac{1}{4} & \text{if } y = x+2 \\ 0 & \text{else} \end{cases}$$

Show that $\{X_t\}$ is positive recurrent, and compute $\pi(0)$, $\pi(1)$, $\pi(2)$ and $\pi(3)$ for the stationary distribution π of $\{X_t\}$.

8.10 Proving the Fundamental Theorem

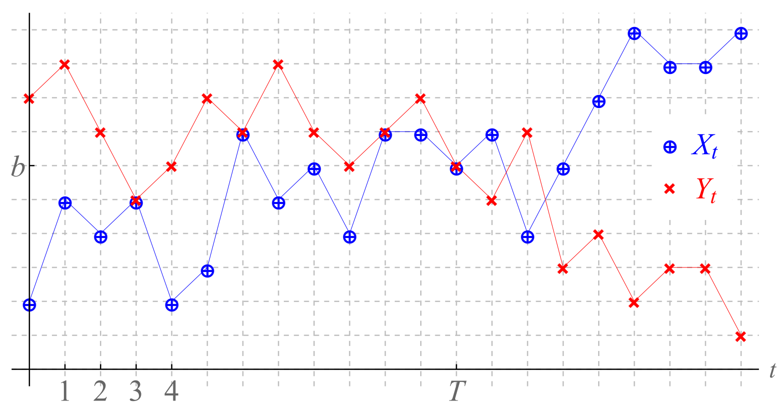
Theorem 8.72 (FTMC) *Let $\{X_t\}$ be an irreducible, aperiodic, and positive recurrent Markov chain. Then the unique stationary distribution of this chain, defined by $\pi(x) = \frac{1}{m_x}$ is steady-state, meaning*

$$\lim_{n\to\infty} \pi_n(x) = \pi(x)$$

for all $x \in \mathcal{S}$, no matter the initial distribution π_0.

PROOF Let $\{Y_t\}$ be a Markov chain, independent of $\{X_t\}$, with the same state space and transition function as $\{X_t\}$, but where the initial distribution of $\{Y_t\}$ is the stationary distribution π.

Pick $b \in \mathcal{S}$ arbitrarily and set $T = \min\{t \geq 1 : X_t = Y_t = b\}$. (If there is no such t, set $T = \infty$.) We call T a "coupling time" because it is the first time at which X_t and Y_t are in the same place b (so X_t and Y_t are "coupled" at state b).



Claim: $P(T < \infty) = 1$.

The proof of this claim is HW. (This is where we use the aperiodicity of $\{X_t\}$, because it applies Theorem 8.39 which says that for any two states, it is possible to get from one to the other in all times greater than or equal to some N.)

Now, define a new process $\{Z_t\}$ which starts out acting like $\{X_t\}$, but switches to acting like $\{Y_t\}$ after the coupling time:

$$Z_t = \begin{cases} X_t & \text{if } t < T \\ Y_t & \text{if } t \geq T \end{cases}$$

$\{Z_t\}$ is a Markov chain with the same initial distribution as $\{X_t\}$ and the same transition function as $\{X_t\}$, therefore $\boxed{\{Z_t\} \sim \{X_t\}}$. Thus

$$\begin{aligned}
&|P(X_t = y) - \pi(y)| \\
&= |P(Z_t = y) - P(Y_t = y)| \\
&= |P(X_t = y \text{ and } t < T) + P(Y_t = y \text{ and } t \geq T) - P(Y_t = y)| \\
&= |P(X_t = y \text{ and } t < T) - P(Y_t = y \text{ and } t < T)| \\
&\leq P(t < T) \to 0 \text{ as } t \to \infty \text{ by the Claim above.}
\end{aligned}$$

Therefore $|P(X_t = y) - \pi(y)| \to 0$ as $t \to \infty$, so

$$\lim_{t\to\infty} \pi_t(y) = \lim_{t\to\infty} \sum_{x\in\mathcal{S}} \pi_0(x) P^t(x,y) = \pi(y)$$

for all x and y.

By choosing π_0 to be $\pi_0(x) = \begin{cases} 1 & \text{if } x = z \\ 0 & \text{else} \end{cases}$, we see that

$$\lim_{t\to\infty} P^t(z,y) = \pi(y)$$

for all $z \in \mathcal{S}$; thus π is steady-state by definition. □

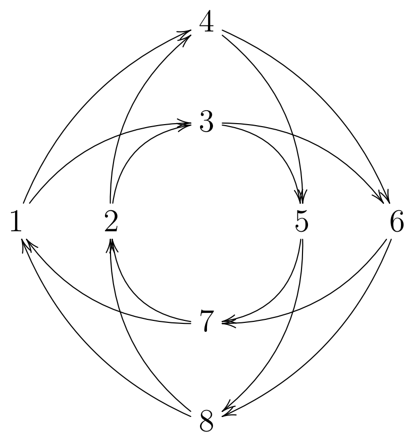
What if the Markov chain is periodic?

We have seen by example that if $\{X_t\}$ is irreducible and positive recurrent but periodic, then the FTMC doesn't hold (the stationary distribution isn't steady state). The reason the argument in the FTMC fails is the claim that $P(T < \infty) = 1$ doesn't hold for periodic Markov chains.

To help understand what happens for a periodic Markov chain, let's consider an example:

EXAMPLE 24

Consider a Markov chain $\{X_t\}$ with state space $\{1, 2, 3, 4, 5, 6, 7, 8\}$ whose directed graph looks like the one below (with unspecified nonzero probabilities on the arrows):



Consider a table of values for $P^n(1, 1)$:

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	$\cdots$
$P^n(1,1)$																	$\cdots$

Similarly, a table of values for $P^n(3, 2)$ looks like

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	$\cdots$
$P^n(3,2)$	0	0	0		0	0	0		0	0	0		0	0	0		$\cdots$

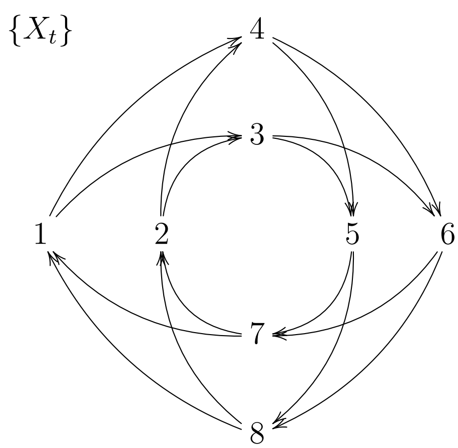
and a table of values for $P^n(8, 4)$ looks like

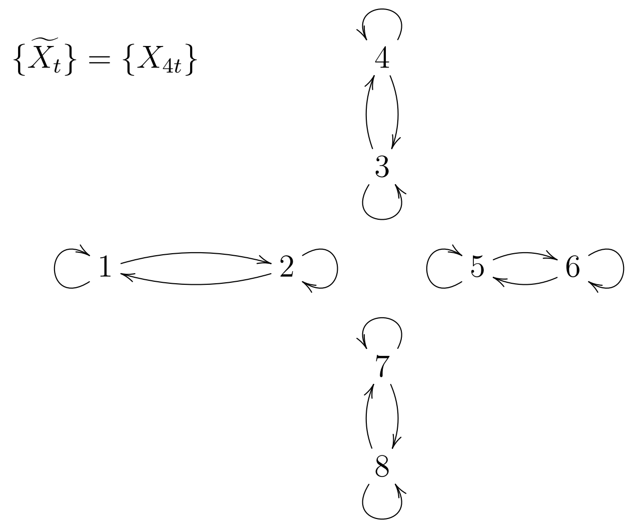
n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	$\cdots$
$P^n(3,2)$	0	0		0	0	0		0	0	0		0	0	0		0	$\cdots$

Observation 1: For every pair x, y of states, there is a number $r = r(x, y)$ such that $P^n(x, y) = 0$ unless n has remainder r when divided by the period n (i.e. unless $n \equiv r \mod d$).

Question: What is the long-term behavior of the *non-zero* terms in these tables (the ones marked with $*$s)?

To determine the long-term behavior of the non-zero $P^n(x, y)$ as $n \to \infty$, we consider a new Markov chain $\{\widetilde{X}_t\}$, where one unit of time in $\{\widetilde{X}_t\}$ corresponds to 4 units of time in $\{X_t\}$, i.e. $\widetilde{X}_t = X_{4t}$, i.e. the transition function for $\widetilde{X}_t$ is $\widetilde{P}(x, y) = P^4(x, y)$. This new chain will be aperiodic but will **not** be irreducible; it has the directed graph shown below at right:





Let m_x and $\widetilde{m}_x$ denote the mean return times to each state x in $\{X_t\}$ and $\{\widetilde{X}_t\}$, respectively. Since one unit of time in $\{\widetilde{X}_t\}$ corresponds to four units of time in $\{X_t\}$, we know that for every $x \in \mathcal{S}$,

$$4\widetilde{m}_x = m_x$$

Now, by the FTMC, we get

$$\begin{aligned}
\lim_{n\to\infty} \widetilde{P}^n(1,1) &= \frac{1}{\widetilde{m}_1} \\
\lim_{n\to\infty} P^{4n}(1,1) &= \frac{1}{m_1/4} \\
\lim_{n\to\infty} P^{4n+0}(1,1) &= \frac{4}{m_1} \\
\text{i.e. } \lim_{n\to\infty} P^{dn+r}(1,1) &= \frac{d}{m_2} = d \cdot \pi(1).
\end{aligned}$$

Observation 2: In the charts on the previous page, the non-zero entries of $P^n(x, y)$ approach $d \cdot \pi(y)$ as $n \to \infty$.

Our observations in the preceding example hold in general:

Theorem 8.73 *Let $\{X_t\}$ be an irreducible, positive recurrent Markov chain with state space $\mathcal{S}$, whose period is $d \geq 2$. Let π denote its unique stationary distribution. Then, for every pair of states x, y, there is a number $r = r(x, y)$ so that :*

1. *$P^n(x, y) = 0$ unless $n = md + r$ for some $m \in \mathbb{N}$ (i.e. unless $n \equiv r \mod d$).*
2. *$\lim_{m\to\infty} P^{md+r}(x, y) = d \cdot \pi(y)$.*

PROOF Let m_x be the mean return time of each state x with respect to the Markov chain $\{X_t\}$. Now consider the Markov chain $\{\widetilde{X}_t\}$ with the same initial distribution as $\{X_t\}$ whose transition function is P^d. Note that the mean return time for each state with respect to $\{\widetilde{X}_t\}$ is $\dfrac{m_x}{d}$.

$\{\widetilde{X}_t\}$ is not irreducible; it has d disjoint communicating classes. Restricting $\{\widetilde{X}_t\}$ to each of these classes gives an aperiodic, pos.recurrent, irreducible chain to which we can apply the FTMC; this gives

$$\lim_{m\to\infty} (P^d)^m(x, x) = \frac{1}{m_x/d} = \frac{d}{m_x}, \quad \text{i.e.} \quad \lim_{m\to\infty} P^{md}(x, x) = d \cdot \pi(x).$$

More generally, if $z \in \mathcal{S}$ is such that $P^d(z, x) > 0$, then z and x belong to the same communicating class of $\{\widetilde{X}_t\}$, so

$$\lim_{m\to\infty} P^{md}(z, x) = d \cdot \pi(x).$$

Now let $x, y \in \mathcal{S}$. If r is such that $P^r(x, y) > 0$, then

$$\begin{aligned}
\lim_{m\to\infty} P^{md+r}(x, y) &= \lim_{m\to\infty} \sum_{z\in\mathcal{S}} P^r(x, z) P^{md}(z, y) \\
&= \sum_{z\in\mathcal{S}} P^r(x, z)\, d \cdot \pi(y) \\
&= 1 \cdot d \cdot \pi(y) \\
&= d \cdot \pi(y)
\end{aligned}$$

as desired. □

8.11 Example computations

DIRECTIONS

For each given Markov chain in Examples 25-28:

1. Classify the states as transient, positive recurrent or null recurrent.
2. Find all communicating classes of the Markov chain.
3. Find the period of each state.
4. Find all stationary distribution(s) of the Markov chain (if any exist) and determine which (if any) of these distributions are steady-state. (If you can't compute the entire stationary distribution, find as many values of the stationary distribution as you can.)
5. Find the mean return time to state 2.

EXAMPLE 25

The Ehrenfest chain with $d = 4$.

EXAMPLE 26

The Markov chain whose transition matrix is

$$\begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 & \frac{3}{4} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

EXAMPLE 27

Let $\{X_t\}$ be a Markov chain with $\mathcal{S} = \{0, 1, 2, 3, 4, 5, 6\}$ such that $P(0, y) = \dfrac{1}{6}$ for all $y \neq 0$; $P(x, 0) = \dfrac{1}{2}$ if $x \neq 0$; $P(x, x+1) = \dfrac{1}{2}$ if $x \in \{1, 2, 3, 4, 5\}$; and $P(6, 1) = \dfrac{1}{2}$.

EXAMPLE 28

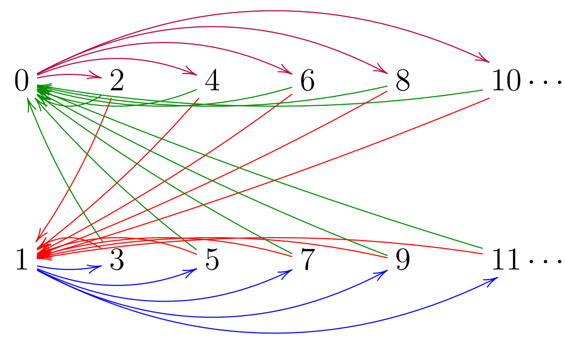
Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S} = \{0, 1, 2, 3, ...\}$ whose transition function is

$$P(0, y) = \begin{cases} 0 & \text{if } y \text{ is odd or } y = 0 \\ \left(\frac{1}{2}\right)^{y/2} & \text{if } y \geq 2 \text{ is even} \end{cases}$$

$$P(1, y) = \begin{cases} 0 & \text{if } y = 1 \text{ or } y \text{ is even} \\ \left(\frac{1}{2}\right)^{(y-1)/2} & \text{if } y \geq 3 \text{ is odd} \end{cases}$$

$$x \geq 2 \Rightarrow P(x, y) = \begin{cases} \frac{1}{2} & \text{if } y = 0 \\ \frac{1}{2} & \text{if } y = 1 \\ 0 & \text{else} \end{cases}$$

Alternate solution:



Define a **factor** of $\{X_t\}$ by grouping states:

8.12 Chapter 8 Homework

Exercise from Section 8.2

1. Suppose we have two boxes and $2d$ marbles, of which d are black and d are red. Initially, d of the balls are placed in Box 1, and the remainder are placed in Box 2. At each trial, a ball is chosen uniformly from each of the boxes; these two balls are put back in the opposite boxes. Let X_0 denote the number of black balls initially in Box 1, and let X_t denote the number of black balls in Box 1 after the t^{th} trial. Find the transition function of the Markov chain $\{X_t\}$.

Exercises from Section 8.3

2. Consider a Markov chain with state space $\mathcal{S} = \{0, 1\}$, where $p = P(0, 1)$ and $q = P(1, 0)$; compute the following quantities in terms of p and q:

   a) $P(X_2 = 0 \,|\, X_1 = 1)$
   b) $P(X_3 = 0 \,|\, X_2 = 0)$
   c) $P(X_2 = 1 \,|\, X_0 = 0)$

3. Continuing with the Markov chain described in Problem 2, suppose the initial distribution is $\pi_0 = (\pi_0(0), \pi_0(1))$. Compute the following quantities in terms of the entries of π_0, p and q:

   a) $P(X_0 = 0 \,|\, X_1 = 0)$
   b) $P(X_1 = 0 \,|\, X_0 = X_2 = 0)$

4. The weather in a city is always one of two types: rainy or dry. If it rains on a given day, then it is 25% likely to rain again on the next day. If it is dry on a given day, then it is 10% likely to rain the next day. If it rains today, what is the probability it rains the day after tomorrow?

5. A Markov chain has state space $\mathcal{S} = \{1, 2, 3, 4, 5\}$ and transition matrix

$$P = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & 0 & \frac{3}{4} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{3}{4} & 0 & \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}.$$

   Assume that the initial distribution is uniform.

a) Sketch the directed graph associated to this Markov chain.
b) Compute the distribution of X_2 (this means compute π_2).
c) Compute $P(X_3 = 5 \mid X_2 = 4)$.
d) Compute $P(X_4 = 2 \mid X_2 = 3)$.
e) Compute $P(X_4 = 5, X_3 = 2, X_1 = 1)$.
f) Compute $P(X_{11} = 1 \mid X_{10} = 4, X_9 = 2, X_8 = 5, X_6 = 1, X_2 = 1)$.
g) Compute $P(X_8 = 3 \mid X_7 = 1 \text{ and } X_9 = 5)$

6. Consider a Markov chain with state space $\{1, 2, 3, 4, 5\}$ and transition matrix

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{2}{7} & 0 & \frac{5}{7} & 0 & 0 \\ 0 & \frac{3}{4} & 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{3} & 0 & \frac{2}{3} \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

a) Compute P^2 and P^3.
b) If the initial distribution is uniform, find the distributions at times $1, 2$ and 3.

7. Consider the Markov chain with $\mathcal{S} = \{1, 2, 3\}$ whose transition matrix is

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1-p & 0 & p \\ 0 & 1 & 0 \end{pmatrix},$$

where $p \in (0, 1)$ is a constant.

a) Compute P^2.
b) Show $P^4 = P^2$.
c) Compute P^n for all $n \geq 1$.
d) If the initial distribution is $\left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right)$, find the time 200 distribution.
e) If the initial distribution is $\left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right)$, find the time 111111 distribution.

8. A dysfunctional family has six members (named Al, Bal, Cal, Dal, Eal, and Fal) who have trouble passing the salt at the dinner table. The family sits around a circular table in clockwise alphabetical order. This family has the following quirks:

- Al is twice as likely to pass the salt to his left than his right.
- Cal and Dal alway pass the salt to their left.
- All other family members pass the salt to their left half the time and to their right half the time.

a) Sketch the directed graph associated to the Markov chain that records the location of the salt after t passes.

b) If Al has the salt now, what is the probability Bal has the salt 3 passes from now?

c) If Al has the salt now, what is the probability that the first time he gets it back is on the 4th pass?

d) If Bal has the salt now, what is the probability that Eal can get it in at most 4 passes?

9. For the Markov chain given in Problem 6, find a distribution π on $\mathcal{S}$ with the property that if the initial distribution is π, then the time 1 distribution is also π.

10. Consider Markov chain with $\mathcal{S} = \{0, 1, 2, ...\}$, where for all $x \in \mathcal{S}$, $P(x, x+1) = \dfrac{1}{2^x}$ and $P(x, 0) = 1 - \dfrac{1}{2^x}$.

    a) Compute $P(X_8 = 9 \,|\, X_7 = 8)$.

    b) Compute $P(X_4 = 7 \,|\, X_2 = 4)$.

    c) Compute $P(X_4 = 7 \,|\, X_2 = 5)$.

    d) Compute $P(X_6 = 4 \,|\, X_0 = 2)$

    e) If the initial distribution π_0 is uniform on $\{0, 1\}$, compute π_2.

Exercises from Section 8.5

11. Consider a Markov chain with state space $\{1, 2, 3\}$ whose transition matrix is

$$\begin{pmatrix} .4 & .4 & .2 \\ .3 & .4 & .3 \\ .2 & .4 & .4 \end{pmatrix}.$$

Find all stationary distributions of this Markov chain.

12. (20 ★ pts) Let $\{X_t\}$ be a Markov chain that has a stationary distribution π. Prove that if $\pi(x) > 0$ and $x \to y$, then $\pi(y) > 0$.

13. Find all stationary distributions of the Markov chain with transition matrix

$$P = \begin{pmatrix} \frac{1}{9} & 0 & \frac{4}{9} & \frac{4}{9} & 0 \\ 0 & 0 & 0 & 0 & 1 \\ \frac{2}{9} & \frac{2}{9} & \frac{2}{9} & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{9} & \frac{8}{9} & 0 \\ 0 & 0 & \frac{7}{9} & \frac{2}{9} & 0 \end{pmatrix}.$$

Hint: The answer is $\pi = \left(\frac{9}{221}, \text{something}, \text{something}, \text{something}, \frac{8}{221}\right)$.

14. Compute all stationary distributions of the Markov chain described in Exercise 1, in the situation where $d = 3$.

15. a) Show that the Markov chain introduced in Exercise 7 has a unique stationary distribution (and compute this stationary distribution, in terms of p).
    b) Is this stationary distribution steady-state? Why or why not?
    Hint: The work you did in Problem 7 should be useful in answering this.

16. A transition matrix of a Markov chain is called **doubly stochastic** if its columns add to 1 (recall that for any transition matrix, the rows must add to 1). Find a stationary distribution of a finite state-space Markov chain with a doubly stochastic transition matrix (the way you do this is by "guessing" the answer, and then showing your guess is stationary).

    NOTE: It is useful to remember the fact you prove in this question.

17. (20 ★ pts) Prove Theorem 8.25 from the notes, which goes like this: let $\pi_1, \pi_2, ...,$ be a finite or countable list of stationary distributions for a Markov chain $\{X_t\}$. Let $\alpha_1, \alpha_2, ...$ be nonnegative numbers whose sum is 1, and let $\pi = \sum_j \alpha_j \pi_j$. Prove that the distribution π is stationary for $\{X_t\}$.

18. (20 ★ pts) Show that for any $d \times d$ stochastic matrix P, 1 is an eigenvalue of P corresponding to eigenvector $(1, 1, 1, ..., 1) \in \mathbb{R}^d$.

    Hint: the crux of this question is to get you to remember what eigenvalues and eigenvectors are (you learned about these creatures in MATH 322).

19. Let

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}.$$

a) Find the eigenvalues of P by solving $\det(P - \lambda I) = 0$.

b) For each eigenvalue you found in part (a), find a corresponding eigenvector (by finding $\mathbf{v} \neq \mathbf{0}$ such that $P\mathbf{v} = \lambda\mathbf{v}$).

c) Diagonalize P (i.e. write $P = S\Lambda S^{-1}$ where Λ is a diagonal matrix whose entries are eigenvalues of P, and S is a matrix whose columns are corresponding eigenvectors of P).

d) Compute P^n (by multiplying out the formula $P^n = S\Lambda^n S^{-1}$).

e) Compute $\lim_{n\to\infty} P^n$.

20. Let $\{X_t\}$ be a Markov chain with state space $\{1,2,3\}$ whose transition matrix is the matrix P given in Exercise 19. Based on your work in Exercise 19, what do you know about stationary and/or steady-state distributions of $\{X_t\}$?

21. Let $\{X_t\}$ be a Markov chain with state space $\{1,2,3,4\}$ whose transition matrix is

$$P = \begin{pmatrix} \frac{1}{7} & \frac{6}{7} & 0 & 0 \\ \frac{11}{14} & \frac{3}{14} & 0 & 0 \\ 0 & 0 & \frac{1}{5} & \frac{4}{5} \\ 0 & 0 & \frac{2}{5} & \frac{3}{5} \end{pmatrix}.$$

a) Find all stationary distributions of $\{X_t\}$.

b) Does $\{X_t\}$ have a steady-state distribution? Explain.

Exercises from Section 8.6

22. Consider a Markov chain with state space $\mathcal{S} = \{0,1\}$, where $p = P(0,1)$ and $q = P(1,0)$. (Assume that neither p nor q are either 0 or 1.) Compute, for each n, the following in terms of p and q:

a) $P_0(T_0 = n)$

Hint: There are two cases: one for $n = 1$, and one for $n > 1$.

b) $P_1(T_0 = n)$

c) $P_0(T_1 = n)$

d) $P_1(T_1 = n)$

23. For the same Markov chain described in Exercise 22, compute these quantities (in terms of p and q):

a) $f_{0,1}$
Hint: Recall that $f_{0,1} = P_0(T_1 < \infty)$. There are two ways to do this: first, you can add up the values of $P_0(T_1 = n)$ from $n = 1$ to ∞; second, you can compute $P_0(T_1 = \infty)$ and use the complement rule.

b) $f_{1,0}$

c) $f_{0,0}$

d) $f_{1,1}$

24. Let $\{X_t\}$ be the Markov chain described in Exercise 6.

a) For each $x \in \mathcal{S}$, compute $P_x(T_1 = 1)$.

b) For each $x \in \mathcal{S}$, compute $P_x(T_1 = 2)$.

c) For each $x \in \mathcal{S}$, compute $P_x(T_1 = 3)$.

25. Consider a Markov chain whose state space is $\mathcal{S} = \{1, 2, 3, 4, 5, 6, 7\}$ and whose transition matrix is

$$\begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{8} & \frac{1}{4} & \frac{1}{8} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

a) List all the closed subsets of this chain.

b) List all the communicating classes of this chain.

c) Write the period of each state that belongs to a communicating class.

26. Let $p \in (0, 1)$ be a constant. Consider a Markov chain with state space $\mathcal{S} = \{0, 1, 2, 3, ...\}$ where $P(x, x+1) = p$ for all $x \in \mathcal{S}$ and $P(x, 0) = 1 - p$ for all $x \in \mathcal{S}$. Explain why this chain is irreducible by showing, for arbitrary states x and y, a sequence of steps which could be followed to get from x to y.

Exercises from Section 8.7

27. For the Markov chain introduced in Exercise 2, compute $E_0(V_{1,3})$.

28. For the Markov chain given in Exercise 25:

a) Determine which states are recurrent and which states are transient.

b) Compute $f_{x,y}$ for all $x, y \in \mathcal{S}$.

c) Compute $E_1(V_1)$.

29. Determine whether the Markov chain described in Exercise 26 is recurrent or transient.

    Hint: Compute f_0 directly by adding up the values of $P_0(T_0 = n)$.

30. Consider a Markov chain with state space $\mathcal{S} = \{0, 1, 2, 3, ...\}$ and transition function defined by

$$P(x,y) = \begin{cases} \frac{1}{2} & \text{if } x = y \\ \frac{1}{2} & \text{if } x > 0 \text{ and } y = x - 1 \\ \left(\frac{1}{2}\right)^{y+1} & \text{if } x = 0 \text{ and } y > 0 \\ 0 & \text{otherwise} \end{cases}.$$

a) Explain why this Markov chain is irreducible.

b) Is this chain recurrent or transient?

31. Consider a Markov chain with state space $\mathcal{S} = \{0, 1, 2, 3, ...\}$ and transition function defined by

$$P(x,y) = \begin{cases} \frac{1}{7} & \text{if } y = 0 \\ \frac{2}{7} & \text{if } y \in \{x+2, x+4, x+6\} \\ 0 & \text{otherwise} \end{cases}.$$

    Classify the states of this Markov chain as recurrent or transient, and find all communicating classes (if any).

32. Consider a Markov chain with state space $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ whose transition matrix is

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{8} & 0 & \frac{7}{8} & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & \frac{3}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{1}{5} & 0 & \frac{1}{5} & \frac{1}{5} & \frac{2}{5} \end{pmatrix}$$

a) Determine which states are recurrent and which states are transient.

b) Compute $f_{x,1}$ for all $x \in \mathcal{S}$.

c) Compute $E_x(V_y)$ for all $x, y \in \mathcal{S}_T$.

33. Compute the stationary distribution of the Ehrenfest chain (introduced in one of the group presentations), in the situation where $d = 4$.

34. a) What is the period of the Ehrenfest chain?

    b) What is the period of the Markov chain introduced in Problem 1?

35. Let $\{X_t\}$ be the Wright-Fisher chain (introduced in one of the group presentations) with $d = 3$. Compute $f_{x,0}$ for all $x \in \mathcal{S}$.

36. Let $\{X_t\}$ be the Wright-Fisher chain with $d = 4$. Compute $E_1(V_2)$.

37. Let $\{X_t\}$ be a Galton-Watson branching chain where each individual has either 0 or 3 offspring, each with probability $\frac{1}{2}$. Compute the extinction probability η.

38. Let $\{X_t\}$ be a Galton-Watson branching chain where the number of offspring of each individual is $Geom(p)$. Compute the extinction probability η.

    Hint: There are two cases, depending on p.

39. Let X_t denote the number of people waiting for service at a fast-food restaurant at time t. Assume $\{X_t\}$ is modeled by a discrete queuing chain where with probability $\dfrac{2}{3}$, two customers enter the queue in each time period, and with probability $\dfrac{1}{3}$, no customers enter the queue in each time period.

    a) If there is initially 1 person being served, what is the probability that at some point in the future, there will be no one in line?

    b) If there are initially 4 people in the queue, what is the probability that the queue never empties?

Exercises from Section 8.8

40. Compute the Cesàro limit of the sequence of numbers $\{0, 1, 0, 1, 0, 1, ...\}$ (justify your answer).

41. Compute (directly, without appealing to any stationary distribution), in terms of p and q, the mean return time to each state for the Markov chain given in Problem 2.

42. For the Markov chain introduced in Exercise 32:

a) Compute $\lim_{n\to\infty} \frac{E_1(V_{2,n})}{n}$.

b) Compute $\lim_{n\to\infty} \frac{E_6(V_{3,n})}{n}$.

c) Compute $\lim_{n\to\infty} \frac{E_6(V_{1,n})}{n}$.

d) Compute $\lim_{n\to\infty} \frac{E_3(V_{1,n})}{n}$.

43. Determine whether the chain introduced in Exercise 26 is positive recurrent or null recurrent.

Hint: Compute the mean return time to state 0 by determining, for each n, $P_0(T_0 = n)$ and then using LOTUS to compute $m_0 = E_0(T_0) = \sum_{n=1}^{\infty} nP_0(T_0 = n)$.

Exercises from Section 8.9

44. Compute all the stationary distributions of the Markov chain introduced in Exercise 32.

45. Fix nonnegative constants $p_0, p_1, ...$ such that $\sum_{y=0}^{\infty} p_y = 1$ and let X_t be a Markov chain on $\mathcal{S} = \{0, 1, 2, \cdots\}$ with transition function P defined by

$$P(x,y) = \begin{cases} p_y & \text{if } x = 0 \\ 1 & \text{if } x > 0, y = x - 1 \\ 0 & \text{else} \end{cases}$$

a) Show this chain is recurrent.

b) Calculate, in terms of the p_y, the mean return time to 0.

c) Under what conditions on the p_y is the chain positive recurrent?

d) Suppose this chain is positive recurrent. Find $\pi(0)$, the value that stationary distribution assigns to state 0.

e) Suppose this chain is positive recurrent. Find the value the stationary distribution π assigns to an arbitrary state x.

Exercises from Section 8.10

46. (40 ★ pts) Complete the proof of Theorem 8.72 by explaining why $P(T < \infty) = 1$.

Hints: To do this,

a) Define a new Markov chain (X_t, Y_t) where the two coordinates act independently. In particular, what is the state space of this chain, and what is its transition function?

b) Prove that the chain (X_t, Y_t) is irreducible, meaning that you can get from any state (a, b) to any other state (x, y) in a finite amount of time. To do this, you will need to apply Theorem 8.39.

c) The event $T < \infty$ corresponds to the chain (X_t, Y_t) hitting a certain state (which one?) in a finite amount of time. By irreducibility of (X_t, Y_t), this probability is 1, which allows you to finish the proof of the claim.

Exercises from Section 8.11

47. Consider the irreducible Markov chain with state space $\mathcal{S} = \{1, 2, 3, 4, 5\}$ whose transition matrix is

$$\begin{pmatrix} 0 & \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} & \frac{3}{4} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

a) Compute the period of this Markov chain.

b) Compute the stationary distribution. Is this distribution steady-state?

c) Describe P^n for n large (there is more than one answer depending on the relationship n and the period d).

d) Suppose the initial distribution is uniform on $\mathcal{S}$. Estimate the time n distribution for large n (there are cases depending on the value of n).

e) Compute $\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} P^k$.

f) Compute m_1 and m_2.

48. Let $\{X_t\}$ be the Ehrenfest chain with $d = 4$ and $X_0 = 0$ (i.e. there are no particles in the left-hand chamber).

a) Estimate the distribution of X_t when t is large and even.

b) Estimate the distribution of X_t when t is large and odd.

c) Compute the expected amount of time until there are again no particles in the left-hand chamber.

49. Consider a Markov chain on $\mathcal{S} = \{0, 1, 2, 3\}$ with transition matrix

$$P = \begin{pmatrix} 0 & \frac{1}{3} & 0 & \frac{2}{3} \\ \frac{1}{5} & 0 & \frac{4}{5} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{2}{3} \\ \frac{1}{5} & 0 & \frac{4}{5} & 0 \end{pmatrix}.$$

   a) Compute the Cesàro limit of P^n.
   b) Compute m_0 and m_2.

50. Consider a Markov chain $\{X_t\}$ on $\mathcal{S} = \{0, 1, 2, ...\}$ with transition function

$$P(x,y) = \begin{cases} 2^{-y-1} & \text{if } x \leq 3 \\ 1/4 & \text{if } x > 3 \text{ and } y \leq 3 \\ 0 & \text{if } x > 3 \text{ and } y > 3 \end{cases}$$

   a) Show the chain is positive recurrent.
   Hint: Consider a factor $\{Y_t\}$ defined by $Y_t = X_t$ if $X_t \leq 3$ and $Y_t = 4$ if $X_t \geq 4$. Show $\{Y_t\}$ is positive recurrent; why does this imply $\{X_t\}$ is positive recurrent?
   b) Find all stationary distributions of $\{X_t\}$.
   Hint: The stationary distribution of Y_t (from part (a)) tells you something about the stationary distribution of X_t.
   c) Suppose you start in state 2. How long would you expect it to take for you to return to state 2 for the fifth time?

51. (20 ★ pts) Suppose a fair die is thrown repeatedly. Let S_n represent the sum of the first n throws. Compute

$$\lim_{n\to\infty} P(S_n \text{ is a multiple of } 13),$$

justifying your reasoning.

52. (30 ★ pts) Your professor owns 3 umbrellas, which at any time may be in his office or at his home. If it is raining when he travels between his home and office, he carries an umbrella (if possible) to keep him from getting wet.

   a) If on every one of his trips, the probability that it is raining is p, what is the long-term proportion of journeys on which he gets wet?
   b) What p as in part (a) causes the professor to get wet most often?

c) In the worst-case scenario described in part (b), on what fraction of his trips will he get wet?

53. (30 ★ pts) A knight is placed in one corner of a chess board. At each step, the knight chooses a square uniformly from the squares that the knight can legally move to (i.e. two squares in one direction, and one to the side). Compute the expected number of moves the knight will make before returning to its starting position.

Chapter 9

Continuous-time Markov chains

9.1 Introducing CTMCs

Goal: study analogues of Markov chains where time is measured *continuously* rather than *discretely*. (The state space $\mathcal{S}$ will remain finite or countable.)

FIRST QUESTION

What "should" a continuous-time Markov chain look like?

	(DISCRETE-TIME) MARKOV CHAIN	**CTMC** (CONTINUOUS-TIME MARKOV CHAIN)
state space $\mathcal{S}$	finite or countable; usually $\mathcal{S} = \{0, 1, ..., d\}$ or $\mathcal{S} = \{0, 1, 2, ...\}$ or $\mathcal{S} = \mathbb{Z}$.	finite or countable; usually $\mathcal{S} \subseteq \mathbb{Z}$ (same)
index set $\mathcal{I}$	X_t = state at time t $t \in \{0, 1, 2, ...\}$ or $t \in \mathbb{Z}$	X_t = state at time t $t \in [0, \infty)$ or $t \in \mathbb{R}$
initial distribution	$\pi_0 : \mathcal{S} \to [0, 1]$; $\sum_{x \in \mathcal{S}} \pi_0(x) = 1$ $\pi_0(x) = P(X_0 = x)$	$\pi_0 : \mathcal{S} \to [0, 1]$; $\sum_{x \in \mathcal{S}} \pi_0(x) = 1$ $\pi_0(x) = P(X_0 = x)$ (same)

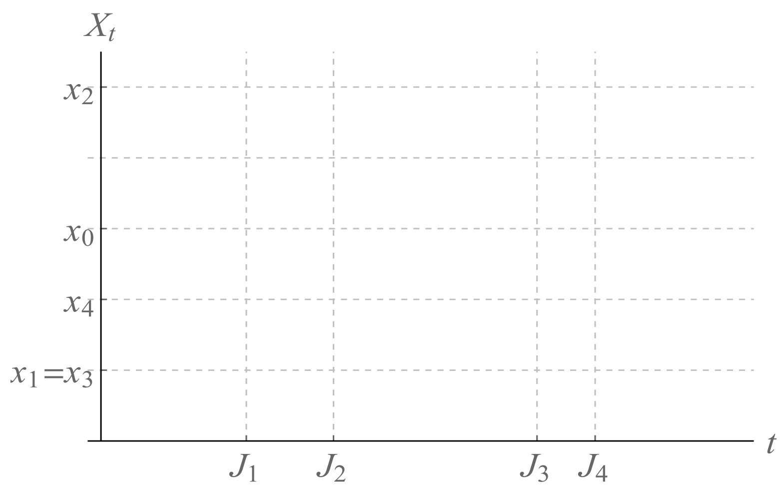
(continued on next page)

	MARKOV CHAIN	CTMC
transition probabilities	we specify time 1 transitions: $P : \mathcal{S} \times \mathcal{S} \to [0, 1]$ $\sum_{y \in \mathcal{S}} P(x, y) = 1 \,\forall\, x \in \mathcal{S}$ $P(x, y) = P(X_{t+1} = y \| X_t = x)$ (we assume these are $\perp$ of t) If $\mathcal{S}$ is finite, write P as a matrix: $P(x, y) \leftrightarrow P_{x,y} = P_{xy}$ From the time 1 transitions, we calculate transition probabilities for any time n: $P^n(x, y) = P(X_{t+n} = y \| X_t = x)$ $= \sum_{z \in \mathcal{S}} P(x, z) P^{n-1}(z, y)$ If $\mathcal{S}$ finite, $P^n(x, y) = (P^n)_{xy}$.	
Markov property	$P(X_t = x_t \| X_0 = x_0, ..., X_{t-1} = x_{t-1})$ $= P(X_t = x_t \| X_{t-1} = x_{t-1})$ $= P(x_{t-1}, x_t)$	

Definition 9.1 *A* **jump process** $\{X_t : t \in \mathcal{I}\}$ *is a stochastic process with index set* $\mathcal{I} = [0, \infty)$ *or* $\mathbb{R}$ *and finite or countable state space* $\mathcal{S}$ *such that with probability 1, the functions* $t \mapsto X_t$ *(these functions are called* **sample functions** *of the process) are right-continuous and piecewise constant.*

That is, there exist times $J_1 < J_2 < J_3 < \ldots$ (these are r.v.s, not constants) and states $x_0, x_1, x_2, \ldots \in \mathcal{S}$ such that

$$X_t = \begin{cases} x_0 & \text{if } 0 \le t < J_1 \\ x_1 & \text{if } J_1 \le t < J_2 \\ x_2 & \text{if } J_2 \le t < J_3 \\ \vdots & \quad \vdots \end{cases}$$



The assumption that the sample functions are right-continuous is necessary for technical reasons (we'll see one of these reasons later).

Definition 9.2 *A* **continuous-time Markov chain (CTMC)** $\{X_t\}$ *is a jump process satisfying the Markov property .*

9.2 General theory of CTMCs

QUESTIONS

1. What properties must the transition functions $P_{xy}(t)$ of a CTMC have?

   (Must they be continuous? Differentiable? Increasing? Decreasing? Do they have a limit as $t \to \infty$? Do they have to go through certain points? Must their formulas be a certain type? Etc.)

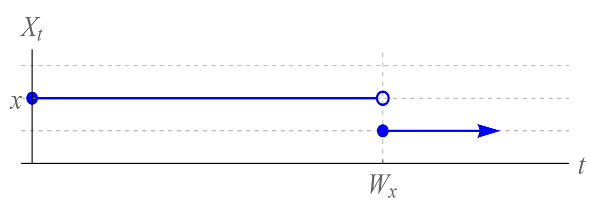
2. What information is really necessary to describe a CTMC?

   (To describe a Markov chain, we only need to write down time 1 transitions–they generate all the time n transitions. For a CTMC, is there something we can write down that sufficiently "generates" all the time t transitions $P_{xy}(t)$?)

Waiting times and holding rates

Definition 9.3 *Let $\{X_t\}$ be a CTMC with state space $\mathcal{S}$. For every $x \in \mathcal{S}$, define:*

$$\begin{aligned} W_x &= \text{ the \textbf{waiting time} at state } x \\ &= \text{ the time until the first jump, if the chain starts at } x \\ &= \min\{t \geq 0 : X_t \neq x, \text{ given that } X_0 = x\} \\ &= X_{J_1} | X_0 = x. \end{aligned}$$



OBSERVATIONS ABOUT WAITING TIMES

1. Waiting times are well-defined because of the assumption that the sample functions are right-continuous.

2. By time homogeneity, the waiting times W_x depend only on state x (and not on exactly what interval of time they are taking place).

3. By the Markov property, the waiting times W_x are ____________________ ,

   and since W_x is continuous, this means each $W_x \sim$ __________________ .

This allows us to define:

Definition 9.4 *Let $\{X_t\}$ be a CTMC with state space $\mathcal{S}$. For every $x \in \mathcal{S}$, define:*

$$\begin{aligned} q_x &= \textit{the } \textbf{holding rate} \textit{ of state } x \\ &= \textit{the parameter of the exponential r.v. } W_x \\ &= \frac{1}{E_x(W_x)}. \end{aligned}$$

Note: If the holding rate of a state is large, you expect to stay in the state for a

____________________ amount of time before jumping.

EXAMPLE 1

Let $\{X_t\}$ be a CTMC. Suppose that the holding rate of state 3 is 4. What is the probability that $X_t = 3$ for all $t \leq 2$?

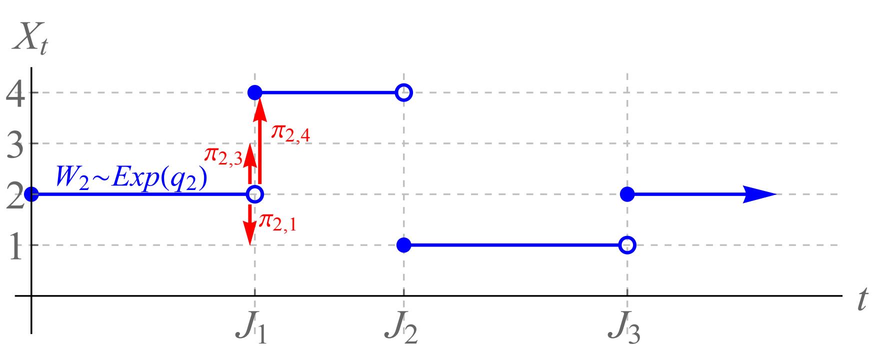
Jump probabilities

Definition 9.5 *Let $\{X_t\}$ be a CTMC. Let $x, y \in \mathcal{S}$. Define*

$$\begin{aligned} \pi_{x,y} &= \textit{the } \textbf{jump probability} \textit{ from } x \textit{ to } y \\ &= \textit{the probability that when the chain first jumps from state } x\textit{, it jumps to state } y \\ &= P_x\left(X_{W_x} = y\right). \end{aligned}$$

Properties of jump probabilities immediate from the definition:

$$\pi_{x,y} \geq 0 \qquad \pi_{x,x} = 0 \qquad \sum_{y \in \mathcal{S}} \pi_{x,y} = 1.$$



Properties of transition functions

Definition 9.6 *Given two quantities x and y, let $\delta_{x,y} = \delta_{xy}$ be the* **Kronecker delta**, *i.e.*

$$\delta_{x,y} = \delta_{xy} = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{else} \end{cases}.$$

Theorem 9.7 (Properties of transition functions) *Let $\{X_t\}$ be a CTMC, and let $x, y \in \mathcal{S}$.*

1. $P_{xy}(0) = \delta_{xy}$.
2. *For all t, the transition functions $P_{x,y}(t)$ satisfy the* **integral equation**
$$P_{x,y}(t) = \delta_{x,y}e^{-q_x t} + \int_0^t q_x e^{-q_x s}\left[\sum_{z\in\mathcal{S}} \pi_{x,z}P_{z,y}(t-s)\right] ds.$$
3. P_{xy} *is a* **continuous** *function of t.*
4. P_{xy} *is a* **differentiable** *function of t, and*
$$P'_{x,y}(t) = -q_x P_{x,y}(t) + q_x \sum_{z\in\mathcal{S}} \pi_{x,z}P_{z,y}(t).$$

PROOF Statement (1) is obvious. Next, statement (2):

$$\begin{aligned}
P_{x,y}(t) &= P_x(X_t = y) \\
&= P_x(X_t = y \cap W_x > t) + P_x(X_t = y \cap W_x \le t) \\
&= P_x(X_t = y \,|\, W_x > t)P(W_x > t) + P_x(X_t = y \cap W_x \le t) \\
&= \delta_{x,y}e^{-q_x t} + \int_0^t P(X_t = y \,|\, W_x = s) f_{W_x}(s)\, ds
\end{aligned}$$

(Law of Total Probability, continuous version)

Now take the conditional probability inside the integral and divvy it up based on the location of the first jump:

$$\begin{aligned}
P_{x,y}(t) &= \delta_{x,y}e^{-q_x t} + \int_0^t f_{W_x}(s) \sum_{z\in\mathcal{S}} P(X_s = z \cap X_t = y \,|\, W_x = s)\, ds \qquad (\diamondsuit) \\
&= \delta_{x,y}e^{-q_x t} + \int_0^t q_x e^{-q_x s}\left[\sum_{z\in\mathcal{S}} \pi_{x,z}P_{z,y}(t-s)\right] ds.
\end{aligned}$$

This finishes the proof of the integral equation.

Next, statement (3). In the integral equation, use the u-sub $u = t - s$, $du = -ds$:

$$\begin{aligned} P_{x,y}(t) &= \delta_{x,y} e^{-q_x t} + \int_0^t q_x e^{-q_x s} \left[\sum_{z \in \mathcal{S}} \pi_{x,z} P_{z,y}(t-s) \right] ds \\ &= \delta_{xy} e^{-q_x t} + - \int_t^0 q_x e^{-q_x (t-u)} \left[\sum_{z \in \mathcal{S}} \pi_{x,z} P_{z,y}(u) \right] du \\ &= \delta_{xy} e^{-q_x t} + q_x e^{-q_x t} \int_0^t e^{q_x u} \left[\sum_{z \in \mathcal{S}} \pi_{x,z} P_{z,y}(u) \right] du \qquad (\heartsuit) \end{aligned}$$

Last, we prove statement (4). By (3), the integrand of the integral in $(\heartsuit)$ is cts. Therefore

$$P_{x,y}(t) = \delta_{xy} e^{-q_x t} + q_x e^{-q_x t} \int_0^t e^{q_x u} \left[\sum_{z \in \mathcal{S}} \pi_{x,z} P_{z,y}(u) \right] du \qquad (\spadesuit)$$

Thus $P_{x,y}$ is differentiable. Finally, we compute the derivative of $P_{x,y}$:

$$\begin{aligned} P'_{x,y}(t) &= \frac{d}{dt} \left[\delta_{xy} e^{-q_x t} + q_x e^{-q_x t} \int_0^t e^{q_x u} \left[\sum_{z \in \mathcal{S}} \pi_{x,z} P_{z,y}(u) \right] du \right] \\ &= \frac{d}{dt} \left[e^{-q_x t} \left(\delta_{xy} + q_x \int_0^t e^{q_x u} \left[\sum_{z \in \mathcal{S}} \pi_{x,z} P_{z,y}(u) \right] du \right) \right]. \end{aligned}$$

Now use the Product Rule:

$$\begin{aligned} &P'_{x,y}(t) \\ &= (\text{first})' \cdot (\text{second}) + (\text{second})' \cdot (\text{first}) \\ &= \left(-q_x e^{-q_x t} \right) \left(\delta_{xy} + q_x \int_0^t e^{q_x u} \left[\sum_{z \in \mathcal{S}} \pi_{x,z} P_{z,y}(u) \right] du \right) \\ &\qquad\qquad + \left(q_x e^{q_x t} \sum_{z \in \mathcal{S}} \pi_{x,z} P_{z,y}(t) \right) \left(e^{-q_x t} \right) \\ &= -q_x \left[e^{-q_x t} \left(\delta_{xy} + q_x \int_0^t e^{q_x u} \left[\sum_{z \in \mathcal{S}} \pi_{x,z} P_{z,y}(u) \right] du \right) \right] + q_x \sum_{z \in \mathcal{S}} \pi_{x,z} P_{z,y}(t) \\ &= -q_x P_{x,y}(t) + q_x \sum_{z \in \mathcal{S}} \pi_{x,z} P_{z,y}(t). \qquad \text{(by } (\spadesuit)\text{)} \end{aligned}$$

This finishes the proof. □

Infinitesimal parameters

It will be convenient to give the values of $P'_{x,y}(0)$ their own names:

Definition 9.8 *Let $\{X_t\}$ be a CTMC. For any $x, y \in \mathcal{S}$, define the* **infinitesimal parameters** *a.k.a.* **generating parameters** *$q_{xy} = q_{x,y}$ to be $q_{xy} = P'_{x,y}(0)$.*

Corollary 9.9 *Let $\{X_t\}$ be a CTMC. Then for any $x, y \in \mathcal{S}$,*

$$q_{xy} = P'_{x,y}(0) = \begin{cases} -q_x & \text{if } x = y \\ q_x \pi_{x,y} & \text{if } x \neq y \end{cases}$$

PROOF Set $t = 0$ in the last statement of Theorem 9.7:

$$\begin{aligned} P'_{x,y}(t) &= -q_x P_{x,y}(t) + q_x \sum_{z \in \mathcal{S}} \pi_{x,z} P_{z,y}(t) \\ P'_{x,y}(0) &= -q_x P_{x,y}(0) + q_x \sum_{z \in \mathcal{S}} \pi_{x,z} P_{z,y}(0) \\ &= -q_x \delta_{xy} + q_x \left[0 + 0 + \dots + 0 + \pi_{x,y} \cdot 1 + 0 + \dots + 0\right] \\ &= -q_x \delta_{xy} + q_x \pi_{x,y} \\ &= \begin{cases} -q_x & \text{if } x = y \\ q_x \pi_{x,y} & \text{if } x \neq y \end{cases} . \square \end{aligned}$$

Note: $q_{xx} \leq 0$ for all x, and if $x \neq y$ then $q_{xy} \geq 0$.

Theorem 9.10 *Let $\{X_t\}$ be a CTMC and let $x \in \mathcal{S}$. Then*

$$\sum_{y \in \mathcal{S}} q_{xy} = 0.$$

PROOF

$$\sum_{y \in \mathcal{S}} q_{xy} = q_{xx} + \sum_{y \neq x} q_{xy} = -q_x + \sum_{y \neq x} q_x \pi_{x,y} = -q_x + q_x \sum_{y \neq x} \pi_{x,y} = -q_x + q_x(1) = 0. \square$$

Why are the q_{xy} called infinitesimal parameters? If t is very small (i.e. infinitesimally small), then by linear approximation (Calculus 1) we have

$$P_{x,y}(t) \approx P_{x,y}(0) + P'_{x,y}(0)t = \delta_{x,y} + q_{xy}t.$$

So these parameters are measuring the infinitesimal rate of change in $P_{x,y}(t)$ when t is small.

EXAMPLE 2

Suppose $\{X_t\}$ is a CTMC with state space $\{1, 2, 3, 4\}$ such that $q_{12} = 3$, $q_{13} = 2$ and $q_{14} = 7$.

1. Compute q_{11}.
2. Use linear approximation to estimate $P_{12}(.08)$.
3. Use linear approximation to estimate $P_{11}(.3)$.
4. Compute the jump probabilities π_{13} and π_{14}.

Backward and forward equations

We are ready to derive two sets of differential equations (actually initial value problems) which the transition functions of a CTMC must satisfy:

Theorem 9.11 (Backward equation) *Let $\{X_t\}$ be a CTMC. Then for all $x, y \in \mathcal{S}$,*

$$P'_{x,y}(t) = \sum_{z \in \mathcal{S}} q_{x,z} P_{z,y}(t) \quad \text{and} \quad P_{x,y}(0) = \delta_{xy}.$$

PROOF From Theorem 9.7,

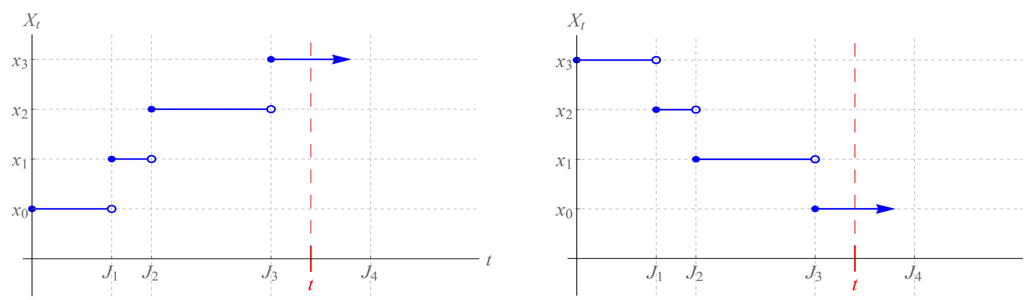
$$\begin{aligned} P'_{x,y}(t) &= -q_x P_{x,y}(t) + q_x \sum_{z \in \mathcal{S}} \pi_{x,z} P_{z,y}(t) \\ &= q_{xx} P_{x,y}(t) + \sum_{z \neq x \in \mathcal{S}} q_x \pi_{x,z} P_{z,y}(t) \\ &= q_{xx} P_{x,y}(t) + \sum_{z \neq x \in \mathcal{S}} q_{x,z} P_{z,y}(t) \\ &= \sum_{z \in \mathcal{S}} q_{x,z} P_{z,y}(t). \ \square \end{aligned}$$

The second set of ODE's the transition functions satisfy is the "forward equation" given in Theorem 9.13. Deriving this equation follows the same line of argument as what we just went through over the last few pages, but instead of conditioning on the first jump in the process (back in line ($\diamondsuit$) of the proof of Theorem 9.7), you condition on the last jump before time t. To make this argument go through, we first need this lemma:

Lemma 9.12 (State reversal identity) *Let* $\{X_t\}$ *be a CTMC. Then*

$$\begin{aligned} &q_{x_n} P\left(J_n \leq t < J_{n+1} \,|\, X_0 = x_0, X_{J_1} = x_1, X_{J_2} = x_2, ..., X_{J_n} = x_n\right) \\ &= q_{x_0} P\left(J_n \leq t < J_{n+1} \,|\, X_0 = x_n, X_{J_1} = x_{n-1}, X_{J_2} = x_{n-2}, ..., X_{J_n} = x_0\right). \end{aligned}$$

What this lemma says: Here's a picture when $n = 3$:



PROOF The event $J_n \leq t < J_{n+1}$ corresponds exactly to

$$J_n = W_{x_0} + W_{x_1} + ... + W_{x_{n-1}} + W_{x_n} > t,$$

i.e.

$$W_{x_n} > t - W_{x_0} - W_{x_1} - ... - W_{x_{n-1}}.$$

Since $W_{x_n} \sim Exp(q_{x_n})$, given values $s_0, ..., s_{n-1}$ of $W_{x_0}, ..., W_{x_{n-1}}$, the probability of this is

$$e^{-q_{x_n}(t - s_0 - s_1 - ... - s_{n-1})} = \exp\left[-q_{x_n}\left(t - \sum_{k=0}^{n-1} s_k\right)\right].$$

So by the continuous LTP, the conditional probability of $J_n \leq t < J_{n+1}$ given $X_{J_j} = x_j$ for $j \in \{1, ..., n\}$ is therefore

$$\int \cdots \int_{\Delta} \exp\left[-q_{x_n}\left(t - \sum_{k=0}^{n-1} s_k\right)\right] f_{W_{x_0}, W_{x_1}, ..., W_{x_n}}(s_0, s_1, ..., s_{n-1})\, dV$$

and since $W_{x_1}, ..., W_{x_n}$ are independent, this is

$$\int \cdots \int_\Delta \exp\left[-q_{x_n}\left(t - \sum_{k=0}^{n-1} s_k\right)\right] f_{W_{x_0}}(s_0) f_{W_{x_1}}(s_1) \cdots f_{W_{x_{n-1}}}(s_{n-1})\, dV$$

and since $W_{x_k} \sim Exp(q_{x_k})$, this is

$$\int \cdots \int_\Delta \exp\left[-q_{x_n}\left(t - \sum_{k=0}^{n-1} s_k\right)\right] \prod_{k=0}^{n-1} q_{x_k} e^{-q_{x_k} s_k}\, dV \tag{9.1}$$

where Δ is the set of $(s_1, ..., s_n) \in \mathbb{R}^n$ with $s_j \geq 0$.

In this last integral, perform a change of variables (with Jacobians) from the variables $(s_0, ..., s_{n-1})$ to $(u_0, ..., u_{n-1})$ by setting $u_0 = t - s_0 - s_1 - ... - s_{n-1}$, $u_1 = s_{n-1}$, $u_2 = s_{n-2}$, $u_3 = s_{n-3}$, ..., $u_{n-1} = s_1$ in (9.1) above to rewrite it as

$$\int \cdots \int_\Delta e^{-q_{x_0} u_0} \prod_{k=0}^{n-1} q_{x_{n-k}} e^{-q_{x_{n-k}} u_k}\, dV; \tag{9.2}$$

this gives the conditional probability in the second expression in the lemma.

Notice that in (9.1), the integral has $q_{x_0}, ..., q_{x_{n-1}}$ but in (9.2), the integral has $q_{x_n}, ..., q_{x_n}$. So multiplying (9.1) by q_{x_n} is equal to what you get when you multiply (9.2) by q_{x_1}, proving the lemma. □

Theorem 9.13 (Forward equation) *Let $\{X_t\}$ be a CTMC. Then for all $x, y \in \mathcal{S}$,*

$$P'_{x,y}(t) = \sum_{z \in \mathcal{S}} P_{x,z}(t) q_{zy} \text{ and } P_{x,y}(0) = \delta_{xy}.$$

PROOF HW (starred problem)

Summary so far

If $\{X_t\}$ is a CTMC, then:

- The transition functions $P_{x,y}(t)$ must be differentiable, and must satisfy two (systems of) differential equations:

  Backward equation: $\begin{cases} P'_{x,y}(t) = \sum_{z \in \mathcal{S}} q_{xz} P_{z,y}(t) \\ P_{x,y}(0) = \delta_{xy} \end{cases}$

  Forward equation: $\begin{cases} P'_{x,y}(t) = \sum_{z \in \mathcal{S}} P_{x,z}(t) q_{zy} \\ P_{x,y}(0) = \delta_{xy} \end{cases}$

- The numbers $q_{xy} = P'_{x,y}(0)$ are called the **infinitesimal parameters** of the CTMC.
  - For small t, we can estimate $P_{x,y}(t)$ by $P_{x,y}(t) \approx \delta_{xy} + q_{xy}(t)$.
  - The infinitesimal parameters satisfy:

    $$q_{xx} \leq 0; \qquad q_{xy} \geq 0 \text{ if } x \neq y; \qquad \sum_{y \in \mathcal{S}} q_{xy} = 0.$$

  - The **holding rate** q_x of state x is $q_x = -q_{xx}$.
    The **waiting time** in state x (until the next jump) is an $Exp(q_x)$ r.v.

- The **jump probabilities** π_{xy} satisfy

  $$\pi_{xx} = 0; \qquad \pi_{xy} = \frac{q_{xy}}{q_x} = -\frac{q_{xy}}{q_{xx}} \text{ if } x \neq y.$$

The infinitesimal parameters **generate** all the other information about the CTMC:

- You can compute the jump probabilities and the holding rates directly from the q_{xy};

- You can write down a system of differential equations that (hypothetically / theoretically) can be solved to produce formulas for the transition functions $P_{x,y}(t)$.

  In practice, these differential equations can be hard to solve unless the CTMC is "nice" (it has a finite state space or is otherwise not too complicated). We'll talk about how to analyze these nice situations later; if the situation isn't nice, in a worst-case scenario you can estimate the values of $P_{xy}(t)$ using a numerical procedure like Euler's method (MATH 330).

9.3 CTMCs with finite state space

In this section we discuss additional machinery to help us analyze CTMCs when the state space $\mathcal{S}$ is finite. As with discrete-time Markov chains, we will use matrices and linear algebra to help us keep track of the relevant information:

Definition 9.14 *Let $\{X_t\}$ be a CTMC with finite state space. For each t, set $P_{xy}(t) = P(X_{s+t} = y \,|\, X_s = x)$ (we assume that $\{X_t\}$ is* **time homogeneous** *so that these probabilities do not depend on s). Then let*

$$P(t) = \begin{pmatrix} P_{11}(t) & \cdots & P_{1d}(t) \\ \vdots & \ddots & \vdots \\ P_{d1}(t) & \cdots & P_{dd}(t) \end{pmatrix};$$

$P(t)$ is called the **time t transition function** *or* **time t transition matrix** *of the CTMC.*

WHAT THE THEORY FROM SECTION 9.2 TELLS US

-
-
-
-

Definition 9.15 *Let $\{X_t\}$ be a CTMC with finite state space $\mathcal{S}$. Then the matrix $Q = P'(0)$ is called the* **infinitesimal matrix** *or the* **generating matrix** *of the CTMC.*

$$Q = \begin{pmatrix} q_{11} & \cdots & q_{1d} \\ \vdots & \ddots & \vdots \\ q_{d1} & \cdots & q_{dd} \end{pmatrix} = \begin{pmatrix} P'_{11}(0) & \cdots & P'_{1d}(0) \\ \vdots & \ddots & \vdots \\ P'_{d1}(0) & \cdots & P'_{dd}(0) \end{pmatrix}.$$

WHAT THE THEORY FROM SECTION 9.2 TELLS US

-
-

Example 3

Suppose the transition matrix of some CTMC $\{X_t\}$ with state space $\{0,1\}$ is

$$P(t) = \begin{pmatrix} \frac{2}{5} + \frac{3}{5}e^{-5t} & \frac{3}{5} - \frac{3}{5}e^{-5t} \\ \frac{2}{5} - \frac{2}{5}e^{-5t} & \frac{3}{5} + \frac{2}{5}e^{-5t} \end{pmatrix}.$$

1. Compute the infinitesimal matrix of $\{X_t\}$.
2. Describe the waiting time to state 1 as a common r.v., giving its parameter(s).

Definition 9.16 *Let $\{X_t\}$ be a CTMC with finite state space $\mathcal{S}$. The* **jump matrix** *of the CTMC is the matrix Π whose entries are the jump probabilities, i.e.*

$$\Pi = \begin{pmatrix} \pi_{1,1} & \cdots & \pi_{1,d} \\ \vdots & \ddots & \vdots \\ \pi_{d,1} & \cdots & \pi_{d,d} \end{pmatrix}.$$

The **jump chain** *of $\{X_t\}$ is the discrete-time Markov chain $\{X_t^{jump}\}$ with initial distribution π_0 and transition matrix Π.*

What the theory from Section 9.2 tells us

•

EXAMPLE 4

Suppose the infinitesimal matrix of some CTMC $\{X_t\}$ with state space $\{1, 2, 3\}$ is

$$Q = \begin{pmatrix} -3 & 2 & 1 \\ 4 & -6 & 2 \\ 0 & 7 & -7 \end{pmatrix}.$$

1. Describe the waiting times for each state. In which state, on the average, would you expect to stay for the longest times before jumping?
2. Compute the jump matrix of the CTMC.

EXAMPLE 5

Consider a CTMC with state space $\{1, 2, 3\}$ and infinitesimal matrix

$$Q = \begin{pmatrix} -2 & 1 & 1 \\ 1 & -5 & 4 \\ 2 & 1 & -3 \end{pmatrix}.$$

1. Sketch the directed graph of this CTMC.
2. Compute the jump matrix of this CTMC.
3. Suppose you start in state 1. What is the probability you stay in state 1 for at most three units of time before jumping?
4. What is the probability that the first three jumps are from state 1 to state 3, then state 3 to state 2, then state 2 to state 3 (given that you start in state 1)?

Interpreting the forward and backward equations

To compute Q from $P(t)$, we have seen that we use the formula

To go the other way (i.e. compute $P(t)$ from Q), what we know so far is that we can set up some differential equations:

Suppose the state space of $\{X_t\}$ is finite and equal to $\{1, 2, ..., d\}$. In this setting, the forward equation becomes

The backward equation works similarly. So, we have:

Theorem 9.17 *Let $\{X_t\}$ be a CTMC with finite state space. Then, the time t transition function $P(t)$ satisfies both of these differential equations:*

Forward equation: $\left\{ \begin{array}{l} P'(t) = P(t)\, Q \\ P(0) = I \end{array} \right. .$

Backward equation: $\left\{ \begin{array}{l} P'(t) = Q\, P(t) \\ P(0) = I \end{array} \right. .$

Exponentiation of matrices

Consider the backward equation $P'(t) = Q\,P(t)$; $P(0) = I$. Suppose that instead of a matrix $P(t)$, you had a function $y(t)$ in the backward equation, and that instead of constant matrix Q, you had a constant number q. This yields the ODE

$$\begin{cases} y'(t) = q\,y(t) \\ y(0) = 1 \end{cases}$$

In MATH 330, you learn that this equation models ____________________________ , and its solution is

So in the matrix version that we have, the solution ought to be

Recall: the Taylor series of e^t:

$$e^t = \sum_{n=0}^{\infty} \frac{t^n}{n!} = 1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots$$

Definition 9.18 *Given a square matrix A, define the* **matrix exponential** *of A to be the matrix e^A (also denoted $\exp(A)$) defined by*

$$e^A = \exp(A) = \sum_{n=0}^{\infty} \frac{1}{n!} A^n = I + A + \frac{1}{2} A^2 + \frac{1}{3!} A^3 + \ldots$$

There's an issue here with what it means for an infinite series of matrices to converge. Take my word for it: this series converges for all square matrices A, to a matrix $e^A = \exp(A)$ which is the same size as A.

WARNING: If $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, $e^A \neq \begin{pmatrix} e^1 & e^2 \\ e^3 & e^4 \end{pmatrix}$.

More generally, if A is a square matrix, then for any $t \in \mathbb{R}$, we have

$$e^{tA} = e^{At} = \sum_{k=0}^{\infty} \frac{1}{k!} (At)^k = I + At + \frac{A^2}{2} t^2 + \frac{A^3}{3!} t^3 + \ldots$$

Theorem 9.19 (Properties of matrix exponentials) *Let A, B and S be square matrices of the same size, where S is invertible. Let $n \in \{0, 1, 2, 3, ...\}$. Then:*

1. *If A is diagonal (i.e. $A = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{pmatrix}$), then $e^A = \begin{pmatrix} e^{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & e^{\lambda_d} \end{pmatrix}$.*
2. *If $AB = BA$, then $\exp(A + B) = \exp(A)\exp(B)$.*
3. *If $B = \exp(A)$, then $B^n = \exp(An)$.*
4. *For any matrix A, $(e^A)^n = e^{An} = e^{nA}$.*
5. $\exp($*zero matrix*$) = I$.
6. $\exp(SAS^{-1}) = Se^A S^{-1}$.

PROOF MATH 322 or MATH 330. □

Importance: Property (6) above suggests a method to compute the exponential of a matrix A. Diagonalize A (this means write $A = S\Lambda S^{-1}$ where the columns of S are eigenvectors of A and the entries of the diagonal matrix Λ are the corresponding eigenvalues); then $e^A = Se^\Lambda S^{-1}$.

Theorem 9.20 *Let $P(t)$ be a family of square matrices, indexed by t. Then, the following are equivalent:*

1. *$P(t) = e^{Qt} = \exp(Qt)$ for some square matrix Q.*
2. *$P(t)$ solves the forward equation $P'(t) = P(t)\,Q$ and $P(0) = I$;*
3. *$P(t)$ solves the backward equation $P'(t) = Q\,P(t)$ and $P(0) = I$.*

PROOF We start by proving (1) ⇒ (2 and 3): suppose $P(t) = e^{Qt} = \sum_{n=0}^{\infty} \frac{1}{n!} Q^n$.

Clearly $P(0) = e^{Q0} = \exp(\text{zero matrix}) = I$, so the initial condition is satisfied.

Also,

$$P'(t) = \frac{d}{dt} e^{Qt} = \frac{d}{dt} \sum_{n=0}^{\infty} \frac{Q^n}{n!} t^n = \sum_{n=1}^{\infty} \frac{Q^n}{(n-1)!} t^{n-1}$$

$$= \begin{cases} Q\left(\sum_{n=0}^{\infty} \frac{Q^n}{n!}\right) = Qe^{Pt} = Q\,P(t). \\ \left(\sum_{n=0}^{\infty} \frac{Q^n}{n!}\right) Q = e^{Pt}Q = P(t)\,Q. \end{cases}$$

The reverse implications (2 or 3) ⇒ (1) come from the *Existence-Uniqueness Theorem* of differential equations (MATH 330), which says that a system of (ordinary) differential equations with given initial condition has a unique solution (under natural hypotheses that hold here). Since e^{Qt} is a solution of $P'(t) = QP(t)$; $P(0) = I$, it must be the only solution. □

Corollary 9.21

1. *If $\{X_t\}$ is a CTMC with finite state space $\mathcal{S}$, then: the time t transition matrices must satisfy $P(t) = \exp(Qt)$ for some matrix Q where*

$$q_{xx} \leq 0 \qquad q_{xy} \geq 0 \textit{ whenever } x \neq y \qquad \sum_{y \in \mathcal{S}} q_{xy} = 0 \textit{ for all } x \in \mathcal{S}.$$

   In particular, this Q is the generating matrix of the $CTMC$: $Q = P'(0)$.

2. *Any $d \times d$ matrix Q which has the properties*

$$q_{xx} \leq 0 \qquad q_{xy} \geq 0 \textit{ whenever } x \neq y \qquad \sum_{y \in \mathcal{S}} q_{xy} = 0$$

   generates a CTMC by setting $P(t) = \exp(Qt)$ for all t.

EXAMPLE 6

Consider a CTMC with state space $\{1, 2, 3\}$ and infinitesimal matrix

$$Q = \begin{pmatrix} -2 & 1 & 1 \\ 1 & -5 & 4 \\ 2 & 1 & -3 \end{pmatrix}.$$

1. Compute $P(t)$.

2. Recall from the previous page that

$$P(t) = \begin{pmatrix} \frac{11}{24} - \frac{1}{12}e^{-6t} + \frac{5}{8}e^{-4t} & \frac{1}{6} - \frac{1}{6}e^{-6t} & \frac{3}{8} + \frac{1}{4}e^{-6t} - \frac{5}{8}e^{-4t} \\ \frac{11}{24} + \frac{5}{12}e^{-6t} - \frac{7}{8}e^{-4t} & \frac{1}{6} + \frac{5}{6}e^{-6t} & \frac{3}{8} - \frac{5}{4}e^{-6t} + \frac{7}{8}e^{-4t} \\ \frac{11}{24} - \frac{1}{12}e^{-6t} - \frac{3}{8}e^{-4t} & \frac{1}{6} - \frac{1}{6}e^{-6t} & \frac{3}{8} + \frac{1}{4}e^{-6t} + \frac{3}{8}e^{-4t} \end{pmatrix}.$$

Compute $P(X_{3/4} = 0 \,|\, X_{1/2} = 1)$.

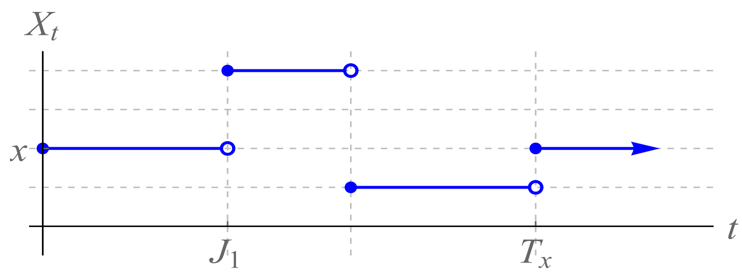
3. If the initial distribution is $\pi_0 = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$, find the distribution at time $t = \ln 2$.

9.4 Class structure, recurrence and transience of CTMCs

Definition 9.22 *Let $\{X_t\}$ be a CTMC and let $y \in \mathcal{S}$. Define the* **hitting time to** y *to be*

$$T_y = \min\{t \geq J_1 : X_t = y\}.$$

(Recall that J_1 is the time of the first jump.)



Definition 9.23 *Let $\{X_t\}$ be a CTMC and let $x, y \in \mathcal{S}$.*

- *Define $f_{x,y} = P_x(T_y < \infty)$. We say $x \to y$ if $f_{x,y} > 0$.*
- *x is called* **recurrent** *if $f_{x,x} = 1$ and* **transient** *otherwise.*
- *x is called* **positive recurrent** *if x is recurrent $m_x = E_x(T_x) < \infty$.*
- *x is called* **null recurrent** *if x is recurrent and $m_x = E_x(T_x) = \infty$.*
- *$\{X_t\}$ is* **irreducible** *if $x \to y$ for all $x, y \in \mathcal{S}$.*

Definition 9.24 *Let $\{X_t\}$ be a CTMC with state space $\mathcal{S}$. The* **embedded chain** *or* **jump chain** *of the CTMC is the (discrete-time) Markov chain $\{X_t^{jump}\}$ whose transition probabilities are given by the jump probabilities $\pi_{x,y}$.*

Notice: ($f_{x,y}$ for a CTMC$\{X_t\}$) $=$ ($f_{x,y}$ for its jump chain $\{X_t^{\text{jump}}\}$). Therefore:

- a CTMC is recurrent, transient, etc. if and only if its jump chain is recurrent, transient, etc., respectively;
- irreducible CTMCs are either positive recurrent, null recurrent, or transient (and must be positive recurrent if their state space is finite); and
- all the same theorems regarding class structure for discrete-time Markov chains hold for CTMCs.

Stationary distributions

Definition 9.25 *Let $\{X_t\}$ be a CTMC with state space $\mathcal{S}$. A distribution π on $\mathcal{S}$ is called* **stationary** *if for all $y \in \mathcal{S}$ and all $t \geq 0$,*

$$\sum_{x \in \mathcal{S}} \pi(x) P_{x,y}(t) = \pi(y).$$

If $\mathcal{S}$ is finite, this means $\pi P(t) = \pi$ in matrix multiplication language.

Theorem 9.26 (Stationarity equation for CTMCs) *Let $\{X_t\}$ be a CTMC with state space $\mathcal{S}$. A distribution π on $\mathcal{S}$ is stationary if and only if*

$$\sum_{x \in \mathcal{S}} \pi(x) q_{xy} = 0 \text{ for all } y \in \mathcal{S}.$$

Note: If $\mathcal{S}$ is finite, this means $\pi Q = \mathbf{0}$ in matrix multiplication language. This gives you a good way to find stationary distributions of CTMCs.

PROOF HW

EXAMPLE 7

Compute the stationary distribution of the CTMC with state space $\mathcal{S} = \{1, 2, 3\}$ whose infinitesimal matrix is

$$Q = \begin{pmatrix} -3 & 2 & 1 \\ 0 & -4 & 4 \\ 1 & 1 & -2 \end{pmatrix}.$$

Solution: Write $\pi = (a, b, c)$; then

$$\pi Q = \mathbf{0} \;\Rightarrow \begin{cases} -3a + c &= 0 \\ 2a - 4b + c &= 0 \\ a + 4b - 2c &= 0 \\ a + b + c &= 1 \end{cases} \Rightarrow \boxed{\pi = \left(\frac{4}{21}, \frac{5}{21}, \frac{12}{21} \right)}.$$

Remark: For the CTMC in Example 7, the jump matrix is $\Pi = \begin{pmatrix} 0 & \frac{2}{3} & \frac{1}{3} \\ 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$. If you solve the equation $\pi^{jump}\Pi = \pi^{jump}$ to find the stationary distribution of the jump chain, you will get

$$\pi^{jump} = \left(\frac{3}{14}, \frac{5}{14}, \frac{6}{14} \right).$$

Question: Is there a connection between π and π^{jump}?

Theorem 9.27 *Let $\{X_t\}$ be a CTMC with state space $\mathcal{S}$.*

1. *Suppose π is a stationary distribution for $\{X_t\}$. For each $x \in \mathcal{S}$, set $\pi'(x) = \pi(x)q_x$. Then*
$$\sum_{x\in\mathcal{S}} \pi'(x)\pi_{x,y} = \pi'(y).$$
2. *Suppose $\pi' : \mathcal{S} \to [0,\infty)$ is a function such that*
$$\sum_{x\in\mathcal{S}} \pi'(x)\pi_{x,y} = \pi'(y).$$
Then if we set, for each $y \in \mathcal{S}$, $\pi^*(y) = \frac{1}{q_y}\pi'(y)$, then for all $y \in \mathcal{S}$,
$$\sum_{x\in\mathcal{S}} \pi^*(x)P_{x,y}(t) = \pi^*(y).$$

PROOF HW

WARNING: The π' defined in Theorem 9.27 may not be a distribution on $\mathcal{S}$ (because its values may not sum to 1). But Theorem 9.27 says π' satisfies the stationarity equation for the jump chain, which in matrix language would be

$$\pi'\,\Pi = \pi'.$$

This means that if $\sum_{y\in\mathcal{S}} \pi'(y) = C$, then by *normalizing π'*, which means setting $\pi^{jump}(y) = \dfrac{1}{C}\pi'(y)$ for all $y \in \mathcal{S}$, we get a stat. dist. π^{jump} for the jump chain.

Similarly, the π^* obtained in statement (2) of Theorem 9.27, *once normalized*, would give the stat. dist. of $\{X_t\}$.

Consequence: if the jump chain of an irred. CTMC has a stat. dist., so does the CTMC, so if the jump chain is pos. recurrent, so is the original CTMC.

The content of Theorem 9.27 can be summarized with this diagram:

$$\begin{tikzcd}[column sep=huge]
\begin{array}{c}\text{stat. dist. } \pi \\ \text{of CTMC } \{X_t\}\end{array} \arrow[r, bend left, "\begin{array}{c}\text{multiply by holding rates,}\\ \text{then normalize}\end{array}"] & \begin{array}{c}\text{stat. dist } \pi^{jump} \\ \text{of jump chain } \{X_t^{jump}\}\end{array} \arrow[l, bend left, "\begin{array}{c}\text{divide by holding rates,}\\ \text{then normalize}\end{array}"]
\end{tikzcd}$$

EXAMPLE 8

Suppose $\{X_t\}$ is a CTMC with state space $\{1, 2, 3, 4\}$. If the stationary distribution of the jump chain of $\{X_t\}$ is uniform, and the holding rates are $q_1 = 2$, $q_2 = 3$, $q_3 = 6$, $q_4 = 4$, compute the stationary distribution of $\{X_t\}$.

Theorem 9.28 (Steady-state distribution of CTMCs) *Let $\{X_t\}$ be an irreducible, positive recurrent CTMC with a stationary distribution π. Then the stationary distribution is steady-state, i.e.*

- $\lim_{t \to \infty} P_{x,y}(t) = \pi(y)$ *for all $x, y \in \mathcal{S}$; and*
- $\lim_{t \to \infty} P(X_t = y) = \pi(y)$ *for all $y \in \mathcal{S}$, regardless of the initial distribution.*

Why is the stationary distribution always steady-state? The short answer is

PROOF Fix $h > 0$ and consider the discrete-time Markov chain $\{Z_n\} = \{X_{hn}\}$ for $n \in \{0, 1, 2, ...\}$.

$\{Z_n\}$ has transition functions $P(x, y) = P_{x,y}(h)$, and since these functions are always positive for $h > 0$, $\{Z_n\}$ is irreducible and aperiodic.

So the FTMC applied to $\{Z_n\}$ gives a stat. dist. π for $\{Z_n\}$ (which must be the stat. dist. for $\{X_t\}$) which is steady-state for $\{Z_n\}$, i.e.

$$\lim_{n \to \infty} P^n(x, y) = \lim_{n \to \infty} P_{x,y}(hn) = \pi(y)$$

for all $x, y \in \mathcal{S}$.

So for t that are multiples of h, $P_{x,y}(t) \to \pi(y)$.

Since h can be chosen arbitarily small and since $t \mapsto P_{x,y}(t)$ is (uniformly) cts, it follows (from a MATH 430 argument) that $\lim_{t \to \infty} P_{x,y}(t) = \pi(y)$. $\square$

Corollary 9.29 *An irreducible, positive recurrent CTMC cannot have more than one stationary distribution.*

PROOF If π and π' are both stationary, then they would both be steady-state for $\{Z_n\}$ as described in the previous theorem. This is impossible. □

Theorem 9.30 (Stationary distributions of CTMCs) *Let $\{X_t\}$ be an irreducible CTMC with state space $\mathcal{S}$.*

1. *If $\{X_t\}$ is transient or null recurrent, then it has no stationary distributions.*
2. *If $\{X_t\}$ is positive recurrent, then it has one stationary distribution π given by $\pi(x) = \dfrac{1}{m_x q_x}$ for all $x \in \mathcal{S}$.*

PROOF The only thing left to prove is the formula for the stat. dist. in the positive recurrent case.

Suppose $\{X_t\}$ is irreducible and positive recurrent, and fix $x \in \mathcal{S}$.

For each $y \in \mathcal{S}$, define

$$\tau_x(y) = E_x\left[\int_0^{T_x} \mathbb{1}_{\{X_s=y\}}\, ds\right];$$

this is the expected amount of time the chain spends in state y before it first returns to x.

A preview: This $\tau_x(y)$ will turn out not to depend on x at all.

Notice $\sum_{y\in\mathcal{S}} \tau_x(y) = E_x(T_x) = m_x$, so by setting $\pi_x(y) = \dfrac{1}{m_x}\tau_x(y)$, we get a distribution π_x on $\mathcal{S}$. This $\pi_x(y)$ measures the *fraction* of the time in the CTMC spent in state y before the first return to x.

A preview: This $\pi_x(y)$ will turn out not to depend on x at all.

Now let $\{Y_n\}$ denote the jump chain associated to $\{X_t\}$ and let T_x^{jump} be the first return time to x in $\{Y_n\}$ (this is the number of jumps it takes x to return to itself in the CTMC). We have

$$\begin{aligned}
\tau_x(y) &= E_x\left[\sum_{n=0}^{\infty} W_y \mathbb{1}_{\{Y_n=y, n<T_x^{jump}\}}\right] \\
&= \frac{1}{q_y} E_x\left[\sum_{n=0}^{\infty} \mathbb{1}_{\{Y_n=y, n<T_x^{jump}\}}\right] \\
&= \frac{1}{q_y} E_x\left[\sum_{n=0}^{T_x^{jump}-1} \mathbb{1}_{\{Y_n=y\}}\right].
\end{aligned}$$

Define $\gamma_x(y) = E_x\left[\sum_{n=0}^{T_x^{jump}-1} \mathbb{1}_{\{Y_n=y\}}\right]$, so that $\tau_x(y) = \dfrac{1}{q_y}\gamma_x(y)$.

A preview: This $\gamma_x(y)$ will turn out not to depend on x either.

Claim: "$\gamma_x\Pi = \gamma_x$", i.e. $\sum_{y\in\mathcal{S}} \gamma_x(y)\,\pi(y,z) = \gamma_x(z)$.

To prove this claim, first note that since the jump chain is positive recurrent, $T_x^{jump} < \infty$) with probability 1, so

$$\begin{aligned}
\gamma_x(z) &= E_x\left[\sum_{n=1}^{T_x^{jump}} \mathbb{1}_{\{Y_n=z\}}\right] \\
&= E_x\left[\sum_{n=1}^{\infty} \mathbb{1}_{\{Y_n=z \text{ and } n<T_x^{jump}\}}\right] \\
&= \sum_{n=1}^{\infty} E_x\left[\mathbb{1}_{\{Y_n=z \text{ and } n<T_x^{jump}\}}\right] \\
&= \sum_{n=1}^{\infty} P(Y_n = z \text{ and } n < T_x^{jump}) \\
&= \sum_{y\in\mathcal{S}}\sum_{n=1}^{\infty} P(Y_n = z, Y_{n-1} = y \text{ and } n < T_x^{jump}) \\
&= \sum_{y\in\mathcal{S}} \pi(y,z) \sum_{n=1}^{\infty} P(Y_{n-1} = y \text{ and } n < T_x^{jump}) \\
&= \sum_{y\in\mathcal{S}} \pi(y,z) E_x\left[\sum_{m=0}^{\infty} \mathbb{1}_{\{Y_m=y \text{ and } n<T_x^{jump}-1\}}\right] \\
&= \sum_{y\in\mathcal{S}} \pi(y,z) E_x\left[\sum_{m=0}^{T_x^{jump}-1} \mathbb{1}_{\{Y_m=y\}}\right] \\
&= \sum_{y\in\mathcal{S}} \pi(y,z)\gamma_x(y).
\end{aligned}$$

Having proven the claim, by (2) of Theorem 9.27, τ_x is a multiple of a stationary distribution of $\{X_t\}$. But the only multiple of τ_x which is a distribution is the π_x we defined earlier, i.e.

$$\pi_x(y) = \frac{1}{m_x}\tau_x(y) = \frac{1}{m_x q_y}\gamma_x(y).$$

So for each $x \in \mathcal{S}$, this π_x must be stationary, and since there is at most one stationary distribution, we know $\pi_x = \pi$ for all $x \in \mathcal{S}$ (this verifies that none of $\pi_x(y)$, $\tau_x(y)$ and $\gamma_x(y)$ actually depend on x). In particular,

$$\begin{aligned}
\pi(x) = \pi_x(x) = \frac{1}{m_x}\tau_x(x) = \frac{1}{m_x q_x}\gamma_x(x) &= \frac{1}{m_x q_x} E_x\left[\sum_{n=0}^{T_x^{jump}-1} \mathbb{1}_{\{Y_n=x\}}\right] \\
&= \frac{1}{m_x q_x}(1) = \frac{1}{m_x q_x}. \ \square
\end{aligned}$$

Corollary 9.31 *Let $\{X_t\}$ be an irreducible, positive recurrent CTMC and let $x, y \in \mathcal{S}$. If we define $\tau_x(y)$ to be the expected amount of time spent in state y before the first return to x, given that $X_0 = x$, then*

$$\tau_x(y) = m_x \pi(y).$$

PROOF From above, $\dfrac{1}{m_y q_y} = \pi_y(y) = \pi_x(y) = \dfrac{1}{m_x q_y}\gamma_x(y)$, it must be that

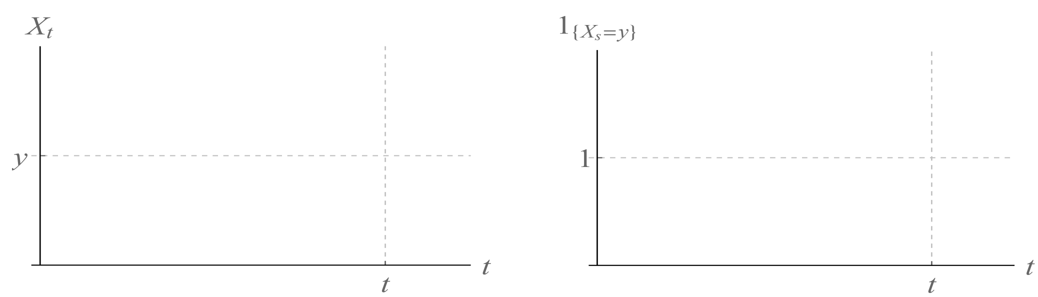
$$\gamma_x(y) = \frac{m_x}{m_y} \quad \Rightarrow \quad \tau_x(y) = \frac{1}{q_y}\gamma_x(y) = \frac{m_x}{q_y m_y} = m_x\pi(y). \ \square$$

Ergodic theorem

We finish this section with a theorem that says time averages (the proportion of time spent in state x in a CTMC) converges to the space average (the value that the stationary distribution assigns x).

Theorem 9.32 (Ergodic theorem for CTMCs) *Let $\{X_t\}$ be an irreducible, positive recurrent CTMC, and let π be the stationary distribution of $\{X_t\}$. Then for all $y \in \mathcal{S}$,*

$$P\left[\lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbb{1}_{\{X_s = y\}}\, ds = \pi(y)\right] = 1.$$



SKETCH OF PROOF Let $x = X_0$. The expected length of each block of time in $[0, t]$ between successive returns to x is m_x, so by the Strong Law of Large Numbers, the average length of the blocks approaches m_x with probability 1.

In each block, the expected amount of time spent in state y is $\tau_x(y) = m_x\pi(y)$, so by the SLLN the average time spent in y in each block approaches $m_x\pi(y)$ with probability 1.

Therefore the proportion of time spent in state y in each block approaches $\dfrac{m_x\pi(y)}{m_x} = \pi(y)$ with probability 1. The result follows. $\square$

9.5 Specific examples of CTMCs

Two-state CTMC

EXAMPLE 9

Let $\{X_t\}$ be a birth-death CTMC on $\mathcal{S} = \{0, 1\} = \{OFF, ON\}$. Let $q_{01} = \lambda$ and $q_{10} = \mu$.

1. Sketch the directed graph of $\{X_t\}$.
2. Compute the infinitesimal matrix of $\{X_t\}$.
3. Compute the stationary distribution of $\{X_t\}$.
4. Compute the transition matrices $P_{x,y}(t)$.

Poisson processes

EXAMPLE 10

A **Poisson process** is a CTMC $\{X_t\}$on $\mathcal{S} = \{0, 1, 2, ...\}$ that starts at 0 ($X_0 = 0$), has constant holding rates ($q_x = \lambda$ for all x) and no deaths or simultaneous births ($\pi_{x,x+1} = 1$ for all x.

1. Sketch the directed graph of a Poisson process.
2. Compute all the infinitesimal parameters of a Poisson process.
3. Use your answer to # 2 to simplify the forward equation for a Poisson process.
4. Compute $P_{x,y}(t)$ when $y < x$.
5. Compute $P_{x,x}(t)$.
6. Use your answer to # 3 to derive a recursive formula for $P_{x,y+1}(t)$ in terms of $P_{x,y}(t)$.
7. Compute $P_{x,x+1}(t)$.
8. Compute $P_{x,x+2}(t)$.
9. Based on your answers to # 7 and # 8, give a formula for $P_{x,x+n}(t)$. (This should match a formula associated to a Poisson process we derived in MATH 414.)

The infinite server queue

EXAMPLE 11

Let X_t denote the number of people in line for some service (including those being served).

Assume that the people arrive at rate λ (i.e. that the number of arrivals in line follows a Poisson process with rate λ) and that the time it takes each customer to be served is exponential with parameter μ.

Assume that there are an infinite number of servers (so no one has to wait in line before being served).

The resulting CTMC $\{X_t\}$ is called the **infinite server queue** or the $M/M/\infty$ **queue**.

Question 1: Compute the time t transition function for the $M/M/\infty$ queue.

Solution: We want $P_{x,y}(t) = P(X_t = y \,|\, X_0 = x)$.

To compute this, note that of the y customers that are supposed to be in the chain at time t, some of them (say k of the y) would have been in the chain at time 0. Call these folks **original customers** and denote their number by X_t^{orig}.

The remaining $y-k$ of them would have to have joined the queue after time 0, but not be served by time t. Call these **new arrivals still in the queue** and denote their number by X_t^{new}.

We have

$$\begin{aligned} P_{x,y}(t) &= \sum_{k=0}^{\min(x,y)} P_x\left(\begin{array}{c}\text{there are } k \text{ original} \\ \text{customers still in the} \\ \text{queue at time } t\end{array}\right) \cdot P\left(\begin{array}{c}\text{there are } y-k \text{ new} \\ \text{arrivals in the queue} \\ \text{at time } t\end{array}\right) \\ &= \sum_{k=0}^{\min(x,y)} P_x\left(X_t^{\text{orig}} = k\right) P\left(X_t^{\text{new}} = y-k\right). \end{aligned} \tag{\$}$$

Next, let's compute $P_x\left(X_t^{\text{orig}} = k\right)$. Each customer has a service time which is $Exp(\mu)$, so the probability the customer is still in the queue at time t is

$$P(\text{service time } \geq t) = P(Exp(\mu) \geq t) = e^{-\mu t}.$$

Since the customers act independently, we count the number still in the queue by a ________________________ r.v., i.e.

$$P_x\left(X_t^{\text{orig}} = k\right) = b(e^{-\mu t}, x, k) = \boxed{\binom{x}{k} e^{-\mu k t}\left(1 - e^{-\mu t}\right)^{x-k}}.$$

Now, we need to compute $P\left(X_t^{\text{new}} = y - k\right)$. More generally, let's compute $P\left(X_t^{\text{new}} = w\right)$ and substitute in $w = y - k$ later.

Let C_t be the number of customers arriving in $(0, t]$. By the LTP, we have

$$P(X_t^{\text{new}} = w) = \sum_{c=0}^{\infty} P(X_t^{\text{new}} = w \mid C_t = c)\, P(C_t = c).$$

Since there have to be at least w arrivals for X_t^{new} to be w, this reduces to

$$P(X_t^{\text{new}} = w) = \sum_{c=w}^{\infty} \boxed{P(X_t^{\text{new}} = w \mid C_t = c)}\, P(C_t = c)$$

$$P\left(\begin{array}{c}\text{of the } c \text{ arrivals,}\\ w \text{ are still in the queue}\end{array}\right) \qquad P(Pois(\lambda t) = c)$$

$$b\left(c, \boxed{P\left(\begin{array}{c}\text{any one arriving customer}\\ \text{is still in the queue}\end{array}\right)}, w\right) \qquad \frac{e^{-\lambda t}(\lambda t)^c}{c!}$$

$$\int_0^t \boxed{P\left(\begin{array}{c|c}\text{customer is still} & \text{customer arrives}\\ \text{in queue at time } t & \text{at time } s\end{array}\right)} \boxed{f_S(s)}\, ds$$

$$P(Exp(\mu) \geq t - s) \qquad \frac{1}{t}$$

$$e^{-\mu(t-s)}$$

All together, this ends up being

$$P(X_t^{\text{new}} = w) = \sum_{c=w}^{\infty} \binom{c}{w} \left[\int_0^t e^{-\mu(t-s)} \frac{1}{t}\, ds\right]^w \left[1 - \int_0^t e^{-\mu(t-s)} \frac{1}{t}\, ds\right]^{c-w} \frac{e^{\lambda t}(\lambda t)^c}{c!}.$$

If you work this out (starred HW problem), you will get

$$P\left(X_t^{\text{new}} = w\right) = \frac{e^{\frac{\lambda}{\mu}(1 - e^{-\mu t})} \left[\frac{\lambda}{\mu}(1 - e^{-\mu t})\right]^w}{w!}.$$

In other words, $X_t^{\text{new}} \sim Pois\left(\frac{\lambda}{\mu}\left(1-e^{\mu t}\right)\right)$. Therefore

$$\begin{aligned} P\left(X_t^{\text{new}}=y-k\right) &= P\left(Pois\left(\frac{\lambda}{\mu}\left(1-e^{\mu t}\right)\right)=y-k\right) \\ &= \boxed{\frac{\left[\frac{\lambda}{\mu}\left(1-e^{-\mu t}\right)\right]^{y-k}}{(y-k)!}\exp\left(-\frac{\lambda}{\mu}(1-e^{-\mu t})\right)}. \end{aligned}$$

Finally, substituting back into ($), from two pages ago, we get

$$\begin{aligned} P_{x,y}(t) &= \sum_{k=0}^{\min(x,y)} P_x(X_t^{\text{orig}}=k)\, P(X_t^{\text{new}}=y-k) \qquad (\$) \\ &= \sum_{k=0}^{\min(x,y)} \left[\binom{x}{k} e^{-\mu k t}(1-e^{-\mu t})^{x-k}\right]\left[\frac{\left[\frac{\lambda}{\mu}(1-e^{-\mu t})\right]^{y-k}}{(y-k)!}\exp\left(\frac{-\lambda}{\mu}(1-e^{-\mu t})\right)\right]. \end{aligned}$$

Question 2: Show the $M/M/\infty$ queue is positive recurrent, and compute its stationary distribution.

Solution: Since the stationary distribution would be steady-state, let's use our formula for $P_{x,y}(t)$ and see what happens when $t \to \infty$:

$$\begin{aligned} \lim_{t\to\infty} P_{x,y}(t) &= \lim_{t\to\infty} (k=0 \text{ term of the above sum}) \\ &= \lim_{t\to\infty} \binom{x}{0} e^{-0}\left(1-e^{-\mu t}\right)^x \left[\frac{\left[\frac{\lambda}{\mu}\left(1-e^{-\mu t}\right)\right]^y}{y!}\exp\left(\frac{-\lambda}{\mu}\left(1-e^{-\mu t}\right)\right)\right] \\ &= (1)1(1)^x\left[\frac{\left[\frac{\lambda}{\mu}(1)\right]^y}{y!}\exp\left(\frac{-\lambda}{\mu}(1)\right)\right] \\ &= \frac{\left(\frac{\lambda}{\mu}\right)^y}{y!} e^{-(\lambda/\mu)}. \end{aligned}$$

We have proven:

Theorem 9.33 (Stat. dist. of the $M/M/\infty$ queue) *The $M/M/\infty$ queue with arrival rate λ and service rate μ is positive recurrent, and its stationary distribution is $Pois\left(\frac{\lambda}{\mu}\right)$.*

9.6 Chapter 9 Homework

Exercises from Section 9.2

1. Consider a continuous-time Markov chain $\{X_t\}$ with state space $\mathcal{S} = \{1, 2, 3, 4\}$ with holding rates $q_1 = q_2 = 1, q_3 = 3, q_4 = 2$ and jump probabilities $\pi_{12} = \frac{1}{8}$, $\pi_{13} = \frac{1}{2}$, $\pi_{21} = 0$, $\pi_{23} = \frac{1}{4}$, $\pi_{31} = \frac{1}{6}$, $\pi_{32} = \frac{2}{5}$, $\pi_{41} = \pi_{42} = \frac{1}{6}$.
   a) Compute π_{34} and π_{44}.
   b) Compute the infinitesimal parameters q_{xy} for all $x, y \in \mathcal{S}$.
   c) What is the probability that $X_t = 2$ for all $t < 4$, given that $X_0 = 2$?
   d) What is the probability that your first two jumps are first to state 3 and then to state 2, given that you start in state 1?
   e) What is $P'_{13}(0)$?
   f) Use linear approximation to estimate $P_{31}(.001)$ and $P_{22}(.06)$.

2. Given infinitesimal probabilities
   a) What is probability of jump before time blah?
   b) What is probability that first jump is to a certain state
   c) Write forward equation for $P'_{12}(t)$
   d) Write backward equation for $P'_{12}(t)$.

3. Consider the CTMC $\{X_t\}$ with state space $\{0, 1\}$ where $q_9 = 3$ and $q_1 = 4$.
   a) Write out the system of differential equations which constitute the backward equation of $\{X_t\}$.
   b) Write out the system of differential equations which constitute the forward equation of $\{X_t\}$.

4. (60 ★ pts) In this problem, we prove Theorem 9.13, which asserts that a CTMC $\{X_t\}$ satisfies the forward equation.
   a) Take a look at this equation (I hope you can convince yourself that this equation is true):

$$P_{x,y}(t) = P_x(X_t = y) = \sum_{n=0}^{\infty} \sum_{z \neq y} P_x(J_n \leq t < J_{n+1}, X_{J_{n-1}} = z, X_{J_n} = y).$$

      i. In this equation, describe in English what the n is referring to.
      ii. In this equation, describe in English what the z is referring to.

b) Using the time reversal identity proven in Lemma 9.12, prove

$$\begin{aligned}&P_x(J_n \le t < J_{n+1} \,|\, X_{J_{n-1}} = z, X_{J_n} = y)\\ &= q_x \int_0^t e^{-q_y s} \frac{q_z}{q_x} P_x(J_{n-1} \le t - s < J_n \,|\, X_{J_{n-1}} = z)\, ds.\end{aligned}$$

c) Use the multiplication principle and substitute in the formula you found in part (b) to the equation from part (a) to derive the following "forward integral equation":

$$P_{x,y}(t) = \delta_{x,y} e^{-q_x t} + \int_0^t \sum_{z \neq x} P_{x,z}(t-s) q_{zy} e^{-q_y s}\, ds$$

d) Perform the u-sub $u = t - s$ in the forward integral equation of part (c) and simplify what you get to obtain

$$P_{x,y}(t) = \delta_{x,y} e^{-q_y t} + e^{-q_y t} \int_0^t \sum_{z \neq x} P_{x,z}(u) q_{zy} e^{q_y u}\, du.$$

e) Explain why you know from the formula of part (d) that $P_{x,y}$ is a differentiable function of t.

f) Differentiate both sides of the equation in (d), and rewrite the equation you obtain to get the forward equation

$$P'_{x,y}(t) = \sum_{z \in \mathcal{S}} P_{x,z}(t) q_{zy}.$$

Hint: This should resemble the computation done in the proof of Theorems 9.7 and 9.11.

Exercises from Section 9.3

5. Let $\{X_t\}$ be a CTMC with time t transition matrix

$$P(t) = \frac{1}{22}\begin{pmatrix} 12 - 11e^{-13t/2} + 21e^{-11t/2} & 4 + 11e^{-13t/2} - 15e^{-11t/2} & b(t) \\ 12 - 33e^{-13t/2} + 21e^{-11t/2} & 4 + 33e^{-13t/2} - 15e^{-11t/2} & 6 - 6e^{-11t/2} \\ 12 + 44e^{-13t/2} - Ke^{-11t/2} & 4 - 44e^{-13t/2} + 40e^{-11t/2} & 6 + 16e^{-11t/2} \end{pmatrix},$$

where $b(t)$ is a function and K is a constant.

a) Compute $b(t)$.

b) Compute K.

c) Compute the infinitesimal matrix Q.

d) Compute the jump matrix Π.

6. Consider a continuous-time Markov chain $\{X_t\}$ with with state space $\{1, 2, 3\}$ and infinitesimal matrix

$$Q = \begin{pmatrix} -5 & 3 & b \\ 4 & -6 & 2 \\ 2 & 1 & -3 \end{pmatrix}.$$

where b is a constant.

a) What is b?

b) Compute the jump matrix Π.

c) Compute the holding rate of state 2.

d) Suppose $X_0 = 2$. What is the probability that when the chain jumps, the next state is 3?

e) Suppose $X_0 = 3$. What is the probability that $X_t = 3$ for all $t \in [0, 4]$?

f) Suppose $X_0 = 3$. What is the expected amount of time before the first jump?

7. Let $\{X_t\}$ be the CTMC of Problem 6.

a) Compute the time t transition matrix $P(t)$.

b) Compute the probability that $X_5 = 2$ given that $X_2 = 2$.

c) If the initial distribution is $\left(\frac{1}{5}, \frac{3}{5}, \frac{1}{5}\right)$, find the distribution at time $t = \ln 6$. Simplify this answer to remove any exponentials and logarithms.

d) Show that $\lim_{t \to \infty} P(X_t = 2 \,|\, X_0 = x)$ does not depend on x, and find its value.

Exercises from Section 9.4

8. (20 ★ pts) (In this problem, we prove Theorem 9.26 from the lecture notes.) Suppose $\{X_t\}$ is a continuous-time Markov chain with finite state space $\mathcal{S}$ and infinitesimal matrix Q.

a) Prove that if π is stationary (i.e. $\pi P(t) = \pi$ for all $t \geq 0$), then $\pi Q = \mathbf{0}$.

b) Prove that if $\pi Q = \mathbf{0}$, then π is stationary.

9. (40 ★ pts) Prove Theorem 9.27 from the lecture notes.

10. Consider the CTMC $\{X_t\}$ from Problem 6.

a) Compute the stationary distribution of $\{X_t\}$.

b) Compute the mean return time to each state.

c) Compute $\tau_2(3)$, where $\tau_2(3)$ is as defined in the proof of Theorem 9.30 (it is the expected amount of time the chain spends in state 3 before it returns to state 2).

11. Compute the stationary distribution of the CTMC described in Problem 1.

12. Suppose $\{X_t\}$ is a continuous-time Markov chain with state space $\{1,2,3\}$ and time t transition matrix

$$P(t) = \frac{1}{9}\begin{pmatrix} 1+6te^{-3t}+8e^{-3t} & 6-6e^{-3t} & 2-6te^{-3t}-2e^{-3t} \\ 1-3te^{-3t}-e^{-3t} & 6+3e^{-3t} & 2+3te^{-3t}-2e^{-3t} \\ 1+6te^{-3t}-e^{-3t} & 6-6e^{-3t} & 2-6te^{-3t}+7e^{-3t} \end{pmatrix}.$$

a) Compute the infinitesimal matrix of this process.

b) What is the probability that $X_2 = 1$, given that $X_0 = 1$?

c) What is the probability that $X_t = 1$ for all $t < 2$, given that $X_0 = 1$?

d) Compute the steady-state distribution π.

e) Compute the mean return time to each state.

f) Suppose you let time pass from $t = 0$ to $t = 1\,200\,000$. What is the expected amount of time in this interval for which $X_t = 3$?

g) Suppose $X_0 = 2$. What is the expected amount of time spent in state 3 before the first time the chain returns to state 2?

h) Suppose $X_0 = 2$. What is the expected amount of time spent in state 3 before the eleventh time the chain returns to state 2?

13. Consider a continuous-time Markov chain $\{X_t\}$ with with state space $\{1,2,3\}$ and infinitesimal matrix

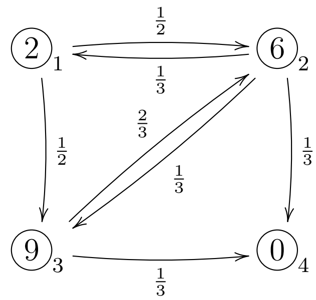
$$Q = \begin{pmatrix} -4 & 1 & 3 \\ 0 & -1 & 1 \\ 0 & 2 & -2 \end{pmatrix}.$$

a) Classify the states as recurrent or transient.

b) Are the recurrent states positive recurrent or null recurrent? Explain.

c) Find all stationary distributions of $\{X_t\}$. Are any of them steady-state?

14. Consider a CTMC $\{X_t\}$, whose directed graph is as given below (the fact that the holding rate of state 4 is 0 means that once you are in state 4, you stay in state 4 forever):



Compute $E_1(T_4)$.

Hint: For $i = 1, 2, 3$, let $k_i = E_i(T_4)$. Set up a system of equations that will enable you to solve for all of the k_i.

Exercises from Section 9.5

15. Let $\{X_t\}$ be an irreducible CTMC with state space $\mathcal{S} = \{0, 1\}$ where $P_{0,1}(t) = \frac{7}{10} + \frac{3}{10}e^{-4t}$. Compute $P(X_5 = 0 \,|\, X_0 = 1)$.

16. (30 ★ pts) Let $\{X_t\}$ be the infinite server queue. Finish the calculation in the middle of Example 11, starting with

$$P(X_t^{\text{new}} = w) = \sum_{c=w}^{\infty} \binom{c}{w} \left[\int_0^t e^{-\mu(t-s)} \frac{1}{t}\, ds\right]^w \left[1 - \int_0^t e^{-\mu(t-s)} \frac{1}{t}\, ds\right]^{c-w} \frac{e^{\lambda t}(\lambda t)^c}{c!},$$

and working out this expression to show that

$$P(X_t^{\text{new}} = w) = \frac{e^{\frac{\lambda}{\mu}\left(1-e^{-\mu t}\right)} \left[\frac{\lambda}{\mu}\left(1 - e^{-\mu t}\right)\right]^w}{w!}.$$

17. A **pure birth process** is a CTMC with state space $\{0, 1, 2, ...\}$ where $\pi_{x,x+1} = 1$ for all $x \in \mathcal{S}$ (this is like a Poisson process, but without the assumption that the holding rates are constant, so hypothetically the q_x's might be the same or different).

    a) Sketch the directed graph of a pure birth process.
    b) Compute all the infinitesimal parameters of a pure birth process.
    c) Use your answer to part (b) to simplify the forward equation of this process.
    d) What is $P_{x,y}(t)$ when $y < x$?
    e) Compute $P_{x,x}(t)$.

f) Solve the differential equation from part (a) to obtain a recursive formula for $P_{x,y+1}(t)$ in terms of $P_{x,y}(t)$.

Hint: there are two cases, depending on whether or not $q_x = q_{x+1}$.

g) Compute $P_{x,x+1}(t)$, under the assumption that $q_x \neq q_{x+1}$.

h) Compute $P_{x,x+1}(t)$, under the assumption that q_x, q_{x+1} and q_{x+2} are all distinct.

18. (40 ★ pts) A **pure death process** is a CTMC with state space $\{0, 1, 2, ...\}$ where $q_0 = 0$ (so 0 is absorbing) and $\pi_{x,x-1} = 0$ for all $x \geq 1$.

    a) Simplify the forward equation of this process as in part (c) of the previous exercise.
    b) Let $x \geq 1$. Compute $P_{x,x}(t)$.
    c) Solve the differential equation from part (a) to obtain a recursive formula for $P_{x,y-1}(t)$ in terms of $P_{x,y}(t)$.
    d) Compute $P_{x,x-1}(t)$, under the assumption that $q_x \neq q_{x-1}$.
    e) Compute $P_{x,x-1}(t)$, under the assumption that $q_x = q_{x-1}$.

19. Let $\{X_t\}$ be an $(M/M/\infty)$-queue where the arrivals follow a Poisson process with mean 12 arrivals per hour, and the service times are exponential with parameter 4 (i.e. the mean service time is $\frac{1}{4}$ hour).

    a) If there are currently 8 individuals in the queue, what is the probability that exactly 5 of those 8 individuals are still in the queue at time $\frac{3}{4}$?
    b) What is the probability that there will be 7 individuals in the queue at time $\frac{3}{2}$ who were not originally in the queue?
    c) Compute and simplify $P_{2,2}(t)$.
    d) Use the steady-state distribution to estimate the probability that at some distant time in the future, there are 7 individuals in the queue.

20. (40 ★ pts) A large triangle has three vertices A, B and C. You and your friend initially stand at vertex A. You decide to move from one vertex to the next clockwise vertex at random times, where the times between your movements are i.i.d. exponential r.v.s, each having parameter λ. Your friend decides to move from one vertex to the next counterclockwise vertex at random times, where the times between their movements are i.i.d. exponential r.v.s, each having parameter μ.

    Here is the catch: you and your friend are handcuffed, so when you move, your friend moves with you, and when your friend moves, you move with them. Derive a formula which gives the probability you and your friend are both on vertex A at time t.

21. Consider a CTMC $\{X_t\}$ with $\mathcal{S} = \{0, 1, 2, 3, ...\}$ where

$$q_{xy} = \begin{cases} \lambda x & \text{if } y = x + 1 \\ \mu x & \text{if } y = x - 1 \\ -(\lambda + \mu)x & \text{if } y = x \\ 0 & \text{else} \end{cases}$$

for positive constants λ and μ.

a) Simplify the forward equation of this process.

b) Let $g_x(t) = E_x(X_t)$. Use the forward equation to show

$$g_x'(t) = (\lambda - \mu)g_x(t).$$

c) Based on part (b), derive a formula for $g_x(t)$.

d) Compute $E_0(X_8)$.

Chapter 10

Martingales

10.1 Background: a gambling problem

Suppose you are playing a game with your friend where you bet \$1 on each flip of a fair coin (fair means the coin flips heads with probability $\frac{1}{2}$ and tails with probability $\frac{1}{2}$). If the coin flips heads, you win, and if the coin flips tails, you lose (mathematically, this is the same as "calling" the flip and winning if your call was correct).

Suppose you come to this game with \$10. What will be the situation after three coin flips?

Let X_t be your bankroll after playing the game t times; this gives a stochastic process $\{X_t\}_{t\in\mathbb{N}}$. We know $X_0 = 10$, for example.

Sequence of flips (in order)	Probability of that sequence	X_3 = bankroll after three flips
H H H	$\frac{1}{8}$	13
H H T	$\frac{1}{8}$	11
H T H	$\frac{1}{8}$	11
H T T	$\frac{1}{8}$	9
T H H	$\frac{1}{8}$	11
T H T	$\frac{1}{8}$	9
T T H	$\frac{1}{8}$	9
T T T	$\frac{1}{8}$	7

To summarize, your bankroll after three flips, i.e. X_3, has the following density:

x	7	9	11	13
$P(X_3 = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

so your expected bankroll is

$$EX_3 = \frac{1}{8}(7) + \frac{3}{8}(9) + \frac{3}{8}(11) + \frac{1}{8}(13) = \frac{7 + 27 + 33 + 13}{8} = \frac{80}{8} = 10.$$

Notice that your expected bankroll after 3 rolls is the amount you started with:

$$EX_3 = X_0.$$

The major question: can you beat a fair game?

Suppose that instead of betting \$1 on each flip, that you varied your bets from one flip to the next. Suppose you think of a method of betting as a *betting strategy*. Here are some things you might try:

Strategy 1: Bet \$1 on each flip.

Strategy 2: Alternate between betting \$1 and betting \$2.

Strategy 3: Start by betting \$1 on the first flip.
After that, bet \$2 if you lost the previous flip, and bet \$1 otherwise.

Strategy 4: Bet \$1 on the first flip.
If you lose, double your bet after each flip you lose until you win once.
Then go back to betting \$1 and repeat the procedure.

QUESTION

Is there a strategy (especially one with bounded bet sizes) you can implement such that your expected bankroll after the 20^{th} flip is greater than your initial bankroll X_0? If so, what is it? If not, what about if you flip 100 times? Or 1000 times? Or any finite number of times?

Furthermore, suppose that instead of planning beforehand to flip a fixed number of times, decide that you will *stop at a random time* depending on the results of the flips. For instance, you might stop when you win five straight bets. Or you might stop when you are ahead \$3.

MORE GENERAL QUESTION

Is there is a betting strategy and a (possibly random) time you can plan to stop, so that if you implement that strategy and stop when you plan to, you will expect to have a greater bankroll than what you start with (even though you are playing a fair game)?

The idea of a martingale

Let's return to the setup of the previous section, where you were wagering $\$1$ on each flip of a fair coin. We saw that in this setting, $E[X_3] = X_0$.

QUESTION

What happens if we condition on some additional information? For example, suppose that the first flip is heads (so that you win your first bet, so that $X_1 = 11$). Given this, what is $E[X_3]$? In other words, **what is $E[X_3 \,|\, X_1 = 11]$?**

Repeating the argument from the previous section, we see

Sequence of flips (in order)	Probability of that sequence	Resulting bankroll after four flips
H H H	$\frac{1}{4}$	13
H H T	$\frac{1}{4}$	11
H T H	$\frac{1}{4}$	11
H T T	$\frac{1}{4}$	9

Therefore $X_3 \,|\, X_1 = 11$ has conditional density

x	13	11	9
$P(X_3 = x \,\|\, X_1 = 11)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

and therefore $E[X_3 \,|\, X_1 = 11] = 0(6) + \frac{1}{4}(13) + \frac{1}{2}(11) + \frac{1}{4}(9) = 11.$

A similar calculation would show that if the first flip was tails, then

$$E[X_3 \,|\, X_1 = 9] = 9.$$

Either way,

In fact, something more general holds. For this Markov chain $\{X_t\}$, we have for any $s \leq t$ that

$$E[X_t \,|\, X_s] = X_s.$$

To see why, let's define another sequence of random variables coming from the process $\{X_t\}$. For each $t \in \{1, 2, 3, ...\}$, define

$$\begin{aligned} S_t &= X_t - X_{t-1} \\ &= \text{the } t^{th} \textbf{ step} \text{ of the process} \\ &= \begin{cases} +1 & \text{if the } t^{th} \text{ flip is H (i.e. you win \$1 on the } t^{th} \text{ game)} \\ -1 & \text{if the } t^{th} \text{ flip is T (i.e. you lose \$1 on the } t^{th} \text{ game).} \end{cases} \end{aligned}$$

Notice:

- $E[S_t] = \frac{1}{2}(1) + \frac{1}{2}(-1) = 0$;
- for any $s \leq t$, $S_t \perp X_s$;
- for any $s \leq t$, $X_t = X_s + S_{s+1} + S_{s+2} + ... + S_t = X_s + \sum_{j=s+1}^{t} S_j$.

Therefore

$$\begin{aligned} E[X_t \,|\, X_s] &= E\left[X_s + \sum_{j=s+1}^{t} S_j \middle| X_s \right] \\ &= E[X_s \,|\, X_s] + \sum_{j=s+1}^{t} E\left[S_j \,|\, X_s\right] \\ &= X_s + \sum_{j=s+1}^{t} E[S_j] \\ &= X_s + 0 \\ &= X_s. \end{aligned}$$

What we have proven is that the process $\{X_t\}$ defined by this game is something called a *martingale*. Informally, a process is a martingale if, given the state(s) of the process up to and including some time s (you think of time s as the "present time"), the expected state of the process at a time $t \geq s$ (think of t as a "future time") is equal to X_s.

Unfortunately, to define this formally in a way that is useful for deriving formulas, proving theorems, etc., we need some additional machinery.

10.2 Filtrations

σ-algebras

<u>Goal of this section</u>: We want to develop mathematical language and notation that will allow us to rigorously define what is meant in general by a *strategy*, and what is meant in general by a *stopping time*.

To do this, we will need to further explore the language of σ-algebras first encountered back in Chapter 1:

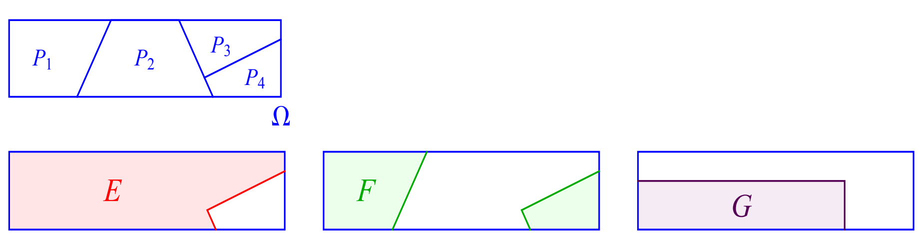
Definition 10.1 *Let Ω be a set. A nonempty collection $\mathcal{F}$ of subsets of Ω is called a $\sigma-$**algebra** (a.k.a. $\sigma-$**field**) if*

1. *$\mathcal{F}$ is "closed under complements", i.e. whenever $E \in \mathcal{F}$, $E^C \in \mathcal{F}$.*
2. *$\mathcal{F}$ is "closed under finite and countable unions and intersections", i.e. whenever $E_1, E_2, E_3, ... \in \mathcal{F}$, both $\bigcup\limits_j E_j$ and $\bigcap\limits_j E_j$ belong to $\mathcal{F}$ as well.*

Theorem 10.2 *Let $\mathcal{F}$ be a $\sigma-$algebra on set Ω. Then $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$.*

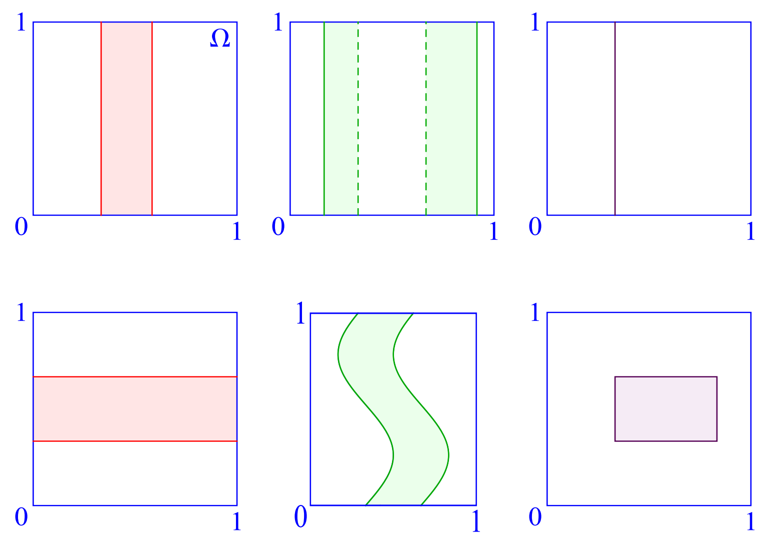
EXAMPLES OF σ-ALGEBRAS

1. Let Ω be any set. Let $\mathcal{F} = \{\emptyset, \Omega\}$. This is called the **trivial σ-algebra** of Ω.
2. Let Ω be any set. Let $\mathcal{F} = 2^\Omega$ be the set of all subsets of Ω. This is called the **power set** of Ω.
3. Let Ω be any set and let $\mathcal{P} = \{P_1, P_2, ..., P_n\}$ be any partition of Ω (that is, that $P_i \cap P_j = \emptyset$ for all $i \neq j$ and $\bigcup_j P_j = \Omega$). Then let $\mathcal{F}$ be the collection of all sets which are unions of some number of the P_j. This $\mathcal{F}$ is called the **σ-algebra generated by** $\mathcal{P}$.



4. Let $\Omega = [0,1] \times [0,1]$.

   - Let $\mathcal{F}_1$ be the trivial σ-algebra of Ω.
   - Let $\mathcal{F}_2$ be the collection of all subsets of Ω of the form $A \times [0,1]$ where $A \subset [0,1]$.
   - Let $\mathcal{F}_3$ be the power set of Ω.



Suppose $\omega = (x, y) \in \Omega$.

1. If you know all the sets in $\mathcal{F}_1$ to which ω belongs, what do you know about ω?

2. If you know all the sets in $\mathcal{F}_2$ to which ω belongs, what do you know about ω?

3. If you know all the sets in $\mathcal{F}_3$ to which ω belongs, what do you know about ω?

Measurability

Definition 10.3 *Let Ω be a set and let $\mathcal{F}$ be a σ-algebra on Ω.*

*A subset E of Ω is called $\mathcal{F}-$***measurable** *(or just* **measurable***) if $E \in \mathcal{F}$.*

*A function (i.e. a r.v.) $X : \Omega \to \mathbb{R}$ is called $\mathcal{F}$***-measurable** *if for any open interval $(a, b) \subseteq \mathbb{R}$, the set*

$$X^{-1}(a, b) = \{\omega \in \Omega : X(\omega) \in (a, b)\}$$

is $\mathcal{F}$-measurable.

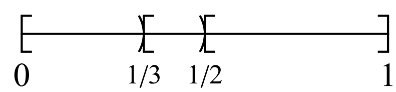
How to interpret this: Think of a σ-algebra $\mathcal{F}$ as revealing some partial information about an ω (i.e. it tells you which sets in $\mathcal{F}$ to which ω belongs, but not necessarily exactly what ω is).

To say that a function X is $\mathcal{F}$-measurable means that the evaluation of $X(\omega)$ depends only on the information contained in $\mathcal{F}$.

EXAMPLE

Let $\Omega = [0, 1]$ and let $\mathcal{F}$ be the σ-algebra generated by the partition

$$\mathcal{P} = \{[0, 1/3), [1/3, 1/2), [1/2, 1]\}.$$



Determine whether each of these functions X is $\mathcal{F}$-measurable:

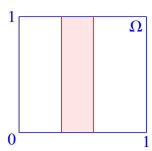
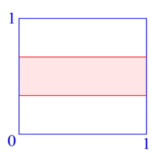
1. $X : \Omega \to \mathbb{R}$ defined by $X(\omega) = 2\omega$.

2. $X : \Omega \to \mathbb{R}$ defined by $X(\omega) = 2$.

3. $X : \Omega \to \mathbb{R}$ defined by $X(\omega) = \begin{cases} 1 & \text{if } \omega < \frac{1}{2} \\ 0 & \text{else} \end{cases}$.

More generally: if $\mathcal{F}$ is generated by a partition $\mathcal{P}$, a r.v. X is measurable if and only if it is constant on each of the partition elements; in other words, if $X(\omega)$ depends not on ω but only on which partition element ω belongs to.

Throughout this chart, let $\Omega = [0,1] \times [0,1]$, so $\omega \in \Omega \leftrightarrow \omega = (x, y)$.

σ-algebra $\mathcal{F}$	information $\mathcal{F}$ reveals about ω	description of $\mathcal{F}$-measurable functions
trivial σ-algebra $\mathcal{F}_{\text{triv}} = \{\emptyset, \Omega\}$	nothing	
$\mathcal{F}_x$ = sets of form $A \times [0,1]$ 	the x-coordinate of ω	
$\mathcal{F}_y$ = sets of form $[0,1] \times A$ 		
power set $\mathcal{F} = 2^{\Omega}$	everything (x and y)	

Note:

Filtrations

Definition 10.4 *Let Ω be a set and let $\mathcal{I} \subseteq [0, \infty)$.*

A **filtration** $\{\mathcal{F}_t\}_{t \in \mathcal{I}}$ *on Ω is a sequence of σ-algebras indexed by elements of $\mathcal{I}$ which is increasing, i.e. if $s, t \in \mathcal{I}$, then*

$$s \leq t \Rightarrow \mathcal{F}_s \subseteq \mathcal{F}_t.$$

Idea: for any filtration $\{\mathcal{F}_t\}$, when $s \leq t$, each $\mathcal{F}_s$-measurable set is also $\mathcal{F}_t$-measurable, so *as t increases, there are more $\mathcal{F}_t$-measurable sets*. Put another way, as t increases you get more information about the points in Ω.

Definition 10.5 *Let $\{X_t\}_{t \in \mathcal{I}}$ be a stochastic process with index set $\mathcal{I}$.*

The **natural filtration** *of $\{X_t\}$ is described by setting*

$$\mathcal{F}_t = \{\textit{events which are characterized only by the values of } X_s \textit{ for } 0 \leq s \leq t\}.$$

Idea: in the context of gambling, think of points in Ω as a list which records the outcome of every bet you make. $\mathcal{F}_t$ is the σ-algebra that *gives you the result of the first t bets*; as t increases, you get more information about what happens.

EXAMPLE

Flip a fair coin twice, start with $\$10$ and bet $\$1$ on the first flip and $\$3$ on the second flip. Let X_t be your bankroll after the t^{th} flip (where $t \in \mathcal{I} = \{0, 1, 2\}$). Describe the filtration $\{\mathcal{F}_0, \mathcal{F}_1, \mathcal{F}_2\}$.

Strategies

Definition 10.6 *Let $\{\mathcal{F}_t\}$ be the natural filtration of stochastic process $\{X_t\}$.*

A **strategy** *for $\{X_t\}$ is another stochastic process $\{B_t\}$ such that for all $s < t$, B_t is $\mathcal{F}_s$-measurable.*

Idea: Suppose you are betting on repeated coin flips and you decide to implement a strategy where B_t is the amount you are going to bet on the t^{th} flip.

- If you own a time machine, you would just go forward in time to see what the coin flips to, bet on that, and win.
- But if you don't own a time machine, the amount B_t you bet on the t^{th} flip is **only allowed to depend on information coming from flips before the t^{th} flip**, i.e. B_t is only allowed to depend on information coming from X_s for $s < t$, i.e. B_t must be $\mathcal{F}_s$-measurable for all $s < t$.

Remark: If the index set $\mathcal{I}$ is discrete, then a process $\{B_t\}$ is a strategy for $\{X_t\}$ if and only if for every t, B_t is $\mathcal{F}_{t-1}$-measurable.

Examples of strategies

Suppose you are betting on repeated coin flips. Throughout these examples, let's use the following notation to keep track of whether you win or lose each game:

$$X_0 = \text{your initial bankroll}$$
$$X_t = \begin{cases} X_{t-1} + 1 & \text{if you win the } t^{th} \text{ game} \\ X_{t-1} - 1 & \text{if you lose the } t^{th} \text{ game} \end{cases}$$
$$S_t = X_t - X_{t-1} = \begin{cases} 1 & \text{if you win the } t^{th} \text{ game} \\ -1 & \text{if you lose the } t^{th} \text{ game} \end{cases}$$

So $\{X_t\}$ would measure your bankroll after t games, **if you are betting $\$1$ on each game**. However, you may want to bet more or less than $\$1$ on each game (varying your bets according to some "strategy"). The idea is that B_t will be the amount you bet on the t^{th} game.

Strategy 1: Bet \$1 on each flip.

Strategy 2: Alternate between betting \$1 and betting \$2.

Strategy 3: Start by betting \$1 on the first flip.
After that, bet \$2 if you lost the previous flip, and bet \$1 otherwise.

Strategy 4: Bet \$1 on the first flip.
If you lose, double your bet after each flip you lose until you win once.
Then go back to betting \$1 and repeat the procedure.

"Strategy" 5: Bet \$5 on the n^{th} flip if you are going to win the n^{th} flip, and bet \$1 otherwise.

Suppose we implement arbitrary strategy $\{B_t\}$ when playing this game. Then our bankroll after t games isn't measured by $\{X_t\}$ any longer; it is

Definition 10.7 *Let $\{X_t\}_{t\in\mathcal{I}}$ be a discrete-time stochastic process.*

The t^{th} **step** *of $\{X_t\}$ is $S_t = X_t - X_{t-1}$ for all t.*

Given a strategy $\{B_t\}$ for $\{X_t\}$, the **transform of** $\{X_t\}$ **by** $\{B_t\}$ *is the process $\{(B\cdot X)_t\}_{t\in\mathbb{N}}$ defined by*

$$(B\cdot X)_t = X_0 + B_1S_1 + B_2S_2 + ... + B_tS_t = X_0 + \sum_{j=1}^{t} B_jS_j.$$

Idea: If you use strategy $\{B_t\}$ to play game $\{X_t\}$, then your bankroll after t games is $(B\cdot X)_t$.

Note: $(B\cdot X)_0 = X_0$.

EXAMPLE

Suppose you implement Strategy 4 as described above (double bet when you lose; reset bet size to $\$1$ when you win). If your initial bankroll is \$50, and the results of the first eight flips are H T T H T T T H, give the values of B_t, X_t, S_t and $(B\cdot X)_t$ for $0\le t\le 8$.

time t	bet size B_t	"W-L" record X_t	result of the t^{th} game S_t	bankroll using strategy $\{B_t\}$ $(B\cdot X)_t$
0	DNE	50	DNE	50
1				
2				
3				
4				
5				
6				
7				
8				
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Stopping times

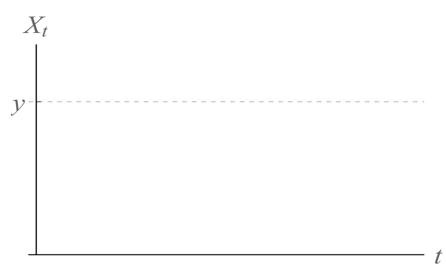
Definition 10.8 *Let $\{X_t\}_{t\in\mathcal{I}}$ be a stochastic process with standard filtration $\{\mathcal{F}_t\}$. A r.v. $T : \Omega \to \mathbb{R} \cup \{\infty\}$ is called a* **stopping time (for $\{X_t\}$)** *if for every $a \in \mathbb{R}$, the set of sample functions satisfying $T \leq a$ is $\mathcal{F}_a$-measurable.*

Idea: In other words, T is a stopping time if you can determine whether or not $T \leq a$ solely by looking at the values of X_t for $t \leq a$.

In the context of playing a game over and over, think of T as a "trigger" which causes you to stop playing the game. Thus you would walk away from the table with winnings given by X_T (or, if you are employing strategy $\{B_t\}$, your winnings would be $(B \cdot X)_T$).

EXAMPLES

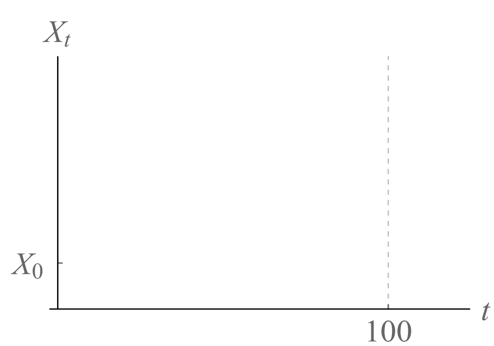
- $T = T_y = \min\{t \geq 0 : X_t = y\}$



- $T = \min\{t > 0 : X_t = X_0\}$

NON-EXAMPLE

- $T = \min\{t \geq 0 : X_t = \max\{X_s : 0 \leq s \leq 100\}\}$



RECALL OUR BIG PICTURE QUESTION

Is there a strategy under which you can beat a fair game?

Using the language of this section, we can restate this more formally:

BIG PICTURE QUESTION (PHRASED WITH MORE PRECISE LANGUAGE)

Suppose process $\{X_t\}$ represents a fair game.

- Is there a predictable sequence $\{B_t\}$ for this process, and a stopping time T so that $E[(B \cdot X)_T] > X_0$?
- If so, what $\{B_t\}$ and what T maximizes $E[(B \cdot X)_T]$?

10.3 Conditional expectation and martingales

Goal of this section: We want to give a formal definition of what it means for a process to represent a "fair game". Such a process will be called a *martingale*.

RECALL FROM CHAPTER 5

We learned how to define the *conditional expectation* of one r.v. X given another Y:

We learned several properties of conditional expectation in MATH 414:

- it's linear;
- it preserves constants and inequalities;
- you can pull out what's known (a.k.a. stability);
- it simplifies for independent r.v.s;
- etc.

Here is a new property we didn't discuss back in Chapter 5:

Theorem 10.9 *Given any bounded, continuous function $\phi : \mathbb{R} \to \mathbb{R}$,*

$$E[X\,\phi(Y)] = E\left[E(X|Y)\,\phi(Y)\right].$$

PROOF (when X, Y continuous):

$$\begin{aligned}
E[X\,\phi(Y)] &= \int\int x\,\phi(y) f_{X,Y}(x,y)\,dA && \text{(LOTUS)} \\
&= \int\int x\,\phi(y) f_{X|Y}(x|y) f_Y(y)\,dA && \text{(multiplication principle)} \\
&= \int\int x f_{X|Y}(x|y)\phi(y) f_Y(y)\,dx\,dy \\
&= \int \left(\int x f_{X|Y}(x|y)\,dx\right) \phi(y) f_Y(y)\,dy \\
&= \int E(X|Y)(y)\,\phi(y) f_Y(y)\,dy && \text{(def'n of cond'l expectation)} \\
&= E\left[E(X|Y)\,\phi(Y)\right] && \text{(LOTUS).}
\end{aligned}$$

If X, Y are discrete, the proof is similar, but has sums instead of integrals. □

Now, we are going to define something new called the conditional expectation of a r.v. *with respect to a σ-algebra* (as opposed to a second r.v.). To do this, we use the property of Theorem 10.9 to motivate a definition:

Definition 10.10 *Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $X : \Omega \to \mathbb{R}$ be a $\mathcal{F}$-measurable r.v. and let $\mathcal{G} \subseteq \mathcal{F}$ be a sub σ-algebra.*

The **conditional expectation of X given** *$\mathcal{G}$ is a function $E(X|\mathcal{G}) : \Omega \to \mathbb{R}$ with the following two properties:*

1. *$E(X|\mathcal{G})$ is $\mathcal{G}$-measurable, and*

2. *for any bounded, $\mathcal{G}$-measurable r.v. $Z : \Omega \to \mathbb{R}$, $E[XZ] = E\left[E(X|\mathcal{G})\,Z\right]$.*

Remarks:

1. These conditional expectations always exist, and are always unique up to sets of probability zero.

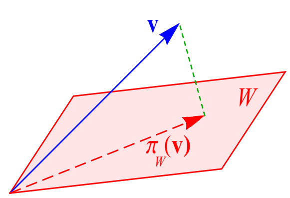
2. By setting $Z = 1$, we see that

$$E[X] = E[E[X|\mathcal{G}]].$$

   This gives you the idea behind this type of conditional expectation: *$E[X|\mathcal{G}]$ is a $\mathcal{G}$-mble r.v. with the same expected value(s) as the original r.v. X.*

Properties of conditional expectation

To understand conditional expectation, it is useful to understand that this is an idea parallel to a concept from linear algebra: *projection*.

	PROJECTION	CONDITIONAL EXPECTATION
vector space	$\mathbb{R}^n$	
description of vectors	traditional vectors $\mathbf{v} = (v_1, \ldots, v_n)$	
notion of “dot product” of two vectors	$\mathbf{v} \cdot \mathbf{w} = \sum_{j=1}^{n} v_j w_j$	
notion of norm/size	$\|\|\mathbf{v} \cdot \mathbf{w}\|\| = \sqrt{\mathbf{v} \cdot \mathbf{v}} = \sqrt{\sum_{j=1}^{n} v_j^2}$	
notion of orthogonality	perpendicularity: $\mathbf{v} \perp \mathbf{w} \Leftrightarrow \mathbf{v} \cdot \mathbf{w} = 0$ $\mathbf{v} \perp W \Leftrightarrow$ $\mathbf{v} \perp \mathbf{w}$ for all $\mathbf{w} \in W$	
subspace	W	
picture		
object under consideration	$\pi_W(\mathbf{v}) =$ vector in W that is closest to $\mathbf{v}$	

	PROJECTION	CONDITIONAL EXPECTATION
formula for the object	$\pi_W(\mathbf{v}) = \sum_j (\mathbf{v} \cdot \mathbf{x}_j)\mathbf{x}_j$ where $\{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ is an orthonormal basis of W (the orthonormal basis comes from Gram-Schmidt)	
orthogonal decomposition	$\mathbf{v} = \pi_W(\mathbf{v}) + \pi_{W^\perp}(\mathbf{v})$ where $\pi_W(\mathbf{v}) \in W$ and $\pi_{W^\perp}(\mathbf{v}) \in W^\perp$. Equivalently, $\mathbf{v} - \pi_W(\mathbf{v}) \in W^\perp$.	
dot product with vector in subspace	If $\mathbf{w} \in W$, then $\mathbf{v} \cdot \mathbf{w}$ $= (\pi_W(\mathbf{v}) + \pi_{W^\perp}(\mathbf{v})) \cdot \mathbf{w}$ $= \pi_W(\mathbf{v}) \cdot \mathbf{w} + \pi_W(\mathbf{v}) \cdot \mathbf{w}$ $= \pi_W(\mathbf{v}) \cdot \mathbf{w} + 0$ $= \pi_W(\mathbf{v}) \cdot \mathbf{w}.$	
linearity properties	$\pi_W(\mathbf{v}_1 + \mathbf{v}_2)$ $= \pi_W(\mathbf{v}_1) + \pi_W(\mathbf{v}_2)$ $\pi_W(r\,\mathbf{v}_1) = r\,\pi_W(\mathbf{v}_1)$	
tower property	If $W_2 \subseteq W_2$, then $\pi_{W_2} \circ \pi_{W_1}(\mathbf{v}) = \pi_{W_2}(\mathbf{v})$.	
stability	If $\mathbf{w} \in W$, then $\pi_W(\mathbf{w}) = \mathbf{w}$ and $\pi_W(\mathbf{v} + \mathbf{w}) = \pi_W(\mathbf{v}) + \mathbf{w}$.	
orthogonality property	If $\mathbf{v} \in W^\perp$, then $\pi_W(\mathbf{v}) = \mathbf{0}$.	

Our observations in the previous two pages lead us to some of these properties of conditional expectaiton; others are new. These properties are widely used (but proving them is beyond the scope of this class).

Theorem 10.11 (Properties of conditional expectation) *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\mathcal{G}$ and $\mathcal{H}$ be $\sigma-$algebras on Ω.*

Suppose $X, Y : \Omega \to \mathbb{R}$ are $\mathcal{F}$-measurable r.v.s, and let $a, b \in \mathbb{R}$. Then, with probability one, all these statements hold:

Positivity: *If $X \geq c$, then $E(X|\mathcal{G}) \geq c$.*

Linearity: $E[aX + bY|\mathcal{G}] = aE[X|\mathcal{G}] + bE[Y|\mathcal{G}]$.

Stability (pulling out what's known): *If X is $\mathcal{G}$-measurable, then*

$$E[X|\mathcal{G}] = X \quad \text{and} \quad E[XY|\mathcal{G}] = X\, E[Y|\mathcal{G}].$$

Independence property: *If $X \perp \mathcal{G}$, then $E[X|\mathcal{G}] = EX$.*

Tower property: *If $\mathcal{H} \subseteq \mathcal{G}$, then $E[E(X|\mathcal{G})|\mathcal{H}] = E[X|\mathcal{H}]$.*

Law of Total Expectation: $E[E(X|\mathcal{G})] = EX$.

Constants are preserved: $E[a|\mathcal{G}] = a$.

An important use of conditional expectation is to define a *martingale*, which is a mathematical formulation of a fair game:

Definition 10.12 *Let $\{X_t\}_{t\in\mathcal{I}}$ be a stochastic process with natural filtration $\{\mathcal{F}_t\}$. The process $\{X_t\}$ is called a* **martingale** *if for every $s \leq t$ in $\mathcal{I}$,*

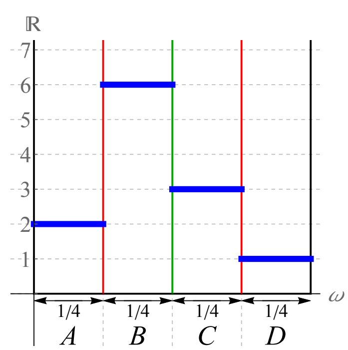
$$E[X_t \,|\, \mathcal{F}_s] = X_s.$$

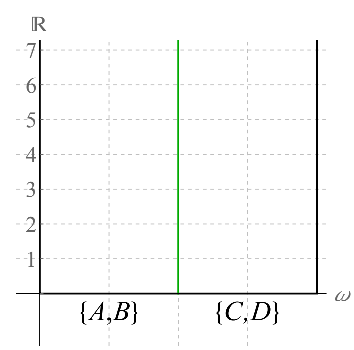
Note: To prove that a process $\{X_t\}$ is a martingale, typically you need to show that the equation in Definition 10.12 holds. This involves applying the properties given in Theorem 10.11.

Example computations

EXAMPLE 1

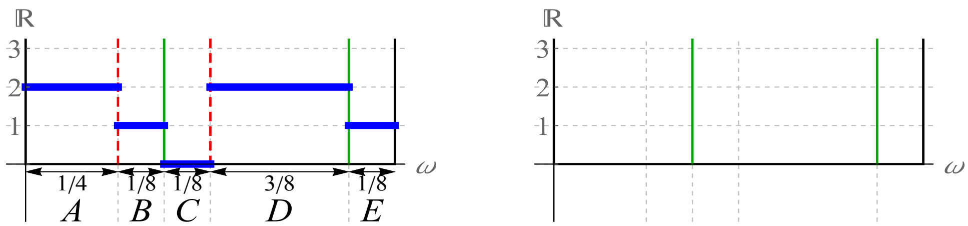
Let $\Omega = \{A, B, C, D\}$; let $\mathcal{F} = 2^\Omega$; let $P \sim Unif(\Omega)$. Let $\mathcal{G}$ be the σ-algebra generated by the partition $\mathcal{P} = \{\{A, B\}, \{C, D\}\}$. Let $X : \Omega \to \mathbb{R}$ be defined by $X(A) = 2$; $X(B) = 6$; $X(C) = 3$; $X(D) = 1$. Compute $E[X|\mathcal{G}]$.





Example 2

Let $\Omega = \{A, B, C, D, E\}$; let $\mathcal{F} = 2^{\Omega}$; let $P(A) = \frac{1}{4}$; $P(B) = P(C) = P(E) = \frac{1}{8}$; $P(D) = \frac{3}{8}$. Let $\mathcal{G}$ be generated by the partition $\mathcal{P} = \{\{A, B\}, \{C, D\}, \{E\}\}$. Let $X(A) = X(D) = 2$; $X(C) = 0$; $X(B) = X(E) = 1$. Compute $E[X|\mathcal{G}]$.



Example 3

Let $\Omega = [0, 1] \times [0, 1]$; let $\mathcal{F} = 2^{\Omega}$; let P be the uniform distribution. Let $\mathcal{G}$ be the σ-algebra of vertical sets (i.e. sets of the form $A \times [0, 1]$). Let $X : \Omega \to \mathbb{R}$ be $X(x, y) = x + y$. Compute $E[X|\mathcal{G}]$.

Properties of martingales

Theorem 10.13 (Characterization of discrete-time martingales) *A discrete-time process $\{X_t\}_{t\in\mathbb{N}}$ is a martingale if and only if $E[X_{t+1}|\mathcal{F}_t] = X_t$ for every $t \in \mathbb{N}$.*

PROOF We use the tower property of conditional expectation. Let $s \leq t$. Then

$$\begin{aligned} E[X_t|\mathcal{F}_s] &= E[E[\cdots E[E[E[X_t|\mathcal{F}_{t-1}]|\mathcal{F}_{t-2}]|\mathcal{F}_{t-3}]\cdots|\mathcal{F}_{s+1}]|\mathcal{F}_s] \\ &= E[E[\cdots E[E[X_{t-1}|\mathcal{F}_{t-2}]|\mathcal{F}_{t-3}]\cdots|\mathcal{F}_{s+1}]|\mathcal{F}_s] \\ &= E[E[\cdots E[X_{t-2}|\mathcal{F}_{t-3}]\cdots|\mathcal{F}_{s+1}]|\mathcal{F}_s] \\ &= \ldots \\ &= E[X_{s+1}|\mathcal{F}_s] \\ &= X_s. \end{aligned}$$

By definition, $\{X_t\}$ is a martingale. □

Theorem 10.14 (Properties of the steps of a discrete-time martingale) *Suppose $\{X_t\}_{t\in\mathbb{N}}$ is a martingale with natural filtration $\{\mathcal{F}_t\}$. Define $S_t = X_t - X_{t-1}$; S_t is called the t^{th}* **step** *of the martingale $\{X_t\}$. Then, for all t:*

1. $X_t = X_0 + \sum\limits_{j=1}^{t} S_t$;
2. S_t *is* $\mathcal{F}_t$*-measurable;*
3. $E[S_{t+1}|\mathcal{F}_t] = 0$.

PROOF First, statement (1):

$$\begin{aligned} X_t &= X_0 + (X_1 - X_0) + (X_2 - X_1) + \ldots + (X_t - X_{t-1}) \\ &= X_0 + S_1 + \ldots + S_t \\ &= X_0 + \sum_{j=1}^{t} S_j \end{aligned}$$

Statement (2) is clear, since both X_t and X_{t-1} are $\mathcal{F}_t$-measurable.

Last, statement (3):

$$\begin{aligned}
E[S_{t+1}|\mathcal{F}_t] &= E[X_{t+1} - X_t|\mathcal{F}_t] \\
&= E[X_{t+1}|\mathcal{F}_t] - E[X_t|\mathcal{F}_t] \\
&= X_t - E[X_t|\mathcal{F}_t] \qquad \text{(since } \{X_t\} \text{ is a martingale)} \\
&= X_t - X_t \qquad \text{(by stability)} \\
&= 0. \ \square
\end{aligned}$$

Theorem 10.15 (Transforms of martingales are martingales) *Let $\{X_t\}_{t\in\mathbb{N}}$ be a martingale and suppose that $\{B_t\}$ is a strategy for $\{X_t\}$. Then the transform $\{(B \cdot X)_t\}$ is also a martingale.*

PROOF

$$\begin{aligned}
&E\left[(B \cdot X)_{t+1}|\mathcal{F}_t\right] \\
&= E\left[X_0 + \sum_{j=1}^{t+1} B_j S_j|\mathcal{F}_t\right] \qquad \text{(by the definition of } (B \cdot X)) \\
&= E[X_0|\mathcal{F}_t] + \sum_{j=1}^{t} E[B_j S_j|\mathcal{F}_t] + E[B_{t+1}S_{t+1}|\mathcal{F}_t] \qquad \text{(by linearity)} \\
&= X_0 + \sum_{j=1}^{t} B_j S_j + B_{t+1}E[S_{t+1}|\mathcal{F}_t] \qquad \text{(by stability)} \\
&= X_0 + \sum_{j=1}^{t} B_j S_j + B_{t+1}(0) \qquad \text{(by (3) of Thm 10.14)} \\
&= X_0 + \sum_{j=1}^{t} B_j S_j \\
&= (B \cdot X)_t. \qquad \text{(by the definition of } (B \cdot X))
\end{aligned}$$

By Theorem 10.13, $\{(B \cdot X)_t\}$ is a discrete-time martingale. $\square$

10.4 Optional Stopping Theorem

Our "big picture" questions

Suppose $\{X_t\}$ is a martingale.

- Is there a predictable sequence $\{B_t\}$ for this process, and a stopping time T so that $E[(B \cdot X)_T] > X_0$?
- If so, what $\{B_t\}$ and what T maximizes $E[(B \cdot X)_T]$?

We are ready to prove a theorem that says the answer to our first question is

Theorem 10.16 (Optional Stopping Theorem (OST)) *Let $\{X_t\}$ be a martingale and let T be a bounded[a] stopping time. Then*

$$E[X_T] = E[X_0].$$

[a] To say T is **bounded** means there is a constant n such that $P(T \leq n) = 1$.

The OST is also called the **Optional Sampling Theorem** because of its applications in statistics.

Proof Let $B_t = \begin{cases} 1 & \text{if } t \leq T \\ 0 & \text{else} \end{cases}$.

(This defines a strategy(?) in which you bet $\$1$ on each game game until the stopping time T hits, and then stop playing.)

$$\begin{aligned} T \text{ is a stopping time} \quad &\Rightarrow \quad G = \{T \leq t-1\} = \{B_t = 0\} \text{ is } \mathcal{F}_{t-1}\text{-measurable} \\ &\Rightarrow \quad G^C = \{T \geq t\} = \{B_t = 1\} \text{ is } \mathcal{F}_{t-1}\text{-measurable} \\ &\Rightarrow \quad \text{each } B_t \text{ is } \mathcal{F}_{t-1}\text{-measurable} \\ &\Rightarrow \quad \{B_t\} \text{ is a strategy for } \{X_t\}. \end{aligned}$$

Now, we are assuming T is bounded; let n be such that $P(T \leq n) = 1$.
For any $t \geq n$, we have

$$\begin{aligned} (B \cdot X)_t &= X_0 + \sum_{j=1}^{t} B_t S_t \\ &= X_0 + \sum_{j=1}^{t} B_t (X_t - X_{t-1}) \\ &= X_0 + 1(X_1 - X_0) + 1(X_2 - X_1) + ... + 1(X_T - X_{T-1}) \\ &\quad + 0(X_{T+1} - X_t) + 0(X_{T+2} - X_{T+1}) + ... \\ &= X_T. \end{aligned}$$

$$\text{Finally,} \quad \begin{aligned} EX_T &= E[(B \cdot X)_t] \\ &= E[(B \cdot X)_0] \quad \text{(since } \{(B \cdot X)_t\} \text{ is a martingale)} \\ &= EX_0. \ \square \end{aligned}$$

We will need the following "tweaked version" of the OST, which requires a little less about T (it only has to be finite rather than bounded) but a little more about $\{X_t\}$ (the values of X_t have to be bounded until T hits):

Theorem 10.17 (OST (tweaked version)) *Let $\{X_t\}$ be a martingale. Let T be a stopping time for $\{X_t\}$ which is finite with probability one. If there is a fixed constant C such that for sufficiently large n, $T \geq n$ implies $|X_n| \leq C$, then*

$$E[X_T] = E[X_0].$$

PROOF Choose a sufficiently large n and let $\overline{T} = \min(T, n)$. $\overline{T}$ is a stopping time which is bounded by n, so the original OST applies to $\overline{T}$, i.e. $\boxed{EX_{\overline{T}} = EX_0}$.

Now

$$\begin{aligned} |EX_T - EX_0| &= |EX_T - EX_{\overline{T}}| \\ &\leq E\,|X_T - X_{\overline{T}}| \qquad \text{(by } \triangle \text{ inequality)} \end{aligned}$$

Corollary 10.18 (You can't beat a fair game) *Let $\{X_t\}$ be a martingale. Let T be a finite stopping time for $\{X_t\}$ and let $\{B_t\}$ be any bounded strategy for $\{X_t\}$. Then*

$$E(B \cdot X)_T = EX_0.$$

PROOF If $\{X_t\}$ is a martingale, so is $(B \cdot X)_t$. Therefore by the tweaked OST,

$$E(B \cdot X)_T = E(B \cdot X)_0 = EX_0. \square$$

Catch: If you are willing to play forever, and/or you are willing to lose a possibly unbounded amount of money first, the OST doesn't apply, and you can beat a fair game using Strategy 4 described several pages ago. But this isn't realistic if you are a human with a finite lifespan and finite wealth.

APPLICATION

Suppose a gambler has $\$50$ and chooses to play a fair game repeatedly until either the gambler's bankroll is up to $\$100$, or until the gambler is broke.

If the gambler bets all $\$50$ on one game, then the probability he leaves a winner is $\frac{1}{2}$. What if the gambler bets in some other way?

10.5 Escape problems

MOTIVATING PROBLEM

You own a share of stock. Currently, it is worth $\$50$, and you expect that as time passes, the value of this stock will go up and down "randomly" according to some mathematical model. You have an investment strategy where, if at some point in the future the stock is worth $\$80$, you will sell it and "cash out while you are ahead", but if at some point in the future the stock price drops to $\$35$, you will sell it, "cutting your losses".

Relevant questions in this setting:

- How long are you going to hold the stock? More specifically:
  - What is the probability that you actually sell the stock?
  - What is the expected amount of time you hold the stock before selling? (also the variance, MGF, density function, etc.)
- When you sell the stock, what is the distribution of the price you sell it for?
  - What is the probability you sell the stock for $\$80$ (as opposed to $\$35$)?
  - What is the expected amount you sell the stock for? (also the variance, MGF, etc.)

Escaping processes

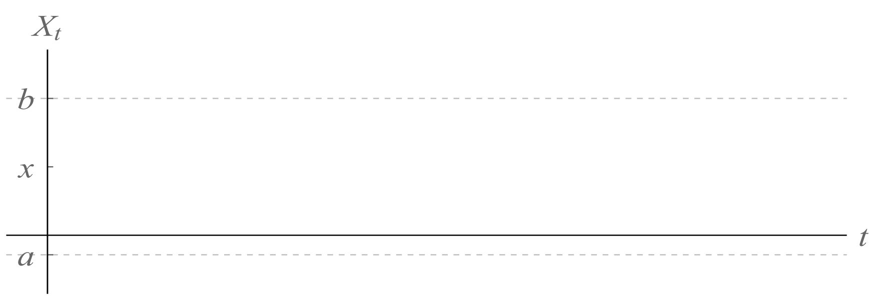
Definition 10.19 *Let $\{X_t\}$ be a stochastic process with state space $\mathcal{S} \subseteq \mathbb{R}$. We say that $\{X_t\}$ is* **escaping** *if for every $a, x, b \in \mathcal{S}$ with $a < x < b$,*

1. $P_x(T_{\{a,b\}} < \infty) = 1$; *and*
2. $P_x(X_t \in (a, b) \,|\, t < T_{\{a,b\}}) = 1$.

Recall: $T_A = \min\{t \geq 1 : X_t \in A\}$. is the **hitting time to** A. (If $A = \{a\}$, then T_a is short for $T_{\{a\}}$.)

Interpretation: When $\{X_t\}$ first hits state a or b, we think of the process as having "escaped" the interval (a, b) that it starts in. To say that a process is *escaping* means

- with probability 1, the process escapes (this is condition (1) above), and
- when the process escapes, it hits a or b (and doesn't jump over them... this is condition (2) above).



Questions we care about: Let $\{X_t\}$ be an escaping process with $a < x < b$. Let $T = T_{\{a,b\}} = \min\{T_a, T_b\}$. T is called an **escape time**.

$$P_x(X_T = a) = P_x(T_a < T_b) = ? \qquad E[X_T] = ? \qquad ET = ?$$

Remark: For an escaping process starting at $x \in (a, b)$, the events

$$T_a < T_b \quad \text{and} \quad T_b < T_a$$

are complements, so

$$P_x(T_a < T_b) = 1 - P_x(T_b < T_a). \tag{10.1}$$

Escape Time Theorem

We can apply martingale theory to the questions we are interested in:

Theorem 10.20 (Escape Time Theorem) *Let $\{X_t\}$ be an escaping process. If $\psi : \mathbb{R} \to \mathbb{R}$ is a function so that $\{\psi(X_t)\}$ is a martingale, then for all $a < x < b$,*

$$P_x(T_a < T_b) = \frac{\psi(b) - \psi(x)}{\psi(b) - \psi(a)} \quad \textit{and} \quad P_x(T_b < T_a) = \frac{\psi(x) - \psi(a)}{\psi(b) - \psi(a)}.$$

PROOF Let $T = T_{\{a,b\}}$. This is a hitting time that is finite with probability 1. Also, whenever $t < T_{\{a,b\}}$, $X_t \in (a, b)$. Therefore, the (tweaked version of the) OST applies to the martingale $\{\psi(X_t)\}$ to give

$$E[\psi(X_T)] = E[\psi(X_T)] \overset{\text{OST}}{=} E[\psi(X_0)] = E[\psi(x)] = \psi(x). \tag{10.2}$$

At the same time, given $X_0 = x$, X_T has the following density:

y		
$P_x(X_T = y)$		

Therefore, we can directly compute $E[\psi(X_t)]$ by LOTUS:

$$\begin{aligned}
\psi(x) &= E[\psi(X_t)] && \text{(by eqn (10.2) above)} \\
&= \sum_y \psi(y) P_x(X_T = y) && \text{(LOTUS)} \\
&= \psi(a) \cdot P_x(X_T = a) + \psi(b) \cdot P_x(X_T = b) \\
&= \psi(a) \cdot P_x(T_a < T_b) + \psi(b) \cdot P_x(T_b < T_a) && \text{(by the density above)} \\
&= \psi(a)\, P_x(T_a < T_b) + \psi(b)\, [1 - P_x(T_a < T_b)] && \text{(by eqn (10.1) above)} \\
&= [\psi(a) - \psi(b)] P_x(T_a < T_b) + \psi(b).
\end{aligned}$$

Solve for $P_x(T_a < T_b)$ to get

$$P_x(T_a < T_b) = \frac{\psi(x) - \psi(b)}{\psi(a) - \psi(b)} = \frac{\psi(b) - \psi(x)}{\psi(b) - \psi(a)};$$

by the complement rule

$$P_x(T_b < T_a) = 1 - P_x(T_a < T_b) = \frac{\psi(x) - \psi(a)}{\psi(b) - \psi(a)}. \square$$

In the next few sections we will apply Theorem 10.20 to a variety of problems that involve computing $P_x(T_a < T_b)$ for escaping processes.

Applications to classification of Markov chains

RECALL

Let $\{X_t\}$ be a Markov chain with state space $\mathcal{S}$. Then:

$$\begin{aligned} f_{x,y} = P_x(T_y < \infty) &= P(\text{chain hits } y\text{, given that it starts at } x) \\ f_x = f_{x,x} &= P(\text{chain returns to } x) \\ x \text{ is recurrent } &\Leftrightarrow f_x = 1 \Leftrightarrow x \text{ must return to itself} \\ x \text{ is transient } &\Leftrightarrow f_x < 1 \Leftrightarrow x \text{ is not recurrent} \\ \{X_t\} \text{ is recurrent (transient) } &\Leftrightarrow \text{ every state } x \in \mathcal{S} \text{ is recurrent (transient)} \\ \{X_t\} \text{ is irreducible } &\Rightarrow \text{ either } \{X_t\} \text{ is recurrent, or } \{X_t\} \text{ is transient.} \end{aligned}$$

We will see that ideas associated to escape times will help us determine the class structure of certain Markov chains.

Theorem 10.21 (Escape Time Corollary) *Let $\{X_t\}$ be a Markov chain with $\mathcal{S} \subseteq \mathbb{Z}$, which is also an escaping process. If $\psi : \mathbb{R} \to \mathbb{R}$ is a function so that $\{\psi(X_t)\}$ is a martingale, then*

- *if $x < y$, then* $f_{x,y} = \lim_{a \to -\infty} P_x(T_y < T_a) = \lim_{a \to \infty} \dfrac{\psi(x) - \psi(a)}{\psi(y) - \psi(a)}$.
- *if $x > y$, then* $f_{x,y} = \lim_{b \to \infty} P_x(T_y < T_b) = \lim_{b \to \infty} \dfrac{\psi(b) - \psi(x)}{\psi(b) - \psi(y)}$.
- $f_{x,x} = P(x,x) + \sum_{y \neq x} f_{y,x}\, P(x,y)$.

PROOF For the first statement, let $x < y$ and $X_0 = x$.

Notice that as a decreases toward $-\infty$, the events $(T_y < T_a)$ increase:

$$(T_y < T_a) \subseteq (T_y < T_{a+1}).$$

Therefore, by continuity of probability measures (all the way back in Chapter 1), this means

$$\lim_{a \to -\infty} P_x(T_y < T_a) = P_x\left[\bigcup_{a<x} (T_y < T_a)\right].$$

But if the chain hits state y, then it must hit y before it hits some $a < x$, so $T_y < \infty$ if and only if $T_y < T_a$ for some $a < x$. So

$$f_{x,y} = P_x(T_y < \infty) = P_x\left[\bigcup_{a<x} (T_y < T_a)\right] = \lim_{a \to -\infty} P_x(T_y < T_a)$$

as wanted.

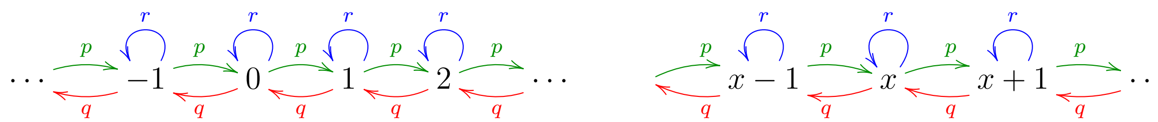
The second statement is a HW problem (the proof is similar to the proof of the first statement).

For the last statement, use the LTP based on the first step in the chain (similar to how we computed absorption probabilities back in Chapter 9):

$$\begin{aligned}
f_{x,x} &= P_x(T_x < \infty) \\
&= \sum_{y \in \mathcal{S}} P_x(T_x < \infty \,|\, X_1 = y)\, P_x(X_1 = y) \\
&= 1\, P_x(X_1 = x) + \sum_{y \neq x} P_x(T_x < \infty \,|\, X_1 = y)\, P(x,y) \\
&= P(x,x) + \sum_{y \neq x} f_{y,x}\, P(x,y). \ \square
\end{aligned}$$

10.6 Simple random walk on $\mathbb{Z}$

A simple random walk models a repeated game where you bet \$1 on each play; simple random walk is a Markov chain which has the following directed graph:



> **Definition 10.22** *A* **simple random walk on** $\mathbb{Z}$ *is a Markov chain* $\{X_t\}$ *with state space* $\mathbb{Z}$ *and transition probabilities*
>
> $$P(x,y) = \begin{cases} p & \text{if } y = x+1 \\ q & \text{if } y = x-1 \\ r & \text{if } y = x \end{cases}$$
>
> *where* p, q *and* r *are non-negative constants so that* $p + q + r = 1$.
>
> *Given a simple random walk, for each* $j \geq 1$ *we let the* j^{th} **step** *of the walk be the r.v.*
>
> $$S_j = X_j - X_{j-1}.$$

In a simple random walk, the steps S_j form an i.i.d. sequence $\{S_j\}$ of r.v.s, each having density

s	-1	0	1
$f_{S_j}(s) = P(S_j = s)$	q	r	p

.

We denote the mean and variance of S_j by

$$\mu = ES_j \quad \text{and} \quad \sigma^2 = Var(S_j).$$

We can write $X_t = X_{t-1} + S_t$; notice $S_t \perp \mathcal{F}_{t-1}$ where $\{\mathcal{F}_t\}$ is the natural filtration.

> **Definition 10.23** *A simple random walk on* $\mathbb{Z}$ *is called* **unbiased** *if* $p = q$ *and is called* **biased** *if* $p \neq q$. *A biased random walk is called* **positively biased** *if* $p > q$ *and* **negatively biased** *if* $p < q$.

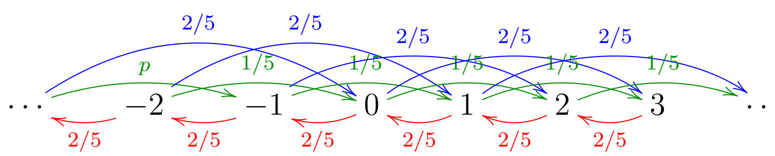
> **Lemma 10.24** *For a simple random walk,* $\mu = ES_j = p - q$. *If the simple random walk is unbiased, then* $\mu = 0$ *and* $\sigma^2 = Var(S_j) = p + q$.

PROOF HW

Remark: The reason these walks are called *simple* is that the steps take only the values $-1, 0$ and 1. For example, here is the directed graph of a (non-simple) random walk where the steps have the density

s	-1	2	3
$f_{S_j}(s) = P(S_j = s)$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{2}{5}$

:



Properties of simple random walk

Class structure

A simple random walk is irreducible if and only if $p > 0$ and $q > 0$. If $r = 0$, the walk has period 2; otherwise the walk is aperiodic.

Escaping property

Lemma 10.25 *Let $\{X_t\}$ be an irreducible simple random walk. Let $A = \{a, b\} \subseteq \mathbb{Z}$ and suppose $X_0 = x$ where $a < x < b$. Then $P(T_A < \infty) = 1$.*

PROOF Since $\{X_t\}$ is irreducible, $p > 0$.

Now let G_n be the event that between times $(n-1)(b-a)$ and $n(b-a)$, the chain always steps in the positive direction. In precise math notation,

$$G_n = \{S_j = 1 \, \forall j \in \{(n-1)(b-a)+1, (n-1)(b-a)+2, ..., n(b-a)\}\}.$$

Note that

1. $P(G_n) \geq p^{b-a} > 0$.
2. since G_j and G_k refer to disjoint blocks of time in the chain, $G_j \perp G_k$.

Thus

$$\begin{aligned} P(\text{no } G_n \text{ occurs}) = P\left(\bigcap_{n=1}^{\infty} G_n^C\right) &= \prod_{n=1}^{\infty} P(G_n^C) \qquad \text{(since the } G_n\text{s are } \perp) \\ &= \lim_{N\to\infty} \prod_{n=1}^{N} P(G_n^C) \\ &= \lim_{N\to\infty} (1 - p^{b-a})^N \\ &= 0 \qquad \text{(since } 1 - p^{b-a} \in (0,1)) \end{aligned}$$

So with probability 1, at least one G_n occurs.

This means that with probability 1, at some time in the future there will be $b-a$ consecutive steps in the positive direction, and that means that unless T_a has already occurred, after those $b-a$ consecutive steps, X_t will be $\geq b$.

Thus either T_a or T_b is finite, and therefore $P(T_A < \infty) = 1$. □

Corollary 10.26 *Simple random walks are escaping (unless $r = 1$).*

PROOF If $r \neq 1$, then either $p > 0$ or $q > 0$. To show condition (1) in the definition of escaping, there are three cases:

Case 1: $p > 0$ and $q > 0$. In this case, the walk is irreducible, so it is escaping by Lemma 10.25.

Case 2: $p > 0$ and $q = 0$. Here, the walk only steps to the right, so $f_{x,x+1} = 1$ so $f_{x,b} = 1$ for any $b > x$. Therefore $P_x(T_b < \infty) = 1$, so $P_x(T_{\{a,b\}} < \infty) = 1$.

Case 3: If $p = 0$ and $q > 0$. Here, the walk only steps to the left, so, $f_{x,x-1} = 1$ so $f_{x,a} = 1$ for any $a < x$. Therefore $P_x(T_a < \infty) = 1$, so $P_x(T_{\{a,b\}} < \infty) = 1$.

Condition (2) in the definition of escaping (when the walk leaves the interval (a, b), it must do so either at a or at b) follows from the fact that the random walk is simple (so it can't "jump over" a or b).

Associated martingales

Lemma 10.27 *Let $\{X_t\}$ be an irreducible simple random walk on $\mathbb{Z}$. Then, each of the following three processes is a martingale:*

- $\{X_t - t\mu\}$;
- $\{(X_t - t\mu)^2 - t\sigma^2\}$;
- $\left\{\left(\frac{q}{p}\right)^{X_t}\right\}$;

Remark: If $\{X_t\}$ is unbiased, the first bullet point tells you that $\{X_t\}$ is itself a martingale (since $\mu = 0$).

PROOF Throughout this proof, let $\{\mathcal{F}_t\}$ be the natural filtration of $\{X_t\}$ (thus also the natural filtration of all the processes in the lemma, since they are each a formula of $\{X_t\}$).

To show $\{X_t - t\mu\}$ is a martingale, let $Y_t = X_t - t\mu$. Since the index set is discrete, it is sufficient to show $E[Y_{t+1} \,|\, \mathcal{F}_t] = Y_t$. Toward that end:

$$\begin{aligned}
E[Y_{t+1} \,|\, \mathcal{F}_t] &= E[X_{t+1} - (t+1)\mu \,|\, \mathcal{F}_t] && \text{(definition of } Y_{t+1}) \\
&= E[X_t + S_{t+1} - (t+1)\mu \,|\, \mathcal{F}_t] && \text{(this is a standard way of splitting up } X_{t+1}\text{, when proving things are martingales)} \\
&= E[X_t \,|\, \mathcal{F}_t] + E[S_{t+1} \,|\, \mathcal{F}_t] - E[(t+1)\mu \,|\, \mathcal{F}_t] && \text{(linearity)} \\
&= X_t + E[S_{t+1} \,|\, \mathcal{F}_t] - (t+1)\mu && \text{(since } X_t \text{ is } \mathcal{F}_t\text{-measurable and } (t+1)\mu \text{ is constant)} \\
&= X_t + E[S_{t+1}] - (t+1)\mu && \text{(since } S_{t+1} \perp \mathcal{F}_t) \\
&= X_t + \mu - t\mu - \mu && \text{(since } E[S_{t+1}] = \mu) \\
&= X_t - t\mu \\
&= Y_t.
\end{aligned}$$

Thus $\{Y_t\} = \{X_t - t\mu\}$ is a martingale.

The proof that $\{(X_t - t\mu)^2 - t\sigma^2\}$ is a HW problem.

Last, let $Y_t = \left(\frac{q}{p}\right)^{X_t}$. We will verify that $E[Y_{t+1} \,|\, \mathcal{F}_t] = Y_t$:

$$\begin{aligned}
E[Y_{t+1} \,|\, \mathcal{F}_t] &= E\left[\left(\frac{q}{p}\right)^{X_{t+1}} \middle|\, \mathcal{F}_t\right] \\
&= E\left[\left(\frac{q}{p}\right)^{X_t + S_{t+1}} \middle|\, \mathcal{F}_t\right] \\
&= E\left[\left(\frac{q}{p}\right)^{X_t} \left(\frac{q}{p}\right)^{S_{t+1}} \middle|\, \mathcal{F}_t\right] \\
&= \left(\frac{q}{p}\right)^{X_t} E\left[\left(\frac{q}{p}\right)^{S_{t+1}} \middle|\, \mathcal{F}_t\right] && \text{(stability)} \\
&= \left(\frac{q}{p}\right)^{X_t} E\left[\left(\frac{q}{p}\right)^{S_{t+1}}\right] && \text{(since } S_{t+1} \perp \mathcal{F}_t).
\end{aligned}$$

To compute the remaining expected value remember that S_{t+1} has density

s	-1	0	1
$f_{S_{t+1}}(s) = P(S_{t+1} = s)$	q	r	p

.

Therefore, by LOTUS,

$$E\left[\left(\frac{q}{p}\right)^{S_{t+1}}\right] = \left(\frac{q}{p}\right)^{1} p + \left(\frac{q}{p}\right)^{0} r + \left(\frac{q}{p}\right)^{-1} q = q + r + p = 1.$$

Substituting into the previous page,

$$E[Y_{t+1} \,|\, \mathcal{F}_t] = \left(\frac{q}{p}\right)^{X_t} (1) = Y_t,$$

verifying that $\{Y_t\} = \left\{\left(\frac{q}{p}\right)^{X_t}\right\}$ is a martingale. □

Escape probabilities

Theorem 10.28 *Let $\{X_t\}$ be an irreducible simple random walk on $\mathbb{Z}$. Let $a < x < b$ be integers. Then:*

- *if $p = q$ (i.e the random walk is unbiased), then*

$$P_x(T_a < T_b) = \frac{b-x}{b-a} \qquad \textit{and} \qquad P_x(T_b < T_a) = \frac{x-a}{b-a}.$$

- *if $p \neq q$ (i.e. the random walk is biased), then*

$$P_x(T_a < T_b) = \frac{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^x}{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^a} \qquad \textit{and} \qquad P_x(T_b < T_a) = \frac{\left(\frac{q}{p}\right)^x - \left(\frac{q}{p}\right)^a}{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^a}$$

PROOF **Case 1:** Suppose the random walk is unbiased. Define $\psi(x) = x$; by the preceding lemma, $\{\psi(X_t)\} = \{X_t\} = \{X_t - t(0)\} = \{X_t - t\mu\}$ is a martingale. By the Escape Time Theorem,

$$P_x(T_a < T_b) = \frac{\psi(b) - \psi(x)}{\psi(b) - \psi(a)} = \frac{b-x}{b-a}$$

$$\text{and } P_x(T_b < T_a) = \frac{\psi(x) - \psi(a)}{\psi(b) - \psi(a)} = \frac{x-a}{b-a}.$$

Case 2: Suppose the random walk is biased. Let $\psi(x) = \left(\frac{q}{p}\right)^x$; by the preceding lemma, $\{\psi(X_t)\} = \left\{\left(\frac{q}{p}\right)^{X_t}\right\}$ is a martingale. Apply the Escape Time Theorem to get

$$P_x(T_a < T_b) = \frac{\psi(b) - \psi(x)}{\psi(b) - \psi(a)} = \frac{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^x}{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^a}$$

$$\text{and } P_x(T_b < T_a) = \frac{\psi(x) - \psi(a)}{\psi(b) - \psi(a)} = \frac{\left(\frac{q}{p}\right)^x - \left(\frac{q}{p}\right)^a}{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^a}. \square$$

EXAMPLE 4

I have \$20 and you have \$15. We each make a series of \$1 bets until one of us goes broke.

1. If we are equally likely to win each bet, what is the probability that you go broke? What amount of money should I expect to end up with?
2. Suppose you are twice as likely as me to win each bet (assume no ties are possible). In this setting, what is the probability you go broke?
3. How long (how many bets), on the average, will it take for one of us to go broke?

Theorem 10.29 (Wald's First Identity) *Let $\{X_t\}$ be an irreducible, simple random walk. Let $a < x < b$ be integers and suppose $X_0 = x$. Let $T = \min\{T_a, T_b\} = T_{\{a,b\}}$. Then*

$$E[X_T] = x + \mu\, ET = x + (p-q)ET.$$

PROOF By Lemma 10.27, we know that $\{X_t - t\mu\}$ is a martingale.

By the OST,

$$x = E[X_0] = E[X_0 - 0\mu] \overset{OST}{=} E[X_T - T\mu] = EX_T - \mu\, ET.$$

$$x = EX_t - \mu\, ET$$

Add $\mu\, ET$ to both sides to get the result. □

Applying Wald's First Identity: If the walk $\{X_t\}$ is biased, we know

$$P(X_T = a) = P_x(T_a < T_b) = \frac{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^x}{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^a}$$

$$P(X_T = b) = P_x(T_b < T_a) = \frac{\left(\frac{q}{p}\right)^x - \left(\frac{q}{p}\right)^a}{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^a}$$

so

$$E[X_T] =$$

By Wald's First Identity, since $EX_T = x + (p-q)ET$,

$$ET = \frac{E[X_T] - x}{p-q} = \frac{\text{above} - x}{p-q}.$$

QUESTION 3 OF EXAMPLE 4, REVISITED

(Recall that I have $\$20$ and you have $\$15$; we each make a series of $\$1$ bets until one of us goes broke.) How long will it take one of us to go broke, if you are twice as likely as I am to win each bet?

Solution: We previously showed that the amount of money I expect to end up with is $E[X_T] = 35\left(\frac{1-2^{20}}{1-2^{35}}\right) \approx .001$. Thus

Follow-up question: What if we are equally likely to win each bet?

Repeating the same logic doesn't work:

So in this setting, we need another fact to answer the question:

Theorem 10.30 (Wald's Second Identity) *Let $\{X_t\}$ be an* ***unbiased*** *simple, irreducible random walk. Let $a < x < b$ be integers and suppose $X_0 = x$. Let $T = \min\{T_a, T_b\} = T_{\{a,b\}}$. Then*

$$Var(X_T) = Var(S_j) \cdot ET = \sigma^2 ET.$$

PROOF By Wald's First Identity, $EX_T = x + \mu ET = x + 0\,ET = x$. So

$$Var(X_T) = E[X_T^2] - (EX_T)^2 = E[X_T^2] - x^2,$$

and we can rearrange this to get

$$x^2 = E[X_T^2] - Var(X_T). \tag{10.3}$$

By Lemma 10.27, we know that $\{Y_t\} = \{(X_t - t\mu)^2 - t\sigma^2\}$ is a martingale. Observe that $EY_0 = E[(X_0 - 0\mu)^2 - 0\sigma^2] = E[X_0^2] = x^2$. So by the OST,

$$x^2 = EY_0 \overset{OST}{=} EY_T = E[X_T^2 - T\sigma^2] \tag{10.4}$$
$$= E[X_T^2] - \sigma^2 ET. \tag{10.5}$$

Equations (10.3) and (10.5) above give two different quantities both equal to x^2, so those quantities are equal, i.e.

$$E[X_T^2] - Var(X_T) = x^2 = E[X_T^2] - \sigma^2 ET.$$

Subtract $E[X_T^2]$ from both sides and multiply through by (-1) to obtain Wald's Second Identity. □

Applying Wald's Second Identity: Suppose $\{X_t\}$ is a simple, unbiased random walk with $r \neq 1$. From the escape probabilities, we know

$$P(X_T = a) = P_x(T_a < T_b) = \frac{b-x}{b-a} \qquad P(X_T = b) = P_x(T_b < T_a) = \frac{x-a}{b-a}$$

so

$$E[X_T] =$$

$$E[X_T^2] =$$

$$Var(X_T) = E[X_T^2] - (E[X_T])^2 =$$

Also,

$$Var(S_j) = E[S_j^2] - E[S_j] = E[S_j^2] =$$

and therefore

$$ET = \frac{Var(X_T)}{Var(S_j)} =$$

Gambler's Ruin

The formulas for $f_{x,y}$ in a simple random walk are known as *Gambler's Ruin*:

Theorem 10.31 (Gambler's Ruin) *Let $\{X_t\}$ be an irreducible, simple random walk on $\mathbb{Z}$. Let x and y be distinct integers ($x \neq y$). Then*

$$f_{x,y} = \begin{cases} 1 & \text{if the walk is unbiased} \\ 1 & \text{if the walk is biased toward walking from } x \text{ toward } y \\ \left(\dfrac{\min\{p,q\}}{\max\{p,q\}}\right)^{|x-y|} & \text{if the walk is biased against walking from } x \text{ toward } y \end{cases}$$

Applying Gambler's Ruin: Suppose $\{X_t\}$ is a simple random walk with $p = .2$ and $q = .7$.

$\{X_t\}$ is negatively biased, i.e. , meaning it is biased towards walking to smaller y from x, and biased against walking to larger y from x. Applying the formulas of Gambler's Ruin,

$f_{-2,5} =$ $f_{8,0} =$

Why is this called "Gambler's Ruin"? Suppose a gambler brings $\$50$ to a casino and makes a series of $\$1$ bets in a game where he has a 50% chance of winning each bet, and a 50% chance of losing each bet. The Gambler's Ruin Theorem says

PROOF We prove Gambler's Ruin with several cases:

If the walk is unbiased and $x < y$:

From the Escape Time Corollary

$$f_{x,y} = \lim_{a \to -\infty} P_x(T_y < T_a)$$

which by the escape probability formulas for unbiased walks is

$$\lim_{a \to -\infty} \frac{x-a}{y-a} \overset{L}{=} \lim_{a \to -\infty} \frac{-1}{-1} = \boxed{1}.$$

If the walk is unbiased and $x > y$:

If $x > y$, then from the Escape Time Corollary

$$f_{x,y} = \lim_{b \to \infty} P_x(T_y < T_b)$$

which by the escape probability formulas for unbiased walks is

$$\lim_{b \to \infty} \frac{b-x}{b-a} \overset{L}{=} \lim_{b \to \infty} \frac{1}{1} = \boxed{1}.$$

If the walk is positively biased ($p > q$) and $x < y$:

From the Escape Time Corollary

$$f_{x,y} = \lim_{a \to -\infty} P_x(T_y < T_a),$$

which by the escape probability formulas for biased walks is

$$\lim_{a \to -\infty} \frac{\left(\frac{q}{p}\right)^x - \left(\frac{q}{p}\right)^a}{\left(\frac{q}{p}\right)^y - \left(\frac{q}{p}\right)^a} = \frac{\left(\frac{q}{p}\right)^x - \infty}{\left(\frac{q}{p}\right)^y - \infty} = \frac{\infty}{\infty}$$

$$\stackrel{L}{=} \lim_{a \to -\infty} \frac{0 - \left(\frac{q}{p}\right)^a \ln\left(\frac{q}{p}\right)}{0 - \left(\frac{q}{p}\right)^a \ln\left(\frac{q}{p}\right)} = \boxed{1}.$$

If the walk is positively biased ($p > q$) and $x > y$:

From the Escape Time Corollary

$$f_{x,y} = \lim_{b \to \infty} P_x(T_y < T_b),$$

which by the escape probability formulas for biased walks is

$$\lim_{b \to \infty} \frac{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^x}{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^y} = \frac{0 - \left(\frac{q}{p}\right)^x}{0 - \left(\frac{q}{p}\right)^y}$$

$$= \left(\frac{q}{p}\right)^{x-y} = \boxed{\left(\frac{\min\{p,q\}}{\max\{p,q\}}\right)^{|x-y|}}.$$

The situation where the walk is negatively biased is a HW problem. □

Recurrence/transience

Theorem 10.32 *Let $\{X_t\}$ be an irreducible, simple random walk on $\mathbb{Z}$. Then $\{X_t\}$ is recurrent if and only if the random walk is unbiased.*

PROOF Since $\{X_t\}$ is irreducible,

$$\{X_t\} \text{ is recurrent} \iff 0 \in \mathcal{S}_R \iff f_0 = 1.$$

If the walk is unbiased:

From the Escape Time Corollary,

$$\begin{aligned} f_0 = f_{0,0} &= P(0,0) + \sum_{y \neq 0} f_{y,0}\, P(0,y) \\ &= P(0,0) + f_{-1,0} P(0,-1) + f_{1,0} P(0,1) \\ &= r + f_{-1,0}\, q + f_{1,0}\, p \\ &= r + (1)q + (1)p \qquad \text{(by Gambler's Ruin)} \\ &= r + q + p = \boxed{1}. \end{aligned}$$

Therefore unbiased simple walks are recurrent.

If the walk is positively biased ($p > q$):

Again, from the Escape Time Corollary,

$$\begin{aligned} f_0 = f_{0,0} &= P(0,0) + \sum_{y \neq 0} f_{y,0}\, P(0,y) \\ &= P(0,0) + f_{-1,0}P(0,-1) + f_{1,0}P(0,1) \\ &= r + f_{-1,0}\, q + f_{1,0}\, p \\ &= r + (1)q + \left(\frac{\min\{p,q\}}{\max\{p,q\}}\right)^{|1-0|} p \qquad \text{(by Gambler's Ruin)} \\ &= r + q + \left(\tfrac{q}{p}\right) p \qquad \text{(since } p > q\text{)} \\ &= r + q + q \\ &< r + q + p \qquad \text{(since } p > q\text{)} \\ &= 1. \end{aligned}$$

Therefore $f_0 < 1$, so positively biased simple walks are transient.

The situation where the walk is negatively biased is similar and left as HW. □

Remark: In the process of proving this theorem we have proven that for any simple random walk,

$$f_0 = r + 2\min(p,q).$$

By symmetry, it must be that this formula also equals f_x for any $x \in \mathbb{Z}$.

Random walk in higher dimensions

In this section we discuss simple, unbiased random walks in $\mathbb{Z}^d$. This means that we assume $\{X_t\}$ is a Markov chain taking values in $\mathbb{Z}^d$ with

- $X_0 = (0,0,...,0) = \mathbf{0}$;
- $P(\mathbf{x},\mathbf{y}) = \begin{cases} \frac{1}{2d} & \text{if } \mathbf{x} - \mathbf{y} = \pm\mathbf{e}_j \text{ for some } j \\ 0 & \text{else} \end{cases}$.

Notation: The vector $\mathbf{e}_j \in \mathbb{R}^d$ is the vector $(0,0,0,...,0,1,0,...,0)$ which has a 1 in the j^{th} place and zeros everywhere else. (Thus $-\mathbf{e}_j$ is $(0,0,...,0,-1,0,...,0)$.)

So in these higher-dimensional walks, you start at the origin and at each step, you move one unit in one of the coordinate directions, choosing the direction you move in uniformly.

These random walks are all irreducible and have period 2.

EXAMPLE: SIMPLE RANDOM WALK ON $\mathbb{Z}$

When $d = 1$, this is a description of simple, unbiased random walk on $\mathbb{Z}$ with $p = q = \frac{1}{2}$. This Markov chain is

EXAMPLE: DRUNKARD'S WALK (RANDOM WALK ON $\mathbb{Z}^2$)

EXAMPLE: DRUNK BIRD'S FLIGHT (RANDOM WALK ON $\mathbb{Z}^3$)

Main question: Will the drunk person ever make it home? Will they make it back to the bar? What about the inebriated bird? In other words, **is unbiased random walk on $\mathbb{Z}^d$ recurrent or transient**?

In a group activity, you have proven (or will prove):

Theorem 10.33 (Pólya's Theorem) *Let $\{X_t\}$ be simple, unbiased random walk on $\mathbb{Z}^d$ as described in this section. Then:*

1. *If $d = 1$ or 2, then $\{X_t\}$ is null recurrent.*
2. *If $d > 2$, then $\{X_t\}$ is transient.*

Mathematician Shizuo Kakutani famously described this theorem by saying "A drunk man will find his way home, but a drunk bird may be lost forever."

Summary of simple random walks on $\mathbb{Z}$

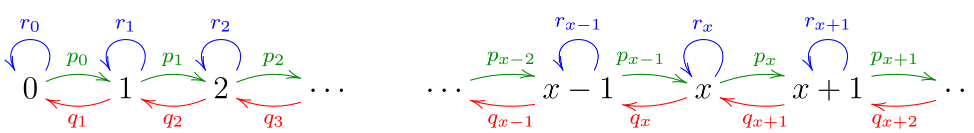
process $\{X_t\}$	UNBIASED SIMPLE RANDOM WALK	BIASED SIMPLE RANDOM WALK
State space	$\mathbb{Z}$	$\mathbb{Z}$
Process determined by	p and q, with $p = q$ ($r = 1 - p - q$)	p and q, with $p \neq q$ ($r = 1 - p - q$)
Other quantities	$\mu = 0$ $\sigma^2 = p + q$	$\mu = p - q$ $\sigma^2 = p + q - (p - q)^2$
Associated martingales	$\{X_t\}$ $\{X_t^2 - t\sigma^2\}$	$\left\{\left(\frac{q}{p}\right)^{X_t}\right\}$ $\{(X_t - \mu t)^2 - t\sigma^2\}$
$P_x(T_a < T_b)$	$\dfrac{b-x}{b-a}$	$\dfrac{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^x}{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^a}$
$P_x(T_b < T_a)$	$\dfrac{x-a}{b-a}$	$\dfrac{\left(\frac{q}{p}\right)^x - \left(\frac{q}{p}\right)^a}{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^a}$
$E_x[T_{\{a,b\}}]$	$\dfrac{(b-x)(x-a)}{p+q}$ (comes from **Wald's** 2^{nd} **Id**: $Var_x(X_T) = \sigma^2\, ET$)	Solve for this using **Wald's** 1^{st} **Id**: $EX_T = x + \mu\, ET$
$f_{x,y}$ (for $x \neq y$)	**Gambler's Ruin:** $f_{x,y} = 1$	**Gambler's Ruin:** $f_{x,y} = 1$ if walk tends toward y; $f_{x,y} = \left(\dfrac{\min\{p,q\}}{\max\{p,q\}}\right)^{\lvert x-y \rvert}$ if walk tends away from y
f_x	1	$r + 2\min(p,q)$
Recurrence/ transience	null recurrent	transient

10.7 Birth and death chains

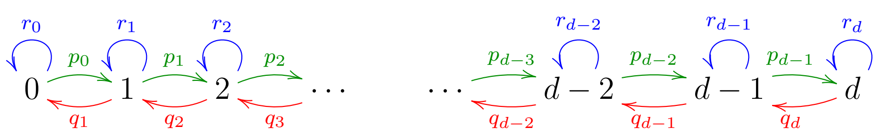
Definition 10.34 *A Markov chain is called a* **birth and death chain** *if its state space is* $\mathcal{S} = \{0, 1, ..., d\}$ *or* $\mathcal{S} = \{0, 1, 2, ...\}$ *and for every* $x \in \mathcal{S}$*, there are nonnegative numbers* p_x, q_x *and* r_x *so that*

1. *for all* $x \in \mathcal{S}$*,* $p_x + q_x + r_x = 1$*;*
2. $q_0 = 0$*;*
3. *if* $\mathcal{S} = \{0, 1, ..., d\}$*, then* $p_d = 0$*; and*
4. *for all* $x \in \mathcal{S}$*,* $\begin{cases} P(x, x+1) = p_x \\ P(x, x) = r_x \\ P(x, x-1) = q_x \end{cases}$.

DIRECTED GRAPHS OF BIRTH AND DEATH CHAINS



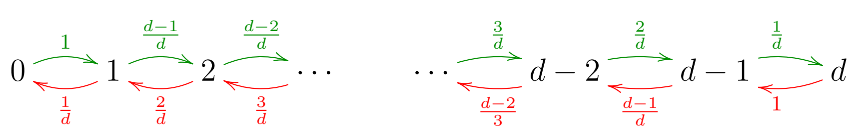
if $\mathcal{S} = \{0, 1, 2, ...\}$, or



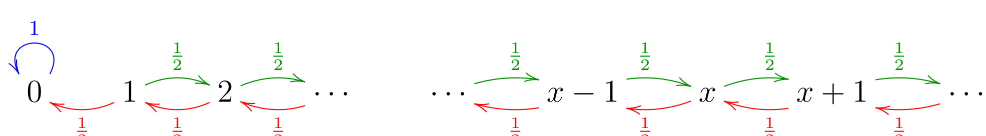
if $\mathcal{S} = \{0, 1, ..., d\}$.

EXAMPLES OF BIRTH AND DEATH CHAINS

- The Ehrenfest chain



- The **gambler's ruin** chain



Properties of birth and death chains

Class structure

A birth and death chain is irreducible if and only if no p_x nor q_x is 0 (other than q_0 or p_d).

If a birth and death chain is not irreducible, then the communicating classes of the chain are themselves birth and death chains (after perhaps relabeling the state space).

If all the $r_x = 0$, the birth and death chain has period 2; otherwise the chain is aperiodic.

Escaping property

Lemma 10.35 *Irreducible birth and death chains are escaping.*

PROOF This proof is essentially the same as the proof that random walks are escaping.

Let $p = \min\{p_a, p_{a+1}, ..., p_b\}$; since $\{X_t\}$ is irreducible, $p > 0$.

Now let G_n be the event that between times $(n-1)(b-a)$ and $n(b-a)$, there are only births in the birth-death chain.

Note that $P(G_n) \geq p^{b-a} > 0$, so by repeating the rest of the proof given for random walks, we see $P(T_A = \infty) \leq P(\text{no } G_n \text{ occurs}) = 0$. □

Steps

Definition 10.36 *Let $\{X_t\}$ be a birth and death chain. For each $t \geq 1$, define the t^{th}* **step** *of the chain to be*

$$S_t = X_t - X_{t-1}.$$

PROPERTIES

- $X_t = X_s + \sum_{j=s+1}^{t} S_t$;
- the S_t are $\perp$ (but in general, not identically distributed) with cond'l density

s	-1	0	1
$P(S_t = s \,\vert\, X_{t-1} = x)$	q_x	r_x	p_x

;

- $S_t \perp \mathcal{F}_s$ for all $s < t$.

Associated martingale

Lemma 10.37 *Let $\{X_t\}$ be an irreducible birth and death chain. Let $\gamma_0 = 1$, and for each $y > 0$, let*

$$\gamma_y = \frac{q_y q_{y-1} q_{y-2} \cdots q_2 q_1}{p_y p_{y-1} p_{y-1} \cdots p_2 p_1}.$$

Define the function $\psi : \mathcal{S} \to \mathbb{R}$ by setting $\psi(0) = 1$ and for $y \geq 1$, setting

$$\begin{aligned} \psi(y) &= 1 + \frac{q_1}{p_1} + \frac{q_2 q_1}{p_2 p_1} + \ldots + \frac{q_{y-1} q_{y-2} \cdots q_2 q_1}{p_{y-1} p_{y-2} \cdots p_2 p_1} \\ &= \gamma_0 + \gamma_1 + \gamma_2 + \ldots + \gamma_{y-1} \\ &= \sum_{j=0}^{y-1} \gamma_j. \end{aligned}$$

Then the stochastic process $\{\psi(X_t)\}$ is a martingale.

PROOF Since $\{X_t\}$ is discrete-time, so is $\{\psi(X_t)\}$, so it is sufficient to verify

$$E[\psi(X_t) \,|\, \mathcal{F}_{t-1}] = \psi(X_{t-1}).$$

We'll show this with a slightly unusual type of computation:

$$\begin{aligned} E[\psi(X_t) \,|\, \mathcal{F}_{t-1}] &= E\left[\psi(X_t) - \psi(X_{t-1}) + \psi(X_{t-1}) \,|\, \mathcal{F}_{t-1}\right] \\ &= E\left[\psi(X_t) - \psi(X_{t-1}) \,|\, \mathcal{F}_{t-1}\right] + E\left[\psi(X_{t-1}) \,|\, \mathcal{F}_{t-1}\right] \\ &= E\left[\psi(X_t) - \psi(X_{t-1}) \,|\, \mathcal{F}_{t-1}\right] + \psi(X_{t-1}) \qquad \text{(stability)} \end{aligned}$$

What's left to show is that the blue term above is zero. To do this, let $x = X_{t-1}$ (this value is information included in $\mathcal{F}_{t-1}$). Then:

- with probability p_x, the t^{th} step is a birth, so $X_t = x + 1$, so

$$\psi(X_t) - \psi(X_{t-1}) = \psi(x+1) - \psi(x) = \sum_{j=0}^{(x+1)-1} \gamma_j - \sum_{j=0}^{x-1} \gamma_j = \gamma_x.$$

- with probability q_x, the t^{th} step is a death, so $X_t = x - 1$, so

$$\psi(X_t) - \psi(X_{t-1}) = \psi(x-1) - \psi(x) = \sum_{j=0}^{(x-1)-1} \gamma_j - \sum_{j=0}^{x-1} \gamma_j = -\gamma_{x-1}.$$

- with probability r_x, the t^{th} step is a loop, so $X_t = x$, so

$$\psi(X_t) - \psi(X_{t-1}) = \psi(x) - \psi(x) = 0.$$

Therefore, by LOTUS, the blue term above is

$$\begin{aligned} E\left[\psi(X_t) - \psi(X_{t-1}) \mid \mathcal{F}_{t-1}\right] &= p_x\gamma_x - q_x\gamma_{x-1} + r_x(0) \\ &= p_x\left(\frac{q_xq_{x-1}\cdots q_1}{p_xp_{x-1}\cdots p_1}\right) - q_x\left(\frac{q_{x-1}\cdots q_1}{p_{x-1}\cdots p_1}\right) \\ &= \frac{q_xq_{x-1}\cdots q_1}{p_{x-1}\cdots p_1} - \frac{q_xq_{x-1}\cdots q_1}{p_{x-1}\cdots p_1} = 0. \end{aligned}$$

This finishes the proof. □

Escape probabilities

Theorem 10.38 *Let $\{X_t\}$ be an irreducible birth and death chain. For any $a < x < b$,*

$$P_x(T_a < T_b) = \frac{\sum_{y=x}^{b-1} \gamma_y}{\sum_{y=a}^{b-1} \gamma_y} \quad \textit{and} \quad P_x(T_b < T_a) = \frac{\sum_{y=a}^{x-1} \gamma_y}{\sum_{y=a}^{b-1} \gamma_y}$$

where the γ_y are as defined in Lemma 10.37:

$$\gamma_0 = 1; \qquad \gamma_y = \frac{q_yq_{y-1}\cdots q_1}{p_yp_{y-1}\cdots p_1} \textit{ for } y > 0.$$

PROOF By the escape time theorem (Theorem 10.20), since $\{\psi(X_t)\}$ is a martingale,

$$P_x(T_a < T_b) = \frac{\psi(b) - \psi(x)}{\psi(b) - \psi(a)} = \frac{\sum_{y=0}^{b-1} \gamma_y - \sum_{y=0}^{x-1} \gamma_y}{\sum_{y=0}^{b-1} \gamma_y - \sum_{y=0}^{a-1} \gamma_y} = \frac{\sum_{y=x}^{b-1} \gamma_y}{\sum_{y=a}^{b-1} \gamma_y}$$

(and the other formula comes from the complement rule). □

Recurrence / transience

Lemma 10.39 *Let $\{X_t\}$ be an irreducible birth and death chain with infinite state space. Then $\{X_t\}$ is recurrent if and only if $f_{1,0} = 1$.*

PROOF $\{X_t\}$ is irreducible, so $\{X_t\}$ is recurrent $\Leftrightarrow$ 0 is recurrent $\Leftrightarrow$ $f_{0,0} = 1$. Now

$$\begin{aligned} f_{0,0} &= P_0(T_0 < \infty) \\ &= P_0(T_0 = 1) + P_0(T_0 \in [2, \infty)) \\ \\ &= \end{aligned}$$

Theorem 10.40 *Let $\{X_t\}$ be an irreducible birth and death chain with $\mathcal{S} = \{0, 1, 2, ...\}$. Then defining γ_y as in the previous theorem,*

$$\{X_t\} \text{ is recurrent } \iff \sum_{y=0}^{\infty} \gamma_y = \infty.$$

PROOF By the preceding lemma, $\{X_t\}$ is recurrent if and only if $f_{1,0} = 1$:

$$\begin{aligned}
f_{1,0} &= \lim_{b\to\infty} P_1(T_0 < T_b) && \text{(by Corollary 10.21)} \\
&= \lim_{b\to\infty} \left(\frac{\sum_{y=1}^{b-1} \gamma_y}{\sum_{y=0}^{b-1} \gamma_y} \right) && \text{(by Theorem 10.38 with } x = 1, a = 0, b = b) \\
&= \lim_{b\to\infty} \left(\frac{\sum_{y=0}^{b-1} \gamma_y - \gamma_0}{\sum_{y=0}^{b-1} \gamma_y} \right) \\
&= \lim_{b\to\infty} \left(\frac{\sum_{y=0}^{b-1} \gamma_y - 1}{\sum_{y=0}^{b-1} \gamma_y} \right) && (\gamma_0 = 1 \text{ by definition}) \\
&= \lim_{b\to\infty} \left(1 - \frac{1}{\sum_{y=0}^{b-1} \gamma_y} \right) \\
&= 1 - \frac{1}{\sum_{y=0}^{\infty} \gamma_y} \\
&= \begin{cases} & \text{if } \sum_{y=0}^{\infty} \gamma_y \text{ diverges} \\ & \text{if } \sum_{y=0}^{\infty} \gamma_y \text{ converges to } C \end{cases}
\end{aligned}$$

This proves the theorem. □

EXAMPLE 5

Let $\{X_t\}$ be a birth and death chain on $\mathcal{S} = \{0, 1, 2, 3, ...\}$ such that

$$p_x = \frac{x+2}{2(x+1)} \quad \text{and} \quad q_x = \frac{x}{2(x+1)}.$$

Is this chain recurrent or transient?

Positive recurrence / stationary distribution

Theorem 10.41 *Let $\{X_t\}$ be an irreducible birth and death chain. Define, for each $y \in \mathcal{S}$,*

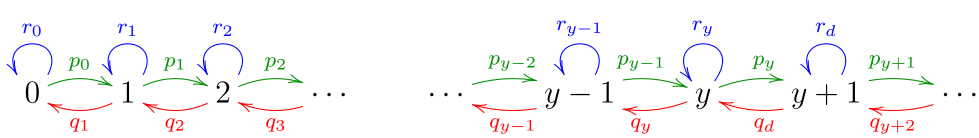
$$\zeta_0 = 1; \qquad \zeta_y = \frac{p_0 p_1 \cdots p_{y-1}}{q_1 q_2 \cdots q_y} \text{ for } y > 0.$$

Then:

1. *If $\sum_{y \in \mathcal{S}} \zeta_y$ converges, then $\{X_t\}$ is positive recurrent and has one stationary distribution π defined by*
$$\pi(x) = \frac{\zeta_x}{\sum_{y \in \mathcal{S}} \zeta_y}.$$
(This includes all situations where $\mathcal{S}$ is finite.)

2. *If $\sum_{y \in \mathcal{S}} \zeta_y$ diverges, then $\{X_t\}$ has no stationary distributions (so it is either null recurrent or transient).*

Think of this mysterious ζ_y as "the product of all the ps to the left of y over the product of all the qs to the left of y in the directed graph:



PROOF Suppose $\{X_t\}$ has a stat. dist. π. Such a π must satisfy the stationarity equation, which can be rearranged as follows:

$$\begin{aligned}
\pi(y) &= \sum_{x \in \mathcal{S}} \pi(x) \boxed{P(x,y)} \leftarrow = 0 \text{ unless } x \in \{y-1, y, y+1\} \\
&= \pi(y-1)P(y-1,y) + \pi(y)P(y,y) + \pi(y+1)P(y+1,y) \\
&= \pi(y-1)p_{y-1} + \pi(y)r_y + \pi(y+1)q_{y+1} \\
&= \pi(y-1)p_{y-1} + \pi(y)[1 - p_y - q_y] + \pi(y+1)q_{y+1} \\
0 &= \pi(y-1)p_{y-1} - \pi(y)p_y - \pi(y)q_y + \pi(y+1)q_{y+1}.
\end{aligned}$$

Moving some terms over, this becomes

$$\pi(y)q_y - \pi(y+1)q_{y+1} = \pi(y-1)p_{y-1} - \pi(y)p_y. \tag{10.6}$$

Let $y = 0$. Equation (10.6) reduces to

$$\begin{aligned} \pi(0)q_0 - \pi(1)q_1 &= \pi(-1)p_{-1} - \pi(0)p_0 \\ -\pi(1)q_1 &= -\pi(0)p_0 \\ \pi(1) &= \frac{p_0}{q_1}\pi(0) \qquad (10.7) \\ \pi(1) &= \zeta_1\,\pi(0). \end{aligned}$$

Similarly, if you plug in $y = 1$ into (10.6) and use (10.7), you will get

$$\pi(2) = \frac{p_1}{q_2}\pi(1) = \frac{p_1 p_0}{q_2 q_1}\pi(0) = \zeta_2\,\pi(0),$$

and by induction you can prove

$$\pi(y) = \zeta_y\,\pi(0).$$

This shows: if π is stationary, then $\pi(y) = \zeta_y\,\pi(0)$ for all $y \in \mathcal{S}$.

Case 1: $\sum\limits_{y\in\mathcal{S}} \zeta_y$ diverges. That would force $\sum\limits_{y\in\mathcal{S}} \pi(y)$ to diverge, so there is no stat. dist., so $\{X_t\}$ is **not** positive recurrent.

Case 2: $\sum\limits_{y\in\mathcal{S}} \zeta_y$ converges. Then $\sum\limits_{y\in\mathcal{S}} \pi(y) = \pi(0)\sum\limits_{y\in\mathcal{S}} \zeta_y$, so for this sum to equal 1 we need $\pi(0) = \dfrac{1}{\sum\limits_{y\in\mathcal{S}} \zeta_y}$. Therefore we get a distribution π defined by

$$\pi(y) = \zeta_y\,\pi(0) = \frac{\zeta_y}{\sum\limits_{y\in\mathcal{S}} \zeta_y}$$

which is stationary, making the chain positive recurrent. $\square$

EXAMPLE 6

Let $\{X_t\}$ be a birth-death chain on $\{0, 1, 2, 3, ...\}$ with $p_0 = 1$; $p_x = \dfrac{1}{x+1}$ for all $x \geq 1$; $q_x = \dfrac{x}{x+1}$ for all $x \geq 1$. Find the stationary distribution of $\{X_t\}$, if one exists.

Solution: First, compute the ζ_y: $\zeta_0 = 1$ and for $y \geq 1$,

$$\zeta_y = \frac{p_0 p_1 \cdots p_{y-1}}{q_1 q_2 \cdots q_y} =$$

Then apply Theorem 10.41:

$$\sum_{y \in \mathcal{S}} \zeta_y = \sum_{y=0}^{\infty} \zeta_y =$$

So the stationary distribution π satisfies

$$\pi(x) = \frac{\zeta_x}{\sum_y \zeta_y} = \frac{\zeta_x}{2e} = \frac{x+1}{2ex!}.$$

10.8 Birth and death CTMCs

A *birth and death CTMC* is a CTMC whose jump chain is a birth and death chain. Equivalently:

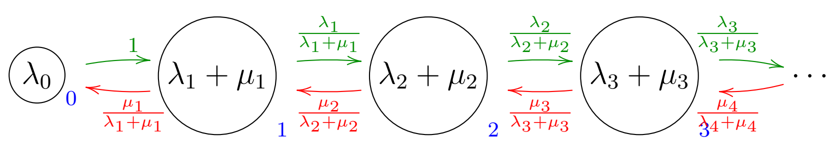
Definition 10.42 *A* **birth and death CTMC** *(or* **birth-death CTMC***) is a CTMC $\{X_t\}$ whose state space is either $\mathcal{S} = \{0, 1, ..., d\}$ or $\mathcal{S} = \{0, 1, 2, ...\}$, such that $q_{x,y} = 0$ whenever $|x - y| > 1$.*

The numbers $\lambda_x = q_{x,x+1}$ are called the **birth rates** *of the process and the numbers $\mu_x = q_{x,x-1}$ are called the* **death rates***.*

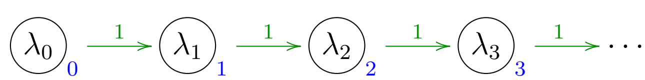
A birth-death CTMC is called a **pure birth process** *if $\mu_x = 0$ for all x, and is called a* **pure death process** *if $\lambda_x = 0$ for all x.*

In a birth and death CTMC, we are given $\begin{cases} \text{birth rates} & \lambda_x = q_{x,x+1} \\ \text{death rates} & \mu_x = q_{x,x-1} \end{cases}$

So the directed graph of a birth-death CTMC looks like



and if the CTMC was a pure birth process, the directed graph would be



$$\begin{array}{c}\text{We know that}\\ \text{a CTMC is}\end{array} \left\{\begin{array}{c}\text{irreducible}\\ \text{recurrent}\\ \text{transient}\\ \text{pos. recurrent}\\ \text{null recurrent}\end{array}\right\} \iff \text{its jump chain is} \left\{\begin{array}{c}\text{irreducible}\\ \text{recurrent}\\ \text{transient}\\ \text{pos. recurrent}\\ \text{null recurrent}\end{array}\right\}.$$

For a birth and death CTMC, the jump chain is a (discrete-time) birth and death chain with

$$p_x = \frac{\lambda_x}{\lambda_x + \mu_x} = \frac{\lambda_x}{q_x} \quad \text{and} \quad q_x = \frac{\mu_x}{\lambda_x + \mu_x} = \frac{\mu_x}{q_x}.$$

We learned in Section 10.7 that the jump chain (and hence the birth and death CTMC) is transient if and only if $\sum_{y=0}^{\infty} \gamma_y < \infty$ where

$$\gamma_0 = 1 \qquad \text{and} \qquad \gamma_y = \frac{q_y \cdots q_1}{p_y \cdots p_1} \text{ for } y > 0.$$

This happens if and only if

$$\sum_{y=1}^{\infty} \frac{q_y \cdots q_1}{p_y \cdots p_y} < \infty$$

$$\text{i.e.} \qquad \sum_{y=1}^{\infty} \frac{\frac{\mu_y}{q_y} \cdot \frac{\mu_{y-1}}{q_{y-1}} \cdots \frac{\mu_1}{q_1}}{\frac{\lambda_y}{q_y} \frac{\lambda_{y-1}}{q_{y-1}} \cdots \frac{\lambda_1}{q_1}} < \infty$$

$$\text{i.e.} \qquad \sum_{y=1}^{\infty} \frac{\mu_y \mu_{y-1} \cdots \mu_1}{\lambda_y \lambda_{y-1} \cdots \lambda_1} < \infty.$$

We have proven:

Theorem 10.43 *An irreducible birth-death CTMC with state space $\mathcal{S} = \{0, 1, ..., \}$ is transient if and only if*

$$\sum_{y \in \mathcal{S}} \gamma_y < \infty, \text{ where } \gamma_0 = 1 \text{ and } \gamma_y = \frac{\mu_y \cdots \mu_q}{\lambda_y \cdots \lambda_1} \text{ for } y > 0.$$

This is the same result as we had for discrete-time birth and death chains, with μs instead of qs and λs instead of ps.

Similarly, one can classify birth-death CTMCs as positive recurrent or not, and compute their stationary distribution, using the following machinery:

Definition 10.44 *Let $\{X_t\}$ be an irreducible birth-death CTMC. Define*

$$\zeta_0 = 1 \text{ and } \zeta_y = \frac{\lambda_0 \cdots \lambda_{y-1}}{\mu_1 \cdots \mu_y}$$

for every $y > 0$ in $\mathcal{S}$.

Think of ζ_y as "the product of all the λs to the left of y over the product of all the μs to the left of y" (so this is the same idea as the ζ_y we cooked up for discrete-time birth and death chains).

Theorem 10.45 *An irreducible birth and death CTMC on $\mathcal{S} = \{0, 1, ..., \}$ is positive recurrent if and only if*

$$\sum_{y \in \mathcal{S}} \zeta_y < \infty.$$

In that situation, the stationary distribution of the birth and death CTMC is given by

$$\pi(x) = \frac{\zeta_x}{\sum_{y \in \mathcal{S}} \zeta_y}.$$

EXAMPLE 7

Show that the birth-death CTMC $\{X_t\}$ with $\lambda_x = 1$ and $\mu_x = 2$ for all x is positive recurrent, and compute its stationary distribution.

Solution: Compute ζ_y for each y. $\zeta_0 = 1$ and for $y \geq 1$, we have

$$\zeta_y = \frac{\lambda_0 \cdots \lambda_{y-1}}{\mu_1 \cdots \mu_y} = \frac{1(1) \cdots 1}{2(2) \cdots 2} = \frac{1}{2^y}.$$

Since

$$\sum_{y=0}^{\infty} \zeta_y = \sum_{y=0}^{\infty} \frac{1}{2^y} = 2 < \infty,$$

$\{X_t\}$ is positive recurrent. The stationary distribution is therefore given by

$$\pi(x) = \frac{\zeta_x}{\sum_y \zeta_y} = \frac{\zeta_x}{2} = \frac{\frac{1}{2^x}}{2} = \boxed{\frac{1}{2^{x+1}}}.$$

(In other words, $\boxed{\pi = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, ...\right)}$.)

Branching processes

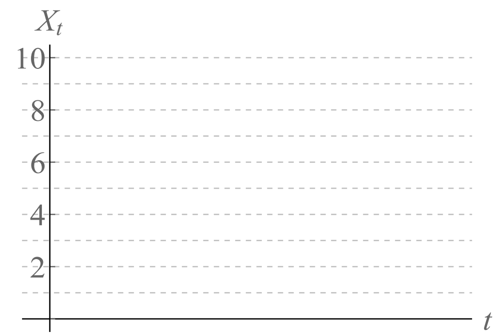
Suppose that at time $t = 0$ you have a population of $X_0 = x$ beings, where X_0 is a r.v. taking values in $\{0, 1, 2, ...\}$.

Each being does nothing for time A ($A : \Omega \to [0, \infty)$ is an exponential r.v. with parameter λ) and then either splits into two beings (with probability $p \in (0, 1)$) or dies (with probability $1 - p$). Each being behaves independently of other beings.

For $t \in [0, \infty)$, let X_t be the number of particles at time t. $\{X_t\}$ is a CTMC called a **(Markov) branching process**.

“population picture”

process $\{X_t\}$



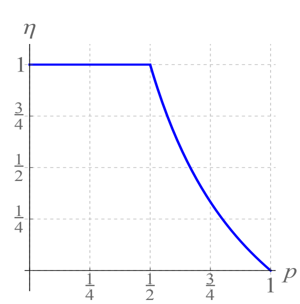
PROPERTIES OF BRANCHING PROCESSES

In a Markov branching process,

1. 0 is **absorbing** (this means $P_{0,0}(t) = 1$ for every $t \geq 0$);
2. every nonzero state in $\mathcal{S}$ is transient (because that state leads to 0 with positive probability, but 0 doesn't lead back); and
3. the directed graph looks like

Theorem 10.46 *Let $\{X_t\}$ be a branching process. Then the extinction probability $\eta = f_{1,0}$ satisfies*

$$\eta = \begin{cases} \dfrac{1-p}{p} & \textit{if } p > \frac{1}{2} \\ 1 & \textit{if } p < \frac{1}{2} \end{cases}.$$



PROOF First, note $\eta = f_{1,0}$ in the branching process is the same as $\eta = f_{1,0}$ in the associated jump chain. Now use the formulas derived in the proof of Theorem 10.40. First, $\gamma_0 = 1$ and if $y > 0$,

$$\gamma_y = \frac{q_1 q_2 \cdots q_y}{p_1 \cdots p_y} = \frac{(1-p)(1-p)\cdots(1-p)}{p\,p \cdots p} = \left(\frac{1-p}{p}\right)^y,$$

so

$$\begin{aligned}
f_{1,0} &= 1 - \frac{1}{\sum_{y=0}^{\infty} \gamma_y} \qquad \text{(from the proof of Thm 10.40)} \\
&= 1 - \left[\sum_{y=0}^{\infty} \left(\frac{1-p}{p}\right)^y\right]^{-1} \\
&= \begin{cases} 1 - \left[\dfrac{1}{1-\left(\frac{1-p}{p}\right)}\right]^{-1} & \text{if } \dfrac{1-p}{p} < 1 \\ 1 - [\infty]^{-1} & \text{else} \end{cases} \\
&= \begin{cases} 1 - \left[1 - \dfrac{1-p}{p}\right] & \text{if } 1-p < p \\ 1 & \text{else} \end{cases} \\
&= \begin{cases} \dfrac{1-p}{p} & \text{if } p > \dfrac{1}{2} \\ 1 & \text{else} \end{cases} \qquad \square
\end{aligned}$$

Note: As with a Galton-Watson branching chain, $f_{x,0} = \eta^x$ for all $x \in \{0, 1, 2, ...\}$.

EXAMPLE 8

Suppose $\{X_t\}$ is a continuous-time Markov branching process with $p = \frac{7}{11}$. If $X_0 = 6$, what is the probability $X_t \neq 0$ for all t?

Summary of birth and death processes

	DISCRETE-TIME BIRTH AND DEATH CHAIN	BIRTH AND DEATH CTMC
State space	$\{0, 1, ..., d\}$ or $\{0, 1, 2, ...\}$	$\{0, 1, ..., d\}$ or $\{0, 1, 2, ...\}$
Process determined by	p_x and q_x for each $x \in \mathcal{S}$ $(r_x = 1 - p_x - q_x)$	λ_x and μ_x for each $x \in \mathcal{S}$ $(q_x = \lambda_x + \mu_x)$
Other quantities	$\gamma_j = \dfrac{q_j q_{j-1} \cdots q_1}{p_j p_{j-1} \cdots p_1}$ $\zeta_j = \dfrac{p_{j-1} p_{j-2} \cdots p_0}{q_j q_{j-1} \cdots q_1}$ $(\gamma_0 = \zeta_0 = 1)$	$\gamma_j = \dfrac{\mu_j \mu_{j-1} \cdots \mu_1}{\lambda_j \lambda_{j-1} \cdots \lambda_1}$ $\zeta_j = \dfrac{\lambda_{j-1} \lambda_{j-2} \cdots \lambda_0}{\mu_j \mu_{j-1} \cdots \mu_1}$ $(\gamma_0 = \zeta_0 = 1)$
Associated martingale	$\{\psi(X_t)\}$, where $\psi(0) = 1$ and $\psi(y) = \sum_{j=0}^{y-1} \gamma_j$	N/A
$P_x(T_a < T_b)$	$\dfrac{\psi(b) - \psi(x)}{\psi(b) - \psi(a)}$	N/A
$P_x(T_b < T_a)$	$\dfrac{\psi(x) - \psi(a)}{\psi(b) - \psi(a)}$	N/A
Recurrence/ transience test	transient if $\sum_y \gamma_y < \infty$; recurrent if $\sum_y \gamma_y = \infty$	transient if $\sum_y \gamma_y < \infty$; recurrent if $\sum_y \gamma_y = \infty$
Positive recurrence test	pos. recurrent if and only if $\sum_y \zeta_y < \infty$	pos. recurrent if and only if $\sum_y \zeta_y < \infty$
stationary distribution (if positive recurrent)	$\pi(x) = \dfrac{\zeta_x}{\sum_y \zeta_y}$	$\pi(x) = \dfrac{\zeta_x}{\sum_y \zeta_y}$

10.9 Chapter 10 Homework

Exercises from Section 10.2

1. In each part of this question, you are given a set Ω and a collection $\mathcal{F}$ of subsets of Ω. Determine, with brief justification, whether or not $\mathcal{F}$ is a σ-algebra on Ω:

   a) $\Omega = \{0, 1, 2, 3, ...\}$; $\mathcal{F} = \{\emptyset, \{0, 2, 4, 6, 8, ...\}, \{1, 3, 5, 7, 9, ...\}, \Omega\}$.

   b) $\Omega = \mathbb{Z}$; $\mathcal{F}$ is the collection of all bounded subsets of Ω (a set is bounded if it is a subset of the interval $[-N, N]$ for some N).

   c) $\Omega = [0, 1]^3$; $\mathcal{F}$ is the collection of sets of the form $E \times [0, 1]$, where $E \subseteq [0, 1]^2$.

   d) $\Omega = [0, 1]^3$; $\mathcal{F}$ is the collection of sets of the form $A \times B \times [0, 1]$, where A and B are subsets of $[0, 1]$.

2. In each part of this problem, you are given a set Ω, a σ-algebra $\mathcal{F}$, and a r.v. $X : \Omega \to \mathbb{R}$. Determine if the given r.v. X is $\mathcal{F}$-measurable.

   a) $\Omega = \{1, 2, 3, 4\}$; $\mathcal{F}$ is generated by the partition of Ω into even and odd numbers; $X(\omega) = \omega^2$.

   b) $\Omega = \{1, 2, 3, 4\}$; $\mathcal{F}$ is generated by the partition of Ω into even and odd numbers; $X(\omega) = \frac{1}{4}\omega^4 - 10\omega^2 + 30\omega$.

   c) $\Omega = [0, 1] \times [0, 1]$; $\mathcal{F}$ is the $\sigma-$algebra of vertical sets (i.e. sets of the form $A \times [0, 1]$); $X(x, y) = xy$.

   d) $\Omega = [0, 1] \times [0, 1]$; $\mathcal{F}$ is the $\sigma-$algebra of vertical sets (i.e. sets of the form $A \times [0, 1]$); $X(x, y) = y^2 - y + 3$.

   e) $\Omega = [0, 1] \times [0, 1]$; $\mathcal{F}$ is the $\sigma-$algebra of vertical sets (i.e. sets of the form $A \times [0, 1]$); $X(x, y) = x^3 - x$.

3. Suppose you are betting on fair coin flips (as usual, you win if you flip heads, and lose if you flip tails), and that you implement Strategy 3 as described in Section 10.2 of the notes (bet $\$1$ on the first flip; afterwards, bet $\$2$ if you lost the previous flip and $\$1$ if you won the previous flip). If the first eight flips are H T T H T T H H, compute the amount you have won or lost in the first eight flips.

4. Suppose you are betting on fair coin flips (as usual, you win if you flip heads, and lose if you flip tails), and that you implement Strategy 4 as described in Section 10.2 of the notes. If your initial bankroll is $\$100$, compute the expected amount of your bankroll after 3 flips.

5. Suppose you are betting on fair coin flips (as usual, you win if you flip heads, and lose if you flip tails), and you implement a strategy described as follows: on the first flip, bet 1. On even numbered flips (the second, fourth, sixth, etc.), bet 3 if you won the previous flip, and bet 1 if you lost the previous flip. On odd numbered flips (other than the first flip), bet 2 if the preceding two flips were the same, and bet 1 if the preceding two flips were different.

   a) Let B_t be the size of your bet on the t^{th} flip. Define B_t using mathematical notation.
   b) Suppose the results of the first ten flips are H T T H H T H H H T. Assuming $X_0 = 0$, compute $(B \cdot X)_t$ for $0 \leq t \leq 10$.

6. Suppose $\{X_t\}$ is a discrete-time process with state space $\mathbb{Z}$. Determine, with brief justification, whether or not the following random variables are stopping times:

   a) $T_a = \min\{t \geq 0 : X_t = X_0\}$.
   b) $T_b = \min\{t \geq 0 : X_t \geq 20\}$.
   c) $T = \min\{t \geq 0 : X_t = X_{10}\}$.
   d) $T = \max\{t \geq 0 : X_t = X_0\}$.
   e) $T = \min\{T_a, T_b\}$, where T_a and T_b are as in parts (a) and (b).
   f) $T = \max\{T_a, T_b\}$, where T_a and T_b are as in parts (a) and (b).

Exercises from Section 10.3

7. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ have the uniform distribution and suppose $\mathcal{F}$ is the $\sigma-$algebra generated by the partition $\{\{1, 2\}, \{3, 4, 5\}, \{6\}$ of Ω. Let X be the random variable defined by $X(1) = 5$, $X(2) = X(3) = X(4) = 1$, $X(5) = X(6) = 9$. Compute $E[X|\mathcal{F}]$.

8. Let $\Omega = [0, 1] \times [0, 1]$ have the uniform distribution, and let $X : \Omega \to \mathbb{R}$ be $X(x, y) = x^2y + x$.

   a) Compute $E[X|\mathcal{F}_x]$, where $\mathcal{F}_x$ is the σ-algebra of vertical sets (i.e. sets of the form $A \times [0, 1]$).
   b) Compute $E[X|\mathcal{F}_y]$, where $\mathcal{F}_y$ is the σ-algebra of horizontal sets (i.e. sets of the form $[0, 1] \times B$).

9. Suppose $\{X_t\}$ is a discrete-time stochastic process in which you flip a coin that flips heads with probability $\frac{1}{3}$ and tails with probability $\frac{2}{3}$. Let $\{\mathcal{F}_t\}$

be the natural filtration of $\{X_t\}$, and let X be a random variable defined by setting

$$X = \begin{cases} 0 & \text{if the first three flips are heads} \\ 10 & \text{if the first two flips are heads but the third is tails} \\ 4 & \text{if the first flip is heads but the second flip is tails} \\ -7 & \text{if the first flip is tails but the second and third flips are heads} \\ -1 & \text{if the first flip is tails and the second and third flips have} \\ & \qquad \text{opposite results} \\ 3 & \text{if the first three flips are tails} \end{cases}$$

Compute $E[X|\mathcal{F}_1]$ and $E[X|\mathcal{F}_2]$.

Exercises from Section 10.4

10. Let $\{X_t\}$ be the Wright-Fisher chain (introduced in a group presentation). Prove that $\{X_t\}$ is a martingale.

11. Let $\{X_t\}$ be the Pólya urn model. For each t, let M_t be the fraction of balls in the urn which are red. Prove that $\{M_t\}$ is a martingale.

12. (20 ★ pts) Modify the Pólya urn model so that you add $c \geq 2$ balls of the color you most recently drew to the urn after each draw (instead of adding one marble of the color you drew). Is the $\{M_t\}$ described in Problem 11 still a martingale?

Exercises from Section 10.5

13. (20 ★ pts) Finish the proof of the Escape Time Corollary (Theorem 10.21), by writing a proof of the second statement of that theorem.

14. Suppose $\{X_t\}$ is an escaping process so that $\{\sqrt{X_t}\}$ is a martingale. Compute $P_{16}(T_{25} < T_9)$.

15. Suppose $\{X_t\}$ is an irreducible, escaping Markov chain with state space $\mathcal{S} = \mathbb{Z}$. Suppose also that $\{e^{-X_t}\}$ is a martingale.

    a) Compute $f_{2,3}$.
    b) Compute $f_{2,1}$.
    c) Classify this Markov chain as recurrent or transient.

Exercises from Section 10.6

16. Prove Lemma 10.24 from the notes, which says that for a simple, random walk, $\mu = p - q$, and that if the walk is unbiased, then $\sigma^2 = p + q$.

17. Prove the second statement of Lemma 10.27 from the notes, which says that if $\{X_t\}$ is an irreducible, simple random walk then $\{(X_t - t\mu)^2 - t\sigma^2\}$ is a martingale (M_{S_j} is the MGF of the step size S_j).

18. (20 ★ pts) Suppose $\{X_t\}$ is an irreducible, simple random walk and let t be any constant. Prove that $\left\{\frac{e^{\theta X_t}}{[M_{S_j}(t)]^t}\right\}$ is a martingale.

19. (20 ★ pts) Prove Wald's Third Identity, which says: let $\{X_t\}$ is an irreducible, simple random walk starting at 0. Let $a < 0 < b$ be integers and set $T = \min\{T_a, T_b\} = T_{\{a,b\}}$. Then
$$E\left[\frac{e^{\theta X_T}}{[M_{S_j}(\theta)]^T}\right] = 1.$$

20. (20 ★ pts) Finish the proof of Gambler's Ruin (Theorem 10.31) by writing out the cases where the walk is negatively biased.

21. (20 ★ pts) Finish the proof of Theorem 10.32 by writing out a proof that negatively biased, irreducible, simple random walks are transient.

22. A gambler makes a series of independent $\$1$ bets. He decides to quit betting as soon as his net winnings reach $\$25$ or his net losses reach $\$50$. Suppose the probabilities of his winning and losing each bet are each equal to $\frac{1}{2}$.
    a) Find the probability that when he quits, he will have lost $\$50$.
    b) Find the expected amount he wins or loses.
    c) Find the expected number of bets he will make before quitting.

23. A typical roulette wheel has 38 numbered spaces, of which 18 are black, 18 are red, and 2 are green. A gambler makes a series of independent $\$1$ bets, betting on red each time (such a bet pays him $\$1$ if the ball in the roulette wheel ends up on a red number). He decides to quit betting as soon as his net winnings reach $\$25$ or his net losses reach $\$50$.
    a) Find the probability that when he quits, he will have lost $\$50$.
    b) Find the expected amount he wins or loses.
    c) Find the expected number of bets he will make before quitting.

24. Suppose two friends, George the Genius and Ichabod the Idiot, play a game that has some elements of skill and luck in it. Because George is better at the game than Ichabod, George wins 55% of the games they play and Ichabod wins the other 45% (the result of each game is independent of each other game). Suppose George and Ichabod both bring $\$100$ to bet with, and they agree to play until one of them is broke.

    a) Suppose George and Ichabod wager $\$1$ on each game. What is the probability that George ends up with all the money?
    b) Suppose George and Ichabod wager $\$5$ on each game. What is the probability that George ends up with all the money?
    c) Suppose George and Ichabod wager $\$25$ on each game. What is the probability that George ends up with all the money?
    d) Suppose George and Ichabod wager $\$100$ on each game. What is the probability that George ends up with all the money?
    e) Based on the answers to parts (a), (b) and (c), determine which of the following statements is true:

       Statement I: The more skilled player benefits when the amount wagered on each game increases.

       Statement II: The more skilled player is harmed when the amount wagered on each game increases.
    f) Suppose you had $\$1000$ and needed $\$2000$ right away, and you therefore decided to go to a casino and turn your $\$1000$ into $\$2000$ by gambling on roulette. In light of your answer to the previous question, which of these strategies gives you the highest probability of ending up with $\$2000$: betting $\$1000$ on red on one spin of the wheel, or betting $\$1$ on red repeatedly, trying to work your way up to $\$2000$ without going broke first?

25. Consider an irreducible, simple random walk X_t starting at zero, where $r = 0$.

    a) Find the probability that $X_t = -2$ for some $t > 0$.
    b) Find p such that $P(X_t = 4 \text{ for some } t > 0) = \dfrac{1}{2}$.

Exercises from Section 10.7

26. Let $\{X_t\}$ be an irreducible birth and death chain with $\mathcal{S} = \{0, 1, 2, 3, ...\}$. Prove that if for all $x \geq 1$, $p_x \leq q_x$, then the chain is recurrent.

27. Let $\{X_t\}$ be an irreducible birth and death chain with $\mathcal{S} = \{0, 1, 2, 3, ...\}$ such that
$$\frac{q_x}{p_x} = \left(\frac{x}{x+1}\right)^2 \qquad \text{for all } x \geq 1.$$
   a) Is this chain recurrent or transient?
   b) Compute $f_{x,0}$ for all $x \geq 1$.
   Hint: $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$.

28. Compute the stationary distribution of the Ehrenfest chain, for arbitrary d.
Hint: the Ehrenfest chain is a birth and death chain.

29. Compute all stationary distributions of the Markov chain with state space $\mathcal{S} = \{0, 1, ..., d\}$ and transition function
$$\begin{cases} P(x, x+1) &= \dfrac{(d-x)^2}{d^2} \\ P(x, x) &= \dfrac{2x(d-x)}{d^2} \\ P(x, x-1) &= \dfrac{x^2}{d^2} \end{cases}$$
(This chain was introduced in Exercise 4 of the Chapter 8 Homework... it counts the number of black balls in one of two boxes, when randomly chosen balls are exchanged between two boxes at each step of the chain.)
Hint: You may need the following identity, which can be assumed without proof:
$$\sum_{j=0}^{d} \binom{d}{j}^2 = \binom{2d}{d}.$$

30. Consider a birth and death chain on $\mathcal{S} = \{0, 1, 2, ...\}$ with
$$p_x = \frac{1}{2^{x+1}}\ \forall x;\ q_x = \frac{1}{2^{x-1}}\ \forall x > 1;\ q_1 = \frac{1}{2}.$$
   a) Prove this chain is positive recurrent.
   b) Compute its stationary distribution.
   c) Compute the mean return time to state 2.

Exercises from Section 10.8

31. Consider a birth and death CTMC $\{X_t\}$ on $\{0, 1, 2, 3, ...\}$ whose death rates are given by $\mu_x = x$ for all $x \in \mathcal{S}$.

    a) Determine whether the process is transient, null recurrent or positive recurrent, if the birth rates are $\lambda_x = x + 1$ for all $x \in \mathcal{S}$.

    b) Determine whether the process is transient, null recurrent or positive recurrent, if the birth rates are $\lambda_x = x + 2$ for all $x \in \mathcal{S}$.

32. Let $\{X_t\}$ be a continuous-time Markov branching process with $p = \dfrac{2}{3}$.

    a) Compute the extinction probability η.

    b) Compute $f_{5,0}$.

    c) What is the minimum value of x so that $P_x(T_0 = \infty) > \dfrac{99}{100}$?

33. If $\{X_t\}$ is a continuous-time Markov branching process with extinction probability $\eta = \dfrac{3}{8}$, what is the probability that each particle splits into two "children" (as opposed to dying)?

34. In Chapter 9, we derived the fact that the stationary distribution of the $(M/M/\infty)$-queue was $Pois\left(\frac{\lambda}{\mu}\right)$ by computing the time t transitions and taking their limit as $t \to \infty$. We can also determine this stationary distribution using the machinery of Chapter 10:

    a) Let $\{X_t\}$ be the $(M/M/\infty)$-queue. Explain why $\{X_t\}$ is a birth and death CTMC.

    b) Determine the birth and death rates of $\{X_t\}$.

    c) Compute ζ_y for each $y \in \{0, 1, 2, 3, ...\}$.

    d) Prove using Theorem 10.45 that $\{X_t\}$ is positive recurrent.

    e) Verify using Theorem 10.45 that the stationary distribution of $\{X_t\}$ is $Pois\left(\dfrac{\lambda}{\mu}\right)$.

35. Suppose customers call a technical support line according to a Poisson process with parameter $\lambda > 0$. They are provided with technical support by n agents where n is a positive integer (n is a constant, not a r.v.). Suppose that the amount of time it takes an agent to solve a customer's problem is exponentially distributed with parameter μ (and that these times are independent of the Poisson process and all independent of one another). Last, assume that whenever there are more than n customers calling the technical support line,

the excess customers get placed on hold until one of the n agents is available. Let X_t represent the number of people on the phone with technical support (including those on hold) at time t. $\{X_t\}$ is called the $n-$*server queue* or the $(M/M/n)-$*queue*.

a) Explain why $\{X_t\}$ is a birth and death CTMC.
b) Compute the birth and death rates of $\{X_t\}$.
c) Compute ζ_y for each $y \in \{0, 1, 2, 3, ...\}$.
d) Show that $\lambda < n\mu$ if and only if $\{X_t\}$ is positive recurrent.
e) Show that $\lambda > n\mu$ if and only if $\{X_t\}$ is transient.

36. (60 ★ pts) Suppose d particles are distributed into two boxes, A and B. Each particle in box A remains in that box for a random length of time that is exponentially distributed with parameter μ before moving to box B. Each particle in box B remains in that box for a random length of time that is exponentially distributed with parameter λ before moving to box A. All particles act independently of one another. For each $t \geq 0$, let X_t be the number of particles in box A at time t. Then $\{X_t\}$ is a birth and death CTMC on $\mathcal{S} = \{0, 1, 2, ..., d\}$.

a) This setup be thought of as a continuous version of what discrete-time Markov chain?
b) Find the birth and death rates.
c) Find $P_{x,d}(t)$ for all $x \in \mathcal{S}$. *Hint:* Think of each particle as generating its own CTMC, where state zero corresponds to being in box B and state 1 corresponds to being in box A. This is a two-state CTMC, so its transition probabilities were derived in Chapter 9. From these transition probabilities, you can get the probability that any one fixed particle is in box A at time t. Multiply these together to get $P_{x,d}(t)$.
d) Find $E_x(X_t)$. *Hint:* Write $X_t = A_t + B_t$ where A_t is the number of particles in box A that started in box A and B_t is the number of particles in box A at time t that started in box B. If $X_0 = x$, then A_t and B_t are both binomial, defined in terms of x and the transition function of the two-state birth-death process described in the hint for part (c).
e) Compute the steady-state distribution for this process; identify this distribution as a common r.v. (stating the parameters).
f) Verify that as $t \to \infty$, $E_x(X_t)$ converges to the expected value of the steady-state distribution.

Chapter 11

Brownian motion

11.1 Definition and construction

Goal: Develop a model for "continuous random movement", i.e. a version of simple random walk where both the index set (the set of times) and the state space are continuous: we want $\mathcal{I} = [0, \infty)$ and $\mathcal{S} = \mathbb{R}$.

Construction of this process

Let $\{S_t\}_{t=1}^{\infty}$ be the steps of a random walk $\{X_t\}$ starting at x with $r = 0$ (no loops), i.e. $\{S_t\}$ is an i.i.d. process with

s	-1	1
$f_{S_t}(s) = P(S_t = s)$	$1-p$	p

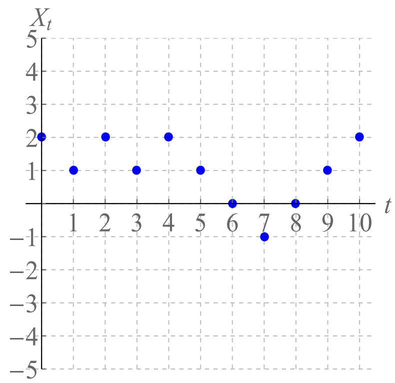
.

OBSERVATIONS

- the step size $\{S_t\}$ is ± 1 unit; and

$$ES_j = p - (1-p) = 2p - 1; \qquad Var(S_j) = E[S_j^2] = 1 - (2p-1)^2 = 4p(1-p).$$

- the random walk $\{X_t\}$ satisfies $X_t = X_0 + \sum_{j=1}^{t} S_t = x + \sum_{j=1}^{t} S_t$.



- $E[X_t] = x + t\,E[S_j] - t = x + t(2p-1)$, so $\boxed{E[X_t] \text{ is linear in } t}$.
- $Var[X_t] = Var[X_0] + Var\left[\sum_{j=1}^{t} S_j\right] = 0 + t\,Var(s_j) = 4p(1-p)t$, so

  $\boxed{Var(X_t) \text{ is proportional to } t}$.
- By the CLT, the sum of t independent S_j's is approximately normal

  $$n((2p-1)t, 4p(1-p)t).$$

  Restated, this means that for large t,

  $$\begin{aligned} X_t = X_0 + \sum_{j=1}^{t} S_t &\approx x + n((2p-1)t, 4p(1-p)t) \\ &= n\left(x + (2p-1)t, 4p(1-p)t\right). \end{aligned}$$

We are going to build a new process by considering "random walks" where the steps take place more and more frequently, and where the sizes of the steps are shrunk.

The eventual goal is to define a process $\{W_t\}$ with index set $[0,\infty)$ and state space $\mathbb{R}$, where like simple random walk,

$$\boxed{E[W_t] \text{ is linear in } t} \text{ and } \boxed{Var(W_t) \text{ is proportional to } t},$$

and the process has the additional property that

$$\boxed{\text{the sample functions } t \mapsto W_t \text{ are continuous}}.$$

Toward that end, for each $n \in \mathbb{N}$, think of a "random walk" that steps every $\Delta t = \dfrac{t}{n}$ units of time (so that the walk steps n times between times 0 and n).

A preview: eventually we want to let $n \to \infty$, so that there will be an "infinite number of steps" that are "infinitely close together".

First try: suppose $X_t = x + \sum_{j=1}^{n} S_j$.

Problem! If we do this, then $Var(X_t) = n\,Var(S_j) = n\,[4p(1-p)]$.

As $n \to \infty$, this variance tends to ∞, so we would end up with a process $\{X_t\}$ for which $Var(X_t) = \infty$ for all t. This is bad; we want $Var(X_t)$ to be proportional to the time t (not the number of steps n).

Fix: change the size of each step from 1 unit to Δx units, and let $\left\{W_t^{(n)}\right\}$ be the process defined by adding n independent steps of size Δx:

$$W_t^{(n)} = x + \sum_{j=1}^{n} (\Delta x)\, S_j.$$

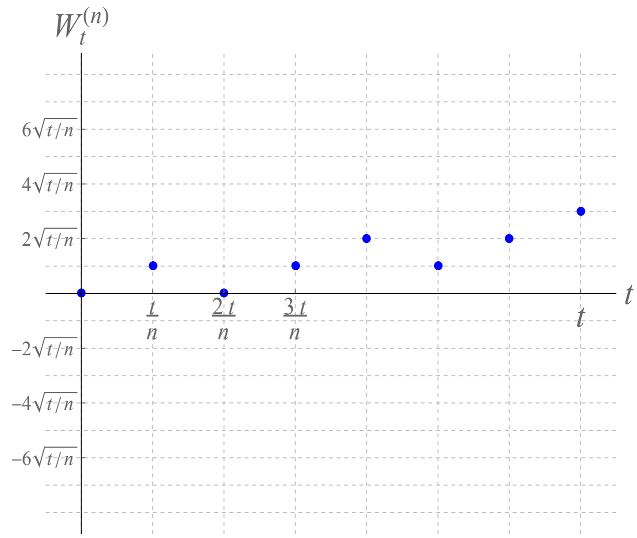
Now,

$$\begin{aligned}
Var\left(W_t^{(n)}\right) &= Var\left(x + \sum_{j=1}^{n} (\Delta x)\, S_j\right) \\
&= \sum_{j=1}^{n} Var\left(\Delta x\, S_j\right) \\
&= \sum_{j=1}^{n} (\Delta x)^2\, Var(S_j) \\
&= (\Delta x)^2 \sum_{j=1}^{n} Var(S_j) \\
&= (\Delta x)^2\, n\, Var(S_j) \\
&= (\Delta x)^2\, n\, \left[4p(1-p)\right].
\end{aligned}$$

We want this variance to be proportional to t, so we need $(\Delta x)^2\, n$ proportional to t, so we need a constant $\sigma^2 > 0$ so that

$$(\Delta x)^2 n = \sigma^2 t, \quad \text{i.e.} \quad (\Delta x)^2 = \sigma^2 \frac{t}{n} = \sigma^2 \Delta t, \quad \text{i.e.} \quad \boxed{\Delta x = \sigma\sqrt{\Delta t}}.$$

In other words, we need **the size of the jumps to be proportional to the square root of the time between jumps**.



We're not done yet; we need to make the mean of $W_t^{(n)}$ work out. At this point,

$$
\begin{aligned}
E\left[W_t^{(n)}\right] &= E\left[x + \sum_{j=1}^{n} (\Delta x)\, S_j\right] \\
&= x + (\Delta x)\, E\left[\sum_{j=1}^{n} S_j\right] \\
&= x + (\Delta x)\, n(2p-1) \\
&= x + \left(\sigma\sqrt{\Delta t}\right)\, n\,(2p-1) \\
&= x + \sigma\sqrt{\frac{t}{n}}\, n\,(2p-1) \\
&= x + \sigma(2p-1)\,\sqrt{t}\sqrt{n}.
\end{aligned}
$$

To make this a linear function of t, we need $2p-1$ proportional to $\frac{\sqrt{t}}{\sqrt{n}}$, i.e. we need a constant μ so that

$$
\sigma(2p-1) = \mu\left(\frac{\sqrt{t}}{\sqrt{n}}\right), \quad \text{i.e.} \quad \boxed{p = \frac{1}{2}\left[1 + \frac{\mu\sqrt{t}}{\sigma\sqrt{n}}\right]}.
$$

That makes $E\left[W_t^{(n)}\right] = x + \frac{\sqrt{t}}{\sqrt{n}} \cdot \sqrt{t}\,\sqrt{n}\mu = x + \mu\, t$ which is the kind of formula we want (the mean is linear in t and doesn't depend on n).

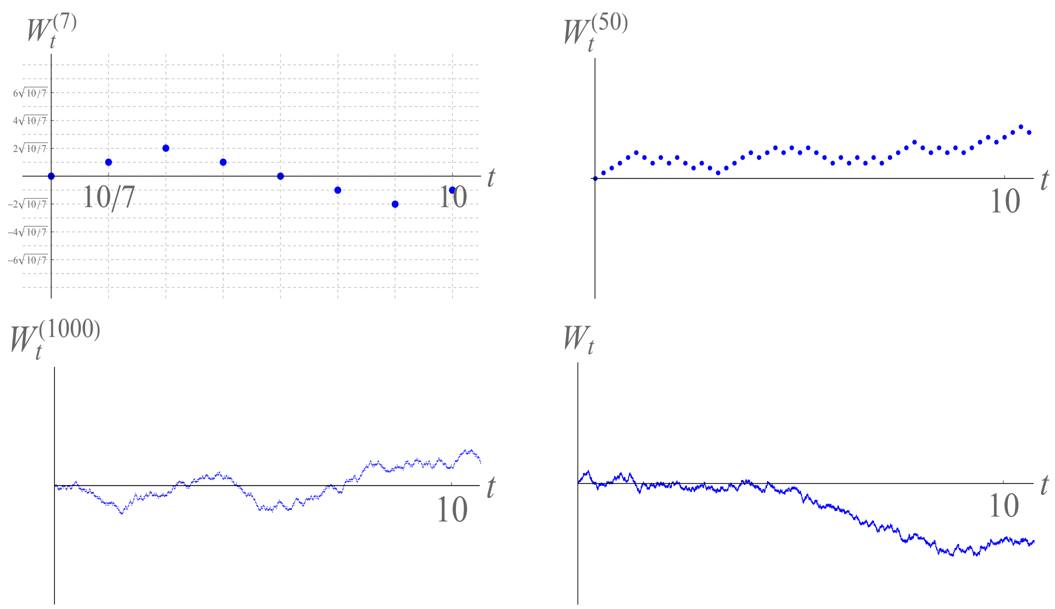
Unfortunately, this choice of p screws with the variance, because now,

$$\begin{aligned} Var(W_t^{(n)}) &= (\Delta x)^2 \, n \, [4p(1-p)] \\ &= \left(\sqrt{\sigma \, \Delta t}\right)^2 \, n \, [4p(1-p)] \\ &= \left(\sigma^2 \Delta t\right) n \, [4p(1-p)] \\ &= \sigma^2 \, t \left[1 - (2p-1)^2\right] \qquad\qquad \text{(since } \Delta t = \frac{t}{n}\text{)} \\ &= \sigma^2 \, t \left(1 - \mu^2 \frac{t}{n}\right). \end{aligned}$$

This quantity depends on n, which seems bad. **BUT!** as $n \to \infty$, this variance tends to $\sigma^2 t$, which is proportional to t and doesn't depend on n.

Since the $W_t^{(n)}$ are (x plus) the sum of the tn i.i.d. r.v.s $\{\Delta x) \, S_j$, as $n \to \infty$ they will tend to a normal r.v. by the Central Limit Theorem. From above, the parameters of this normal r.v. must be μt and $\sigma^2 t$, so we can conclude that

$$W_t = \lim_{n \to \infty} W_t^{(n)} \sim n(x + \mu t, \sigma^2 t).$$



This produces a stochastic process $\{W_t\}$ called a *Brownian motion* (or a *Weiner process*). To summarize, the way we constructed this $\{W_t\}$ is

$$\begin{aligned} W_t &= \lim_{n \to \infty} W_t^{(n)} \\ &= \lim_{n \to \infty} \left(\begin{array}{c} \text{random walks with} \\ \text{steps that occur at times separated by gaps } \Delta t = \frac{t}{n}, \\ \text{where the size of each step is } \pm \Delta x = \pm \sigma \sqrt{\Delta t} = \pm \sigma \sqrt{\frac{t}{n}} \end{array} \right). \end{aligned}$$

Definition of Brownian motion

Definition 11.1 *A stochastic process* $\{W_t\}$ *is called a* **Brownian motion (BM) with drift** *if there are numbers* $\mu \in \mathbb{R}$, $\sigma^2 > 0$ *and* $x \in \mathbb{R}$ *so that:*

the process starts at x**:** $W_0 = x$ *(assume* $x = 0$ *unless told otherwise).*

increments of the process are normal *(with parameters proportional to the elapsed time): for every* $s < t$,

$$W_t - W_s \sim n(\mu(t-s), \sigma^2(t-s)).$$

the process has independent increments: *for any* $0 \le t_1 < t_2 < \cdots < t_n$*, the r.v.s*

$$W_{t_2} - W_{t_1},\ W_{t_3} - W_{t_2},\ \ldots,\ W_{t_n} - W_{t_{n-1}}$$

are independent.

sample functions are continuous: *with probability* 1*, the functions* $t \mapsto W_t$ *are continuous.*

The number μ *is called the* **drift parameter** *of* $\{W_t\}$*; we say* $\{W_t\}$ *has* **positive drift** *if* $\mu > 0$ *and has* **negative drift** *if* $\mu < 0$.

If the process has no drift, i.e. $\mu = 0$*, we call* $\{W_t\}$ *a* **Brownian motion**.

The number σ^2 *is called the* **variance parameter** *of* $\{W_t\}$.

A **standard Brownian motion** *is a BM starting at* $x = 0$ *with no drift (*$\mu = 0$*) and variance parameter* $\sigma^2 = 1$.

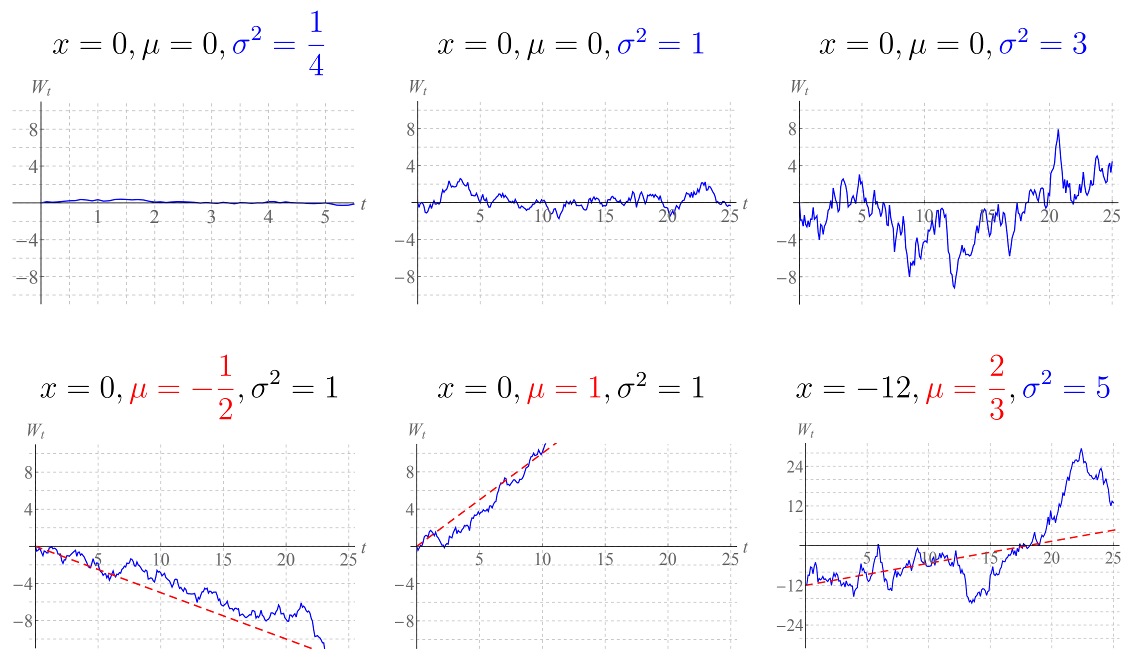
Theorem 11.2 *The process* $\{W_t\}$ *we constructed earlier is, in fact, a BM.*

Real-world applications of Brownian motion

- movements of particles suspended in a liquid
  (first noticed by Robert Brown (1827), for whom the process is named);
- fluctuations in the stock market;
- the path-integral formulation of quantum mechanics;
- option pricing models (Black-Scholes equations);
- cosmology models

Why is BM so prevalent? As we have seen, Brownian motions approximate random walks with small but frequent jumps (so long as the size of the jump is proportional to the square root of the time between jumps).

Example sample functions of BMs



You can simulate a sample function of a BM with this *Mathematica* code:

ListLinePlot[x+RandomFunction[WienerProcess[μ, σ^2], {$xmin$, $xmax$, .01}]]

(Change x, μ and σ^2 to match the parameters of the given BM.)

An interesting question

Assume the price of a share of Coca-Cola is modeled by a BM with $\mu = \frac{1}{2}$ and $\sigma^2 = 4$, and that the price of a share of McClendon Soft Drink Corporation is modeled by a BM with $\mu = -1$ and $\sigma^2 = 16$. Both stocks are currently valued at $\$100$ per share.

Would you rather buy a share of Coca-Cola or a share of McClendon Soft Drink Corporation?

What do we know about Brownian motion so far?

EXAMPLE 1

Suppose $\{W_t\}$ is a BM starting at $x = 2$ with parameters $\mu = \frac{1}{3}$ and $\sigma^2 = 9$.

1. Describe the random variable W_3.
2. Describe the random variable $W_8 - W_2$.
3. Compute the probability that $W_8 > 1$.
4. Compute the probability that $W_7 - W_5 \leq 2$.
5. Compute the probability that $W_8 - W_7 < 1$ and $W_{14} - W_{12} > -3$.
6. Compute the expected value of W_6.
7. Compute the covariance between W_6 and W_{11}.

11.2 Symmetries and scaling laws

This section is about identifying properties of a Brownian motion that come from the fact that "everything is normal". More specifically, we will discover formulas of a Brownian motion that turn out to be themselves BMs.

To prove that these formulas are actually BMs, we could check the four criteria of Definition 11.1. However, there is a better way that involves a key property of BMs: namely, that they are *Gaussian*:

Gaussian processes

Definition 11.3 *Let $\{X_t\}$ be a stochastic process with index set $\mathcal{I}$. We say that $\{X_t\}$ is* **Gaussian** *if any finite linear combination of the X_t's is joint normal.*

More precisely, we require that for any times $t_1, ..., t_n \in \mathcal{I}$ and any constants $c_1, ..., c_n \in \mathbb{R}$, the random variable

$$\sum_{j=1}^{n} c_j X_{t_j} = c_1 X_{t_1} + c_2 X_{t_2} + ... + c_n X_{t_n}$$

is normal.

Theorem 11.4 *Any BM $\{W_t\}$ is a Gaussian process.*

PROOF To show this, let $c_1, ..., c_n \in \mathbb{R}$ and let $t_1, ..., t_n \in [0, \infty)$; without loss of generality $t_1 < t_2 < ... < t_n$. Let $t_0 = 0$ (for notational purposes only). The goal is to verify that

$$\sum_{i=1}^{n} c_j W_{t_i} = c_1 W_{t_1} + c_2 W_{t_2} + ... + c_n W_{t_n}$$

is normal. To show this, break this sum into independent increments:

$$\begin{aligned}
&\sum_{i=1}^{n} c_j W_{t_i} \\
&= c_1 W_{t_1} + c_2 W_{t_2} + c_3 W_{t_3} + ... + c_n W_{t_n} \\
&= c_1 W_{t_1} + c_2 \left[W_{t_1} + (W_{t_2} - W_{t_1})\right] + c_3 \left[W_{t_1} + (W_{t_2} - W_{t_1}) + (W_{t_3} - W_{t_2})\right] + ... \\
&= (c_1 + ... + c_n) W_{t_1} + (c_2 + ... + c_n)(W_{t_2} - W_{t_1}) + (c_3 + ... + c_n)(W_{t_3} - W_{t_2}) + ... \\
&= \left[\sum_{j=1}^{n} c_j\right] W_{t_j} + \left[\sum_{j=2}^{n} c_j\right] (W_{t_2} - W_{t_1}) + \left[\sum_{j=3}^{n}\right] (W_{t_3} - W_{t_2}) + ...
\end{aligned}$$

From the previous page,

$$\sum_{i=1}^{n} c_j W_{t_i} = \sum_{i=1}^{n} \left[\sum_{j=i}^{n} c_j \right] (W_{t_i} - W_{t_{i-1}}).$$

Since $\{W_t\}$ is a BM, each of the terms inside the parentheses above are normal and independent.

That means any linear combination of them is normal, so $\sum_{j=1}^{n} b_j W_{t_j}$ is normal.

By definition, $\{W_t\}$ is Gaussian. $\square$

Lemma 11.5 *Let $\{W_t\}$ be a BM. For any function $f : [0, \infty) \to \mathbb{R}$ and any function $g : [0, \infty) \to [0, \infty)$ and any constant $b \in \mathbb{R}$, if we define the process $\{X_t\}$ by*

$$X_t = f(t) W_{g(t)} + b,$$

then $\{X_t\}$ is Gaussian.

PROOF Let $c_1, ..., c_n \in \mathbb{R}$ and let $t_1, ..., t_n \in [0, \infty)$. Then,

$$\sum_{j=1}^{n} c_j X_{t_j} = \sum_{j=1}^{n} c_j \left[f(t_j) W_{g(t_j)} + b \right] = b\, n + \sum_{j=1}^{n} \left[c_j f(t_j) \right] W_{g(t_j)}$$

is a linear combination of the W_t's (which is normal since $\{W_t\}$ is Gaussian) plus a constant, which is normal.

That means $\{X_t\}$ is Gaussian. $\square$

Criteria for a process being a BM

Now, we can give a criteria for checking whether a process is a BM. The idea is that *any Gaussian process with the same means and covariances as a BM must be a BM*.

Theorem 11.6 *Suppose $\{X_t : t \in [0, \infty)\}$ is a stochastic process such that:*

1. *$\{X_t\}$ is Gaussian;*
2. *there is a constant a so that $E[X_t] = x + at$; and*
3. *there is a constant $b > 0$ so that $Cov(X_s, X_t) = b \min(s, t)$.*

Then $\{X_t\} \sim \{W_t\}$, where $\{W_t\}$ is a BM starting at x with drift $\mu = a$ and variance parameter $\sigma^2 = b$.

PROOF We check the parts of Definition 11.1 one-by-one:

Starts at x: by assumption (3), $Var[X_0] = Cov(X_0, X_0) = b\min(0,0) = 0$.

Therefore X_0 is a constant, which must be x since $E[X_0] = x + a(0) = x$.

Normal increments: this follows from $\{X_t\}$ being assumed Gaussian, since any increment $X_{t_1} - X_{t_0}$ is a linear combination of the X_t's.

$\perp$ increments: Let $t_1 < t_2 < t_3 < t_4$.

$$\begin{aligned} Cov(X_{t_2} - X_{t_1}, Cov(X_{t_4} - X_{t_3}) &= Cov(X_{t_2}, X_{t_4}) - Cov(X_{t_2}, X_{t_3}) \\ &\quad - Cov(X_{t_1}, X_{t_4}) + Cov(X_{t_1}, X_{t_3}) \\ &= b\min(t_2, t_4) - b\min(t_2, t_3) \\ &\quad - b\min(t_1, t_4) + b\min(t_1, t_3) \\ &= b\,t_2 - b\,t_2 - b\,t_1 + b\,t_1 \\ &= 0. \end{aligned}$$

Therefore $X_{t_2} - X_{t_1}$ and $X_{t_4} - X_{t_3}$ are uncorrelated. But since $\{X_t\}$ is Gaussian, any combination of the X_t's is normal. That means $(X_{t_1}, X_{t_2}, X_{t_3}, X_{t_4})$ has a joint normal distribution, which implies that uncorrelated combinations of those variables are independent (see Chapter 6). Thus $\{X_t\}$ has independent increments.

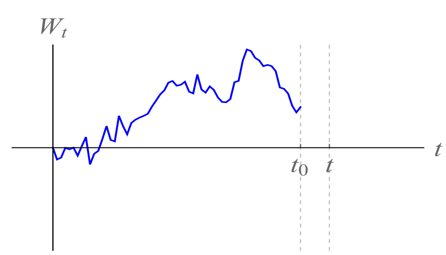
Cts sample functions (sketch of proof): fix $t_0 \geq 0$. Since $\{X_t\}$ is Gaussian, for each t, $X_t - X_{t_0}$ is normal with mean

$$E[X_t - X_{t_0}] = E[X_t] - E[X_{t_0}] = x + at - (x + at_0) = a(t - t_0)$$

and variance

$$\begin{aligned} Var(X_t - X_{t_0}) &= Cov(X_t - X_{t_0}, X_t - X_{t_0}) \\ &= Cov(X_t, X_t) - Cov(X_t, X_{t_0}) - Cov(X_{t_0}, X_t) + Cov(X_{t_0}, X_{t_0}) \\ &= b\,t - 2b\min(t, t_0) + b\,t_0 \\ &= b(t + t_0 - 2\min(t, t_0)). \end{aligned}$$

If t is close enough to t_0, both the mean and variance of $X_t - X_{t_0} \sim n(a(t - t_0), b(t + t_0 - 2\min(t, t_0))$ will be small. This will force $X_t - X_{t_0}$ to be small with high probability, which will force the sample functions to be continuous (with probability 1). $\square$



Examples

Theorem 11.7 *Let $\{W_t\}$ be a BM. Then, for any constants a and b, $\{aW_t + b\}$ is a BM.*

PROOF Suppose $\{W_t\}$ starts at x, has drift μ and variance parameter σ^2.

First,

Second,

Third,

So by Theorem 11.6, $\{-W_t\}$ is a BM starting at $ax+b$, with drift $a\mu$ and variance parameter $a^2\sigma^2$. □

Theorem 11.8 (Universal scaling law) *Let $\{W_t\}$ be a BM with zero drift, starting at zero. Then, for any constant a, $\{aW_{t/a^2}\}$ is a BM with zero drift and the same variance parameter as $\{W_t\}$.*

PROOF Suppose $\{W_t\}$ has variance parameter σ^2.

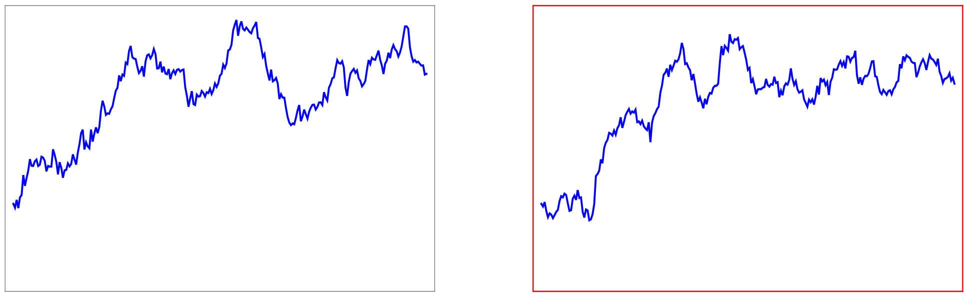
First, by Lemma 11.5 with $f(t) =$, $g(t) =$ and $b =$, $\{aW_{t/a^2}\}$ is Gaussian.

Second, $E[aW_{t/a^2}] =$

Third, $Cov(aW_{s/a^2}, aW_{t/a^2}) =$

Theorem 11.6 gives the result. □

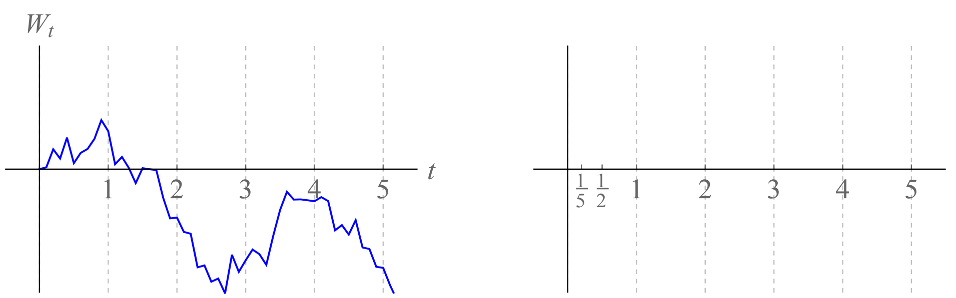
The universal scaling law tells us that if we take a trajectory of a zero drift BM that starts at 0, and zoom in on part of it (zooming in faster horizontally than we do vertically), we will see the same thing no matter how much we zoom in, i.e. the trajectories are "self-similar". Thus the trajectories in a BM are objects called **fractals**.



Theorem 11.9 (Inversion symmetry) *Let $\{W_t\}$ be a BM with zero drift, starting at zero. Then $\{t\,W_{1/t}\}$ is a zero drift BM starting at zero.*

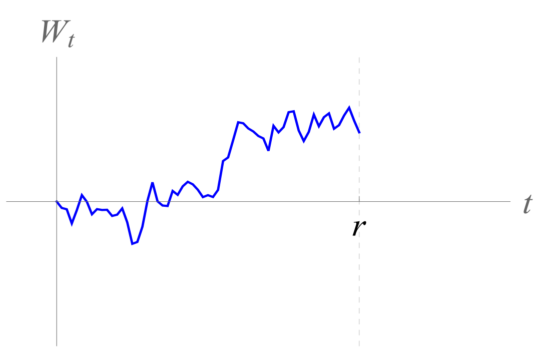
PROOF HW

A picture explaining inversion symmetry:



Markov properties

Theorem 11.10 (Markov property for BM) *Let $\{W_t\}$ be a BM. For any constant $r \geq 0$, $\{W_{r+t} - W_r\}$ is a BM, independent of W_r.*



PROOF Let $X_t = W_{r+t} - W_r$.

First, we show $\{X_t\}$ is Gaussian: let $c_1, ..., c_n \in \mathbb{R}$ and $t_1, ..., t_n \in [0, \infty)$:

$$\begin{aligned}
\sum_{j=1}^{n} c_j X_j &= \sum_{j=1}^{n} c_j \left[W_{r+t_j} - W_r \right] \\
&= \sum_{j=1}^{n} \left[c_j W_{r+t_j} - c_j W_r \right] \\
&= \sum_{j=1}^{n} c_j W_{r+t_j} - \left(\sum_{j=1}^{n} c_j \right) W_r
\end{aligned}$$

The first term is normal since $\{W_t\}$ is Gaussian; the second term is normal and independent of the first, so the whole sum is normal. This makes $\{X_t\}$ Gaussian by defintion.

Second, we compute the mean of $\{X_t\}$:

$$EX_t = E[W_{r+t} - W_r] = EW_{r+t} - EW_r = x + \mu(r+t) - (x + \mu r) = \mu t.$$

Third, we compute the covariances of $\{X_t\}$:

$$\begin{aligned}
Cov(X_s, X_t) &= Cov(W_{s+r} - W_r, W_{t+r} - W_r) \\
&= Cov(W_{s+r}, W_{t+r}) - Cov(W_r, W_{t+r}) - Cov(W_{s+r}, W_r) + Cov(W_r, W_r) \\
&= \sigma^2 \min(s+r, t+r) - \sigma^2 r - \sigma^2 r + \sigma^2 r \\
&= \sigma^2 [\min(s,t) + r] - \sigma^2 r \\
&= \sigma^2 \min(s,t).
\end{aligned}$$

By Theorem 11.6, $\{X_t\}$ is a BM starting at 0 with drift μ and variance parameter σ^2. It is independent of W_r by the independent increment property. □

A stronger version of the Markov property is this result (which, by the way, also holds for Markov chains and CTMCs). Its proof is beyond the scope of this class:

Theorem 11.11 (Strong Markov property) *Let $\{W_t\}$ be a BM and let T be a stopping time for $\{W_t\}$. Define $X_t = W_{T+t} - W_T$. Then $\{X_t\}$ is a BM, independent of $\{W_t : t \leq T\}$.*

11.3 Martingales and escape problems

In this section, we look at properties of BM that resemble properties of random walks we studied in the previous chapter.

Escaping property

Theorem 11.12 *Let $\{W_t\}$ be a BM. Then $\{W_t\}$ is escaping.*

PROOF Let $a < x < b$ and let $T = T_{\{a,b\}}$. We need to verify two properties: the first is that $P_x(T < \infty)$. To do this, for $n \in \{0, 1, 2, ...\}$, let E_n be the event that between times n and $n+1$, the value of the BM goes up by at least $b-a$. This event has probability

$$p = P(E_n) = P(W_{n+1} - W_n > b - a) = P(n(\mu, \sigma^2) > b - a)$$
$$= 1 - \Phi\left(\frac{b - a - \mu}{\sigma}\right) \in (0, 1),$$

and since the E_n are $\perp$,

$$P(\text{no } E_n \text{ occurs}) = P\left(\bigcap_{n=0}^{\infty} E_n^C\right) = \prod_{n=0}^{\infty} P(E_n^C) = \prod_{n=0}^{\infty}(1-p) = (1-p)^{\infty} = 0.$$

Thus, for some n, E_n occurs. That means the BM goes up by at least $b-a$ between times n and $n+1$, so it must be that for the n where E_n occurs, either $W_n \leq a$ or $W_{n+1} \geq b$. Either way, $T < \infty$.

The second property, that $P_x(W_t \in (a, b) \,|\, t < T_{\{a,b\}}) = 1$, follows from the fact that the sample functions are continuous. □

Associated martingales

Theorem 11.13 *Let $\{W_t\}$ be a BM. Then these processes are all martingales:*

- $\{W_t - \mu t\}$
- $\{(W_t - \mu t)^2 - \sigma^2 t\}$
- $\left\{\exp\left(\frac{-2\mu}{\sigma^2} W_t\right)\right\}$

PROOF We start with the proof that $\{W_t - \mu t\}$ is a martingale.
Let $\{\mathcal{F}_t\}$ be the natural filtration of $\{W_t\}$, and let $0 < s < t$.

Our goal is to show $E[W_t - \mu t \,|\, \mathcal{F}_s] = E[W_s - \mu s]$:

$$E[W_t - \mu t \,|\, \mathcal{F}_s]$$

$$= E[W_s + (W_t - W_s) - \mu t \,|\, \mathcal{F}_s]$$

(this is a typical way of breaking up W_t when proving things are martingales)

$$\begin{aligned}
&= E[W_s \,|\, \mathcal{F}_s] + E[W_t - W_s \,|\, \mathcal{F}_s] - E[\mu t \,|\, \mathcal{F}_s] \\
&= W_s + E[W_t - W_s \,|\, \mathcal{F}_s] - \mu t \\
&= W_s + E[W_t - W_s] - \mu t \\
&= W_s + \mu(t-s) - \mu t \\
&= W_s - \mu s.
\end{aligned}$$

Thus $\{W_t - \mu t\}$ is a martingale.

The other proofs are HW problems. □

Escape probabilities

Theorem 11.14 *Suppose $\{W_t\}$ is a BM, and let $a < x < b$.*

- *If the BM has zero drift, then*

$$P_x(T_a < T_b) = \frac{b-x}{b-a} \qquad \textit{and} \qquad P_x(T_b < T_a) = \frac{x-a}{b-a}.$$

- *If the BM has drift $\mu \neq 0$, then*

$$P_x(T_a < T_b) = \frac{\exp\left(\frac{-2\mu b}{\sigma^2}\right) - \exp\left(\frac{-2\mu x}{\sigma^2}\right)}{\exp\left(\frac{-2\mu b}{\sigma^2}\right) - \exp\left(\frac{-2\mu a}{\sigma^2}\right)}$$

and

$$P_x(T_b < T_a) = \frac{\exp\left(\frac{-2\mu x}{\sigma^2}\right) - \exp\left(\frac{-2\mu a}{\sigma^2}\right)}{\exp\left(\frac{-2\mu b}{\sigma^2}\right) - \exp\left(\frac{-2\mu a}{\sigma^2}\right)}.$$

PROOF Suppose that $\{W_t\}$ has zero drift. Then $\{W_t\}$ is a martingale, and since $\{W_t\}$ is escaping, the Escape Time Theorem (Theorem 10.20) applies with $\psi(x) = x$ to give

$$P_x(T_a < T_b) = \frac{\psi(b) - \psi(x)}{\psi(b) - \psi(a)} = \frac{b-x}{b-a}$$

and

$$P_x(T_b < T_a) = \frac{\psi(x) - \psi(a)}{\psi(b) - \psi(a)} = \frac{x-a}{b-a}.$$

The proof when the BM has drift is HW. □

EXAMPLE 2

Suppose the price of a stock is currently $\$70$. If the price is modeled with a BM with drift with $\mu = \frac{1}{2}$ and $\sigma^2 = 8$, what is the probability the price of the stock hits $\$80$ before it hits $\$60$?

Theorem 11.15 (Wald's First Identity for BM) *Let $\{W_t\}$ be a BM and suppose $a < x < b$. Let $T = T_{\{a,b\}}$. Then*

$$E_x[W_T] = x + \mu\, ET.$$

PROOF From a previous theorem, we know that $\{W_t - \mu t\}$ is a martingale. Therefore

$$E_x[W_T] - \mu\, ET = E_x[W_T - \mu\, T] \stackrel{\text{OST}}{=} E_x[W_0 - \mu(0)] = x - 0 = x.$$

Add $\mu\, ET$ to both sides to get the result. □

Theorem 11.16 (Wald's Second Identity for BM) *Let $\{W_t\}$ be a BM with zero drift and suppose $a < x < b$. Let $T = T_{\{a,b\}}$. Then:*

$$E_x[W_T^2] = x^2 + \sigma^2\, E_x[T].$$

PROOF From earlier, we know $\{(W_t - \mu t)^2 - \sigma^2 t\}$ is a martingale; since the BM has zero drift this reduces to $\{W_t^2 - \sigma^2 t\}$. Therefore

$$E_x(W_T^2 - \sigma^2 T) \stackrel{\text{OST}}{=} E_x(W_0^2 - \sigma^2(0)) = x^2.$$

Also,

$$E_x(W_T^2 - \sigma^2 T) = E_x(W_T^2) - E_x(\sigma^2 T) = E_x[W_T^2] - \sigma^2 E_x[T].$$

Equate these two formulas and add $\sigma^2 E_x[T]$ to both sides to get Wald's Second Identity. □

Theorem 11.17 *Let $\{W_t\}$ be a BM with zero drift and suppose $a < x < b$. Then*

$$E_x[T_{\{a,b\}}] = \frac{(x-a)(b-x)}{\sigma^2}.$$

PROOF Let $T = T_{\{a,b\}}$. From earlier, we can compute

$$\begin{aligned}
E_x[W_T^2] &= a^2 P_x(T_a < T_b) + b^2 P_x(T_b < T_a) && \text{(LOTUS)} \\
&= a^2 \left(\frac{b-x}{b-a}\right) + b^2 \left(\frac{x-a}{b-a}\right) && \text{(escape probabilities)} \\
&= \frac{a^2 b - a^2 x + b^2 x - b^2 a}{b-a} \\
&= \frac{ab(a-b) + (b^2 - a^2)x}{b-a} \\
&= \frac{-ab(b-a) + (b-a)(b+a)x}{b-a} = -ab + (b+a)x = ax + bx - ab.
\end{aligned}$$

By Wald's Second Identity, we therefore have

$$ax + bx - ab = x^2 + \sigma^2 E_x[T];$$

solve for $E_x[T]$ to get

$$E_x[T] = \frac{ax + bx - ab - x^2}{\sigma^2} = \frac{(x-a)(b-x)}{\sigma^2}. \square$$

EXAMPLE 3

Suppose the price of a stock is modeled by a BM with no drift and $\sigma^2 = 5$. If the price of the stock is initially 40,

1. What is the probability that the stock price hits 60 before it hits 30?
2. How long should expect to wait until the first instant where the stock price is either 30 or 60?

Recurrence and transience

Theorem 11.18 *Suppose $\{W_t\}$ is a BM, and let $x \neq y$ be real numbers. Then:*

$$f_{x,y} = P_x(T_y < \infty) = \begin{cases} 1 & \text{if the BM has zero drift} \\ 1 & \text{if the BM drifts from } x \text{ towards } y \\ \exp\left(\dfrac{-2\mu|x-y|}{\sigma^2}\right) & \text{if the BM drifts from } x \text{ away from } y \end{cases}$$

PROOF This proof splits into several cases, all of which apply the Escape Time Corollary (Theorem 10.21) and the escape probabilities of BM (Theorem 11.14).

First, if the BM has zero drift, then for $\psi(x) = x$, $\{\psi(W_t)\}$ is a martingale.

Case 1: if $x < y$, then $f_{x,y} = \lim_{a \to -\infty} P_x(T_y < T_a) = \lim_{a \to -\infty} \dfrac{x-a}{y-a} \stackrel{L}{=} \lim_{a \to -\infty} \dfrac{-1}{-1} = 1$.

Case 2: if $x > y$, then $f_{x,y} = \lim_{b \to \infty} P_x(T_y < T_b) = \lim_{b \to \infty} \dfrac{b-x}{b-y} \stackrel{L}{=} \lim_{b \to \infty} \dfrac{1}{1} = 1$.

On the other hand, if the BM has drift $\mu \neq 0$, then for $\psi(x) = \exp\left(\frac{-2\mu}{\sigma^2}x\right)$, $\{\psi(W_t)\}$ is a martingale.

Case 3: if $x < y$ and $\mu > 0$ (so the BM drifts towards y), then

$$\begin{aligned} f_{x,y} = \lim_{a \to -\infty} P_x(T_y < T_a) &= \lim_{a \to -\infty} \frac{\exp\left(\frac{-2\mu x}{\sigma^2}\right) - \exp\left(\frac{-2\mu a}{\sigma^2}\right)}{\exp\left(\frac{-2\mu y}{\sigma^2}\right) - \exp\left(\frac{-2\mu a}{\sigma^2}\right)} \\ &\stackrel{L}{=} \lim_{a \to -\infty} \frac{0 - \exp\left(\frac{-2\mu a}{\sigma^2}\right) \cdot \frac{-2\mu}{\sigma^2}}{0 - \exp\left(\frac{-2\mu a}{\sigma^2}\right) \cdot \frac{-2\mu}{\sigma^2}} = 1. \end{aligned}$$

Case 4: if $x < y$ and $\mu < 0$ (so the BM drifts away from y), then

$$\begin{aligned} f_{x,y} = \lim_{a \to -\infty} P_x(T_y < T_a) &= \lim_{a \to -\infty} \frac{\exp\left(\frac{-2\mu x}{\sigma^2}\right) - \exp\left(\frac{-2\mu a}{\sigma^2}\right)}{\exp\left(\frac{-2\mu y}{\sigma^2}\right) - \exp\left(\frac{-2\mu a}{\sigma^2}\right)} \\ &= \frac{\exp\left(\frac{-2\mu x}{\sigma^2}\right) - \exp(-\infty)}{\exp\left(\frac{-2\mu y}{\sigma^2}\right) - \exp(-\infty)} \\ &= \exp\left(\frac{-2\mu x}{\sigma^2} + \frac{2\mu y}{\sigma^2}\right) \\ &= \exp\left(\frac{-2\mu(y-x)}{\sigma^2}\right) = \exp\left(\frac{-2\mu|x-y|}{\sigma^2}\right). \end{aligned}$$

This leaves two cases corresponding to when $x > y$. These are left as HW. $\square$

Theorem 11.19 *Let $\{W_t\}$ be a BM.*

- *If $\{W_t\}$ has zero drift, then $\{W_t\}$ is recurrent.*
- *If $\{W_t\}$ has nonzero drift, then $\{W_t\}$ is transient.*

PROOF Assume first that the BM has zero drift. Then, by the continuous LTP (conditioning on the value of W_1), we have

$$\begin{aligned} f_x = f_{x,x} = P_x(T_x < \infty) &= \int_{-\infty}^{\infty} P_y(T_x < \infty)\, f_{W_1}(y)\, dy \\ &= \int_{-\infty}^{\infty} f_{y,x}\, f_{n(x+\mu,\sigma^2)}(y)\, dy \\ &= \int_{-\infty}^{\infty} 1\, f_{n(x+\mu,\sigma^2)}(y)\, dy \\ &= 1 \end{aligned}$$

so x is recurrent, making the BM $\{W_t\}$ recurrent.

Now, assume that the BM has drift $\mu \geq 0$ (otherwise consider $\{-W_t\}$). Then,

$$\begin{aligned} f_x = f_{x,x} &= P_x(T_x < \infty) \\ &= \int_{-\infty}^{\infty} P_y(T_x < \infty)\, f_{W_1}(y)\, dy \\ &= \int_{-\infty}^{\infty} f_{y,x} f_{n(x+\mu,\sigma^2)}(y)\, dy. \end{aligned}$$

Now split this integral into the part where $y \leq x$ and the part where $y > x$:

$$f_x = \int_{-\infty}^{x} f_{y,x}\, f_{n(x+\mu,\sigma^2)}(y)\, dy + \int_{x}^{\infty} f_{y,x}\, f_{n(x+\mu,\sigma^2)}(y)\, dy.$$

For the first integral, since $y < x$ and the drift is positive, $f_{y,x} = 1$ so we get

$$\int_{-\infty}^{x} (1)\, f_{n(x+\mu,\sigma^2)}(y)\, dy = P(n(x+\mu,\sigma^2) \leq x) = \Phi\left(\frac{x+\mu-x}{\sigma}\right) = \Phi(0) = \frac{1}{2}.$$

For the second integral, since $y > x$ and the drift is positive, $f_{y,x} = \exp\left(\frac{-2\mu|x-y|}{\sigma^2}\right)$ so we get

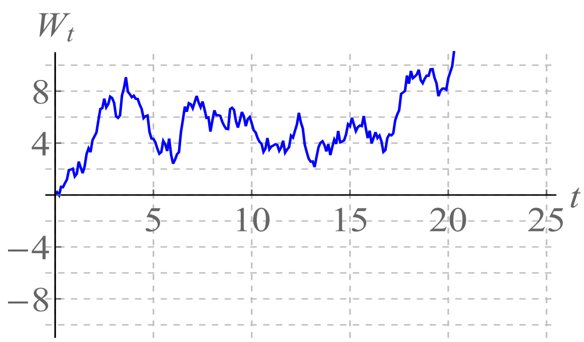
$$\begin{aligned} &\int_{x}^{\infty} \exp\left(\frac{-2\mu|x-y|}{\sigma^2}\right) f_{n(x+\mu,\sigma^2)}(y)\, dy \\ &= \int_{x}^{\infty} \exp\left(\frac{-2\mu(y-x)}{\sigma^2}\right) \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(y-x-\mu)^2}{2\sigma^2}\right) dy \\ &= \int_{x}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(y-x-\mu)^2 - 4\mu(y-x)}{2\sigma^2}\right] dy. \end{aligned}$$

Now use the u-sub $u = y - x$, $du = dy$ to get

$$\begin{aligned}
&\int_0^\infty \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(u-\mu)^2 - 4\mu(u)}{2\sigma^2}\right] du \\
&= \int_0^\infty \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(u+\mu)^2}{2\sigma^2}\right] du \\
&= \int_0^\infty f_{n(-\mu,\sigma^2)}(u)\, du \\
&= P(n(-\mu,\sigma^2) > 0) < \frac{1}{2}.
\end{aligned}$$

Adding the blue and red parts together, we see $f_x < \frac{1}{2} + \frac{1}{2} = 1$.
Therefore x is transient, meaning $\{W_t\}$ is transient. $\square$

But there's actually more to show here - we need to know that $\{W_t\}$ returns to its starting value *at an unbounded set of times in the future*.



We'll address this question (and other things) in the next section.

11.4 Reflection principle

QUESTION

Let $\{W_t\}$ be a BM with zero drift. From the previous section, we know W_t hits b, i.e. $f_{x,b} = 1$, i.e. $P_x(T_b < \infty) = 1$.

What is the distribution of the r.v T_b which measures the time it takes to hit b?

Theorem 11.20 (Reflection principle) *Let $\{W_t\}$ be a BM with zero drift, starting at x. Fix $b \neq x$. Then*

$$F_{T_b}(t) = P(T_b \leq t) = 2 - 2\Phi\left(\frac{|x-b|}{\sigma\sqrt{t}}\right).$$

PROOF *Case 1: $b > x$.* We observe first that $W_t \geq b$ only if $T_b \leq t$:

Therefore

$$P(W_t \geq b) = P(W_t \geq b \bigcap T_b \leq t) = P(W_t \geq b \,|\, T_b \leq t)P(T_b \leq t)$$

which implies

$$F_{T_b}(t) = P(T_b \le t) = \frac{P(W_t \ge b)}{P(W_t \ge b \,|\, T_b \le t)}$$

$$=$$

$$=$$

$$=$$

$$= 2 - 2\Phi\left(\frac{b-x}{\sigma\sqrt{t}}\right)$$

$$= 2 - 2\Phi\left(\frac{|x-b|}{\sigma\sqrt{t}}\right).$$

Case 2: $b < x$. Here, $W_t \le b$ only if , so

Corollary 11.21 *Let* $\{W_t\}$ *be a zero drift BM with parameter* σ^2 *starting at* x. *Fix* $b > 0$ *and let* $T_b = \min\{t \ge 0 : W_t = b\}$. *Then* T_b *has density*

$$f_{T_b}(t) = \frac{|x-b|}{\sigma\sqrt{2\,\pi\,t^3}} \exp\left[\frac{-(x-b)^2}{2\,t\,\sigma^2}\right].$$

PROOF HW (just differentiate F_{T_b} with respect to t and simplify). □

EXAMPLE 4

Let $\{W_t\}$ be a BM with parameter 6. If $W_1 = 2$, what is the probability $W_t = -5$ for some $t \in [1, 5]$?

Consequences of the reflection principle

Theorem 11.22 (Strong recurrence of BM) *Let $\{W_t\}$ be a BM with zero drift. With probability 1, there is an unbounded set of times t such that $W_t = W_0$.*

PROOF It is sufficient to show $P_0(W_s = 0 \text{ for some } s \geq 1) = 1$. We have

$$\begin{aligned}
&P_0(W_s = 0 \text{ for some } s \geq 1) \\
&= \lim_{t\to\infty} P_0(W_s = 0 \text{ for some } s \in [1,t]) \\
&= \lim_{t\to\infty} \int_{-\infty}^{\infty} f_{W_1}(b) P_0(W_s = 0 \text{ for some } s \in [1,t] \,|\, W_1 = b)\, db
\end{aligned}$$

$$= \lim_{t\to\infty} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-b^2}{2\sigma^2}\right) \left[2 - 2\Phi\left(\frac{b}{\sigma\sqrt{t-1}}\right)\right]\, db$$

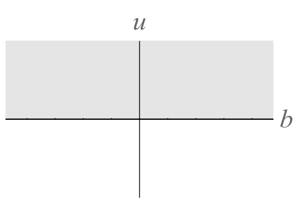
From the previous page,

$$
\begin{aligned}
&P_0(W_s = 0 \text{ for some } s \geq 1) \\
&= \lim_{t\to\infty} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-b^2}{2\sigma^2}\right) \left[2 - 2\Phi\left(\frac{b}{\sigma\sqrt{t-1}}\right)\right] db
\end{aligned}
$$

$$
\begin{aligned}
&= \lim_{t\to\infty} \frac{2}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(\frac{-b^2}{2\sigma^2}\right) \int_{\frac{b}{\sigma\sqrt{t-1}}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) dx\, db
\end{aligned}
$$

$$
\begin{aligned}
&= \lim_{t\to\infty} \frac{2}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(\frac{-b^2}{2\sigma^2}\right) \int_{\frac{b}{\sqrt{t-1}}}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-u^2}{2\sigma^2}\right) du\, db \\
&= \frac{1}{\pi\sigma^2} \int_{-\infty}^{\infty} \int_0^{\infty} \exp\left(\frac{-(u^2+b^2)}{2\sigma^2}\right) du\, db
\end{aligned}
$$

(now change the integral to polar coordinates)



$$
\begin{aligned}
&= \frac{1}{\pi\sigma^2} \int_{r=0}^{\infty} \int_0^{\pi} \exp\left(\frac{-r^2}{2\sigma^2}\right) r\, d\theta\, dr \\
&= \frac{1}{\sigma^2} \int_0^{\infty} e^{-r^2/2\sigma^2} r\, dr \\
&\qquad \left(\text{let } v = -\frac{r^2}{2\sigma^2};\ dv = -\frac{r}{\sigma^2} dr, \text{ i.e. } -\sigma^2\, dv = r\, dr\right) \\
&= \frac{1}{\sigma^2}(-\sigma^2) \int_0^{-\infty} e^v\, dv \\
&= \int_{-\infty}^{0} e^v\, dv \\
&= e^0 - e^{-\infty} = 1. \ \square
\end{aligned}
$$

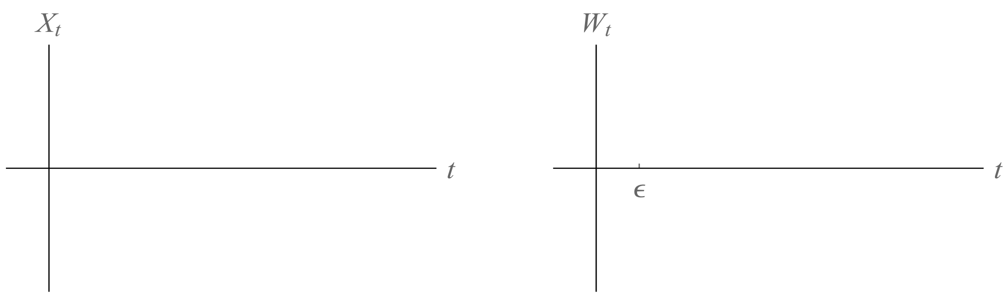
Theorem 11.23 *Let $\{W_t\}$ be a BM with zero drift. For every $\epsilon > 0$ (no matter how small), there are infinitely many times $t \in (0, \epsilon)$ so that $W_t = W_0$.*

PROOF First, we can assume $W_0 = 0$ (otherwise shift $\{W_t\}$ by a constant so that it starts at 0).

Next, let $X_t = \begin{cases} tW_{1/t} & \text{if } t > 0 \\ 0 & \text{if } t = 0 \end{cases}$. From Theorem 11.9, $\{X_t\}$ is also a BM with zero drift.

By strong recurrence, there is an unbounded set of times $t_1, t_2, t_3, \ldots$ such that

$$X_{t_1} = X_{t_2} = \ldots = X_0 = 0.$$

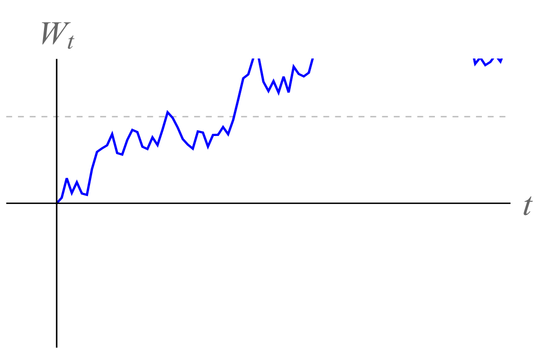


But that means $W_{1/t_1}, W_{1/t_2}, \ldots$ must also all be zero.

And given any $\epsilon > 0$, there will be infinitely many of the times $\frac{1}{t_1}, \frac{1}{t_2}, \frac{1}{t_3} \ldots$ in the interval $(0, \epsilon)$ (since the t_j are unbounded). □

CONSEQUENCE

If the trajectory of a zero drift BM crosses a horizontal line, then it actually crosses that horizontal line *infinitely many times* that are arbitrarily close to any one of the times it crosses the line:



Theorem 11.24 (Nondifferentiability of paths) *Let $\{W_t\}$ be a BM. With probability 1, a Brownian sample function $t \mapsto W_t$ is nowhere differentiable (i.e. not differentiable at any time t).*

CONSEQUENCE

With probability 1, the sample functions of a Brownian motion are "infinitely jagged", i.e. nowhere smooth.

PROOF We proceed with two cases:

Case 1: $\{W_t\}$ has zero drift.

In this case, we will first prove the sample function isn't differentiable at 0. To do this, by the definition of derivative,

$$\begin{aligned}
\left.\frac{d}{dt}W_t\right|_{t=0} \text{ exists} \quad &\iff \lim_{h\to 0}\frac{W_h - W_0}{h} \text{ exists} \\
&\iff \lim_{h\to 0}\frac{W_h}{h} \text{ exists} \\
&\Rightarrow \frac{W_h}{h} < A \quad \begin{array}{l}\text{for some fixed constant } A, \\ \text{and for all } h \in (0,\epsilon)\end{array} \\
&\iff W_h < Ah \text{ for all } h \in (0,\epsilon).
\end{aligned}$$

But by the reflection principle,

$$\begin{aligned}
\lim_{h\to 0} P(W_h < Ah) &= \lim_{h\to 0}\left[1 - \left(2 - 2\Phi\left(\frac{Ah}{\sqrt{h}}\right)\right)\right] \\
&= \lim_{h\to 0}\left[2\Phi(A\sqrt{h}) - 1\right] \\
&= 2\Phi(0) - 1 \\
&= 2\left(\tfrac{1}{2}\right) - 1 = 0.
\end{aligned}$$

Therefore $P\left(\left.\frac{d}{dt}W_t\right|_{t=0} \text{ exists}\right) = 0.$

Now, if $\{W_t\}$ is a BM with zero drift that is differentiable at t_0, $\{W_{t+t_0} - W_{t_0}\}$ would be a BM with zero drift that was differentiable at 0, contradicting the above argument. Therefore $\{W_t\}$ is nowhere differentiable (with probability 1).

Case 2: $\{W_t\}$ has nonzero drift μ.

If such a BM is differentiable at time t, then $\{W_t - \mu t\}$ is differentiable at time t (as it is the difference of two differentiable functions).

But $\{W_t - \mu t\}$ is a BM with zero drift, so this would violate Case 1. □

11.5 Brownian motion in higher dimensions

Definition 11.25 *Let $\{\mathbf{W}_t\}$ (a.k.a. $\{\overrightarrow{W_t}\}$) be a stochastic process with state space $\mathbb{R}^d$. $\{\mathbf{W}_t\}$ is called a d-**dimensional Brownian motion** if each coordinate of the process is a BM, and the coordinates are independent.*

*If each of the coordinates is a standard BM (they have zero drift, start at 0 and variance parameter 1), then we call $\{\mathbf{W}_t\}$ a **standard** d-**dimensional BM**.*

Escape probabilities

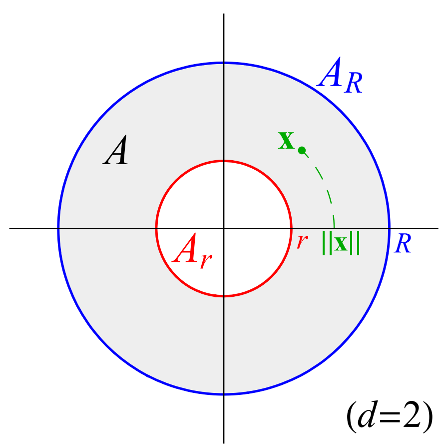
Let $\{\mathbf{W}_t\}$ be a standard $d-$dim'l BM and fix $0 < r < R < \infty$.

Define the sets

$$\begin{aligned} A_r &= \{\mathbf{x} \in \mathbb{R}^d : ||\mathbf{x}|| = r\}; \\ A_R &= \{\mathbf{x} \in \mathbb{R}^d : ||\mathbf{x}|| = R\}; \\ A &= \{\mathbf{x} \in \mathbb{R}^d : ||\mathbf{x}|| \in (r, R)\}; \end{aligned}$$

and also let

$$\begin{aligned} T_{A_r} &= \min\{t \geq 0 : \mathbf{W}_t \in A_r\}; \\ T_{A_R} &= \min\{t \geq 0 : \mathbf{W}_t \in A_R\}; \\ T &= \min\{T_1, T_2\}. \end{aligned}$$

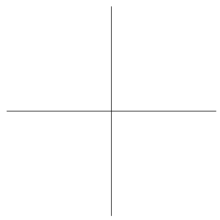


Our goal is to determine the escape probabilities $P_{\mathbf{x}}(T_{A_r} < T_{A_R})$ and $P_{\mathbf{x}}(T_{A_R} < T_{A_r})$.

To do this, for $\mathbf{x} \in A$, define $f(\mathbf{x}) = P_{\mathbf{x}}(T_{A_R} < T_{A_r})$.

By rotational symmetry, we can write $f(\mathbf{x}) = g(||\mathbf{x}||)$ for some function $g : \mathbb{R} \to \mathbb{R}$ such that $g(r) = 0$ and $g(R) = 1$.

f has another important property: the value of f at $\mathbf{x}$ is equal to the average value of f along any circle of small radius centered at $\mathbf{x}$:



Therefore $f : A \to \mathbb{R}$ is what is called a *harmonic* function, meaning it satisfies the following equation, which is called the *heat equation* (Google the "Dirichlet problem" or "heat equation" for more on this):

$$\sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2} f(\mathbf{x}) = 0 \qquad \text{for all } \mathbf{x} \in A.$$

To analyze this equation, first observe that for any x_j, we can use the Chain Rule to obtain

$$\begin{aligned} \frac{\partial}{\partial x_j}(||\mathbf{x}||) &= \frac{\partial}{\partial x_j}\left(\sqrt{x_1^2 + \ldots + x_d^2}\right) \\ &= \frac{1}{2\sqrt{x_1^2 + \ldots + x_d^2}} \cdot 2x_j \\ &= \frac{x_j}{\sqrt{x_1^2 + \ldots + x_d^2}} \\ &= \frac{x_j}{||\mathbf{x}||}. \end{aligned} \tag{11.1}$$

Therefore

$$\begin{aligned} 0 &= \sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2} f(\mathbf{x}) \\ &= \sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2} g(||\mathbf{x}||) \\ &= \sum_{j=1}^{d} \frac{\partial}{\partial x_j}\left[g'(||\mathbf{x}||)\frac{x_j}{||\mathbf{x}||}\right] \\ &\quad \text{(using the Chain Rule with the above computation)} \\ &= \sum_{j=1}^{d}\left[g''(||\mathbf{x}||)\frac{x_j}{||\mathbf{x}||} \cdot \frac{x_j}{||\mathbf{x}||} + g'(||\mathbf{x}||)\frac{1 \cdot ||\mathbf{x}|| - \frac{x_j}{||\mathbf{x}||}x_j}{||\mathbf{x}||^2}\right] \\ &\quad \text{(Product and Quotient Rules)} \end{aligned}$$

From the previous page, we have

$$
\begin{aligned}
0 &= \sum_{j=1}^{d} \left[\frac{x_j^2 g''(||\mathbf{x}||)}{||\mathbf{x}||^2} + \frac{g'(||\mathbf{x}||)}{||\mathbf{x}||} - \frac{g'(||\mathbf{x}||)x_j^2}{||\mathbf{x}||^3} \right] \\
&= \frac{g''(||\mathbf{x}||)}{||\mathbf{x}||^2} \sum_{j=1}^{d} x_j^2 + \sum_{j=1}^{d} \frac{g'(||\mathbf{x}||)}{||\mathbf{x}||} - \frac{g'(||\mathbf{x}||)}{||\mathbf{x}||^3} \sum_{j=1}^{d} x_j^2 \\
&= \frac{g''(||\mathbf{x}||)}{||\mathbf{x}||^2} ||\mathbf{x}||^2 + d\, \frac{g'(||\mathbf{x}||)}{||\mathbf{x}||} - \frac{g'(||\mathbf{x}||)}{||\mathbf{x}||^3} ||\mathbf{x}||^2 \\
&\qquad \text{(since } \sum_{j=1}^{d} x_j^2 = ||\mathbf{x}||^2\text{)} \\
0 &= g''(||\mathbf{x}|| + d\, \frac{g'(||\mathbf{x}||)}{||\mathbf{x}||} - \frac{g'(||\mathbf{x}||)}{||\mathbf{x}||}
\end{aligned}
$$

Multiply through by $||\mathbf{x}||$ to obtain

$$0 = ||\mathbf{x}|| g''(||\mathbf{x}||) + (d-1) g'(||\mathbf{x}||). \tag{11.2}$$

Thinking of $||\mathbf{x}||$ as an independent variable "t", this is the second-order ODE

$$0 = t g''(t) + (d-1) g'(t). \tag{11.3}$$

which has no g in it (only t, g' and g''); therefore it can be solved with MATH 330 methods:

Integrate $g'(t) = Ct^{1-d}$ to get

$$g(t) = \begin{cases} & \text{if } d = 2 \\ & \text{if } d \geq 3 \end{cases}$$

If you plug in the known values of g (i.e. $g(r) = 0$ and $g(R) = 1$) and solve for the constants (HW), you will obtain:

Theorem 11.26 (Annular escape probabilities) *Let $\{\mathbf{W}_t\}$ be a standard d-dim'l BM. Suppose $r \le ||\mathbf{x}|| \le R$. Then, if A_r and A_R are the spheres of radius r and R centered at the origin, we have*

$$P_{\mathbf{x}}(T_{A_R} < T_{A_r}) = \begin{cases} \dfrac{x-r}{R-r} & \text{if } d = 1 \\ \dfrac{\ln ||\mathbf{x}|| - \ln r}{\ln R - \ln r} & \text{if } d = 2 \\ \dfrac{r^{2-d} - ||\mathbf{x}||^{2-d}}{r^{2-d} - R^{2-d}} & \text{if } d \ge 3 \end{cases}$$

In all cases, $P_{\mathbf{x}}(T_{A_r} < T_{A_R}) = 1 - P_{\mathbf{x}}(T_{A_R} < T_{A_r})$.

Recurrence/transience

Dimension 3 (or higher):

Suppose $r > 0$ is the radius of a small sphere centered at the origin. If a 3-dimensional BM travels to $\mathbf{x}$ with $||\mathbf{x}|| > r$, then

$$\begin{aligned} P_{\mathbf{x}}(T_{A_r} < \infty) &= \lim_{R\to\infty} P_{\mathbf{x}}(T_{A_r} < T_{A_R}) \\ &= 1 - \lim_{R\to\infty} P_{\mathbf{x}}(T_{A_R} < T_{A_r}) \\ &= \\ &= \\ &= \\ &= (\text{something bigger than } 1)^{\text{negative number}} \\ &< 1. \end{aligned}$$

So there is a chance that the BM never comes back to within r of the origin. Thus we say that in dimension 3 or higher, BM is transient.

Dimension 2:

(more interesting) Repeating the above calculation when $d = 2$, we get

$$\begin{aligned} P_{\mathbf{x}}(T_{A_r} < \infty) &= \lim_{R \to \infty} P_{\mathbf{x}}(T_{A_r} < T_{A_R}) \\ \\ &= \\ &= \\ \\ &= \\ \\ &= \end{aligned}$$

This time, it is assured that the BM will return to within r of the origin, so in dimension 2, BM is "neighborhood recurrent", because it returns to any "neighborhood" (i.e. within any positive distance) of where it was.

BUT: does a 2-dim'l BM get back exactly to where it was? Suppose a 2-dim'l BM starts at $\mathbf{0}$ and then travels distance $||\mathbf{x}||$ away. The probability that it returns to $\mathbf{0}$ is

$$\begin{aligned} P_{\mathbf{x}}(T_{A_0} < \infty) &= \lim_{r \to 0} P_{\mathbf{x}}(T_{A_r} < T_{A_R}) \\ &= 1 - \lim_{r \to 0} P_{\mathbf{x}}(T_{A_R} < T_{A_r}) \\ &= 1 - \lim_{r \to 0} \frac{\ln ||\mathbf{x}|| - \ln r}{\ln R - \ln r} \\ \\ &= \\ \\ &= \\ \\ &= \end{aligned}$$

Therefore, with probability 1, 2-dim'l BMs do <u>not</u> return to where they start, so 2-dim'l BM is "point transient".

Dimension 1:

We already proved 1-dim'l BM with zero drift is point recurrent in Theorem 11.19.

Putting everything together, we have shown the following set of facts:

Theorem 11.27 *Let $\{\mathbf{W}_t\}$ be a standard d-dimensional BM.*

1. *If $d = 1$, then $\{\mathbf{W}_t\}$ is point recurrent.*
2. *If $d = 2$, then $\{\mathbf{W}_t\}$ is point transient, but neighborhood recurrent.*
3. *If $d \geq 3$, then $\{\mathbf{W}_t\}$ is transient.*

EXAMPLE 5

Suppose a standard 3-dimensional BM starts at the point $(1, 1, 1)$. What is the probability that the point strikes the sphere of radius 1 centered at the origin before it strikes the sphere of radius 2 centered at the origin?

11.6 Chapter 11 Homework

Exercises from Section 11.1

1. Suppose $\{W_t\}$ is a Brownian motion starting at 0 with variance parameter $\sigma^2 = 3$.

   Note: if no drift is specified in a BM, that means the BM has zero drift.

   a) Compute $P(W_4 \geq 1)$.
   b) Compute $P(W_9 - W_2 \leq -2)$.
   c) Compute $P(W_7 > W_5)$.
   d) Compute the variance of W_8.
   e) Compute $Cov(W_3, W_7)$.
   f) Compute $Var(W_8 + W_9)$.

2. Suppose $\{X_t\}$ is a BM with drift $\mu = 5$ starting at $x = -1$ that has variance parameter $\sigma^2 = 4$.

   a) Compute $P(X_1 \geq 6)$.
   b) Compute $P(X_9 - X_7 \leq 3)$.
   c) Compute $P(X_4 > 15 \,|\, X_2 = 7, X_1 = -1)$.
   d) Compute $P(X_{13} > X_7)$.
   e) Compute the mean and variance of X_8.
   f) Compute $Cov(X_{11}, X_{16})$.
   g) Compute $Var(X_2 + X_5)$.

3. Suppose $\{W_t\}$ is a BM with $\sigma^2 = 5$. Compute

$$P(W_8 < 2 \,|\, W_2 = W_1 = 3).$$

Exercises from Section 11.2

4. Let $\{W_t\}$ be standard BM and let $X_t = (W_t)^2$ for all t.

   Recall: "standard" means $x = 0$, $\mu = 0$ and $\sigma^2 = 1$.

   a) Is $\{X_t\}$ a Gaussian process? Explain your answer.
   b) Find the mean function of $\{X_t\}$.
   c) Find the joint moment generating function of W_s and W_t.
   d) Use your answer to part (b) to find $E[W_s^2 W_t^2]$.

e) Find the covariance function of $\{X_t\}$.

5. Let $\{W_t\}$ be a BM with parameter σ^2 and let $a \leq s$ and $a \leq t$. Prove that
$$E[(W_s - W_a)(W_t - W_a)] = \sigma^2 \min(s-a, t-a).$$

6. Prove the **inversion symmetry** of BM (Theorem 11.9), which says that if $\{W_t\}$ is a Brownian motion with zero drift, starting at zero, then $\{tW_{1/t}\}$ is a standard Brownian motion.

7. Prove that if $\{W_t\}$ and $\{\widehat{W}_t\}$ are independent Brownian motions, then for any constants b_1 and b_2, the process $\{b_1 W_t + b_2 \widehat{W}_t\}$ is also a Brownian motion. Determine formulas for its parameters, in terms of b_1, b_2, and the parameters of $\{W_t\}$ and $\{\widehat{W}_t\}$.

   Note: The result of this problem generalizes: any linear combination of a finite number of independent BMs is also a BM (although you don't have to prove this).

Exercises from Section 11.3

8. Let $\{W_t\}$ be a BM. Prove that $\{(W_t - \mu t)^2 - \sigma^2 t\}$ is a martingale.

9. (20 ★ pts) Let $\{W_t\}$ be a BM. Prove that $\left\{\exp\left(\frac{-2\mu}{\sigma^2} W_t\right)\right\}$ is a martingale.

10. (30 ★ pts) Prove the second statement of Theorem 11.14 in the lecture notes, in which escape probabilities for BM with drift are derived.

11. (20 ★ pts) Let $\{W_t\}$ be a BM. Finish the proof of Theorem 11.18 by verifiying that if $x > y$, then
$$f_{x,y} = \begin{cases} 1 & \text{if } \mu < 0 \text{ (the BM drifts from } x \text{ towards } y) \\ \exp\left(\dfrac{-2\mu|x-y|}{\sigma^2}\right) & \text{if } \mu > 0 \text{ (the BM drifts from } x \text{ away from } y) \end{cases}.$$

12. You own one share of stock whose price is approximated by a BM with $\sigma^2 = 10$ (and time is measured in days). You bought the stock when its price was $\$15$, but now it is worth $\$25$.

    a) Suppose you decide to sell the stock when the price of the stock next reaches either $\$28$ or $\$20$:

       i. What is the probability you sell the stock for $\$28$?
       ii. What is the expected amount you will sell the stock for?
       iii. How much longer should you expect to hold the stock before selling?

b) Suppose you change your mind and decide to sell the stock when the price of the stock reaches $\$27$. What is the probability you hold the stock forever?

13. Suppose that you own a collectible item whose value at time t is modeled by a BM with drift where $\sigma^2 = 2$ and $\mu = \frac{1}{5}$. The item is presently valued at $\$30$, and you plan to sell the item when the value of the item reaches $\$45$ or $\$20$, whichever happens first.

    a) What is the probability that you sell the item for $\$45$?

    b) What is the expected value at which you will sell the item?

    c) How long should you expect to keep the item before you sell it?

14. The value of a painting is a BM with drift $\mu = 4$ and variance parameter $\sigma^2 = 8$.

    a) If the painting is currently worth 30, what is the probability that it is eventually worth 200?

    b) If the painting is currently worth 30, what is the probability that it is eventually worthless?

Exercises from Section 11.4

15. Suppose instead that you decide to sell the stock of Exercise 12 the next time its price hits $\$15$ or after ten days, whichever happens first. What is the probability that when you sell your stock, you will have to sell it for $\$15$?

16. Prove Corollary 11.21 from the notes, which says that if $\{W_t\}$ is a zero drift BM with parameter σ^2 starting at x, and if $T_b = \min\{t \geq 0 : W_t = b\}$, then T_b has density
$$f_{T_b}(t) = \frac{|x-b|}{\sigma\sqrt{2\,\pi\,t^3}} \exp\left[\frac{-(x-b)^2}{2\,t\,\sigma^2}\right].$$

17. Let $\{W_t\}$ be a Brownian motion with parameter σ^2 and fix $t \geq 0$. Let $M = \max\{W_s : 0 \leq s \leq t\}$. Prove M is a continuous r.v. (this implies $M > 0$ with probability one) and compute the density function of M.

    Note: This is a transformation problem. The way you show $M(t)$ is continuous is by computing its CDF and recognizing that this CDF is a continuous function.

18. (40 ★ pts) Let $\{W_t\}$ be a standard Brownian motion, and let $0 < t_0 < t_1$. Show
$$P(W_t = 0 \text{ for some } t \in (t_0, t_1)) = \frac{2}{\pi} \arctan\sqrt{\frac{t_1 - t_0}{t_0}}.$$

Hint: Condition on the value of W_{t_0} and use the result of Corollary 11.21. You will get a double integral; reverse the order of the integrals and then evaluate using a u-sub on each integral.

19. Let $\{W_t\}$ be a standard Brownian motion, and let L be the largest time $t \in [0,1]$ such that $W_t = 0$.

    a) Compute the density function of L. *Hint:* Use the fact proved in Exercise 18.
    b) Use a computer or graphing calculator to graph the density function you found in part (a)..
    c) Based on the graph you produce in part (b), describe qualitatively what is true about L (i.e. which values of L are most likely)?

Exercises from Section 11.5

20. In the lecture, we saw that for a 2-dimensional Brownian motion, the function g described in Section 11.5 had the form

$$g(t) = C \ln t + D$$

    for unknown constants C and D. Use the fact that $g(r) = 0$ and $g(R) = 1$ to solve for C and D, and therefore write g in terms of r and R. (You should get the formula stated in Theorem 11.26.)

21. In the lecture, we saw that for a d-dimensional Brownian motion where $d \geq 3$, the function g described in Section 11.5 had the form

$$g(t) = \frac{C}{2-d} t^{2-d} + D$$

    for unknown constants C and D. Use the fact that $g(r) = 0$ and $g(R) = 1$ to solve for C and D, and therefore write g in terms of r and R. (You should get the formula stated in Theorem 11.26.)

22. Let $\{\mathbf{W}_t\}$ be a standard 2-dimensional BM, starting at the origin.

    a) What is the probability that $\mathbf{W}_t = (0,0)$ for some $t > 0$?
    b) What is the probability that $\mathbf{W}_t \in \{(x,y) : x^2 + y^2 < 1\}$ for some $t > 100$?
    c) Suppose $\mathbf{W}_{15} = (3,4)$. What is the probability that, after time 15, $\mathbf{W}_t$ strikes the circle $\{(x,y) : x^2 + y^2 = 4\}$ before it strikes the circle $\{(x,y) : x^2 + y^2 = 49\}$?

23. Suppose that the position of a particle of pollen suspended in a liquid is modeled by a standard 3-dimensional BM, and that at time 4, the pollen is at position $(1,2,3)$.

a) What is the probability that the pollen particle eventually reaches $(0, 4, -1)$?

b) What is the probability that the pollen particle strikes the sphere of radius 4 centered at the origin before it strikes the sphere of radius 2 centered at the origin?

24. Suppose $\{\mathbf{W}_t\}$ is a standard 5-dimensional BM with

$$\mathbf{W}_0 = (1, 2, 1, -3, 1).$$

What is the probability that $||\mathbf{W}_t|| = 2||\mathbf{W}_0||$ before $||\mathbf{W}_t|| = \frac{1}{2}||\mathbf{W}_0||$?

25. (30 ★ pts) Let $\{W_t\}$ and $\{\widehat{W}_t\}$ be independent, standard BMs and let a be a positive constant.

a) Prove that $P\left(W_t = a\widehat{W}_t \text{ for infinitely many } t\right) = 1$.

b) What is the probability that $W_t = \widehat{W}_t + a$ for infinitely many t? Prove your answer.

26. (30 ★ pts) Let $\{W_t\}, \{\widehat{W}_t\}, \{\widetilde{W}_t\}$ be independent, standard BMs.

a) Is $P\left(W_t = \widehat{W}_t \text{ for infinitely many } t\right) = 1$? Why or why not?

b) Is $P\left(W_t = \widehat{W}_t = \widetilde{W}_t \text{ for infinitely many } t\right) = 1$? Why or why not?

27. (50 ★ pts) Let $\{(X_t, Y_t)\}$ be a standard 2-dimensional Brownian motion. Let $T = \min\{t : X_t = 1\}$. Compute the density function of Y_T, and identify Y_T as a common random variable.

Appendix A

Tables

A.1 Charts of properties of common r.v.s (the "blue sheet")

The next page has a chart listing relevant properties of the common discrete random variables.

The following page has a chart listing relevant properties of the common continuous random variables.

Discrete Distribution X	Density Function $f_X(x)$	$E(X)$	Var(X)	PGF $G_X(t)$ MGF $M_X(t)$
uniform on $\{1, ..., n\}$	$f(x) = \frac{1}{n}$ for $x = 1, 2, ..., n$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$G_X(t) = \frac{t(t^n-1)}{n(t-1)}$ $M_X(t) = \frac{e^t(e^{nt}-1)}{n(e^t-1)}$
binomial(n, p)	$f(x) = \binom{n}{x} p^x(1-p)^{n-x}$ for $x = 0, 1, ..., n$	np	$np(1-p)$	$G_X(t) = (1-p+pt)^n$ $M_X(t) = (1-p+pe^t)^n$
$Geom(p)$ $0 < p < 1$	$f(x) = p(1-p)^x$ for $x = 0, 1, 2, ...$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$	$G_X(t) = \frac{p}{1-(1-p)t}$ $M_X(t) = \frac{p}{1-(1-p)e^t}$
negative binomial $NB(r, p)$	$f(x) = \binom{r+x-1}{x} p^r(1-p)^x$ for $x = 0, 1, 2, ...$	$r\frac{1-p}{p}$	$r\frac{1-p}{p^2}$	$G_X(t) = \left(\frac{p}{1-(1-p)t}\right)^r$ $M_X(t) = \left(\frac{p}{1-(1-p)e^t}\right)^r$
$Pois(\lambda)$	$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ for $x = 0, 1, 2, ...$	λ	λ	$G_X(t) = e^{\lambda(t-1)}$ $M_X(t) = e^{\lambda(e^t-1)}$
hypergeometric $Hyp(n, r, k)$	$f(x) = \frac{\binom{r}{x}\binom{n-r}{k-x}}{\binom{n}{k}}$ for $x = 0, 1, ..., k$	$\frac{kr}{n}$	$\frac{kr}{n}\left(\frac{n-r}{n}\right)\frac{n-k}{n-1}$	not given here
$d-$dimensional hypergeometric with parameters $n, (n_1, ..., n_d), k$	$f(x_1, ..., x_d) = \frac{\binom{n_1}{x_1}\binom{n_2}{x_2}\cdots\binom{n_d}{x_d}}{\binom{n}{k}}$ for $x_1 + x_2 + ... + x_d = k$	N/A	N/A	N/A
multinomial $n, (p_1, ..., p_d)$	$f(x_1, ..., x_d) = \frac{n!}{x_1!x_2!\cdots x_d!}p_1^{x_1}p_2^{x_2}\cdots p_d^{x_d}$ for $x_1 + x_2 + ... + x_d = n$	N/A	N/A	N/A

Continuous Distribution X	Density Function $f_X(x)$ Distribution Function $F_X(x)$	Expected Value EX Variance $Var(X)$ MGF $M_X(t)$
uniform on $[a,b]$	$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{else} \end{cases}$ $F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a,b] \\ 1 & x > b \end{cases}$	$EX = \frac{a+b}{2}$ $Var(X) = \frac{(b-a)^2}{12}$ $M_X(t) = \frac{e^{tb}-e^{ta}}{t(b-a)}$
exponential with parameter $\lambda > 0$	$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$ $F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$	$EX = \frac{1}{\lambda}$ $Var(X) = \frac{1}{\lambda^2}$ $M_X(t) = \frac{\lambda}{\lambda - t}$ for $t < \lambda$
Cauchy	$f(x) = \frac{1}{\pi(1+x^2)}$ $F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan x$	$EX = \infty$ $Var(X)$ DNE $M_X(t)$ DNE
std. normal $n(0,1)$	$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ $F(x) = \Phi(x)$	$EX = 0$ $Var(X) = 1$ $M_X(t) = e^{t^2/2}$
normal $n(\mu, \sigma^2)$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$ $F(x) = \Phi(\frac{x-\mu}{\sigma})$	$EX = \mu$ $Var(X) = \sigma^2$ $M_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$
gamma $\Gamma(r, \lambda)$	$f(x) = \begin{cases} \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$ F_X not given here	$EX = \frac{r}{\lambda}$ $Var(X) = \frac{r}{\lambda^2}$ $M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^r$ for $t < \lambda$
joint normal with mean vector μ; covariance matrix Σ	$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma}} \exp\left[\frac{-1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)\right]$	$E\mathbf{X}$ and $Var(\mathbf{X})$ DNE $M_{\mathbf{X}}(\mathbf{t}) = \exp\left(\mathbf{t} \cdot \mu + \frac{1}{2}\mathbf{t}^T \Sigma \mathbf{t}\right)$

A.2 Useful sum and integral formulas (the "pink sheet")

Triangular Number Formula: For all $n \in \{1, 2, 3, ...\}$,

$$1 + 2 + 3 + ... + n = \sum_{j=0}^{n} j = \frac{n(n+1)}{2}.$$

Finite Geometric Series Formula: for all $r \in \mathbb{R}$,

$$\sum_{n=0}^{N} r^n = \frac{1 - r^{N+1}}{1 - r}.$$

Infinite Geometric Series Formulas: for all $r \in \mathbb{R}$ such that $|r| < 1$,

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r} \qquad \sum_{n=N}^{\infty} r^n = \frac{r^N}{1-r}.$$

Derivative of the Geometric Series Formula: for all $r \in \mathbb{R}$ such that $|r| < 1$,

$$\sum_{n=0}^{\infty} n r^n = \frac{r}{(1-r)^2}.$$

Exponential Series Formula: for all $r \in \mathbb{R}$,

$$\sum_{n=0}^{\infty} \frac{r^n}{n!} = e^r.$$

Binomial Theorem: for all $n \in \mathbb{N}$, and all $x, y \in \mathbb{R}$,

$$\sum_{k=0}^{n} \begin{pmatrix} n \\ k \end{pmatrix} x^k y^{n-k} = (x+y)^n.$$

Vandermonde Identity: for all $n, k, r \in \mathbb{N}$,

$$\sum_{x=0}^{n} \begin{pmatrix} r \\ x \end{pmatrix} \begin{pmatrix} n-r \\ k-x \end{pmatrix} = \begin{pmatrix} n \\ k \end{pmatrix}.$$

Gamma Integral Formula: for all $r > 0$, $\lambda > 0$,

$$\int_0^{\infty} x^{r-1} e^{-\lambda x} \, dx = \frac{\Gamma(r)}{\lambda^r}.$$

Normal Integral Formula: for all $\mu \in \mathbb{R}$ and all $\sigma > 0$,

$$\int_{-\infty}^{\infty} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx = \sigma\sqrt{2\pi}.$$

Beta Integral Formula: for all $r > 0$, $\lambda > 0$,

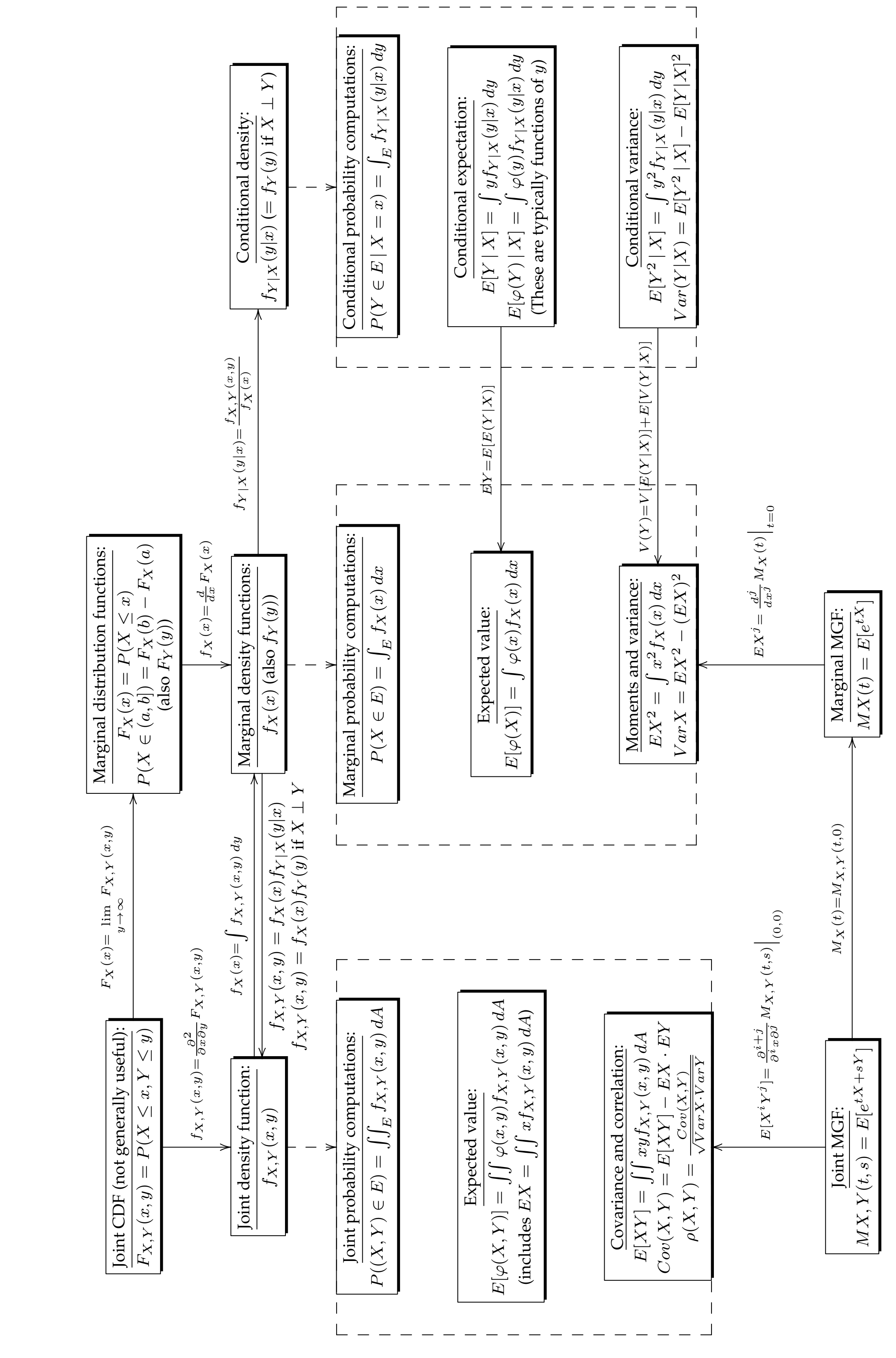
$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} \, dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

A.3 Table of values for the cdf of the standard normal

Entries represent $\Phi(z) = P(n(0,1) \leq z)$. The value of z to the first decimal is in the left column. The second decimal place is given in the top row.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8436	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

A.4 Road map of standard computations with joint distributions



A.5 Facts associated to escape probabilities (the "orange sheet")

PROCESS $\{X_t\}$	SIMPLE RANDOM WALK	BIRTH AND DEATH PROCESS: DISCRETE-TIME	BIRTH AND DEATH PROCESS: CTMC	BROWNIAN MOTION
Important auxiliary quantities	$\mu = p - q$ $\sigma^2 = p + q - (p-q)^2$	$\gamma_j = \frac{q_j q_{j-1} \cdots q_1}{p_j p_{j-1} \cdots p_1}$ $\zeta_j = \frac{p_{j-1} p_{j-2} \cdots p_0}{q_j q_{j-1} \cdots q_1}$ $(\gamma_0 = \zeta_0 = 1)$	$\gamma_j = \frac{\mu_j \mu_{j-1} \cdots \mu_1}{\lambda_j \lambda_{j-1} \cdots \lambda_1}$ $\zeta_j = \frac{\lambda_{j-1} \lambda_{j-2} \cdots \lambda_0}{\mu_j \mu_{j-1} \cdots \mu_1}$ $(\gamma_0 = \zeta_0 = 1)$ $(q_x = \lambda_x + \mu_x)$	
Associated martingale(s)	$\{X_t - \mu t\}$ $\{(X_t - \mu t)^2 - t\sigma^2\}$ $\left\{ \left(\frac{q}{p}\right)^{X_t} \right\}$	$\{\psi(X_t)\}$, where $\psi(0) = 1$ and $\psi(y) = \sum_{j=0}^{y-1} \gamma_j$	N/A	$\{W_t - \mu t\}$ $\{(W_t - \mu t)^2 - t\sigma^2\}$ $\left\{ \exp\left(\frac{-2\mu}{\sigma^2} W_t\right) \right\}$
$P_x(T_a < T_b)$ ($P_x(T_b < T_a)$ is 1− this)	$\frac{b-x}{b-a}$ if unbiased; $\frac{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^x}{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^a}$ if biased	$\frac{\psi(b) - \psi(x)}{\psi(b) - \psi(a)}$	N/A	$\frac{b-x}{b-a}$ if $\mu = 0$; $\frac{\exp\left(\frac{-2\mu b}{\sigma^2}\right) - \exp\left(\frac{-2\mu x}{\sigma^2}\right)}{\exp\left(\frac{-2\mu b}{\sigma^2}\right) - \exp\left(\frac{-2\mu a}{\sigma^2}\right)}$ if $\mu \neq 0$
$E_x[T_{\{a,b\}}]$	$\frac{(b-x)(x-a)}{p+q}$ if unbiased; if biased, solve for this using $EX_T = x + \mu ET$	N/A		$\frac{(x-a)(b-x)}{\sigma^2}$ if $\mu = 0$; solve using $EW_T = x + \mu ET$ if $\mu \neq 0$
$f_{x,y}$ (for $x \neq y$)	1 if walk is unbiased or tends toward y; $\left(\frac{\min\{p,q\}}{\max\{p,q\}}\right)^{\lvert x-y \rvert}$ if walk tends away from y	N/A		1 if $\mu = 0$; $\exp\left(\frac{-2\mu \lvert x-y \rvert}{\sigma^2}\right)$ if $\mu \neq 0$
f_x	$r + 2\min(p,q)$	$f_0 = 1 - \left[\sum_y \gamma_y\right]^{-1}$		1 if $\mu = 0$; $\frac{1}{2} + \Phi\left(\frac{-\mu}{\sigma}\right)$ if $\mu \neq 0$
Recurrence/ transience	null recurrent $\Leftrightarrow$ unbiased transient $\Leftrightarrow$ biased	recurrent $\Leftrightarrow \sum_y \gamma_y = \infty$ positive recurrent $\Leftrightarrow \sum_y \zeta_y < \infty$		null recurrent $\Leftrightarrow \mu = 0$ transient $\Leftrightarrow \mu \neq 0$
stationary distribution (if positive recurrent)	N/A	$\pi(x) = \frac{\zeta_x}{\sum_y \zeta_y}$		N/A