

Probability Notes

Justin Burruss

2026-09-12

Contents

1	Background	2
2	Basics	2
2.1	Interpretations	2
2.2	Set Theory and Probability Spaces	3
2.3	Basic Theorems	8
2.4	Conditional Probability	13
2.5	Independence	15
2.6	Worked Problems	16
3	Random Variables	17
3.1	Discrete Random Variables	17
3.2	Probability Mass Function (PMF)	18
3.3	Cumulative Distribution Function (CDF)	19
3.4	Continuous Random Variables	20
3.5	Probability Density Function (PDF)	23
3.6	Expectation	26
3.7	Conditional Expectation	29
3.8	Conditional PMFs and PDFs	29
3.9	Variance and Standard Deviation	30
4	Families of Random Variables	32
4.1	Bernoulli (Discrete)	35
4.2	Beta (Continuous)	36
4.3	Binomial (Discrete)	38
4.4	Cauchy (Continuous)	38
4.5	Discrete Uniform (Discrete)	39
4.6	Erlang (Continuous)	41

4.7	Exponential (Continuous)	41
4.8	Gamma (Continuous)	42
4.9	Gaussian (Continuous)	43
4.10	Geometric (Discrete)	43
4.11	Laplace (Continuous)	43
4.12	Multinomial (Discrete)	43
4.13	Poisson (Discrete)	44
4.14	Uniform (Continuous)	45
5	Multiple Random Variables	45
5.1	Definitions and Notation	46
5.2	Worked Problems	48
6	Stochastic Processes	54
6.1	Definitions and Notation	54
6.2	Poisson Processes	55
7	Markov Chains	57
7.1	Discrete-Time Markov Chains	58
7.2	Structure and Properties	65
8	Source Notes	71
8.1	Notation	71
8.2	Quotes	71

1 Background

These are general notes for probability. The document was created in RMarkdown with R plus a little $\text{\LaTeX}$. *TikZ* (“*TikZ* ist kein Zeichenprogramm”) was used for the diagrams.

2 Basics

2.1 Interpretations

When working through probability theory it can be useful to motivate definitions and theorems by appealing to intuitions about what probability “really means”. With probability there’s no single accepted interpretation, though. Like other useful mathematical abstractions, the mathematical abstraction that is probability gets applied in many different contexts. This applicability leads to a number of interpretations of what probability means. Interpretations include:

- as a generalization of deductive logic

- extends by allowing degrees of implication between false (0) and true (1)
- examine the probability space and evidence logically
- as a degree of belief
 - probabilities quantify what may be subjective (but rational) degrees of belief
 - update *a priori* beliefs given new evidence
- as a relative frequency
 - probabilities are objective properties to be estimated
 - frequency in the long-run
- as a propensity or disposition
 - also an objective property
 - effects from causes (propensity or tendency)
- “classical” approach of reducing to equally likely cases
 - may use principle of indifference

Nonetheless, to quote from *Introduction to Probability Theory* by Hoel, Port, and Stone:

For the mathematical theory of probability the interpretation of probabilities is irrelevant, just as in geometry the interpretation of points, lines, and planes is irrelevant.

What is broadly accepted is that mathematical probability ought to be developed from axioms of probability independent of any specific interpretation. Quoting A. N. Kolmogorov:

The theory of probability, as a mathematical discipline, can and should be developed from axioms in exactly the same way as Geometry and Algebra.

As Casella and Berger put it in the second edition of *Statistical Inference*:

...the axiomatic approach is not concerned with the interpretations of probabilities, but is concerned only that the probabilities are defined by a function satisfying the axioms.

Probability theory starts with using the theory of sets to construct probability spaces.

2.2 Set Theory and Probability Spaces

Quoting from *Probability and Stochastic Processes* by Yates and Goodman:

The mathematical basis of probability is the theory of sets.

2.2.1 Outcomes and the Sample Space

We can define an **outcome** (or **elementary event**) as any possible “thing that happens” in a given trial, i.e., any possible observation of an experiment. We then define a universal set of all the possible outcomes named the **sample space**. Yates and Goodman give a nice definition in the second edition of *Probability and Stochastic Processes*:

The sample space of an experiment is the finest-grain, mutually exclusive, collectively exhaustive set of all possible outcomes.

By “finest-grain, mutually exclusive, collectively exhaustive” we mean:

- finest-grain as in all possible distinguishable outcomes are identified separately
- mutually exclusive in that if one outcome occurs, no other outcome occurs
- collectively exhaustive in that every outcome is included

Authors typically denote the sample space with a capital letter, e.g., Ω (Greek letter capital Omega), E (as in “elements”), or S (as in “samples” or “sample points”) and often use the corresponding lowercase for individual outcomes, e.g., ω_i for an outcome in Ω .

Thus we put the “things that happen” into 1:1 correspondence with elements of our sample space.

By using sets to set up probability, we inherit set notation:

- Brackets $\{\}$ indicate a set
 - e.g., read $\{x \mid 0 \leq x \leq 1\}$ as “the set of x such that x is between 0 and 1 inclusive”
 - sometimes using $:$ instead of $\mid$, e.g., $\{x : 0 \leq x \leq 1\}$
 - either $\{\}$ or $\emptyset$ to indicate the empty set
- $x \in A$ indicates x is in A
 - can be read as “is an element” or “is a member of”
- $x \notin A$ means x is not in A
- $A \subseteq B$ indicates A is a subset of B (possibly equal subset)
 - subset as in $A \subseteq B \iff x \in A \implies x \in B$
 - some sources may use $\subset$ for possibly equal subset
 - will use $\subset$ for proper subset in these notes
 - reflexive: $A \subseteq A$
- $A = B$ indicates A equals B
 - equals as in $A = B \iff A \subseteq B$ and $B \subseteq A$
- $A \cup B$ denotes A union B
 - union operating like a logical “or”, i.e., $A \cup B = \{x \mid (x \in A) \vee (x \in B)\}$
 - idempotent: $A \cup A = A$
 - commutative: $A \cup B = B \cup A$
 - associative: $A \cup (B \cup C) = (A \cup B) \cup C$
- $A \cap B$ denotes A intersect B
 - intersection corresponding to logical “and”, i.e., $A \cap B = \{x \mid (x \in A) \wedge (x \in B)\}$
 - idempotent: $A \cap A = A$
 - commutative: $A \cap B = B \cap A$
 - associative: $A \cap (B \cap C) = (A \cap B) \cap C$
- $A \setminus B$ indicates A minus B , i.e., the difference
 - $A \setminus B = \{x \mid x \in A \text{ and } x \notin B\}$
 - $A \setminus A = \emptyset$
 - $A \setminus \emptyset = A$
 - $\emptyset \setminus A = \emptyset$
- A^c indicates the complement of A , i.e., $A^c = \Omega \setminus A$

- some authors use $\overline{A}$ to indicate A^c

The $\cap$ and $\cup$ operations have the distributive property as well, e.g., we can distribute the $A \cup$ of $A \cup (B \cap C)$ across the other terms as $(A \cup B) \cap (A \cup C)$.

The set theoretic terms “disjoint” and “partition” come up in the context of probability.

Disjoint A and B are disjoint if $A \cap B = \emptyset$.

Partition A partition of a set A is a set containing one or more empty subsets of A such that every element of A is in exactly one of these subsets. For example, if $\bigcup_{i=1}^{\infty} A_i = \Omega$ and $A_1, A_2, \dots$ are all pairwise disjoint, then the set $A_1, A_2, \dots$ forms a partition of Ω .

There are other basic set theory identities and “laws”. One example are de Morgan’s laws:

- $(A \cup B)^c = A^c \cap B^c$ “the complement of the union is the intersection of the complements”
- $(A \cap B)^c = A^c \cup B^c$ “the complement of the intersection is the union of the complements”

Note that the same procedure can yield different samples spaces depending on what it is we’re observing. Say the procedure is to flip a coin twice. If our observations are to simply count heads, then our sample space $\Omega = \{0, 1, 2\}$, i.e., just the three outcomes. If on the other hand we’re observing heads or tails (including the order) then our sample space $\Omega = \{HH, HT, TH, TT\}$ and we have four outcomes.

2.2.2 Events and Event Spaces

After defining what we mean by outcomes, our next move is to create a structure against which we will assign probabilities. In other words, we will separate the “things that happen” from the “statements we can evaluate about those things”. To that end we define an **event** (or **random event**) to be a set of outcomes of an experiment, i.e., some subset of Ω . We collect events together into an **event space**. Again following Yates and Goodman:

An **event space** is a collectively exhaustive, mutually exclusive set of events.

Popular notation choices for an event space are capital letters such as $\mathfrak{F}$ or $\mathcal{F}$ (as in “field” in the mathematical sense), $\mathcal{B}$ (as in “Borel field”), $\mathcal{E}$ (as in “event”), or $\mathcal{A}$. We’ll use $\mathcal{F}$ as the default label for an event space. Sources may define $\mathcal{F}$ as a sigma-field (also called sigma-algebra or Borel field, or in German-language sources as Borelsche Körper or Borelsche σ -Algebra), this indicating that:

- $\mathcal{F}$ is a field of subsets of Ω
- the empty set is a member of the field $\mathcal{F}$, i.e., $\emptyset \in \mathcal{F}$
- $\mathcal{F}$ is closed under complements, i.e., $A \in \mathcal{F} \implies A^c \in \mathcal{F}$
- $\mathcal{F}$ is closed under countable unions

Having these properties makes $\mathcal{F}$ closed under countable intersections as well.

Notice that we had “an” event space, not “the” event space. We have options. When choosing $\mathcal{F}$ a sensible approach is to go with the smallest set that satisfies. So for example if the procedure is to flip a coin twice and our observations are to simply count heads, thus yielding $\Omega = \{0, 1, 2\}$, what makes sense as the small set that satisfies? Arguably it makes the most sense to use $\mathcal{F} = \mathcal{P}(\Omega)$ (the power set of Ω), which would be $\mathcal{P}(\Omega) = \{\emptyset, \{0\}, \{1\}, \{2\}, \{0, 1\}, \{0, 2\}, \{1, 2\}, \{0, 1, 2\}\}$.

We do *not* require that every subset of Ω be in $\mathcal{F}$. Here’s an example adapted from *A First Course in Probability and Markov Chains* by Modica and Poggiolini. Say we have as a sample space five playing cards $\Omega = \{A\spadesuit, K\spadesuit, Q\spadesuit, J\heartsuit, J\diamondsuit\}$. It is completely valid to define our $\mathcal{F}$ as the algebra generated by just the subsets $A = \{A\spadesuit, K\spadesuit\}$ and $B = \{K\spadesuit, Q\spadesuit\}$. Applying our requirements gives us:

- $\emptyset \in \mathcal{F}$
- $\Omega \in \mathcal{F}$
- $A \in \mathcal{F}$
- $B \in \mathcal{F}$
- $A^c \in \mathcal{F}$
- $B^c \in \mathcal{F}$
- $A \cap B \in \mathcal{F}$
- $A \cap B^c \in \mathcal{F}$
- $A^c \cap B \in \mathcal{F}$
- all the finite unions of the above

Thus we get $\{A\spadesuit\} = A \cap B^c$, $\{K\spadesuit\} = A \cap B$, and so forth. Adding these up, we get just 16 events in $\mathcal{F}$, half the number we would have had with the power set of Ω :

1. empty set $\emptyset$
2. singleton $\{A\spadesuit\}$
3. singleton $\{K\spadesuit\}$
4. singleton $\{Q\spadesuit\}$
5. pair $\{A\spadesuit, K\spadesuit\}$
6. pair $\{A\spadesuit, Q\spadesuit\}$
7. pair $\{K\spadesuit, Q\spadesuit\}$
8. pair $\{J\heartsuit, J\diamondsuit\}$
9. trio $\{A\spadesuit, K\spadesuit, Q\spadesuit\}$
10. trio $\{A\spadesuit, J\heartsuit, J\diamondsuit\}$
11. trio $\{K\spadesuit, J\heartsuit, J\diamondsuit\}$
12. trio $\{Q\spadesuit, J\heartsuit, J\diamondsuit\}$
13. quartet $\{A\spadesuit, K\spadesuit, J\heartsuit, J\diamondsuit\}$
14. quartet $\{A\spadesuit, Q\spadesuit, J\heartsuit, J\diamondsuit\}$
15. quartet $\{K\spadesuit, Q\spadesuit, J\heartsuit, J\diamondsuit\}$
16. sample space $\{A\spadesuit, K\spadesuit, Q\spadesuit, J\heartsuit, J\diamondsuit\}$

This $\mathcal{F}$ leaves out 16 possibilities—for example the singletons $\{J\heartsuit\}$ and $\{J\diamondsuit\}$ —but is a completely valid event space as it is closed under complements, countable unions, etc.

We can accommodate countably infinite outcomes. For example, if our procedure was “count the number of flips of a coin until you get heads”, then $\Omega = \{0, 1, 2, \dots\}$ (often denoted as $\mathbb{N}$ or $\mathbb{N}_0$), so we would let $\mathcal{F} = \mathcal{P}(\Omega)$.

What if we have uncountably infinite outcomes? In that case we choose $\mathcal{F}$ to contain any outcome of interest. If for instance outcomes are real numbers, then we can set up $\mathcal{F}$ to contain all sets of the form $[a, b]$, $(a, b]$, $[a, b)$, (a, b) (plus their complements and countable unions) for $a, b \in \mathbb{R}$, or perhaps only the half-open intervals $[a, b)$ or only the open intervals (a, b) . Intervals like $[-\infty, a)$ can be useful for cumulative distribution functions.

Some authors consider $\mathcal{F}$ to be a class as opposed to a set where by “class” we mean a more general notion from the foundations of set theory.

Upshot is we have many possible event spaces. For these notes, when not specifying otherwise, for finite and countably infinite let’s assume we’re using $\mathcal{P}(\Omega)$ and for uncountably infinite we choose a $\mathcal{F}$ that satisfies the sigma-field requirements.

One other detail—to the standard set notation we also add:

- AB indicates A and B both occurring, i.e., AB is short for $A \cap B$

Many authors use this notation, going at least as far back as Kolmogorov’s 1933 work, where he indicates he follows the 1927 text *Mengenlehre* by Felix Hausdorff—*Mengenlehre* does indeed use AB to indicate *die Menge der Elemente, die zugleich zu A und zu B gehören*, and you can look even further back to Hausdorff’s 1914 *Grundzüge der Mengenlehre* where he uses AB to indicate *das Produkt* of A and B . For these notes we will use both AB and $A \cap B$.

2.2.3 Probability Measure

Next up is to bring in probabilities. To each set A in $\mathcal{F}$ is assigned a non-negative real number $P[A]$ (sometimes written $P(A)$, $\mathbb{P}(A)$, or $Pr(A)$) such that $P[\Omega] = 1$, i.e., $P[\cdot]$ is a function that maps events to non-negative real numbers such that the total over all outcomes comes out to 1. We read $P[A]$ as “the probability of A ”.

Together, the sample space Ω , event space $\mathcal{F}$, and probability measure P give us a model for working with random processes. Many sources group these together into a triplet $(\Omega, \mathcal{F}, P)$.

2.2.4 Countable Additivity

Then we need just one more rule to define how probabilities interact. One common choice is countable additivity (sometimes called sigma-additivity): for any countable collection $A_1, A_2, \dots$ of mutually exclusive events, the probability of their union is the same as the sum of their probabilities, i.e.

$$P[A_1 \cup A_2 \cup \dots] = P[A_1] + P[A_2] + \dots, \quad \text{for pairwise disjoint } A_1, A_2, \dots$$

Countable additivity is the typical but not the only choice.

2.2.5 Alternatives

Alternatively, some sources go the finite route and use $P[A_1 \cup A_2] = P[A_1] + P[A_2]$, $A_1 \cap A_2 = \emptyset$ as a third axiom. In instances where we have a finite sample space then really all we need is finite additivity, so arguably in those cases we ought to be parsimonious and avoid the other route having the infinity.

Another interesting special case is when each outcome in Ω is equally likely, meaning we have:

$$P[\omega_i] = \frac{1}{n}, \quad 1 \leq i \leq n, \quad \Omega = \{\omega_1, \omega_2, \dots, \omega_n\} \text{ all equally likely outcomes}$$

This configuration gives us a uniform probability space. Much of the classic work in probability involved uniform probabilities. Many games of chance involving dice rolls or the like fit nicely under a uniform probability space. Another way to express this is:

$$P[A] = \frac{|A|}{|\Omega|}, \quad \forall A \subseteq \Omega$$

Where $|A|$ indicates the cardinality of set A and $\forall$ means “for all”.

This gives us a uniform distribution on a finite set and is sufficient for many games of chance involving die rolls and the like. For these notes, though, we’ll stick with countable additivity.

2.2.6 Axioms of Probability

Putting this all together, if we’re going the countable additivity route we get these three axioms:

1. $P[A] \geq 0$, for any event $A \in \mathcal{F}$

2. $P[\Omega] = 1$
3. (countable additivity) $P[\bigcup_{i=1}^{\infty} A_i] = \sum_{i=1}^{\infty} P[A_i]$ for pairwise disjoint $A_1, A_2, \dots$

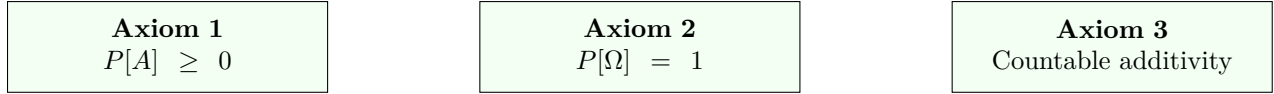


Figure 1: Axioms of Probability

This leads us into a number of theorems.

2.3 Basic Theorems

2.3.1 Probability of the empty set

First off, we can use axioms 1 and 3 to identify the probability of the empty set.

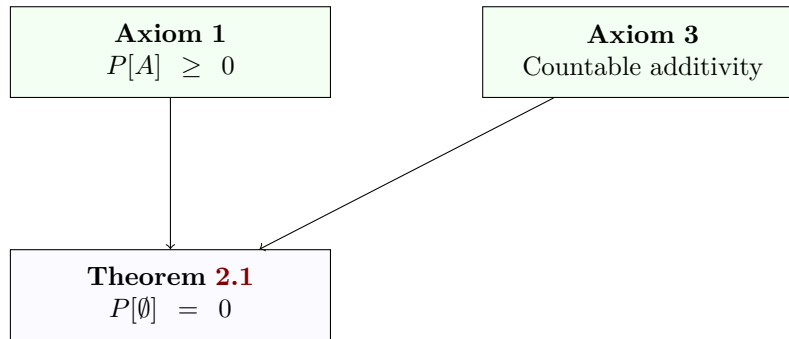


Figure 2: Probability of the empty set

Theorem 2.1. Probability of the empty set

$$P[\emptyset] = 0$$

Proof. Let A_1 be some arbitrary event in $\mathcal{F}$, and for $n = 2, 3, \dots$ let $A_n = \emptyset$ ($\emptyset$ is also in $\mathcal{F}$). In set theory, $A \cup \emptyset = A$, $A \cap \emptyset = \emptyset$, and $\emptyset \cap \emptyset = \emptyset$, hence the events $A_1, A_2, \dots$ are pairwise disjoint, and we can invoke the

countable additivity axiom to write:

$$\begin{aligned}
P[A_1] &= P\left[\bigcup_{i=1}^{\infty} A_i\right] \\
&= \sum_{i=1}^{\infty} P[A_i] && \text{by the countable additivity axiom} \\
&= P[A_1] + P[A_2] + \sum_{i=3}^{\infty} P[A_i] && \text{peeling away first two terms of the sum} \\
P[A_1] - P[A_1] &= P[A_1] + P[A_2] + \sum_{i=3}^{\infty} P[A_i] - P[A_1] && \text{subtract } P[A_1] \text{ from both sides} \\
0 &= P[A_2] + \sum_{i=3}^{\infty} P[A_i] \\
&= P[\emptyset] + \sum_{i=3}^{\infty} P[A_i] && A_2 = \emptyset
\end{aligned}$$

Then we put our first axiom to work. $P[A] \geq 0$ for any event $A \in \mathcal{F}$, so the terms on the right-hand side above are non-negative, hence the right-hand side is a sum of non-negative real numbers equal to 0. This is only possible if each term is 0. Therefore the first term on the right $P[\emptyset] = 0$. $\square$

Another approach would be to work in the other direction by proving $P[A^c] = 1 - P[A]$ using $P[\Omega] = 1$ (axiom 1), then applying axiom 3.

2.3.2 Finite additivity

Next theorem—since $P[\emptyset] = 0$ we can plug into axiom 3 a special case for just mutually exclusive sets A_1 and A_2 in $\mathcal{F}$ to find that:

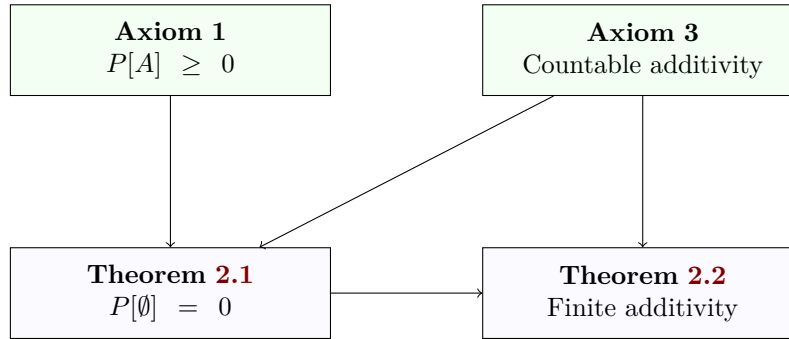


Figure 3: Finite additivity for two disjoint events

Theorem 2.2. Finite additivity for two disjoint events

If $A_1 \cap A_2 = \emptyset$, $A_1, A_2 \in \mathcal{F}$, then

$$P[A_1 \cup A_2] = P[A_1] + P[A_2]$$

Proof. For $A_1, A_2, \dots \in \mathcal{F}$ let A_1 and A_2 be arbitrary disjoint events and A_n be $\emptyset$ for all $n > 2$.

$$\begin{aligned}
P[A_1 \cup A_2] &= P[A_1 \cup A_2 \cup \dots] \\
&= \sum_{i=1}^{\infty} P[A_i] && \text{countable additivity (Axiom 3) since } A_1, A_2, \dots \text{ pairwise disjoint} \\
&= P[A_1] + P[A_2] + \sum_{i=3}^{\infty} P[A_i] \\
&= P[A_1] + P[A_2] + \sum_{i=3}^{\infty} P[\emptyset] \\
&= P[A_1] + P[A_2] + \sum_{i=3}^{\infty} 0 && P[\emptyset] = 0 \text{ (Theorem 2.1)} \\
&= P[A_1] + P[A_2]
\end{aligned}$$

□

More generally, for $A \in \mathcal{F}$ where $A = A_1 \cup A_2 \cup \dots \cup A_n$ when $A_i \cap A_j = \emptyset$, $i \neq j$ we get:

Theorem 2.3. Finite additivity (n disjoint events)

Let

$$A = \bigcup_{i=1}^n A_i$$

where the events $A_1, \dots, A_n \in \mathcal{F}$ are pairwise disjoint. Then

$$P[A] = \sum_{i=1}^n P[A_i]$$

In other words, the probability of the union of disjoint events is one and the same as the sum of the probabilities of the disjoint events. A good way to prove this would be via induction on Theorem 2.2.

Proof. The case $n = 2$ is Theorem 2.2. To set up our induction, suppose we have a set of $m = n + 1$ pairwise disjoint events $\{A_1, \dots, A_{n+1}\}$. Then $\{A_1, \dots, A_n\}$ is also a set of pairwise disjoint events, and $\{A_1, \dots, A_n\} \cap A_{n+1} = \emptyset$, so applying Theorem 2.2 we get $P[\{A_1, \dots, A_n\} \cup A_{n+1}] = P[\{A_1, \dots, A_n\}] + P[A_{n+1}]$. Now define $A = \bigcup_{i=1}^m A_i$ and we get the form

$$\begin{aligned}
P[A] &= \sum_{i=1}^{m-1} P[A_i] + P[A_m] && P[\{A_1, \dots, A_n\} \cup A_{n+1}] = P[\{A_1, \dots, A_n\}] + P[A_{n+1}] \\
&= \sum_{i=1}^m P[A_i]
\end{aligned}$$

□

2.3.3 Probability of the complement of an event

From the axioms and finite additivity we can identify the probability of the complement of an event.

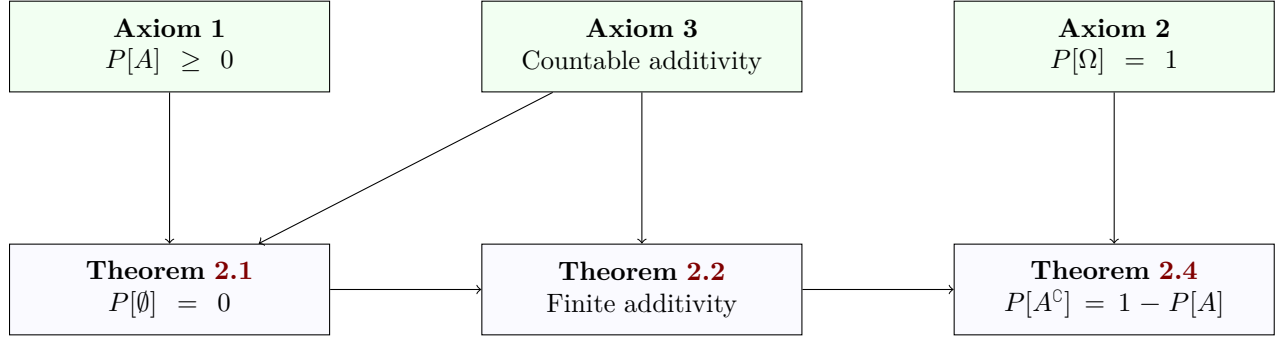


Figure 4: Probability of the complement of an event

Theorem 2.4. Probability of the complement of an event

$$P[A^c] = 1 - P[A]$$

Proof. Let $A \in \mathcal{F}$ be some arbitrary event.

$$\begin{aligned}
 P[A \cup A^c] &= P[\Omega] \\
 &= 1 && \text{axiom 2} \\
 P[A] + P[A^c] &= 1 && \text{since } A \text{ and } A^c \text{ are disjoint (Theorem 2.2)} \\
 P[A] - P[A] + P[A^c] &= 1 - P[A] \\
 P[A^c] &= 1 - P[A]
 \end{aligned}$$

□

2.3.4 Probability of the union of two events

We saw finite additivity already for the case where we have disjoint events. What about more generally for any $A_1, A_2 \in \mathcal{F}$ where the events are not necessarily disjoint? From set theory we inherit (or ought to inherit) the principle of inclusion-exclusion.

Theorem 2.5. Probability of the union of two events

For $A_1, A_2 \in \mathcal{F}$:

$$P[A_1 \cup A_2] = P[A_1] + P[A_2] - P[A_1 \cap A_2]$$

Rather than “inheriting without proof” maybe it’s best to show the reasoning behind the principle of inclusion-exclusion.

Proof.

$$\begin{aligned}
 P[A_1 \cup A_2] &= P[A_1 \setminus A_2] + P[A_1 \cap A_2] + P[A_2 \setminus A_1] \\
 &= P[A_1] - P[A_1 \cap A_2] + P[A_1 \cap A_2] + P[A_2 \setminus A_1] && P[A_1 \setminus B] = P[A_1] - P[A_1 \cap B] \\
 &= P[A_1] + P[A_2 \setminus A_1] && P[A_1 \cap A_2] \text{ cancellation} \\
 &= P[A_1] + P[A_2] - P[A_1 \cap A_2] && P[A_2 \setminus A_1] = P[A_2] - P[A_1 \cap A_2]
 \end{aligned}$$

□

2.3.5 Complete probability

Another basic item. Say we have some set B that partitions the sample space Ω , i.e., $\Omega = \bigcup_{i=1}^n B_i$. Then by applying finite additivity we can know the probability of the sum of intersections of $A \in \mathcal{F}$ with the elements of B .

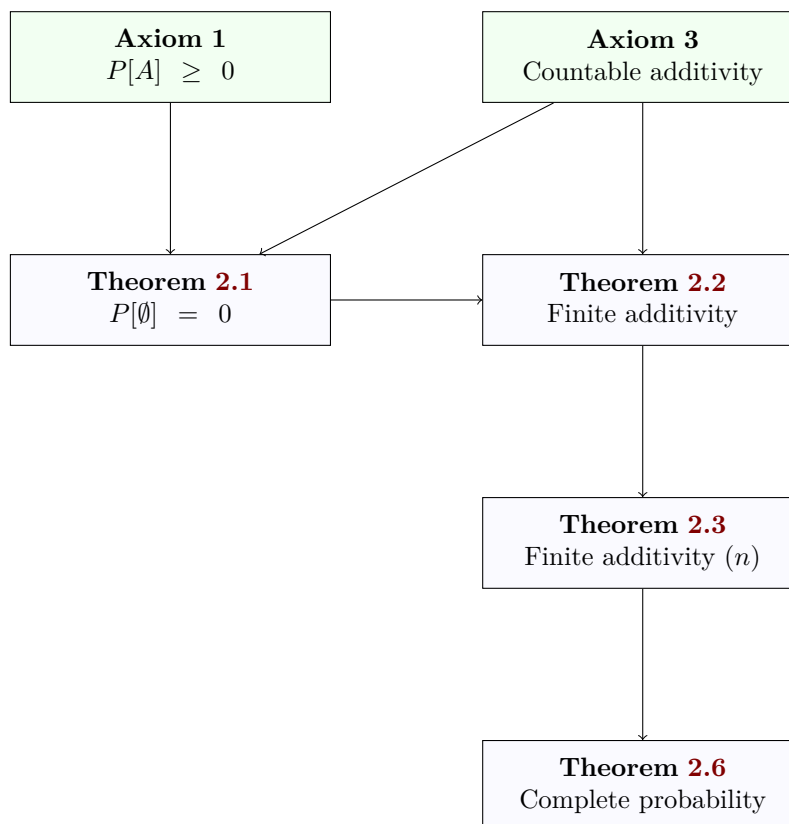


Figure 5: Complete probability

Theorem 2.6. Complete probability

For a partition of the sample space $\{B_1, B_2, \dots, B_n\}$ and any $A \in \mathcal{F}$:

$$P[A] = \sum_{i=1}^n P[A \cap B_i]$$

Proof.

$$\begin{aligned}
 A &= A \cap \Omega \\
 &= A \cap \bigcup_{i=1}^n B_i \\
 &= \bigcup_{i=1}^n (A \cap B_i)
 \end{aligned}$$

distributing

Then seeing as how each $A \cap B_i$ is pairwise disjoint, we apply Theorem 2.3 ($P[A] = \sum_{i=1}^n P[A_i]$) to get $P[A] = \sum_{i=1}^n P[A \cap B_i]$. $\square$

Kolmogorov named this *Satz über die vollständige Wahrscheinlichkeit* (Theorem about the complete probability) in his original 1933 text with the subsequent English translation naming it as “The Theorem on Total Probability”. Sometimes this is known as the “Law of Total Probability”, whereas other sources reserve that label for a theorem that uses the definition of conditional probability. Let’s follow Kolmogorov and use “complete probability” for this one and reserve “Law of Total Probability” for the theorem that relies on the next topic to cover: conditional probability.

2.4 Conditional Probability

We use the notation $P[A|B]$ to indicate conditional probability (read $P[A|B]$ as “the probability of A given B ”). The conditional probability of event A given the occurrence of event B is defined as:

$$P[A|B] = \frac{P[AB]}{P[B]}, \quad P[B] > 0$$

The $|$ symbol calls to mind the “such that” of set notation. An alternative notation is $P_B(A)$ to indicate the conditional probability of event A under the condition B .

As noted earlier, the notation AB means $A \cap B$.

Observe that this $P[A|B]$ was only defined when $P[B] > 0$. If $P[B] > 0$ it means B never occurs, making it illogical to discuss $P[A|B]$. We only care about $P[A|B]$ when it is a valid probability measure, i.e., relative to the sample space consisting of outcomes in B . Note also that $P[A|B]$ follows all three axioms of probability provided $P[B] > 0$, e.g., $P[A] \geq 0$ (Axiom 1) leads us to $P[A|B] \geq 0$.

One example of conditional probability to demonstrate. Consider the procedure that we draw one card from the standard 52-card deck. Our sample space $\Omega = \{A\spadesuit, 2\spadesuit, 3\spadesuit, \dots, K\clubsuit\}$, which is to say the 52 possible 1-card draws from the 13 faces and four suits $\spadesuit, \heartsuit, \diamondsuit, \clubsuit$. Let event $A = \{K\diamondsuit\}$ and event $B = \{\text{any card of the red suits}\}$. We have $P[A] = \frac{1}{52}$ and $P[B] = \frac{26}{52} = \frac{1}{2}$. We find $P[A|B] = \frac{P[AB]}{P[B]} = \frac{\frac{1}{52}}{\frac{1}{2}} = \frac{1}{26}$. Likewise, let $C = \{K\spadesuit\}$, then $P[C|B] = \frac{P[CB]}{P[B]} = \frac{0}{\frac{1}{2}} = 0$.

Once we’ve introduced the definition of conditional probability we quickly find another version of Theorem 2.6 that often goes by the name The Law of Total Probability.

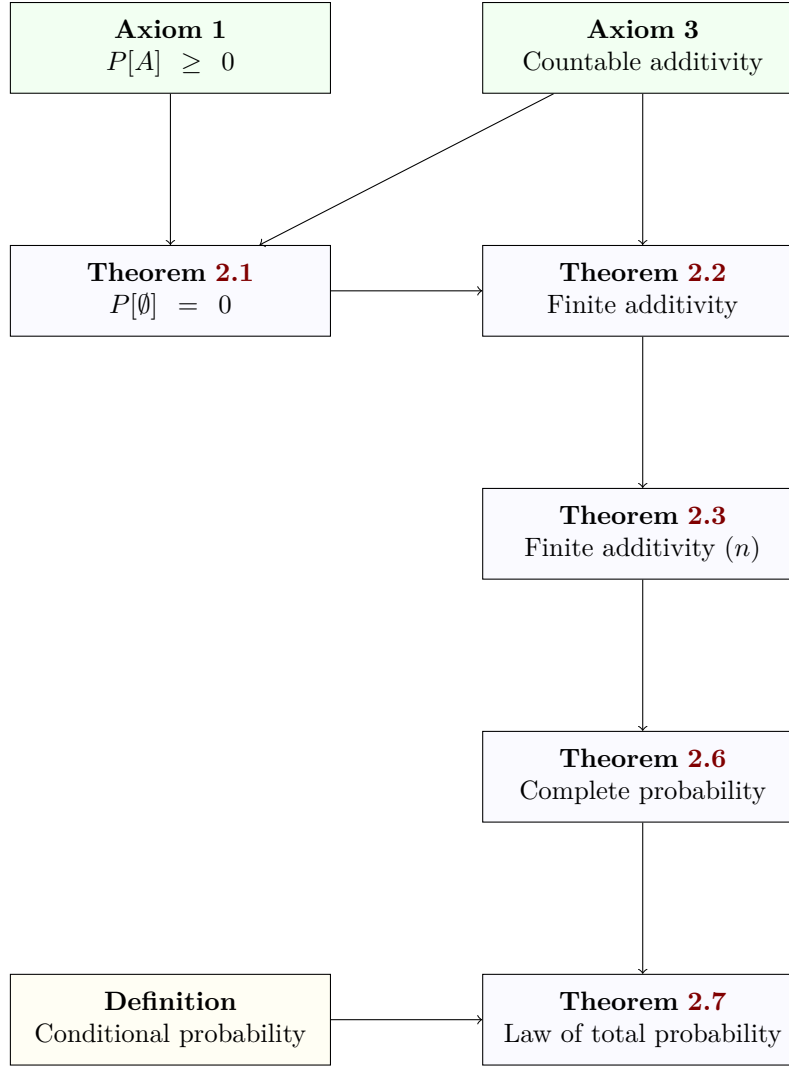


Figure 6: Law of total probability

Theorem 2.7. Law of total probability

For set B that partitions the sample space Ω , i.e., $\Omega = \bigcup_{i=1}^n B_i$ with $P[B_i] > 0$ for all i and arbitrary event $A \in \mathcal{F}$:

$$P[A] = \sum_{i=1}^n P[A|B_i]P[B_i]$$

Proof.

$$\begin{aligned}
P[A|B] &= \frac{P[AB]}{P[B]} \\
P[A|B]P[B] &= \frac{P[AB]}{P[B]}P[B] && \text{multiply both sides by } P[B] \\
P[A|B]P[B] &= P[AB] && P[B] \text{ cancellation} \\
\sum_{i=1}^n P[A|B_i]P[B_i] &= \sum_{i=1}^n P[A \cap B_i] && \text{writing out the sums} \\
&= P[A] && \text{Theorem 2.6}
\end{aligned}$$

□

Next, we do a little algebra on the definition of conditional probability to get Bayes' Theorem.

Theorem 2.8. Bayes' Theorem

For $A, B \in \mathcal{F}$ when $P[A] > 0$ and $P[B] > 0$:

$$P[B|A] = \frac{P[A|B]P[B]}{P[A]}$$

Proof. Starting with the definition of conditional probability:

$$\begin{aligned}
P[A|B] &= \frac{P[AB]}{P[B]} && P[B] > 0 \\
P[A|B]P[B] &= \frac{P[AB]}{P[B]}P[B] && \text{multiply both sides by } P[B] \\
P[A|B]P[B] &= P[AB] && P[B] \text{ cancellation}
\end{aligned}$$

Substituting into the definition of conditional probability:

$$\begin{aligned}
P[B|A] &= \frac{P[AB]}{P[A]} && P[A] > 0 \\
&= \frac{P[A|B]P[B]}{P[A]} && P[AB] = P[A|B]P[B]
\end{aligned}$$

□

This can be applicable for example when we want to reason about causes having observed effects, i.e., we have $P[A|B]$ but want it in terms of $P[B|A]$.

If $P[A|B] = P[A]$ and $P[B|A] = P[B]$ then we see that A and B are not really dependent on each other, a property defined as independence.

2.5 Independence

Events A and B are independent if and only if:

$$P[AB] = P[A]P[B]$$

Independence and disjointedness are not the same. If A and B are disjoint, then $P[AB] = 0$, whereas if A and B are independent $P[AB] = P[A]P[B]$.

Some authors will distinguish between “independence” and “mutual independence”, for example Kolmogorov (*gegenseitigen unabhängigkeit* and *gegenseitig unabhängig* vs. just *unabhängigkeit* and *unabhängig*), but for these notes we will always mean “mutual independence” when we write “independence”. The distinction is necessary when we have more than one event. Three events $A_1, A_2, A_3 \in \mathcal{F}$ are independent if and only if:

1. $P[A_1 \cap A_2] = P[A_1]P[A_2]$
2. $P[A_1 \cap A_3] = P[A_1]P[A_3]$
3. $P[A_2 \cap A_3] = P[A_2]P[A_3]$
4. $P[A_1 \cap A_2 \cap A_3] = P[A_1]P[A_2]P[A_3]$

We can't simply check $P[A_1 \cap A_2 \cap A_3] = P[A_1]P[A_2]P[A_3]$ and be done with it. In other words, we need “mutual independence”. Yates and Goodman give a good illustrative example. Say $\Omega = \{1, 2, 3, 4\}$ and we have a uniform probability space where $P[\omega] = \frac{1}{4}$ for all $\omega \in \Omega$. Are the events $A_1 = \{1, 3, 4\}$, $A_2 = \{2, 3, 4\}$, $A_3 = \emptyset$ independent?

- $P[A_1] = P[\{1, 3, 4\}] = \frac{3}{4}$
- $P[A_2] = P[\{2, 3, 4\}] = \frac{3}{4}$
- $P[A_3] = P[\emptyset] = 0$
- $P[A_1]P[A_2]P[A_3] = (\frac{3}{4})(\frac{3}{4})(0) = 0$
- $P[A_1 \cap A_2 \cap A_3] = P[\{1, 3, 4\} \cap \{2, 3, 4\} \cap \emptyset] = P[\emptyset] = 0$
- therefore $P[A_1 \cap A_2 \cap A_3] = P[A_1]P[A_2]P[A_3]$
- but $P[A_1 \cap A_2] = P[\{1, 3, 4\} \cap \{2, 3, 4\}] = P[\{3, 4\}] = \frac{2}{4} = \frac{1}{2}$
- and $P[A_1]P[A_2] = (\frac{3}{4})(\frac{3}{4}) = \frac{9}{16}$
- so $P[A_1 \cap A_2] \neq P[A_1]P[A_2]$

2.6 Worked Problems

Here's a worked problem to illustrate. We're working in our probability space $(\Omega, \mathcal{F}, P)$ and have two events E and F where:

$$P[E] = \frac{1}{3}, \quad P[F] = \frac{1}{6}$$

It is known that events E and F are disjoint. Find $P[E \cap F]$ and $P[E \cup F^c]$ and determine whether E and F are independent. First, $P[E \cap F]$:

$$E \cap F = \emptyset \implies P[E \cap F] = 0$$

Next, $P[F^c]$:

$$\begin{aligned} P[F^c] &= 1 - P[F] && \text{Theorem 2.4 } P[A^c] = 1 - P[A] \\ &= 1 - \frac{1}{6} \\ &= \frac{5}{6} \\ E \cap F &= \emptyset \\ &\implies E \subseteq F^c \\ &\implies E \cup F^c = F^c \\ &\implies P[E \cup F^c] = P[F^c] = \frac{5}{6} \end{aligned}$$

Lastly, are E and F independent? Suppose they are independent, then by definition $P[EF] = P[E]P[F]$, but $E \cap F = \emptyset \implies P[EF] = 0$, and $P[E]P[F] = \frac{1}{3} \cdot \frac{1}{6} \neq 0$, a contradiction. Therefore E and F are not independent.

3 Random Variables

Typically we want to assign numbers to outcomes of the sample space, and to that end we use what are called **random variables**. Random variables are defined in the context of a probability space. A random variable X on a probability space $(\Omega, \mathcal{F}, P)$ is a function X with domain Ω such that $\{\omega \in \Omega \mid X(\omega) \leq x\}$ is an event in $\mathcal{F}$ for every $x \in \mathbb{R}$ (assuming we're defining random variables to be real-valued).

It is common to define random variables to be real-valued, there being interest in computing averages and other statistics, with some sources using “random variable” to mean real-valued random variable specifically. Alternatives include integer-valued random variable ($\mathbb{Z}$ -valued), non-negative integer-valued (i.e., $\mathbb{N}_0$ -valued), rational-valued ($\mathbb{Q}$ -valued), Boolean-valued, some kind of categorical-valued random variable, or anything else measurable that made sense for the given application. So for instance, if we defined our random variable to be $\mathbb{N}_0$ -valued, then instead of requiring that $\forall x \in \mathbb{R}, \{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F}$ we would instead require that $\forall n \in \mathbb{N}_0, \{\omega \in \Omega \mid X(\omega) = n\} \in \mathcal{F}$.

The most common and broadly-applicable approach is to define the random variables as real-valued. For instance, P. Billingsley in *Probability and Measure* defines a random variable as a real-valued function $X = X(\omega)$. In these notes we will use real-valued random variables unless specifying otherwise.

So in the context of a probability space $(\Omega, \mathcal{F}, P)$ we define our random variable X as some measurable function $X : \Omega \rightarrow \mathbb{R}$, or perhaps we have random vectors and we're mapping to $\mathbb{R}^n$. To say more about random variables, let's start with the case when the range of X is finite or countably infinite: in this case we would say X is a **discrete random variable**.

3.1 Discrete Random Variables

A discrete real-valued random variable X on a probability space $(\Omega, \mathcal{F}, P)$ is a function X with domain Ω and range a finite or countably infinite subset $\{x_1, x_2, \dots\}$ of the real numbers $\mathbb{R}$ such that for $\omega \in \Omega$, $\{\omega \mid X(\omega) = x_i\}$ is an event for all i .

How do we make our random values $\mathbb{R}$ -valued when our outcomes are not obviously numeric? For categorical outcomes this may be facilitated by way of a mapping. For example, if our outcomes are letter grades, and $\Omega = \{A, A-, B+, B, \dots, F\}$, then one possible way to map each discrete ω to a real-valued random variable is:

ω	$X(\omega)$
A	4.00
A-	3.67
B+	3.33
B	3.00
...	...
F	0.00

Another example: the Boolean values **TRUE** could be mapped to 1, **FALSE** to 0.

Recall that we have flexibility in defining our event space $\mathcal{F}$, so we can't assume that $\mathcal{F}$ contains all possible subsets of Ω . However, here we've specified in our definition of discrete random variables the requirement that for $\omega \in \Omega$, $\{\omega \mid X(\omega) = x_i\}$ is an event for all i . Consequently, we can talk about the probability of $\{\omega \mid X(\omega) = x_i\}$ since by how we've defined random variables $A = \{\omega \mid X(\omega) = x_i\}$ is an event in $\mathcal{F}$ and thus gets a non-negative real value from our probability function $P[\cdot]$.

Most authors use capital Latin characters such as X for random variables and their corresponding lowercase Latin letter for particular realized values. Thus we might see $\{X = x\} = \{\omega \in \Omega \mid X(\omega) = x\}$ or particular values like $\{X = 0\}$.

Some sources indicate the range of X using Ω_X while others use $\mathcal{X}$. Let's use Ω_X here in these notes to remind us of the linkage between domain Ω and function X .

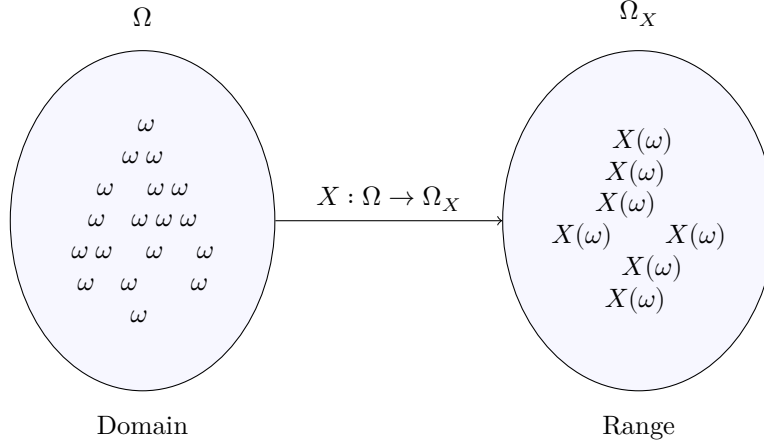


Figure 7: Random variable X is a function from domain Ω to range Ω_X

So how exactly do we get a probability using $P[\cdot]$ on discrete random variables? The first approach is to use probability mass functions.

3.2 Probability Mass Function (PMF)

The probability mass function (PMF) of the discrete random variable X is

$$P_X(x) = P[X = x]$$

Note that $\{X = x\}$ is an event consisting of all outcomes ω from our sample space Ω where $X(\omega) = x$. For any value x in Ω_X , $P_X(x)$ is the probability of the event $\{X = x\}$. Consequently:

- $P_X(x) \geq 0$ for any x
- $\sum_{x \in \Omega_X} P_X(x) = 1$
- for event $A \subseteq \Omega_X$, $P[A] = \sum_{x \in A} P[X = x]$

Some authors refer to a probability mass function as a “discrete density function” and use the notation f_X , but for these notes we will reserve f_X for probability density functions, i.e., continuous density.

If desired, we could define the PMF $P_X(\cdot)$ more broadly so we spell out what happens if given an $x \notin \Omega_X$. For instance, if X was set up to be a real-valued random variable, then when might define P_X more robustly as:

$$P_X : \mathbb{R} \rightarrow [0, 1], \quad P_X(x) = \begin{cases} P[X = x], & x \in \Omega_X \\ 0, & \text{otherwise} \end{cases}$$

In any case, Ω_X is our “minimal domain”, so it’s fine if we think of things as shown in [Figure 8](#).

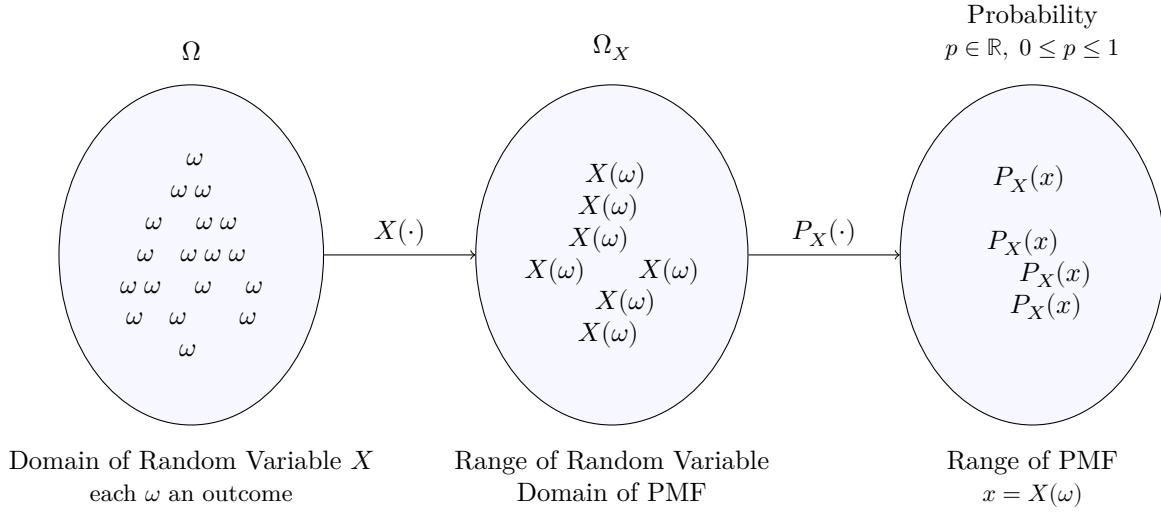


Figure 8: $P_X(\cdot)$ is a function from domain Ω_X to range $[0, 1]$

So for example, the PMF for $\mathbb{R}$ -valued random variable X for the outcome of the roll of a fair six-sided die would be:

$$P_X(x) = \begin{cases} \frac{1}{6}, & x \in \{1, 2, 3, 4, 5, 6\} \\ 0, & \text{otherwise} \end{cases}$$

Alternatively we could set up the outcome of a six-sided die as a categorical-valued random variable having just the six categories $\{\square, \square, \square, \square, \square, \square\}$, but this would require us to create some kind of mapping function if we wanted to compute averages or other statistics. Here since our sample space was $\Omega = \{\square, \square, \square, \square, \square, \square\}$ it was arguably better to define our random variable to be $\mathbb{R}$ -valued and use a transformation from Ω to Ω_X such that $X(\square) = 1$, $X(\square) = 2$, etc.

Where we're basically guaranteed to want to go the real-valued route is when we have continuous random variables. To define continuous random variables we first have to define what we mean by **cumulative distribution function**.

3.3 Cumulative Distribution Function (CDF)

The cumulative distribution function (CDF) of random variable X is

$$F_X(x) = P[X \leq x]$$

So for instance, for our fair six-sided die roll we should find $F_X(1) = \frac{1}{6}$, $F_X(2) = \frac{1}{3}$, etc.

Some sources refer to cumulative distribution functions simply as “distribution functions”.

We immediately inherit multiple interesting properties since $F_X(x)$ describes the probability of an event.

- $F_X(-\infty) = 0$
- $F_X(\infty) = 1$
- $F_X(\cdot)$ is non-decreasing

Bringing in finite additivity we can find the probability of a random variable falling in a given interval.

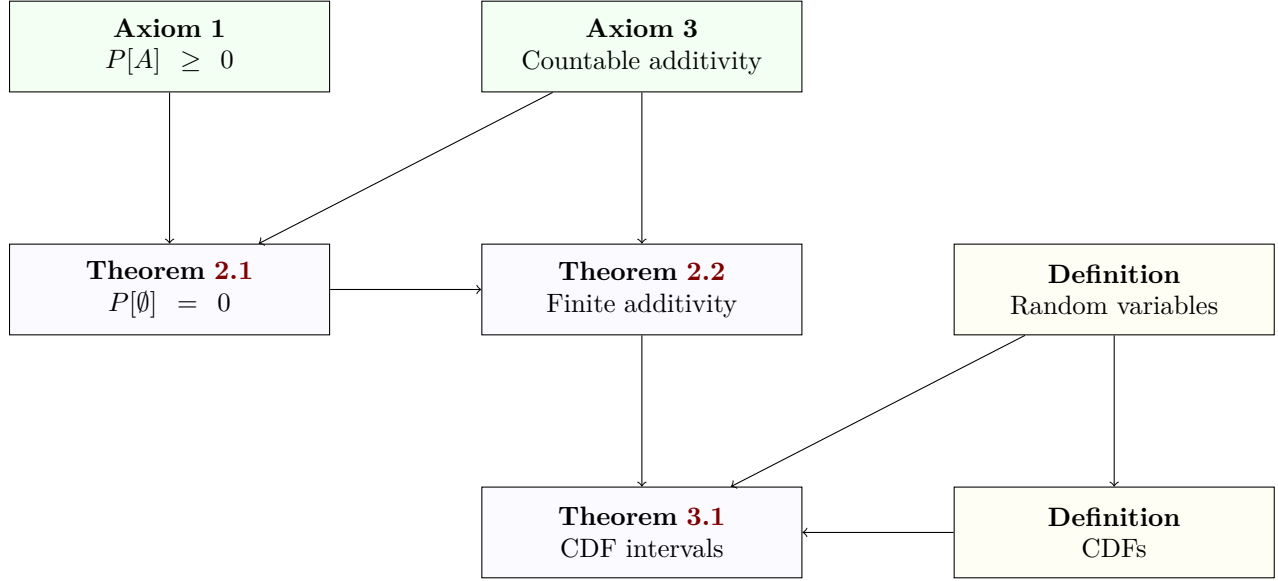


Figure 9: Probability of an interval from the CDF

Theorem 3.1. Probability of an interval from the CDF

For $a \leq b$:

$$F_X(b) - F_X(a) = P[a < X \leq b]$$

Proof. We observe that the event $\{X \leq a\}$ is contained in $\{X \leq b\}$, which we can write as a disjoint union of $\{X \leq a\}$ and $\{a < X \leq b\}$:

$$\begin{aligned}
 P[\{X \leq b\}] &= P[\{X \leq a\} \cup \{a < X \leq b\}] \\
 &= P[\{X \leq a\}] + P[\{a < X \leq b\}] && \text{Theorem 2.2} \\
 P[\{X \leq b\}] - P[\{X \leq a\}] &= P[\{X \leq a\}] - P[\{X \leq a\}] + P[\{a < X \leq b\}] \\
 P[\{X \leq b\}] - P[\{X \leq a\}] &= P[\{a < X \leq b\}] \\
 F_X(b) - F_X(a) &= F_X(a < X \leq b) && \text{definition of } F_X(\cdot)
 \end{aligned}$$

□

Trying this out on our fair six-sided die roll, $P[1 < X \leq 3] = F_X(3) - F_X(1) = \frac{1}{2} - \frac{1}{6} = \frac{1}{3}$, as required.

With CDFs defined we can finally define what we mean by continuous random variable.

3.4 Continuous Random Variables

A **continuous random variable** is a random variable where the CDF $F_x(x)$ is a continuous function.

For instance, the CDF for our fair six-sided die roll under real-valued random variable X is:

$$F_X(x) = \begin{cases} 0, & x < 1 \\ \frac{\lfloor x \rfloor}{6}, & 1 \leq x \leq 6 \\ 1, & x > 6 \end{cases}$$

The notation $\lfloor x \rfloor$ indicates the floor of x . Observe that this $F_X(x)$ is **not continuous** because we have hops at each of the whole number values 1 ... 6. Perhaps rather than writing “observe” without showing the picture it’s best here to set up a side-by-side plot so we can literally observe the difference. First, let’s load our fair die roll into an R `data.frame`.

```
dfFairDie <- data.frame(
  x = 0:6,
  xend = 1:7,
  F = (0:6)/6
)
```

Then let’s pick something continuous to illustrate the distinction. An example of a continuous random variable is when X is an exponential random variable with parameter λ (where $\lambda > 0$):

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

Let’s load that into another `data.frame` and load libraries for plotting the two items. We’ll use $\lambda = 1$.

```
library(ggplot2, quietly=TRUE)
library(patchwork, quietly=TRUE)

x <- seq(0, 6.5, length.out = 651)

dfExp <- data.frame(
  x = x,
  F = 1 - exp(-x) # i.e., lambda=1
)
```

Our leftmost plot: show $F_X(x)$ for the fair die roll. Let’s add open circles (the `shape = 1` part) on right ends to remind us that the intervals are open on that side as they jump up there and filled circles on the left ends to remind us that the intervals are closed on that side.

```
# this will be our left plot
p1 <- ggplot(dfFairDie, aes(x=x, y=F, xend=xend)) +

  geom_point(color = "steelblue") +
  geom_segment(color = "steelblue", linewidth = 1.1) +
  geom_point(aes(xend, F), shape = 1, color = "steelblue") +

  labs(
    x = expression(x),
    y = expression(F[X](x)),
    title = "Die Roll CDF",
    subtitle = "(Discrete Random Variable)"
  ) +

  coord_cartesian(
    xlim = c(0, 6.5),
    ylim = c(0, 1.1)
  ) +

  # include y axis labels at 1/6 intervals
```

```

scale_y_continuous(
  breaks = (0:6)/6,
  labels = parse(text = c(
    "0",
    "frac(1,6)",
    "frac(1,3)",
    "frac(1,2)",
    "frac(2,3)",
    "frac(5,6)",
    "1"
  ))
) +

theme_minimal(base_size = 12)

```

Our rightmost plot: the exponential.

```

# right plot
p2 <- ggplot(dfExp, aes(x, F)) +

  geom_line(color = "steelblue", linewidth = 1.1) +

  labs(
    x = expression(x),
    y = expression(F[X](x)),
    title = "Exponential CDF",
    subtitle = "(Continuous Random Variable)"
  ) +

  coord_cartesian(
    xlim = c(0, 6.5),
    ylim = c(0, 1.1)
  ) +

  scale_y_continuous(
    breaks = (0:6)/6,
  ) +

  theme_minimal(base_size = 12) +

  # we only need the left y-axis
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank(),
        axis.title.y = element_blank())

```

We use | to tell patchwork we want the plots to be side-by-side (it would have been / to stack them top-to-bottom).

```

# display
(p1 | p2) + plot_layout(guides = "collect")

```

Now we can directly observe the difference: the continuous random variable is smooth and uninterrupted on our interval of interest $[0, 6.5]$ (or if we wanted, on $[0, \infty)$). Our other random variable with the die roll is not continuous as we have discontinuities at each of the whole numbers $1, 2, \dots, 6$.

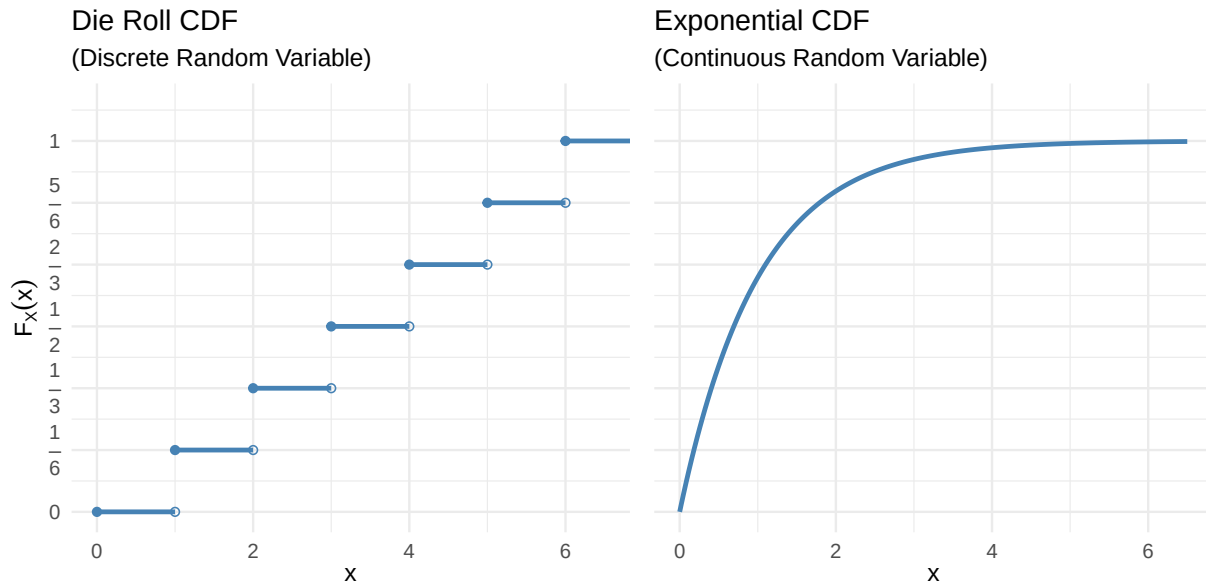


Figure 10: CDFs of two Random Variables

Now that we've defined what we mean by continuous random variable we can define probability density functions.

3.5 Probability Density Function (PDF)

The probability density function (PDF) of a continuous random variable X is defined as the derivative of the CDF $F_X(x)$ with respect to x :

$$f_X(x) = \frac{dF_X(x)}{dx}$$

Here we adopt the Leibniz-style notation where $\frac{d}{dx}$ indicates the derivative with respect to x .

Notice that we immediately get the following properties:

- since CDFs are non-decreasing, $f_X(x) \geq 0$ for all x
- $F_X(x) = \int_{-\infty}^x f_X(u) du$
- $\int_{-\infty}^{\infty} f_X(x) dx = 1$

Now that we've defined both PMFs and PDFs we can do another side-by-side plot for our fair six-sided die roll (PMF for a discrete random variable) versus our exponential (PDF for a continuous random variable).

First we add the probability for each die roll to that `data.frame` we set up earlier. As discussed above, our PMF $P_X(x)$ is $\frac{1}{6}$ for each outcome 1...6.

```
# possible faces are 1 through 6
faces <- 1:6

# P[X=x] is 1/n for each face (where n is however many faces), 0 otherwise
dfFairDie$P_X <- as.numeric(dfFairDie$x %in% faces) / sum(dfFairDie$x %in% faces)
```

For our continuous random variable, we need to find the PDF—our starting place is the CDF.

When $x \geq 0$ and $\lambda > 0$:

$$\begin{aligned}
 f_X(x) &= \frac{d}{dx} [F_X(x)] \\
 &= \frac{d}{dx} [1 - e^{-\lambda x}] \\
 &= \frac{d}{dx} [1] - \frac{d}{dx} [e^{-\lambda x}] && \text{linearity of differentiation} \\
 &= 0 - \frac{d}{dx} [e^{-\lambda x}] && \text{constant term rule} \\
 &= 0 - -\lambda e^{-\lambda x} && \text{chain rule + exponential rule} \\
 &= \lambda e^{-\lambda x} && - \text{sign cancellation}
 \end{aligned}$$

Then if $x < 0$ or $\lambda \leq 0$ we have density 0. Put these two cases together and we define the PDF of the exponential for any x or λ :

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \lambda > 0 \\ 0, & \text{otherwise} \end{cases}$$

It's a one-liner to add $f_X(x)$ to the `data.frame`, taking care to use the same parameter setting ($\lambda = 1$) as before.

```
# f_x(x) = exp(-x) for x >= 0 and lambda=1
dfExp$f <- exp(-dfExp$x)
```

As before, the die roll can go on the left. Let's use closed circles (`shape = 16`) for the discrete values with droplines to emphasize the discrete nature of $P_X(x)$.

```
# this will be our left plot
p1 <- ggplot(dfFairDie, aes(x=x, y=P_X)) +

  geom_point(color = "steelblue", shape = 16) +
  geom_segment(aes(xend = x, yend = 0),
               color = "steelblue",
               linewidth = 1.0) +

  labs(
    x = expression(x),
    y = expression(P[X](x)),
    title = "Die Roll PMF",
    subtitle = "(Discrete Random Variable)"
  ) +

  coord_cartesian(
    xlim = c(0, 6.5),
    ylim = c(0, 1.1)
  ) +

  # include y axis labels at 1/6 intervals
  scale_y_continuous(
    breaks = (0:6)/6,
```

```

labels = parse(text = c(
  "0",
  "frac(1,6)",
  "frac(1,3)",
  "frac(1,2)",
  "frac(2,3)",
  "frac(5,6)",
  "1"
))
) +

theme_minimal(base_size = 12)

```

Our rightmost plot: the exponential.

```

# right plot
p2 <- ggplot(dfExp, aes(x, f)) +

  geom_line(color = "steelblue", linewidth = 1.1) +

  labs(
    x = expression(x),
    y = expression(f[X](x)),
    title = "Exponential PDF",
    subtitle = "(Continuous Random Variable)"
  ) +

  coord_cartesian(
    xlim = c(0, 6.5),
    ylim = c(0, 1.1)
  ) +

  scale_y_continuous(
    breaks = (0:6)/6,
  ) +

  theme_minimal(base_size = 12) +

  # we only need the title on the left y-axis, no need to repeat the ticks
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank())

```

As before we use `|` to tell patchwork we want the plots to be side-by-side.

```

# display
(p1 | p2) + plot_layout(guides = "collect")

```

So we've defined random variables and three basic functions of random variables: PMFs, CDFs, and PDFs. Now we can move on to the topic of expectation.

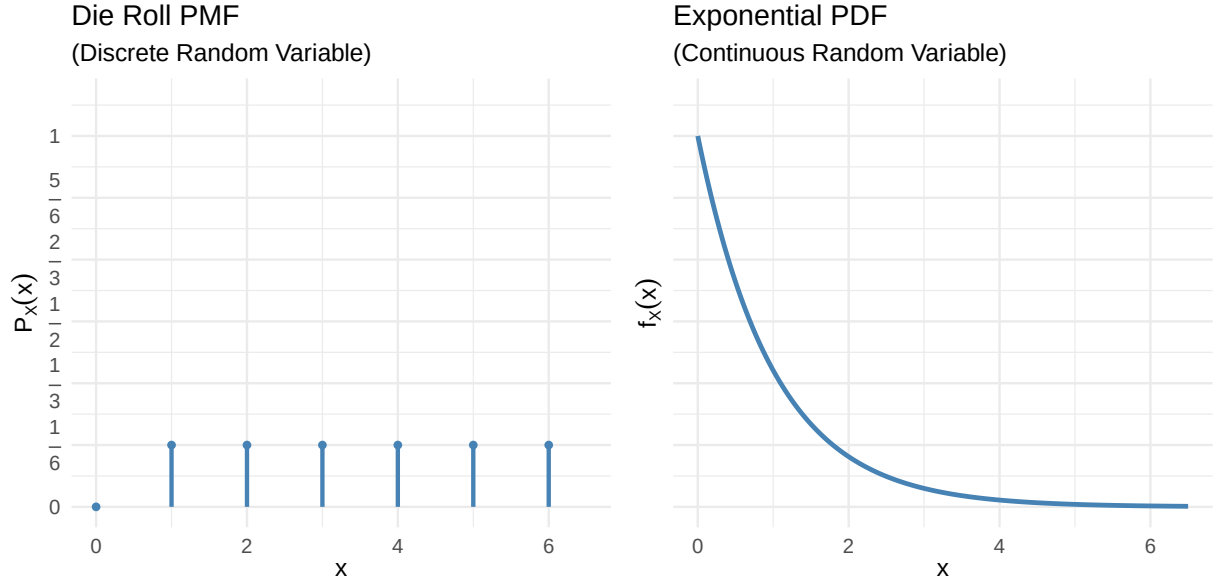


Figure 11: PMF and PDF for two Random Variables

3.6 Expectation

Quoting Abraham de Moivre from his first edition of *The Doctrine of Chances* (1718):

The Expectation of obtaining any Thing, is estimated by the Value of that Thing multiplied by the Probability of obtaining it.

Kolmogorov refers to this as **mathematical expectation** (*mathematische Erwartung* in the original German in 1933). Arguably that’s a more descriptive term less susceptible to misunderstandings, but as most subsequent sources use just “expectation” we will follow this convention in these notes.

The notation $E[\cdot]$ indicates expectation. Alternatives to the $E[\cdot]$ notation include $E(X)$, $\mathbf{E}[X]$, $\mathbb{E}[X]$, $\mathbf{E}(x)$, enclosing the random variable in angle brackets as in $\langle X \rangle$, or just putting the E right up next to the random variable as in EX . A common convention is to write μ_X (with lowercase Greek Mu) to indicate the mean value of random variable X , i.e., $E[X]$, the Mu calling to mind the “m” sound of “mean”.

For our axiomatic framework, Billingsley gives a nice definition of **expectation** in *Probability and Measure*:

The expected value of a [real-valued] random variable X on $(\Omega, \mathcal{F}, P)$ is the integral of X with respect to the measure P

$$E[X] = \int X dP = \int_{\Omega} X(\omega) P[d\omega]$$

Other sources may give the definition (equivalently) as $\int_{\Omega} X(\omega) dP[\omega]$.

Putting this elegant abstract definition from *Probability and Measure* to work we get the following specific formulations for real-valued random variable X .

Discrete	Continuous
$E[X] = \sum_{x \in \Omega_X} x P_X(x)$	$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$

Expectation for some function of X works the same way, e.g., if we wanted the expectation of X^2 then it's:

$E[X^2] = \sum_{x \in \Omega_X} x^2 P_X(x)$	$E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx$
---	--

Some sources refer to this property as the “expected value of a derived random variable”. Ross and some other sources have referred to this the “law of the unconscious statistician” (LoTUS). For discrete random variables the “law” is:

Theorem 3.2. Discrete LoTUS

For PMF P_X and function g of real-valued random variable X having possible outcomes Ω_X :

$$E[g(X)] = \sum_{x \in \Omega_X} g(x) P_X(x)$$

Likewise we have a continuous version.

Theorem 3.3. Continuous LoTUS

For PDF f_X and function g of real-valued random variable X :

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Here's a proof for the discrete case adapted from Yates and Goodman.

Proof. By the definition of expectation for a discrete random variable:

$$E[g(X)] = E[Y] = \sum_{y \in \Omega_Y} y P_Y(y)$$

For any $y \in \Omega_Y$, the event $\{Y = y\}$ is equivalent to the event that $g(X) = y$. Thus:

$$P_Y(y) = P[Y = y] = P[g(X) = y]$$

If g is not necessarily one-to-one, there may be multiple values of x for which $g(x) = y$. We'll use the "stack" notation $\sum_{\substack{x \in \Omega_X \\ g(x)=y}}$ to indicate multiple conditions, in this case that $x \in \Omega_X$ (x is one of the possible values of X) and $g(x) = y$ (among the possible values, also require that they get mapped to y by g), i.e., $x \in \{x \in \Omega_X \mid g(x) = y\}$. Therefore:

$$P[g(X) = y] = \sum_{\substack{x \in \Omega_X \\ g(x)=y}} P[X = x] = \sum_{\substack{x \in \Omega_X \\ g(x)=y}} P_X(x)$$

Substituting this into the expression for $E[Y]$ gives

$$\begin{aligned} E[g(X)] &= \sum_{y \in \Omega_Y} y \sum_{\substack{x \in \Omega_X \\ g(x)=y}} P_X(x) \\ &= \sum_{y \in \Omega_Y} \sum_{\substack{x \in \Omega_X \\ g(x)=y}} g(x) P_X(x) \end{aligned} \quad y = g(x) \text{ for every } x \text{ in the inner sum}$$

The sets $\{x \in \Omega_X : g(x) = y\}$ and $y \in \Omega_Y$ partition Ω_X , so every $x \in \Omega_X$ appears exactly once in the double sum, yielding:

$$E[g(X)] = \sum_{x \in \Omega_X} g(x) P_X(x).$$

□

So for discrete random variables, $E[\cdot]$ is the weighted average, i.e., the weighted arithmetic mean. We can try that out on our fair six-sided die we set up earlier using the `weighted.mean` function of R. We ought to get a value of 3.5.

```
# E[X] ought to be 3.5
weighted.mean(dfFairDie$x, dfFairDie$p)
```

```
## [1] 3.5
```

Next, for our continuous example (the exponential) we apply the PDF we derived above, setting $\lambda = 1$. The expectation of X is over the full support $[0, \infty)$, and if we use $x \geq 0$ and set $\lambda = 1$ we find $E[X] = 1$ as follows:

$$\begin{aligned}
 E[X] &= \int_{-\infty}^{\infty} x f_X(x) dx \\
 &= \int_0^{\infty} x e^{-x} dx && \text{our PDF when } \lambda = 1 \\
 &\quad \text{let } u = x, \, dv = e^{-x} dx \\
 &\quad \Rightarrow du = dx, \, v = -e^{-x} \\
 \int_0^{\infty} x e^{-x} dx &= [uv]_0^{\infty} - \int_0^{\infty} v du && \int u dv = uv - \int v du \\
 &= [-x e^{-x}]_0^{\infty} - \int_0^{\infty} -e^{-x} dx && \text{plugging in } u, v, du \\
 &= [-x e^{-x}]_0^{\infty} + \int_0^{\infty} e^{-x} dx && \text{pull out the minus sign} \\
 &= [-x e^{-x}]_0^{\infty} - [e^{-x}]_0^{\infty} \\
 &= [0 - 0] - [0 - 1] \\
 &= 1
 \end{aligned}$$

We're using the $[x]_a^b$ notation above to indicate “evaluated from $x = a$ to $x = b$ ”.

Recognizing that $[-x e^{-x}]$ evaluated at $x = \infty$ equals 0 was necessary above, maybe that's worth showing—short story is exponential e^x grows faster than x (or any power of x), and we can apply L'Hôp once to see how e^{-x} dominates in the limit:

$$\begin{aligned}
 \lim_{x \rightarrow \infty} x e^{-x} &= \lim_{x \rightarrow \infty} \frac{x}{e^x} && e^{-x} = \frac{1}{e^x} \\
 &= \lim_{x \rightarrow \infty} \frac{\frac{d}{dx}[x]}{\frac{d}{dx}[e^x]} && \text{L'Hôpital's rule} \\
 &= \lim_{x \rightarrow \infty} \frac{1}{\frac{d}{dx}[e^x]} && \frac{d}{dx}[x] = 1 \\
 &= \lim_{x \rightarrow \infty} \frac{1}{e^x} && \frac{d}{dx}[e^x] = e^x \\
 &= 0
 \end{aligned}$$

That demonstrates finding $E[X]$ over the full support. What if we are considering some narrower window? This brings us to the topic of **conditional expectation**.

3.7 Conditional Expectation

Under conditional expectation we are evaluating $E[\cdot]$ given some other event, so we need to normalize by the probability of that conditioning event. One way to do this is to define conditional versions of our PMFs and PDFs.

3.8 Conditional PMFs and PDFs

If event B has positive probability (i.e., $P[B] > 0$), the **conditional probability mass function** of event X given B is:

$$P_{X|B}(x) = P[X = x|B] = \frac{P[\{X = x\} \cap B]}{P[B]}, \quad P[B] > 0$$

We can write that as:

$$P_{X|B}(x) = \begin{cases} \frac{P_X(x)}{P[B]} & x \in B \\ 0 & \text{otherwise} \end{cases}$$

Then our conditional expected value for the discrete random variable is defined using the conditional PMF as:

$$E[X|B] = \sum_{x \in B} x P_{X|B}(x)$$

For the fair die roll case we set up earlier, if we conditioned on $B = \{1, 2\}$ then our $P_{X|B}(x)$ is just 50/50 between 1 and 2, and $E[X|B]$ comes out to 1.5.

What about for the continuous example, the exponential? Earlier we set up a `data.frame` for points on the interval $[0, 6.5]$. Let's use just that interval and find $E[X|0 \leq X \leq 6.5]$, again where $\lambda = 1$. We've already seen that $P[A|B] = \frac{P[AB]}{P[B]}$, and similarly with **conditional density** we need to normalize by the probability of the given event.

If event B has positive probability (i.e., $P[B] > 0$), we have the conditional density of event X given B as:

$$f_X(x|B) = \begin{cases} \frac{f_X(x)}{P[B]}, & \text{if } x \in B \\ 0, & \text{otherwise} \end{cases}$$

Here for $E[X|0 \leq X \leq 6.5]$ let's first find $P[0 \leq X \leq 6.5]$ using the CDF we wrote down earlier. This will be our denominator.

$$\begin{aligned} P[0 \leq X \leq 6.5] &= F_X(6.5) - F_X(0) \\ &= 1 - e^{-\lambda 6.5} - (1 - e^{-\lambda 0}) \\ &= 1 - e^{-6.5} - (1 - e^0) & \lambda = 1 \\ &= 1 - e^{-6.5} \\ &= 1 - e^{-13/2} \end{aligned}$$

Then the numerator, i.e., the unnormalized (and truncated at $x = 6.5$) first moment:

$$\begin{aligned}
 \int_{x=0}^{x=6.5} x f_X(x) dx &= \int_0^{6.5} x e^{-x} dx && \lambda = 1 \\
 &\text{let } u = x, \, dv = e^{-x} dx \\
 &\implies du = dx, \, v = -e^{-x} \\
 \int_0^{6.5} x e^{-x} dx &= [uv]_0^{6.5} - \int_0^{6.5} v du && \text{as before, integration by parts} \\
 &= [-x e^{-x}]_0^{6.5} - \int_0^{6.5} -e^{-x} dx \\
 &= [-x e^{-x}]_0^{6.5} + \int_0^{6.5} e^{-x} dx \\
 &= [-x e^{-x}]_0^{6.5} - [e^{-x}]_0^{6.5} \\
 &= [e^{-x}(-x - 1)]_0^{6.5} && \text{combining} \\
 &= e^{-6.5}(-6.5 - 1) - e^0(-0 - 1) \\
 &= -7.5e^{-6.5} - (-1) \\
 &= 1 - 7.5e^{-6.5} \\
 &= 1 - \frac{15}{2}e^{-13/2}
 \end{aligned}$$

Assembling these we get:

$$E[X|0 \leq X \leq 6.5] = \frac{1 - \frac{15}{2}e^{-13/2}}{1 - e^{-13/2}}$$

We should get something less than but pretty close to 1.

```
(1-(15/2)*exp(-13/2))/(1-exp(-13/2))
```

```
## [1] 0.9902129
```

The built-in numeric integration of R yields the same result.

```
integrate(function(x) x * exp(-x), lower = 0, upper = 6.5)$value/(1-exp(-13/2))
```

```
## [1] 0.9902129
```

Expectation is the **first moment** (“moment” in the mathematical sense). Two other statistics come from the second moment: variance and standard deviation.

3.9 Variance and Standard Deviation

For a real-valued random variable X :

- the n th moment is $E[X^n]$
- the n th central moment is $E[(X - E[X])^n]$

Variance is the 2nd central moment and standard deviation follows from variance.

The **variance** of random variable X is defined as:

$$\text{Var}[X] = E[(X - E[X])^2]$$

Sometimes variance is styled as $^2(X)$ (using lowercase Greek sigma), $\text{Var}[X]$, $\text{Var}(X)$, or just $\text{Var } X$.

For discrete random variable X , the definition means:

$$\text{Var}[X] = \sum_{x \in \Omega_X} P_X(x) (x - E[X])^2$$

Whereas for continuous real-valued random variable Y that means:

$$\text{Var}[Y] = \int_{y=-\infty}^{y=\infty} f_Y(y) (y - E[Y])^2 dy$$

The **standard deviation** of random variable X is defined as:

$$\sigma_X = \sqrt{\text{Var}[X]}$$

Or alternatively, if we were starting from σ_X :

$$\text{Var}[X] = \sigma_X^2$$

The choice of σ (lowercase Greek Sigma) for standard deviation is very common with the σ calling to mind the initial “s” sound of “standard deviation”. For instance, K. Pearson uses σ (in a few places, a Latin capital S) for what he calls “standard-deviation” in his 1893 work *On the Mathematical Theory of Evolution*.

There’s another way to find $\text{Var}[X]$ that can be more convenient:

Theorem 3.4. $\text{Var}[X] = E[X^2] - E[X]^2$

$$\text{Var}[X] = E[X^2] - (E[X])^2$$

Proof. For discrete random variable X :

$$\begin{aligned} \text{Var}[X] &= \sum_{x \in \Omega_X} P_X(x) (x - E[X])^2 && \text{definition} \\ &= \sum_{x \in \Omega_X} P_X(x) (x^2 - 2x E[X] + (E[X])^2) && \text{expand the square} \\ &= \sum_{x \in \Omega_X} P_X(x) x^2 - \sum_{x \in \Omega_X} P_X(x) 2x E[X] + \sum_{x \in \Omega_X} P_X(x) (E[X])^2 && \text{linearity} \\ &= \sum_{x \in \Omega_X} P_X(x) x^2 - \sum_{x \in \Omega_X} P_X(x) 2x E[X] + (E[X])^2 && \sum_{x \in \Omega_X} P_X(x) = 1 \\ &= \sum_{x \in \Omega_X} P_X(x) x^2 - 2 \sum_{x \in \Omega_X} P_X(x) x E[X] + (E[X])^2 && \text{pull out the 2} \\ &= \sum_{x \in \Omega_X} P_X(x) x^2 - 2E[X] E[X] + (E[X])^2 && \sum_{x \in \Omega_X} P_X(x) x = E[X] \\ &= \sum_{x \in \Omega_X} P_X(x) x^2 - 2(E[X])^2 + (E[X])^2 \\ &= \sum_{x \in \Omega_X} P_X(x) x^2 - (E[X])^2 && (E[X])^2 \text{ cancellation} \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

□

For continuous random variables we have a similar proof.

Proof. For continuous random variable X :

$$\begin{aligned}
 \text{Var}[X] &= \int_{x=-\infty}^{x=\infty} f_X(x) (x - E[X])^2 dx && \text{definition} \\
 &= \int_{-\infty}^{\infty} f_X(x) (x^2 - 2x E[X] + (E[X])^2) dx && \text{expand the square} \\
 &= \int_{-\infty}^{\infty} f_X(x) x^2 dx - \int_{-\infty}^{\infty} f_X(x) 2x E[X] dx + \int_{-\infty}^{\infty} f_X(x) (E[X])^2 dx && \text{linearity} \\
 &= \int_{-\infty}^{\infty} f_X(x) x^2 dx - 2 \int_{-\infty}^{\infty} f_X(x) x E[X] dx + \int_{-\infty}^{\infty} f_X(x) (E[X])^2 dx && \text{pull out the 2} \\
 &= \int_{-\infty}^{\infty} f_X(x) x^2 dx - 2 \cdot E[X] \cdot E[X] + \int_{-\infty}^{\infty} f_X(x) (E[X])^2 dx && \int_{-\infty}^{\infty} f_X(x) x = E[X] \\
 &= \int_{-\infty}^{\infty} f_X(x) x^2 dx - 2 \cdot E[X] \cdot E[X] + 1 \cdot (E[X])^2 && \int_{-\infty}^{\infty} f_X(x) = 1 \\
 &= \int_{-\infty}^{\infty} f_X(x) x^2 dx - 2(E[X])^2 + (E[X])^2 \\
 &= \int_{-\infty}^{\infty} f_X(x) x^2 dx - (E[X])^2 \\
 &= E[X^2] - (E[X])^2
 \end{aligned}$$

□

Looking back at our example random variables for the fair six-sided die roll and the exponential, both random variables have distributions that belong to what some sources refer to as the “families of random variables”, which is to say certain categories of distributions famous enough that get their own name, arguably the most “famous” of which is the normal distribution. In the next section we collect the families together alphabetically.

4 Families of Random Variables

The plots for the distribution of the random variable for the fair six-sided die were informative; let’s write a little helper function to facilitate plotting the other discrete distributions. In R we can use an ellipsis for a variable number of arguments, which for ggplot we can tack on at the end after putting into a list.

```

PlotPMF <- function(df,
  title = "PMF",
  subtitle = "(Discrete Random Variable)",
  ...) {

  p <- ggplot(df, aes(x = x, y = P_X)) +

    geom_point(color = "steelblue", shape = 16) +

    geom_segment(
      aes(xend = x, yend = 0),
      color = "steelblue",
      linewidth = 1.0
    )
}

```

```

) +

labs(
  x = expression(x),
  y = expression(P[X](x)),
  title = title,
  subtitle = subtitle
) +

theme_minimal(base_size = 12)

# add any var args to the end, e.g., coord_cartesian(xlim=c(0,1),ylim=c(0,1))
p + list(...)
}

```

Similarly for plotting discrete CDFs.

```

PlotDiscreteCDF <- function(df,
                             title = "CDF",
                             subtitle = "(Discrete Random Variable)",
                             ...) {

  # automatically add the xend; for last one use diff between prior
  df$xend <- c(df$x[-1], tail(df$x, 1) + diff(tail(df$x, 2)))

  p <- ggplot(df, aes(x = x, y = F, xend = xend)) +

    # horizontal step
    geom_segment(
      color = "steelblue",
      linewidth = 1.1
    ) +

    # closed left endpoint
    geom_point(
      color = "steelblue",
      shape = 16
    ) +

    # open right endpoint
    geom_point(
      aes(x = xend, y = F),
      color = "steelblue",
      shape = 1
    ) +

    labs(
      x = expression(x),
      y = expression(F[X](x)),
      title = title,
      subtitle = subtitle
    ) +

```

```

    theme_minimal(base_size = 12)

    # add any var args to the end
    p + list(...)
  }

```

Now for each discrete distribution family of interest we can plot a PMF and CDF.

What about for continuous distributions? For PDFs we'll use `geom_line` for line plots:

```

PlotPDF <- function(df,
                    title = "PDF",
                    subtitle = "(Continuous Random Variable)",
                    ...) {

  p <- ggplot(df, aes(x = x, y = f_X)) +

    geom_line(
      color = "steelblue",
      linewidth = 1.1
    ) +

    labs(
      x = expression(x),
      y = expression(f[X](x)),
      title = title,
      subtitle = subtitle
    ) +

    theme_minimal(base_size = 12)

  # add any var args to the end
  p + list(...)
}

```

Likewise for CDFs for continuous distributions:

```

PlotContinuousCDF <- function(df,
                             title = "CDF",
                             subtitle = "(Continuous Random Variable)",
                             ...) {

  p <- ggplot(df, aes(x = x, y = F)) +

    geom_line(
      color = "steelblue",
      linewidth = 1.1
    ) +

    labs(
      x = expression(x),
      y = expression(F[X](x)),
      title = title,

```

```

    subtitle = subtitle
  ) +

  theme_minimal(base_size = 12)

# add any var args to the end
p + list(...)
}

```

Now on the the families.

4.1 Bernoulli (Discrete)

Alphabetically our first family of distributions, the Bernoulli distribution is a particularly simple one.

For $0 \leq p \leq 1$:

Bernoulli

$P_X(x)$	=	$\begin{cases} 1-p & x=0 \\ p & x=1 \\ 0 & \text{otherwise} \end{cases}$
$E[X]$	=	p
$\text{Var}[X]$	=	$p(1-p)$

It's so basic that R doesn't even bother with a separate function; instead we treat it as a 1-trial case of the binomial distribution. With Bernoulli we have one trial with outcome success or failure, often coded as 1 or TRUE for success and 0 or FALSE for failure.

To plot the PMF and CDF, first we set up the data using the R built-ins `dbinom` and `pbinom`.

```

dfBernoulli <- data.frame(
  x = 0:1,
  P_X = dbinom(0:1, size = 1, prob = 0.5),
  F = pbinom(0:1, size = 1, prob = 0.5)
)

```

For our Bernoulli plots we want the same limits and labels—we can put those in a list.

```

listPlotSettingsBernoulli <- list(
  coord_cartesian(
    xlim = c(-0.1, 1.1),
    ylim = c(0, 1.05)
  ),
  scale_y_continuous(
    breaks = (0:2)/2,
    labels = parse(text = c("0", "frac(1,2)", "1"))
  ),
  scale_x_continuous(
    breaks = c(0, 1),
    labels = c("0", "1")
  )
)

```

For the left plot we want the PMF.

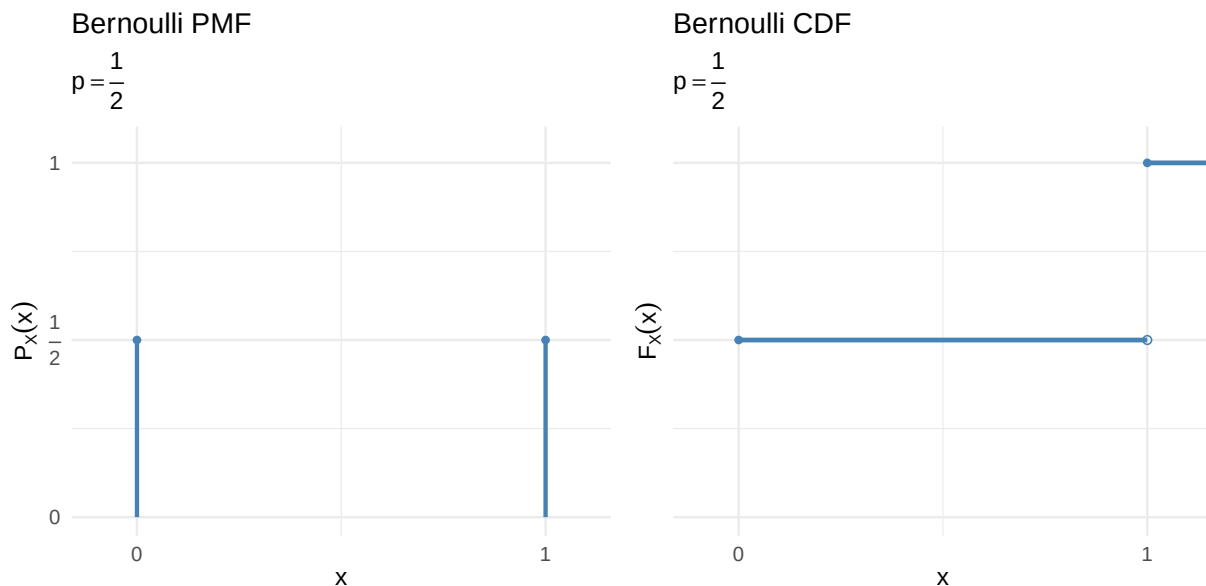
```
p1 <- PlotPMF(dfBernoulli,
  title = "Bernoulli PMF",
  subtitle = expression(p==frac(1,2)),
  listPlotSettingsBernoulli
)
```

The right plot will be the CDF; here we suppress repeating the y-axis text and ticks as we're keeping the same axis.

```
p2 <- PlotDiscreteCDF(dfBernoulli,
  title = "Bernoulli CDF",
  subtitle = expression(p==frac(1,2)),
  listPlotSettingsBernoulli,
  theme(axis.text.y = element_blank(),
    axis.ticks.y = element_blank())
)
```

As before, the | tells patchwork we want the plots to be side-by-side.

```
# display
(p1 | p2) + plot_layout(guides = "collect")
```



The Bernoulli distribution is named for Jacob Bernoulli.

4.2 Beta (Continuous)

For $i, j \in \mathbb{Z}_+$ the beta function is defined as:

$$\beta(i, j) = \frac{(i + j - 1)!}{(i - 1)!(j - 1)!}$$

Where ! is the symbol for factorial. Then for a $\beta(i, j)$ random variable X we have:

Beta

$$\begin{aligned}
 f_X(x) &= \begin{cases} \beta(i, j) x^{i-1} (1-x)^{j-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \\
 E[X] &= \frac{i}{i+j} \\
 \text{Var}[X] &= \frac{ij}{(i+j)^2 (i+j+1)}
 \end{aligned}$$

The $\beta(i, j)$ is necessary to normalize (i.e., so that total probability is 1).

The parameters i and j determine the shape of the density. To illustrate the distribution, we'll use $i = 2$ and $j = 5$. R provides the built-in functions `dbeta` and `pbeta` for the PDF and CDF. The Beta distribution functionality in R supports any positive real numbers for i and j , not just positive integers.

```

i <- 2
j <- 5
dfBeta <- data.frame(
  x = seq(0, 1, length.out = 500),
  f_x = dbeta(seq(0, 1, length.out = 500), shape1 = i, shape2 = j),
  F = pbeta(seq(0, 1, length.out = 500), shape1 = i, shape2 = j)
)

```

Plotting as with the other distributions (no need to print the code).

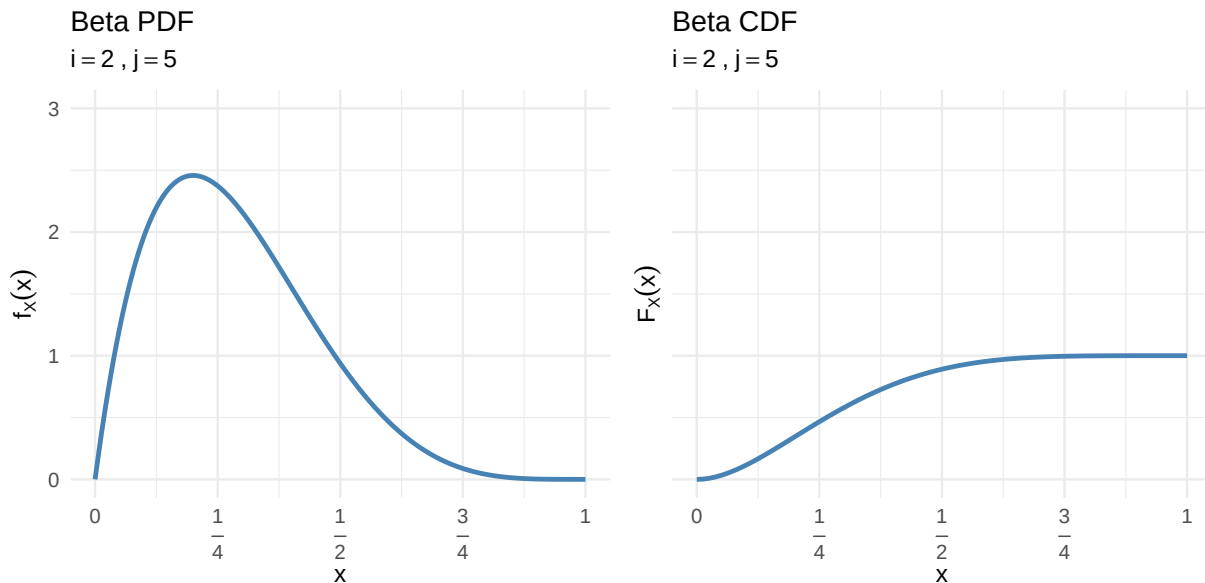


Figure 12: PMF and CDF for Beta-distributed Random Variable

The Beta distribution is closely related to the binomial distribution and is commonly used to model quantities that are restricted to the interval $(0, 1)$ or $[0, 1]$.

Setting $i = j = 1$ gives us a uniform distribution. If both i and j are less than 1 we get a bimodal distribution with spikes at 0 and 1. When i and j are both greater than 1 we get a unimodal distribution: balanced when $i = j$, off-center otherwise (left-skewed when $j > i$, right-skewed when $j < i$).

4.3 Binomial (Discrete)

For $n \in \mathbb{Z}_+$, $x \in \{0, 1, \dots, n\}$ and $0 \leq p \leq 1$:

Binomial

$$\begin{aligned} P_X(x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ E[X] &= np \\ \text{Var}[X] &= np(1-p) \end{aligned}$$

The $\binom{n}{x}$ notation indicates “ n choose x ”.

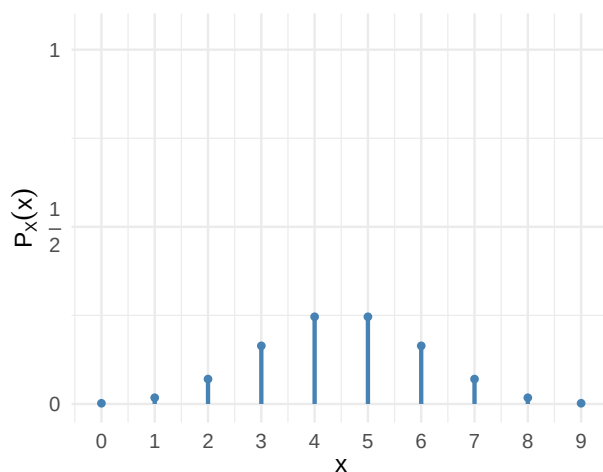
To plot the PMF and CDF, first we set up the data using the R built-ins `dbinom` and `pbinom`.

```
dfBinomial <- data.frame(  
  x = 0:9,  
  P_X = dbinom(0:9, size = 9, prob = 0.5),  
  F = pbinom(0:9, size = 9, prob = 0.5)  
)
```

No need to print out necessary plotting code.

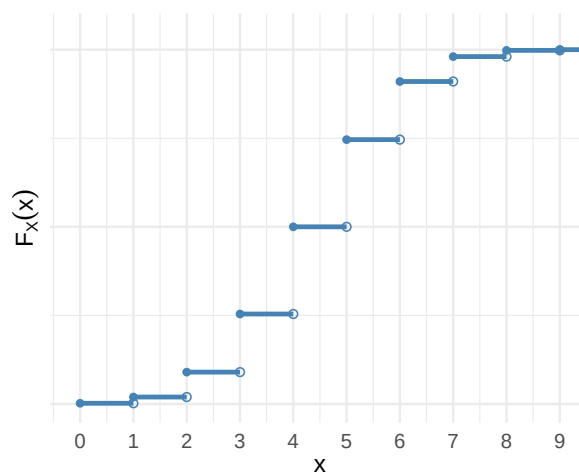
Binomial PMF

$$p = \frac{1}{2}$$



Binomial CDF

$$p = \frac{1}{2}$$



With a Binomial distribution we have some number n of trials, each with outcome success (at probability p) or failure (at probability $1-p$).

4.4 Cauchy (Continuous)

For scale $\gamma > 0$ (lowercase Greek gamma), $x \in \mathbb{R}$, and location $-\infty < x_0 < \infty$:

Cauchy

$$\begin{aligned}
 f_X(x) &= \frac{1}{\pi\gamma \left[1 + \left(\frac{x-x_0}{\gamma}\right)^2\right]} \\
 &= \frac{1}{\pi} \frac{\gamma}{\gamma^2 + (x-x_0)^2} \\
 E[X] &= \text{undefined} \\
 \text{Var}[X] &= \text{undefined}
 \end{aligned}$$

Our R built-ins are `dcauchy` and `pcauchy`.

```
dfCauchy <- data.frame(
  x = seq(-4, 4, length.out = 500),
  f_x = dcauchy(seq(-4, 4, length.out = 500), location=0, scale=0.5),
  F = pcauchy(seq(-4, 4, length.out = 500), location=0, scale=0.5)
)
```

Plotting.

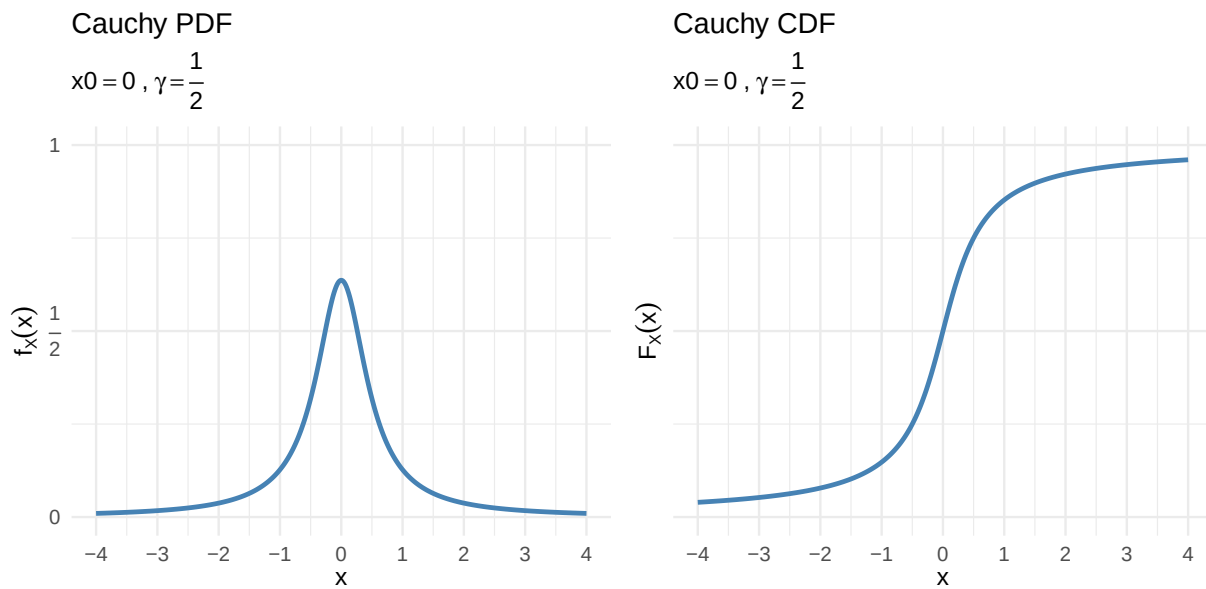


Figure 13: PMF and CDF for Cauchy-distributed Random Variable

4.5 Discrete Uniform (Discrete)

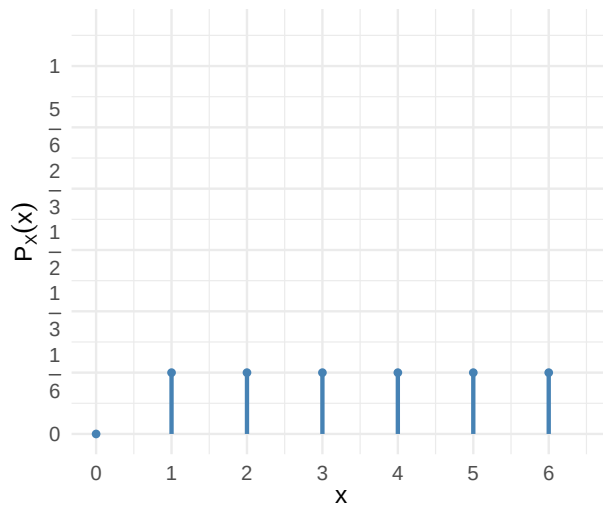
For $a, b \in \mathbb{Z}$ with $a \leq b$:

Discrete Uniform

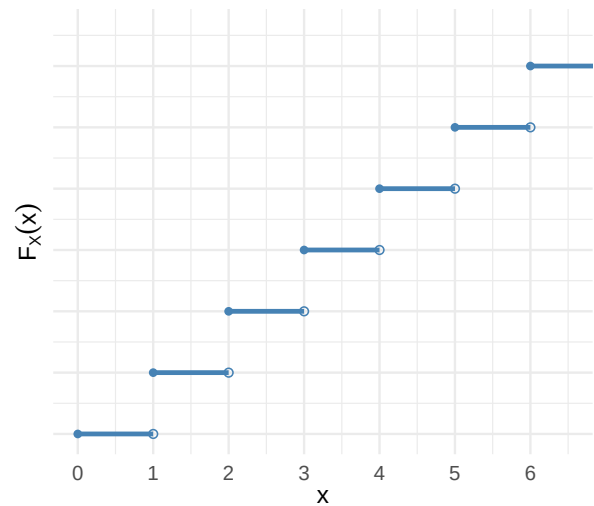
$$\begin{aligned}
 P_X(x) &= \begin{cases} \frac{1}{b-a+1} & x \in \{a, a+1, \dots, b\} \\ 0 & \text{otherwise} \end{cases} \\
 E[X] &= \frac{a+b}{2} \\
 \text{Var}[X] &= \frac{(b-a+1)^2 - 1}{12} \\
 &= \frac{(b-a)(b-a+2)}{12}
 \end{aligned}$$

We visualized this one already with the fair six-sided die.

Discrete Uniform PMF
a=1, b=6



Discrete Uniform CDF
a=1, b=6



Where do we get that 12 in the denominator of the variance? Let $n = b - a + 1$. Shifting a random variable does not change its variance, so define $Y = X - a + 1$. Now Our expectation of discrete uniform random variable Y is:

$$E[Y] = \frac{1}{n} \sum_{k=1}^n k = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2},$$

and the expected value of Y^2 is:

$$E[Y^2] = \frac{1}{n} \sum_{k=1}^n k^2 = \frac{1}{n} \frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}.$$

Now we can apply Theorem 3.4.

$$\begin{aligned}
\text{Var}[Y] &= E[Y^2] - (E[Y])^2 && \text{Theorem 3.4} \\
&= \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 \\
&= \frac{2(n+1)(2n+1)}{2 \cdot 6} - \frac{3(n+1)^2}{3 \cdot 2^2} \\
&= \frac{2(n+1)(2n+1) - 3(n+1)^2}{12} && \text{combining} \\
&= \frac{(n+1)[2(2n+1) - 3(n+1)]}{12} && \text{factor out } (n+1) \\
&= \frac{(n+1)(4n+2-3n-3)}{12} \\
&= \frac{(n+1)(n-1)}{12} \\
&= \frac{n^2-1}{12}.
\end{aligned}$$

We set $Y = X - a + 1$, so $\text{Var}[X] = \text{Var}[Y]$, and we have:

$$\text{Var}[X] = \frac{n^2-1}{12}$$

Substituting $n = b - a + 1$,

$$\begin{aligned}
\text{Var}[X] &= \frac{(b-a+1)^2-1}{12} \\
&= \frac{((b-a)+1)^2-1^2}{12} && \text{recognize a difference of squares} \\
&= \frac{((b-a)+1-1)((b-a)+1+1)}{12} \\
&= \frac{(b-a)(b-a+2)}{12}
\end{aligned}$$

$u^2 - v^2 = (u-v)(u+v)$

4.6 Erlang (Continuous)

For $k \in \mathbb{Z}_+$ and $\lambda > 0$:

Erlang

$ \begin{aligned} f_X(x) &= \begin{cases} \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} & x > 0 \\ 0 & \text{otherwise} \end{cases} \\ E[X] &= \frac{k}{\lambda} \\ \text{Var}[X] &= \frac{k}{\lambda^2} \end{aligned} $
--

4.7 Exponential (Continuous)

For $\lambda > 0$:

Exponential

$$\begin{aligned} f_X(x) &= \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \\ E[X] &= \frac{1}{\lambda} \\ \text{Var}[X] &= \frac{1}{\lambda^2} \end{aligned}$$

We saw this earlier with $\lambda = 1$; this time let's try plotting it with $\lambda = \frac{3}{4}$. Rather than crafting our own function with `exp` we will use the `dexp` and `pexp` built-ins.

```
dfExponential <- data.frame(  
  x = seq(0, 9, length.out = 500),  
  f_X = dexp(seq(0, 9, length.out = 500), rate = 0.75),  
  F = pexp(seq(0, 9, length.out = 500), rate = 0.75)  
)
```

No need to print out necessary plotting code. In RMarkdown we suppress the code printout using `echo=FALSE` in the chunk options.

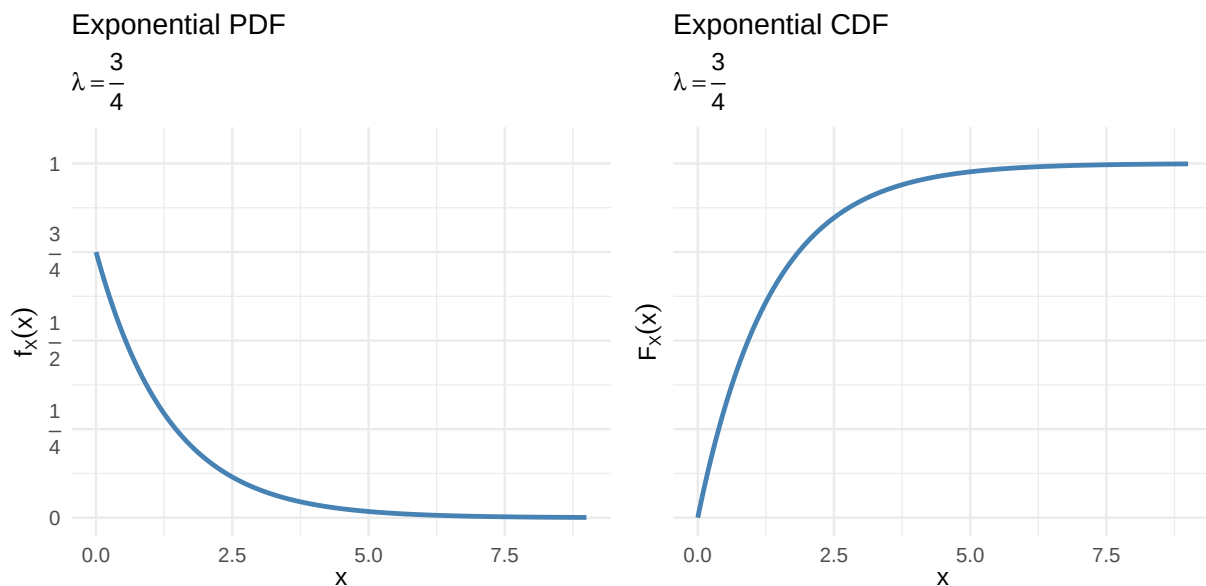


Figure 14: PMF and CDF for Exponentially-distributed Random Variable

4.8 Gamma (Continuous)

For shape parameter $\alpha > 0$ and rate parameter $\beta > 0$

Gamma

$$\begin{aligned}
f_X(x) &= \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \\
E[X] &= \frac{\alpha}{\beta} \\
\text{Var}[X] &= \frac{\alpha}{\beta^2}
\end{aligned}$$

4.9 Gaussian (Continuous)

For mean $\mu \in \mathbb{R}$ (lowercase Greek mu) and standard deviation $\sigma > 0$ (lowercase Greek sigma):

Gaussian

$$\begin{aligned}
f_X(x) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\
E[X] &= \mu \\
\text{Var}[X] &= \sigma^2
\end{aligned}$$

4.10 Geometric (Discrete)

For $0 < p \leq 1$:

Geometric

$$\begin{aligned}
P_X(x) &= \begin{cases} p(1-p)^{x-1} & x = 1, 2, \dots \\ 0 & \text{otherwise} \end{cases} \\
E[X] &= \frac{1}{p} \\
\text{Var}[X] &= \frac{1-p}{p^2}
\end{aligned}$$

4.11 Laplace (Continuous)

For location (mean) $\mu \in \mathbb{R}$ and scale $b > 0$:

Laplace

$$\begin{aligned}
f_X(x) &= \frac{1}{2b} e^{-\frac{|x-\mu|}{b}} \\
E[X] &= \mu \\
\text{Var}[X] &= 2b^2
\end{aligned}$$

4.12 Multinomial (Discrete)

For $n \in \mathbb{Z}_+$ and $p_1, \dots, p_k \geq 0$ such that $\sum_{i=1}^k p_i = 1$:

Multinomial

$$\begin{array}{rcl} P_X(x_1, \dots, x_k) & = & \begin{cases} \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k} & \sum_{i=1}^k x_i = n \\ 0 & \text{otherwise} \end{cases} \\ E[X_i] & = & np_i \\ \text{Var}[X_i] & = & np_i(1 - p_i) \end{array}$$

So in other words, each X_i has an $E[X_i]$. The complete $E[X]$ would be a vector:

$$E[X] = \begin{bmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_k] \end{bmatrix}$$

Likewise, each X_i has a $\text{Var}[X_i]$.

4.13 Poisson (Discrete)

For rate $\lambda \in \mathbb{R}_+$:

Poisson

$$\begin{array}{rcl} P_X(x) & = & \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!} & x = 0, 1, 2, \dots \\ 0 & \text{otherwise} \end{cases} \\ E[X] & = & \lambda \\ \text{Var}[X] & = & \lambda \end{array}$$

The parameter λ represents the average number of events occurring in a fixed interval. Alternatively we can parameterize it with average rate r and time interval $T = t_1 - t_0$ so that $\lambda = rT$.

To plot the PMF and CDF, first we set up the data using the R built-ins `dpois` and `ppois`. Let's choose to set $\lambda = 4$.

```
dfPoisson <- data.frame(
  x = 0:9,
  p_x = dpois(0:9, lambda = 4),
  F = ppois(0:9, lambda = 4)
)
```

Plotting is the same as the other discrete distributions (not repeating the code here).

With a Poisson distribution the random variable X counts the number of events occurring in a fixed interval.

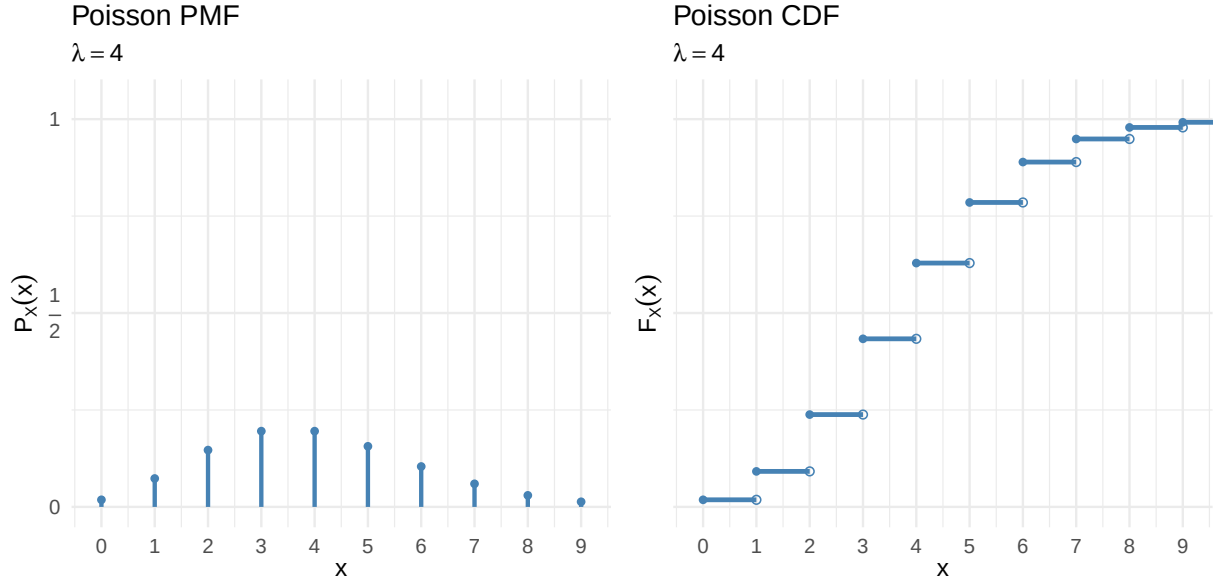


Figure 15: PMF and CDF for Poisson-distributed Random Variable

4.14 Uniform (Continuous)

For $a, b \in \mathbb{R}$ with $a \leq b$:

Discrete Uniform

$$\begin{aligned}
 P_X(x) &= \begin{cases} \frac{1}{b-a+1} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \\
 E[X] &= \frac{a+b}{2} \\
 \text{Var}[X] &= \frac{(b-a)^2}{12}
 \end{aligned}$$

5 Multiple Random Variables

As mentioned earlier,

in the context of a probability space $(\Omega, \mathcal{F}, P)$ we define our random variable X as some measurable function $X : \Omega \rightarrow \mathbb{R}$, or perhaps we have random vectors and we're mapping to $\mathbb{R}^n$.

When we have $X : \Omega \rightarrow \mathbb{R}$ the random variable is said to be **univariate**, whereas an n -dimensional random vector is some measurable function $X : \Omega \rightarrow \mathbb{R}^n$. A common scenario is when $n = 2$, in which case we have a **bivariate** random vector. More generally we can say that we have a **multivariate** random vector.

5.1 Definitions and Notation

5.1.1 Multivariate CDFs

The **joint cumulative distribution function** of random variables X and Y is:

$$F_{X,Y}(x, y) = P[X \leq x, Y \leq y]$$

More generally, the joint cumulative distribution function for an n -dimensional random vector $X = \{X_1, \dots, X_n\}$ is:

$$F_{X_1, \dots, X_n}(x_1, \dots, x_n) = P[X_1 \leq x_1, \dots, X_n \leq x_n]$$

Multivariate PMFs

If we're working with a discrete random vector we define a **joint probability mass function** as:

$$P_{X,Y}(x, y) = P[X = x, Y = y]$$

or more generally as:

$$P_{X_1, \dots, X_n}(x_1, \dots, x_n) = P[X_1 = x_1, \dots, X_n = x_n]$$

Each set $\{X_1 = x_1, \dots, X_n = x_n\}$ is an event, and we might name our joint range $\Omega_{X_1, \dots, X_n}$.

5.1.2 Multivariate PDFs

For continuous multivariate random variables we define a **joint probability density function** as:

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y}$$

or more generally as:

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \frac{\partial^n F_{X_1, \dots, X_n}(x_1, \dots, x_n)}{\partial x_1 \dots \partial x_n}$$

The $\frac{\partial^n}{\partial x \partial y}$ is Leibniz-style notation for second order mixed partial derivatives, i.e., $\frac{\partial}{\partial x} \left(\frac{\partial}{\partial y} \right)$.

5.1.3 Vector Notation

The $X_1, \dots, X_n$ notation is a little tiresome. Let's permit bold letters to represent vectors so that for instance n -dimensional random vector $\mathbf{X}$ is:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

Then we have:

- X and X_i are random variables
- x and x_i are realizations/values
- $\mathbf{X}$ is a random vector
- $\mathbf{x}$ is a realization of a random vector

Now we can express the multivariate functions more concisely as:

$$\begin{aligned} F_{\mathbf{X}}(\mathbf{x}) &= F_{X_1, \dots, X_n}(x_1, \dots, x_n) && \text{CDF of random vector } \mathbf{X} \\ P_{\mathbf{X}}(\mathbf{x}) &= P_{X_1, \dots, X_n}(x_1, \dots, x_n) && \text{PMF of discrete random vector } \mathbf{X} \\ f_{\mathbf{X}}(\mathbf{x}) &= f_{X_1, \dots, X_n}(x_1, \dots, x_n) && \text{PDF of continuous random vector } \mathbf{X} \end{aligned}$$

Furthermore, let's write more compactly as $\mathbf{X} = [X_1 \ X_2 \ \dots \ X_n]^\top$ where $^\top$ is the transpose operator.

We can also extend this to pairs of random vectors. So if X has n components and Y has m components we get:

$$\begin{aligned} F_{\mathbf{X}, \mathbf{Y}}(\mathbf{x}, \mathbf{y}) &= F_{X_1, \dots, X_n, Y_1, \dots, Y_m}(x_1, \dots, x_n, y_1, \dots, y_m) && \text{joint CDF of random vectors } \mathbf{X}, \mathbf{Y} \\ P_{\mathbf{X}, \mathbf{Y}}(\mathbf{x}, \mathbf{y}) &= P_{X_1, \dots, X_n, Y_1, \dots, Y_m}(x_1, \dots, x_n, y_1, \dots, y_m) && \text{joint PMF of discrete random vectors } \mathbf{X}, \mathbf{Y} \\ f_{\mathbf{X}, \mathbf{Y}}(\mathbf{x}, \mathbf{y}) &= f_{X_1, \dots, X_n, Y_1, \dots, Y_m}(x_1, \dots, x_n, y_1, \dots, y_m) && \text{joint PDF of continuous random vectors } \mathbf{X}, \mathbf{Y} \end{aligned}$$

5.1.4 Multivariate Expectation

For expectation, just as lowercase Greek Mu (μ) is often used to evoke the “m” in “mean”, $\mu_{\mathbf{X}}$ is often chosen to name the column vector having expected values of an $m \times n$ matrix $\mathbf{X}$ of random vectors $X_1 \dots X_n$:

$$E[\mathbf{X}] = \mu_{\mathbf{X}} = \begin{bmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{bmatrix} = [E[X_1] \ E[X_2] \ \dots \ E[X_n]]^\top$$

5.1.5 Correlation and Covariance

The correlation of $\mathbf{X}$ —called the autocorrelation matrix, the correlation matrix, or the second moment matrix—is the symmetric matrix where each i, j element is the expected value of $X_i X_j$.

$$R_{\mathbf{X}} = E[\mathbf{X}\mathbf{X}^\top] = \begin{bmatrix} E[X_1 X_1] & E[X_1 X_2] & \dots & E[X_1 X_n] \\ E[X_2 X_1] & E[X_2 X_2] & \dots & E[X_2 X_n] \\ \vdots & \vdots & \ddots & \vdots \\ E[X_n X_1] & E[X_n X_2] & \dots & E[X_n X_n] \end{bmatrix}$$

Here the choice of capital R can bring to mind the use of r or ρ (lowercase Greek Rho) for the Pearson correlation coefficient.

From $\mu_{\mathbf{X}}$ and $R_{\mathbf{X}}$ we construct the covariance matrix $C_{\mathbf{X}}$ (sometimes named $K_{\mathbf{X}}$).

$$C_{\mathbf{X}} = E[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^\top] = R_{\mathbf{X}} - \mu_{\mathbf{X}}\mu_{\mathbf{X}}^\top$$

The choice of “C” or “K” can remind us of the hard C sound in “Covariance matrix” (in German, the literal K of *Kovarianzmatrix*). The matrix looks like this:

$$C_{\mathbf{X}} = \begin{bmatrix} \text{Cov}[X_1 X_1] & \text{Cov}[X_1 X_2] & \dots & \text{Cov}[X_1 X_n] \\ \text{Cov}[X_2 X_1] & \text{Cov}[X_2 X_2] & \dots & \text{Cov}[X_2 X_n] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[X_n X_1] & \text{Cov}[X_n X_2] & \dots & \text{Cov}[X_n X_n] \end{bmatrix} = \begin{bmatrix} \text{Var}[X_1] & \text{Cov}[X_1 X_2] & \dots & \text{Cov}[X_1 X_n] \\ \text{Cov}[X_2 X_1] & \text{Var}[X_2] & \dots & \text{Cov}[X_2 X_n] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[X_n X_1] & \text{Cov}[X_n X_2] & \dots & \text{Var}[X_n] \end{bmatrix}$$

The covariance matrix is symmetric and positive semi-definite.

From using $R_{\mathbf{X}}$ for autocorrelation it's a natural extension to use $R_{\mathbf{XY}}$ for the cross-correlation between $\mathbf{X}$ and $\mathbf{Y}$ and $C_{\mathbf{XY}}$ for the cross-covariance between $\mathbf{X}$ and $\mathbf{Y}$.

One last notation addition: before we used $[x]_a^b$ notation above to indicate “evaluated from $x = a$ to $x = b$ ”, but as we have many brackets in play with multivariate random variables, let's use right vertical bars so that $x \Big|_a^b$ indicates “evaluated from $x = a$ to $x = b$ ”.

5.2 Worked Problems

Let's take a tour through that notation by working some problems. Suppose we have a three-dimensional random vector $\mathbf{X} = [X_1 \ X_2 \ X_3]^\top$ having PDF:

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} 6 & 0 \leq x_1 \leq x_2 \leq x_3 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

What is our expected value vector $E[\mathbf{X}]$?

One thing to check is whether this is already normalized, i.e., the total probability is already unity.

$$\begin{aligned} \int_{\mathbb{R}^3} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} &= \int_{x_3=0}^1 \int_{x_2=0}^{x_3} \int_{x_1=0}^{x_2} 6 dx_1 dx_2 dx_3 \\ &= \int_{x_3=0}^1 \int_{x_2=0}^{x_3} 6x_2 dx_2 dx_3 \\ &= \int_{x_3=0}^1 6 \cdot \frac{1}{2} \cdot x_3^2 dx_3 \\ &= 6 \cdot \frac{1}{2} \cdot \frac{1}{3} \cdot x_3^3 \Big|_{x_3=0}^{x_3=1} = 1 \end{aligned}$$

That checks out. Next, since we're looking to build $E[\mathbf{X}] = [E[X_1] \ E[X_2] \ E[X_3]]^\top$ perhaps a good approach is to roll up the $f_{\mathbf{X}}(\mathbf{x})$ to get the three joint PDFs f_{X_1, X_2} , f_{X_2, X_3} , f_{X_1, X_3} , then roll those up to get the three univariate PDFs f_{X_1} , f_{X_2} , f_{X_3} , taking care to keep track of the bounds.

First:

$$f_{X_1, X_2}(x_1, x_2) = \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) dx_3 = \int_{x_2}^1 6 dx_3 = 6x_3 \Big|_{x_3=x_2}^{x_3=1} = 6(1 - x_2)$$

Being careful with the bounds we get a new PDF:

$$f_{X_1, X_2}(x_1, x_2) = \begin{cases} 6(1 - x_2) & 0 \leq x_1 \leq x_2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

A second round of marginalization:

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1, X_2}(x_1, x_2) dx_2 = \int_{x_1}^1 6(1 - x_2) dx_2 = -3(1 - x_2)^2 \Big|_{x_2=x_1}^{x_2=1} = 0 - (-3(1 - x_1)^2) = 3(1 - x_1)^2$$

Now we get:

$$f_{X_1}(x_1) = \begin{cases} 3(1 - x_1)^2 & 0 \leq x_1 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Then for $E[X_1]$ integrate:

$$\begin{aligned}
E[X_1] &= \int_{-\infty}^{\infty} x_1 f_{X_1}(x_1) dx_1 \\
&= \int_0^1 x_1 3(1-x_1)^2 dx_1 \\
&= 3 \int_0^1 x_1(1-x_1)^2 dx_1 \\
&= 3 \int_0^1 (x_1 - 2x_1^2 + x_1^3) dx_1 \\
&= 3 \left(\frac{1}{2}x_1^2 - \frac{2}{3}x_1^3 + \frac{1}{4}x_1^4 \right) \Big|_{x_1=0}^{x_1=1} \\
&= 3 \left(\frac{1}{2} - \frac{2}{3} + \frac{1}{4} \right) \\
&= 3 \left(\frac{6}{12} - \frac{8}{12} + \frac{3}{12} \right) \\
&= 3 \left(\frac{1}{12} \right) \\
&= \frac{1}{4}
\end{aligned}$$

That's one down. Next, let's marginalize to get $f_{X_2, X_3}(x_2, x_3)$ and again to get $f_{X_2}(x_2)$.

$$f_{X_2, X_3}(x_2, x_3) = \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) dx_1 = \int_0^{x_2} 6 dx_1 = 6x_1 \Big|_{x_1=0}^{x_1=x_2} = 6x_2$$

Being careful with the bounds we get a new PDF:

$$f_{X_2, X_3}(x_2, x_3) = \begin{cases} 6x_2 & 0 \leq x_2 \leq x_3 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Rolling up:

$$f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_{X_2, X_3}(x_2, x_3) dx_3 = \int_{x_2}^1 6x_2 dx_3 = 6x_2 x_3 \Big|_{x_3=x_2}^{x_3=1} = 6x_2 - 6x_2^2 = 6x_2(1-x_2)$$

Now we get:

$$f_{X_2}(x_2) = \begin{cases} 6x_2(1-x_2) & 0 \leq x_2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Then for $E[X_2]$ integrate:

$$\begin{aligned}
E[X_2] &= \int_{-\infty}^{\infty} x_2 f_{X_2}(x_2) dx_2 \\
&= \int_0^1 x_2 (6x_2 - 6x_2^2) dx_2 \\
&= 6 \int_0^1 (x_2^2 - x_2^3) dx_2 \\
&= \left(\frac{6}{3} x_2^3 - \frac{6}{4} x_2^4 \right) \Big|_{x_2=0}^{x_2=1} \\
&= \frac{6}{3} - \frac{6}{4} \\
&= 2 - \frac{3}{2} \\
&= \frac{1}{2}
\end{aligned}$$

That's two down. Finally, we marginalize to get $f_{X_1, X_3}(x_1, x_3)$ and again to get $f_{X_3}(x_3)$.

$$f_{X_1, X_3}(x_1, x_3) = \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) dx_2 = \int_{x_1}^{x_3} 6 dx_2 = 6x_2 \Big|_{x_2=x_1}^{x_2=x_3} = 6(x_3 - x_1)$$

Being careful with the bounds we get a new PDF:

$$f_{X_1, X_3}(x_1, x_3) = \begin{cases} 6(x_3 - x_1) & 0 \leq x_1 \leq x_3 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Rolling up:

$$\begin{aligned}
f_{X_3}(x_3) &= \int_{-\infty}^{\infty} f_{X_1, X_3}(x_1, x_3) dx_1 = \int_0^{x_3} 6(x_3 - x_1) dx_1 \\
&= (6x_3 x_1 - 3x_1^2) \Big|_{x_1=0}^{x_1=x_3} = 6x_3^2 - 3x_3^2 = 3x_3^2
\end{aligned}$$

Now we get:

$$f_{X_3}(x_3) = \begin{cases} 3x_3^2 & 0 \leq x_3 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Then for $E[X_3]$ integrate:

$$\begin{aligned}
E[X_3] &= \int_{-\infty}^{\infty} x_3 f_{X_3}(x_3) dx_3 \\
&= \int_0^1 x_3 3x_3^2 dx_3 \\
&= 3 \int_0^1 x_3^3 dx_3 \\
&= \frac{3}{4} x_3^4 \Big|_{x_3=0}^{x_3=1} \\
&= \frac{3}{4}
\end{aligned}$$

Thus our expected value vector is:

$$E[\mathbf{X}] = [E[X_1] \quad E[X_2] \quad E[X_3]]^T = \left[\frac{1}{4} \quad \frac{1}{2} \quad \frac{3}{4} \right]^T$$

Next question: what is the covariance matrix $R_{\mathbf{X}} = E[\mathbf{X}\mathbf{X}^T]$? We'll need to find:

- $E[X_1X_1], E[X_1X_2], E[X_1X_3]$
- $E[X_2X_2], E[X_2X_3]$
- $E[X_3X_3]$

First up:

$$\begin{aligned}
 E[X_1X_1] &= \int_0^1 x_1^2 f_{X_1}(x_1) dx_1 \\
 &= \int_0^1 x_1^2 3(1-x_1)^2 dx_1 && \text{plugging in what we found earlier} \\
 &= 3 \int_0^1 x_1^2 (1-x_1)^2 dx_1 \\
 &= 3 \int_0^1 (x_1^2 - 2x_1^3 + x_1^4) dx_1 \\
 &= 3 \left(\frac{1}{3}x_1^3 - \frac{1}{2}x_1^4 + \frac{1}{5}x_1^5 \right) \Big|_{x_1=0}^{x_1=1} \\
 &= 3 \left(\frac{1}{3} - \frac{1}{2} + \frac{1}{5} \right) \\
 &= 3 \left(\frac{10}{30} - \frac{15}{30} + \frac{6}{30} \right) \\
 &= 3 \left(\frac{1}{30} \right) \\
 &= \frac{1}{10}
 \end{aligned}$$

Then:

$$\begin{aligned}
E[X_1 X_2] &= \int_0^1 \int_{x_2=x_1}^1 x_1 x_2 f_{X_1, X_2}(x_1, x_2) dx_2 dx_1 \\
&= \int_0^1 \int_{x_1}^1 x_1 x_2, 6(1 - x_2) dx_2 dx_1 \\
&= 6 \int_0^1 x_1 \int_{x_1}^1 (x_2 - x_2^2) dx_2 dx_1 \\
&= 6 \int_0^1 x_1 \left(\frac{1}{2} x_2^2 - \frac{1}{3} x_2^3 \right) \Big|_{x_2=x_1}^{x_2=1} dx_1 \\
&= 6 \int_0^1 x_1 \left(\frac{1}{2} - \frac{1}{3} - \frac{1}{2} x_1^2 + \frac{1}{3} x_1^3 \right) dx_1 \\
&= 6 \int_0^1 \left(\frac{1}{6} x_1 - \frac{1}{2} x_1^3 + \frac{1}{3} x_1^4 \right) dx_1 \\
&= 6 \left(\frac{1}{12} x_1^2 - \frac{1}{8} x_1^4 + \frac{1}{15} x_1^5 \right) \Big|_{x_1=0}^{x_1=1} \\
&= 6 \left(\frac{1}{12} - \frac{1}{8} + \frac{1}{15} \right) \\
&= 6 \left(\frac{10}{120} - \frac{15}{120} + \frac{8}{120} \right) \\
&= 6 \left(\frac{3}{120} \right) \\
&= \frac{3}{20}
\end{aligned}$$

Next:

$$\begin{aligned}
E[X_1 X_3] &= \int_0^1 \int_{x_3=x_1}^1 x_1 x_3 f_{X_1, X_3}(x_1, x_3) dx_3 dx_1 \\
&= \int_0^1 \int_{x_1}^1 x_1 x_3, 6(x_3 - x_1) dx_3 dx_1 \\
&= 6 \int_0^1 x_1 \int_{x_1}^1 (x_3^2 - x_1 x_3) dx_3 dx_1 \\
&= 6 \int_0^1 x_1 \left(\frac{1}{3} x_3^3 - \frac{x_1}{2} x_3^2 \right) \Big|_{x_3=x_1}^{x_3=1} dx_1 \\
&= 6 \int_0^1 x_1 \left(\frac{1}{3} - \frac{x_1}{2} - \frac{1}{3} x_1^3 + \frac{1}{2} x_1^3 \right) dx_1 \\
&= 6 \int_0^1 \left(\frac{1}{3} x_1 - \frac{1}{2} x_1^2 + \frac{1}{6} x_1^4 \right) dx_1 \\
&= 6 \left(\frac{1}{6} x_1^2 - \frac{1}{6} x_1^3 + \frac{1}{30} x_1^5 \right) \Big|_{x_1=0}^{x_1=1} \\
&= 6 \left(\frac{1}{6} - \frac{1}{6} + \frac{1}{30} \right) \\
&= \frac{1}{5}
\end{aligned}$$

That's it for the X_1 cases. Now for $E[X_2X_2]$:

$$\begin{aligned}
E[X_2X_2] &= \int_0^1 x_2^2 f_{X_2}(x_2) dx_2 \\
&= \int_0^1 x_2^2, 6x_2(1-x_2) dx_2 \\
&= 6 \int_0^1 (x_2^3 - x_2^4) dx_2 \\
&= \left(\frac{6}{4}x_2^4 - \frac{6}{5}x_2^5 \right) \Big|_{x_2=0}^{x_2=1} \\
&= \frac{6}{4} - \frac{6}{5} = \frac{15}{10} - \frac{12}{10} \\
&= \frac{3}{10}
\end{aligned}$$

Then:

$$\begin{aligned}
E[X_2X_3] &= \int_0^1 \int_{x_3=x_2}^1 x_2x_3 f_{X_2,X_3}(x_2, x_3) dx_3 dx_2 \\
&= \int_0^1 \int_{x_2}^1 x_2x_3, 6x_2 dx_3 dx_2 \\
&= 6 \int_0^1 x_2^2 \int_{x_2}^1 x_3 dx_3 dx_2 \\
&= 6 \int_0^1 x_2^2 \left(\frac{1}{2}x_3^2 \right) \Big|_{x_3=x_2}^{x_3=1} dx_2 \\
&= 3 \int_0^1 x_2^2(1-x_2^2) dx_2 \\
&= 3 \int_0^1 (x_2^2 - x_2^4) dx_2 \\
&= 3 \left(\frac{1}{3}x_2^3 - \frac{1}{5}x_2^5 \right) \Big|_{x_2=0}^{x_2=1} \\
&= 3 \left(\frac{1}{3} - \frac{1}{5} \right) = 3 \left(\frac{5}{15} - \frac{3}{15} \right) = \frac{5}{5} - \frac{3}{5} \\
&= \frac{2}{5}
\end{aligned}$$

Finally:

$$\begin{aligned}
E[X_3X_3] &= \int_0^1 x_3^2 f_{X_3}(x_3) dx_3 \\
&= \int_0^1 x_3^2, 3x_3^2 dx_3 \\
&= 3 \int_0^1 x_3^4 dx_3 \\
&= \frac{3}{5}x_3^5 \Big|_{x_3=0}^{x_3=1} \\
&= \frac{3}{5}
\end{aligned}$$

Multiplication is commutative, so $E[X_i X_j] = E[X_j X_i]$ and we have all the entries we need to complete the correlation matrix (aka second-moment matrix):

$$R_{\mathbf{X}} = E[\mathbf{X}\mathbf{X}^T] = \begin{bmatrix} E[X_1 X_1] & E[X_1 X_2] & E[X_1 X_3] \\ E[X_2 X_1] & E[X_2 X_2] & E[X_2 X_3] \\ E[X_3 X_1] & E[X_3 X_2] & E[X_3 X_3] \end{bmatrix} = \begin{bmatrix} \frac{1}{10} & \frac{3}{20} & \frac{1}{5} \\ \frac{3}{20} & \frac{3}{10} & \frac{2}{5} \\ \frac{1}{5} & \frac{2}{5} & \frac{3}{5} \end{bmatrix}$$

One last problem: find the covariance matrix $C_{\mathbf{X}}$.

$$\begin{aligned} C_{\mathbf{X}} &= R_{\mathbf{X}} - \mu_{\mathbf{X}} \mu_{\mathbf{X}}^T \\ &= \begin{bmatrix} \frac{1}{10} & \frac{3}{20} & \frac{1}{5} \\ \frac{3}{20} & \frac{3}{10} & \frac{2}{5} \\ \frac{1}{5} & \frac{2}{5} & \frac{3}{5} \end{bmatrix} - \begin{bmatrix} \frac{1}{4} \\ \frac{1}{2} \\ \frac{3}{4} \end{bmatrix} \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{3}{4} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{10} & \frac{3}{20} & \frac{1}{5} \\ \frac{3}{20} & \frac{3}{10} & \frac{2}{5} \\ \frac{1}{5} & \frac{2}{5} & \frac{3}{5} \end{bmatrix} - \begin{bmatrix} \frac{1}{16} & \frac{1}{8} & \frac{3}{16} \\ \frac{1}{8} & \frac{1}{4} & \frac{3}{8} \\ \frac{3}{16} & \frac{3}{8} & \frac{9}{16} \end{bmatrix} \\ &= \begin{bmatrix} \frac{8}{80} - \frac{5}{80} & \frac{6}{40} - \frac{5}{40} & \frac{16}{80} - \frac{15}{80} \\ \frac{6}{40} - \frac{5}{40} & \frac{6}{20} - \frac{5}{20} & \frac{16}{40} - \frac{15}{40} \\ \frac{16}{80} - \frac{15}{80} & \frac{16}{40} - \frac{15}{40} & \frac{48}{80} - \frac{45}{80} \end{bmatrix} \\ &= \begin{bmatrix} \frac{3}{80} & \frac{1}{40} & \frac{1}{80} \\ \frac{1}{40} & \frac{1}{20} & \frac{1}{40} \\ \frac{1}{80} & \frac{1}{40} & \frac{3}{80} \end{bmatrix} \end{aligned}$$

Thus our covariance matrix is:

$$C_{\mathbf{X}} = \begin{bmatrix} \frac{3}{80} & \frac{1}{40} & \frac{1}{80} \\ \frac{1}{40} & \frac{1}{20} & \frac{1}{40} \\ \frac{1}{80} & \frac{1}{40} & \frac{3}{80} \end{bmatrix}$$

6 Stochastic Processes

Quoting Yates and Goodman:

When we study random variables, each observation corresponds to one or more numbers. When we study stochastic processes, each observation corresponds to a function of time.

6.1 Definitions and Notation

A **stochastic process** $X(t)$ consists of an experiment with a probability measure $P[\cdot]$ defined on a sample space Ω and a function that assigns a sample function $x(t, \omega)$ to each outcome ω in Ω .

The **ensemble** of a stochastic process is the set of all possible time functions that can result from an experiment.

The stochastic process $X(t)$ may be:

- continuous-time, continuous-value
- continuous-time, discrete-value
- discrete-time, continuous-value
- discrete-time, discrete-value

A **random sequence** X_n is an ordered sequence of random variables $X_0, X_1, \dots$

So for example, if our stochastic process $X(t)$ consists of repeated rolls of a fair six-sided die where t refers to the roll number, then we have a discrete-time, discrete-value stochastic process. In this case, each roll is independent and identically distributed (iid), so $P_{X(0)}(x) = P_{X(1)}(x) = P_{X(2)}(x) = \dots$

A **counting process** $N(t)$ starts at time 0 and counts occurrences of events, often referred to as *arrivals*. Starting at time $t = 0$ also means the sample function $n(t, \omega)$ must equal 0 for any $t \leq 0$. Since we are counting, each value is a natural number, and each increase is by one, i.e., we don't skip counts or allow tied arrival times.

6.2 Poisson Processes

A counting process $N(t)$ is a **Poisson process** of rate λ if the number of arrivals in any interval $(t_0, t_1]$ is a Poisson random variable having parameter (and expected value) $\lambda(t_1 - t_0)$ and for any pair of non-overlapping intervals the number of arrivals in each interval are independent random variables. The interarrival times of a Poisson processes are Exponential(λ) and therefore “memoryless”, i.e., no matter how long we have waited for an arrival, the remaining time until the arrival remains an exponential λ random variable.

As a counting process, our initial condition is $N(t) = 0$ for all $t \leq 0$. The time of the first arrival is X_1 and each $X_2, X_3, \dots$ are subsequent arrival times, so in general the random time X_n between arrival $n - 1$ and arrival n is called the n th interarrival time.

This λ stays constant and is used for each arrival. In other words, we have a single and common experiment defined on a common sample space. One interarrival time is distributed using the same settings as any other interarrival time. This λ parameter may be referred to as the intensity or arrival rate.

Stepping through an example:

1. We start at $t = 0$, there are not yet any arrivals, so $N(0) = 0$.
2. $N(t)$ remains zero until our first arrival.
3. Our first arrival happens at $X_1 = x_1$, so $N(x_1) = 1$.
4. $N(t)$ remains one until our next arrival.
5. We have an interarrival time of $X_2 = x_2$ happening at $t = x_1 + x_2$, so $N(x_1 + x_2) = 2$.
6. Next we have an interarrival time of $X_3 = x_3$ happening at $t = x_1 + x_2 + x_3$, so $N(t) = 3$.
7. ...and so forth. Thus, arrival n occurs at time $t = x_1 + \dots + x_n$.

Let's create a Poisson process and step through the first five arrivals. Each interarrival time is exponentially distributed; let's choose $\lambda = 1$ for our example. The R built-in for this is `rexp`. We'll visualize it using a step plot following the same convention that we put the closed circles on the closed left sides (plotting code not shown).

```
set.seed(2026)

# start at zero, then add exponentially distributed interarrivals
dfPoissonProcess <- data.frame(x = c(0, rexp(5, 1)))
```

```

# t is the cumulative sum of the interarrival times
dfPoissonProcess$t <- cumsum(dfPoissonProcess$x)

# add a t_end for plotting purposes
dfPoissonProcess$t_end <- c(
  dfPoissonProcess$t[-1],
  tail(dfPoissonProcess$t, 1) + diff(tail(dfPoissonProcess$t, 2))
)

# start at zero, then show N(t) at each arrival time
dfPoissonProcess$N <- 0:(nrow(dfPoissonProcess)-1)

```

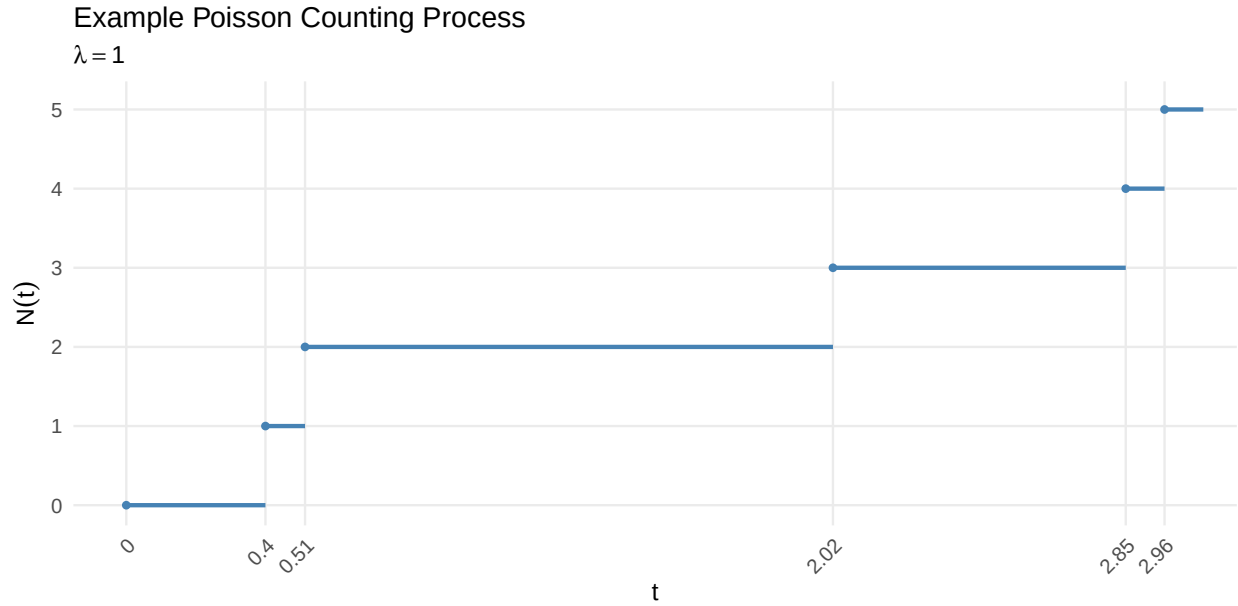


Figure 16: Example Poisson Process

Recall that a Poisson random variable has these properties for rate $\lambda \in \mathbb{R}_+$:

Poisson

$$\begin{aligned}
 P_X(x) &= \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!} & x = 0, 1, 2, \dots \\ 0 & \text{otherwise} \end{cases} \\
 E[X] &= \lambda
 \end{aligned}$$

Since $N(t) \sim \text{Poisson}(\lambda t)$ we have $E[N(1)] = \lambda$, then $E[N(2)] = 2\lambda$, and so forth. Thus $E[N(t)] = \lambda t$ for $t \geq 0$.

We can express a PMF for a Poisson process if we let Poisson random variable $M = N(t_1) - N(t_0)$ where t_1 is some time after t_0 . $M = N(t_1) - N(t_0) \implies M \sim \text{Poisson}(\lambda(t_1 - t_0))$, meaning increments over that interval $t_1 - t_0$ are Poisson distributed with mean $\lambda(t_1 - t_0)$. Plugging in $\lambda(t_1 - t_0)$ for the Poisson parameter in the Poisson random variable PMF we get this PMF:

$$P_M(m) = \begin{cases} \frac{e^{-\lambda(t_1 - t_0)} [\lambda(t_1 - t_0)]^m}{m!} & m = 0, 1, \dots \\ 0 & \text{otherwise} \end{cases}$$

If we let $T = t_1 - t_0$ we can re-write this as:

$$P_M(m) = \begin{cases} \frac{(\lambda T)^m}{m!} e^{-\lambda T} & m = 0, 1, \dots \\ 0 & \text{otherwise} \end{cases}$$

From the PMF and the definition of a Poisson process we can immediately find the CDF for the interarrival times $X_1, X_2, \dots$ of a Poisson process of rate λ :

Theorem 6.1. Poisson process interarrival times CDF

$$F_{X_n}(x) = P[X_n \leq x] = 1 - P[X_n > x] = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Proof. Given $X_1 = x_1, X_2 = x_2, \dots, X_{n-1} = x_{n-1}$, arrival $n-1$ occurs at time $t_{n-1} = x_1 + \dots + x_{n-1}$. Then for some $x > 0$ the n th interarrival time X_n is greater than x if and only if there are no arrivals in the interval $(t_{n-1}, t_{n-1} + x]$. By definition of a Poisson process, the number of arrivals in the interval $(t_{n-1}, t_{n-1} + x]$ is independent of the past history in that non-overlapping interval described by $X_1, \dots, X_{n-1}$. Thus we can write:

$$\begin{aligned} P[X_n > x \mid X_1 = x_1, \dots, X_{n-1} = x_{n-1}] &= P[N(t_{n-1} + x) - N(t_{n-1}) = 0] \\ &= \frac{e^{-\lambda x} (\lambda x)^0}{0!} && \text{PMF when } m=0 \\ &= e^{-\lambda x} \end{aligned}$$

Then we plug in $P[X_n \leq x] = 1 - P[X_n > x] = 1 - e^{-\lambda x}$ to get the CDF. □

Having the CDF, we can quickly find our PDF for the interarrival times $X_1, X_2, \dots$ by taking the derivative of the CDF, yielding:

Theorem 6.2. Poisson process interarrival times PDF

$$f_{X_n}(x) = f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

So to recap:

- start at $t = 0$ with initial $N(0) = 0$
- X_n is the time between arrivals $n-1$ and n
- $X_1, X_2, \dots$ are $\sim \text{iidExponential}(\lambda)$
- $N(t_1) - N(t_0) \sim \text{Poisson}(\lambda(t_1 - t_0))$
- $P[X_n > x] = 1 - e^{-\lambda x}$ for $x \geq 0$
- CDF $F_{X_n}(x) = 1 - P[X_n > x] = 1 - e^{-\lambda x}$ for $x \geq 0$
- PDF $f_{X_n}(x) = \lambda e^{-\lambda x}$ for $x \geq 0$

7 Markov Chains

The **Markov property** is a modeling assumption that future state depends only on current state. This property allows us to construct **Markov chains**, the first type of which is the **Discrete-Time Markov Chain**.

7.1 Discrete-Time Markov Chains

Let Ω be a finite or countably infinite set where for each pair i, j in Ω there is assigned a non-negative real number p_{ij} satisfying the constraint that the sum over j of p_{ij} equals 1:

$$\sum_{j \in \Omega} p_{ij} = 1$$

Let $X_n = X_1, X_2, \dots$ be a sequence of random variables whose ranges are each contained in Ω . Each ω in Ω is a **state**, and our p_{ij} numbers give us the **transition probabilities** from state i to state j , i.e., $p_{ij} = P[X_{n+1} = j \mid X_n = i]$. The sequence X_n is a Markov chain if

$$P[X_{n+1} = j \mid X_1 = i_1, \dots, X_n = i_n] = P[X_{n+1} = j \mid X_n = i_n] = p_{i_n j}$$

for every n and every sequence $i_1, \dots, i_n$ for which $P[X_1 = i_1, \dots, X_n = i_n] > 0$.

7.1.1 Finite Example

For instance, say we have a set Ω of three states of a home office printer “Working”, “Paper Jam”, and “Out of Ink”, we can represent our state transition probabilities in a matrix $\mathbf{P}$ where each row totals to 1, for example with the following probabilities:

$$\mathbf{P} = \begin{matrix} & \begin{matrix} j=1 & j=2 & j=3 \end{matrix} \\ \begin{matrix} i=1 \\ i=2 \\ i=3 \end{matrix} & \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} \end{matrix} = \begin{matrix} & \begin{matrix} \text{Working} & \text{Paper Jam} & \text{Out of Ink} \end{matrix} \\ \begin{matrix} \text{Working} \\ \text{Paper Jam} \\ \text{Out of Ink} \end{matrix} & \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.6 & 0.3 & 0.1 \\ 0.5 & 0.3 & 0.2 \end{bmatrix} \end{matrix}$$

The above example matrix $\mathbf{P}$ indicates for state 1 (i.e., $i = 1$, the “Working” row) that the probabilities are 0.7 of staying in state 1 (i.e., $j = 1$, the “Working” column), 0.2 of transitioning to state 2, and 0.1 of moving to state 3. Or for instance looking in the second row, $P[X_{n+1} = \text{Working} \mid X_n = \text{Paper Jam}] = p_{21} = 0.6$.

A matrix having the properties that each element is non-negative and each row sums to 1 is known as a **right-stochastic matrix**, often just called a stochastic matrix.

Some sources will use P_{ij} to mean the element in the i th row and j th column of matrix $\mathbf{P}$. Thus, $P_{ij} = p_{ij}$.

In R we might use the built-in `matrix` functionality to represent the transition matrix (which creates a two-dimensional array):

```
# create the transition matrix P
P <- matrix(
  c(0.7, 0.2, 0.1,
    0.6, 0.3, 0.1,
    0.5, 0.3, 0.2),
  nrow=3, byrow=TRUE)
```

The transition matrix $\mathbf{P}$ can be represented as a graph where the states are graphs and the transition probabilities arrows between nodes.

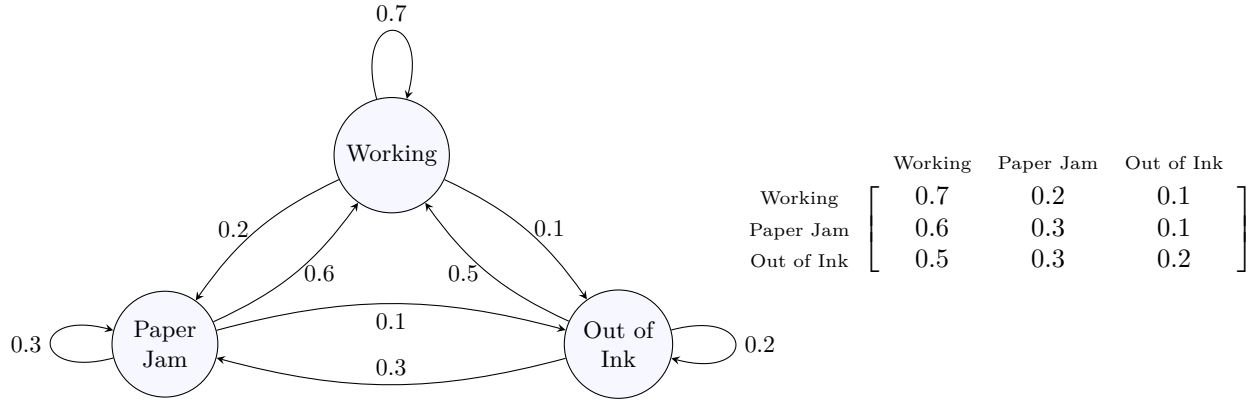


Figure 17: Example Markov Chain

We can seed the system by specifying the vector $\mathbf{p}(0)$ of initial probabilities. Let's set $\mathbf{p}(0) = [1 \ 0 \ 0]^T$ so that we always start in state 1.

```
# create the initial state vector p0^T (p0 transposed)
p0T <- matrix(c(1.0, 0.0, 0.0), nrow=1)
```

So after one step, our probabilities for each of the three states are given by the vector $\mathbf{p}(1) = \mathbf{p}(0)^T \mathbf{P}$ which we calculate by applying the matrix multiplication operator `%%`:

```
p0T %% P
```

```
##      [,1] [,2] [,3]
## [1,] 0.7 0.2 0.1
```

Likewise, after two steps, our probabilities for each of the three states are given by the vector $\mathbf{p}(2) = \mathbf{p}(1)^T \mathbf{P} = \mathbf{p}(0)^T \mathbf{P}^2$:

```
p0T %% P %% P
```

```
##      [,1] [,2] [,3]
## [1,] 0.66 0.23 0.11
```

The properties of the system are entirely determined by these transition probabilities and initial probabilities. The value of X_n summarizes all the past history of the system, and X_{n+1} depends only on the X_n . The **n-step transition probabilities** are given by the matrix $\mathbf{P}(n)$ which has elements:

$$P_{ij}(n) = p_{ij}^{(n)} = P[X_{m+n} = j \mid X_m = i]$$

The n th-order transition probabilities may be derived using the Chapman-Kolmogorov equations—for a finite Markov chain, the n -step transition probabilities satisfy:

$$p_{ij}^{(n+m)} = \sum_{k \in \Omega} p_{ik}^{(n)} p_{kj}^{(m)}, \quad n, m \geq 0$$

In matrix form:

$$\mathbf{P}(n+m) = \mathbf{P}(n)\mathbf{P}(m)$$

We have $\mathbf{P}(n) = \mathbf{P}^n$, so for example $\mathbf{P}(n+1) = \mathbf{P}(n)\mathbf{P}$. Note that that $\mathbf{P}(0) = \mathbf{P}^0 = \mathbf{I}$ (the identity matrix). Because $\mathbf{P}(n) = \mathbf{P}^n$ we also have:

$$\mathbf{P}^{n+m} = \mathbf{P}^n \mathbf{P}^m$$

So for instance, if we wanted to find the probabilities for being in each of the three states after 12 steps we would apply matrix multiplication 12 times. There's no built-in matrix exponentiation operator in R, but we can simply use `rep` to write the multiplication 12 times. We can use `fractions` from the MASS library to display the final result. We'll also define a helper function to print the $\text{\LaTeX}$.

```
library(MASS, quietly=TRUE) # for fractions()

# helper function to build LaTeX matrix rows
mat_to_latex<-function(M) {
  M_str<-apply(M, 1, function(x) {
    paste(
      paste(x, collapse = "&"),
      "\\\\"
    )
  })
  return(M_str)
}

# find the 12th step (no matrix power operator for this in Base R)
p12 = MASS::fractions(
  matrix(
    eval(
      parse(
        text=paste("p0T %%% P",
                    paste(rep("%% P", 12),
                          collapse=" "),
                          collapse=" ")
      )
    ), nrow=1)
)
```

To display the result we need to use `results='asis'` in the RMarkdown chunk. We did our work using a row vector, but by convention these are column vectors, hence we write $\mathbf{p}^T(12)$.

```
# collect the LaTeX code into a block, otherwise between
# writeLines and RMarkdown the codes can be escaped inconsistently.
begin<-"\\begin{bmatrix}"
end<-"\\end{bmatrix}"
block<-c("$",
  "\\mathbf{p}^\\mathsf{T}(12) = \\mathbf{p}^\\mathsf{T}(0) \\mathbf{P}^{12} = ",
  begin, mat_to_latex(as.character(p12)), end,
  "$")
writeLines(block)
```

$$\mathbf{p}^T(12) = \mathbf{p}^T(0)\mathbf{P}^{12} = [53/81 \quad 19/81 \quad 1/9]$$

For a finite Markov chain with initial probability vector $\mathbf{p}(0)$, the **limiting state probabilities**, when they exist, are defined to be the vector:

$$\boldsymbol{\pi} = \lim_{n \rightarrow \infty} \mathbf{p}(n)$$

If we do have a limiting state probability vector π , then:

$$\pi^T = \pi^T \mathbf{P}$$

Was the $\mathbf{p}^T(12)$ we calculated earlier a limiting state probability vector?

```
# print = or != depending on whether left hand side equals right hand side
check <- function(statement_true, lhs, rhs) {
  relation <- if (statement_true) "=" else "\\ne"
  cat("\\[" , lhs, relation, rhs, "\\]\\n")
}

check(
  isTRUE(all.equal(p12, fractions(p12 %*% P))),
  "\\mathbf{p}^\\mathsf{T}(12)",
  "\\mathbf{p}^\\mathsf{T}(12) \\mathbf{P}"
)
```

$$\mathbf{p}^T(12) = \mathbf{p}^T(12)\mathbf{P}$$

Somewhere at or prior to the 12th iteration we reached a limiting state. How might we find the limiting state by hand? We want to find $\pi = [\pi_1 \ \pi_2 \ \pi_3]^T$, so we could set $\pi^T = \pi^T \mathbf{P}$ subject to the constraint that $\pi_1 + \pi_2 + \pi_3 = 1$. So we have:

$$\begin{aligned} \pi_1 &= .7\pi_1 + .6\pi_2 + .5\pi_3 \\ \pi_2 &= .2\pi_1 + .3\pi_2 + .3\pi_3 \\ \pi_3 &= .1\pi_1 + .1\pi_2 + .2\pi_3 \\ 1 &= \pi_1 + \pi_2 + \pi_3 \end{aligned}$$

Rewriting this system (and multiplying those first three by 10 for readability) we have:

$$\begin{aligned} 3\pi_1 - 6\pi_2 - 5\pi_3 &= 0 \\ -2\pi_1 + 7\pi_2 - 3\pi_3 &= 0 \\ -\pi_1 - \pi_2 + 8\pi_3 &= 0 \\ \pi_1 + \pi_2 + \pi_3 &= 1 \end{aligned}$$

Using the third equation to write π_1 in terms of π_2 and π_3 :

$$\pi_1 + \pi_2 = 8\pi_3 \implies \pi_1 = 8\pi_3 - \pi_2$$

Substituting this into the second equation gives:

$$\begin{aligned} -2(8\pi_3 - \pi_2) + 7\pi_2 - 3\pi_3 &= 0 \implies -16\pi_3 + 9\pi_2 - 3\pi_3 = 0 \\ &\implies 9\pi_2 - 19\pi_3 = 0 \\ &\implies 9\pi_2 = 19\pi_3 \\ &\implies \pi_2 = \frac{19}{9}\pi_3 \end{aligned}$$

Substituting this back into our expression for π_1 :

$$\pi_1 = 8\pi_3 - \frac{19}{9}\pi_3 = \frac{53}{9}\pi_3$$

Using the constraint $\pi_1 + \pi_2 + \pi_3 = 1$:

$$\begin{aligned}\frac{53}{9}\pi_3 + \frac{19}{9}\pi_3 + \pi_3 &= 1 \implies \frac{81}{9}\pi_3 = 1 \\ \implies \pi_3 &= \frac{9}{81} = \frac{1}{9}\end{aligned}$$

Substituting back into the expression for π_2 :

$$\pi_2 = \frac{19}{9} \left(\frac{1}{9} \right) = \frac{19}{81}$$

Likewise for π_1 :

$$\pi_1 = \frac{53}{9} \left(\frac{1}{9} \right) = \frac{53}{81}$$

The limiting state probabilities are therefore:

$$\boldsymbol{\pi} = [\pi_1 \quad \pi_2 \quad \pi_3]^\top = \left[\frac{53}{81} \quad \frac{19}{81} \quad \frac{1}{9} \right]^\top$$

7.1.2 Countably Infinite Example

The same ideas apply when the state space is countably infinite rather than finite. Consider the state space $\Omega = \{1, 2, 3, \dots\}$. Suppose that when the system is in state $i > 1$, it moves to the next higher state (“moves to the right”) with probability p , and moves to the next lower state (“moves to the left”) with probability $1 - p$. At state 1, there is no state 0, so we instead let the chain remain at 1 with probability $1 - p$ (i.e., we are bounded on one side).

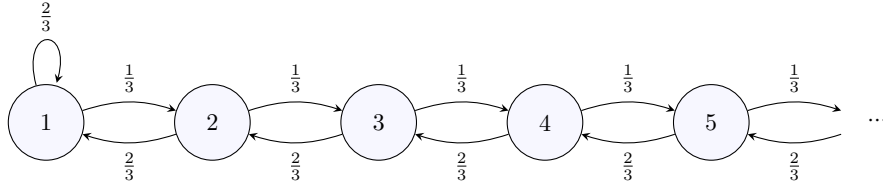


Figure 18: Example countably infinite Markov chain

For a concrete example, let $p = \frac{1}{3}$. The chain moves to the right with probability $1/3$ and to the left with probability $2/3$. Unlike our finite example, there is now no finite transition matrix. We can nevertheless display the first of the infinite transition matrix:

$$\mathbf{P} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & \dots \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ \vdots \end{matrix} & \left[\begin{array}{cccccc} 2/3 & 1/3 & 0 & 0 & 0 & \dots \\ 2/3 & 0 & 1/3 & 0 & 0 & \dots \\ 0 & 2/3 & 0 & 1/3 & 0 & \dots \\ 0 & 0 & 2/3 & 0 & 1/3 & \dots \\ 0 & 0 & 0 & 2/3 & 0 & \ddots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \end{array} \right] \end{matrix}$$

There is no final state, but every row still sums to one, just as in the finite example.

As before, we can specify an initial state. Suppose that initially the system is in state 1 with probability one:

$$\mathbf{p}(0) = [1 \quad 0 \quad 0 \quad 0 \quad \dots]^\top$$

After one step we have:

$$\mathbf{p}(1) = \mathbf{p}^T(0)\mathbf{P} = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} & 0 & 0 & \dots \end{bmatrix}^T$$

After two steps:

$$\mathbf{p}(2) = \mathbf{p}^T(1)\mathbf{P} = \begin{bmatrix} \frac{2}{3}(\frac{2}{3}) + \frac{1}{3}(\frac{2}{3}) & \frac{2}{3}(\frac{1}{3}) & \frac{1}{3}(\frac{1}{3}) & 0 & \dots \end{bmatrix}^T = \begin{bmatrix} \frac{2}{3} & \frac{2}{9} & \frac{1}{9} & 0 & \dots \end{bmatrix}^T$$

This chain has limiting probabilities $\boldsymbol{\pi} = [\pi_1 \quad \pi_2 \quad \pi_3 \quad \dots]^T$.

We know that the probabilities must sum to 1, so we have:

$$\sum_{j=1}^{\infty} \pi_j = 1$$

For state 1, there are two ways to be in state 1 after one step: we can already be in state 1 and remain there, or we can be in state 2 and move to the left. Therefore:

$$\pi_1 = \frac{2}{3}\pi_1 + \frac{2}{3}\pi_2$$

Combining and rearranging:

$$\frac{1}{3}\pi_1 = \frac{2}{3}\pi_2$$

Thus:

$$\pi_1 = 2\pi_2 \implies \pi_2 = \frac{1}{2}\pi_1$$

Then for subsequent elements we can arrive at state j either by moving right from $j-1$ or by moving left from $j+1$. Thus:

$$\pi_j = \frac{1}{3}\pi_{j-1} + \frac{2}{3}\pi_{j+1}, \quad j = 2, 3, \dots$$

Plugging that in for π_3 :

$$\begin{aligned} \pi_3 &= \frac{3}{2}\pi_2 - \frac{1}{2}\pi_1 \\ &= \frac{3}{2}\left(\frac{1}{2}\pi_1\right) - \frac{1}{2}\pi_1 && \text{substituting} \\ &= \frac{1}{4}\pi_1 \end{aligned}$$

Likewise for π_4 :

$$\begin{aligned} \pi_4 &= \frac{3}{2}\pi_3 - \frac{1}{2}\pi_2 \\ &= \frac{3}{2}\left(\frac{1}{4}\pi_1\right) - \frac{1}{2}\left(\frac{1}{2}\pi_1\right) && \text{substituting} \\ &= \frac{1}{8}\pi_1 \end{aligned}$$

Does this geometric pattern continue for every subsequent j ? In other words, is it true that:

$$\pi_j = \frac{1}{2^j}\pi_1, \quad j = 2, 3, \dots$$

Let's rearrange what we have so far to put in terms of $j+1$:

$$\begin{aligned} \pi_j &= \frac{1}{3}\pi_{j-1} + \frac{2}{3}\pi_{j+1} \quad j = 2, 3, \dots \\ \implies \frac{2}{3}\pi_{j+1} &= \pi_j - \frac{1}{3}\pi_{j-1} \quad j = 2, 3, \dots \\ \implies \pi_{j+1} &= \frac{3}{2}\pi_j - \frac{1}{2}\pi_{j-1} \quad j = 2, 3, \dots \end{aligned}$$

Define a difference—if it's true that the geometric pattern continues, then this difference will always be zero:

$$d_j = \pi_j - \frac{1}{2}\pi_{j-1}, \quad j = 2, 3, \dots$$

For $j = 2$ we already have that the difference is zero:

$$\begin{aligned} d_2 &= \pi_2 - \frac{1}{2}\pi_1 \\ &= \frac{1}{2}\pi_1 - \frac{1}{2}\pi_1 && \text{plug in } \pi_2 = \frac{1}{2}\pi_1 \\ &= 0 \end{aligned}$$

We can see that for $j + 1$ our difference is the same as j :

$$\begin{aligned} d_{j+1} &= \pi_{j+1} - \frac{1}{2}\pi_j \\ &= \left(\frac{3}{2}\pi_j - \frac{1}{2}\pi_{j-1} \right) - \frac{1}{2}\pi_j \\ &= \pi_j - \frac{1}{2}\pi_{j-1} \\ &= d_j \end{aligned}$$

So $d_{2+1} = d_2 = 0$, $d_{3+1} = d_3 = 0$, and so forth; by induction we've established that $\pi_j = \frac{1}{2}\pi_{j-1}$ for $j = 2, 3, \dots$. We can rewrite the relation to put π_j in terms of π_1 :

$$\begin{aligned} \pi_j &= \frac{1}{2}\pi_{j-1} \quad j = 2, 3, \dots \\ \implies \pi_j &= \frac{1}{2^{j-1}}\pi_1 \quad j = 1, 2, \dots \end{aligned}$$

Finally we put the $\sum_{j=1}^{\infty} \pi_j = 1$ to work:

$$\begin{aligned} \sum_{j=1}^{\infty} \pi_j &= 1 \\ \sum_{j=1}^{\infty} \frac{1}{2^{j-1}}\pi_1 &= 1 \\ \pi_1 \sum_{j=1}^{\infty} \frac{1}{2^{j-1}} &= 1 \\ \pi_1 \sum_{j=1}^{\infty} \left(\frac{1}{2} \right)^{j-1} &= 1 \end{aligned}$$

We recognize a geometric series with a ratio r between terms of $\frac{1}{2}$, so $\sum_{j=1}^{\infty} \left(\frac{1}{2} \right)^{j-1} = \frac{1}{1-\frac{1}{2}} = 2$, and we find:

$$\pi_1 \cdot 2 = 1 \implies \pi_1 = \frac{1}{2}$$

Plugging that in we get:

$$\pi_j = \left(\frac{1}{2} \right) \left(\frac{1}{2} \right)^{j-1} = \left(\frac{1}{2} \right)^j$$

We can display our first few elements of $\boldsymbol{\pi}^\top$ using the R on the sequence $1:\mathbf{N}$ for some $\mathbf{N}$ of our choosing:

```

N <- 8
block<-c("$$",
  "\\sympf{\\pi} = ",
  begin,
  " \\, ",
  paste(
    paste(
      MASS::fractions( (0.5)^(1:N) ),
      collapse=" & ",
      " & \\dots \\, " ),
    end,
    "\\mathsf{T}",
    "$$")
writeLines(block)

```

$$\pi = [1/2 \quad 1/4 \quad 1/8 \quad 1/16 \quad 1/32 \quad 1/64 \quad 1/128 \quad 1/256 \quad \dots]^\top$$

7.2 Structure and Properties

With Markov chains we can inherit some properties from graph theory and linear algebra. If a weighted directed graph G is a pair (V, E) of a set of vertices $V = \{v_1, v_2, \dots\}$ and set of edges $E = \{(v_i, v_j) \mid v_i, v_j \in V\}$ together with a set of non-negative weights $W = \{w_{ij} \in \mathbb{R} \mid w_{ij} \geq 0 \wedge (v_i, v_j) \in E\}$ then we can see how a finite Markov chain is a special case in which, for each vertex v_i , the weights of all outgoing edges sum to one:

$$\sum_{j \mid (v_i, v_j) \in E} w_{ij} = 1$$

Thus the “states” of a Markov chain correspond to the “nodes” of a graph and the “transition probabilities” correspond to graph “weights”.

Some definitions:

State i has **period** d if d is the largest natural number such that $P_{ij}(n) = 0$ whenever n is not divisible by d . A state is **aperiodic** if $d = 1$. To illustrate, Figure 19 has $d = 2$. Furthermore, this chain has $P_{ij}(n) = 0$ when n is not divisible by 2.

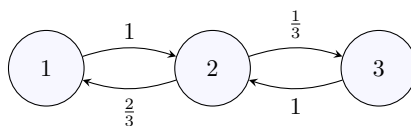


Figure 19: Example Markov chain having period $d = 2$

State j is **accessible** from state i when $P_{ij}(n) > 0$ for some $n > 0$. Read $i \rightarrow j$ as “ j is accessible from i ” and $i \nrightarrow j$ as “ j is not accessible from i ”.

States i and j **communicate** when $i \rightarrow j$ and $j \rightarrow i$. Read $i \leftrightarrow j$ as “ i communicates with j ”. A state always communicates with itself since in zero steps we are still at that state.

A **communicating class** is a nonempty subset of states C such that for $i \in C, j \in C \iff i \leftrightarrow j$. For example, in Figure 20 we have two communicating classes: C_1 and C_2 .

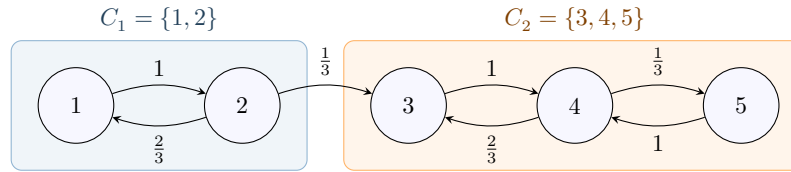


Figure 20: Example Markov chain having two communicating classes

Figure 20 also illustrates **transient** versus **recurrent** states. A state i of a finite Markov chain is transient if there exists a state j such that $i \rightarrow j$ and $j \nrightarrow i$. We see that structure here in that $v_2 \rightarrow v_3$ and $v_3 \nrightarrow v_2$, meaning that v_2 is transient. Likewise, $v_1 \rightarrow v_3$ and $v_3 \nrightarrow v_1$ (if one state of a communicating class is transient, all states of that communicating class are transient). The states on class C_2 are all recurrent.

In a finite Markov chain, not *all* classes can be transient—there must be *at least one* recurrent state (and thus at least one recurrent class). If there are multiple recurrent classes, then in some sense each of these recurrent classes a separate sub-system.

A Markov chain is **irreducible** when there is only one communicating class.

7.2.1 Irreducible Markov Chains

An irreducible Markov chain has some interesting properties vis-à-vis matrix decomposition. If $\mathbf{P}$ is the transition matrix of an irreducible Markov chain, then:

- $\mathbf{P}$ will have $\lambda = 1$ as one eigenvalue
- all other eigenvalues will have $|\lambda| \leq 1$
- the presence of $|\lambda| = 1$ eigenvalues (other than the guaranteed $\lambda = 1$ eigenvalue) indicates persistent oscillations/cycles
- the left eigenvector that goes with the $\lambda = 1$ eigenvalue can be normalized to be equal to the stationary distribution π

We can try this out in R using the built-in `eigen` to get the eigenvalues and eigenvectors. First, let's built a two-state chain that oscillates between state 1 and 2. We should get one $\lambda = 1$ and another $\lambda = -1$.

```
# create a 2x2 transition probability matrix having an oscillation
P = matrix(
  c(
    0, 1,
    1, 0
  ),
  nrow=2, byrow=TRUE)

# use R to get the eigenvalues and eigenvectors
e <- eigen(P)

# display our two eigenvalues
e$values
```

```
## [1] 1 -1
```

We get two $|\lambda| = 1$ eigenvalues for our $d = 2$ irreducible Markov chain.

Similarly, a 3-state cycle:

```

# create a 3x3 transition probability matrix having a cycle
P = matrix(
  c(
    0, 1, 0,
    0, 0, 1,
    1, 0, 0
  ),
  nrow=3, byrow=TRUE)

# use R to get the eigenvalues and eigenvectors
e <- eigen(P)

# display the three eigenvalues
e$values

```

```
## [1] -0.5+0.8660254i -0.5-0.8660254i 1.0+0.0000000i
```

We see the guaranteed $\lambda = 1$. For the two eigenvalues that have imaginary components, is $|\lambda| = 1$?

```
abs(e$values)
```

```
## [1] 1 1 1
```

Thus we find three $|\lambda| = 1$ eigenvalues for our $d = 3$ irreducible Markov chain. It's worth plotting ([Figure 21](#)). First we use the R built-ins `Re` and `Im` to extract the real and imaginary parts of the eigenvalues, saving to a `data.frame` for plotting.

```

eig_df <- data.frame(
  real = Re(e$values),
  imag = Im(e$values)
)

```

The imaginary parts make more sense if we print the square root part rather than leave as decimals—a quick helper function should do the trick.

```

ToRadicalChr <- function(x, tol = 1e-10) {
  sapply(x, function(x) {

    if (!is.finite(x))
      return(as.character(x))

    if (abs(x) < tol)
      return("0")

    neg <- x < 0
    ax <- abs(x)

    # hard-code some known values
    special <- c(
      "0" = 0,
      "1" = 1,
      "(sqrt(5)-1)/4" = (sqrt(5) - 1) / 4,

```

```

"(sqrt(5)+1)/4" = (sqrt(5) + 1) / 4,
"sqrt(10+2*sqrt(5))/4" = sqrt(10 + 2 * sqrt(5)) / 4,
"sqrt(10-2*sqrt(5))/4" = sqrt(10 - 2 * sqrt(5)) / 4
)

# match known forms
for (label in names(special)) {
  if (abs(ax - special[[label]]) < tol) {
    return(if (neg) paste0("-", label) else label)
  }
}

# generic sqrt(n) / d search
for (d in 1:20) {
  n <- ax^2 * d^2
  n0 <- round(n)

  if (abs(n - n0) < tol * max(1, n0)) {
    result <- if (n0 == 1) {
      if (d == 1) "1" else paste0("1/", d)
    } else {
      paste0("sqrt(", n0, ")/", d)
    }

    return(if (neg) paste0("-", result) else result)
  }
}

as.character(x)
}, USE.NAMES = FALSE)
}

```

We can plot the eigenvalues over the unit circle; surprisingly, there's no ggplot built-in for circles, but we can use `geom_function` to plot top and bottom.

```

p1 <- ggplot(eig_df, aes(x = real, y = imag)) +
  geom_function(
    fun = function(x) sqrt(1 - x^2),
    xlim = c(-1, 1),
    color = "grey70",
    linewidth = 0.8
  ) +
  geom_function(
    fun = function(x) -sqrt(1 - x^2),
    xlim = c(-1, 1),
    color = "grey70",
    linewidth = 0.8
  ) +
  geom_point(
    color = "steelblue",
    fill = "steelblue",
    shape = 21,
    size = 2
  )

```

```

) +
coord_equal(xlim = c(-1.15, 1.15), ylim = c(-1.15, 1.15)) +
scale_x_continuous(
  breaks = sort(unique(c(0, eig_df$real))),
  labels = MASS::fractions(sort(unique(c(0, eig_df$real))))
) +
scale_y_continuous(
  breaks = sort(unique(eig_df$imag)),
  labels = parse(text = ToRadicalChr(sort(unique(eig_df$imag))))
) +
labs(
  x = expression(Re(lambda)),
  y = expression(Im(lambda))
) +
theme_minimal(base_size = 5) +
theme(panel.grid.minor.x = element_blank(),
      panel.grid.minor.y = element_blank())

ggsave(
  "markov_chain_eigenvals_d3.pdf",
  plot = p1,
  dpi = 300,
  width = 2.5,
  height = 1.1)

```

The `ggsave` call is so we can embed the image via $\text{\LaTeX}$. Finally, we output the plot (Figure 21) next to the transition probability matrix $\mathbf{P}$.

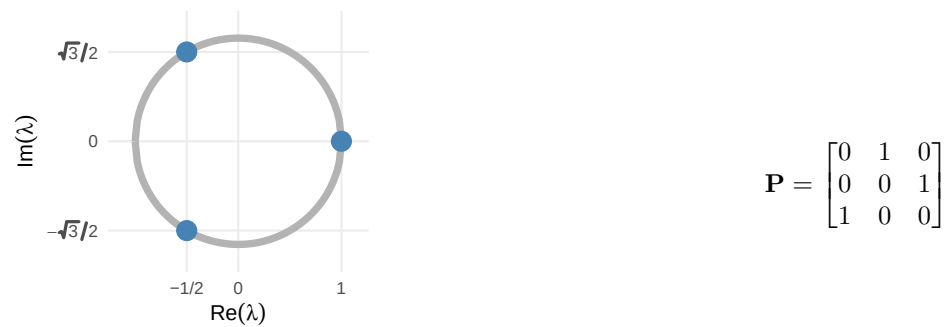
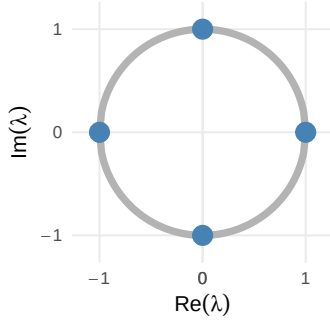


Figure 21: Three $|\lambda| = 1$ eigenvalues for a $d = 3$ irreducible Markov chain

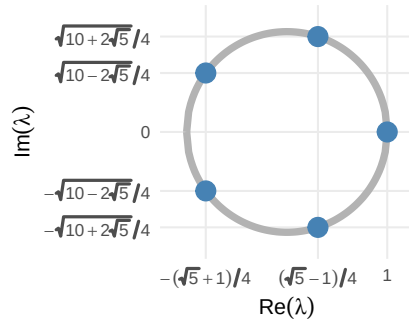
Likewise, we can plot a $d = 4$ case (code not shown).



$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Figure 22: Four $|\lambda| = 1$ eigenvalues for a $d = 4$ irreducible Markov chain

Now a $d = 5$ case (code not shown).



$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure 23: Four $|\lambda| = 1$ eigenvalues for a $d = 5$ irreducible Markov chain

Note that the d distinct $|\lambda| = 1$ eigenvalues for these periodic irreducible finite Markov chains are given by:

$$\lambda_1 = 1, \lambda_2 = e^{\frac{2\pi i}{d}}, \lambda_3 = e^{\frac{4\pi i}{d}}, \lambda_4 = e^{\frac{6\pi i}{d}}, \dots$$

Or to put it another way, if we let $\Phi = e^{\frac{2\pi i}{d}}$, then the $|\lambda| = 1$ eigenvalues are:

$$\lambda_1 = \Phi^0, \lambda_2 = \Phi^1, \lambda_3 = \Phi^2, \lambda_4 = \Phi^3, \dots, \lambda_n = \Phi^{n-1}$$

Now to try out the property that relates the eigenvalues and eigenvectors to the stable distribution $\boldsymbol{\pi}$. Let's reuse that example Markov chain where we already calculated the stable distribution to be $\begin{bmatrix} \frac{53}{81} & \frac{19}{81} & \frac{1}{9} \end{bmatrix}^T$. We want the *left* eigenvectors, so we use the R built-in transpose function `t` to get $\mathbf{P}^T$, then pass that to `eigen` to get the eigenvectors and eigenvalues. Will we get the same results?

```
# create the transition probability matrix
P = matrix(
  c(0.7, 0.2, 0.1,
    0.6, 0.3, 0.1,
    0.5, 0.3, 0.2),
  nrow=3, byrow=TRUE)

# find the eigenvector for the λ=1 eigenvalue
eT <- eigen(t(P))
idx <- which.min(abs(eT$values - 1))
```

```
pi_unnorm <- Re(eT$eigenvectors[, idx])

# normalize the eigenvector for the  $\lambda=1$  eigenvalue
pi_norm <- pi_unnorm / sum(pi_unnorm)

# display as fractions
MASS::fractions(pi_norm)
```

```
## [1] 53/81 19/81 1/9
```

They match.

8 Source Notes

8.1 Notation

Notation conventions used by different sources are given in Table 1.

Source	Sample Space	Event Space	Probability	Expectation
Kolmogorov (1933)	E	$\mathfrak{F}$	$P(\cdot)$	$E(X)$
Hoel/Port/Stone (1971)	Ω	$\mathcal{A}$	$P(\cdot)$	EX
Ross (1972)	S		$P(\cdot)$	$E[X]$
Casella/Berger (1990)	S	$\mathcal{B}$	$P(\cdot)$	EX
Billingsley (1995)	Ω	$\mathcal{F}$	$P(\cdot)$	$E[X]$
Yates/Goodman (1998)	S		$P[\cdot]$	$E[X]$
Bertsekas/Tsitsiklis (2002)	Ω		$\mathbf{P}[\cdot]$	$\mathbf{E}[X]$

Table 1: Notation Conventions by Source

8.2 Quotes

(1718) *The Doctrine of Chances*

The first edition of *The Doctrine of Chances* by Abraham de Moivre was published in 1718 in English. Some quotes:

The Probability of an Event is greater, or less, according to the number of Chances by which it may Happen, compared with the number of all Chances, by which it may either Happen or Fail.

The Expectation of obtaining any Thing, is estimated by the Value of that Thing multiplied by the Probability of obtaining it.

Permutations are the Changes which several things can receive in the different Orders in which they may be placed... Combinations are the various Conjunctions which several things may receive without any respect to

Order...

The text is full of gaming-related problems with their solutions. It's a classic for Classical Probability.

(1933) *Grundbegriffe der Wahrscheinlichkeitsrechnung*

Andrey Kolmogorov's first edition of *Grundbegriffe der Wahrscheinlichkeitsrechnung* was published in 1933 in German; later a second edition was published in English in 1956 under the title *Foundations of the Theory of Probability*. His name is transliterated as Kolmogoroff in the original German language publication.

Sample spaces, events, the probability measure P , and the three axioms are all present in the 1933 edition.

The theory of probability, as a mathematical discipline, can and should be developed from axioms in exactly the same way as Geometry and Algebra.

Die Wahrscheinlichkeitstheorie als mathematische Disziplin soll und kann genau in demselben Sinne axiomatisiert werden wie die Geometrie oder die Algebra.

(1971) *Introduction to Probability Theory*

The 1971 book *Introduction to Probability Theory* by Paul Hoel, Sidney Port, and Charles Stone is the first volume of a 3-volume series.

For the mathematical theory of probability the interpretation of probabilities is irrelevant, just as in geometry the interpretation of points, lines, and planes is irrelevant.

(1972) *Introduction to Probability Models*

The first edition of Shelon Ross' *Introduction to Probability Models* was published in 1972; the thirteenth edition was published in 2024.

From the second edition:

The following proposition, known as the "law of the unconscious statistician," shows how we can calculate the expectation of $g(X)$ without first determining the distribution of $g(X)$. [...] The law got its name from "unconscious" statisticians who have used it as if it were the definition of $E[g(X)]$.

Quoting from the 2024 edition:

Any realistic model of a real-world phenomenon must take into account the possibility of randomness. That is, more often than not, the quantities we are interested in will not be predictable in advance but, rather, will exhibit an inherent variation that should be taken into account by the model.

(1976) *A First Course in Probability*

The first edition of Shelon Ross' *A First Course in Probability* was published in 1976; subsequent editions were published in 1984, 1988, 1994, 1998, 2002, 2006, 2010, 2014, and 2019 with a "global" adaptation of the 10th being published in 2020.

(1995) *Probability and Measure*

Patrick Billingsley's *Probability and Measure* was first published in 1995; the third edition dates to 2011.

The expected value of a [real-valued] random variable X on $(\Omega, \mathcal{F}, P)$ is the integral of X with respect to the measure P

(1998) *Probability and Stochastic Processes*

Authors Roy Yates and David Goodman have a second edition published in 2005 and a third in 2014.

Quotes from the second edition:

The mathematical basis of probability is the theory of sets.

The sample space of an experiment is the finest-grain, mutually exclusive, collectively exhaustive set of all possible outcomes.

Everyone accepts that the probability of an event is a number between 0 and 1. Some people interpret probability as a physical property (like mass or volume or temperature) that can be measured. [...] Another interpretation of probability relates to the knowledge that we have about something. [...] Both views are useful when we apply probability theory to practical problems.

The mathematics of probability begins when the situation is so complex that we just can't replicate everything important exactly... In this case, repetitions of the same procedure yield different results.

(2002) *Statistical Inference 2nd edition*

Quotes from the second edition *Statistical Inference* by George Casella and Roger Berger:

...the axiomatic approach is not concerned with the interpretations of probabilities, but is concerned only that the probabilities are defined by a function satisfying the axioms.

The subject of probability theory is the foundation upon which all of statistics is built, providing a means for modeling populations, experiments, or almost anything else that could be considered a random phenomenon.

Aside from the mean, $E X$, of a random variable, perhaps the most important moment is the second central moment, more commonly known as the variance.