Probability Lectures and Labs 2023 Edition

Mark Huber

Probability: Lectures and Labs

2023 Edition

Mark Huber

Version 2023-11-02

ISBN: 9798674518877

Cover art:
Image by Yuri_B from Pixabay

Contents

Preface

Purpose   This book covers a one semester course in probability for students who have had the traditional sequence of single variable through multivariable Calculus and a course in linear algebra. It is intended to prepare students for advanced work in probability. For example, students who are intending to go on to take mathematical statistics, stochastic processes, or any other advanced probability course.

The structure of the course is designed to be 2/3 traditional lecture, and 1/3 inquiry based learning, where students complete labs primarily on their own to gain intuition about the nature of various probability laws and concepts.

Organization   Part I is the bulk of the course, containing the main ideas, concepts, and theorems of probability. Part III collects some extra useful facts about probability and distributions, and can be integrated into the course as needed. Part IV contains some background information about sets, logic, functions, combinatorics, and integration and students have often forgotten before taking their first mathematical probability course. Finally, Part IV contains worked solutions to many of the exercises in the text found at the end of each chapter.

Summary   In Part I, I present a summary of the ideas in each chapter at the beginning of the chapter. The first time you read this this summary will probably not make much sense, but when you understand the material of the chapter, the words and notation in the summary should be understood completely. So a good way to check that you are understanding a chapter is to go back and reread the summary.

My approach   When I teach the course, I leave everything in Part III to be read as needed by students, and start lecturing immediately with Part I of the text. But whether you start with Chapter 1 or a later chapter, Part III should serve as a valuable reference for students as they delve into Part I.

The course as I teach it has three meeting sessions per week. I alternate between lectures on Monday and Friday, followed by a lab from Part II where students generate random variables and study their properties on Wednesday. These lab exercises are implemented using R.

The classroom I teach in has a computer available for every student, but most students prefer to bring their own laptop. The labs are structured as a main lab followed by an extended lab. Students who finish the main lab before the class session is up are required in my course to then complete the extended lab, as the time to finish varies considerably between students based on their familiarity with computers.

Formulas and Definitions

$\int_{s \in A} f(s)\, d\# = \sum_{s \in A} f(s)$

$\mathbb{E}[g(X)] = \int_s g(s) f_X(s)\, d\mu$

$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB)$

$f_X(x) = \int_y f_{(X,Y)}(x, y)$

$\mathrm{cdf}_X(a) = F_X(a) = \mathbb{P}(X \leq a)$ $\mathbb{V}(A) = \mathbb{E}[A^2] - \mathbb{E}[A]^2$

$f_{A|B=b}(a) \propto f_A(a) f_{B|A=a}(b)$

$\mathrm{mgf}_X(t) = \mathbb{E}[e^{tX}]$

$\mathrm{gf}_X(s) = \mathbb{E}(s^X)$

$\mathrm{Cor}(A, B) = \dfrac{\mathrm{Cov}(A, B)}{\mathrm{SD}(A)\,\mathrm{SD}(B)}$

$\mathrm{MAD}(X) = \mathbb{E}[|X - \mathbb{E}[X]|]$

$\mathrm{Cov}(A, B) = \mathbb{E}[AB] - \mathbb{E}[A]\mathbb{E}[B]$

$\mathbb{P}(A|B) = \dfrac{\mathbb{P}(AB)}{\mathbb{P}(B)}$

$\mathbb{P}(X \in A) = \int_{s \in A} f_X(s)\, d\mu$

Distributions (quick guide)

Uniform random variables

X	$f_X(s)$
$\mathsf{Unif}(A)$	$\mathbb{1}(s \in A)/\mu(A)$

Standard random variables

X	$f_X(s)$	$\mathbb{E}[X]$	$\mathbb{V}(X)$
$\mathsf{Unif}([0,1])$	$\mathbb{1}(s \in [0,1])$	$1/2$	$1/12$
$\mathsf{Exp}(1)$	$\exp(-s)\mathbb{1}(s \geq 0)$	1	1
$\mathsf{N}(0,1)$	$\tau^{-1/2}\exp(-s^2/2)$	0	1
$\mathsf{Cauchy}(0)$	$\frac{2}{\tau} \cdot \frac{1}{s^2+1}$	does not exist	does not exist

Shifting and scaling the standard random variables

X	From standard	$f_X(s)$	$\mathbb{E}[X]$	$\mathbb{V}(X)$
$\mathsf{Unif}([a,b])$	$(b-a)U + a$	$\mathbb{1}(s \in [a,b])/(b-a)$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$
$\mathsf{Exp}(\lambda)$	T/λ	$\lambda\exp(-\lambda s)\mathbb{1}(s \geq 0)$	$1/\lambda$	$1/\lambda^2$
$\mathsf{N}(\mu,\sigma^2)$	$\mu + \sigma Z$	$(\sigma^2\tau)^{-1/2}\exp\left(-\frac{(s-\mu)^2}{2\sigma^2}\right)$	μ	σ

More common random variables

X	$f_X(s)$	$\mathbb{E}[X]$	$\mathbb{V}(X)$
$\mathsf{Bern}(p)$	$p\mathbb{1}(s=1) + (1-p)\mathbb{1}(s=0)$	p	$p(1-p)$
$\mathsf{Bin}(n,p)$	$\binom{n}{i}p^i(1-p)^{n-1}\mathbb{1}(i \in \{0,\ldots,n\})$	np	$np(1-p)$
$\mathsf{Geo}(p)$	$p(1-p)^{i-1}\mathbb{1}(i \in \{1,2,\ldots\})$	$1/p$	$(1-p)/p^2$
$\mathsf{NegBin}(r,p)$	$\binom{i-1}{r-1}p^r(1-p)^{i-r}\mathbb{1}(i \in \{1,2,\ldots\})$	r/p	$r(1-p)/p^2$
$\mathsf{Gamma}(n,\lambda)$	$\lambda^n s^{n-i}\exp(-\lambda s)\Gamma(n)^{-1}\mathbb{1}(s \geq 0)$	n/λ	n/λ^2
$\mathsf{Pois}(\mu)$	$\frac{\exp(-\mu)\mu^i}{i!}\mathbb{1}(i \in \{0,1,2,\ldots\}$	μ	μ
$\mathsf{Beta}(a,b)$	$s^{a-1}(1-s)^{b-1}\mathbb{1}(s \in [0,1])$	$a/(a+b)$	$\frac{ab}{(a+b)^2(a+b+1)}$

Part I

PROBABILITY

Chapter 1

What is probability?

> **Question of the Day** Suppose that I know that it will either rain today, snow today, or neither. The chance of rain is 30%, the chance of snow is 15%. What is the chance that neither happens?

> **Summary** The mathematics of partial information is called **probability**, and is used to make calculations about the chance that certain outcomes occur.

Sometimes everything is known about a mathematical model. For instance, we might know that the length of a room is exactly 30 feet or the pressure in a tank is exactly 112 psi. In other cases, we might only have partial information. We might know that the length of a room is somewhere between 20 and 40 feet, or the pressure is at most 120 in the tank. Then we need a way to model the length and pressure that handles the fact that we only have partial information about the true value.

Consider the stock market. The value of the stock market is known today, and that gives us some partial information about what the stock market will be tomorrow. Because tomorrow's market has not happened yet, we do not have complete information, and so we need a way to model the incomplete information that we do have.

Another way that partial information can come about is when it is possible to get total information, but that would be too expensive. For instance, it does not make sense for a survey team to ask every person in the United States what their preferred candidate for president is. Instead, if a small sample from the population is chosen randomly and then surveyed, that only gives partial information about the preferences of everyone.

It could also be that the goal is to understand random physical processes that are too complicated to have a complete model. Flipping a coin, throwing dice, or shuffling of cards are all examples of physical processes that can lead to a lack of information. In fact, this type of randomness is often the first kind that people come into contact. This leads some people to think that randomness equals physical randomness. However, that is far from true, and physical systems are only a special case of the notion that randomness means partial information.

The branch of mathematics that deals with partial information is called *probability*. This word comes from the Latin *probabilis* which means plausible, and first appeared when talking about the uncertainty of evidence in court cases. Probabilities are still used in court today, and are also are used to model everything from how many friends will attend a party to the future state of the planet. It is essential to be able to calculate correctly with partial information, and to combines sources of uncertainty correctly. That is what this course is all about.

Here the blackboard boldface capital letter P, written $\mathbb{P}$, is used to denote the probability that an event occurs. For instance, if I think the probability that it will rain today is 30%, I would write $\mathbb{P}(\text{rain}) = 0.3$.

Another term for probability is chance, a word that comes from the Latin *cadere*, which means cases. In the example there two cases: either it rains, or it does not rain. In modern parlance, say that the *outcomes* are $\{\text{rain}, \text{no rain}\}$.

A *space* is the name for a set that is special in some way. If you have not seen sets before, now is a good time to check out Chapter 43. For probability, call the set of outcomes that have probabilities associated with events the *outcome space*.

> **Definition 1**
> The **outcome space** is the set of possible outcomes when we have complete information about something. This is also sometimes called the **sample space** or **state space**.

> **Notation 1**
> Often Ω (the capital Greek letter Omega) or $\mathcal{S}$ is used to denote the outcome space.

1.1   Probability as a measure

In mathematics, a *measure* tells us the size of a set. *Lebesgue* measure is one of the most commonly used measures. In one dimension, Lebesgue measure is what we think of as length. For example, the Lebesgue measure of the interval $[3, 7]$ is the length of the interval, or $7 - 3 = 4$.

In two dimensions, Lebesgue measure is the area of a region. The Lebesgue measure of the unit circle is the area of the unit circle, which is $(1/2)\tau = \pi$. Here $\pi = 3.141\ldots$ is the *half circle constant*, the length halfway around a circle of radius 1, and $\tau = 6.283\ldots$ is the *full circle constant*, the length all the way around a circle of radius 1.

Probabilities are also a measure, but they measure the more abstract notion of the chance that the event occurs. For instance, with the question of the day, the outcome space is $\Omega = \{r, s, n\}$ where r stands for rain, s stands for snow, and n stands for neither.

> **Notation 2**
> When writing the probability of a set of outcomes of size 1, it is common to omit the curly braces around the set. For instance,
>
> $$\mathbb{P}(r) = \mathbb{P}(\{r\}).$$

Since probability is a measure, it shares certain properties that all measures have. The length of a stick is an example of a measure, and has the following properties.

1. The length of a stick is nonnegative.

2. If I break a stick into two disjoint pieces, the lengths of the two pieces add to the original length of the stick.
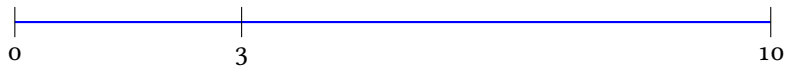
$$m([0, 10]) = m([0, 3) \cup [3, 10]) = m([0, 3)) + m([3, 10]) = 3 + 7 = 10.$$

For probabilities, these two ideas translate as

1. The probability that an event occurs is always nonnegative.

2. If I have two events A and B that are disjoint (so $A \cap B = \emptyset$), then the probability that at least one of the events occurs equals the sum of the probabilities that they occur.

We can also write these ideas mathematically. For a measure m,

1. For any measurable set A, $m(A) \geq 0$.

2. For any measurable sets A and B such that $A \cap B = \emptyset$, then $m(A \cup B) = m(A) + m(B)$.

When using counting measure, if I break up the set into two smaller sets that do not overlap, again the overall size is the sum of the sizes of the smaller sets.

$$\#(\{a, b, c, d, e\}) = \#(\{a, b\} \cup \{c, d, e\}) = \#(\{a, b\}) + \#(\{c, d, e\}) = 2 + 3 = 5.$$

For probabilities, this means

$$(\forall A, B : AB = \emptyset)(\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)).$$

In fact, this works for any finite number of sets, not just 2. Nonoverlapping sets come up often enough that we give this property a special name.

> **Definition 2**
> Say that sets A and B are **disjoint** if $A \cap B = \emptyset$. A collection $\{A_\alpha\}$ of sets is **disjoint** if every pair of sets in the collection are disjoint.

1.2 Total probability

Probabilities have an extra property that not all measures have. The total amount of probability that we work with is $100\% = 1$. In words, the probability that something happens is 1. Mathematically, for outcome space Ω,

$$\mathbb{P}(\Omega) = 1.$$

There is nothing special about the constant 1 here: we could have set it to any positive number and built a theory of probability around it. Historically, using 1 as the total probability seemed most natural to mathematicians, and so 1 is used as the total probability constant by everyone today.

Solving the question of the day   With these properties, we can now solve the question of the day:

$$\mathbb{P}(\{r, s, n\}) = \mathbb{P}(\{r\} \cup \{s\} \cup \{n\}) = 30\% + 15\% + \mathbb{P}(\{n\}) = 1,$$

hence $\mathbb{P}(\{n\}) = 1 - 0.3 - .15 = \boxed{55\%}$.

1.3 Zeno's paradox

Zeno had several paradoxes, but in the most famous a line is broken in half, then the right half is broken in half, and so on. This breaking in half is done an infinite number of times, and still the sum of the measures of the pieces equals the overall measure of the set.
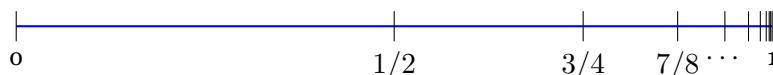
$$m([0,1]) = m([0,1/2)) + m([1/2,3/4)) + m([3/4,7/8)) + \cdots$$
$$1 = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots$$

In other words, for a measure we want the additivity property to hold not just for a finite collection of disjoint sets, but even for an infinite sequence of disjoint sets.

1.4   Formal definition of probability

Now let us make the discussion of the last session precise by giving a formal definition of probability.

> **Definition 3**
> Let $\mathbb{P}$ have outcome space Ω. If $\mathbb{P}(A)$ is defined for a set $A \subseteq \Omega$, then call A **measurable** or an **event**.

Saying that A is measurable means that $\mathbb{P}(A)$ is defined. Then we want to make sure the probability that the outcome is not in A is also defined. In other words $\mathbb{P}(A^C)$ where A^C is the complement of A should also be defined.

Similarly, if $A_1, A_2, \ldots$ are a sequence of subsets of Ω, and $\mathbb{P}(A_i)$ is defined for each one, we would also like the probability of the union $\cup_{i=1}^{\infty} A_i$ to be defined. By this infinite union we mean

$$\cup_{i=1}^{\infty} A_i = \{a : (\exists i)(a \in A_i)\}.$$

Hence we make the following definition.

> **Definition 4**
> Let $\mathcal{F}$ be a collection of subsets of Ω such that
>
>   1. Closed under complements: If $A \in \mathcal{F}$, then $A^C \in \mathcal{F}$.
>
>   2. Closed under countable unions: If $A_1, A_2, \ldots$ are all in $\mathcal{F}$ then
>
> $$\cup_{i=1}^{\infty} A_i \in \mathcal{F}.$$
>
> Call $\mathcal{F}$ a σ**-algebra**, and the elements of $\mathcal{F}$ **measurable sets**.

A couple notes:

1. The symbol σ is the Greek letter sigma, so the term σ-algebra is read out loud as "sigma-algebra".

2. A synonym for σ-algebra is σ-field.

3. We often refer to the elements of $\mathcal{F}$ as *measurable sets* because they are the sets that we assign probabilities to.

4. Whenever Ω is a finite set, $\mathcal{F}$ will be the set of all subsets of Ω.

5. When $\Omega = \mathbb{R}$, the σ-algebra $\mathcal{F}$ is usually taken to be the *Borel sets*. In this first course in probability we will not define the Borel sets precisely, except to note that any interval (open or closed, finite or infinite) is a member of the Borel sets.

Example 1

If $[0,3]$ and $[2,9]$ are measurable sets, show that $[0,9]$ is as well.

Answer. Measurable sets are closed under countable unions (which includes finite unions). So because $[0,3]$ and $[2,9]$ are measurable, so is

$$[0,3] \cup [2,9] = [0,9].$$

Definition 5

A function $\mathbb{P} : \mathcal{F} \to [0,1]$ is a **probability distribution** if the following hold.

1. Total probability: $\Omega \in \mathcal{F}$ and $\mathbb{P}(\Omega) = 1$.

2. Countable additivity: If $A_1, A_2, \ldots$ are all subsets of Ω so that for all $i \neq j$, $A_i A_j = \emptyset$, then
$$\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Note: sometimes mathematicians will call these the axioms of probability instead of a definition.

Example 2

Suppose $\mathbb{P}(i) = (1/3)^{|i|}$ for $i \in \{1, 2, \ldots\} \cup \{-1, -2, \ldots\}$. What is $\mathbb{P}(\{1, 2, 3, \ldots\})$?

Answer Using our rule for probability measures

$$\mathbb{P}(\{1, 2, \ldots\}) = \mathbb{P}(\{1\} \cup \{2\} \cup \{3\} \cup \cdots)$$
$$= \sum_{i=1}^{\infty} \mathbb{P}(\{i\})$$
$$= \sum_{i=1}^{\infty} \left(\frac{1}{3}\right)^i$$
$$= \frac{1/3}{1 - 1/3} = \frac{1}{3 - 1} = \boxed{0.5000}.$$

Note the the numerical result in this example was given to *4 significant figures*, or *4 sig figs* for short. For numbers such as 0.5, the 4 sig fig version 0.5000 is identical, but for numbers like $\sqrt{2}$, the 4 sig fig version 1.414 will be slightly different. However, the relative error between the two versions will be at most 0.01%, which is a small enough error to fall below most needs in practice, while increasing readability drastically. Therefore, throughout this text, 4 sig figs will be the standard way to approximate numerical results.

Problems

1.1 For $A \in \mathcal{F}$, find $\mathbb{P}(A \cup A^C)$.

1.2 Find $\mathbb{P}((A \cup B \cup C) \cup (A \cup B \cup C)^C)$.

1.3 Prove that if the state space Ω is measurable, so is $\emptyset$.

1.4 What is
$$[0, 1/2) \cup [1/2, 3/4) \cup [3/4, 7/8) \cup \cdots?$$

1.5 If $[0, 1 - 1/n]$ is measurable for every $n \geq 2$, show that the interval $[0, 1)$ is measurable.

1.6 If $\{i\}$ is measurable for every positive integer i, show that $\{2, 4, 6, \ldots\}$ is measurable.

1.7 A *partition* of a set Ω is a collection of sets that are disjoint whose union is Ω. Suppose A, B, and C partition Ω. What is $\mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C)$?

1.8 Show that for the Borel sets the set
$$[0, 1] \cup [2, 3] \cup [4, 5] \cup \cdots$$
is measurable.

1.9 Suppose for $i \in \{0, 1, 2, \ldots\}$,
$$\mathbb{P}([i, i + 1)) = (1/3)^{i+1}.$$
What is $\mathbb{P}([0, \infty)$?

1.10 Suppose for $i \in \{1, 2, 3 \ldots\}$, $\mathbb{P}(i) = (1/4)^i$. What is
$$\mathbb{P}(\{1, 2, 3, \ldots\})?$$

Properties of probability distributions

Question of the Day Given that $\mathbb{P}(A_1) = 0.2$, $\mathbb{P}(A_2) = 0.9$ and $\mathbb{P}(A_1 A_2) = 0.15$, what is $\mathbb{P}(A_1 \cup A_2)$?

Summary We can derive several properties of probability distributions from the basic definition.

$$\mathbb{P}(\emptyset) = 0 \qquad \text{Empty set probability}$$
$$\mathbb{P}(A^C) = 1 - \mathbb{P}(A) \qquad \text{Complements}$$
$$A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B) \qquad \text{Monotonicity}$$
$$\mathbb{P}(A) = 1 \Rightarrow (\forall B)(\mathbb{P}(B) = \mathbb{P}(AB)) \qquad \text{Certain events}$$
$$\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i) \qquad \text{Union bound}$$
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB) \qquad \text{Inclusion/exclusion for two events}$$

The key objects in probability are *random variables*. A regular variable takes on a specific value, but we do not have any information about what that value is. For instance, say

$$x \in \mathbb{R},$$

Here x is a *variable*, $\in$ means *is an element of* and $\mathbb{R}$ means *the real numbers*. Saying that x is a real-valued variable means x is taking on some specific value, but we have no idea what that value is.

Statements can be made about x, but the best that we can say about the statements is that they are either true or false. For instance,

$$\{x \in [10, 30]\} \in \{\text{TRUE}, \text{FALSE}\}$$

since the statement: x is an element of the closed interval $[10, 30]$ is either a true statement, a false statement, but not both.

In order to turn this into numbers, we use the *indicator function.* This function takes as input an expression that is either true or false. It returns 1 if the input is true, and 0 otherwise.

> **Definition 6**
> Using T for a true statement and F for a false statement, the **indicator function** $\mathbb{1}$: $\{F, T\} \to \{0, 1\}$ is defined as
>
> $$\mathbb{1}(T) = 1$$
> $$\mathbb{1}(F) = 0.$$

Using indicator notation,

$$\mathbb{1}(x \in [0, 30]) \in \{0, 1\}.$$

Still this is all or nothing: either we know the statement is true and return a 1, or we know that the statement is false and return a 0.

So now suppose we have a random variable X that represents the height of a building that we see in the distance. Without measuring the building exactly, we don't know for sure what the height is, but we can assign a probability to it.

$$\mathbb{P}(X \in [10, 30]) \in [0, 1].$$

The probability that we assign now is not just 0 or 1, but any number from 0 up to 1 that indicates our degree of belief that the statement is true. For a set A, if $\mathbb{P}_X(A) = \mathbb{P}(X \in A)$ is going to be a probability measure (aka probability distribution), it is necessary to satisfy some properties.

Last time, we said that a probability distribution (probability measure) was a function $\mathbb{P}$ from the set of events $\mathcal{F}$ to $[0, 1]$ that satisfied two properties:

1. Total probability: $\mathbb{P}(\Omega) = 1$.

2. Countable additivity: If $A_1, A_2, \ldots$ are all subsets of Ω so that for all $i \neq j$, $A_i A_j = \emptyset$, then

$$\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Start with these two facts about probability measures. There are many more facts about distributions, but mathematicians like to try and see if they can get away with assuming only a few things, and then proving the other properties are logical consequences of the few things that they assumed were true.

2.1 Complements and the empty set

Recall that $\emptyset$ is the *empty set* where for any variable x, the statement $x \in \emptyset$ is always false. Given that, what should $\mathbb{P}(X \in \emptyset) = \mathbb{P}_X(\emptyset)$ be? Well, we know that $X \in \emptyset$ is always false, so that should make the probability 0, but that is not one of the things that we assumed was true about a probability measure. We can, however, derive it from definition of a distribution.

> **Fact 1** (The empty set has probability 0.)
> $\mathbb{P}(\emptyset) = 0.$

In words, this means that the probability that the outcome is nothing is 0. Let's do the proof!

Proof. Let $A_i = \emptyset$ for all i. Then the A_i are disjoint and their union also equals A_1, so

$$\mathbb{P}(A_1) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) + \cdots .$$

Canceling $\mathbb{P}(A_1)$ from both sides gives

$$0 = \mathbb{P}(A_2) + \mathbb{P}(A_3) + \cdots \geq \mathbb{P}(A_2) \geq 0,$$

so $\mathbb{P}(A_2) = \mathbb{P}(\emptyset) = 0$. □

This type of proof is typically in this type of definition driven mathematics. It does not add to our intuition, what it does is to verify that we do not need to make a separate assumption that $\mathbb{P}(\emptyset) = 0$. Instead, this fact is already logically implied from our definitions.

Now that $\mathbb{P}(\emptyset) = 0$ is a proven fact, it can be used to prove other facts. For instance, in the definition of a distribution, we said that a countable sequence of measurable events had the probability of the union equal to the sum of the probabilities. What if we only have a finite sequence of events?

Fact 2
Let $A_1, \ldots, A_n$ be disjoint and measurable. Then

$$\mathbb{P}\left(A_1 \cup A_2 \cup \cdots \cup A_n\right) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots + \mathbb{P}(A_n).$$

Proof. Extend our finite set of events $A_1, \ldots, A_n$ out to a sequence by making everything else in the sequence the empty set. That is, consider the sequence

$$A_1, A_2, \ldots, A_n, \emptyset, \emptyset, \ldots .$$

The entire sequence is disjoint since the intersection of any A_i with $\emptyset$ is $\emptyset$, the intersection of any two empty sets is empty, and the intersection of any two nonidentical A_i is empty.

Hence

$$\mathbb{P}(A_1 \cup \cdots \cup A_n \cup \emptyset \cup \emptyset \cup \cdots) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n) + \sum_{i=1}^{\infty} \mathbb{P}(\emptyset)$$
$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n) + 0,$$

and that completes the proof. □

This last fact immediately leads in to our next important fact.

Fact 3 (probability of complement of an event is one minus the probability of the event.)
For an event A
$$\mathbb{P}(A^C) = 1 - \mathbb{P}(A).$$

Proof. Since A and A^C are disjoint

$$\mathbb{P}(A) + \mathbb{P}(A^C) = \mathbb{P}(A \cup A^C) = \mathbb{P}(\Omega) = 1.$$

Bringing the $\mathbb{P}(A)$ over to the other side completes the proof. □

Example 3
Each of ten million outcomes $\omega_1, \omega_2, \dots, \omega_{10^7}$ is equally likely. What is the chance that the outcome is not ω_1.

Answer We could say that it is

$$\mathbb{P}(\{\omega_1\}^C) = \mathbb{P}(\omega_2) + \mathbb{P}(\omega_3) + \cdots + \mathbb{P}(\omega_{10^7}),$$

or more succinctly,

$$\mathbb{P}(\{\omega_1\}^C) = 1 - \mathbb{P}(\{\omega_1\}) = 1 - 1/10^7 = \boxed{0.9999999}.$$

Fact 4 (Probabilities are strictly increasing with respect to subsets.)
Suppose $A \subseteq B$ for two events A and B. Then

$$\mathbb{P}(A) \leq \mathbb{P}(B).$$

Proof. Let $A \subseteq B$. Then $B = A \cup A^C B$ (in words, B consists of those elements that are either in A or in B but not in A.) Also, A and $A^C B \subseteq A^C$ are disjoint. Hence

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(A^C B) \geq \mathbb{P}(A).$$

$\square$

The fact means that we need only worry about the region of probability that has probability 1.

Fact 5
Suppose $\mathbb{P}(A) = 1$. Then $\mathbb{P}(B) = \mathbb{P}(A \cap B)$.

Proof. Note that $A \cap B$ and $A^C \cap B$ are disjoint and have union B. Hence

$$\mathbb{P}(A \cap B) + \mathbb{P}(A^C \cap B) = \mathbb{P}(B).$$

Also, since $A^C \cap B \subseteq A^C$, $\mathbb{P}(A^C \cap B) \leq \mathbb{P}(A^C) = 0$, so $\mathbb{P}(A^C \cap B) = 0$ which gives the result. $\square$

2.2 The union bound

Now consider finding the probability that at least one of a sequence of events occurs. In set notation, that means we are looking for the probability of the union of the events.

$$\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) + \cdots.$$

Because the events could overlap, our simple addition rule does not apply. To make the sets disjoint, consider first the union of two events.

Fact 6
For any A_1, A_2,
$$A_1 \cup A_2 = A_1 \cup A_1^C A_2.$$

Proof. Since $A_1^C A_2 \subseteq A_2$, both $A_1 \subseteq A_1 \cup A_2$ and $A_1^C A_2 \subseteq A_2$. Taken together,

$$A_1 \cup A_1^C A_2 \subseteq A_1 \cup A_2.$$

For the other direction, let $a \in A_1 \cup A_2$. Then a is either in A_1, in A_2, or in both. Case I: $a \in A_1$, then $a \in A_1 \cup A_1^C A_2$. Case II: $a \notin A_1$, then since it is not in A_1 or both in A_1 and A_2, it must be in A_2. Since $a \notin A_1$, then $a \in A_1^C$ by definition, so $a \in A_1^C A_2$ and $a \in A_1 \cup A_1^C A_2$. □

Moreover, A_1 and $A_1^C A_2 \subseteq A_1^C$ are disjoint. Also, $A_1^C A_2 \subseteq A_2$. So

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_1^C A_2) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2).$$

This fact about two sets can be generalized to what is called the *union bound* or sometimes *Bonferroni's inequality.*

Fact 7 (The union bound.)
Let $A_1, A_2, A_3, \ldots$ be a sequence of events. Then

$$\mathbb{P}((\cup_{i=1}^\infty A_i) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) + \cdots.$$

Proof. Let $A_1, A_2, \ldots$ be a sequence of events. Then

$$A_1 \cup A_2 \cup \cdots = A_1 \cup A_1^C A_2 \cup A_1^C A_2^C A_3 \cup \cdots,$$

where each event in the right hand side is disjoint from the others. Hence

$$\mathbb{P}((\cup_{i=1}^\infty A_i) = \mathbb{P}(A_1) + \mathbb{P}(A_1^C A_2) + \mathbb{P}(A_1^C A_2^C A_3) + \cdots \leq \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots.$$

□

2.3 Inclusion/exclusion

While the union bound can give an upper bound on a probability, it will not always give the precise probability. To accomplish this, first look at the union of two events. Consider a Venn diagram with two events.
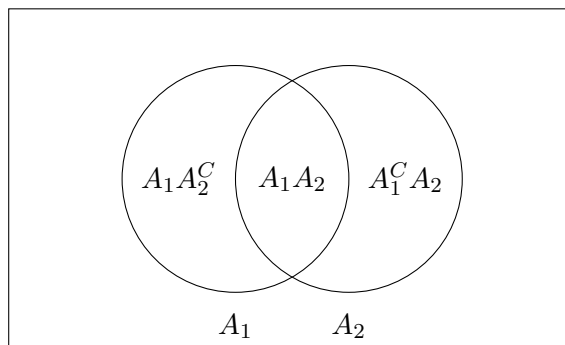
Notice that

$$A_1 \cup A_2 = A_1 A_2^C \cup A_1 A_2 \cup A_1 A_2^C$$
$$A_1 = A_1 A_2 \cup A_1 A_2^C$$
$$A_2 = A_1 A_2 \cup A_1^C A_2.$$

We won't go through the formal proof here, because it is similar to that of the last few facts. Using this breakdown of $A_1 \cup A_2$ means that $\mathbb{P}(A_1) = \mathbb{P}(A_1 A_2) + \mathbb{P}(A_1 A_2^2)$, or rearranging:

$$\mathbb{P}(A_1 A_2^C) = \mathbb{P}(A_1) - \mathbb{P}(A_1 A_2).$$

Similarly,

$$\mathbb{P}(A_1^C A_2) = \mathbb{P}(A_2) - \mathbb{P}(A_1 A_2).$$

This allows us to find the probability of the union of two events exactly.

Fact 8 (Inclusion/exclusion for two events.)
For any two events A_1 and A_2,

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 A_2).$$

Proof.

$$\begin{aligned}
\mathbb{P}(A_1 \cup A_2) &= \mathbb{P}(A_1 A_2^C) + \mathbb{P}(A_1 A_2) + \mathbb{P}(A_1 A_2^C) \\
&= \mathbb{P}(A_1) - \mathbb{P}(A_1 A_2) + \mathbb{P}(A_1 A_2) + \mathbb{P}(A_2) - \mathbb{P}(A_1 A_2) \\
&= \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 A_2).
\end{aligned}$$

$\square$

The above fact is called *inclusion/exlcusion* because we are including A_1 and A_2 and then excluding $A_1 A_2$ by subtracting the probability.

What's more, we can extend this to the union of any finite number of events. What happens is intersections of odd numbers of events get added, while intersections of even numbers of events get subtracted.

Fact 9 (Inclusion/exclusion principle)
For sets $A_1, \ldots, A_n$, let $[n] = \{1, 2, 3 \ldots, n\}$.

$$\mathbb{P}\left(\cup_{i=1}^{n} A_i\right) = \sum_{i \in n \text{ and odd}} \sum_{\{a_1, a_2, \ldots, a_i\} \subseteq [n]} \mathbb{P}(A_{a_1} A_{a_2} \cdots A_{a_i}) -$$
$$\sum_{i \in n \text{ and even}} \sum_{\{a_1, a_2, \ldots, a_i\} \subseteq [n]} \mathbb{P}(A_{a_1} A_{a_2} \cdots A_{a_i})$$

So for three sets, we have

$$\begin{aligned}
\mathbb{P}(A_1 \cup A_2 \cup A_3) = \ &\mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) - \mathbb{P}(A_1 A_2) - \mathbb{P}(A_1 A_3) - \mathbb{P}(A_2 A_3) + \\
&\mathbb{P}(A_1 A_2 A_3).
\end{aligned}$$

Problems

2.1 Suppose A and B are disjoint events, $\mathbb{P}(A) = 0.1$ and $\mathbb{P}(B) = 0.7$. What is $\mathbb{P}(A \cup B)$?

2.2 Suppose A_1, A_2 and A_3 are disjoint sets, $\mathbb{P}(A_1) = \mathbb{P}(A_2) = \mathbb{P}(A_3) = 0.3$.

  a) What is $\mathbb{P}(A_1 \cap A_2 \cap A_3)$?

b) What is $\mathbb{P}(A_1 \cup A_2)$?

2.3 Suppose $\mathbb{P}(A) = 0.4$, $\mathbb{P}(B) = 0.8$ and $\mathbb{P}(AB) = 0.3$. What is $\mathbb{P}(A \cup B)$?

2.4 Suppose $\mathbb{P}(A) = 0.7$ and $\mathbb{P}(B) = 0.4$. Is it possible for A and B to be disjoint? If so, give an example, otherwise prove that this is not possible.

2.5 If $\mathbb{P}([0,3]) = 0.3$ and $\mathbb{P}([5,9]) = 0.6$, what is $\mathbb{P}([0,3] \cup [5,9])$?

2.6 If $\mathbb{P}(\{a,b,c\}) = 0.2$ and $\mathbb{P}(\{d,e\}) = 0.4$, what is $\mathbb{P}(\{a,b,c,d,e\}$?

2.7 Say $\mathbb{P}(A_1) = \mathbb{P}(A_2) = \mathbb{P}(A_3) = 0.2$. Give an upper bound for $\mathbb{P}(A_1 \cup A_2 \cup A_3)$.

2.8 If $\mathbb{P}(A_i) \le (1/3)^i$ for $i \in \{1,2,3,\ldots\}$, give an upper bound on

$$\mathbb{P}(A_1 \cup A_2 \cup A_3 \cup \cdots).$$

2.9 Suppose a fair six sided die with sides labeled $\{1,2,\ldots,6\}$ is rolled three times. There are many possible outcomes, for instance, $(2,3,3)$ is one possible outcome.

a) How many possible outcomes are there?

b) If each outcome is equally likely, what must the probability of each outcome be?

c) What is the chance of getting all 6's on the three rolls?

d) What is the chance of not getting all 6's on the three rolls.

2.10 A department store models every person entering the store as either no spend, mid spend, or high spend. If the probability a person is no spend is 0.15 and mid spend is 0.4. What is the probability a person is high spend?

2.11 $\mathbb{P}(A \cup B) = 0.3$. What is $\mathbb{P}(A^C B^C)$?

2.12 Suppose $\mathbb{P}(A \cup B) = 0.5$. What is $\mathbb{P}(A^C B^C)$?

2.13 Suppose $\mathbb{P}(A \in [0,3]) = 1$, $\mathbb{P}(A \in [1,2]) = 0.3$ and $\mathbb{P}(A \in [2,3]) = 0.6$. What is $\mathbb{P}(A \in [2,5])$?

Discrete random variables

> **Question of the Day** Two fair six sided dice are rolled independently. What is the chance that they add to 5?

> **Summary** **Random variables** such as X represent values which are not completely unknown, so there is some partial information. The **distribution** of X is the probability measure $\mathbb{P}_X(A) = \mathbb{P}(X \in A)$. When the random variable is known to lie in a countably infinite set with probability 1, call the random variable **discrete**. Random variables X that are **uniformly distributed** over a finite set A (write $X \sim \mathsf{Unif}(A)$) are equally likely to be any element of A. Say that $(X, Y) \sim \mathsf{Unif}(A \times B)$ if and only if $X \sim \mathsf{Unif}(A)$, $Y \sim \mathsf{Unif}(B)$ where X and Y are **independent**.

3.1 Variables and random variables

If I tell you that $x_1 = 3$ and $x_2 = 2$, then $x_1 + x_2 = 5$. We typically call x_1 and x_2 *variables* and you have learned over the years many rules for dealing with variables. But now consider what happens when x_1 and x_2 are not known completely. Suppose we still know something about them, namely, that they are the outcomes of a roll of a fair six sided die, but we do not know any more than that.

Then we typically use capital letters to denote our values. For instance, we could use X_1 and X_2 for the dice rolls. (By the way, "dice" is the plural of "die". Writing in 2022, English is trending towards using the word "dice" for both one die and the plural form, but I will use "die" for one and "dice" for more than one die throughout.)

So now the question of the day reads: What is the chance that $X_1 + X_2 = 5$? Because we have only partial information about X_1 and X_2, we say that they are *random variables*.

For instance, we know that X_1 is either 1, 2, 3, 4, 5, or 6. Mathematically, we use *set notation* and write

$$X_1 \in \{1, 2, 3, 4, 5, 6\}.$$

The curly brackets { and } indicate that this is set. In a set, order does not matter, so $\{1, 2, 3, 4, 5, 6\} = \{6, 5, 4, 3, 2, 1\}$. We often use an ellipsis ... to indicate that the reader should mentally complete the in between parts, so

$$\{1, \ldots, 6\} = \{1, 2, 3, 4, 5, 6\}.$$

3.2   The uniform distribution

Now consider the question of the day. For this problem, we have two dice. Each die is a fair six sided die. That means that if X_1 represents the first die roll, and X_2 the second, then both X_1 and X_2 are elements of the set $\{1, 2, 3, 4, 5, 6\}$.

We use the word *fair* to indicate that we initially have no knowledge about the state of each die. That is, each of the possibilities are equally likely. This is the situation when you have the least amount of information about the value of a random variable.

To be more precise about what we mean by fair, we can talk about the distribution of the random variable.

Intuition 1

For a random variable X, the function $\mathbb{P}_X$ is defined as

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A)$$

is called the **distribution** of X.

That means that a distribution is a function. Now, real valued functions that we use a lot get special names, such as the exponential function, or the sine function. In the same way, we give special distributions that we use often their own name.

For instance, for X a fair six sided die roll,

$$\mathbb{P}_X(\{1, 3, 5\}) = \mathbb{P}(X \in \{1, 3, 5\}) = 3/6 = 1/2.$$

Notice that for a single outcome,

$$\mathbb{P}_X(\{1\}) = \mathbb{P}_X(\{2\}) = \cdots = \mathbb{P}_X(\{6\}) = 1/6.$$

In other words, there is a single probability that X equals each of the possible outcomes. The Latin prefix for one is *uni*, which is why unicycles have one wheel and unicorns have one horn. Because each outcome has this single probability, we name this distribution *uniform*. Formally, it can be defined as follows.

Definition 7

Let B be a set that satisfies $\#(B) > 0$ and $\#(B) < \infty$. Then X is **uniform over** B, write $X \sim \mathsf{Unif}(B)$, if for all measurable $A \subseteq B$,

$$\mathbb{P}(X \in A) = \frac{\#(A)}{\#(B)}.$$

Note that in particular, if $A = \{a\}$, then $\#(A) = 1$, and

$$\mathbb{P}(X \in A) = \mathbb{P}(X = a) = 1/\#(B).$$

Theoretically, that definition requires that we check the probability for all sets $A \subseteq B$. There are many such sets, so it is helpful to note that actually, we only have to check sets of size 1!

Fact 10

Suppose $\mathbb{P}(X = b) = 1/\#(B)$ for all $b \in B$. Then $X \sim \mathsf{Unif}(B)$.

Proof. Let $A \subseteq B$. Then

$$\mathbb{P}(X \in A) = \mathbb{P}\left(\cup_{a \in A}\{a\}\right) = \sum_{a \in A} \mathbb{P}(X \in \{a\}) = \sum_{a \in A} \frac{1}{\#(B)} = \frac{\#(A)}{\#(B)},$$

so $X \sim \mathsf{Unif}(B)$. $\qquad\square$

Example 4

Suppose $Y \sim \mathsf{Unif}(\{1, \ldots, 10\})$. What is the probability that $Y \in \{3, 4, 5\}$?

The answer is $\#(\{3, 4, 5\})/\#(\{1, \ldots, 10\}) = 3/10 = \boxed{0.3000}$.

3.3 Independent uniform random variables

We say that two random variables are *independent* if knowing the value of one random variable does not give us any information about the other. In the question of the day, knowing X_1 does not tell us anything about X_2. Mathematically, independence can be defined as follows.

Definition 8

Events A and B are **independent** if

$$\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B).$$

Random variables X and Y are **independent** if for all C and D, the events that $X \in C$ and $Y \in D$ are independent.

Definition 9

Random variables X and Y are **independent** if for all measurable A and B,

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

Example 5

Say $X \sim \mathsf{Unif}(\{1, 2, 3\})$ and $Y \sim \mathsf{Unif}(\{1, 2, 3, 4\})$ are independent. Find

$$\mathbb{P}(X = 1, Y \in \{2, 3\}).$$

Answer Because the random variables are independent,

$$\mathbb{P}(X = 1, Y \in \{2, 3\}) = \mathbb{P}(X = 1)\mathbb{P}(Y \in \{2, 3\})$$
$$= \frac{1}{3} \cdot \frac{2}{4} = \frac{1}{6} = 0.1666\ldots$$

For uniform random variables, independence gives us the following.

Fact 11

It hold that $(X, Y) \sim \mathsf{Unif}(\Omega_X \times \Omega_Y)$ if and only if $X \sim \mathsf{Unif}(A)$, $Y \sim \mathsf{Unif}(B)$, and X and Y are independent random variables.

Notation check: remember that $A \times B$ (read as A cross B) means the set of two dimensional vectors (a, b) such that $a \in A$ and $b \in B$.

Proof. Let $(a, b) \in A \times B$, so $a \in A$ and $b \in B$. Then by the independence of X and Y

$$\mathbb{P}(X = a, Y = b) = \mathbb{P}(X = a)\mathbb{P}(Y = b) = \frac{1}{\#(A)} \cdot \frac{1}{\#(B)} = \frac{1}{\#(A \times B)}.$$

Hence $(X, Y) \sim \mathsf{Unif}(A \times B)$.   □

In the question of the day, $X_1 \in \{1, \ldots, 6\}$ and $X_2 \in \{1, \ldots, 6\}$. Since each is uniform and independent,

$$(X_1, X_2) \sim \mathsf{Unif}(\{(1, 1), (1, 2), \ldots, (6, 5), (6, 6)\}).$$

There are 36 elements in $\{1, \ldots, 6\} \times \{1, \ldots, 6\}$, and exactly 4 of them

$$(1, 4)(2, 3), (3, 2), (4, 1)$$

have $X_1 + X_2 = 5$. Hence

$$\mathbb{P}(X_1 + X_2 = 5) = \mathbb{P}((X_1, X_2) \in \{(1, 4), (2, 3), (3, 2), (4, 1)\}) = \frac{4}{36} \approx \boxed{0.1111}.$$

3.4   *What makes a random variable discrete*

A random variable that is uniform over a finite set is an example of a *discrete random variable*. What makes it discrete? In mathematics, discrete refers to sets that are *finite* or *countably infinite*.

> **Definition 10**
> A set A is **finite** if there exists $n \in \{1, 2, 3, \ldots\}$ and an onto function $f : \{1, \ldots, n\} \to A$.

So a set is finite if we can count the elements of A using numbers $\{1, \ldots, n\}$ for some n. For instance, for the set $\{a, b, c\}$, I can assign element a the number 1, element b the number 2, and c the number 3. Since 3 is a positive integer, the set is finite.

> **Definition 11**
> A set A is **countably infinite** (a.k.a. **discrete**) if there exists an onto function $f : \{1, 2, 3, \ldots\} \to A$.

> **Definition 12**
> A random variable is **discrete** if there exists a discrete set Ω such that $\mathbb{P}(X \in \Omega) = 1$.

> **Example 6**
> Suppose $\mathbb{P}(X = i) = (1/2)^i$ for all $i \in \{1, 2, \ldots\}$. Prove that X is discrete.
>
> **Answer**   By the countable additivity property,
>
> $$\mathbb{P}(X \in \{1, 2, 3, \ldots\}) = \sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^i = 1,$$
>
> so X falls into a countably infinite set with probability 1, and so is discrete.

Problems

3.1 Let $U \sim \mathsf{Unif}(\{1, 2, 3, 4, 5, 6\})$. What is $\mathbb{P}(U \leq 4)$?

3.2 Let X be uniform over the positive even numbers that are at most 100. What is the chance that X is a multiple of 4?

3.3 Let $A = \{a, b, c\}$ and $B = \{d, e\}$. What is $A \times B$?

3.4 What is $\{1, 2, 3\} \times \{2, 3, 4\}$?

3.5 Let $W \sim \mathsf{Unif}(\{a, b, c, d\})$. What is $\mathbb{P}(W \in \{a, c\})$?

3.6 Let $W \sim \mathsf{Unif}(\{a, b\} \times \{c, d\})$. What is $\mathbb{P}(W = (a, c))$?

3.7 Let $X_1 \sim \mathsf{Unif}(\{1, \ldots, 6\})$ and $X_2 \sim \mathsf{Unif}(\{1, \ldots, 6\})$ be independent. Then what is $\mathbb{P}(X_1 + X_2 = 6)$?

3.8 Suppose that (X, Y) is drawn uniformly from $\{1, 2, 3\} \times \{1, 2\}$. What is the chance of picking $X = Y = 2$?

3.9 Suppose I roll three fair six sided dice so that each outcome is equally likely, and call the result (X_1, X_2, X_3). Let S be the smallest value showing on the dice. For $i \in \{1, 2, 3, 4, 5, 6\}$, find $\mathbb{P}(S = i)$.

3.10 Suppose I roll three fair ten sided dice where each die is marked $\{0, 1, 2, \ldots, 9\}$. What is the chance of rolling $(0, 0, 7)$?

3.11 Prove that $\{2, 3, 4, 5, \ldots\}$ is a discrete set.

3.12 Suppose that for $i \in \{1, 2, 3, \ldots\}$, $\mathbb{P}(X = i) = (1/3)(2/3)^{i-1}$. Prove that X is a discrete random variable.

Continuous random variables

Question of the Day Suppose U_1 and U_2 are independent and uniform over $[0, 1]$. What is the chance that $U_1 \leq U_2^2$?

Summary **Uniform random variables** X over a set B of nonzero finite measure have $\mathbb{P}(X \in A) = m(A)/m(B)$ for all measurable $A \subset B$. **Continuous random variables** have $\mathbb{P}(X = a) = 0$ for all $a \in \mathbb{R}$. Uniform random variables over a continuous set are continuous random variables.

4.1   Continuous Uniform random variables

Recall that for discrete random variables, and $A \subset B$, if $X \sim \mathsf{Unif}(B)$, then

$$\mathbb{P}(X \in A) = \frac{\#(A)}{\#(B)}.$$

So for $X \in \{a, b, c, d, e\}$,

$$\mathbb{P}(X \in \{b, c\}) = \frac{2}{5} = 0.4000.$$

This works because $\#(B) > 0$ and $\#(B) < \infty$, so the denominator makes sense.

For a continuous uniform random variable, instead of using counting measure, we use Lebesgue measure.

Definition 13
Let m denote Lebesgue measure, and suppose B is a set such that $m(B) \in (0, \infty)$. Say that X is **uniform over B** if for all measurable $A \subseteq B$,

$$\mathbb{P}(X \in A) = \frac{m(A)}{m(B)}.$$

Note: this definition gives a probability distribution for any measure where $m(B)$ is a positive finite number, not just counting or Lebesgue! However, these are usually the only two measures for which the resulting distribution is called uniform.

Example 7
Suppose $Y \sim \mathsf{Unif}([2, 8])$. What is $\mathbb{P}(Y \in [3, 4])$?

Answer   Since $[3, 4] \subseteq [2, 8]$,

$$\mathbb{P}(Y \in [3, 4]) = \frac{m([3, 4])}{m([2, 8])} = \frac{4 - 3}{8 - 2} = \frac{1}{6} \approx \boxed{0.1666}.$$

For variables uniform over B, the chance that the variable lands outside of the set B is 0. Formally, we have the following.

Fact 12
For $X \sim \mathsf{Unif}(B)$, if $CB = \emptyset$, then $\mathbb{P}(X \in C) = 0$.

Proof. Let C be such that $CB = \emptyset$. Then CB and B are disjoint, so

$$\mathbb{P}(X \in CB \cup B) = \mathbb{P}(X \in CB) + \mathbb{P}(X \in B).$$

But $\mathbb{P}(X \in B) = 1$ and $\mathbb{P}(X \in CB \cup B) \le 1$. Hence $\mathbb{P}(X \in CB) = 0$.   □

Example 8
Suppose $Y \sim \mathsf{Unif}([2, 8])$. What is $\mathbb{P}(Y \in [0, 4])$?
   Since $[0, 4] = [0, 2) \cup [2, 4]$ and $[2, 4] \subseteq [2, 8]$,

$$\mathbb{P}(Y \in [0, 4]) = \mathbb{P}(Y \in [0, 2)) + \mathbb{P}(Y \in [2, 4]) = 0 + \frac{4 - 2}{8 - 2} = \frac{2}{6} \approx \boxed{0.3333}.$$

4.2   Independent random variables

Now let us consider two dimensions. Suppose that $(U_1, U_2) \sim \mathsf{Unif}([0, 1] \times [0, 1])$ so that we are drawing a point uniformly from the unit square.
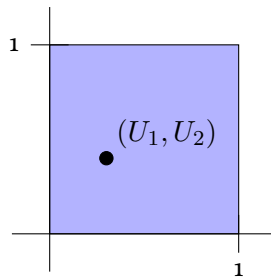
The same definition applies to two, three, or higher numbers of dimensions.

Example 9

What is the chance that $U_1 \geq 2U_2$?

If U_1 is on the horizontal axis and U_2 is on the vertical, then we want

$$U_2 \geq (1/2)U_1,$$

and so the region where $U_2 \geq (1/2)U_1$ looks like this:
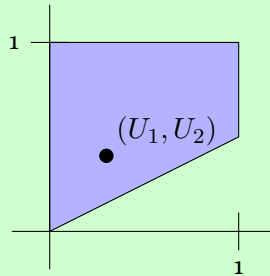
This is a trapezoid with area

$$\int_{u_1=0}^{1} 1 - (1/2)u_1 \; du_1 = u_1 - u_1^2/4|_0^1 = 3/4 = \boxed{0.7500}.$$

The same rule (Fact 11) for independent uniform random variables that worked for the discrete case also applies in the continuous case. That is,

$$(X, Y) \sim \mathsf{Unif}(A \times B) \Leftrightarrow X \sim \mathsf{Unif}(A), Y \sim \mathsf{Unif}(B), X \text{ and } Y \text{ are independent.}$$

Question of the Day Here $U_1 \sim \mathsf{Unif}([0, 1])$ and $U_2 \sim \mathsf{Unif}([0, 1])$ are independent, so $(U_1, U_2) \sim \mathsf{Unif}([0, 1] \times [0, 1])$. So to find $\mathbb{P}(U_1 \leq U_2^2)$, we need to graph that region in U_1, U_2 space.
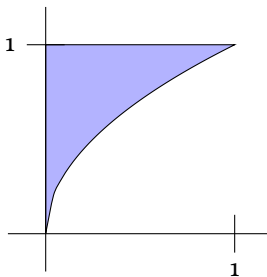
To find the area:

$$\int_{y=0}^{1} \int_{x=0}^{y^2} 1 \; dx \; dy = \int_{y=0}^{1} y^2 \; dy = y^3/3|_0^1 = 1/3 \approx \boxed{0.3333}.$$

Example 10

Suppose $(U_1, U_2) \sim \mathsf{Unif}(A)$, where A is the triangular region with vertices $(0,0)$, $(1,0)$, and $(0,1)$. Then prove that U_1 and U_2 are not independent.

Answer   Consider the events $U_1 > 0.5$ and $U_2 > 0.5$. Then if both these things are true, $U_1 + U_2 > 1$, which is outside of the trianglular region A:
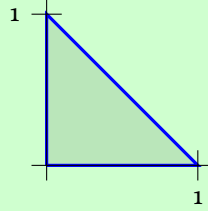
So $\mathbb{P}(U_1 > 0.5, U_2 > 0.5) = 0$.
Now consider $\mathbb{P}(U_1 > 0.5)$. This is the shaded region of the triangle.
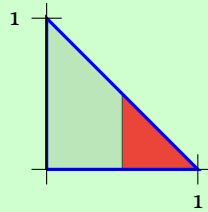
This shaded region has area $(1/2)(1/2)(1/2) = 1/8$, and the overall triangle has area $(1/2)(1)(1) = 1/2$, so the probability that the uniform lands in the shaded region is

$$\mathbb{P}(U_1 > 0.5) = \frac{1/8}{1/2} = \frac{1}{4}.$$

Similarly, the probability that $U_2 > 0.5$ is the probability that the uniform lands in the blue shaded region.
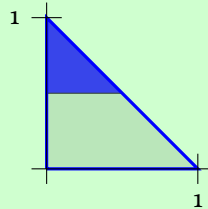
This is also $1/4$. Hence

$$\mathbb{P}(U_1 > 0.5)\mathbb{P}(U_2 > 0.5) = (1/4)(1/4) = 1/16,$$

which is not $\mathbb{P}(U_1 > 0.5, U_2 > 0.5) = 0$. So U_1 and U_2 are not independent.

4.3 What makes a random variable continuous

Suppose $X \sim \mathsf{Unif}([0,1])$. Then

$$\mathbb{P}(X = 0.3) = \mathbb{P}(X \in [0.3, 0.3]) = \frac{0.3 - 0.3}{1 - 0} = 0.$$

In fact, for any single point x in $[0,1]$, $\mathbb{P}(X = x) = 0$. That turns out to be the way we define a continuous random variable.

> **Definition 14**
> Say that X is a **continuous random variable** if for all x, $\mathbb{P}(X = x) = 0$.

> **Example 11**
> Show that for $U \sim \mathsf{Unif}([0,1])$, U^2 is a continuous random variable.
>   Let $x \in \mathbb{R}$. Suppose $x \neq 0$. Then
>
> $$\mathbb{P}(U^2 = x) = \mathbb{P}(U = \sqrt{x}) + \mathbb{P}(U = -\sqrt{x}) = 0 + 0 = 0$$
>
>   Now suppose $x = 0$. Then $\mathbb{P}(U^2 = 0) = \mathbb{P}(U = 0) = 0$. Hence U^2 is continuous.

From the last example, you might be tempted to think that any function of a continuous random variable is also a continuous random variable. But that is not true!

Let $f(x) = 1$ when $x \in [0, 0.5]$, and $f(x) = 2$ when $x \in (0.5, 1]$. Then for $U \sim \mathsf{Unif}([0,1])$,

$$\mathbb{P}(f(U) = 1) = \mathbb{P}(U \in [0, 0.5]) = 0.5$$
$$\mathbb{P}(f(U) = 2) = \mathbb{P}(U \in (0.5, 1]) = 1 - 0.5 = 0.5$$

So $f(U) \sim \mathsf{Unif}(\{1, 2\})$ is a discrete random variable.

4.4 Constructing a continuous uniform random variable

So we said what a continuous uniform random variable is, but how can we build such a variable from discrete uniforms? The answer works as follows.

Suppose that $X_i \sim \mathsf{Unif}(\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\})$, and for every n, $\{X_1, \ldots, X_n\}$ are independent random variables. Then we call

$$X_1, X_2, \ldots$$

an *independent, identically distributed* or *iid* sequence of random variables.

Given this sequence, we treat the sequence as digits of a real number in $[0,1]$. Formally, we let

$$U = \sum_{i=1}^{\infty} X_i / 10^i.$$

So for instance, if

$$X_1 = 6, X_2 = 0, X_3 = 4, X_4 = 7,$$

then

$$U = 0.6047\ldots$$

It turns out (with a fair amount of work), that one can prove that a random variable U generated in this way is actually uniform over the interval $[0,1]$

There's another, more geometric way of thinking about this as well. Start with the interval $[0, 1]$. Then flip a fair coin whose sides are labeled either R for right or L for left. Then if the coin is left, we consider the left half of the interval $[0, 1/2)$, and if it is right we consider the right half of the interval $[1/2, 1)$. Then flip the coin again (independently) and select the right or left half, and so on.

So if the first four flips were $RLLR$, then the interval has been reduced down to $[5/16, 6/16)$. Each time the length of the interval gets chopped in half. If we take the limit of the left endpoint of the interval as the number of flips goes to infinity, it approaches a single value, and that will be our $U \sim \mathsf{Unif}([0, 1])$.

For most values, there is one sequence that reaches that value. For instance,

$$RLRLRLRLRL \cdots$$

leads to $2/3$. For some values, there are two sequences that reach that value, for instance both

$$RLLLLLL \ldots \text{ and } LRRRRRRR \ldots$$

lead to $1/2$.

Either way, the *probability* of having either that single or double sequence of flips is equal to 0. And that is why the continuous uniforms have probability 0 of hitting any particular state.

4.5   *The paradox of the real numbers*

Recall that for a countably infinite sequence of disjoint sets, the sum of the probabilities is equal to the probability of the union of the sets. Every set is the disjoint union of the singleton sets that each contain a single element. That is,

$$A = \cup_{a \in A} \{a\}.$$

So for $[0, 1]$,

$$[0, 1] = \cup_{a \in [0,1]} \{a\}.$$

However, for $U \sim \mathsf{Unif}([0, 1])$, $\mathbb{P}(U \in [0, 1]) = 1$, while $\mathbb{P}(U = a) = \mathbb{P}(U \in \{a\}) = 0$. So

$$1 \neq \sum_{a \in [0,1]} \mathbb{P}(U = \{a\}).$$

That means that the set of numbers in $[0, 1]$ do *not* form a countable set! For this reason, we say that the interval $[0, 1]$ is an *uncountable* set.

That was a bit of a shock to the first person to discover this fact: Georg Cantor. Many mathematicians refused to believe it for a time, but today it is a commonly accepted fact about the real numbers that we just have to live with.

Problems

4.1 Suppose $W \sim \mathsf{Unif}([-3, 3])$.

    a) What is $\mathbb{P}(W \in [-1, 2])$?

    b) What is $\mathbb{P}(W \in [-5, 0])$?

4.2 Suppose $Y \sim \mathsf{Unif}[0, 10]$.

     a) Find $\mathbb{P}(Y \in [3, 7])$.

     b) Find $\mathbb{P}(Y \in [6, 12])$.

4.3 Suppose (U_1, U_2) is uniformly chosen over the unit circle

$$\{(x, y) : x^2 + y^2 \leq 1\}.$$

What is the chance that $|U_1| \geq U_2$?

4.4 Suppose that $U = (U_1, U_2)$ is uniformly chosen over the region:

$$\{(x, y) : x \geq 2, 0 \leq y \leq 1/x^2\}.$$

     a) What is $\mathbb{P}(U_1 \leq 5)$?

     b) What is $\mathbb{P}(U_2 \geq .01)$?

4.5 Let U_1 and U_2 be independent uniform random variables over $[0, 1]$. What is the chance that $U_2 \geq 3U_1$?

4.6 Let $\Omega = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq x^2\}$ and suppose $(X, Y) \sim \mathsf{Unif}(\Omega)$.

     a) Find $\mathbb{P}(X \leq 0.3)$.

     b) For $a \in \mathbb{R}$, find $\mathbb{P}(X \leq a)$. Write your answer using indicator functions.

     c) Suppose X_1, X_2, X_3 are iid draws from X. Find $\mathbb{P}(\min\{X_1, X_2, X_3\} \leq 0.3)$.

4.7 Say that $R \sim \mathsf{Unif}([0, 1])$.

     a) What is $\mathbb{P}(R \leq 0.4)$?

     b) What is $\mathbb{P}(R \leq 1.4)$?

     c) What is $\mathbb{P}(R \leq -0.4)$?

4.8 Suppose that (U_1, U_2) is uniform over the quadrilateral region with vertices $(0, 0), (0, 1), (2, 2), (2, 0)$. Prove that U_1 and U_2 are not independent.

Functions of random variables

Question of the Day Suppose $U \sim \text{Unif}([0,1])$. Let $T = -\ln(U)$. What is $\mathbb{P}(T \le a)$ for $a \ge 0$?

Summary The **cumulative distribution function** or **cdf** of a random variable X is $\text{cdf}_X(a) = \mathbb{P}(X \le a)$. Two random variables with the same cdf have the same distribution. Intuitively, a random variable is any function of a uniform random variable. Several such random variables important enough to be given names are as follows.

- The Bernoulli distribution is a random variable that is either 0 or 1.

- The exponential distribution with rate λ is the negative the natural logarithm of a uniform over $[0,1]$ divided by λ.

- The geometric distribution is the number of flips of a (possibly unfair) coin needed to obtain the first head.

Suppose I start with a random variable X that is uniform over $[-1,1]$. Now I take the absolute value of that random variable. So $Y = f(X) = |X|$. Note that Y contains less information than X. The value of X told us both the distance from 0, and whether the value was positive or negative. The value of Y only tells us the distance from 0.

What is the distribution of Y? A useful fact is that the distribution of a real-valued random variable is completely determined by the *cumulative distribution function* or *cdf* of the variable.

> **Definition 15**
> For a random variable Y, the **cumulative distribution function** or **cdf** of Y is defined as
> $$\text{cdf}_Y(a) = \mathbb{P}(Y \le a).$$

> **Fact 13**
> Let X and Y have the same cdf. Then $\mathbb{P}_X \sim \mathbb{P}_Y$.

That is, if X and Y have the same cdf, then they have the same distribution. The proof of this important fact is usually given in a second course in real analysis.

As our first example, let's find the cdf for uniform random variables over $[0, 1]$.

Fact 14
Let $U \sim \mathsf{Unif}([0, 1])$. Then

$$\mathrm{cdf}_U(a) = \mathbb{P}(U \le a) = a\mathbb{1}(a \in [0, 1]) + \mathbb{1}(a > 1).$$

Proof. If $a < 0$, then $\mathbb{P}(U < a) = 0$. If $a \in [0, 1]$ then

$$\mathbb{P}(U \le a) = \frac{a - 0}{1 - 0} = a,$$

and if $a > 1$, then $\mathbb{P}(U \le a) = 1$.   □

Let's graph this cdf function.
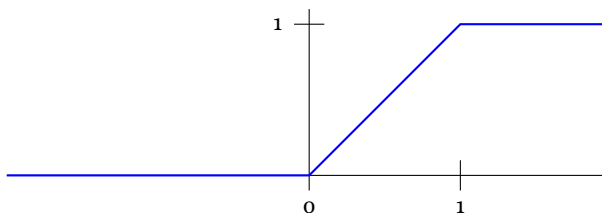
Notation 3
Another common notation for the cdf of a random variable is to use a capital letter F, so

$$F_Y(a) = \mathrm{cdf}_Y(a).$$

Now work out our example from earlier.

Example 12
Suppose $X \sim \mathsf{Unif}([-1, 1])$. What is the distribution of $Y = |X|$?

Answer Consider the cdf of Y. $|X| \ge 0$ so for $a < 0$, $\mathbb{P}(Y < a) = 0$. Also, for $X \in [-1, 1]$, $|X| \le 1$, so for $a > 1$, $\mathbb{P}(Y < a) = 1$. Last, if $a \in [0, 1]$, then

$$\begin{aligned}
\mathbb{P}(Y \le a) &= \mathbb{P}(|X| \le a) \\
&= \mathbb{P}(-a \le X \le a) \\
&= \frac{a - (-a)}{1 - (-1)} \\
&= \frac{2a}{2} = a.
\end{aligned}$$

Hence the cdf of Y is the same as a cdf of a uniform over $[0, 1]$, and must have the same distribution. That is, $\boxed{Y \sim \mathsf{Unif}([0, 1]).}$

Sometimes a function of a continuous random variable is a discrete random variable.

> **Example 13**
> Suppose $U \sim \mathsf{Unif}([0,1])$. Then consider $W = \mathbb{1}(U \leq 0.3) + 4\mathbb{1}(U > 0.3)$. What is the distribution of W?
>
> **Answer**   Here $\mathbb{P}(W = 1) = \mathbb{P}(U \leq 0.3) = 0.3$ and $\mathbb{P}(W = 4) = (1 - 0.3) = 0.7$, so the distribution is
> $$\boxed{\mathbb{P}(W = 1) = 0.3, \ \mathbb{P}(W = 4) = 0.7}.$$

It turns out that *every* real-valued random variable can be written as a function of one or more uniform random variables!

> **Intuition 2**
> Let $U \sim \mathsf{Unif}([0,1])$ and $X = f(U)$ for some function f that can be computed. Then X is a **random variable.**

Some remarks on this idea of a random variable.

1. In more advanced probability, random variables are taken to be *measurable functions*. These functions expand upon the notion of computable functions. For this more advanced definition, see Chapter 32.

2. Which functions can be computed depends on the model we are using for computation. For our purposes, all commonly used functions (squaring, multiplying, adding, et cetera) are computable functions.

3. For *computer simulation*, this definition means that any stream of uniform $[0,1]$ random variables can be used as the source of randomness in the simulation.

4. Note that if $X = f(U_1)$ for $U_1 \sim \mathsf{Unif}([0,1])$, and $Y = g(X)$, then $Y = g(f(U_1))$. This means that any computable function of a random variable is another random variable.

Given this definition, we can now look at some common distributions.

5.1   The Bernoulli distribution

Our first distribution is called the *Bernoulli distribution* after the Swiss mathematician Jacob Bernoulli who did pioneering work in probability. This is in some sense the simplest nontrivial distribution, as a random variable B has a Bernoulli distribution if $\mathbb{P}(B = 1) = p$ and $\mathbb{P}(B = 0) = 1 - p$ for some $p \in [0,1]$.

> **Definition 16**
> Say that B has the **Bernoulli distribution with parameter** p, and write $B \sim \mathsf{Bern}(p)$, if $B = \mathbb{1}(U \leq p)$, where $U \sim \mathsf{Unif}([0,1])$.

5.2   The exponential distribution

The random variable $T = -\ln(U)$ is an important one in probability. We say that this random variable has an *exponential* distribution. More generally, this is defined as follows.

> **Definition 17**
> Say that T has the **exponential distribution** with rate λ, and write $T \sim \mathsf{Exp}(\lambda)$, where $\lambda > 0$ is a parameter, if
> $$T = -\frac{1}{\lambda}\ln(U),$$
> where $U \sim \mathsf{Unif}([0,1])$.

Note that since $U \in (0,1)$ with probability 1, $\ln(U) < 0$, so $T > 0$ with probability 1.

5.3   *The magic of a uniform over* $[0,1]$

One of the awesome things about a uniform over $[0,1]$ is that you can use it to get more than one uniform over $[0,1]$!

Remember that a uniform random variable over $[0,1]$ can be thought of as an infinite stream of uniformly random digits over $\{0,1,\ldots,9\}$. For instance,

$$U = 0.66113356083398457 2979351\ldots$$

Gives the stream of digits $6,6,1,1,3,3,5,6,0,8,3,9,8,\ldots$.

Now suppose that we wanted two uniforms over $[0,1]$. Then we could just use the odd digits for the first uniform, and the even for the second. That is,

$$U_1 = 0.613503947995\ldots$$
$$U_2 = 0.613683852731\ldots.$$

So $X = f(U) = U_1 - U_2$ is also a random variable!

We can go further! Split U_2 into U_2 and U_3, split U_3 into U_3 and U_4, and so on. In the end, the result is an infinite sequence of uniform random variables that are uniform over $[0,1]$.

$$U_1, U_2, U_3, U_4, \ldots.$$

Since each uniform is composed of independent digits, each one will also be independent of the others. Such a sequence of independent, identically distributed random variables are called *iid*.

> **Definition 18**
> Suppose $X_1, X_2, \ldots$ all have the same distribution, and for all n, $(X_1, \ldots, X_n)$ are independent random variables. Say that the $\{X_i\}$ form an **independent, identically distributed** or **iid** sequence of random variables.

Then the following holds.

> **Fact 15**
> Let $U_1, U_2, \ldots$ be iid $\mathsf{Unif}([0,1])$. Then any $f(U_1, U_2, \ldots)$ that can be computed gives a random variable.

This can be used to give the *geometric distribution*. Suppose that G is the smallest value of i such that $U_i \leq p$. Then we say that G has the *geometric distribution* with parameter i.

Mathematically, the smallest value of a nonempty set A is called the *infimum* (write $\inf(A)$) of the set. For instance,

$$\inf\{1, 2, 3\} = 1$$
$$\inf\{4, 6, 8, \ldots\} = 4$$
$$\inf\{\} = \infty.$$

That last equation is by convention: $\inf(\emptyset) = \infty$.

Now if $U_1, U_2, \ldots$ is a sequence of iid uniforms over $[0, 1]$, then $\mathbb{1}(U_1 \leq p), \mathbb{1}(U_2 \leq p), \mathbb{1}(U_3 \leq p), \ldots$ give a sequence of 0's and 1's. For instance, the sequence might be

$$0, 0, 1, 0, 1, 1, \ldots.$$

Then the locations with 1's would be $\{3, 5, 6, \ldots\}$. Taking the infimum of this set gives 3, which is the first position of a 1. This motivates the definition of a geometric distributed random variable, which is the number of flips of a coin with probability p of heads needed to get the first head.

> **Definition 19**
> Let $G = \inf\{i : U_i \leq p\}$. Then write $G \sim \text{Geo}(p)$, and say that G is a **geometric random variable with parameter** p where U_1, U_2, U_3 are iid.

You can think of a geometric random variable as the number of flips of a coin needed to see the first head, where the probability that the coin is heads is p.

> **Example 14**
> Let $G \sim \text{Geo}(0.3)$. What is the chance that $G = 3$?
>
> **Answer**   For G to equal 3, we must have $U_1 > 0.3, U_2 > 0.3$ and $U_3 \leq 0.3$. Hence
>
> $$\mathbb{P}(G = 3) = (1 - 0.3)(1 - 0.3)(0.3) = \boxed{0.1470}.$$

Problems

5.1 Suppose $U \sim \text{Unif}([0, 1])$ and $A = -\ln(U)/2$.

    a) Find $\mathbb{P}(A \geq 2)$.

    b) Find $\mathbb{P}(A \geq -2)$.

    c) For $a \geq 0$, find $\mathbb{P}(A \geq a)$.

    d) For $a < 0$, find $\mathbb{P}(A \geq a)$.

5.2 Suppose $U \sim \text{Unif}([0, 1])$ and $W = 1/U$.

    a) Find $\mathbb{P}(W \geq 2)$.

    b) Find $\mathbb{P}(W \geq -2)$.

5.3 Let $U \sim \text{Unif}([-1, 1])$. Find the cdf of $1 - U^2$.

5.4 Let $U \sim \text{Unif}([-1, 1])$. Find the cdf of U^3.

5.5 Let ω be uniform over $[0, 1]$, and suppose $X(\omega) = 2\omega + 3$. Find

    a) $\mathbb{P}(X \in [3.5, 4.7])$.

    b) $\mathbb{P}(X \in [0, 1])$.

    c) $\mathbb{P}(X^2 \le 10)$.

5.6 Suppose $U \sim \mathsf{Unif}([-1, 0])$. Prove that $-U \sim \mathsf{Unif}([0, 1])$ by showing that $\mathrm{cdf}_{-U}(a) = a\mathbb{1}(a \in [0, 1]) + \mathbb{1}(a > 1)$.

5.7 Suppose $U \sim \mathsf{Unif}([-1, 0])$.

    a) Let $X = U^2$. Find the cdf of X.

    b) Find the cdf of U.

5.8 Suppose $U \sim \mathsf{Unif}([0, 1])$ and $X = U^2$. Find cdf_U.

5.9 Let $G \sim \mathsf{Geo}(p)$. For i a value in $\{1, 2, 3 \ldots\}$, what is $\mathbb{P}(G = i)$?

5.10 Suppose that (U_1, U_2) is uniform over the quadrilateral region with vertices $(0, 0), (0, 1), (2, 2)$, and $(2, 0)$. Find the cdf of U_1.

5.11 Let $U \in [-1, 1]$. What is $\mathbb{P}(U^2 \ge 0.6)$?

5.12 Suppose $T \sim \mathsf{Exp}(2)$. Find and graph cdf_T.

5.13 Consider the probability that for $\mathsf{Exp}(1)$ and $\mathsf{Unif}([0, 1])$ random variables drawn independently, that the second is bigger than the first. To find this, let U_1 and U_2 be iid $\mathsf{Unif}([0, 1])$. Then set $T = -\ln(U_2)$. Then find $\mathbb{P}(U_1 \ge T)$.

5.14 Let $T_1 \sim \mathsf{Exp}(1)$ and $T_2 \sim \mathsf{Exp}(2)$ be independent. Find $\mathbb{P}(T_1 \ge T_2)$.

5.15 Let $B \sim \mathsf{Bern}(p)$ and $T \sim \mathsf{Exp}(1)$ be independent random variables. Find $\mathbb{P}(T \ge B)$.

5.16 Suppose $\mathbb{P}(X = 1) = 0.2$, $\mathbb{P}(X = 2) = 0.3$, and $\mathbb{P}(X = 3) = 0.5$. For $T \sim \mathsf{Exp}(1)$ independent of X, find $\mathbb{P}(T \ge X)$.

5.17 The time until radioactive decay of a single atom is exponentially distributed with rate λ. If T is the time until the particle decays, the half-life t_{hl} is the time such that $\mathbb{P}(T \ge t_{\mathrm{hl}}) = 1/2$. The half-life for an atom of uranium 238 is 4.5 billion years.

    a) What is λ?

    b) If the Earth is 4.2 billion years old, what is the chance that an atom of U-238 present at the birth of the planet is still intact?

5.18 Plutonium 241 has a half life of 14.4 years. What is the chance that a single atom of Pu-241 survives at least 20 years?

Chapter 6

Conditioning

Question of the Day Suppose $T \sim \mathsf{Exp}(2)$. What is the probability that T is at least 4 given that it is at least 1?

Summary Conditioning is a way of saying what extra information about a random variable we are given. We use a vertical | to separate the random variable (to the left of the bar), and the information (to the right of the bar.) So $\mathbb{P}(X \in A | Y \in B)$ means the probability that the random variable X falls into A given the information that the random variable Y falls into B. As long as $\mathbb{P}(Y \in B) > 0$, then

$$\mathbb{P}(X \in A | Y \in B) = \frac{\mathbb{P}(X \in A, Y \in B)}{\mathbb{P}(Y \in B)}.$$

When we have extra knowledge about a random variable, we typically use a vertical bar | to separate the event we want the probability of (on the left) and the event that we know happened (on the right).

So for the question of the day, want we want to find is

$$\mathbb{P}(T \geq 4 | T \geq 1).$$

To the right of the bar goes $\{T \geq 1\}$, the event that we are told (given) happened. To the left of the bar goes the event $\{T \geq 4\}$ that we are trying to find the probability of. We say we are trying to find the probability $T \geq 4$ conditioned on $T \geq 1$.

To understand what this probability is, it helps to first step back and consider a simple example. Suppose that $U \sim \mathsf{Unif}(\{1, 2, 3, 4, 5, 6\})$, and then further information is obtained that $U \leq 4$. What should the distribution of U given that $U \leq 4$, written $[U | U \leq 4]$, be?

Well, saying $U \sim \mathsf{Unif}(\{1, 2, 3, 4, 5, 6\})$ means that we have no information about which possibility is more likely. The extra information that $U \leq 4$ eliminates 5 and 6 as possible values, but does not tell us anything about whether or not (say) 3 is more likely than 1. This leads to the following intuition.

Intuition 3
Suppose $A \subseteq B$ where A has positive measure. Then for $U \sim \mathsf{Unif}(B)$,

$$[U | U \in A] \sim \mathsf{Unif}(A).$$

In particular, suppose $C \subseteq A \subseteq B$, and $U \sim \mathsf{Unif}(B)$. Then

$$\mathbb{P}(U \in C | U \in A) = \mathbb{P}(W \in C),$$

where $W \sim \mathsf{Unif}(A)$. Hence

$$\mathbb{P}(U \in C | U \in A) = \frac{m(C)}{m(A)} = \frac{m(C)/m(B)}{m(A)/m(B)} = \frac{\mathbb{P}(U \in C)}{\mathbb{P}(U \in A)}.$$

Now consider conditioning when C is not a subset of A. So the Venn diagram might look something like
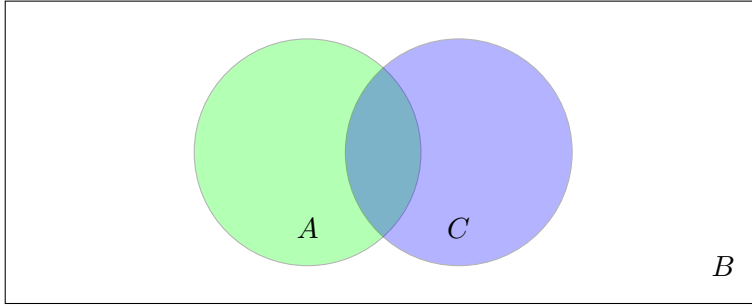
Remember that $U \sim \mathsf{Unif}(B)$, we are told that we fall into A, and now are asking what is the chance that we also fell into C? Well, of course, most of C does not happen. The only way for U to fall into C at this point is if U falls into $A \cap C$. So

$$\mathbb{P}(U \in C | U \in A) = \mathbb{P}(U \in A \cap C | U \in A) = \frac{m(A \cap C)}{m(A)}.$$

Now divide top and bottom by $m(B)$ to get

$$\mathbb{P}(U \in C | U \in A) = \frac{m(AC)/m(B)}{m(A)/m(B)} = \frac{\mathbb{P}(U \in A \cap C)}{\mathbb{P}(U \in A)}.$$

At this point remember that all probability distributions are based on the uniform distribution over $[0, 1]$. That means for any two events E_1 and E_1, there exists C and A such that

$$E_1 = \{U \in C\}, \ E_2 = \{U \in A\}.$$

Hence

$$\mathbb{P}(E_1 | E_2) = \mathbb{P}(U \in C | U \in A) = \frac{\mathbb{P}(U \in A, U \in C)}{\mathbb{P}(U \in A)} = \frac{\mathbb{P}(E_1, E_2)}{\mathbb{P}(E_2)}.$$

This argument motivates the following definition of conditional probability.

> **Definition 20**
> For events A and B where $\mathbb{P}(B) > 0$, the **conditional probability of event A given event B** is
> $$\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}.$$

In particular, if there are random variables X and Y so that the events of interest are $A = \{X \in C\}$ and $B = \{Y \in D\}$, then

$$\mathbb{P}(X \in C | Y \in D) = \frac{X \in C, Y \in D}{\mathbb{P}(Y \in D)}.$$

Question of the Day   With this in mind, we can now tackle the question of the day. In this problem, we are given that $T \sim \text{Exp}(2)$. Recall that this means that $T = -(1/2)\ln(U)$, where U is uniform over $[0, 1]$.

Now to solve the Question of the Day, we need only apply the conditional probability formula.

$$
\begin{aligned}
\mathbb{P}(T \geq 4 | T \geq 1) &= \frac{\mathbb{P}(T \geq 4, T \geq 1)}{\mathbb{P}(T \geq 1)} \\
&= \frac{\mathbb{P}(T \geq 4)}{\mathbb{P}(T \geq 1)} \\
&= \frac{\mathbb{P}(-(1/2)\ln(U) \geq 4)}{\mathbb{P}(-(1/2)\ln(U) \geq 1)} \\
&= \frac{\mathbb{P}(\ln(U) \leq -8)}{\mathbb{P}(\ln(U) \leq -2)} \\
&= \frac{\mathbb{P}(U \leq \exp(-8))}{\mathbb{P}(U \leq \exp(-2))} \\
&= \frac{\exp(-8)}{\exp(-2)} = \exp(-6) \approx \boxed{0.002478}.
\end{aligned}
$$

Independence can be viewed in terms of conditional probabilities.

> **Fact 16**
> Two random variables X and Y are independent if and only if for all A and B with $\mathbb{P}(Y \in B) > 0$,
> $$\mathbb{P}(X \in A | Y \in B) = \mathbb{P}(X \in A). \tag{6.1}$$

Proof. Suppose X and Y are independent. Let A and B be such that $\mathbb{P}(Y \in B) > 0$. Then

$$
\mathbb{P}(X \in A | Y \in B) = \frac{\mathbb{P}(X \in A, Y \in B)}{\mathbb{P}(Y \in B)} = \frac{\mathbb{P}(X \in A)\mathbb{P}(Y \in B)}{\mathbb{P}(Y \in B)} = \mathbb{P}(X \in A).
$$

On the other hand, suppose (6.1) holds for all A and B such that $\mathbb{P}(Y \in B) > 0$. Let A be any measurable set.

Let B be any measurable set with $\mathbb{P}(Y \in B) = 0$. Then $\{X \in A, Y \in B\} \subseteq \{Y \in B\}$.

$$
\mathbb{P}(X \in A, Y \in B) \leq \mathbb{P}(Y \in B) = 0 = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).
$$

Now let B be any measurable set with $\mathbb{P}(Y \in B) > 0$. Then

$$
\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A | Y \in B)\mathbb{P}(Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).
$$

In either case, the probability of the intersection is the product of the probabilities, and so X and Y are independent. $\qquad \square$

In other words, X and Y are independent if knowing some information about Y does not change the distribution of X.

6.1   Other ways of viewing conditioning

As long as $\mathbb{P}(A > 0)$,

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

A different way of viewing this formula is to say that given the information that A occurred, we are no longer working in outcome space Ω, we are instead working within A. So that becomes our new probability outcome space. But remember that one of our rules of probability is that the probability of the event that the outcome is in the space is 1. So we need to renormalize our probabilities to make that happen. That is why we divide $\mathbb{P}(A, B)$ by $\mathbb{P}(A)$. That way,

$$\mathbb{P}(A|A) = \frac{\mathbb{P}(A, A)}{\mathbb{P}(A)} = 1.$$

A third way of looking at conditional probability is as a two-stage experiment. Suppose that we have a product that might have property A, and might have property B. What is the chance that the product has both properties?

Well, suppose we send the product to two rooms for testing. The first room tests to see if the product has property A. If it does have property A, then the product goes on to the second room, which tests to see if it has property B. If the product does not have property A, then it is immediately destroyed by throwing it into the nearest volcano.

The chance that the first room passes the product along is $\mathbb{P}(A)$. The chance that the second room passes the product along is $\mathbb{P}(B|A)$ (since room two only tests for property B if the product passed the first test.) The chance that the product passes both rooms is $\mathbb{P}(A)\mathbb{P}(B|A)$. This must equal the chance that the product has both properties, so

$$\mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B|A).$$

6.2   Reminder: the difference between disjoint and independent

There are two important properties that a pair of events can have. The first is that they are disjoint. This means that both events cannot occur at the same time. For instance, it cannot both rain and not rain on a particular day.

Disjoint events with positive probability can never be independent, since

$$\mathbb{P}(AB) = \mathbb{P}(\emptyset) = 0,$$

while $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$ imply that $\mathbb{P}(A)\mathbb{P}(B) > 0$.

If two events A and B are disjoint

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$

Independent means that knowing that one event occurred does not change the probability that the other event occurred. So

$$\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B|A) = \mathbb{P}(A)\mathbb{P}(B).$$

Our shorthand is that the probability function turns the union of disjoint sets into a sum. It turns the intersection of independent sets into a product. Note that this is similar (but not exactly the same) as how unions and intersections work with indicator functions:

$$\mathbb{1}(A \cup B) = \mathbb{1}(\mathbb{1}(A) + \mathbb{1}(B) > 0), \ \mathbb{1}(AB) = \mathbb{1}(A)\mathbb{1}(B).$$

Problems

6.1 Suppose $\mathbb{P}(A|B) = 0.3$ and $\mathbb{P}(B) = 0.8$. What is $\mathbb{P}(AB)$?

6.2 The chance of rain on Tuesday is 40%. Given that it rains on Tuesday, the chance of rain on Wednesday is 50%. What is the chance that it rains on both Tuesday and Wednesday?

6.3 Suppose $\mathbb{P}(A) = 0.3$ and $\mathbb{P}(B) = 0.5$, and we know that A and B are independent. What is $\mathbb{P}(A|B)$?

6.4 Suppose A and B are independent with $\mathbb{P}(A) = 0.35$ and $\mathbb{P}(B) = 0.21$. What is $\mathbb{P}(A|B)$?

6.5 Let $X \sim \mathsf{Unif}(\{1, 2, 3, 4, 5, 6\})$.

    a) What is $\mathbb{P}(X = 5|X \geq 3)$?

    b) What is $\mathbb{P}(X = 5|X \geq 6)$?

6.6 The chance of being diagnosed with non-Hodgkin lymphoma in a year is approximately 0.000194. What is the chance that two friends (assuming independence) being diagnosed with lymphoma in the same year?

6.7 Let $X \sim \mathsf{Unif}(\Omega)$, where Ω is a finite set. Let $A \subseteq \Omega$. Let Y have the same distribution as X conditioned on $X \in A$. Prove that $Y \sim \mathsf{Unif}(A)$.

6.8 Suppose that (U_1, U_2) is uniform over $[0, 1] \times [0, 1]$. Find

$$\mathbb{P}(U_1 \geq 0.5|U_1 \geq 3U_2).$$

6.9 A lab occasionally has small leaks of chemicals in the experimental space. Each leak is independent of the others and has a 90% chance of being benign, and a 10% chance of being toxic. The lab director has two drones at her disposal. The first drone can detect whether or not any toxic leaks are in the lab. The second drone can count the number of leaks present in the lab.

The drones are sent in: the first reports that yes, there is at least one toxic leak in the lab. The second drone reports there are exactly three leaks in the lab.

Conditioned on this information, what is the chance that there is exactly one toxic leak, and two benign leaks?

6.10 Suppose $\mathbb{P}(X \in A) = 0.2$, $\mathbb{P}(X \in B) = 0.7$, $\mathbb{P}(X \in C) = 0.4$, and $\mathbb{P}(X \in AC) = 0.15$. What is $\mathbb{P}(X \in A|X \in C)$?

6.11 For $U \sim \mathsf{Unif}([2, 10]$, what is $\mathbb{P}(U \leq 3|U \leq 5)$?

6.12 Let X_1, X_2, X_3 be iid $\mathsf{Unif}(\{1, 2, \ldots, 6\})$. Let $R = \min\{X_1, X_2, X_3\}$. What is $\mathbb{P}(R = 6|R \geq 3)$?

Binomials and Bayes' Rule

Question of the Day Let p denote a chance that an experiment is a success. Initially, say that $p \sim \mathsf{Unif}(\{0, 0.1, 0.2, \ldots, 0.8, 0.9, 1\})$. The experiment is conducted independently six times, with the result that there were two successes and four failures. What is the new distribution of p given this information?

Summary The **binomial distribution** (write $N \sim \mathsf{Bin}(n, p)$) is the number of successful experiments when n independent experiments with probability of success p are conducted. For $i \in \{0, 1, \ldots, n\}$,

$$\mathbb{P}(N = i) = \binom{n}{i} p^i (1-p)^{n-i},$$

where $\binom{n}{i}$ (read n choose i) is $n!/[i!(n-i)!]$.
Bayes' Rule is a way of turning around conditional probabilities. It says that as long as A and B are events with probability greater than 0,

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

In the question of the day, an experiment is being run independently multiple times. The number of successes N in such a situation is said to have a *binomial distribution.* Once we understand the binomial distribution, we will understand what $[N|p]$ is, that is, the distribution of N given p.

But in the qotd, we want the reverse, we want to know $[p|N]$. *Bayes' Rule* uses the conditional probability formula to accomplish this turnaround, and gives us a methodical way of handling such problems.

7.1   The Binomial distribution

To understand the binomial distribution, start with a concrete example.

Example 15

Suppose $p = 0.4$, and N is the number of times that an experiment that is run independently 6 times is a success. What is $\mathbb{P}(N = 4)$?.

Answer   Consider a sequence of experiments with 4 successes. For example, SSFSFS is such a sequence. This particular sequence (by independence) has chance

$$p \cdot p \cdot (1 - p) \cdot p \cdot (1 - p) \cdot p = p^4 (1 - p)^2$$

chance of happening. Note that we put a factor of p every time there is a success, and a factor of $1 - p$ every time there is a failure.

Every sequence with 4 S's and 2 F's will have chance $p^4 (1 - p)^2$ chance of occurring. How many such sequences are there? Well, the first F could appear in the first spot, which leaves 5 places for the second F. Or it could be in the second spot, leaving 4 places for the second F. Adding up all the possibilities gives $5 + 4 + 3 + 2 + 1 = 15$ places. Hence the total probability is

$$15(0.4)^4 (0.6)^2 = 0.13824 \approx \boxed{0.1382}.$$

We can generalize this to obtain the binomial distribution.

Definition 21

Suppose that an experiment with probability p of success is repeated independently n times. If N is the number of successes, say that N has the **binomial distribution** with parameters n and p. Write $N \sim \mathsf{Bin}(n, p)$.

The term binomial comes from the binomial coefficients $\binom{n}{i}$.

Definition 22

The **binomial coefficient** $\binom{n}{i}$ (read n choose i) is the number of sequences in $\{S, F\}^n$ that have exactly i components labeled S.

Fact 17

The formula for the binomial coefficient is

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

In Example 15 earlier, we needed $\binom{6}{4} = 6!/[4!2!] = 6 \cdot 5/[1 \cdot 2] = 15$ which we found directly, so at least the formula works in that case!

Proof. Consider the number of permutations of $\{1, 2, \ldots, n\}$. This is a sequence in $\{1, \ldots, n\}^n$ where each component has a different label. A permutation can be found by first choosing which of n elements is first, leaving $n - 1$ elements, and so on down to 1. So the number of permutations is $n \cdot (n - 1) \cdot (n - 2) \cdots 1 = n!$.

Or we could have chosen which spots the elements $\{1, \ldots, i\}$ go in, then permuted these elements in $i!$ ways, and then permuted the remaining elements in $(n - i)!$ ways. These both count the same

thing, so they are equal, and

$$n! = \binom{n}{i} \cdot i! \cdot (n-i)!.$$

Rearranging the terms completes the proof. □

Fact 18
For $N \sim \text{Bin}(n, p)$ and $i \in \{0, 1, \ldots, n\}$

$$\mathbb{P}(N = i) = \binom{n}{i} p^i (1-p)^{n-i}$$

where $\binom{n}{i}$ is the number of sequences in $\{S, F\}^n$ with i components labeled S.

Proof. Let $i \in \{0, 1, \ldots, n\}$. Then every sequence with i S's will have $n - i$ F's, and so will have probability $p^i (1-p)^{n-i}$ by independence. The number of such sequences is $\binom{n}{i}$ by definition. □

So for the question of the day, we now know how to calculate the distribution of N given p. But can we do the reverse? Can we calculate p given N? That's what Bayes' Rule is all about

7.2 Bayes' Rule

Bayes' Rule (or Bayes' Theorem or Bayes' formula) tells us how to "flip" conditional probabilities around.

Theorem 1 (Bayes' Rule)
Suppose A and B are events where $\mathbb{P}(A)$ and $\mathbb{P}(B)$ are both greater than 0. Then

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

Proof. Note that
$$\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B|A) = \mathbb{P}(B)\mathbb{P}(A|B).$$

Dividing both sides by $\mathbb{P}(B)$ gives the result. □

Armed with Bayes' Rule, we are now ready to take on the qotd.

Qotd We are given that $[N|p] \sim \text{Bin}(6, p)$, and want to find $[p|N = 2]$. Let $\Omega = n\{0, 0.1, \ldots, 0.9, 1\}$. Then since $p \in \Omega$ with probability 1, even conditioned on $N = 2$, p will still be in Ω. So want we want to find is

$$\mathbb{P}(p = \alpha | N = 2)$$

for all $\alpha \in \Omega$

Using Bayes' Rule gives

$$\mathbb{P}(p = \alpha | N = 2) = \frac{\mathbb{P}(N = 2 | p = \alpha)\mathbb{P}(p = \alpha)}{\mathbb{P}(N = 2)} = \frac{\binom{6}{2}\alpha^2(1-\alpha)^4(1/11)\mathbb{1}(\alpha \in \Omega)}{\mathbb{P}(N = 2)}$$
$$= C\alpha^4(1-\alpha)^2\mathbb{1}(\alpha \in \Omega).$$

So now all we need to do is find C. To do this, note that from the total probability rule,

$$\sum_{a \in \Omega} \mathbb{P}(p = \alpha | N = 2) = \sum_{a \in \Omega} C\alpha^2 (1 - \alpha)^4 = 1$$

so solving for C gives

$$C = \left[\sum_{a \in \Omega} \alpha^2 (1 - \alpha)^4\right]^{-1}.$$

If the size of Ω is small, this can be done by hand. Here $\#(\Omega) = 11$, so the following R code does the work for us.

```
alpha <- seq(0, 1, by = 0.1)
print(alpha^2 * (1 - alpha)^4)
C = 1 / sum(alpha^2 * (1 - alpha)^4)
print(C * alpha^2 * (1 - alpha)^4)
```

The result is given in the following table. The second column is the unnormalized probabilities for p given $N = 2$. The third column is the second column divided by the sum of the entries in the second column to make sure that the entries sum to 1. The third column is the answer to the question.

α	$\alpha^2(1 - \alpha)^4$	$\mathbb{P}(p = \alpha | N = 2)$
0	0	0
0.1	0.006561	0.06891
0.2	0.016384	0.1720
0.3	0.21609	0.2269
0.4	0.020736	0.2178
0.5	0.015625	0.1641
0.6	0.009216	0.09680
0.7	0.003969	0.04168
0.8	0.001024	0.01075
0.9	.000081	0.0.0008507
1	0	0
sum	0.0925	1

7.3   Variants of Bayes' Rule

In using $\mathbb{P}(X \in A | Y \in B) = \mathbb{P}(Y \in B, X \in A)\mathbb{P}(X \in A)/\mathbb{P}(Y \in B)$, it is often helpful to break B into two disjoint sets, $B = A^C B + AB$. Then we get the following variant of Bayes' Rule:

Fact 19

Suppose A and B are events with $\mathbb{P}(A)$, $\mathbb{P}(A^C)$, and $\mathbb{P}(B)$ all positive. Then

$$\mathbb{P}(A | B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^C)\mathbb{P}(A^C)}.$$

Example 16

Students who study for an exam (which they do with probability 30%) are 95% likely to pass. Students who do not study are only 80% likely to pass. Given that a student passes, what is the chance that they studied?

Answer Let S be the event that they study, and P be the event that they pass. Then

$$\mathbb{P}(S|P) = \frac{\mathbb{P}(P|S)\mathbb{P}(S)}{\mathbb{P}(P|S)\mathbb{P}(S) + \mathbb{P}(P|S^C)\mathbb{P}(S^C)}$$

$$= \frac{(0.95)(0.3)}{(0.95)(0.3) + (0.8)(0.7)} \approx \boxed{33.72\%}.$$

Because the studying does not change the passing rate much in a relative sense, it does not affect the conditional probability much. Now suppose that we consider not just passing, but who gets an A on the exam.

Example 17

Students who study for an exam (which they do with probability 30%) are 60% likely to get an A. Students who do not study are are only 10% likely to get an A. Given that a student gets an A, what is the chance that they studied?

Answer Let S be the event that they study, and A the event that they get an A on the exam. Then

$$\mathbb{P}(S|A) = \frac{\mathbb{P}(A|S)\mathbb{P}(S)}{\mathbb{P}(A|S)\mathbb{P}(S) + \mathbb{P}(A|S^C)\mathbb{P}(S^C)}$$

$$= \frac{(0.6)(0.3)}{(0.6)(0.3) + (0.1)(0.7)} \approx \boxed{72\%}.$$

The set $\{A, A^C\}$ form what is called a *partition* of the set Ω.

Definition 23

Sets $A_1, \ldots, A_n$ **partition** Ω if they are disjoint and their union is Ω.

Fact 20

For a partition $A_1, \ldots, A_n$ of Ω and any event B,

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

Proof. Note that

$$B = B \cap \Omega = B \cap (A_1 \cup A_2 \cup \cdots \cup A_n) = (B \cap A_1) \cup \cdots \cup (B \cap A_n).$$

Since the A_i are disjoint, the events $B \cap A_i$ are also disjoint. Hence

$$\mathbb{P}(B) = \mathbb{P}(B \cap A_1) + \cdots + \mathbb{P}(B \cap A_n).$$

Now we can use the conditional probability formula on each of these to say

$$\mathbb{P}(B \cap A_i) = \mathbb{P}(A_i)\mathbb{P}(B|A_i).$$

The result follows.

☐

Problems

7.1 Suppose $X \sim \mathsf{Bin}(10, 0.2)$. What is $\mathbb{P}(X \geq 2)$?

7.2 Suppose $X \sim \mathsf{Bin}(10, 0.23)$. What is $\mathbb{P}(X \leq 2)$?

7.3   a) What is 5 choose 2?

   b) How many ways are there to arrange the letters AABBB?

7.4 Each letter in a DNA sequence is equally likely to be from $\{A, G, C, T\}$. What is the chance that exactly 10 out of 40 letters in a sequence are A?

7.5 How many sequences using letters F and S are of length 10 and have exactly 8 S letters?

7.6 Consider the number of sequences of 10 letters using F and S that have exactly 8 S letters. The sequence must start with either an F or an S.

   a) If it starts with an F, then the remaining 9 letters must have exactly 8 S letters. How many ways can this happen?

   b) If the sequence starts with an S, then the remaining 9 letters must have exactly 7 S letters. How many ways can this happen?

   c) Add the results from the last two parts to find the total number of 10 letter sequences with exactly 8 S letters.

7.7 Suppose $N \sim \mathsf{Bin}(10, 0.3)$. What is $\mathbb{P}(N = 8)$?

7.8 A drug trial has 18 participants, each of which is expected (independently) to be a success with probability 0.2. What is that chance that one or zero participants achieves success?

7.9 Suppose that $[X|N] \sim \mathsf{Unif}(\{1, 2, 3, \ldots, N\})$ and $N \sim \mathsf{Unif}(\{1, 2, 3, 4, 5, 6\})$. What is

$$\mathbb{P}(N = 3|X = 2)?$$

7.10 Dimer Pharmaceuticals creates 3 types of drugs for a particular illness. The first is effective in 50% of patients, the second in 37%, and the third in 5%.

   a) If a patient is equally likely to receive any of the three drugs, what is the probability that the drug is effective on their illness?

   b) If the drug is effective for the patient, what is the probability that the drug was of the third type.

   c) If the first drug is given to 300 people, approximate the probability that it is effective for exactly 145 people using Stirling's formula.

7.11 Autotomic Industries produces two types of pain relievers that here we will call A and B for simplicity. Type A relieves pain in 40% of patients, while type B relieves pain in 20% of patients.

A patient takes one of the painkillers (they do not know which type) and relieves their pain. What is the chance that they used type A?

7.12 Archytas Manufacturing has four factories for their new laptops. Each laptop manufactured has a small chance of failure. Factory 1 has a 0.03% chance of failure, Factory 2 has a 0.02% chance, Factory 3 has a 0.07% chance, and Factory 4 has a 0.01% chance.

    a) If a laptop is equally likely to come from each of the four factories, what is the overall chance that it is defective?

    b) In a laptop is defective, what is the chance that it came from Factory 1?

    c) Investigation reveals that the defective laptop came from Factory 1 or 2. Now what is the probability that it came from Factory 1?

7.13 Bets on red and black on a roulette table pay even odds, which means if you bet x dollars and win, you get back your x dollar bet plus x more dollars. If you lose, then you lose your x dollar bet.

Suppose you repeatedly bet the same amount of money on red at a roulette table for twenty spins of the wheel. On an American Roulette wheel there are 18 out of 38 spaces that are red, and the ball is equally likely to land in any of the spaces.

    a) Find the probability that at the end of the twenty games you are ahead (so you have more money than when you started.)

    b) Find the probability that at the end of the twenty games you are behind (so you have less money than when you started.)

    c) Find the probability that at the end of the twenty games you have broken even.

7.14 The Happy Eyes LASIK medical center owns three machines for performing surgery. Use of the first machine in surgery results in a successful operation with 95% of patients, the second is successful 97% of the time, and the third machine results in successful surgery 99% of the time.

Incoming patients are randomly assigned a machine for surgery: 50% have their surgery done on the first machine, while 20% have it done on the second, and 30% on the third.

    a) Given that the surgery is not a success for a patient, what is the chance that it was done using the first machine?

    b) Given that the surgery is not a success, and either the first or second machine was used, what is the chance that it was the second machine that was used?

7.15 A psychology experiment is trying to determine if soothing music played before an exam increases test scores by 10% or more. They believe that there is a 40% chance that playing the music will improve the score. If their hypothesis is valid, and they run the experiment on 20 students, what is the chance that at least 10 show improvement?

Densities for continuous random variables

Question of the Day Suppose $T \sim \text{Exp}(3)$. Find the density (pdf) and cumulative distribution function (cdf) of T.

Summary Some random variables have a probability density function, also known as the density or the pdf. A discrete random variable X has density f with respect to counting measure if for all countable sets A

$$\mathbb{P}(X \in A) = \sum_{a \in A} f(a).$$

A continuous random variable X has density with respect to Lebesgue measure f if for all measurable A,

$$\mathbb{P}(X \in A) = \int_{a \in A} f(a)\, dA.$$

For both discrete and continuous random variables, the cumulative density function, or cdf is defined as

$$F_X(a) = \mathbb{P}(X \leq a).$$

8.1 *Differentials*

When dealing with continuous random variables, it will be helpful to have the notion of a differential. Intuitively, given a random variable t, the differential of t, written dt, is an infinitesimally small change in the value of the variable t.

Differentials can be used both for setting up derivatives and setting up integrals. The derivative of a function f that maps the variable x to the variable y is often written

$$f'(x) = \frac{dy}{dx},$$

indicating that the derivative is the small chance in y resulting from a small change in x. We can write the small change in y is

$$dy = f(x + dx) - f(x),$$

making

$$f'(x) = \frac{f(x + dx) - f(x)}{dx}.$$

This type of equation above is an informal way of thinking about derivatives. There are several ways to make this thinking precise, one way is to use limits, in which case

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

Here h which is approaching zero is a stand in for dx, the infinitesimally small change in x.

For the integral application, the differential will also mean a small interval or set that surrounds t. That is, use dt to refer to an interval around the variable t that is infinitesimally small.

With this, we can create the notion of a *differential rectangle* that measures the area under a curve $g(t)$ that lies in the differential interval dt about t.
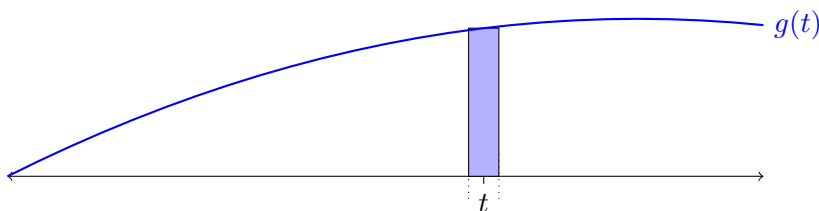
The area of the differential rectangle will be $g(t)\,dt$ since the height is $g(t)$ and the width is dt.

To find the total area under the curve from a to b, we have to sum up the area under all the differential rectangles. This is called integration (integral means whole, and we want the whole area) and is represented by a stretched out S for sum:

$$\text{area} = \int_{t=a}^{b} g(t)\,dt.$$
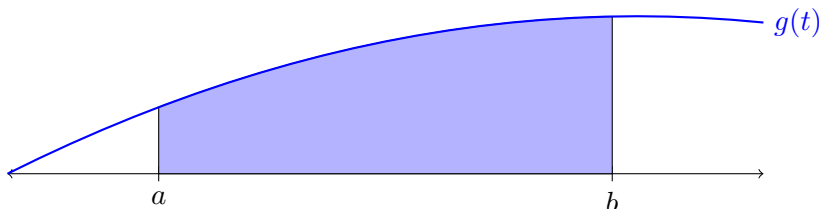
8.2   Differentials and probability

Write $\mathbb{P}(X \in dt)$ to indicate the probability that the random variable X falls into this infinitesimally small set dt around t.

Then to find $\mathbb{P}(X \in A)$, integrate $\mathbb{P}(X \in dt)$ for $t \in A$.

Example 18

Suppose $\mathbb{P}(X \in dx) = 3x^2 \mathbb{1}(x \in [0,1])\,dx$. Find $\mathbb{P}(X \in [0.4, 0.6])$.

Answer   This gives rise to the integral

$$\mathbb{P}(X \in [0.4, 0.6]) = \int_{x=0.4}^{0.6} 3x^2 \mathbb{1}(x \in [0,1])\,dx = x^3 \big|_{0.4}^{0.6} = \boxed{0.1520}.$$

In the example, note that

$$\frac{\mathbb{P}(X \in dt)}{dt} = 3x^2 \mathbb{1}(x \in [0,1])$$

forms a type of derivative. This is a generalization of the derivative called a Radon-Nikodym derivative, or as it is more commonly known, a *density*.

Of course, the total probability rule must still be true, so integrating a density over $\mathbb{R} = (-\infty, \infty)$ should give you 1.

Example 19
Suppose $\mathbb{P}(X \in dx) = C \exp(-2x)\mathbb{1}(x \geq 0) \, dx$. What is C?

Answer   Here

$$\mathbb{P}(X \in \mathbb{R}) = \int_{\mathbb{R}} C \exp(-2x)\mathbb{1}(x \geq 0) \, dx$$
$$= \int_{x=0}^{\infty} C \exp(-2x) \, dx$$
$$= C \exp(-2x)/(-2)|_0^{\infty}$$
$$= C/2.$$

So $C/2 = 1$, and $\boxed{C = 2}$.

Definition 24
Say that f_X is the **density** (a.k.a. **probability density function** or **pdf**) of continuous random variable X if for all measurable events A,

$$\mathbb{P}(X \in A) = \int_{s \in A} f(s) \, ds.$$

Remark   Perhaps the most common confusion in probability is between the terms *density* and *distribution*. The distribution of a random variable is a function $\mathbb{P}_X$ that maps an event A into $\mathbb{P}(X \in A)$. That is to say, $\mathbb{P}_X(A) = \mathbb{P}(X \in A)$. The density of X is a different function f that can be used to calculate the distribution using integration, but they are definitely not the same function!

Confusing the density and the distribution is like confusing the *integrand*, which is the function $f(s)$ being integrated, and the *integral*, which is $\int_A f(s) \, ds$. This distinction becomes important later on in statistics, where the distribution is the *statistical model* while the density is the *likelihood*.

8.3   The cdf and densities

Recall that the *cumulative distribution function*, or *cdf*, of a random variable Y is

$$\text{cdf}_Y(a) = F_Y(a) = \mathbb{P}(Y \leq a).$$

Fact 21

If X is a continuous random variable, then the cdf of X, F_X, is differentiable at all but a countable number of places. Furthermore, if

$$f_X(x) = F'_X(x)$$

everywhere $F_X(x)$ is differentiable, then $f_X(x)$ is a density for the random variable X.

The fact that $F_X(x)$ is differentiable in all but a countable number of places is an advanced fact from real analysis. Another advanced fact is that if $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for all sets A of the form $(-\infty, a]$, then X and Y have the same distribution. This is known as *Caratheodory's Theorem*. With these two ideas, the fact can be shown.

Proof. Let $A = (-\infty, a]$. Then

$$\mathbb{P}(X \in A) = F_X(x) = \int_{-\infty}^{a} F'_X(x) \, dx = \int_{-\infty} f_X(x) \, dx.$$

Hence $f_X(x)$ is a density for a random variable whose probabilities on closed infinite intervals matches that of X. Hence $f_X(x)$ is a density for X. □

With this fact in hand, the question of the day can be answered! Recall that a random variable X has the exponential distribution with parameter 3 if

$$X = -\frac{1}{3} \ln(U),$$

where U is uniform over $[0, 1]$. Since $U \in (0, 1)$ with probability 1, $\ln(U)$ is negative, and so with probability 1 $X \geq 0$.

Hence $\text{cdf}_X(a) = \mathbb{P}(X \leq a) = 0$ for all $a \leq 0$. Now assume $a > 0$. Then

$$\begin{aligned}
\text{cdf}_X(a) &= \mathbb{P}(X \leq a) \\
&= \mathbb{P}(-(1/3) \ln(U) \leq a) \\
&= \mathbb{P}(\ln(U) \geq -3a) \\
&= \mathbb{P}(U \geq \exp(-3a)),
\end{aligned}$$

which is $1 - \exp(-3a)$ since for $a > 0 \exp(-3a) \in (0, 1)$. That means that

$$\boxed{\text{cdf}_X(a) = [1 - \exp(-3a)]\mathbb{1}(a \geq 0)}.$$

The graph of the cdf looks like
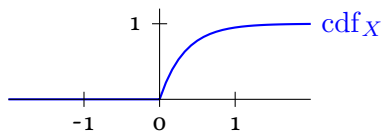
Differentiating then gives the density

$$\boxed{f_X(a) = 3\exp(-3a)\mathbb{1}(a \geq 0).}$$
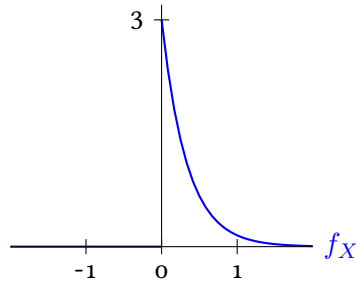
Remarks

- Since the cdf is $\mathbb{P}(X \in (-\infty, a])$ increases as a increases, the density (slope) of the cdf is always nonnegative.

- The cdf is a probability is always lies between 0 and 1. However, the pdf is the derivative of this function, and so might be larger than 1, as in our example.

- For a and b, statements like $\mathbb{P}(X \leq a)$ and $\mathbb{P}(X \geq b)$ are called *tails* of the distribution. The probability of these tails has to go to 0 as $a \to -\infty$ and $b \to \infty$. In terms of the cdf, this means

$$\lim_{a \to -\infty} \mathrm{cdf}_X(a) = 0, \quad \lim_{a \to \infty} \mathrm{cdf}_X(a) = 1.$$

- When you differentiate a function that has an indicator function as a function, the indicator function can be brought outside of the derivative. That is because the derivative of the constant 0 function is 0, and the derivative of the constant 1 function is 1 times the derivative of the rest of the factors in the function.

- The function $\mathrm{cdf}_X(a) = 3 \exp(-3a)\mathbb{1}(a \geq 0)$ does not have a derivative at 0, as there is a sharp bend in the function. That is okay: you can redefine (or define arbitrarily) the density of a continuous random variable at a countable number of places without changing the distribution of the function it represents. That is because

$$\int_a^a f(s)\,ds = 0$$

for any value of a.

- That means that the density of continuous random variables is not unique, there are an infinite number of functions f that could be the density.

8.4 Normalizing densities

Suppose we only know a density up to a normalizing constant. For instance, suppose

$$f_X(s) = Cs^2\mathbb{1}(s \in [0, 1]).$$

Can we figure out what the constant C must be?

  Yes! We use the following fact.

Fact 22

For a continuous random variable X with density f_X,

$$\int_{-\infty}^{\infty} f_X(s)\, ds = 1.$$

Proof. Note

$$\int_{-\infty}^{\infty} f_X(s)\, ds = \mathbb{P}(X \in (-\infty, \infty)) = 1.$$

$\square$

Example 20

Suppose X has density $f_X(s) = Cs^2 \mathbb{1}(s \in [0,1])$. Find C.

Answer. Note

$$1 = \int_{-\infty}^{\infty} Cs^2 \mathbb{1}(s \in [0,1])\, ds$$

$$= C \int_0^1 s^2\, ds$$

$$= C \cdot \left. \frac{s^3}{3} \right|_0^1$$

$$= C/3.$$

Solving then gives $\boxed{C = 3}$.

8.5 Scaling and shifting random variables

Often we need to move (shift) random variables and stretch (scale) them out to better model our needs

Definition 25

For a random variable A, say that $B = a + bA$ has been **shifted** by a and **scaled** by b.

How does that affect the density? In a pretty straightforward fashion.

Fact 23

Let X have density $f_X(s)$ with respect to Lebesgue measure. Then for all $a \neq 0$ and $b \in \mathbb{R}$,

$$f_{aX+b}(s) = |1/a| f_X((s-b)/a).$$

Proof. Find the cdf of $aX + b$. First consider when $a > 0$.

$$\mathrm{cdf}_{aX+b}(s) = \mathbb{P}(aX + b \leq s) = \mathbb{P}(X \leq (s-b)/a) = \mathrm{cdf}_X((s-b)/a).$$

Differentiating then gives the result.

The case when $a < 0$ is similar. $\square$

Problems

8.1 Suppose $X = \sqrt{U}$ where $U \sim \mathsf{Unif}([0, 1])$. Find the density of X.

8.2 Suppose that X has density $f_X(s) = (x^2/9)\mathbb{1}(x \in [0, 3])$.

    a) Find $\mathbb{P}(X \in [0, 1])$.

    b) Find a value m such that $\mathbb{P}(X \leq m) = 0.5$. (For continuous random variables, such a value m is called a *median* of the distribution of X or more simply a median of X.)

8.3 Suppose $f_X(s) = \exp(-s)[1 - \exp(-2)]^{-1}\mathbb{1}(s \in [0, 2])$.

    a) What is $\mathbb{P}(X \geq 1.1)$?

    b) What is $\mathbb{P}(X \leq -0.5)$?

    c) Graph F_X, the cdf of X.

8.4 The average weight of chickens (in kg) on a poultry farm is modeled as having density

$$f(s) = 25(x - 1.8)\mathbb{1}(x \in [1.8, 2]) + 25(2.2 - x)\mathbb{1}(x \in [2, 2.2])$$

    a) What is the probability that a chicken weighs more than 2.1 kilos?

    b) What is the probability that a chicken weighs more than 2.5 kilos?

8.5 Suppose W has density $f_W(x) = 3x^2\mathbb{1}[x \in [0, 1])$. What is the density of $Y = 3W + 2$?

8.6 Suppose U has distribution $\mathsf{Unif}([-1, 1])$.

    a) Find the density of U.

    b) Find the density of $-2U + 1$.

8.7 Suppose X has density $f_X(x) = C/(1 + x^2)$. What is C?

8.8 Suppose $\mathbb{P}(Y \in dy) = Cy \exp(-2y)\mathbb{1}(y \geq 0)\,dy$. What is C?

8.9 Suppose $f_T(t) = 2\exp(-2t)\mathbb{1}(t \geq 0)$. Find the density of $2T + 1$.

8.10 Suppose $f_Z(z) = \tau^{-1/2}\exp(-z^2/2)$. For $\sigma > 0$ and $\mu \in \mathbb{R}$, find the density of

$$\mu + \sigma Z.$$

8.11 Let $U \sim \mathsf{Unif}([-2, 2])$.

    a) Let $T = U^3$. What is the density of T?

    b) Let $V = U^4$. What is the density of V?

8.12 Suppose $U \sim \mathsf{Unif}([-\tau/2, \tau/2])$ and $X = \arctan(U)$. Find the density of X.

8.13 Show that if T has an exponential distribution with rate λ, then $\lfloor T \rfloor + 1$ has a geometric distribution and find the parameter p of the geometric as a function of λ.

8.14 Show that if T has an exponential distribution with rate λ, and $\alpha > 0$, then $\alpha T \sim \mathsf{Exp}(\lambda/\alpha)$.

Chapter 9

Densities for discrete random variables

Question of the Day Suppose $U \sim \mathsf{Unif}(\{1, 2, 3, 4\})$. Find the density and cumulative distribution function of U.

Summary For a discrete random variable X, the density of X is $f_X(i) = \mathbb{P}(X = i)$. The density is given with respect to counting measure. However, while the density of continuous random variables is called the probability density function or pdf, the density of discrete random variable is often called the probability mass function or pmf instead.

The cdf of X is defined in the same way as for continuous random variables:

$$\mathrm{cdf}_X(a) = F_X(a) = \mathbb{P}(X \leq a).$$

The last section introduced the notion of a density that was a derivative of probability with respect to Lebesgue measure. This applied to continuous random variables.

$$\frac{\mathbb{P}(X \in ds)}{ds} = f_X(s).$$

What is the situation when the random variable is discrete?

Consider a concrete example. Suppose

$$\mathbb{P}(X = 0) = 0.2, \ \mathbb{P}(X = 1) = 0.3, \ \mathbb{P}(X = 2) = 0.5. \tag{9.1}$$

Now suppose we put a tiny infinitesimal interval around a point 2. Then no matter how small the interval is, the chance that X falls in the interval is 0.5. So $\mathbb{P}(X \in ds) = 0.5$ for $s = 2$.

Similarly, if $s = 2.4$, then for an infinitesimally small interval around 2.4, there is 0 chance that X falls into the interval. In general, we have

$$\mathbb{P}(X \in ds) = \mathbb{P}(X = s).$$

What about ds? Well, ds only contains the point s in a discrete measure, and the counting measure of $\{s\}$ is exactly 1. Hence

$$\frac{\mathbb{P}(X \in ds)}{ds} = \frac{\mathbb{P}(X = s)}{1} = \mathbb{P}(X = s)$$

when we are working with discrete random variables.

> **Definition 26**
> For X a discrete random variable, the **density** (a.k.a. **pdf** or **probability mass function** or **pmf**) is
> $$f_X(i) = \mathbb{P}(X = i).$$

Note that for continuous random variables we only have the term probability density function, or pdf for the density, but for discrete distributions it is also sometimes called the probability mass function or pmf. This goes back to the view of 1 unit of probability as 1 unit of mass, say a kilogram of clay. This mass is then broken up and spread out over the possible values attained by the random variable.

For the random variable X defined in (9.1), the density will be

$$f_X(i) = 0.2\mathbb{1}(X = 0) + 0.3\mathbb{1}(X = 1) + 0.5\mathbb{1}(X = 2).$$

In the Question of the Day, we can write the density of $U \sim \mathsf{Unif}(\{1, 2, 3, 4\})$ as

$$f_U(i) = (1/4)\mathbb{1}(i \in \{1, 2, 3, 4\})$$

since all the probabilities are the same.

9.1   CDF for discrete random variables

The cdf for discrete random variables is defined in exactly the same way as for continuous ones:

$$\mathrm{cdf}_X(a) = F_X(a) = \mathbb{P}(X \le a).$$

The difference is that while the cdf of a continuous random variable is continuous, the cdf for a discrete random variable has jumps.

Consider again $U \sim \mathsf{Unif}(\{1, 2, 3, 4\})$. For $a < 1$, $\mathbb{P}(U \le a < 1) = 0$, so the cdf stays flat at 0. But when $a = 1$, $\mathbb{P}(U \le 1) = 1/4$. There is a jump of size $1/4$ at $a = 0$.

Then $\mathbb{P}(U \le 1.1) = \mathbb{P}(U \le 1.5) = \mathbb{P}(U \le 1.9999) = 1/4$: the cdf stays flat until we hit $a = 2$, at which point $\mathbb{P}(U \le 2) = (1/4) + (1/4) = 1/2$. There is again a jump of size $1/4$.

The graph looks as follows.
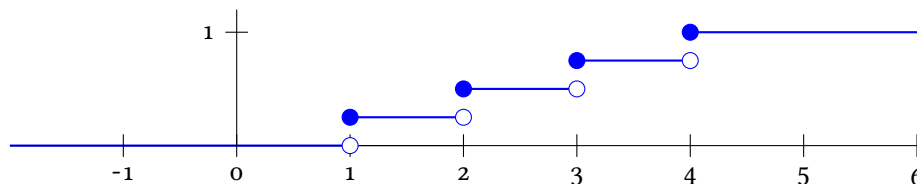
The filled circle indicates what the function value is at a jump. So at 2, the filled circle is at height 0.5 and the empty circle is at height 0.25. Hence the function value at that point is 0.5.

If we have the cdf, how can we determine $\mathbb{P}(X = a)$? Well, that is just the size of the jump in the function at a. For continuous random variables, the cdf is continuous and so all the "jumps" have size 0. In general, we can calculate this as follows.

> **Fact 24**
> For a random variable X with cdf F_X,
> $$\mathbb{P}(X = s) = F_X(s) - \lim_{h \to 0} F_X(s - |h|).$$

Shifting and scaling work slightly differently for densities of discrete random variables.

> **Fact 25**
> Let X have density $f_X(s)$ with respect to counting measure. Then for all $a \neq 0$ and $b \in \mathbb{R}$,
> $$f_{aX+b}(s) = f_X((s-b)/a).$$

Proof. Here
$$f_{aX+b}(s) = \mathbb{P}(aX + b = s) = \mathbb{P}(X = (s-b)/a) = f_X((s-b)/a).$$

$\square$

Note that there is no $|1/a|$ factor as in the Lebesgue measure case.

9.2 The maximum function and cdf's

Consider independent random variables X and Y with cdf functions F_X and F_Y. What is the cdf of $\max\{X, Y\}$?

Consider a special case: for $\max\{X, Y\} \leq 3$, then both X and Y have to be at most 3. That happens with probability equal to $\mathrm{cdf}_X(3)\,\mathrm{cdf}_Y(3)$. In other words, for the maximum operator, the cdf is the product of the cdf's of the individual random variables.

> **Fact 26**
> Suppose that X and Y are independent random variables. Then
> $$\mathrm{cdf}_{\max\{X,Y\}}(a) = \mathrm{cdf}_X(a)\,\mathrm{cdf}_Y(a).$$

What if we want to get a handle on minimums? In this case we use that
$$\mathbb{P}(\min\{X, Y\} > a) = \mathbb{P}(X > a)\mathbb{P}(Y > b).$$

> **Definition 27**
> The function $S_X(t) = \mathbb{P}(X > t)$ is called the **survival function** for X.

It is called the survival function because if X measures the amount of time an object survives before breaking down, $S_X(t)$ measures the probability the item lives longer than t time.

> **Fact 27**
> For independent random variables X and Y,
> $$S_{\min\{X,Y\}}(t) = S_X(t)S_Y(t).$$

9.3 Medians and Modes

Whether dealing with a continuous or discrete density, there are several places of interest in the density.

For a random variable X, the *median* is a value m such that $\mathbb{P}(X \leq m)$ and $\mathbb{P}(X \geq m)$ are both at least $1/2$.

Definition 28

A random variable X has a **median** m if $\mathbb{P}(X \leq m) \geq 1/2$ and $\mathbb{P}(X \geq m) \geq 1/2$. The set of m that are medians is the **median set**.

Fact 28

If X has a continuous cdf function then the median set consists of solutions to $\mathrm{cdf}_X(m) = 1/2$.

This leads to the following relationship to densities.

Fact 29

If X has density $f_X(s)$ with respect to μ, then any solution to

$$\int_{-\infty}^{m} f_X(s)\, d\mu = \frac{1}{2}.$$

is a median.

Example 21

Let $X \sim \mathsf{Unif}(\{1,2,3,4,5\})$. then 3 is the unique median of X (here $\mathbb{P}(X \geq 3) = 0.6 = \mathbb{P}(X \leq 3)$.)

Let $Y \sim \mathsf{Unif}(\{1,2,3,4\})$ then any $m \in [2,3]$ is a median of Y.

Another place of interest for a density is the place where the density is as large as possible. This is called a *mode* of the density.

Definition 29

The **mode set** of a density $f(x)$ is $\arg\max(f(x))$. Elements of the mode set are called **modes**.

In other words, the set of modes is the set of arguments x that make the function $f(x)$ as large as possible.

Example 22

Suppose $X \in \{1, 2, \ldots\}$ has density $f(i) = (24/\tau^2)i^{-2}$. Find the mode.

Answer   Since $(24/\tau^2)i^{-2}$ is strictly decreasing for $i \in \{1, 2, \ldots\}$, to find the mode we must look for i as small as possible, which in this case makes the mode set $\boxed{\{1\}}$.

Now for a continuous example.

Example 23
Find the mode(s) of $f(x) = x \exp(-x)\mathbb{1}(x \geq 0)$.

Answer When $x < 0$, $f(x) = 0$, so there are no modes there.

$$[x \exp(-x)]' = x' \exp(-x) + x[\exp(-x)]' = (1 - x) \exp(-x),$$

which has a unique solution at $x = 1$. Since $f'(x) < 0$ for $x > 1$ and $f'(x) > 0$ for $x < 1$, this critical point is a global maximum over $(0, \infty)$, and the mode set is $\boxed{\{1\}}$.

Unfortunately, not all densities are differentiable everywhere. In fact, not all densities are continuous everywhere!

Example 24
Find the mode of X where $X \sim \mathsf{Exp}(\lambda)$.

Answer The density of X is $\lambda \exp(-\lambda s)\mathbb{1}(s \geq 0)$. When $s < 0$, $f_X(s) = 0$. For $s > 0$, $[f_X(x)]' = -\lambda^2 \exp(-\lambda s) < 0$. Hence $f_X(x)$ is decreasing on $[0, \infty)$, and $\arg\max f_X(x) = \boxed{\{0\}}$, and that is the only mode.

Note that if we change the density slightly to $g(x) = \lambda \exp(-\lambda s)\mathbb{1}(s > 0)$, the distribution is unchanged. However, this density does not have a mode! At $x = 0$ we have $g(0) = 0$, but $g(x)$ gets larger and larger the closer we approach o from the right. So technically g does not have a maximizing value, and so there is no mode for this density.

So unlike the mean or median of a random variable, the mode is not a function of the distribution, but is a function of the density.

Problems

9.1 For X with density $f_X(i) = 0.3\mathbb{1}(i = 1) + 0.7\mathbb{1}(i = 4)$, what is $\mathbb{P}(X \leq 2)$?

9.2 Suppose $f_X(i) = 0.3\mathbb{1}(i = 2) + 0.2\mathbb{1}(i = 4) + 0.5\mathbb{1}(i = 5)$.

    a) What is $\mathbb{P}(X \geq 2.5)$?

    b) Graph the cdf of X.

9.3 Let U_1 and U_2 be iid $\mathsf{Unif}(\{1, 2, 3, 4\})$. Find the density of $U_1 + U_2$.

9.4 Let U_1, U_2, U_3 be iid $\mathsf{Unif}(\{1, 2, 3, 4, 5, 6\})$, and $X = \max\{U_1, U_2, U_3\}$.

    a) Find $\mathrm{cdf}_F(a)$.

    b) What is $\mathbb{P}(X = 4)$?

9.5 Suppose $X \sim \mathsf{Unif}(\{1, \dots, 10\})$. What is the mode set of X?

9.6 Suppose X has density

$$f_X(i) = 0.3\mathbb{1}(i = 1) + 0.4\mathbb{1}(i = 7) + 0.3\mathbb{1}(i = 10).$$

Find the mode set of X.

9.7 Suppose X has density $x^2 \exp(-x)\mathbb{1}(x \geq 0)$. Find the mode(s) of X.

9.8 Suppose Y has density $105x^2(1-x)^4\mathbb{1}(x \in [0,1])$. Find the mode(s) of Y.

9.9 Suppose $X \sim \mathsf{Exp}(1)$ and $Y \sim \mathsf{Exp}(2)$ are independent.

    a) Find the survival function of X.

    b) Find the survival function of Y.

    c) Find $\mathbb{P}(\min(X,Y) \geq 2)$.

9.10 Let U_1, U_2, U_3 be iid $\mathsf{Unif}([0,1])$. Find the cdf of $\min(U_1, U_2, U_3)$.

Mean of a random variable

Question of the Day What is the average value of a random variable that equals 2 with probability 0.3, 3 with probability 0.5, and 6 with probability 0.2?

Summary For some random variables, when you take the sample average of many independent draws from the same distribution, it converges towards a real number called the **mean**, **average**, **expectation**, or **expected value** of the random variable. When X is discrete with density f_X, the mean is defined to be

$$\mathbb{E}[X] = \sum_{a \in A} a f_X(a).$$

The **Strong Law of Large Numbers** states that for a random variable X where $\mathbb{E}[|X|] < \infty$,

$$\mathbb{P}\left(\lim_{n \to \infty} \frac{X_1 + \cdots + X_n}{n} = \mathbb{E}[X]\right) = 1.$$

Consider a random variable B that is $\text{Bern}(p)$, so it equals 1 with probability p, and 0 otherwise.

Now let $B_1, B_2, \ldots$ be an iid stream of draws from B. Then consider the *sample average* of the first n draws:

$$S_n = \frac{B_1 + B_2 + B_3 + \cdots + B_n}{n}.$$

The sum $B_1 + B_2 + \cdots + B_n$ counts the total number of 1's in the first n draws. For instance, $1 + 0 + 0 + 1 + 0 = 2$, since there are 2 ones in $(1, 0, 0, 1, 0)$.

From our intuitive understanding of probability, it seems reasonable that this sample average should converge to the value of p. One of the great breakthroughs in probability was when Jacob Bernoulli proved in 1713 that this intuition is correct: S_n converges in some sense to p. This result was so important that Bernoulli random variables are named in his honor. Today, we have an even stronger version of his result.

Fact 30

For $B_1, B_2, \ldots$ an iid stream of $\text{Bern}(p)$ random variables,

$$\mathbb{P}\left(\lim_{n \to \infty} \frac{B_1 + B_2 + \cdots + B_n}{n} = p\right) = 1.$$

This idea can be used to define the *expected value* of a Bernoulli random variable.

> **Definition 30**
> The **expected value** (aka **expectation** aka **mean** aka **average** of $B \sim \text{Bern}(p)$ is $\mathbb{E}[B] = p$.

Mathematically, a limit is an example of a *linear operator*.

> **Definition 31**
> For a vector space V with scalars S, say that $\mathcal{L}$ is a **linear operator** if for all $v, w \in V$ and $s, t \in S$,
> $$\mathcal{L}(sv + tw) = s\mathcal{L}(v) + t\mathcal{L}(w).$$

Examples of linear operators include:

- **Matrix-point multiplication** Here vectors are points $v \in \mathbb{R}^n$, and $\mathcal{L}(v) = Av$ for some matrix A. Then for real values s, t and vectors v and w,
  $$\mathcal{L}(sv + tw) = A(sv + tw) = s(Av) + t(Aw) = s\mathcal{L}(v) + t\mathcal{L}(w).$$

- **Differentiation** Here the vectors are differentiable functions, the scalars are real numbers, and
  $$[sf + tg]' = sf' + tg'.$$

- **Integration** Here vectors are integrable functions, scalars are real numbers, and
  $$\int_{x \in A} [sf + tg](x)\, dx = s \int_{x \in A} f(x)\, dx + t \int_{x \in A} g(x)\, dx.$$

- **Limits of Sequences** Here vectors are sequences, scalars are real numbers, and provided $\{a_n\}$ and $\{b_n\}$ are sequences with limits,
  $$\lim_{n \to \infty} (sa_n + tb_n) = s \lim_{n \to \infty} a_n + t \lim_{n \to \infty} b_n.$$

So suppose that $A \sim \text{Bern}(p)$ and $B \sim \text{Bern}(q)$. Say $A_1, A_2, \ldots \sim A$ are iid and $B_1, B_2, \ldots \sim B$ are iid. But for each i, A_i and B_i might not be independent of each other.

For instance, suppose $U_1, U_2, \ldots$ are iid $\text{Unif}([0, 1])$. Then if $A_i = \mathbb{1}(U_i \leq 0.2)$ then $A_i \sim \text{Bern}(0.2)$. If $B_i = \mathbb{1}(U_i \leq 0.3)$ then $B_i \sim \text{Bern}(0.3)$. However, A_i and B_i are *not* independent, since if $A_i = 1$ then B_i must equal 1 as well.

Let $C_i = A_i + B_i$, and consider the limit of the sample averages of the C_i. Then
$$\frac{C_1 + \cdots + C_n}{n} = \frac{A_1 + B_1 + A_2 + B_2 + \cdots + A_n + B_n}{n}$$
$$= \frac{A_1 + \cdots + A_n}{n} + \frac{B_1 + \cdots + B_n}{n}.$$

Then if we take the limit as n approaches infinity,
$$\lim n \to \infty \frac{C_1 + \cdots + C_n}{n} = p + q$$

with probability 1.

This establishes a useful fact about expected value that we will use over and over again: expected value is also also a linear operator.

Definition 32

Say that a random variable is **integrable** if it has a finite expected value.

Fact 31

Consider as vectors the set of integrable random variables, and let real numbers be scalars. Then expected value is a linear operator.

Hence for any two random variables X and Y and scalars a and b,

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

For instance, this allows us to find the expected value of a binomial random variable.

Fact 32

For $X \sim \mathsf{Bin}(n, p)$, $\mathbb{E}[X] = np$.

Proof. Note for $B_1, \ldots, B_n$ iid $\mathsf{Bern}(p)$,

$$B = B_1 + \cdots + B_n \sim \mathsf{Bin}(n, p).$$

Taking the expectation of both sides gives

$$\mathbb{E}[B] = \mathbb{E}[B_1] + \cdots + \mathbb{E}[B_n] = p + \cdots + p = np.$$

$\square$

For the question of the day, to use our rule about Bernoulli random variables, we must write X as the sum of indicator random variables. That is very easy to do!

For $X \in \{2, 3, 6\}$,

$$X = 2\mathbb{1}(X = 2) + 3\mathbb{1}(X = 3) + 6\mathbb{1}(X = 6).$$

For instance, when $X = 3$, we get 3 on the left hand side, and $(2)(0) + (3)(1) + (6)(0) = 3$ on the right hand side. The $X = 2$ and $X = 6$ cases are similar.

Now take the expected value of both sides:

$$\mathbb{E}[X] = 2\mathbb{E}[\mathbb{1}(X = 2)] + 3\mathbb{E}[\mathbb{1}(X = 3)] + 6\mathbb{E}[\mathbb{1}(X = 6)].$$

The mean of an indicator function is the probability that the indicator function equals 1. So for instance, $\mathbb{E}[\mathbb{1}(X = 2)] = \mathbb{P}(X = 2) = 0.3$.

Hence

$$\mathbb{E}[X] = 2\mathbb{P}(X = 2) + 3\mathbb{P}(X = 3) + 6\mathbb{P}(X = 6) = 2(0.3) + 3(0.5) + 6(0.2) = 3.3.$$

This procedure could be followed every time it is necessary to calculate an expected value. But this idea can be easily generalized, and inspires our definition of expected value.

Definition 33

Suppose Ω satisfies $\sum_{x \in \Omega} \mathbb{P}(X = x) = 1$. Then the **expected value** (aka **mean**, **average**, **expectation**) of X is

$$\mathbb{E}[X] = \sum_{x \in \Omega} x\mathbb{P}(X = x)$$

provided this limit exists.

Notation 4

We can also write this as an integral with respect to counting measure.

$$\sum_{x \in \Omega} x\mathbb{P}(X = x) = \int_{x \in \Omega} x\mathbb{P}(X \in x)\, d\# = \int_{x \in \Omega} x\mathbb{P}(X \in dx).$$

The fact that the limit of the sample averages equals the expected value of the random variable is called the Strong Law of Large Numbers.

Theorem 2 (Strong Law of Large Numbers)

Let X have finite expectation, and $X_1, X_2, \ldots \sim X$ be an iid sequence. Then

$$\mathbb{P}\left(\frac{X_1 + \cdots + X_n}{n} = \mathbb{E}[X]\right) = 1.$$

Remarks:

- Nothing is said about how quickly the sample average converges to the expected value. It could be very slow or very fast. We will learn more about this later when we study the variance of a random variable.

- The convergence only happens with probability 1. This means that there do exist sequences where the convergence does not happen. But the probability of any of these unusual sequences happening is zero.

- For example, suppose we have a random variable $X \sim \mathsf{Unif}(\{1, 2, 3, 4, 5, 6\})$. Consider an iid sequence of die rolls $X_1, X_2, \ldots$ where $X_i \sim X$ for all i. One possible sequence of die rolls is $1, 3, 1, 3, 1, 3, 1, 3, \ldots$. For this sequence, the sample average converges to 2 rather than 3.5. But the chance of getting this particular sequence is 0. The SLLN says that when you sum over all sequences where the sample average does not converge to 3.5, the total probability of all those bad sequences is still 0.

10.1   Symmetry

Another useful property of means is that if the density is symmetric around a number, the mean of the random variable equals that number.

Definition 34

A function f is **symmetric** around m if

$$(\forall \delta \in \mathbb{R})(f(m + \delta) = f(m - \delta)).$$

For example, $f(x) = (b - a + 1)^{-1}\mathbb{1}(x \in \{a, a + 1, \ldots, b\}$ is symmetric around $(a + b)/2$. Note that if $(a + b)/2$ is an integer, then $f(m + \delta)$ and $f(m - \delta)$ are both 1 if and only if δ is an integer at most $(b - a)/2$. If $(a + b)/2$ is not an integer, then $f(m + \delta)$ and $f(m - \delta)$ are both 1 if and only if $\delta = k/2$ where k is an integer at most $b - a$.

Definition 35

A random variable is **symmetric** around m if X and $2m - X$ (which is also $m - (X - m)$) have the same distribution.

> **Fact 33**
> If a random variable X has a density which is symmetric around m then X is symmetric around m.

Proof. We will show that they have the same cdf. Let $a \in \mathbb{R}$. Then

$$\begin{aligned}
\mathbb{P}(X \leq a) &= \sum_{s \leq a} f_X(s) \\
&= \sum_{s \leq a} f_X(s + m - m) \\
&= \sum_{s \leq a} f_X(m + (s - m)) \\
&= \sum_{s \leq a} f_X(m - (s - m)) \qquad \text{by symmetry} \\
&= \sum_{s \leq a} f_X(2m - s) \\
&= \mathbb{P}(2m - X \leq a).
\end{aligned}$$

$\square$

Now we can use linearity to show that an integrable symmetric random variable has mean equal to the point of symmetry.

> **Fact 34**
> Let X be an integrable random variable symmetric about m. Then
>
> $$\mathbb{E}[X] = m.$$

Proof. By linearity

$$2m = \mathbb{E}[2m] = \mathbb{E}[X + 2m - X] = \mathbb{E}[X] + \mathbb{E}[2m - X].$$

Since X is symmetric, $\mathbb{E}[X] = \mathbb{E}[2m - X]$. Hence $2m = \mathbb{E}[X] + \mathbb{E}[X] \Rightarrow \mathbb{E}[X] = m$. $\square$

In particular, this gives the expected value for discrete uniforms.

> **Fact 35**
> For $U \sim \mathsf{Unif}(\{a, a + 1, \ldots, b\})$, $\mathbb{E}[U] = (a + b)/2$.

Note that this symmetry rule only works if the random variable is integrable. Consider $R \in \{\ldots, -2, -1\} \cup \{1, 2, 3, \ldots\}$ where

$$\mathbb{P}(R = i) = (3/\pi^2)|i|^{-2}.$$

Then the mean of R does not exist, even though R is symmetric about 0.

Problems

10.1 Given that $\mathbb{P}(Y = 2) = 0.4$ and $\mathbb{P}(Y = -1) = 0.6$, what is $\mathbb{E}[Y]$?

10.2 Say that $\mathbb{P}(R = 0) = 0.3$, $\mathbb{P}(R = 2) = 0.45$ and $\mathbb{P}(R = 3) = 0.25$. What is $\mathbb{E}[R]$?

10.3 Suppose $\mathbb{P}(X = 2) = 0.3$, $\mathbb{P}(X = 4) = 0.2$ and $\mathbb{P}(X = 5) = 0.5$. What is $\mathbb{E}[X]$?

10.4 Let W have density

$$W = (1/10)\mathbb{1}(i \in \{1, 2, 3, 4\}) + (2/10)\mathbb{1}(i \in \{5, 6, 7\}).$$

What is $\mathbb{E}[W]$?

10.5 Suppose $\mathbb{E}[X] = 34$. What is $\mathbb{E}[2X - 5]$?

10.6 Let $\mathbb{E}[X] = 2$. What is $\mathbb{E}[15 - 5X]$?

10.7 Say $X \sim \mathsf{Unif}(\{-2, -1, 0, 1, 2\})$. What is $\mathbb{E}[X]$?

10.8 Suppose $\mathbb{P}(X = 0) = 0.15$, and $\mathbb{P}(X = 2) = 0.65$, $\mathbb{P}(X = 7) = 0.2$. What is $\mathbb{E}[X]$?

10.9 Say $\mathbb{E}[R] = 3$ and $\mathbb{E}[S] = 6$. What is $\mathbb{E}[R - S]$?

10.10 If $\mathbb{E}[Z_1] = \mu_1$ and $\mathbb{E}[Z_2] = \mu_2$, what is $\mathbb{E}[2Z_1 + 4Z_2]$?

10.11 Suppose $U_1, U_2, \ldots \sim \mathsf{Unif}(\{1, 2, 3, 4\})$. Show that $\lim_{n \to \infty}(U_1 + \cdots + U_n)/n = 2.5$ with probability 1.

10.12 Suppose $\mathbb{P}(X = i) = (6/\pi^2)i^{-2}$ for all $i \in \{1, 2, \ldots\}$. Show that X is not integrable.

Chapter 11

Expected value of general random variables

Question of the Day Let $T \sim \mathsf{Exp}(4.5)$. What is the average value of T, denoted $\mathbb{E}[T]$?

Summary The **expected value** of a real valued function g of a real valued random variable X is

$$\mathbb{E}[g(X)] = \int_{a \in \mathbb{R}} g(a) \mathbb{P}(X \in da).$$

For X with density $f_X(s)$ with respect to Lebesgue measure, this means

$$\mathbb{E}[g(X)] = \int_{a \in \mathbb{R}} g(a) f_X(a) \, da,$$

and if the density is with respect to counting measure,

$$\mathbb{E}[g(X)] = \sum_a g(a) f_X(a).$$

Basic **Monte Carlo** algorithms operate by constructing a random variable whose mean is the quantity of interest, then simulating that random variable multiple times and averaging the result.

11.1 Integrals with respect to Lebesgue and counting measure

For a random variable X and event A,

$$\mathbb{P}(X \in A) = \mathbb{E}[\mathbb{1}(X \in A)] = \int_{a \in \Omega} \mathbb{1}(a \in A) \mathbb{P}(X \in da).$$

Note that in the $\mathbb{E}$ operator is a function of the random variable, namely, $\mathbb{1}(X \in A)$. In the integral, the function of the variable of integration a is the same function. This gives $\mathbb{1}(a \in A)$ inside the integral.

It turns out that this method of turning a mean into an integral applies to any other function as

71

well! For instance,

$$\mathbb{E}[X^2] = \int_{a\in\mathbb{R}} a^2 \mathbb{P}(X \in da)$$

$$\mathbb{E}[X] = \int_{a\in\mathbb{R}} a\mathbb{P}(X \in da)$$

$$\mathbb{E}[\sin(X)] = \int_{a\in\mathbb{R}} \sin(a)\mathbb{P}(X \in da).$$

When X has density $f_X(a)$ with respect to counting measure, the integral turns into a sum:

$$\mathbb{E}[g(X)] = \int_{a\in\mathbb{R}} g(a)\mathbb{P}(X \in da) = \sum_a g(a)f_X(a)$$

When X has density $f_X(a)$ with respect to Lebesgue measure, the integral acts like a regular Riemann integral:

$$\mathbb{E}[g(X)] = \int_{a\in\mathbb{R}} g(a)\mathbb{P}(X \in da) = \int_{a\in\mathbb{R}} g(a)f_X(a)\, da.$$

These rules are sometimes called *The Law of the Unconscious Statistician* because in creating the integral, just replace the random variable X in the expectation by the variable of integration a in the integral. So easy, it can be done while unconscious!

Theorem 3

Suppose X has density $f_X(s)$ with respect to Lebesgue measure. Then

$$\mathbb{E}[g(X)] = \int_{s\in\mathbb{R}} g(s)f_X(s)\, ds.$$

If X is a discrete random variable with density $f_X(s)$ with respect to counting measure,

$$\mathbb{E}[g(X)] = \sum_s g(s)f_X(s).$$

Definition 36

When $\mathbb{E}[g(X)]$ exists and is finite, call $g(X)$ **integrable**.

Note that we always apply the function g to the dummy variable s, and *not* to the density function. That always stays the same. Let's look at some examples.

Example 25

Suppose $f_X(s) = 12s^2(1 - s)\mathbb{1}(s \in [0, 1])$. Find $\mathbb{E}[X^2]$.

Answer This is

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} s^2 \cdot 12s^2(1 - s)\mathbb{1}(s \in [0, 1]) \, ds$$

$$= \int_0^1 12s^4(1 - s) \, ds$$

$$= 12[s^5/5 - s^6/6]|_0^1$$

$$= 12[1/5 - 1/6] = 12/30 = 4/10 = \boxed{0.4000}$$

This distribution comes from the *Beta* family. The specific parameters for this distribution are parameters 3 and 2. The 3 is one more than the power of s, and the 2 is one more than the power of $1 - s$ in the density. Therefore, we can test the result in R using

```
results <- rbeta(10^6,3,2)
mean(results^2)
```

which returned 0.3997603 when I ran the code.

For discrete random variables, the same rule holds: just square the values, the density stays the same.

Example 26

Suppose $Y \sim \text{Unif}(\{1, 2, 3, 4, 5, 6\})$. Find $\mathbb{E}[Y^2]$.

Answer This will be

$$\mathbb{E}[Y] = 1^2\mathbb{P}(Y = 1) + 2^2\mathbb{P}(Y = 2) + \cdots + 6^2\mathbb{P}(Y = 6)$$

$$= (1/6)[1 + 4 + 9 + 16 + 25 + 36]$$

$$= 91/6 = \boxed{15.16}.$$

We can test this in R with

```
results <- sample(1:6, 10^6, replace = TRUE)
mean(results^2)
```

which returned 15.1608 when I ran it.

Example 27
Question of the day Let $X \sim \text{Exp}(4.5)$. Then X has density

$$f_X(s) = 4.5 \exp(-4.5s)\mathbb{1}(s \geq 0).$$

Hence the expected value of X is

$$\mathbb{E}[X] = \int_{s \in \mathbb{R}} s \cdot 4.5 \exp(-4.5s)\mathbb{1}(s \geq 0)\, ds$$

$$= \int_{s=0}^{\infty} s \cdot 4.5 \exp(-4.5s)\, ds$$

At this point we need to manufacture a derivative to slide over to get rid of the s:

$$\mathbb{E}[X] = \int_{s \in \mathbb{R}} s \cdot [-\exp(-4.5s)]'\mathbb{1}(s \geq 0)\, ds$$

$$= \int_{s=0}^{\infty} [-s \exp(-4.5s)]' - [s]'[-\exp(-4.5s)]\, ds$$

$$= [-s \exp(-4.5s)]\big|_0^{\infty} - \int_{s=0}^{\infty} -\exp(-4.5s)\, ds$$

$$= \lim_{s \to \infty} -s \exp(-4.5s) - 0 - \exp(-4.5s)/4.5\big|_0^s$$

Remember the general rule:

$$\text{logarithms} \ll \text{polynomials} \ll \text{exponentials} \ll \text{factorials.}$$

Here s is growing polynomially, and $\exp(-4.5s) = 1/\exp(4.5s)$ is decreasing exponentially, so $s \exp(-4.5s)$ goes to 0 as s goes to infinity. L'Hopital's Rule can also be used to verify this fact.]
   Hence

$$\mathbb{E}[X] = 0 - 0 - (0 - 1/4.5) = \boxed{0.2222\ldots}.$$

11.2   Properties of continuous means

In Chapter 10, we noted the two most important properties of expected value.

- Linearity. For any random variables X and Y and real numbers a and b,

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

- Strong Law of Large Numbers. If X is integrable, and $X_1, X_2, \ldots$ are iid X, then

$$\mathbb{P}\left(\frac{X_1 + \cdots + X_n}{n} \to \mathbb{E}[X]\right) = 1.$$

Both of these properties hold for the expectation of $g(X)$ as well.

Example 28
Let $X \sim \mathsf{Exp}(4.5)$ and $Y \sim \mathsf{Unif}(\{1, 2, 3\})$. What is $\mathbb{E}[X + Y]$?

Answer   Here $\mathbb{E}[X] = 2/9$ and $\mathbb{E}[Y] = (1/3)(1) + (1/3)(2) + (1/3)(3) = 6/3 = 2$.
Hence
$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] = 2/9 + 2 = 20/9 = \boxed{2.222\ldots}$$
by linearity.

11.3   Applications of the SLLN

Consistent estimators in statistics   Suppose that I have a stream of incoming data $X_1, X_2, \ldots$ that I model as iid X where $\mathbb{E}[X] = \theta$. The value of θ is a parameter that I am trying to find. For instance, I might have a model where

$$X_1, X_2, \ldots \sim \mathsf{Unif}[0, 2\theta].$$

The *sample average*

$$\hat{\theta}_n = \frac{X_1 + \cdots + X_n}{n}$$

gives us a sequence of estimators of the parameter value θ.

Then we know by the SLLN that the sample average satisfies

$$\lim_{n \to \infty} \hat{\theta}_n = \theta$$

with probability 1.

When we have a sequence of estimators $\hat{\theta}_n$ that converge to the true value θ as the number of data points goes to infinity, we say that the estimate is *consistent*, and this is a very good property to have.

Monte Carlo   In Monte Carlo simulation, to evaluate an integral, first it is necessary to build a random variable X such that $\mathbb{E}[X]$ equals the value of the integral.

Example 29

Construct a random variable such that $\mathbb{E}[X] = I$, where

$$I = \int_0^1 \sqrt{x^{2.5} - \ln(x)}\, dx.$$

(Note that this function does not have an elementary antiderivative.)

Answer   Since the region of integration is $[0, 1]$, we can base our random variable off of a uniform $U \sim \mathsf{Unif}([0, 1])$. Let

$$X = g(U) = \sqrt{U^{2.5} - \ln(U)}.$$

By the formula for expectation of a function of a random variable, this gives

$$\mathbb{E}[X] = \mathbb{E}[g(U)]$$
$$= \int_{-\infty}^{\infty} \sqrt{u^{2.5} - \ln(u)}\, \mathbb{1}(u \in [0, 1])$$
$$= I.$$

With this formula created, the following code gives an estimate of the value of I.

```
u <- runif(10^6)
> mean(sqrt(u^2.5 - log(u)))
```

This gave `1.095166` when I ran it, which is pretty close to the true answer of 1.0954.

Problems

11.1 For X with density $12s^2(1 - s)\mathbb{1}(s \in [0, 1])$, find $\mathbb{E}[X]$.

11.2 Let $X \sim \mathsf{Unif}([3, 6])$. Find $\mathbb{E}[X]$.

11.3 Suppose $U_1, U_2, \ldots \sim \mathsf{Unif}([0, 4])$. Show that $\lim_{n \to \infty} (U_1 + \cdots + U_n)/n = 2$ with probability 1.

11.4 Suppose $T_1, T_2, \ldots \sim \mathsf{Exp}(2)$. Show that

$$\lim_{n \to \infty} \frac{T_1 + \cdots + T_n}{n} = 1/2$$

with probability 1.

11.5 For Z with density

$$f_Z(z) = \tau^{-1/2} \exp(-z^2/2),$$

verify using the integral that $\mathbb{E}[Z] = 0$.

11.6 Let $U = \mathsf{Unif}([0, 1])$. Then $1 - U \sim \mathsf{Unif}([0, 1])$ as well. Note that

$$\mathbb{E}[U] + \mathbb{E}[1 - U] = \mathbb{E}[U + 1 - U] = \mathbb{E}[1] = 1.$$

Given that $\mathbb{E}[U] = \mathbb{E}[1 - U]$ (since they have the same distribution), what is $\mathbb{E}[U]$?

11.7 Suppose $\mathbb{P}(X = -1) = 0.3$ and $\mathbb{P}(X = 1) = 0.7$. What is $\mathbb{E}[X^2]$?

11.8 Suppose $Y = 1/U$ where $U \sim \mathsf{Unif}([0, 1])$. Show that Y is not integrable.

11.9 Let X have density $s \exp(-s^2/2)\mathbb{1}(s \geq 0)$. Find $\mathbb{E}[X^2]$.

11.10 Let $U \sim \mathsf{Unif}([0, 1])$. Find the expected value of $\sqrt{U}$.

11.11 Build a random variable W such that $\mathbb{E}[W] = I$, where

$$I = \int_{-1}^{1} 2x^2 \, dx.$$

11.12 Build a random variable Y such that $\mathbb{E}[Y] = I$, where

$$I = \int_0^3 \exp(-x^{2.5}) \, dx.$$

11.13 For a random variable A, the *mean absolute deviation* of A is defined as

$$\mathrm{MAD}(A) = \mathbb{E}\left[|A - \mathbb{E}[A]|\right].$$

Let $A \sim \mathsf{Exp}(\lambda)$. Find $\mathrm{MAD}(A)$.

11.14 For $U \sim \mathsf{Unif}([0, 1])$, find $\mathrm{MAD}(U)$.

11.15 Three zombies are chasing you. Each runs at a speed that is independent of the other, and a continuous uniform between 6.0 and 11.0 miles per hour.

a) If you can run at 10.0 miles per hour, what is the chance that you will get away from the zombies?

b) What is the expected speed of the fastest zombie?

11.16 Two birds are flying with speed (independently of each other) uniform between 21.1 and 32.3 mph. What is the expected speed of the faster bird?

11.17 Let $U \sim \mathsf{Unif}([0, 2])$.

a) Find the cdf of $X = U^3$.

b) Find the density of X.

c) Find $\mathbb{E}[X]$.

11.18 For $A \sim \mathsf{Exp}(2)$, find $\mathbb{E}[A^3]$.

11.19 Suppose $A \sim \mathsf{Exp}(3)$, so A has density

$$f_A(s) = 3 \exp(-3s)\mathbb{1}(s \geq 0).$$

The density of an exponential is the multiplicative inverse of the rate, so $\mathbb{E}[A] = 1/3$.

a) What is $\mathbb{E}[2A - 1]$?

b) What is $\mathbb{E}[\exp(1.5A)]$?

c) What is the density of $2A - 1$?

11.20 A random variable X has the *Beta* distribution with parameters a and b if it has density

$$f_X(s) = s^{a-1}(1 - s)^{b-1}\mathbb{1}(s \in [0, 1]).$$

a) For X Beta with parameters 3 and 1, find $\mathbb{E}[X]$.

b) Find $\mathbb{E}[3X + 6]$.

c) Find $\mathbb{E}[X^2]$.

Chapter 12

Conditional Expectation

Question of the Day Suppose we have two random variables N and X such that the distribution of X depends on the value of N. Specifically, $N \sim \mathsf{Unif}(\{1,2,3,4\})$ and $[X|N] \sim \mathsf{Unif}(\{1,2,\ldots,N\})$. That means $\mathbb{E}[X|N] = (N+1)/2$. So then what is $\mathbb{E}[X]$?

Summary The conditional probability $\mathbb{E}[X|Y]$ is the average of the random variable X given the value of another random variable Y. The result is a function of Y.

If you then take the average of this function of Y over values of Y, we get back the overall average value of X. This result, that

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X],$$

is the **Fundamental Theorem of Probability**.

The Fundamental Theorem of Probability can be used to give us probability trees, expectation trees, and the mean and variance of a geometric random variable.

12.1 Conditioning on a random variable

In the question of the day, we are not directly told the distribution of X. Instead, we are told that the distribution of X given the value of another random variable N is uniform over the number 1 through N.

This information is enough to calculate the density of X. For example, consider the chance that X equals 2. Since $1 \leq X \leq N$, N must equal either 2, 3, or 4 for this to happen.

So we can break up the event:

$$\mathbb{P}(X=2) = \mathbb{P}(X=2, N=2) + \mathbb{P}(X=2, N=3) + \mathbb{P}(X=2, N=4).$$

Each of these we can break down using conditional probability, so for example,

$$\mathbb{P}(X=2, N=3) = \mathbb{P}(N=3)\mathbb{P}(X=2|N=3) = (1/4)(1/3) = 1/12.$$

Combining gives

$$\mathbb{P}(X=2) = \left(\frac{1}{4}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{4}\right)\left(\frac{1}{3}\right) + \left(\frac{1}{4}\right)\left(\frac{1}{4}\right) = \frac{1}{4}\left[\frac{1}{2} + \frac{1}{3} + \frac{1}{4}\right] = \frac{13}{48}.$$

Repeating for $i \in \{1, 2, 3, 4\}$ gives the following density for X:

$$f_X(i) = \frac{25}{48}\mathbb{1}(i = 1) + \frac{13}{48}\mathbb{1}(i = 2) + \frac{7}{48}\mathbb{1}(i = 3) + \frac{3}{48}\mathbb{1}(i = 4).$$

Now that we have a density, we can find the expected value:

$$\mathbb{E}[X] = (1)\frac{25}{48} + (2)\frac{13}{48} + (3)\frac{7}{48} + (4)\frac{3}{48} = \frac{84}{48} = \boxed{1.750}.$$

Okay, so we were able to handle this problem by direct computation, but this quickly becomes very cumbersome. For instance, even if N is just $\mathsf{Unif}(\{1, 2, \ldots, 6\})$, the amount of work almost doubles.

For a faster method, consider finding $\mathbb{E}[X|N]$. What does this mean? This is the average value of X given the value of N is fixed rather than a random variable. Normally N would be a random variable, but here we are saying that the value of N is somehow known to us. In this case, we can treat N as though it was just a regular variable. Then we know that $\mathbb{E}[X|N] = (1 + N)/2$, because $[X|N]$ is a uniform draw from 1 up to N.

Note that for $f(n) = (1 + n)/2$, this says, that

$$\mathbb{E}[X|N] = f(N).$$

That is, the *conditional expectation* $\mathbb{E}[X|N]$ is itself a function of N. And functions of random variables are themselves random variables.

> **Intuition 4**
> The **conditional expectation** $\mathbb{E}[X|Y]$ is a new random variable that equals $g(Y)$ for some function g.

12.2   *The Fundamental Theorem of Probability*

Okay, so we can calculate $\mathbb{E}[X|N] = (1 + N)/2$, but how does that get us closer to what we want, which is $\mathbb{E}[X]$? Well, once we know how the average value of X depends on N, we can get rid of the dependence on N by averaging over the different values of N. That is,

$$\mathbb{E}[\mathbb{E}[X|N]] = \mathbb{E}[X],$$

and in our problem

$$\mathbb{E}[X] = \mathbb{E}[(1 + N)/2] = (1 + \mathbb{E}[N])/2 = (1 + (1 + 4)/2)/2 = 1.75,$$

exactly as we found earlier!

This important result goes by several names, such as the *law of total expectation*. Since it is fundamental to so much of probability, and includes the conditional probability formula as a special case, we will refer to it here as the *Fundamental Theorem of Probability* or FTP.

> **Theorem 4** (Fundamental Theorem of Probability)
> For random variables X and Y where $\mathbb{E}[X]$ and $\mathbb{E}[X|Y]$ are always finite,
>
> $$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X].$$

12.3 Expectation and Probability Trees

In the FTP, $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$. The right hand side has an inner mean, $\mathbb{E}[X|Y]$, and then an outer mean that surrounds it. In the question of the day, it made sense to evaluate the inner mean first, but sometimes it makes more sense to evaluate the outer mean first, then the inner mean. Consider the following example.

Example 30

Suppose that when the economy does well (which happens with probability 0.3), the average increase in a stock price is 20%. When the economy does moderately (chance 0.5), the average increase in stock prices is 5%, and when the economy does poorly (chance 0.2), the average increase in stock prices is -10%. What is the average increase in stock prices?

Answer From the FTP, we do not need to know the distribution of stock prices to answer this question, the average change is enough! Let $S \in \{1, 2, 3\}$ denote the state of the economy (1=poor, 2=moderate, 3=well) and T the change in the stock. Then

$$\mathbb{E}[T] = \mathbb{E}[\mathbb{E}[T|S]] = \mathbb{E}[T|S = 1]\mathbb{P}(S = 1) + \mathbb{E}[T|S = 2]\mathbb{P}(S = 2) +$$
$$\mathbb{E}[T|S = 3]\mathbb{P}(S = 3)$$
$$= (0.2)(0.3) + (0.05)(0.5) + (-0.10)(0.2) = 0.065 = \boxed{6.500\%}.$$

This type of calculation can be represented graphically by using an *expectation tree*.
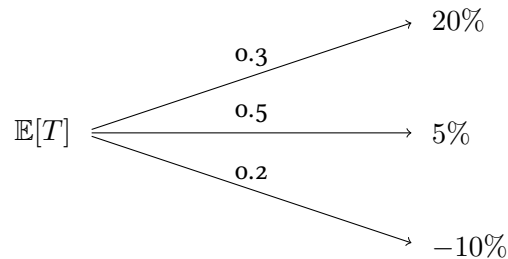
Along the three possible branches we put the probability of each branch. The sum of the weights of the branches should always equal 1. At the end of each branch we put the expected value should that branch occur.

Then to calculate the value of the expectation tree, multiply the weight of the branches times the expectation at the end of the branch, and sum up the result (in this case 6.5%.)

Remember that $\mathbb{E}(\mathbb{1}(X \in A)) = \mathbb{P}(X \in A)$, that is, probabilities are a special case of expectations. When the FTP is applied to probabilities it often goes by the name *the law of total probability*.

Fact 36 (Law of total probability)

Suppose $A_1, A_2, \ldots$ are disjoint events such that one is true with probability 1. Then

$$\mathbb{P}(B) = \sum_{i=1}^{\infty} \mathbb{P}(B|A_i)\mathbb{P}(A_i).$$

Proof. Let T be a random variable that takes on values in $\{1, 2, \ldots\}$. Specifically, let $T = i$ if A_i is true. Then by the FTP,

$$
\begin{aligned}
\mathbb{E}[\mathbb{1}(B)] &= \mathbb{E}[\mathbb{E}[\mathbb{1}(B)|T]] \\
&= \sum_{i=1}^{\infty} \mathbb{E}[\mathbb{1}(B)|T = i]\mathbb{P}(T = i) \\
&= \sum_{i=1}^{\infty} \mathbb{P}(B|A_i)\mathbb{P}(A_i).
\end{aligned}
$$

$\square$

Example 31

Continuing our earlier example, if the probability of event A depends on the value of S, then

$$\mathbb{P}(A) = \mathbb{P}(A|S = 1)\mathbb{P}(S = 1) + \mathbb{P}(A|S = 2)\mathbb{P}(S = 2) + \mathbb{P}(A|S = 3)\mathbb{P}(S = 3).$$

The graphical representation then becomes a *probability tree*.
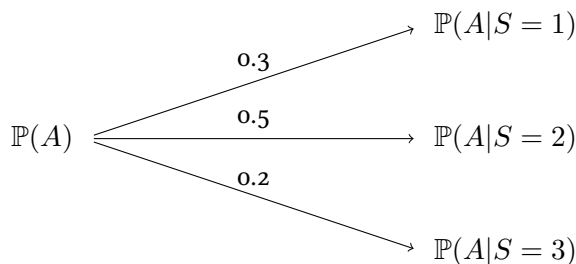
12.4   *Mean of a geometric random variable*

Recall that if $U_1, U_2, \ldots$ are iid $\mathsf{Unif}([0, 1])$, then $B_i = \mathbb{1}(U_i \leq p)$ gives rise to an iid sequence $B_1, B_2, \ldots$ of $\mathsf{Bern}(p)$ random variables.

Furthermore, if we let G be the smallest value of i such that $B_i =$, then we say G has a *geometric distribution* with parameter p.

Consider using the FTP to find $\mathbb{E}[G]$ by conditioning on B_1.

Fact 37

For $G \sim \mathsf{Geo}(p)$, $\mathbb{E}[G] = 1/p$.

Proof. Consider the first Bernoulli random variable B_1. If $B_1 = 1$, then $G = 1$, but if $B_1 = 0$, then the distribution of G is the same as the one wasted draw plus a new geometric random variable. That is,

$$[G|B_1 = 0] \sim 1 + G.$$

This statement can be checked directly:

$$
\begin{aligned}
\mathbb{P}(G = i|B_1 = 0) &= \mathbb{P}(G = i, B_1 = 0)/\mathbb{P}(B_1 = 0) \\
&= p(1 - p)^{i-1}\mathbb{1}(i > 1)/(1 - p) \\
&= p(1 - p)^{i-2}\mathbb{1}(i > 1)
\end{aligned}
$$

and

$$\mathbb{P}(1 + G = i) = \mathbb{P}(G = i - 1) = p(1-p)^{i-1}\mathbb{1}(i - 1 > 0)$$

which are the same function.

Hence

$$\mathbb{E}(G|B_1 = 0) = \mathbb{E}(1 + G) = 1 + \mathbb{E}(G).$$

Putting that into the FTP gives:

$$\begin{aligned}
\mathbb{E}(G) &= \mathbb{E}(\mathbb{E}(G|B_1)) \\
&= \mathbb{E}(G|B_1 = 0)\mathbb{P}(B_1 = 0) + \mathbb{E}(G|B_1 = 1)\mathbb{P}(B_1 = 1) \\
&= (1 + \mathbb{E}(G))(1 - p) + (1)(p).
\end{aligned}$$

Solving for $\mathbb{E}(G)$ then gives $\mathbb{E}(G) = 1/p$. □

12.5 Conditional probability formula

At the start of this chapter I said that the FTP generalizes the conditional probability formula. To see that this is true, first note that

$$\mathbb{P}(AB) = \mathbb{E}(\mathbb{1}(AB)) = \mathbb{E}(\mathbb{1}(A)\mathbb{1}(B)).$$

Then by the FTP

$$\begin{aligned}
\mathbb{P}(AB) &= \mathbb{E}[\mathbb{E}(\mathbb{1}(A)\mathbb{1}(B))|\mathbb{1}(B))] \\
&= \mathbb{E}[\mathbb{1}(A)\mathbb{1}(B)|\mathbb{1}(B) = 1]\mathbb{P}(\mathbb{1}(B) = 1) + \mathbb{E}[\mathbb{1}(A)\mathbb{1}(B)|\mathbb{1}(B) = 0]\mathbb{P}(\mathbb{1}(B) = 0) \\
&= \mathbb{E}[\mathbb{1}(A)|\mathbb{1}(B) = 1]\mathbb{P}(B) + \mathbb{E}[0|\mathbb{1}(B) = 0]\mathbb{P}(\mathbb{1}(B) = 0) \\
&= \mathbb{P}(A|B)\mathbb{P}(B),
\end{aligned}$$

which is the conditional probability formula!

Of course, the proof of the FTP utilizes the conditional probability formula, so this is just an exercise to show that the FTP is a more general form.

Problems

12.1 Suppose that B_1, B_2 are iid Bern(0.3). Say $\mathbb{P}(N = 1) = 0.6$ and $\mathbb{P}(N = 2) = 0.4$.

    a) Find the density of

$$S = \sum_{i=1}^{N} B_i.$$

    b) Find $\mathbb{E}[S]$ using the density.

    c) Find $\mathbb{E}[S]$ using the Fundamental Theorem of Probability.

12.2 Suppose I roll $N \sim \mathsf{Unif}(\{1, 2\})$. Then I roll N dice independently and identically distributed as $\mathsf{Unif}(\{1, 2, 3, 4, 5, 6\})$ and sum them to get S. That is

$$[S|N = 1] = X_1, \ [S|N = 2] = X_1 + X_2.$$

Or more compactly,

$$[S|N] = \sum_{i=1}^{N} X_i.$$

Here $X_i \sim \mathsf{Unif}(\{1, 2, 3, 4, 5, 6\})$

    a) What is the probability $S = 4$?

    b) What is the probability $S = 7$?

    c) Find the density of S, $f_S(i)$ for $i \in \{1, 2, \ldots, 12\}$.

    d) Find $\mathbb{E}[S]$ from $f_S(i)$.

    e) Find $\mathbb{E}[S]$ from the Fundamental Theorem of Probability.

12.3 A party has either low attendance (20% chance), medium attendance (40% chance) or high attendance (40% chance). With low attendance the average revenue for the party is $-\$300$, with medium $\$500$, and with high $\$1000$.

Draw an expectation tree to calculate the average revenue from the party.

12.4 Lisa and Bart go spelunking in a cave, and unfortunately, soon get lost. Each time they try to find the exit, they have a 20% chance of finding the exit in an hour, a 45% of returning back to where they started after an hour, and a 35% of returning back to where they started after three hours.

    a) What is the chance that they find their way out after exactly four hours?

    b) What is the chance that they find their way out after exactly eight hours?

    c) What is the expected amount of time they spend in the cave?

12.5 Suppose the time until arrival of a customer (call it T) is an exponential random variable with rate parameter A (so $[T|A] \sim \mathsf{Exp}(A)$.) A is a random variable that is uniform over the interval $[5, 10]$. What is $\mathbb{E}[T]$?

12.6 The probability p of success for an experiment is modeled as uniform over $[0.4, 0.5]$. Then 27 independent trials of the experiment are run. What is the expected number of successes?

Chapter *13*

Joint densities

Question of the Day Suppose (X_1, X_2) has the joint density

$$f_{(X_1, X_2)}(x_1, x_2) = \exp(-x_1 - x_2)\mathbb{1}(x_1, x_2 \geq 0).$$

What is $\mathbb{P}((X_1, X_2) \in [0, 1] \times [0, 2])$?

Summary The word **joint** is used with densities that are two dimensional or higher. Two dimensional densities are called **bivariate**. You find probabilities for bivariate densities by using a two dimensional integral. So for $A \subseteq \mathbb{R}^2$:

$$\mathbb{P}((X, Y) \in A) = \int_{(x,y) \in A} f_{X,Y}(x, y) \, d\mathbb{R}^2.$$

If $f(x_1, \ldots, x_n)$ is the density of $X_1, \ldots, X_n$, For (X, Y) with a joint density, the density of a particular X_i can be found by **integrating out** the other variables. For bivariate random variables, this means

$$f_X(x) = \int_{y \in \mathbb{R}} f_{(X,Y)}(x, y) \, dy, \ f_Y(y) = \int_{x \in \mathbb{R}} f_{(X,Y)}(x, y) \, dx.$$

For a measurable function $g : \mathbb{R}^2 \to \mathbb{R}$,

$$\mathbb{E}[g(X_1, X_2)] = \int_{(s_1, s_2) \in \mathbb{R}^2} g(s_1, s_2) f_{(X_1, X_2)}(s_1, s_2) \, d\mu$$

Recall that f_X is the *density* of X with respect to μ if we find probabilities by integrating with respect to μ. That is,

$$\mathbb{P}(X \in A) = \int_{x \in A} f_X(x) \, d\mu.$$

When μ is Lebesgue measure ℓ, we say that we have a continuous random variable and

$$\int_{x \in A} f_X(x) \, d\ell = \int_{x \in A} f_X(x) \, dx$$

When μ is counting measure #, we say X is a discrete random variable, and

$$\int_{x \in A} f_X(x) \, d\# = \sum_{x \in A} f_X(x).$$

Random variables in more than one (but still finite) dimension operate the same way. When you integrate the density, though, you are doing it over a higher dimensional space.

In this chapter, we will concentrate on what happens when you have two random variables. We call this situation *bivariate*.

> **Definition 37**
> A random variable that is a point in $\mathbb{R}^2$ is called **bivariate**.

> **Example 32**
> In the question of the day, (X_1, X_2) is a bivariate random variable.
>   If $(U_1, U_2) \sim \mathsf{Unif}([0,1]^2)$, then U_1 and U_2 form a bivariate random variable.

> **Definition 38**
> For (X, Y), say that $f_{X,Y}$ is the **density** of (X, Y) with respect to μ if for all $A \subseteq \mathbb{R}^2$,
>
> $$\mathbb{P}((X, Y) \in A) = \int_{(x,y) \in A} f_{X,Y}(x, y) \, d\mu.$$

Qotd   For the question of the day, this works out as

$$\mathbb{P}((X_1, X_2) \in [0,1] \times [0,2]) = \int_{(x_1,x_2) \in [0,1] \times [0,2]} f_{X_1,X_2}(x_1, x_2) \, d\mathbb{R}^2.$$

Since the integrand is nonnegative, we can use Tonelli's theorem to write it as an iterated integral.

$$\begin{aligned}
\mathbb{P}((X_1, X_2) \in [0,1] \times [0,2]) &= \int_{x_1 \in [0,1]} \int_{x_2 \in [0,2]} \exp(-x_1 - x_2) \mathbb{1}(x_1, x_2 \geq 0) \, dx_2 \, dx_1 \\
&= \int_{x_1 \in [0,1]} -\exp(-x_1 - x_2)|_0^2 \, dx_1 \\
&= \int_{x_1 \in [0,1]} -\exp(-x_1 - 2) - (-\exp(-x_1)) \, dx_1 \\
&= \exp(-x_1 - 2) - \exp(-x_1)|_0^1 \\
&= \exp(-3) - \exp(-1) - (\exp(-2) - \exp(0)) \approx \boxed{0.5465}.
\end{aligned}$$

When dealing with discrete random variables, this turns into a sum.

Example 33
Let (X, Y) have density $f_{X,Y}(x, y) = (1/37)(x^2 + y)$ for $x \in \{1, 2, 3\}$ and $y \in \{1, 2\}$. What is $\mathbb{P}(Y = 1)$?

Answer This is

$$\mathbb{P}(Y = 1) = \mathbb{P}((X, Y) \in \{(1, 1), (2, 1), (3, 1)\}$$
$$= (1/37)[(1 + 1) + (4 + 1) + (9 + 1)] = 17/37 \approx \boxed{0.4594}.$$

Note that this means (by complements) that $\mathbb{P}(Y = 2) = 20/37$. Since these are the only two options,

$$\mathbb{P}(Y = 1) = \frac{17}{37}, \ \mathbb{P}(Y = 2) = \frac{20}{37}$$

completely describes the distribution of Y. When we calculate the distribution of one component of jointly distributed random variables, that is called a *marginal distribution*.

Definition 39
For a random vector (X_1, X_2), the distribution of X_1 or X_2 is called a **marginal distribution**.

You can find the marginal density for a particular variable by *integrating out* the dummy variable for the other random variable.

Fact 38
Let (X, Y) have density $f_{X,Y}$ with respect to $\mu \times \nu$. Then

$$f_X(x) = \int_{y \in \mathbb{R}} f_{X,Y}(x, y) \, d\nu, \quad f_y(y) = \int_{x \in \mathbb{R}} f_{X,Y}(x, y) \, d\mu.$$

Proof. For f_X to be a density of X, it must be true that for all measurable A,

$$\mathbb{P}(X \in A) = \int_A f_X(a) \, d\mu.$$

Since $\mathbb{P}(Y \in \mathbb{R}) = 1$,

$$\mathbb{P}(X \in A) = \mathbb{P}(X \in A, Y \in \mathbb{R})$$
$$= \int_{x \in A} \left[\int_{y \in \mathbb{R}} f_{X,Y}(x, y) \, d\nu \right] d\mu$$

and so the piece inside the brackets satisfies the definition of a density of X. The argument for f_Y is similar. $\square$

Example 34

Let (X, Y) have density $f_{X,Y}(x, y) = (1/37)(x^2 + y)$ for $x \in \{1, 2, 3\}$ and $y \in \{1, 2\}$. What is the marginal distribution of X?

Answer   For $x \in \{1, 2, 3\}$,

$$f_X(x) = \sum_{y \in \{1,2\}} (1/37)(x^2 + y) = (1/37)[(x^2 + 1) + (x^2 + 2)] = \boxed{\frac{2x^2 + 3}{37}}.$$

13.1   Independence and joint densities

Independence means that the probability of multiple events factors into the product of each event. In the same way, random variables have a joint density that is independent means that the density factors into densities for each of the marginals.

Recall, that $\mu \times \nu$ is a product measure if for all A that is μ measurable and B that is ν measurable,

$$(\mu \times \nu)(A \times B) = \mu(A) \times \nu(B).$$

Our two most commonly used measures, counting measure and Lebesgue measure, use product measure in higher dimensions. That is why the area of a rectangle is the product of the lengths of the sides. For instance,

$$\ell([0, 4] \times [3, 5]) = \ell([0, 4]) \cdot \ell([3, 5]) = 4 \cdot 2 = 8.$$

If I pick a ball from a bag which contains one red, one green, and one blue balls, and then roll a six sided die, the number of possible outcomes is

$$\#(\{\text{red}, \text{green}, \text{blue}\} \times \{1, 2, 3, 4, 5, 6\}) = 3 \cdot 6 = 18.$$

Fact 39

Suppose random variables X and Y have a joint density with respect to product measure $\mu \times \nu$ that factors into a piece that only involves one input and another piece that only involves the other input. That is, it has the form

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

where f_X is a density with respect to μ and f_Y is a density with respect to ν. Then f_X is a density of X, f_Y is a density of Y, and X and Y are independent random variables.

Proof. Let $f_{X,Y} = f_X(x) f_Y(y)$. Then $\mathbb{P}(Y \in \mathbb{R}) = 1$, so for any μ measurable set A,

$$\mathbb{P}(X \in A) = \mathbb{P}(X \in A, Y \in \mathbb{R})$$

$$= \int_{x \in A} \int_{y \in \mathbb{R}} f_X(x) f_Y(y) \, d\nu \, d\mu$$

$$= \int_{x \in A} f_X(x) \left[\int_{y \in \mathbb{R}} f_Y(y) \, d\nu \right] d\mu$$

$$= \int_{x \in A} f_X(x) \, d\mu.$$

This is the definition of what it means for X to have density f_X. The proof that f_Y must be the density of Y is similar.

Now for independence. Let A be μ measurable and B be ν measurable. Then

$$
\begin{aligned}
\mathbb{P}(X \in A, Y \in B) &= \int_{x \in A} \int_{y \in B} f_X(x) f_Y(y) \, d\nu \, d\mu \\
&= \int_{x \in A} f_X(x) \left[\int_{y \in B} f_Y(y) \, d\nu \right] d\mu \\
&= \int_{x \in A} f_X(x) \mathbb{P}(Y \in B) \, d\mu \\
&= \mathbb{P}(Y \in B) \int_{x \in A} f_X(x) \, d\mu \\
&= \mathbb{P}(Y \in B) \mathbb{P}(X \in A),
\end{aligned}
$$

so X and Y are independent. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

Note that there are usually multiple choices of joint density for continuous random variables. All that is needed is *one* joint density that factors for the variables to be independent.

The other direction holds as well.

> **Fact 40**
> Say X has density f_X with respect to μ and Y has density f_Y with respect to ν are independent random variables. Then (X, Y) has as a joint density $f_{(X,Y)}(x, y) = f_X(x) \cdot f_Y(y)$ with respect to $\mu \times \nu$.

13.2 Means for joint densities

The expected value of a bivariate random vector is similar to that of a single valued random variable.

> **Definition 40**
> For (X, Y) with density $f_{X,Y}$ with respect to μ, and a computable function $g : \mathbb{R}^2 \to \mathbb{R}$,
> $$
> \mathbb{E}[g(X, Y)] = \int_{x,y} g(x, y) f_{X,Y}(x, y) \, d\mu.
> $$

First a discrete example.

> **Example 35**
> Let (X, Y) have density $f_{X,Y}(x, y) = (1/37)(x^2 + y)$ for $x \in \{1, 2, 3\}$ and $y \in \{1, 2\}$. What is $\mathbb{E}[XY]$?
>
> **Answer** Since these are discrete random variables, this will be a sum
> $$
> \sum_{x=1}^{3} \sum_{y=1}^{2} xy[(1/37)(x^2 + y)] = \frac{138}{37} = \boxed{3.729\ldots}.
> $$

Let's try a continuous example.

Example 36

Suppose (X_1, X_2) has joint density

$$f_{(X_1, X_2)}(x_1, x_2) = \exp(-x_1 - x_2)\mathbb{1}(x_1, x_2 \geq 0).$$

Find $\mathbb{E}[X_1 X_2]$.

Answer This integral will be

$$\mathbb{E}[X_1 X_2] = \int_{(x_1, x_2) \in \mathbb{R}^2} x_1 x_2 \exp(-x_1 - x_2)\mathbb{1}(x_1, x_2 \geq 0) \, d\mathbb{R}^2$$

$$= \int_{x_1 \geq 0, x_2 \geq 0} x_1 x_2 \exp(-x_1 - x_2) \, d\mathbb{R}^2$$

Over the limits of integration, the integrand is nonnegative, so iterated integrals can be used to obtain

$$\mathbb{E}[X_1 X_2] = \int_{x_1 \geq 0} \int_{x_2 \geq 0} x_1 x_2 \exp(-x_1 - x_2) \, dx_2 \, dx_1 = \boxed{1}.$$

Problems

13.1 Suppose (X, Y) has density $(1/60)(x + 2y)\mathbb{1}(x \in [0, 2], y \in [0, 5])$.

    a) Find the marginal density of X.

    b) Find the marginal density of Y.

    c) Find $\mathbb{E}[XY]$.

13.2 Suppose (X, Y) has density $(x^3 + y^2)/150$ for $X \in \{1, 2, 3\}$ and $Y \in \{1, 2, 3\}$.

    a) Find the marginal distribution of X.

    b) Find $\mathbb{E}[XY]$.

13.3 Suppose (X, Y) has density

$$f_{X,Y}(x, y) = (1/1260)x^3 y^2 \mathbb{1}(x \in \{1, 2, 3\})\mathbb{1}(y \in \{1, 3, 5\}).$$

    a) Prove that X and Y are independent.

    b) What is $\mathbb{P}(X = 2)$?

13.4 Suppose (X, Y) has density

$$\frac{2\sqrt{2}}{\tau} \exp(-x^2 - 2y^2).$$

Prove that X and Y are independent. A fact about integrals that is useful for this problem is that

$$\int_{s=-\infty}^{\infty} \exp(-s^2/2) \, ds = \sqrt{\tau}.$$

13.5 Suppose $(X_1, X_2) = (7.314, 2.103)$. What are the order statistics?

13.6 For $(Y_1, Y_2) = (2.3, -0.4, 1.6)$, what are the order statistics of the $\{Y_i\}$?

13.7 Suppose $(X_1, X_2) = (5.623, 5.623)$. What are the order statistics?

13.8 For $(U_1, U_2) \sim \mathsf{Unif}([0, 1]^2)$, what is $\mathbb{P}(U_1 = U_{(1)})$?

13.9 Suppose the order statistics $X_{(1)} = 1.3$ and $X_{(2)} = 3.4$. What possible states could the original vector (X_1, X_2) take on?

Random variables as vectors

Question of the Day Let $T \sim \mathsf{Unif}(\{1, 2, 3\}$. What is the standard deviation of T?

Summary Vector spaces consist of vectors and scalars. Vectors add together to give a new vector, and scalars multiply vectors to give a scaled vector. Inner products can be used to give a norm on the vector space.

For an integrable random variable X, the centered version is $X_c = X - \mathbb{E}[X]$. An important vector space comes from the set of centered random variables. For this vector space, the inner product is the mean of the product of the centered random variables. This is called the **covariance**. The covariance of a random variable with itself is the **variance**, the square root of that is the **standard deviation.**

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$
$$\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$
$$\mathrm{SD}(X) = \sqrt{\mathbb{V}(X)}.$$

14.1   Vector spaces

A *vector space* consists of two types of objects: vectors and scalars. If I add two vectors together, I get a new vector. If I multiply a scalar times a vector, I get a new vector as well.

The type of vectors that most people see first are vectors that measure *displacement*. These spatial vectors consist of two points, a tail and a head. For instance, if the tail is $(3, 2)$, and the head is $(6, 5)$, then the vector looks like this.
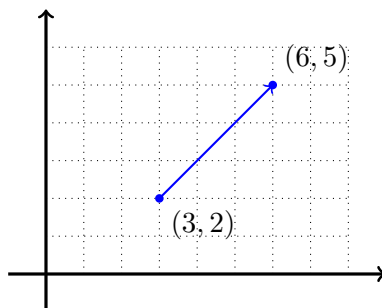
We call $(3, 2)$ the *tail* of the vector and $(6, 5)$ the *head* of the vector. To add two spatial vectors, just place the tail of one vector at the head of the second vector.
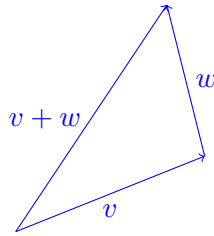
To scale a vector, we just change its length, and if the scale is negative, also reverse its direction.
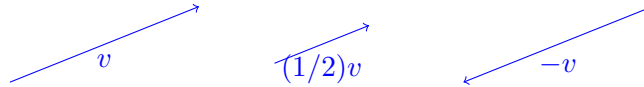
Notice that it doesn't matter where we draw the tail of the spatial vector. Because we are only concerned with the difference between the head and the tail, it is as if the tail was always placed at the origin $(0, 0)$.

To turn random variables into vectors, we will do something similar. To center the random variables, we will subtract the mean of the random variable.

Definition 41

For an integrable random variable X,

$$X_c = X - \mathbb{E}[X]$$

is the **centered** version of X.

The mean of a centered random variable is 0.

Fact 41

The centered version of an integrable random variable has mean 0.

Proof. Note $\mathbb{E}[X]$ is a constant, so

$$\mathbb{E}[X - \mathbb{E}[X]] = \mathbb{E}[X] - \mathbb{E}[X] = 0.$$

$\square$

Note that if A and B are random variables with mean 0, then $\mathbb{E}[A + B] = \mathbb{E}[A] + \mathbb{E}[B] = 0 + 0 = 0$. That means adding two vectors (centered random variables) gives a new vector.

Similarly, for any constant c, $\mathbb{E}[cA] = c\mathbb{E}[A] = 0$, so multiplying by a scalar also returns a vector.

Now to actually make sure that centered random variables form a vector space, we need to verify all the rules of a vector space.

> **Definition 42**
>
> A **vector space** is a set of vectors V together with a set of scalars S, and two operations with the following properties.
>
> 1. There is vector addition, $+$, such that $(\forall v, w \in V)(v + w \in V)$. This addition is associative and commutative. There is a zero vector 0 such that $(\forall v \in V)(v + 0 = v)$. There exist inverses, so that $(\forall v \in V)(\exists w \in V)(v + w = 0)$.
>
> 2. There is scalar multiplication, $\cdot$, such that $(\forall s \in S)(\forall v \in V)(sv \in V)$. This multiplication has an identity element $1 \in S$ so that $(\forall v)(1v = v)$. Also $(\forall a, b \in S)(\forall v \in V)((ab)v = a(bv))$. It is distributive in two ways:
>
> $$(\forall a \in S)(\forall v, w \in V)(a(v + w) = av + aw)$$
>
> and
>
> $$(\forall a, b \in S)(\forall v \in V)((a + b)v = av + bv).$$

It is straightforward to verify that these rules do apply to our centered random variables.

14.2 Norms and Inner products

The *norm* of a vector measures the size of the vector. For spatial vectors, a commonly used norm is the length of the vector. For centered random variable vectors, the norm is measuring the spread, a numerical measure of uncertainty, in the variables. In general, a norm has to satisfy the following rules.

> **Definition 43**
>
> A **norm** of a vector space takes as input a vector and returns a nonnegative real number. Write the norm of v as $\|v\|$. Then a norm must obey the following rules.
>
> 1. For any $c \in \mathbb{R}$ and vector v, $\|cv\| = |c| \, \|v\|$.
>
> 2. For any vectors v and w,
>
> $$\|v + w\| \leq \|v\| + \|w\|.$$
>
> 3. For any vector v, $\|v\| \geq 0$, and $\|v\|$ only equals 0 if v is the zero vector.

For spatial vectors, the second property means that the length of a side of a triangle must be less than the sum of the lengths of the other two sides, so this is also known as the *triangle inequality*.

The *inner product* between two vectors measures how "lined up" the two vectors are. It has the usual properties that we associate with a product such as commutativity and distribution. Because we are only using real numbers in this course, we will stick to the real definition. It is slightly different for complex numbers.

> **Definition 44**
> An **inner product** $\langle x, y \rangle$ maps vectors x, y to a real number while satisfying four properties. (In these properties u, v and w are vectors, and α is a scalar.)
>
> 1. $\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$.
>
> 2. $\langle \alpha v, w \rangle = \alpha \langle v, w \rangle$
>
> 3. $\langle v, w \rangle = \langle w, v \rangle$
>
> 4. $\langle v, v \rangle \geq 0$ where equality holds if and only if $v = 0$.

When $V = \mathbb{R}^n$, $S = \mathbb{R}$, the usual inner product is the *dot product* defined for $v = (v_1, \ldots, v_n)$ and $w = (w_1, \ldots, w_n)$ as

$$\langle v, w \rangle = v \cdot w = \sum_{i=1}^{n} v_i w_i.$$

For instance, $(3, 2, 5) \cdot (-1, 0, 2) = -3 + 0 + 10 = 7$.

We can use any inner product to form a *norm*.

> **Fact 42**
> For an inner product, $\langle \cdot, \cdot \rangle$, the function $\|v\| = \langle v, v \rangle$ is a norm.

Proof. The fourth property of inner products immediately gives the third property of norms.

To show the first property of norms, note

$$\langle cv, cv \rangle = c \langle v, cv \rangle = c \langle cv, v \rangle = c^2 \langle v, v \rangle,$$

and taking the square root of both sides and using $\sqrt{c^2} = |c|$ finishes the proof.

The second property of norms (the triangle inequality) is equivalent to a fact that we will discuss later called the Cauchy-Schwartz inequality. $\square$

> **Definition 45**
> Call $\|v\| = \sqrt{\langle v, v \rangle}$ an **inner product norm**.

For the dot product,

$$\|v\| = (v \cdot v)^{1/2} = \left(\sum_{i=1}^{n} v_i^2 \right)^{1/2}.$$

This is the same as the *Euclidean length* or L_2 *norm* of the vector.

14.3   Covariance, Variance, and Standard Deviation

So for our vector space of centered random variables, what should our inner product be? We use something very simple, the mean of the product of the two random variables. We call this the *covariance*. That is, the covariance (inner product) between two centered random variables is

$$\text{Cov}(X_c, Y_c) = \mathbb{E}[X_c Y_c].$$

If the variables are not already centered, then center them before taking the inner product.

> **Definition 46**
> The **covariance** of two integrable random variables X and Y is
> $$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$
> when this mean exists.

Linearity of expectation gives a simplification of this expression.

> **Fact 43**
> For integrable random variables X and Y with XY integrable,
> $$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

So another way to view covariance is it measures how far away the mean of the product is from the product of the means.

Proof. Note that

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY - X\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]].$$

Taking out the constants and using linearity gives

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y]$$

and simplifying gives the result. $\square$

> **Fact 44**
> Covariance is an inner product.

Proof. Let X_c, Y_c, and W_c be three centered random variables, and $a \in \mathbb{R}$. Then the properties tend to follow from linearity of expectations.

1. Distribution:

$$\begin{aligned}\text{Cov}(X_c + Y_c, W_c) &= \mathbb{E}[(X_c + Y_c)W_c] = \mathbb{E}[X_cW_c + Y_cW_c] \\ &= \mathbb{E}[X_cW_c] + \mathbb{E}[Y_cW_c] = \text{Cov}(X_c, W_c) + \text{Cov}(Y_c, W_c).\end{aligned}$$

2. Scaling:
$$\text{Cov}(aX_c, Y_c) = \mathbb{E}[aX_cY_c] = a\mathbb{E}[X_cY_c] = a\,\text{Cov}(X_c, Y_c).$$

3. Commutativity:
$$\text{Cov}(X_c, Y_c) = \mathbb{E}[X_cY_c] = \mathbb{E}[Y_cX_c] = \text{Cov}(Y_c, X_c).$$

4. Self inner product
$$\text{Cov}(X_c, X_c) = \mathbb{E}[X_c^2] \geq 0$$
since $X_c^2 \geq 0$.
Now assume $\mathbb{E}[X_c^2] = 0$. For any $\epsilon > 0$,
$$\mathbb{E}[X_c^2] \geq \mathbb{E}[X_c^2 \mathbb{1}(X_c > \epsilon] \geq \mathbb{E}[\epsilon^2 \mathbb{1}(|X_c| > \epsilon)] = \epsilon^2 \mathbb{P}(|X_c| > \epsilon)],$$
therefore for all $\epsilon > 0$, $\mathbb{P}(|X_c| > \epsilon) = 0$, and $\mathbb{P}(X_c = 0) = 1$.

□

Because covariance is an inner product, that means that it gives us an inner product norm. This norm is called the *standard deviation.* Recall the norm is the square root of the inner product of the random variable with itself. In probability $\mathrm{Cov}(X, X)$ is called the *variance* of the random variable.

> **Definition 47**
> The **variance** of an integrable random variable X is $\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$. If $\mathbb{V}(X) = \infty$, then we can say that the random variable has infinite variance, or that the variance does not exist.

Plugging $X = Y$ into $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ immediately gives the following characterization of variance.

> **Fact 45**
> The variance of a random variable is $\mathbb{E}[X^2] - \mathbb{E}[X]^2$ when these expectations exist.

Then the inner product norm itself (the square root of the variance) is called the *standard deviation*

> **Definition 48**
> The **standard deviation** of a random variable with finite variance is the nonnegative square root of the variance.

> **Fact 46**
> For a random variable X with finite variance and $c \in \mathbb{R}$,
> $$\mathbb{V}(cX) = c^2\mathbb{V}(X).$$

Proof. $\mathbb{V}(cX) = \mathrm{Cov}(cX, cX) = c \cdot c\,\mathrm{Cov}(X, X) = c^2\mathbb{V}(X).$ □

Because the standard deviation a norm, it is always nonnegative, and we have the following scaling fact.

> **Fact 47**
> For X a random variable with finite variance and $c \in \mathbb{R}$,
> $$\mathrm{SD}(cX) = |c|\,\mathrm{SD}(X).$$

Proof. $\mathrm{SD}(cX) = \sqrt{\mathbb{V}(cX)} = \sqrt{c^2\mathbb{V}(X)} = |c|\sqrt{\mathbb{V}(X)} = |c|\,\mathrm{SD}(X).$ □

Example 37

For the Question of the Day,

$$\mathbb{E}[X^2] = (1/3)(1)^2 + (1/3)2^2 + (1/3)3^2 = 14/3,$$

and $\mathbb{E}[X] = (1+3)/2 = 2$. Hence

$$\mathbb{V}(X) = 14/3 - 2^2 = 2/3 = 0.6666\ldots,$$

and

$$\mathrm{SD}(X) = \sqrt{2/3} = \boxed{0.8164\ldots}.$$

Problems

14.1 Let $v = (-1, -1, 2)$ and $w = (5, 2, -3)$.

    a) What is $v \cdot w$?

    b) What is $\|v\|$?

14.2 Suppose $v \cdot w = 4$, $v \cdot y = 6$, and $v \cdot v = 10$.

    a) What is $v \cdot (w + y)$?

    b) What is $3v \cdot (-2w)$?

    c) What is $\sqrt{(3v) \cdot (3v)}$?

14.3 Say X is discrete with density $f_X(1) = 0.7$, $f_X(5) = 0.2$, $f_X(10) = 0.1$.

    a) Find $\mathbb{E}[X]$.

    b) Find $\mathrm{SD}[X]$.

14.4 Let Y have density

$$f_Y(i) = (1/4)\mathbb{1}(i \in \{3, 5\}) + (1/6)\mathbb{1}(i \in \{7, 9, 11\}).$$

    a) Find $\mathbb{E}[Y]$.

    b) Find $\mathrm{SD}(Y)$.

14.5 Suppose $U \sim \mathsf{Unif}([0, 10])$.

    a) What is the centered random variable U_c as a function of U?

    b) What is the variance of U_c?

14.6 Let $T \sim \mathsf{Exp}(3.1)$. What is the standard deviation of T?

14.7 Let (X, Y) be uniform over

$$A = \{(x, y) : x \geq 0, y \geq 0, x + y \leq 1\}.$$

Find $\mathrm{Cov}(X, Y)$.

14.8 Suppose (X_1, X_2) have joint density $f_{(X_1,X_2)}(x_1, x_2) = (2/3)(x + 2y)\mathbb{1}((x, y) \in [0, 1]^2)$. What is $\mathrm{Cov}(X_1, X_2)$?

14.9 True or false: a random variable with finite mean always has a finite standard deviation.

14.10 Let X_1, X_2, X_3 be iid $\mathsf{Unif}(\{1, 2, \ldots, 6\})$. Let $S = \inf\{X_i\}$.

    a) What is the expected value of S?

    b) What is the variance of S?

14.11 For a random variable with finite mean μ, finite standard deviation σ, and finite third moment $\mathbb{E}[X^3]$, the *skewness* of the random variable is defined as

$$\mathrm{skew}(X) = \mathbb{E}\left[\frac{(X - \mu)^3}{\sigma^3}\right].$$

    a) If X has skewness 3, what is the skewness of $2X$?

    b) What is the skewness of $-2X$?

14.12 For $T \sim \mathsf{Exp}(1)$, find the skewness of T.

14.13 Find the skewness of $U \sim \mathsf{Unif}([0, 1])$.

14.14 Let $Z \sim \mathsf{N}(0, 1)$. Find the skewness of Z.

14.15 Topper Building Co. suffers a number of delays that is uniform over $\{0, 1, 2, 3, 4\}$. Each delay costs the builder an amount of time that is exponential with parameter 0.3 per month. Find the expectation and variance of the total delay time.

14.16 Suppose X_1, X_2, X_3 are iid uniform over $\{1, 2, \ldots, 6\}$ and $S \sim \mathsf{Unif}(\{1, 2, 3\})$. Find the variance of

$$\sum_{i=1}^{S} X_i.$$

Chapter 15

Correlation

Question of the Day Suppose (X, Y) is uniformly drawn from

$$\Omega = \{(0,0), (0,1), (0,2), (1,0), (1,1)\}.$$

What is the correlation between X and Y?

Summary

The Cauchy-Schwarz inequality says that for any inner product and vectors x and y,

$$\langle x, y \rangle^2 \leq \langle x, x \rangle \cdot \langle y, y \rangle.$$

For random variables X and Y such that XY, X^2, and Y^2 are integrable, the **correlation** between X and Y is

$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\,\mathrm{SD}(Y)}.$$

In the qotd, X and Y are not independent: if I condition on $X = 0$ then $Y \sim \mathsf{Unif}(\{0, 1, 2\})$, while if $X = 1$, then $Y \sim \mathsf{Unif}(\{0, 1\})$. Or as another proof:

$$\mathbb{P}(X = 1, Y = 1) = 1/5,$$

but

$$\mathbb{P}(X = 1)\mathbb{P}(Y = 1) = (2/5)(2/5) = 4/25 \neq 1/5.$$

On the other hand, they are not completely dependent on each other. Knowing the value of X does not completely determine Y, and knowing Y does not completely determine X. So we want a way of describing in some sense just how dependent the random variables are on each other.

15.1 The Cauchy-Schwarz inequality

A very important fact about inner products and the inner product norm is the Cauchy-Schwartz inequality.

Theorem 5 (The Cauchy-Schwarz inequality)
For any inner product and vectors x and y:

$$\langle x, y \rangle^2 \leq \langle x, x \rangle \cdot \langle y, y \rangle.$$

or expressed using the inner product norm:

$$|\langle x, y \rangle| \leq \|x\| \cdot \|y\| .$$

Moreover, you only get inequality when there exists a scalar α such that $x = \alpha y$ or $y = \alpha x$.

The Cauchy-Schwarz inequality means that the value of the fraction

$$\frac{\langle v, w \rangle}{\|v\| \cdot \|w\|},$$

lies between -1 and 1. You may recall the following interesting geometric fact. If θ is the angle between $v, w \in \mathbb{R}^n$, then

$$\cos(\theta) = \frac{\langle v, w \rangle}{\|v\| \cdot \|w\|},$$

So if this ratio on the right hand side is 1, then $\theta = 0$ and the vectors are pointing in the same direction. If the ratio is -1, then $\theta = \tau/2$, and the vectors are pointing in opposite directions. If the ratio is 0, then $\theta = \tau/4$ or $\theta = -\tau/4$, and the vectors are perpendicular (orthogonal) to one another.

15.2   Angles and correlation

It does not really make sense to talk about the angle between two random variables. However, we still know from Cauchy-Schwartz that the ratio

$$\frac{\text{Cov}(X, Y)}{\text{SD}(X) \, \text{SD}(Y)} \in [-1, 1].$$

So we give this quantity a name, we call it the *correlation* between the two random variables.

Definition 49
For X and Y such that XY, X^2 and Y^2 are integrable, the **correlation** between X and Y is

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{SD}(X) \, \text{SD}(Y)}.$$

Fact 48
The correlation lies between -1 and 1.

Proof. From Fact 44, we know that covariance is an inner product. So it follows directly from the Cauchy-Schwarz inequality. □

Now let us answer the Question of the Day. To make the calculation, it is necessary to find the values of $\mathbb{E}[X]$, $\mathbb{E}[Y]$, $\mathbb{E}[XY]$, $\mathbb{E}[X^2]$, and $\mathbb{E}[Y^2]$. Because these are discrete random variables, the

expectations can be found using a sum.

$$\mathbb{E}[X] = \frac{1}{5}\left[0 + 0 + 0 + 1 + 1\right] = \frac{2}{5}$$

$$\mathbb{E}[Y] = \frac{1}{5}\left[0 + 1 + 2 + 0 + 1\right] = \frac{4}{5}$$

$$\mathbb{E}[XY] = \frac{1}{5}\left[0 + 0 + 0 + 0 + 1\right] = \frac{1}{5}$$

$$\mathbb{E}[X^2] = \frac{1}{5}\left[0 + 0 + 0 + 1 + 1\right] = \frac{2}{5}$$

$$\mathbb{E}[Y^2] = \frac{1}{5}\left[0 + 1 + 4 + 0 + 1\right] = \frac{6}{5}$$

So

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = (1/5) - (2/5)(4/5) = -3/25$$
$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = (2/5) - (2/5)^2 = 6/25$$
$$\mathbb{V}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = (6/5) - (4/5)^2 = 14/25$$

Putting this together,

$$\mathrm{Cor}(X, Y) = \frac{-3/25}{\sqrt{(6/25)(14/25)}} = \frac{-3}{\sqrt{84}} \approx \boxed{-0.3273}$$

So X and Y are negatively correlated. That means that on average when one is large, the other will be smaller than its average value.

15.3 Independence and correlation

When two random variables have correlation 0, we say that they are *uncorrelated.*

> **Definition 50**
> Two random variables X and Y are **uncorrelated** if their correlation exists and is 0.

If the correlation between two random variables X and Y is 0, that means that knowledge of X does not change the *average* value of Y. However, that does not means that X and Y are independent!

To see why, consider

$$(X, Y) \sim \mathsf{Unif}(\{(1, 1), (1, -1), (-1, 2), (-1, -2)\}).$$

Note that if $X = 1$ then $\mathbb{E}[Y|X = 1] = 0$, and if $X = -1$ then $\mathbb{E}[Y|X = -1] = 0$. So knowledge of X does not change the average value of Y.

More directly:

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = (1/4)[1 - 1 - 2 + 2] - (0)(0) = 0.$$

However, knowing that $X = 1$ means that $Y \in \{1, -1\}$, while if $X = -1$ then $Y \in \{2, -2\}$, so X and Y are not independent. Put another way

$$\mathbb{P}(X = 1, Y = 1) = 1/4, \text{ but } \mathbb{P}(X = 1)\mathbb{P}(Y = 1) = (1/2)(1/4) = 1/8.$$

So while is is not always true that uncorrelated random variables are independent, it is true is that independent random variables are uncorrelated.

> **Fact 49**
> If X and Y are independent and their correlation exists, then they are uncorrelated.

Proof. Suppose that X and Y with finite covariance have joint density $f_{X,Y}$. Then

$$
\begin{aligned}
\mathbb{E}[XY] &= \int f_{X,Y}(x,y)\, d\mu(x,y) \\
&= \int_x \int_y f_X(x) f_Y(y)\, d\mu(y)\, d\mu(x) \qquad \text{by Tonelli} \\
&= \int_x f_X(x) \int_y f_Y(y)\, d\mu(y)\, d\mu(x) \\
&= \int_x f_X(x) \mathbb{E}[Y]\, d\mu(x) \\
&= \mathbb{E}[X]\mathbb{E}[Y],
\end{aligned}
$$

Hence $\mathrm{Cov}(X,Y) = 0$ are they are uncorrelated. $\qquad\square$

Problems

15.1 For $(X,Y) \sim \mathsf{Unif}(\{(0,0),(0,2),(1,2)\})$, find the correlation between X and Y.

15.2 For (W,Z) uniform over $\{(0,1),(1,0),(2,2)\}$, find the correlation between W and Z

15.3 For (A,B) with density

$$
f_{(A,B)}(a,b) = (2/3)(a+2b)\mathbb{1}(a \in [0,1])\mathbb{1}(b \in [0,1]),
$$

find $\mathrm{Cor}(A,B)$.

15.4 For (X,Y) with density $f_{X,Y}(x,y) = 2\exp(-x-2y)\mathbb{1}(x,y \geq 0)$, Find $\mathrm{Cor}(X,Y)$.

15.5 Consider random variables X and Y with joint density

$$
f_{(X,Y)}(x,y) = C\exp(-x-xy-y)\mathbb{1}(x \geq 0, y \geq 0).
$$

a) Numerically find the covariance of X and Y?

b) Numerically find is the correlation of X and Y?

15.6 Let

$$
f_{(X,Y)}(r,s) = \frac{1}{C} \cdot \frac{r+s}{r}\mathbb{1}(r \in [1,2], s \in [0,1]).
$$

a) Find the value of C.

b) Find the density of Y.

c) Find $\mathrm{Cor}(X,Y)$.

Adding random variables together

Question of the Day Suppose $X \sim \mathsf{Unif}([0,1])$ and $Y \sim \mathsf{Exp}(4)$ are independent. What is $\mathbb{V}[X+Y]$?

Summary For any random variable X and Y with finite variance and covariances,

$$\mathbb{V}[X+Y] = \mathbb{V}[X] + \mathbb{V}[Y] + 2\operatorname{Cov}(X,Y).$$

For a finite set of random variables,

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

In general, if X and Y have densities f_X and f_Y with respect to μ, then $f_{X+Y} = f_X * f_Y$, where the $*$ is the convolution operator defined as

$$[f * g](s) = \int_x f(x)g(s-x)\, d\mu.$$

16.1   Variance of sums

Recall that

$$(a+b)^2 = a^2 + b^2 + 2ab.$$

This follows from distributive laws.

Because variance is just the inner product of a vector with itself (like the square of a number), the same formula applies to two random variables.

Fact 50

Let X and Y have finite variance and covariance. Then

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\operatorname{Cov}(X,Y).$$

Proof. Recall that $\mathrm{Cov}(X, Y)$ is an inner product between X and Y. Inner products are distributive and commutative, and variance is the covariance of a random variable with itself, so

$$\begin{aligned}
\mathbb{V}(X + Y) &= \mathrm{Cov}(X + Y, X + Y) \\
&= \mathrm{Cov}(X, X) + \mathrm{Cov}(X, Y) + \mathrm{Cov}(Y, X) + \mathrm{Cov}(Y, Y) \\
&= \mathbb{V}(X) + \mathbb{V}(Y) + 2\,\mathrm{Cov}(X, Y).
\end{aligned}$$

□

This same argument can be extended to the sum of n random variables.

Fact 51

Let $X_1, \ldots, X_n$ have finite variances and all pairs have finite covariances. Then

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

In particular, if $X_1, \ldots, X_n$ are independent, then the variance of the sum is the sum of the variances.

This is also known as the Pythagorean Theorem, because if two displacement vectors a and b are at right angles, then $a \cdot b = 0$, and

$$\|a + b\|^2 = (a + b) \cdot (a + b) = a \cdot a + b \cdot b + 2 \cdot 0 = \|a\|^2 + \|b\|^2.$$

So the fact that for a right triangle, the length of the hypotenuse squared equals the sum of the squares of the lengths of the other sides, is exactly the same as the fact that the variance of the sum of uncorrelated random variables is the sum of the variances!

16.2   Standard deviation of sample averages

This property gives rise to the following result about sample averages of iid random variables.

Fact 52

Let $X_1, \ldots, X_n \sim X$ be iid random variables where X has finite variance, and

$$S = \frac{X_1 + \cdots + X_n}{n}.$$

Then

$$\mathrm{SD}(S) = \frac{\mathrm{SD}(X)}{\sqrt{n}}.$$

Proof. Since each $X_i \sim X$, $\mathrm{SD}(X_i) = \mathrm{SD}(X)$ for all i. That means

$$\mathbb{V}(S) = \frac{1}{n^2}[\mathbb{V}(X_1) + \cdots + \mathbb{V}(X_n)] = \frac{n\mathbb{V}(X)}{n^2} = \frac{\mathbb{V}(X)}{n},$$

and taking the square root of both sides finishes the proof.   □

16.3 Convolutions

Suppose $X \sim \text{Unif}(\{1, 2, 3\})$ and $Y \sim \text{Unif}(\{1, 2, 3, 4, 5, 6\})$. What is the distribution of $X + Y$? Well, we know that $X + Y \in \{2, 3, \ldots, 9\}$, but how do we calculate the probabilities? We sum over the states that lead to that chance.

For instance,

$$
\begin{aligned}
\mathbb{P}(X + Y = 7) &= \sum_{i \in \{1,2,3\}} \mathbb{P}(X = i)\mathbb{P}(X + Y = 7 | X = i) \\
&= \sum_{i \in \{1,2,3\}} \mathbb{P}(X = i)\mathbb{P}(Y = 7 - i) \\
&= (1/3)(1/6) + (1/3)(1/6) + (1/3)(1/6) \\
&= 1/6 = 0.1666 \ldots.
\end{aligned}
$$

More generally, the following holds.

Fact 53

Let X and Y be independent and have density f_X and f_Y with respect to μ. Then

$$
f_{X+Y}(s) = \int_x f_X(x)f_Y(s - x)\, d\mu(x) = \int_y f_Y(y)f_X(s - y)\, d\mu(y).
$$

This integration operation is called the *convolution* of the densities.

Definition 51

The **convolution** of real valued functions f and g with respect to measure μ is

$$
[f * g](s) = \int_x f(x)g(s - x)\, d\mu.
$$

Example 38

Suppose $X \sim \mathsf{Exp}(2)$ and $Y \sim \mathsf{Exp}(1)$ are independent. What is the density of $X + Y$?

Answer   The densities of X and Y are

$$f_X(x) = 2\exp(-2x)\mathbb{1}(x \geq 0)$$
$$f_Y(y) = \exp(-y)\mathbb{1}(y \geq 0)$$

So convolve to get the density of the sum:

$$f_{X+Y}(s) = \int_x f_X(x)f_Y(s - x)\,dx$$
$$= \int_x 2\exp(-2x)\mathbb{1}(x \geq 0)\exp(-(s - x))\mathbb{1}(s - x \geq 0)\,dx$$

If $s < 0$, then $\mathbb{1}(x \geq 0)\mathbb{1}(x \leq s) = 0$, so assume $s \geq 0$. Then

$$f_{X+Y}(s) = \int_{x=0}^{s} 2\exp(-s - x)\,dx$$
$$= \frac{2}{-1}\exp(-s - x)\big|_{x=0}^{s}$$
$$= -2[\exp(-2s) - \exp(-s)].$$

Putting both cases for s together gives

$$\boxed{f_{X+Y}(s) = 2[\exp(-s) - \exp(-2s)]\mathbb{1}(s \geq 0).}$$

More generally, if X and Y are not independent then to find the density of $X + Y$, we need the joint density of X and Y.

Fact 54

Let X and Y have joint density $f_{X,Y}$ with respect to a product measure μ. Then

$$f_{X+Y}(s) = \int_x f_{X,Y}(x, s - x)\,d\mu(x) = \int_y f_{X,Y}(s - y, y)\,d\mu(y).$$

Problems

16.1 Let $A \sim \mathsf{Unif}([0, 1])$ and $B \sim \mathsf{Unif}([0, 2])$ be independent. Find the density of $A + B$.

16.2 Suppose $X \sim \mathsf{Unif}([0, 1])$ and $Y \sim \mathsf{Exp}(2)$. Find the density of $X + Y$.

16.3 Suppose $X \sim \mathsf{Unif}(\{1, 2, 3\})$ and $Y \sim \mathsf{Unif}(\{3, 5\})$ are independent. What is the density of $X + Y$?

16.4 Suppose X_1 and X_2 are iid $\mathsf{Unif}(\{1, 2, 3, 4\})$. Find the density of $X + Y$.

16.5 Suppose R and G are discrete random variables where $R \sim \text{Bern}(0.3)$ and $G \sim \text{Geo}(0.6)$. So

$$f_R(i) = (0.3)\mathbb{1}(i = 1) + (0.7)\mathbb{1}(i = 0), \ \ f_G(i) = (0.6)(0.4)^{i-1}\mathbb{1}(i \in \{1, 2, \ldots\}).$$

Find the density of $R + G$.

16.6 Suppose X and Y are independent, discrete random variables with $f_X(1) = 0.2$, $f_X(2) = 0.3$ and $f_x(3) = 0.5$, while $f_Y(1) = 0.5$, $f_y(2) = 0.1$, $f_Y(3) = 0.4$. Find the density of $X + Y$.

Chapter 17

The moment generating function

> **Question of the Day** Suppose $A_1, \ldots, A_{10}$ are iid uniform over $\{1, 2, 3\}$. What is $\mathbb{P}(A_1 + \cdots + A_{10} = 20)$?

Summary

The **generating function** of a random variable X is

$$\mathrm{gf}_X(s) = \mathbb{E}[s^X].$$

The **moment generating function** of a random variable X is a function such that

$$\mathrm{mgf}_X(t) = \mathbb{E}[\exp(tX)].$$

The moment generating function of X suffices to define the distribution of X. For $X_1, \ldots, X_n$ independent,

$$\mathrm{mgf}_{X_1 + \cdots + X_n}(t) = \mathrm{mgf}_{X_1}(t)\,\mathrm{mgf}_{X_2}(t)\cdots\mathrm{mgf}_{X_n}(t).$$

In this chapter, we will learn about a new way to encode probability distributions. This method does not use probabilities, but instead uses expected value in a particular way. The advantage of this encoding is that if you know the encoding for a finite set of independent random variables, it is easy to find the encoding for their sum.

17.1   Generating functions

Let's start with a simple example. Suppose X and Y are independent random variables with densities

$$f_X(i) = 0.3\mathbb{1}(i = 1) + 0.3\mathbb{1}(i = 2) + 0.4\mathbb{1}(i = 3)$$
$$f_Y(i) = 0.5\mathbb{1}(i = 1) + 0.5\mathbb{1}(i = -1).$$

This means $\mathbb{P}(X \in \{1, 2, 3\}) = \mathbb{P}(Y \in \{-1, 1\}) = 1$. Consider the problem of finding $\mathbb{P}(X + Y = 2)$. There are two ways that X and Y can sum to 2. Either $X = 1$ and $Y = 1$, or

$X = 3$ and $Y = -1$. So

$$\begin{aligned}
\mathbb{P}(X + Y = 2) &= \mathbb{P}(X = 1, Y = 1) + \mathbb{P}(X = 3, Y = -1) \\
&= \mathbb{P}(X = 1)\mathbb{P}(Y = 1) + \mathbb{P}(X = 3)\mathbb{P}(Y = -1) \\
&= (0.3)(0.5) + (0.4)(0.5) = 0.35.
\end{aligned}$$

Now we consider an encoding for X and Y that tells us everything about the values they take on and the probabilities that they take on those values. Make a polynomial expression in s. The coefficient c of a term cs^i in the polynomial will be the probability that $X = i$. Then add all the terms together.

Since $X = 1$ with probability 0.3, I start with a term that reads $0.3s$. The $\mathbb{P}(X = 2) = 0.3$ gives a $0.3s^2$ term, and the $\mathbb{P}(X = 3) = 0.4$ gives a $0.4s^3$ term. Adding gives a polynomial function:

$$f(s) = 0.3s + 0.3s^2 + 0.4s^3.$$

Note that this function completely describes the distribution of X.

We can do something similar for Y.

$$g(s) = 0.5s^1 + 0.5s^{-1}.$$

We restrict ourselves to $s > 0$ just to make sure everything is defined.

Now look at what happens when we multiply these two polynomials together:

$$\begin{aligned}
f(s)g(s) &= (0.3s + 0.3s^2 + 0.4s^3)(0.5s + 0.5s^{-1}) \\
&= (0.3)(0.5)s^2 + (0.3)(0.5)s^3 + (0.4)(0.5)s^4 + (0.3)(0.5)s^0 + (0.3)(0.5)s^1 + (0.3)(0.5)s^2 \\
&= 0.15s^0 + 0.15s + 0.35s^1 + 0.15s^2 + 0.15s^3 + 0.2s^4.
\end{aligned}$$

Now look at the coefficient of s^2. That coefficient came from $(0.3)(0.5) + (0.4)(0.5)$, since the s^1s^1 term combined to give s^2 and the s^3s^{-1} term combined to give s^2.

In other words, the multiplication of the encodings resulted in an s^2 term that exactly replicated how we found $\mathbb{P}(X + Y = 2)$. This is not a coincidence!

There is another way to describe how this polynomial was formed. If you sum the probability that $X = i$ times s^i for each i, this is just

$$\mathbb{E}[s^X]$$

by our rule for the expectation of a function of X. The X is the exponent here because the value i is in the exponent of s.

> **Definition 52**
> The **generating function** of a random variable X is $\text{gf}_X(s) = \mathbb{E}[s^X]$.

First a simple fact.

> **Fact 55**
> For any random variable $\text{gf}_X(1) = 1$.

Note in our example from earlier,

$$f(1) = 0.3 \cdot 1^1 + 0.3 \cdot 1^2 + 0.4 \cdot 1^3 = 0.3 + 0.3 + 0.4,$$

so $\text{gf}_X(1) = 1$ is just another way of stating that the sum of probabilities of a random variable must add to 1.

Second, we have why generating functions are useful.

> **Fact 56**
> Let X and Y be independent random variables. Then $\text{gf}_{X+Y}(s) = \text{gf}_X(s)\,\text{gf}_Y(s)$.

Proof. Suppose X and Y are independent. Then for any $s > 0$, s^X and s^Y are independent as well. Hence

$$\mathbb{E}[s^{X+Y}] = \mathbb{E}[s^X s^Y]$$
$$= \mathbb{E}[s^X]\mathbb{E}[s^Y],$$

and we are done! $\qquad\square$

> **Example 39**
> For the question of the day, since we are adding 10 iid copies of the random variable together, we must multiply the generating function by itself ten times. That is
>
> $$\text{gf}_{A_1+\cdots+A_{10}}(s) = \text{gf}_{A_1}(s)\cdots\text{gf}_{A_{10}}(s) = \text{gf}_A(s)^{10}.$$
>
> Using Wolfram Alpha to expand $[(1/3)s + (1/3)s^2 + (1/3)s^3)]^{10}$, we get something like
>
> $$\frac{1}{3^{10}}s^{30} + \frac{10}{3^{10}}s^{29} + \cdots + \frac{8953}{3^{10}}s^{20} + \cdots + \frac{1}{3^{10}}s^{10},$$
>
> therefore the answer is $8953/3^{10} = \boxed{0.1516\ldots}$.

17.2 Moment generating function

Since $s > 0$, we can say that $s = e^t$ for some t. If we write the generating function as a function of t instead of s, we have the *moment generating function*.

> **Definition 53**
> The **moment generating function** of a random variable X is $\text{mgf}_X(t) = \mathbb{E}[\exp(tX)]$ for all values of t where this expected value exists.

> **Example 40**
> Suppose X has density
>
> $$f_X(i) = 0.2\mathbb{1}(i=3) + 0.7\mathbb{1}(i=6) + 0.1\mathbb{1}(i=8).$$
>
> What is the moment generating function of X?
>
> **Answer** Here $X \in \{3, 6, 8\}$, so the moment generating function is
>
> $$\text{mgf}_X(t) = \mathbb{E}[\exp(tX)] = \sum_{i\in\{3,6,8\}} \exp(ti)f_X(i)$$
> $$= \boxed{0.2\exp(3t) + 0.7\exp(6t) + 0.1\exp(8t).}$$

Note that outcomes now appear as coefficients of t inside the exponential functions, while the probabilities remain as the coefficients outside of the exponential functions.

Example 41
Suppose X has moment generating function

$$0.6 \exp(-2t) + 0.4 \exp(4t).$$

What is the density of X?

Answer   To get this moment generating function, it must be that

$$f_X(i) = 0.6\mathbb{1}(i = -2) + 0.4\mathbb{1}(i = 4).$$

Moment generating functions inherit the nice multiplicative property from generating functions.

Fact 57
Suppose X and Y are independent with moment generating functions at t. Then

$$\operatorname{mgf}_{X+Y}(t) = \operatorname{mgf}_X(t)\operatorname{mgf}_Y(t).$$

17.3   Moment generating functions for continuous random variables

The procedure for finding the moment generating function for a continuous random variable is similar, although we will be using an integral rather than a sum.

Example 42
Find the moment generating function of $U \sim \operatorname{Unif}([0, 1])$.

Answer   Since U is continuous, if $t \neq 0$ this is

$$\mathbb{E}[\exp(tU)] = \int_{\mathbb{R}} \exp(ts)\mathbb{1}(s \in [0, 1]) \, ds = \int_0^1 \exp(ts) \, ds$$

$$= \frac{1}{t} \exp(ts)|_0^1 = \frac{\exp(t) - 1}{t}.$$

if $t = 0$ the result is 1, making the overall answer

$$\operatorname{mgf}_U(t) = \begin{cases} t^{-1}(e^t - 1) & t \neq 0 \\ 1 & t = 0 \end{cases}.$$

Remark   We had to write our answer differently for $t \neq 0$ and $t = 0$. If we wanted to, we could write this for both cases at the same time by using a Taylor series expansion. Recall

$$\exp(t) - 1 = t + t^2/2! + \cdots,$$

so

$$\frac{\exp(t) - 1}{t} = 1 + t/2! + t^2/3! + \cdots$$

which is defined for all t, including $t = 0$ (where it now clearly evaluates to 1).

17.4   How to generate moments

So why is it called the moment generating function? Recall that the ith moment of a random variable X is $\mathbb{E}[X^i]$. The name moment comes from its use in physics.

Consider the case where

$$\mathrm{mgf}_X(t) = 0.2\exp(3t) + 0.7\exp(6t) + 0.1\exp(8t),$$

and we take the derivative with respect to t. For any constant k,

$$[\exp(kt)]' = k\exp(kt),$$

so for the mgf,

$$[\mathrm{mgf}_X(t)]' = (0.2)(3)\exp(3t) + (0.7)(6)\exp(6t) + (0.1)(8)\exp(8t).$$

Now let's plug in $t = 0$ so that $\exp(k(0)) = 1$ for all k.

Then

$$[\mathrm{mgf}_X(t)]'|_{t=0} = (0.2)(3) + (0.7)(6) + (0.1)(8).$$

This is just how we find the expected value of X! We multiply the probability that it takes on values times those values. So

$$[\mathrm{mgf}_X(t)]'|_{t=0} = \mathbb{E}[X].$$

Suppose we took the derivative of $[\mathrm{mgf}_X(t)]'$. This gives

$$[\mathrm{mgf}_X(t)]'' = (0.2)(3)^2\exp(3t) + (0.7)(6)^2\exp(6t) + (0.1)(8)^2\exp(8t).$$

Again evaluating at $t = 0$ turns all the exponential factors to 1, so

$$[\mathrm{mgf}_X(t)]''|_{t=0} = (0.2)(3)^2 + (0.7)(6)^2 + (0.1)(8)^2 = \mathbb{E}[X^2].$$

Why is this happening? Well, suppose that we could bring a derivative inside of an expectation operator. That would mean that

$$\frac{d\mathbb{E}[\exp(tX)]}{dt} = \mathbb{E}\left[\frac{d\exp(tX)}{dt}\right] = \mathbb{E}[X\exp(tX)].$$

If we then plug in $t = 0$, we get $\mathbb{E}[X]$. If we differentiate twice (and again assuming that we can bring the derivative inside the expectation operator) we get

$$\frac{d^2\mathbb{E}[\exp(tX)]}{dt^2} = \mathbb{E}\left[\frac{d^2\exp(tX)}{dt^2}\right] = \mathbb{E}[X^2\exp(tX)].$$

This time if we plug in $t = 0$ we get $\mathbb{E}[X^2]$, the second moment. This pattern continues.

> **Fact 58**
> Suppose X has a moment generating function defined for a nontrivial interval containing $t = 0$. Then the ith derivative of the moment generating function of X evaluated at $t = 0$ is $\mathbb{E}[X^i]$.

Our earlier illustration used a discrete random variable, but this also works perfectly well for continuous distributions.

Example 43
Use the moment generating function of $U \sim \text{Unif}([0, 1])$ to find the mean and variance of U.

Answer   Recall that the Taylor series for the moment generating function is

$$\text{mgf}_U(t) = 1 + \frac{t}{2!} + \frac{t^2}{3!} + \frac{t^3}{4!} + \cdots .$$

So if we differentiate once we get

$$[\text{mgf}_U(t)]' = \frac{1}{2} + \frac{2t}{6} + \frac{3t^2}{24} + \cdots .$$

Hence $[\text{mgf}_U(t)]'|_{t=0} = 1/2 = \mathbb{E}[U] = \boxed{0.5000}$.
  Differentiating again gives

$$[\text{mgf}_U(t)]'' = \frac{2}{6} + \frac{6t}{24} + \cdots$$

so $[\text{mgf}_U(t)]''|_{t=0} = 1/3$.
  Hence the variance of U is $(1/3) - (1/2)^2 = 1/12 = \mathbb{V}(U) = \boxed{0.08333}$.

Note that if we use $\text{mgf}_U(t) = (\exp(t) - 1)/t$, then this is not defined at $t = 0$, but we can still differentiate nearby and take the limit as t approaches o. The derivative is

$$
\begin{aligned}
[\text{mgf}_U(t)]' &= [\exp(t) - 1]' t^{-1} + (\exp(t) - 1)[t^{-1}]' \\
&= t^{-1} \exp(t) - t^{-2}(\exp(t) - 1) \\
&= \frac{t \exp(t) - \exp(t) + 1}{t^2},
\end{aligned}
$$

which has limit $1/2$ as $t \to 0$. (This can be checked by using L'Hopital's rule twice.)

Problems

17.1 Suppose $\mathbb{P}(X = 1) = \mathbb{P}(X = 2) = 0.5$. What is $\text{mgf}_X(t)$?

17.2 Find the moment generating function of X which has density

$$f_X(i) = 0.25\mathbb{1}(i = 1) + 0.6\mathbb{1}(i = 2) + 0.15\mathbb{1}(i = 3).$$

17.3 Suppose X has moment generating function

$$\text{mgf}_X(t) = 0.1 \exp(10t) + 0.9 \exp(-5t).$$

What is the density of X?

17.4 Prove using moment generating functions that if $X \sim \text{Unif}(\{0, 1, 2, \ldots, n - 1\})$ and $Y \sim \text{Unif}([0, 1])$, then $X + Y \sim \text{Unif}([0, n])$.

17.5 Let $Z_1, Z_2, \ldots, Z_n$ be iid $\mathsf{N}(0, 1)$. Recall for $Z \sim \mathsf{N}(0, 1)$, $\text{mgf}_Z(t) = \exp(t^2/2)$.

a) What is the the moment generating function of $Z_1 + Z_2$?

b) What is the moment generating function of:

$$\frac{Z_1 + \cdots + Z_n}{n}.$$

c) What is the moment generating function of:

$$\frac{Z_1 + \cdots + Z_n}{\sqrt{n}}.$$

17.6 Let $U_1, U_2, \ldots, U_n$ be iid $\mathsf{Unif}([0, 1])$. Find the moment generating function of

$$\frac{U_1 + \cdots + U_n}{\sqrt{n}}.$$

17.7 Suppose that X has the following density:

$$f_X(r) = \frac{3}{8}(r^3 - 8r^2 + 19r - 12)\mathbf{1}(r \in [1, 3]).$$

a) Find the mode(s) of X.

b) Find the median(s) of X.

c) Find the mean of X.

d) Find $\mathbb{E}[e^{tX}]$.

17.8 Suppose X is uniform over $[5, 10]$. What is the moment generating function of X?

Normal random variables

Question of the Day Suppose Z_1 and Z_2 are iid standard normal random variables. What is $\mathbb{P}(Z_2 > (1/2)Z_1)$?

Summary The **standard normal** distribution has density with respect to Lebesgue measure
$$f_Z(x) = \frac{1}{\sqrt{\tau}} \exp(-x^2/2).$$
Write $Z \sim \mathsf{N}(0,1)$, as this random variable has mean 0 and standard deviation 1. For constants $\mu \in \mathbb{R}$ and $\sigma > 0$, say $\mu + \sigma Z$ is a normal random variable with mean μ and variance σ^2, write $\mu + \sigma Z \sim \mathsf{N}(\mu, \sigma^2)$.

18.1 The normal distribution

The normal distribution was introduced by Gauss as a way of fitting errors in calculations. (Historically he used it in helping astronomers find the dwarf planet Ceres in the asteroid belt.) In its modern form, the distribution is defined as follows.

Definition 54
Let Z have density
$$f(z) = \frac{1}{\sqrt{\tau}} \exp(-z^2/2).$$
Then say Z has a **standard normal distribution**, write $Z \sim \mathsf{N}(0,1)$.
 Let $W = \mu + \sigma Z$. Then say W has a **normal distribution** or **Gaussian distribution** with parameters μ and σ. Write $W \sim \mathsf{N}(\mu, \sigma^2)$.

Fact 59
For $W \sim \mathsf{N}(\mu, \sigma^2)$, $\mathbb{E}[W] = \mu$ and $\mathrm{SD}(W) = \sigma$.

 Some authors (particularly in the social sciences) have also dubbed this distribution the "bell-shaped curve". This is a truly terrible name for this distribution, as it applies equally (if not more) to symmetric beta distributions and the Cauchy distribution. It pretty much betrays the user as someone who only knows about normal distributions and not the other rich variety of densities that exist in probability theory.

Upon first glance at the density, one might wonder what the full circle constant $\tau = 6.2831\ldots$ is doing there? Recall that $\tau = 2\pi$, where π is the half circle constant. And why is τ inside a square root sign?

To answer this mystery, we need to think about drawing two independent normals. Let $Z_1, Z_2 \sim \mathsf{N}(0,1)$ be independent. Now consider transforming these Cartesian coordinates to polar coordinates.

Recall that polar coordinates for a point $(x, y) \in \mathbb{R}^2$ use the distance from the origin r and the angle counterclockwise from the horizontal axis in the right hand direction θ. This is one of those things more easily seen with a picture:
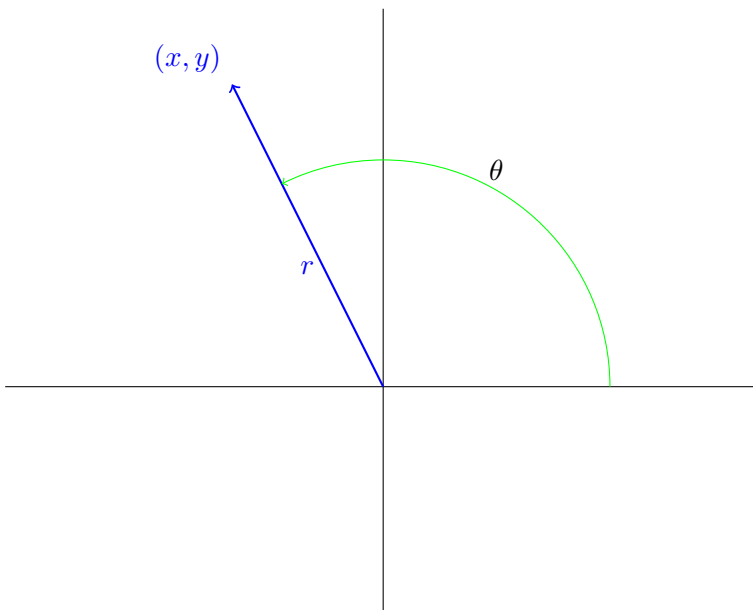
So how do we transform from rectangular to polar coordinates? We use the distance formula and trigonometric rules to get

$$r = x^2 + y^2, \ \theta = \arctan(y/x).$$

For our random variables, we get

$$R = Z_1^2 + Z_2^2, \ \theta = \arctan(Z_2/Z_1).$$

The next question to ask is what is the distribution of R and θ. Are they independent like Z_1 and Z_2? To answer this, we need one more key fact about the polar coordinate transform that you learned when doing multivariate integrals.

$$dx \, dy = r \, dr \, d\theta.$$

Because Z_1 and Z_2 are independent,

$$
\begin{aligned}
\mathbb{P}(Z_1 \in dx, Z_2 \in dy) = f_{Z_1,Z_2}(x,y)\, dx\, dy &= \frac{1}{\tau} \exp(-x^2/2) \exp(-y^2/2)\, dx\, dy \\
&= \frac{1}{\tau} \exp(-[x^2+y^2]/2)\, dx\, dy \\
&= \frac{1}{\tau} \exp(-r^2/2)\, dx\, dy \\
&= \frac{1}{\tau} r \exp(-r^2/2)\, dr\, d\theta \\
&= \mathbb{P}(R \in dr, \theta \in d\theta).
\end{aligned}
$$

Notice that this factors into a piece that only depends on R, and a piece that only depends on θ.

$$
\mathbb{P}(R \in dr, \theta \in d\theta) = \left[\frac{1}{\tau}\, d\theta \right] \left[r \exp(-r^2/2)\, dr \right].
$$

Because $\theta \in [0, \tau]$, this first factor is the density of a uniform over $[0, \tau]$. The second factor is a bit weirder, with density $f_R(r) = r \exp(-r^2/2)$. The distribution which has this density is called the *Rayleigh distribution* or a *chi distribution with two degrees of freedom*. A third way to think about R is that $R \sim \sqrt{A}$ where $A \sim \mathsf{Exp}(1/2)$.

When the distribution of θ is uniform and independent of R, we call the distribution *rotationally symmetric*.

> **Definition 55**
> A pair of random variables (X, Y) is **rotationally symmetric** if when converted to polar coordinates (R, θ), θ is $\mathsf{Unif}([0, \tau])$ and is independent of R.

Note that if we add a fixed angle to θ, then it will still be uniform over $[0, \tau]$ (recognizing that angle s and $s + \tau$ are the same angle.) In practice, this means that we can rotate the region as much as we want in solving the problem.

Example 44
Question of the Day: For $Z_1, Z_2 \sim N(0,1)$ iid, what is the probability that $Z_2 > (1/2)Z_1$?

Answer   The region looks like
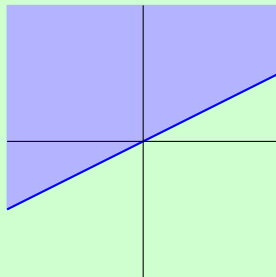
But by rotating it slightly, the region looks like
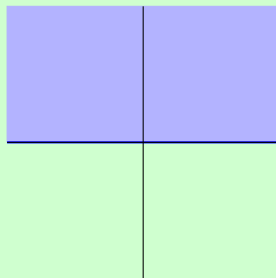
So
$$\mathbb{P}(Z_2 > (1/2)Z_1) = \mathbb{P}(Z_2 > 0) = 1/2 = \boxed{0.5000}.$$

18.2   Scaling and shifting normals

Suppose $N \sim N(3, 2^2)$, and then we consider $3N$. Well, we know that $N = 3 + 2Z$, where Z is a standard normal. Hence
$$3N = 3(3 + 2Z) = 9 + 6Z.$$

Therefore,
$$3N \sim N(9, 6^2).$$

Fact 60
Suppose $N \sim N(\mu, \sigma^2)$. Then $a + bN \sim N(a + b\mu, (b\sigma)^2)$.

Proof. Since $N = \mu + \sigma Z$ where Z is standard normal,
$$a + bN = a + b(\mu + \sigma Z) = a + b\mu + (b\sigma)Z,$$

and the result follows.   □

In other words, a shifted and scaled normal random variable is another shifted and scaled normal random variable!

18.3 Adding independent normal random variables

To understand what happens when we add normal random variables, it helps to know their moment generating function.

> **Fact 61**
> The moment generating function of a standard normal random variable is $\exp(t^2/2)$.

Proof. Recall that $\mathrm{mgf}_Z(t) = \mathbb{E}[\exp(tZ)]$, so

$$
\begin{aligned}
\mathrm{mgf}_Z(t) &= \int_z \exp(tz)\tau^{-1/2}\exp(-z^2/2)\,dz \\
&= \int_z \exp(tz - z^2/2)\tau^{-1/2}\,dz \\
&= \int_z \exp(-(z^2 - 2tz)/2)\tau^{-1/2}\,dz \\
&= \int_z \exp(-(z^2 - 2tz + t^2 - t^2)/2)\tau^{-1/2}\,dz \\
&= \int_z \exp(-(z-t)^2/2)\tau^{-1/2}\exp(-(-t^2)/2)\,dz \\
&= \exp(t^2/2)\int_z \exp(-(t-z)^2/2)\tau^{-1/2}\exp(t^2/2)\,dz \\
&= \exp(t^2/2).
\end{aligned}
$$

Note that the last integral is 1 because it is just the integral over all z of a density of a $\mathsf{N}(-t, 1)$ random variable, and densities always integrate to 1. □

> **Fact 62**
> The moment generating function of $N \sim \mathsf{N}(\mu, \sigma^2)$ is
> $$\mathrm{mgf}_N(t) = \exp[\mu t + (\sigma^2/2)t^2].$$

Proof. Recall that for $Z \sim \mathsf{N}(0,1)$, $\mu + \sigma Z \sim N \sim \mathsf{N}(\mu, \sigma^2)$. Hence

$$
\begin{aligned}
\mathrm{mgf}_N(t) &= \mathrm{mgf}_{\mu+\sigma Z}(t) \\
&= \mathbb{E}[\exp((\mu + \sigma Z)t)] \\
&= \mathbb{E}[\exp(\mu t)\exp((\sigma t)Z)] \\
&= \exp(\mu t)\mathbb{E}[\exp((\sigma t)Z)] \\
&= \exp(\mu t)\exp((\sigma t)^2/2) \\
&= \exp(\mu t + (\sigma^2/2)t^2)
\end{aligned}
$$

□

> **Fact 63**
> If Z_1 and Z_2 are iid normal random variables ($\mathsf{N}(0,1)$), then $Z_1 + Z_2 \sim \mathsf{N}(0,2)$.

Proof. The moment generating function of $Z_1 + Z_2$ is just the product of the moment generating functions. So

$$\operatorname{mgf}_{Z_1+Z_2}(t) = \exp(t^2/2)\exp(t^2/2) = \exp(2t^2/2) = \exp((\sqrt{2}t)^2/2).$$

This is $\mathbb{E}[\exp(\sqrt{2}Z_1)]$. In other words, it is the moment generating function of $Z_3 = \sqrt{2}Z_1$. By the previous fact, this has distribution $\mathsf{N}(0,2)$. □

We can extend this to adding an arbitrary number of normal random variables.

Fact 64
For $i \in \{1,\ldots,n\}$, suppose $X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$ are independent. Then

$$\sum_{i=1}^{n} X_i \sim \mathsf{N}\left(\sum_{i=1}^{n}\mu_i, \sum_{i=1}^{n}\sigma_i^2\right).$$

Problems

18.1 For Z a standard normal, find
$$\mathbb{P}(Z \in [-2,2]).$$

18.2 For Z_1, Z_2 iid $\mathsf{N}(0,1)$, find $\mathbb{P}(Z_1 \leq -Z_2)$.

18.3 The Digital Life conference draws a number of attendees each year that is normally distributed with mean $59\,000$ and standard deviation $10\,000$. Independently, E3 draws a number of attendees that is normally distributed with mean $75\,000$ and standard deviation $5\,000$.

   a) Suppose I average the two numbers. What is the distribution of the average?

   b) What is the chance that the average of the two conferences is greater than $70\,000$?

   c) What is the distribution of the number attending Digital Life minus the number attending E3?

   d) What is the chance that more people attend Digital Life than E3?

18.4 Suppose $A \sim \mathsf{N}(10,9), B \sim \mathsf{N}(-5,5), C \sim \mathsf{N}(7,2)$ are independent. What is the distributon of the average of the three random variables?

18.5 Suppose that $W_1,\ldots,W_n$ are iid standard normal random variables. What is the distribution of $W_1 + \cdots + W_n$?

18.6 Continuing the last problem, what is

$$\mathbb{P}(W_1 + \cdots + W_n \geq 2\sqrt{n})?$$

The Central Limit Theorem

Question of the Day Let $U_1, \dots, U_{10} \sim \mathsf{Unif}([0,1])$ be iid. Approximate $\mathbb{P}(U_1 + \dots + U_{10} \geq 6)$ using an normal approximation.

Summary Suppose that $X_1, X_2, \dots$ are an iid sequence of random variables with finite mean μ and variance σ^2. Then the **Central Limit Theorem** says that

$$(\forall a \in \mathbb{R}) \left(\lim_{n \to \infty} \mathbb{P}\left(\frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n}} \leq a \right) = \mathbb{P}(Z \leq a) \right),$$

where $Z \sim \mathsf{N}(0,1)$.

As noted earlier, Gauss used the normal distribution to model errors in experiments. Why does the normal distribution do so well? The answer lies in our additive property of moment generating functions.

19.1 Standardizing a sum

Recall that if $\mathbb{E}[X_i] = \mu$ and $\mathrm{SD}(X_i) = \sigma$, then

$$S = \frac{X_1 - \mu + X_2 - \mu + \dots + X_n - \mu}{\sigma\sqrt{n}} = \frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n}}$$

is a random variable with mean 0 and standard deviation 1.

Suppose that the $X_i \sim \mathsf{N}(\mu, \sigma^2)$. Then from the previous chapter, we know that

$$X_1 + \dots + X_n \sim \mathsf{N}(n\mu, n\sigma^2),$$

so by the scaling properties of expectation and variance

$$\frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n}} \sim \mathsf{N}(0,1).$$

Consider the following process for a random variable X. Suppose that $X_1, \dots, X_n$ are iid with the same distribution as X, and the X_i are totaled, their mean is subtracted, and then they are divided by their standard deviation. The resulting random variable has a (possibly new) distribution. So this process takes a distribution and returns a (possibly new) distribution. What the above

calculation says that if a standard normal distribution is fed into this process, the output is still a standard normal distribution. Mathematicians call this a *fixed point.*

Here is a simpler example of a fixed point for a process. Suppose that I consider that has input a number x, and returns

$$f(x) = x - \frac{x^2}{2} + 1.$$

Note that $f(\sqrt{2}) = \sqrt{2} - \frac{2}{2} + 1 = \sqrt{2}$. In other words, $\sqrt{2}$ is a fixed point for this process.

Now suppose I start with a value other than $\sqrt{2}$, like 1. Then $f(1) = 1 - 1/2 + 1 = 3/2$. Then $f(3/2) = 3/2 - 9/8 + 1 = 11/8$, and $f(11/8) = 183/128$, and so on.

Then $(183/128)^2 = 33489/16384 = 2.044\ldots$, so $183/128$ must be very close to $\sqrt{2}$. In other words, as we apply the function to x values over and over again, the result converges to the fixed point, which is the square root of two.

You see this type of behavior in many places in mathematics, including attractors in differential equation systems and the ergodic theorem in Markov chain theory. In many cases, applying a process to a mathematical object causes the object to move closer to the fixed point of the process.

19.2   The CLT

This fixed point behavior for normals manifests itself in the following way. Start with a random variable X with mean 0 and standard deviation 1. Then for $X_1, X_2, \ldots$ iid X,

$$\frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}}$$

begins to look more and more like a standard normal the larger n gets.

What if X does not have mean o and standard deviation 1? Well, then standardize the random variables by substracting off their mean and dividing by their standard deviation. This leads to an important result in probability called the *Central Limit Theorem* or CLT.

> **Theorem 6** (Central Limit Theorem)
> Let X be a random variable with finite mean μ and standard deviation σ. Then for any $a \in \mathbb{R}$,
> $$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le a\right) = \mathbb{P}(Z \le a),$$
> where Z is a standard normal random variable.

We can use this fact to approximate the probability that the sum of random variables is smaller or larger than some number.

Example 45
Suppose $U_1, \ldots, U_{10} \sim \mathsf{Unif}([0,1])$. Approximate the probability that $U_1 + \cdots + U_{10} \geq 7$ using the CLT.

Answer In order to use the CLT, we first need to find $\mu = \mathbb{E}[U]$ and $\sigma = \mathrm{SD}(U)$. A straightforward calculation gives us $\mu = 1/2$ and $\sigma = 1/\sqrt{12}$. Remember that anything you do to one side of an inequality you also have to do to the other side, so

$$\mathbb{P}(U_1 + \cdots + U_{10} \geq 7) = \mathbb{P}\left(\frac{U_1 + \cdots + U_{10} - 10(1/2)}{\sqrt{1/12} \cdot \sqrt{10}} \geq \frac{7 - 10(1/2)}{\sqrt{1/12}\sqrt{10}}\right)$$

$$\approx \mathbb{P}\left(Z \geq \frac{7 - 10(1/2)}{\sqrt{1/12}\sqrt{10}}\right)$$

Using the `pnorm` command in R gives this last probability as $\boxed{0.01422\ldots}$.

How good is this approximation? That can be difficult to figure out. There is a result called the Berry-Esseen Theorem, but despite years of improvements, it still is not very useful in practice.

Today the Central Limit Theorem is primarily used as a theoretical tool rather than a practical method of estimating probabilities. There are much better ways using Monte Carlo methods to get an estimate of the probabilities associated with sums. In the example above, a simple Monte Carlo estimate of the true answer is 0.01364 ± 0.00004, so the normal approximation is not too bad here. But for more complex problems, the CLT can give approximate values that are off by orders of magnitude, so it really should only be used as a last resort.

Sometimes the normal distribution does show up in actual data, but it is pretty rare. Oustide of very simple problems, most data sets look nothing like the normal distribution, and overreliance on its use can be a serious problem in some discliplines. An example of this comes from mathematical finance, where an over use of modeling using Gaussians is considered to have helped contribute to the 2008 financial crisis.

19.3   Using the CLT with discrete random variables

As a theorem, The CLT applies equally well to discrete and continuous random variables.

Example 46

Suppose $D_1, \ldots, D_{20}$ have density

$$f_D(i) = 0.7\mathbb{1}(i = -1) + 0.3\mathbb{1}(i = 1).$$

Estimate $\mathbb{P}(D_1 + \cdots + D_{20} \geq 0)$ using the CLT.

Answer Here the mean is

$$\mathbb{E}[D] = 0.7(-1) + 0.3(1) = -0.4$$

and second moment is

$$\mathbb{E}[D^2] = 0.7(-1)^2 + 0.3(1)^2 = 1,$$

so $\mathrm{SD}(D) = \sqrt{1 - (-0.4)^2} = \sqrt{0.84}$. Hence

$$\mathbb{P}(D_1 + \cdots + D_{20} \geq 0) = \mathbb{P}\left(\frac{D_1 + \cdots + D_{20} - 20(-0.4)}{\sqrt{20}\sqrt{0.84}} \geq \frac{(0.4)(20)}{\sqrt{20 \cdot 0.84}}\right)$$

The probability a standard normal is least $0.4\sqrt{20/0.84}$ can be found with `1 - pnorm(0.4 * sqrt(20 / 0.8))`, which gives $\boxed{0.02275\ldots}$.

If the above problem needed an exact solution, let N be the number of times that the $D_i = 1$. Then the number of times the D_i are -1 is $20 - N$. Hence

$$D_1 + \cdots + D_{20} = N(1) + (20 - N)(-1)$$

and this is at least 0 when $2N - 20 \geq 0$ or $N \geq 10$. Therfore this is the chance that a binomial random variable with parameters 20 and 0.3 is at least 10, and `1 - pbinom(9, 20, 0.3)` returns a value of about 0.04796.

Next consider the following example:

$$\mathbb{P}(X = 1) = 0.3, \mathbb{P}(X = 2) = 0.2, \mathbb{P}(X = 5) = 0.5.$$

Could the CLT be used to estimate the probability the sum of 10 iid copies of these random variables are at least 25?

First find the mean

$$0.3(1) + 0.2(2) + 0.5(3) = 2.2,$$

and standard deviation

$$\mathrm{sqrt}(0.3(1)^2 + 0.2(2)^2 + 0.5(3)^2 - 2.2^2) = \mathrm{sqrt}(0.76).$$

So the basic CLT estimate would be

$$\mathbb{P}\left(X_1 + \cdots + X_{10} \geq 25\right) = \mathbb{P}\left(\frac{X_1 + \cdots + X_{10} - 10(2.2)}{\mathrm{sqrt}(10 \cdot 5.6)} \geq \frac{25 - 10(2.2)}{\mathrm{sqrt}(10 \cdot 0.76)}\right)$$

$$\approx \mathbb{P}\left(Z \geq \frac{25 - 10(2.2)}{\mathrm{sqrt}(10 \cdot 5.6)} \geq \frac{25 - 10(2.2)}{\mathrm{sqrt}(10 \cdot 0.76)}\right) \approx \boxed{0.1382}$$

Not great, since the true answer is about 0.186. But, there is a way to do much better! The result of summing up a bunch of integers will itself be an integer. And generally speaking, the area under the normal of the interval $[24.5, 25.5]$ will be better at estimating the probability that the random variable is 25, then then interval $[25, 26]$. This leads to the use of the *half-integer correction*.

> **Definition 56**
> For random variables $X_1, X_2, \ldots, X_n$, the **half-integer correction** to the CLT is the following. For any integer a,
>
> $$\mathbb{P}(X_1 + \cdots + X_n \leq a) = \mathbb{P}(X_1 + \cdots + X_n \leq a + 1/2),$$
>
> and
>
> $$\mathbb{P}(X_1 + \cdots + X_n \geq a) = \mathbb{P}(X_1 + \cdots + X_n \leq a - 1/2),$$

Applying this to our example gives:

$$\mathbb{P}(X_1 + \cdots + X_{10} \geq 25) = \mathbb{P}(X_1 + \cdots + X_{10} \geq 24.5)$$
$$= \mathbb{P}\left(\frac{X_1 + \cdots + X_{10} - 10(2.2)}{\text{sqrt}(10 \cdot 5.6)} \geq \frac{24.5 - 10(2.2)}{\text{sqrt}(10 \cdot 0.76)}\right)$$
$$\approx \mathbb{P}\left(Z \geq \frac{24.5 - 10(2.2)}{\text{sqrt}(10 \cdot 5.6)} \geq \frac{24.5 - 10(2.2)}{\text{sqrt}(10 \cdot 0.76)}\right) \approx \boxed{0.1822}$$

which is much closer to the true answer.

Problems

19.1 Let $D_1, \ldots, D_8$ be iid rolls of a fair eight-sided die. Approximate the probability that $\sum D_i \geq 30$ using the CLT.

19.2 Let $A_1, \ldots, A_{10}$ be iid Exp(2). Approximate $\mathbb{P}(A_1 + \cdots + A_{10} \geq 7)$ using the CLT.

19.3 Suppose that R has density

$$f_R(r) = 2r \cdot \mathbb{1}(r \in [0, 1]).$$

    a) What is the expected value of R?

    b) What is the variance of R?

    c) Say that $R_1, R_2, \ldots$ are independent random variables with the same distribution as R. Using the CLT, approximately what is

$$\mathbb{P}(R_1 + \cdots + R_{100} \geq 70)?$$

    d) What is the expected value of R conditioned on $R \in [0.3, 0.5]$?

19.4 Suppose X has density

$$f_X(x) = (3/4)x(2 - x)\mathbb{1}(x \in [0, 1]).$$

    a) What is $\mathbb{E}[X]$?

b) What is $\mathrm{SD}(X)$?

c) For $X_1, X_2, \ldots, X_{20}$, approximate with the CLT $\mathbb{P}(X_1 + \cdots + X_{20} \geq 13.4)$.

19.5 Let $U_1, U_2, \ldots, U_{12}$ be standard uniform random variables (so uniform over $[0, 1]$.) Then approximate

$$\mathbb{P}(U_1 + \cdots + U_{12} < 7)$$

using the CLT.

19.6 Let $R_1, R_2, \ldots, R_{30}$ be uniform over $\{1, 2, 3, 4, 5, 6\}$. Approximate $\mathbb{P}(R_1 + \cdots R_{30} < 100)$ using the CLT.

The Bernoulli Process

Question of the Day Suppose that $B_1, \ldots, B_{25}$ are independent identical trials that are either 0 or 1, with $\mathbb{P}(B_i = 1) = 0.6$. Let $S = B_1 + \cdots B_{20}$. What is $\mathbb{P}(S = 16)$?

Summary
A **Bernoulli random variable** is either 1 (with probability p) or 0 (with probability $1 - p$. Write $X \sim \mathsf{Bern}(p)$. It represents the number of successes on a single trial that can be considered either a success or a failure.
A **Bernoulli process** is a stream of random variables $B_1, B_2, \ldots$ that are iid $\mathsf{Bern}(p)$. From the Bernoulli process, we can create **binomial**, **geometric**, and **negative binomial random variables**. If you have n trials, then you can think of the number of successes as the sum of n independent Bernoulli random variables. Call this distribution **Binomial**, and write $X = X_1 + \cdots + X_n \sim \mathsf{Bin}(n, p)$. For $X \sim \mathsf{Bin}(n, p)$,

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \mathbb{1}(k \in \{0, 1, \ldots, n\}).$$

20.1   The Bernoulli distribution

Suppose that I have an experiment that has two outcomes, either success or failure. I record successes using a 1 and failures using a 0. Then a single experiment, a single 1 or 0 random variable, is said to have a *Bernoulli distribution*

Definition 57
Say that X has a **Bernoulli distribution** with parameter p, and write $X \sim \mathsf{Bern}(p)$ if

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

Remark   Bernoulli random variables are also called *indicator random variables*, since if $X = \mathbb{1}(Y \in A)$ for any random variable Y, X will also be either 0 or 1, and hence have a Bernoulli distribution with $p = \mathbb{P}(Y \in A)$.

Fact 65
The mean of a Bernoulli random variable is p, and the variance is $p(1 - p)$.

Proof. For $B \sim \text{Bern}(p)$, the mean is $p(1) + (1-p)(0) = p$, while the variance is

$$\mathbb{E}[B^2] - \mathbb{E}[B]^2 = p(1)^1 + (1-p)(0)^2 - p^2 = p - p^2 = p(1-p).$$

$\square$

In general, any collection of random variables is called a *stochastic process.* For the first stochastic process that we will study, all of the variables have the Bernoulli distribution.

> **Definition 58**
> If $B_1, B_2, \ldots \sim \text{Bern}(p)$ are iid, call the $\{B_i\}$ a **Bernoulli process**.

We have already seen another distribution based on the Bernoulli process, the binomial distribution.

> **Fact 66**
> Let $B_1, B_2, \ldots$ be a Bernoulli process with parameter p and n a positive integer. Then
> $$S_n = B_1 + \cdots + B_n \sim \text{Bin}(n, p).$$

In the Question of the Day, each $B_i \sim \text{Bern}(0.6)$, and $n = 20$. Therefore $S \sim \text{Bin}(20, 0.6)$, which means

$$\mathbb{P}(S = 16) = \binom{20}{16} 0.6^{16} 0.4^4 \approx \boxed{0.03499}.$$

Because the Bernoulli random variables are independent, the mean and variance of the random variables add together. This immediately gives the following result.

> **Fact 67**
> Let $B \sim \text{Bin}(n, p)$. Then $\mathbb{E}[B] = np$, $\mathbb{V}(B) = np(1-p)$.

20.2   The Geometric distribution

A Bernoulli process is a sequence of 0's and 1's. A typical run might look like

$$(B_1, B_2, \ldots) = 00001000000011001110011001101011100010010010001 \cdots$$

Consider the positions in the sequence where there is a 1. There is a 1 at position 5, position 13, position 14, and so on. The smallest numbered position that is 1 can be written using the infimum function:

$$G = \inf\{i : B_i = 1\}.$$

As stated earlier in the text, we call this random variable *geometric* with parameter p and write $G \sim \text{Geo}(p)$.

> **Example 47**
> In the qotd, for $G = \inf\{i : B_i = 1\}$, what is $\mathbb{P}(G = 4)$?
>
> **Answer**   In order for $G = 4$, the sequence must start out 0001. Each 0 has probability $1 - p$ of occurring, and the final 1 has probability p. In the qotd, $p = 0.6$, so the answer is
> $$(0.4^3)(0.6) = \boxed{0.03840}.$$

> **Fact 68** (Geometric density)
> A geometric random variable with parameter p has probability $(1-p)^{i-1}p$ of equaling i for any positive integer i.

Remark   There are two common definitions for the geometric random variable in practice. One is what is given here, the other only counts the 0's *before* the first one, so is equivalent to $G-1$. Be sure when you see geometric random variables, that you know which definition is in use!

Note that the Bernoulli process is *memoryless*. This leads to the following fact about geometric random variables.

> **Fact 69**
> For $G \sim \text{Geo}(p)$, and $a \in \{1, 2, \ldots\}$
> $$[G|G > a] \sim a + G.$$

In words, if we wait a tries without seeing a 1, the number of tries we need until the next 1 will also be a geometric random variable.

Proof. This is shown by verifying that the densities are the same. Let a and i be positive integers. Then

$$\mathbb{P}(G = i|G > a) = \frac{\mathbb{P}(G = i, G > a)}{G > a}$$
$$= p(1-p)^{i-1}\mathbb{1}(i > a)/(1-p)^a$$
$$= p(1-p)^{i-a-1}\mathbb{1}(i - a > 0).$$

Similarly,

$$\mathbb{P}(a + G = i) = \mathbb{P}(G = i - a) = p(1-p)^{i-a-1}\mathbb{1}(i - a > 0).$$

Since the densities are the same, the distribution of the random variables are the same.   □

In particular,
$$[G|B_1 = 0] \sim [G|G > 1] \sim 1 + G.$$

Earlier we used this with the Fundamental Theorem of Probability to show that $\mathbb{E}[G] = 1/p$.

> **Example 48**
> A fair six sided die is thrown until a 4 shows up. What is the average number of throws of the die needed?
>
> **Answer**   The number of throws will be geometrically distributed with parameter $1/6$. This is because that is the chance of a success, which in this case is when a 4 comes up. Hence the expected number of throws is
> $$\frac{1}{1/6} = \boxed{6}.$$

Now for the variance of a geometric random variable.

> **Fact 70**
> For $G \sim \mathrm{Geo}(p)$, $\mathbb{V}(G) = (1-p)/p^2$.

Proof. By the FTP

$$
\begin{aligned}
\mathbb{E}[G^2] &= \mathbb{E}[\mathbb{E}[G^2|B_1]] \\
&= \mathbb{P}(B_1 = 1)\mathbb{E}[G^2|B_1 = 1] + \mathbb{P}(B_1 = 0)\mathbb{E}[G^2|B_1 = 0] \\
&= p\mathbb{E}[G^2|B_1 = 1] + (1-p)\mathbb{E}[G^2|B_1 = 0] \\
&= p + (1-p)\mathbb{E}[(1+G)^2] \\
&= p + (1-p)\mathbb{E}[1 + 2G + G^2] \\
&= p + (1-p) + 2(1-p)/p + (1-p)\mathbb{E}[G^2].
\end{aligned}
$$

Bring the $(1-p)\mathbb{E}[G^2]$ over the other side to get

$$
\begin{aligned}
p\mathbb{E}[G^2] &= p + (1-p) + 2(1-p)/p \\
\mathbb{E}[G^2] &= 1 + (1-p)/p + 2(1-p)/p^2.
\end{aligned}
$$

Therefore the variance is

$$
\begin{aligned}
\mathbb{V}(G) &= \mathbb{E}[G^2] - \mathbb{E}[G]^2 \\
&= 1 + (1-p)/p + 2(1-p)/p^2 - 1/p^2 \\
&= (p^2 + p - p^2 + 2(1-p) - 1)/p^2 = (p + 1 - 2p)/p^2 = (1-p)/p^2.
\end{aligned}
$$

$\square$

20.3   The Negative Binomial distribution

The geometric random variable with parameter p is the first time a 1 appears in a Bernoulli process with parameter p. A *Negative Binomial* with parameters k and p is the kth time a 1 shows up in a a Bernoulli process with parameter p.

> **Definition 59**
> Say that N has a **Negative binomial** distribution with parameters k and p and write $N \sim \mathrm{NegBin}(k,p)$ if
> $$N = \inf\{i : B_1 + \cdots + B_i = k\}.$$

In a binomial distribution, the number of trials is fixed by the parameter n and the random variable is the number of 1's that appear in the first n draws. In a negative binomial distribution, the number of 1's is fixed, and the random variable is the number of trials needed for k different 1's to appear in the draws.

> **Fact 71**
> If $N \sim \mathrm{NegBin}(k,p)$, then
> $$\mathbb{P}(N = i) = \binom{i-1}{k-1}(1-p)^{i-k}p^k$$

Proof. If $N = i$ then $B_i = 1$ and there are $k - 1$ differnt 1's appearing in $B_1, \ldots, B_{i-1}$. There are $\binom{i-1}{k-1}$ ways to choose where these 1's appear. Each such sequence contains k different 1's and $i - k$ different 0's, and so has probability $(1-p)^{i-k}p^k$ of appearing. □

Consider $N \sim \text{NegBin}(2, p)$. The first 1 appears in position G_1 where $G_1 \sim \text{Geo}(p)$. Then it is like the process starts over from scratch, and we have to wait a geometric number of times for the next 1 to appear.

Fact 72
Let $G_1, G_2, \ldots, G_k$ be iid $\text{Geo}(p)$. Then

$$G_1 + \cdots + G_k \sim \text{NegBin}(k, p).$$

This immediately gives the mean and variance of a negative binomial.

Fact 73
For $N \sim \text{NegBin}(k, p)$, $\mathbb{E}[N] = k/p$, $\mathbb{V}(N) = k(1-p)/p^2$.

20.4 Point perspective

Instead of keeping track of the Bernoulli random variables, we can just keep track of the *points* where the Bernoulli's are 1.

Definition 60
For a Bernoulli process $\{B_i\}$, let $P = \{i : B_i = 1\}$ be a **Bernoulli point process** and call the elements of P **points**.

Example 49
If the Bernoulli process starts

$$1, 1, 1, 0, 0, 1, 0, 1, 1, 0,$$

The first few points are

$$P = \{1, 2, 3, 6, 8, 9\}.$$

Their graph looks like
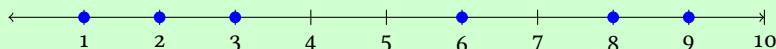
This point process has certain properties.

Fact 74

A Bernoulli point process with parameter p satisfies the following.

1. If A and B are disjoint subsets of $\{1, 2, 3, \ldots\}$, then

$$\#(P \cap A) \text{ and } \#(P \cap B)$$

   are independent random variables.

2. For all $A \subseteq \{1, 2, \ldots\}$,

$$\mathbb{E}[\#(P \cap A)] = p \cdot \#(A).$$

Note that we can recast our distributions in terms of these points.

Fact 75

Consider a Bernoulli point process $P = \{P_1, P_2, \ldots\}$ with parameter p, where $P_1 < P_2 < P_3 < \ldots$.

$$P_1 \sim \mathsf{Geo}(p)$$
$$\forall r \geq 1, P_{r+1} - P_r \sim \mathsf{Geo}(p)$$
$$P_r \sim \mathsf{NegBin}(r, p)$$
$$\#(P \cap A) \sim \mathsf{Bin}(\#(A), p)$$

This fact follows immediately from the fact that the B_i are all iid $\mathsf{Bern}(p)$ random variables.

Problems

20.1 Suppose $X \sim \mathsf{Bin}(34, 0.23)$. What is $\mathbb{E}[X]$?

20.2 Say $W \sim \mathsf{Bin}(10, 0.2)$ and $Y \sim \mathsf{Bin}(10, 0.3)$. What is $\mathbb{E}[W + Y]$?

20.3 Let $G \sim \mathsf{Geo}(0.38)$.

    a) What is $\mathbb{E}[G]$?

    b) What is $\mathbb{V}[G]$

20.4 Suppose I roll a fair six sided die over and over until I get a 5. Let T be the number of rolls that I make. What is $\mathbb{E}[T]$?

20.5 Let $N \sim \mathsf{NegBin}(20, 0.38)$.

    a) What is $\mathbb{E}[N]$?

    b) What is $\mathbb{V}[N]$

20.6 Let B_i be a Bernoulli process with parameter 0.4. What is the chance that $B_1 + \cdots + B_5 = 4$?

20.7 Suppose $X \sim \mathsf{Bin}(13, 0.2)$ and $Y \sim \mathsf{Bin}(27, 0.2)$ are independent. What is the distribution of $X + Y$?

20.8 Let B_i be a Bernoulli process with parameter 0.2.

    a) Find $\mathbb{P}(\inf\{i : B_i = 1\} = 4)$

    b) Find $\mathbb{P}(\inf\{i : B_i = 0\} = 4)$

20.9 Let Y be a positive integer valued random variable with $\mathbb{E}[Y] = 4.2$, and $[X|Y] = \text{Bin}(Y, 0.3)$. Then what is $\mathbb{E}[X]$?

20.10 Find the moment generating function of a geometric random variable with parameter p.

20.11 Find $\mathbb{E}[G^2]$ for a geometric random variable by conditioning on B_1 and taking the expectation again.

Poisson point processes in one dimension

Question of the Day Suppose that a beam 2 meters long has defects modeled as a Poisson point process of rate $\lambda = 0.3/\text{meter}$. What is the probability that there are two or more defects in the beam?

Summary
Constuct a set P that is a **Poisson point process P of rate λ over** $[0, \infty)$ as follows. Let $A_1, A_2, \ldots$ be an iid sequence of exponential random variables of rate λ. Then set

$$P_1 = A_1$$
$$P_2 = A_1 + A_2$$
$$P_3 = A_1 + A_2 + A_3$$
$$\vdots \;\; = \;\; \vdots$$

Then let $P = \{P_1, P_2, P_3, \ldots\}$.
This means that P_i has a **gamma** distribution with parameters i and λ. Write $G \sim \text{Gamma}(\alpha, \beta)$ if G has density

$$f_G(s) = \lambda^r s^{r-1} \exp(-\lambda s) \mathbb{1}(s \geq 0)/\Gamma(r).$$

Here $\Gamma(r)$ is the gamma function that equals $(r - 1)!$ when r is an integer and $\int_0^\infty x^{r-1} \exp(-x) \, dx$ when r is not.

A Bernoulli process $B_1, B_2, B_3, \ldots$ where $B_i \sim \text{Bern}(p)$ is associated with a Bernoulli point process where $P = \{i : B_i = 1\}$.
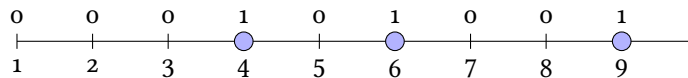
The expected number of points that fall into a set A is just the number of points in A times p. Since the B_i are independent, for $A \cap B = \emptyset$, the number of points that fall into A and the number of points that fall into B are independent of each other. That is,

1. For A, B disjoint subsets of $\{1, 2, \ldots\}$, $\#(P \cap A)$ and $\#(P \cap B)$ are independent random variables.

2. For A a subset of $\{1, 2, \ldots\}$, $\mathbb{E}[\#(P \cap A)] = p \cdot \#(A)$.

Here the points are limited to lie on the integers $\{1, 2, 3, \ldots\}$. Consider letting the points be anywhere in $[0, \infty)$.
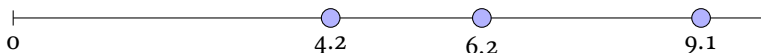
We can build such a point process over $[0, \infty)$ by changing the measure in Property 2 from counting measure to Lebesgue measure.

> **Definition 61**
> Say that P is a **Poisson point process of rate λ over** $[0, \infty)$ if
>
> 1. For A and B disjoint measurable subsets of $[0, \infty)$, $\#(P \cap A)$ and $\#(P \cap B)$ are independent random variables.
>
> 2. For A a measurable subset of $[0, \infty)$, $\mathbb{E}[\#(P \cap A)] = \lambda \cdot \ell(A)$, where ℓ is Lebesgue measure.

Recall that for a Bernoull point process $P = \{P_1, P_2, \ldots\}$ where $P_1 < P_2 < \cdots$,

$$P_1 \sim \mathsf{Geo}(p)$$
$$\forall r \geq 1, P_{r+1} - P_r \sim \mathsf{Geo}(p)$$
$$P_r \sim \mathsf{NegBin}(r, p)$$
$$\#(P \cap A) \sim \mathsf{Bin}(\#(A), p).$$

Similarly, the distribution of the first point, the distribution of the difference between successive points, the distribution of the rth point, and the distribution of the number of points in a given interval can all be described for a Poisson point process as well.

> **Fact 76**
> For $P = \{T_1, T_2, \ldots\}$ a Poisson point process of rate λ over $[0, \infty)$, where $T_1 < T_2 < \cdots$,
>
> $$T_1 \sim \mathsf{Exp}(\lambda)$$
> $$\forall r \geq 1, T_{r+1} - T_r \sim \mathsf{Exp}(\lambda)$$
> $$T_r \sim \mathsf{Gamma}(r, \lambda)$$
> $$\#(P \cap A) \sim \mathsf{Pois}(\ell(A) \cdot \lambda).$$

Here $\mathsf{Pois}(\mu)$ is the Poisson distribution with mean μ. It is defined as follows.

> **Definition 62**
> Say that $X \sim \mathsf{Pois}(\mu)$ if it has density
>
> $$f_X(i) = \exp(-\mu)\frac{\mu^i}{i!}\mathbb{1}\{i \in \{0, 1, 2 \ldots\}\}.$$

So the exponential distribution is the continuous analogue of the geometric distribution, the gamma is the continuous analogue of the negative binomial, and Poisson counts points in continuous space where Binomial counts points in discrete space.

To understand why the Poisson distribution has the form that it does, it helps to consider the question of the day, and what is known as the exponential space.

21.1 The exponential space and Poisson distribution

Let $A = [0, 2]$ represent the beam in the qotd. Then if P is a Poisson point process of rate 0.3 over $[0, \infty)$, then $P \cap [0, 2]$ are the points that fall in the beam's length. So by our fact,

$$\#(P \cap A) \sim \text{Pois}(0.3 \cdot 2).$$

Why is this?

Well, $P \cap A$ could contain no points, or 1 point, or 2 points, and so on. The space with no point is $\emptyset$. The space with one point is $\binom{[0,2]}{1}$. The space with two points is $\binom{[0,2]}{2}$, and so on. Therefore the set of points P is a subset of

$$\emptyset \cup \binom{[0, 2]}{1} \cup \binom{[0, 2]}{2} \cup \binom{[0, 2]}{3} \cup \cdots,$$

which is known as the *exponential space*.

Recall that notation like $\binom{[0,2]}{i}$ means the set of all subsets of $[0, 2]$ of size i. What is the measure of $\binom{[0,2]}{i}$? Well, the measure of a single point from $[0, 2]$ is 2. The measure of $[0, 2]^2$ is 2^2, and in general the measure of $[0, 2]^i$ is 2^i.

However, that is the measure of the i-tuples where the order of the points matters. To get the measure of the *subsets* where the order does not matter, divide by the number of ways to order the points. This is $i!$. For instance, the vectors $(0.3, 1.3)$ and $(1.3, 0.3)$ both map to the subset $\{0.3, 1.3\}$. If I have a subset of size i, there are $i!$ vectors that map to the subset. So we only measure the subsets of size i, that only has measure

$$\text{measure}\left(\binom{[0, 2]}{i}\right) = \frac{2^i}{i!}$$

Now, how does the rate figure in? Well, one way to view the rate λ is that it gives a bonus factor for having more points in the set. With two points, we get a bonus factor of λ^2, with seven points λ^7, and so on.

So if we have three points, they get a bonus factor of λ^3, and since there are $2^3/3!$, they overall contribute

$$\frac{(2\lambda)^3}{3!}$$

so the measure. When we add this up for 0, 1, 2, or any nonnegative integer number of points, we get

$$1 + \frac{(2\lambda)}{1!} + \frac{(2\lambda)^2}{2!} + \frac{(2\lambda)^3}{3!} + \cdots .$$

Look familiar? This is exactly the Taylor series expansion of $\exp(2\lambda)$. So to normalize this expression so that everything sums to 1, we multiply by $\exp(-2\lambda)$.

That means that the probability that we are in the part of the space that has three points is $\exp(-2\lambda)(2\lambda)^3/3!$. The probability that we are in the four part part of the space is $\exp(-2\lambda)(2\lambda)^4/4!$. And so on. If we let N denote the number of points in the process, then that is why we have the distribution of a Poisson that we do. For $N \sim \text{Pois}(2\lambda)$,

$$\mathbb{P}(N = i) = \exp(-2\lambda)\frac{(2\lambda)^i}{i!}\mathbb{1}(i \in \{0, 1, 2, \ldots\}).$$

Note that the 2 for the question of the day comes from the fact that the Lebesgue measure of $[0, 2]$ is 2. With our notation $\ell([0, 2]) = 2$, so $\lambda \cdot \ell([0, 2]) = (0.3)(2) = 0.6$, so $\mathbb{P}(N \geq 2)$ is

$$
\begin{aligned}
\mathbb{P}(N \geq 2) &= 1 - \mathbb{P}(N \leq 1) \\
&= 1 - \mathbb{P}(N = 1) - \mathbb{P}(N = 0) \\
&= 1 - \exp(-0.6)[(0.6)^0/0! + (0.6)^1/1!] \approx \boxed{0.1219\ldots}.
\end{aligned}
$$

So the number of points in $P \cap A$ is Poisson distributed with parameter $\lambda \cdot \ell(A)$. Since the average number of points is also $\lambda \cdot \ell(A)$, that means that the parameter of a Poisson distribution is its own parameter.

> **Fact 77**
> For $N \sim \mathsf{Pois}(\mu)$, $\mathbb{E}[N] = \mu$.

Poissons random variables also have the nice property that the variance is also μ.

> **Fact 78**
> For $N \sim \mathsf{Pois}(\mu)$, $\mathbb{V}(N) = \mu$.

Oftentimes, such processes are used to model arrival times of events, such as a customer arriving. Hence the points T_i is the process are often called *arrival times*. The times between arrivals, $P_i - P_{i-1}$ are often called *interarrival times*.
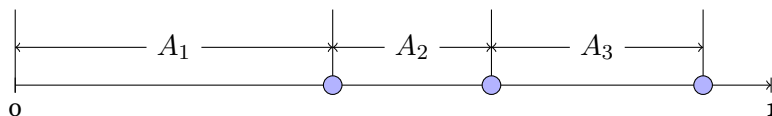
How do we know that the time of the first arrival has an $\mathsf{Exp}(\lambda)$ distribution? Well, consider

$$
\mathbb{P}(T_1 > a) = \mathbb{P}(\#(P \cap [0, a]) = 0) = \exp(-\lambda a).
$$

That is just the survival function of an exponential random variable with rate λ.

The remaining interarrival times are found in a similar fashion.

21.2   The Gamma distribution

Now consider the time of the r arrival, T_r. Let $s > 0$. What has to happen for T_r to fall into a small differential interval around s of width ds? That is, what is $\mathbb{P}(T_r \in ds)$?

Two things have to happen for $T_r \in ds$.

1. There has to be a point in the differential interval around s.

2. There have to be $r - 1$ points in the interval $[0, s]$.

3. The rest of the space has to be empty.

Here's the intuition. By that we mean that this is why it is true, but would not be considered a rigorous proof! For a Poisson point process, a little differential element around s contains either 0 points or 1 point. So it is Bernoulli, and we know the mean is $\lambda \, ds$ since ds is the measure of the differential interval.

Okay, now what is the chance that the interval $[0, s]$ contains $r - 1$ points! That has a Poisson distribution with parameter λs, so

$$\exp(-\lambda s)\frac{(\lambda s)^{r-1}}{(r-1)!}.$$

Multiplying by $\lambda\, ds$ for the probability that there is an rth point near s gives

$$\mathbb{P}(T_r \in ds) = \frac{\lambda^r \exp(-\lambda s)s^{r-1}}{(r-1)!}$$

This inspires the definition of the *gamma distribution*.

Definition 63
Say X has a **gamma distribution** with parameters α and λ if it has density

$$f_X(s) = \frac{\lambda^\alpha \exp(-\lambda s)s^{\alpha-1}}{\Gamma(\alpha)}\mathbb{1}(s \geq 0),$$

where

$$\Gamma(\alpha) = \int_0^\infty \exp(-s)s^{\alpha-1}\, ds$$

is called the **Gamma function** and normalizes the density.

Fact 79
When k is a positive integer, $\Gamma(k) = (k-1)!$.

Fact 80
For $P_1 < P_2 < \cdots$ an ordered Poisson point process of rate λ, $P_i \sim \text{Gamma}(i, \lambda)$.

Remarks

- We motivated the gamma distribution by considering α a positive integer, but the gamma distribution is defined for any $\alpha \geq -1$.

- For k a positive integer, the distribution $\text{Gamma}(k, \lambda)$ is also known as the **Erlang distribution** after the Danish mathematician Agner Erlang who invented much of queuing theory.

Example 50
Suppose $P_1 < P_2 < \cdots$ is an ordered Poisson point process of rate 2.5. What is the chance that $P_3 \in [1, 2]$?

Answer We know $P_3 \sim \text{Gamma}(3, 2.5)$, so

$$\mathbb{P}(P_3 \in [1, 2]) = \int_{s=1}^2 \frac{(2.5)^3 s^2 \exp(-2.5s)}{2!}\, ds \approx \boxed{0.4191}.$$

Problems

21.1 For P a PPP over $[0, \infty)$ of rate 2, what is the distribution of $\inf(P)$?

21.2 For P a PPP over $[0, \infty)$ of rate 3.2, what is the expected value of $\inf(P)$?

21.3 Let P be a Poisson point process over $[0, \infty)$ of rate 1.8, and $P_1 = \inf(P)$. What is $\mathbb{P}(P_1 \leq 1)$?

21.4 Suppose $T_1, T_2, \ldots$ are an iid sequence of $\mathsf{Exp}(2)$ random variable. Let

$$N = \sup\{n : T_1 + \cdots + T_n \leq 4.1\}.$$

What is $\mathbb{P}(N = 8)$?

21.5 Suppose that times that a bus arrives at a stop over one hour is modeled as a Poisson point process of rate $1.4/\text{hr}$.

    a) What is the chance that exactly one bus arrives in the hour?

    b) What is the expected number of buses that arrive in the hour?

    c) What is the expected number of buses that arrive in the first half hour?

21.6 Requests for information at Honnold library during finals week arrive according to a Poisson process at rate 4.2 per hour.

    a) What is the expected number of requests seen during a six hour shift?

    b) What is the chance that the third request arrives before the end of the first hour?

    c) What is the covariance between the time of the third request and the time of the fourth request?

    d) Each request (independently) has a 5% chance of being unsolvable. What is the chance that at least one unsolvable request comes in during a six hour shift?

21.7 For a Poisson point process over $[0, \infty)$ of rate λ, let $N_A = \#(P \cap A)$. Then find $\mathrm{Cov}(N_{[0,2)}, N_{[0,3)})$.

The Poisson point process

Question of the Day Lightning strikes in a forest covering 3 square miles are modeled as occuring at rate 21.2 per square mile using a Poisson point process. What is the expected number of strikes over the whole forest?

Summary When points are placed over a space Ω such that the number of points in disjoint sets is independent, and the mean number of points in a set is given by the measure μ of the set, the points form a **Poisson point process**. Write $P \sim \mathrm{PPP}(\Omega, \mu)$. For A a measurable subset of Ω,

$$N_A = \#(P \cap A) \sim \mathsf{Pois}(\mu(A)).$$

One type of random process that we have not yet considered is when we have *spatial* data, which consists of a set of points in some space chosen uniformly at random. We need a way to model such data, for instance

1. Location of outbreaks of a disease in a community.

2. Flaws in a sheet of metal.

3. Cancerous cells in a tissue sample.

To handle this and more general situations, we now give our most general definition of a Poisson point process over a region Ω together with a measure μ.

Definition 64
A collection P of points in Ω is a **Poisson point process of rate measure μ over Ω** (write $P \sim \mathrm{PPP}(\Omega, \mu)$ if P satisfies two properties.

1. For A and B disjoint measurable subsets of Ω, $\#(P \cap A)$ and $\#(P \cap B)$ are independent random variables.

2. For A a measurable subset of Ω, $\mathbb{E}[\#(P \cap A)] = \mu(A)$.

Consider the following rate measures.

1. Bernoulli point process: $\Omega = \{1, 2, 3, \ldots\}$, $\mu(A) = p \cdot \#(A)$.

2. Poisson point process with rate λ over $[0, \infty)$: $\Omega = [0, \infty)$, $\mu(A) = \lambda \cdot \ell(A)$, where ℓ is Lebesgue measure.

3. In the Question of the Day, $\Omega = \mathbb{R}^2$, $\mu(A) = 21.2 \cdot \ell(A)$.

Note that the final value of $\mu(A)$ should be unitless since it counts the average number of points in region A.

> **Example 51**
> **Question of the day** Lightning strikes in a forest covering 3 square miles occur at rate 21.2 per square mile as a Poisson point process. What is the expected number of strikes over the whole forest?
>
> **Answer**   The overall mean is the measure of the space (3 square miles) times the rate (21.2 per square mile) or $\boxed{63.60}$.

Once we know the mean of the number of points in A, we actually know the entire distribution.

> **Fact 81**
> For $P \sim \text{PPP}(\Omega, \mu)$ and A a measurable subset of A.
>
> $$N_A = \#(P \cap A) \sim \text{Pois}(\mu(A)).$$

22.1   Summing independent Poisson random variables

One of the nice things about Poisson random variables is that if you add two independent Poissons, then the result is still a Poisson random variable!

> **Fact 82**
> Suppose $N_1 \sim \text{Pois}(\mu_1)$ and $N_2 \sim \text{Pois}(\mu_2)$ are independent. Then $N_1 + N_2 \sim \text{Pois}(\mu_1 + \mu_2)$.

Proof. Suppose that we have a Poisson point process P_1 of rate 1 on $[0, \mu_1]$, and P_2 is a PPP of rate 1 on $(\mu_1, \mu_1 + \mu_2]$. Note that $\#(P_1) \sim \text{Pois}(\mu_1)$ and $\#(P_2) \sim \text{Pois}(\mu_2)$.

Let $P = P_1 \cup P_2$ be the combination of these two point processes. The result is a Poisson point process of rate 1 over $[0, \mu_1 + \mu_2]$. It has a number of points equal to $\#(P_1 + P_2) = \#(P_1) + \#(P_2)$. So

$$\#(P_1) + \#(P_2) \sim \text{Pois}(\mu_1 + \mu_2),$$

by the properties of Poisson point processes. $\square$

Recall that if $Z_1 \sim \text{N}(\mu_1, \sigma_1^2)$ and $Z_2 \sim \text{N}(\mu_2, \sigma_2^2)$ are independent, then $Z_1 + Z_2$ is also a normally distributed random variable, with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$. So for both normal random variables and Poisson random variables, adding independent draws togther stays in the same family, with the appropriate parameter choice to make the mean and variance work out.

For independent Poisson random variables with means μ_1 and μ_2, the mean of the sum must be $\mu_1 + \mu_2$, and the variance is $\mu_1 + \mu_2$ because they are independent. This fits nicely with the fact that the sum is a Poisson random variable with parameter (and so mean and variance) equal to $\mu_1 + \mu_2$.

22.2   Thinning

Consider the following problem.

> **Example 52**
> Suppose that arrivals to a queue occur according to a Poisson process at rate 3 per hour.
> Each arrival independently has a 40% chance of requiring a long service, and 60% chance
> of requiring a short service. What is the chance that there are at least two long services
> in the first hour?

This is an example of a problem where the notion of *thinning* is used. Remember that the rate of 3 per hour indicates that in a tiny time interval dt, we expected to have $3\,dt$ arrivals. But since only 40% of those arrivals require a long service, the probability of a long service arrival drops to $3(0.4)\,dt = 1.2\,dt$.

In other words, if we only consider those arrivals that are long service arrivals, they form a new Poisson process at rate 1.2 per hour. So the example above has answer equal to $\mathbb{P}(N \geq 2)$ where $N \sim \mathrm{Pois}((1.2)(1))$.

$$\mathbb{P}(N \geq 2) = 1 - \mathbb{P}(N = 0) - \mathbb{P}(N = 1) = 1 - \exp(-1.2) - \frac{1.2}{2}\exp(-1.2) = \boxed{0.5180\ldots}.$$

> **Definition 65**
> Let P be a Poisson point process of rate measure $\mu(A)$. Let $f : A \to [0,1]$ assign a probability to each point in A. For $P = \{P_1, \ldots, P_n\}$ independently draw $B_1, \ldots, B_n \sim \mathrm{Bern}(p)$. If $P' = \{P_i : B_i = 1\}$, call P' the **thinned Poisson point process**.

> **Fact 83** (Thinning a Poisson point process)
> If you thin a Poisson point process of rate measure μ over A using the same probability p of retaining every $a \in P$, then the result is a Poisson point process of rate $p\mu$ over A.

Proof. Let P' be the thinned version of the Poisson point process P. The independence of $\#(P' \cap A)$ and $\#(P' \cap B)$ for disjoint A and B follows from the independence of $\#(P \cap A)$ and $\#(P \cap B)$.

Let A be a measurable set. Then

$$\begin{aligned}
\mathbb{E}[\#(P' \cap A)] &= \mathbb{E}[\mathbb{E}[\#(P' \cap A)|\#(P \cap A)]] \\
&= \mathbb{E}[p\#(P \cap A)]] \\
&= p\mathbb{E}[\#(P \cap A)] = p\mu(A).
\end{aligned}$$

$\square$

22.3   Conditioning on the number of points

Let A and B be disjoint sets. Suppose I know that a point of the Poisson point process is either in A or B, so $a \in A \cup B$. What is the chance that the point falls into A? Not surprisingly, the odds that it lands in A versus B is the rate measure of A against the rate measure of B.

Fact 84
Let a be a point of a Poisson point process of rate measure μ over $A \cup B$, where A and B are disjoint. Then

$$\mathbb{P}(a \in A) = \frac{\mu(A)}{\mu(A) + \mu(B)}.$$

Example 53
Trains arrive as a Poisson point process of rate 2 per hour. Given that exactly one train arrives in the first three hours, what is the chance that it arrives in the first hour?

Answer   The first hour has measure $2(1 - 0) = 2$. The remaining two hours have measure $2(3 - 1) = 4$. Therefore, the chance that the train arrives in the first hour is

$$\frac{2}{2 + 4} = 1/3 = \boxed{0.3333\ldots}.$$

Moreover, all the points of a Poisson point process over a continuous space are independent of each other.

Fact 85
Let $P = \{P_1, \ldots, P_N\}$ be a Poisson point process over A. Then $P_1, \ldots, P_N$ are independent random variables.

Since each train is independent, the number that fall into a particular region will be binomially distributed.

Fact 86
Let $P = \{P_1, \ldots, P_n\}$ be a Poisson point process over A of rate measure μ. For B a measurable subset of A, given the number of points in the PPP:

$$[\#(P \cap B) | \#(P)] \sim \mathsf{Bin}(\#(P), \mu(B)/\mu(A)).$$

Example 54
Continuing the last example, if there are exactly three trains in the first three hours, what is the chance that there is exactly one arriving in the first hour?

Answer   Each of these three trains is independent, and so the number of trains in the first hour is $N_{[0,1]} \sim \mathsf{Bin}(3, 1/3)$, and

$$\mathbb{P}(N_{[0,1]} = 3) = \binom{3}{1}(1/3)^1(2/3)^2 = 4/9 = \boxed{0.4444\ldots}.$$

Problems

22.1 Suppose N_1 and N_2 are independent Poisson random variables with means 2 and 3 respectively. What is the chance that $N_1 + N_2 = 5$?

22.2 Say $N_1, \ldots, N_{10}$ are Poisson random variables of rate 0.5. What is the chance that their sum is greater than 1?

22.3 EPA clean-up sites in a county are modeled as a Poisson point process of rate $\lambda = 3/\text{mi}^2$.

    a) If the region has an area of 9 square miles, what is the expected number of clean-up sites?

    b) If the region is known to have at least 25 clean up sites, what is the chance that it has at least 30 such sites? (Probably want to use a computer for the calculations on this one.)

22.4 An epidemiological model covers a city of size 4 square miles. Suppose the model is that disease outbreaks form a Poisson point process of rate 10 per square mile.

    a) What is the mean number of infected in the city?

    b) Using a computer, find the probability that more than the average number of people are infected.

    c) What is the mean number of infected in a neighborhood of size 2.3 square miles?

22.5 Suppose that $P \sim \text{Pois}([0, 2], \lambda \cdot \ell)$, where $\lambda > 0$ is a constant and ℓ is Lebesgue measure. If $N_{[0,2]} = 10$, what is the chance that $N_{[0,1]} = 4$?

22.6 Pine trees in a forest are modeled as occurring as a Poisson point process with rate 15.4 per square kilometer. Suppose the forest is divided into two pieces, a slope of size 14 square kilometers, and a flat region of size 23.1. Suppose there are 597 pine trees in the forest.

    a) What is the average number of trees on the slope region?

    b) What is the chance that there are more than the average number of trees on the slope region?

22.7 Outbreaks of a disease are modeled as coming from a Poisson point process with rate 2.3 per square mile.

    a) If the city is 3 square miles, what is the chance that there are exactly 6 outbreaks?

    b) Suppose the part of the city west of the river is 1.2 square miles (leaving 1.8 square miles east of the river). If there are exactly 8 outbreaks across the city, what is the chance that at least 3 of them are on the west side of the river?

22.8 Defects in a steel sheet are modeled as occurring at 6.1 per square meter. If there are 23 defects in a sheet of size 4 square meters, what is the chance that a portion one square meter in size has exactly 6 defects?

Joint densities in higher dimensions

Question of the Day Suppose (X_1, X_2, X_3) have joint density

$$f_{(X_1,X_2,X_3)}(x_1, x_2, x_3) = (1/3)[x_1 + 2x_2 + 3x_3]\mathbb{1}(x_1, x_2, x_3 \in [0,1]),$$

find $\mathbb{E}[X_1 X_2 X_3]$.

Summary

A random vector $(X_1, \ldots, X_n)$ has density $f_{X_1,\ldots,X_n}$ with respect to measure μ if for all events A,

$$\mathbb{P}((X_1, \ldots, X_n) \in A) = \int_{(x_1,\ldots,x_n)\in A} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\, d\mu.$$

To find the marginal distributions for higher dimensional integrals.

$$f_{X_i}(x_i) = \int_{(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)\in\mathbb{R}^{n-1}} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\, d\mu.$$

To find the expected value of a function of the variables:

$$\mathbb{E}[g(X_1, \ldots, X_N)] = \int_{\mathbb{R}^n} g(x_1, \ldots, x_n) f_{X_1,\ldots,X_N}(x_1, \ldots, x_n)\, d\mu$$

23.1   *Finding probabilities*

Recall that densities are used to calculate probabilities and to find expected values. Joint densities in higher dimensions are the same way. The same method holds for finding probabilities.

Fact 87

For A a measurable subset of $\mathbb{R}^n$, and $(X_1, \ldots, X_n)$ with joint density $f_{X_1,\ldots,X_n}$,

$$\mathbb{P}((X_1, \ldots, X_n) \in A) = \int_{(x_1,\ldots,x_n)\in A} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\, d\mu$$

Example 55
Let (X_1, X_2, X_3) have density

$$f_{(X_1,X_2,X_3)}(x_1, x_2, x_3) = (1/3)[x_1 + 2x_2 + 3x_3]\mathbb{1}(x_1, x_2, x_3 \in [0, 1])$$

with respect to Lebesgue measure. Then find $\mathbb{P}(\max\{X_1, X_2, X_3\} \leq 0.5)$.

Answer   The event $\{\max\{X_1, X_2, X_3\} \leq 0.5\}$ is the same as $\{X_1 \in [0, 0.5], X_2 \in [0, 0.5], X_3 \in [0, 0.5]\}$. Therefore the integral is

$$\int_{x_1 \in [0,0.5], x_2 \in [0,0.5], x_3 \in [0,0.5]} (1/3)[x_1 + 2x_2 + 3x_3] \, d\mathbb{R}^3 = 1/16 = \boxed{0.06250}.$$

23.2   Finding means

Expected values are also handled the same way as before, by the Law of the Unconscious Statistician:

Fact 88
For random variables $(X_1, \ldots, X_n)$ with joint density $f_{(X_1,\ldots,X_n)}$ with respect to μ, if $\mathbb{E}[g(X_1, \ldots, X_n)]$ exists then

$$\mathbb{E}[g(X_1, \ldots, X_n)] = \int_{(s_1,\ldots,s_n) \in \mathbb{R}^n} g(s_1, \ldots, s_n) f_{X_1,\ldots,X_n}(s_1, \ldots, s_n) \, d\mu.$$

This can be used to solve the Question of the Day!

Question of the day   Using the Law of the Unconscious Statistician,

$$\mathbb{E}[X_1 X_2 X_3] = \int_{(x_1,x_2,x_3) \in \mathbb{R}^3} x_1 x_2 x_3 (1/3)[x_1 + 2x_2 + 3x_3]\mathbb{1}(x_1, x_2, x_3 \in [0, 1]) \, d\mathbb{R}^3$$

$$= \int_{x_1 \in [0,1], x_2 \in [0,1], x_3 \in [0,1]} x_1 x_2 x_3 (1/3)[x_1 + 2x_2 + 3x_3] \, d\mathbb{R}^3$$

$$= 1/6 = \boxed{0.1666\ldots}.$$

23.3   Testing for independence

Remember that for bivariate random variables, the random variables are independent if the joint density factors into the marginal densities. The same fact holds in the higher dimensional case as well.

Fact 89 (Independence means joint is product of marginal densities)
Consider random variables $X_1, \ldots, X_n$ where each X_i has density f_{X_i} with respect to μ_i. Then the $\{X_i\}$ are independent if and only if $\prod_{i=1}^n f_i$ is a joint density for the $\{X_i\}$ with respect to the product measure $\times_{i=1}^n \mu_i$.

Fact 90
If the joint density f for $X_1, \ldots, X_n$ factors into the product of n densities, then the X_i are independent.

Example 56

Suppose $X_1, \ldots, X_n$ have joint density

$$f_{X_1,\ldots,X_n}(s_1, \ldots, s_n) = \exp\left(-\sum_{i=1}^{n} s_i\right) \mathbb{1}(s_1, \ldots, s_n \geq 0).$$

Show that the X_i are independent.

Answer   Since

$$\exp\left(-\sum_{i=1}^{n} s_i\right) = \prod_{i=1}^{n} \exp(-s_i)$$

$$\mathbb{1}(s_1, \ldots, s_n \geq 0) = \prod_{i=1}^{n} \mathbb{1}(s_i \geq 0),$$

$$f_{(X_1,\ldots,X_n)}(s_1, \ldots, s_n) = \prod_{i=1}^{n} \exp(-s_i)\mathbb{1}(s_i \geq 0)$$

Remember that to show random variables are not independent, all we need are events where the probability that event happens is not the product of the individual events.

Example 57

Let (X_1, X_2, X_3) have density

$$f_{(X_1,X_2,X_3)}(x_1, x_2, x_3) = (1/3)[x_1 + 2x_2 + 3x_3]\mathbb{1}(x_1, x_2, x_3 \in [0, 1])$$

with respect to Lebesgue measure. Show that the X_i are not independent.

Answer   In an earlier example, we showed that

$$\mathbb{P}(X_1 \in [0, 0.5], X_2 \in [0, 0.5], X_3 \in [0, 0.5]) = 1/16.$$

However,

$$\mathbb{P}(X_1 \in [0, 0.5]) = \int_{[0,0.5]\times[0,1]\times[0,1]} (1/3)(x_1 + 2x_2 + 3x_3)\, d\mathbb{R}^3 = 11/24,$$

$$\mathbb{P}(X_2 \in [0, 0.5]) = \int_{[0,1]\times[0,0.5]\times[0,1]} (1/3)(x_1 + 2x_2 + 3x_3)\, d\mathbb{R}^3 = 10/24,$$

$$\mathbb{P}(X_3 \in [0, 0.5]) = \int_{[0,1]\times[0,1]\times[0,0.5]} (1/3)(x_1 + 2x_2 + 3x_3)\, d\mathbb{R}^3 = 9/24,$$

And $(11/24)(5/24)(9/24) = 55/768 = 0.0716\ldots$ which does not equal $1/16 = 0.0625$.

23.4   Finding marginals

To find the marginal density of a random variable, we can *integrate out* the other variables.

Example 58

Let (X_1, X_2, X_3) have density

$$f_{(X_1, X_2, X_3)}(x_1, x_2, x_3) = (1/3)[x_1 + 2x_2 + 3x_3]\mathbb{1}(x_1, x_2, x_3 \in [0, 1])$$

with respect to Lebesgue measure. Find the density of X_1.

Answer

$$\mathbb{P}(X_1 \in A) = \mathbb{P}(X_1 \in A, X_2 \in \mathbb{R}, X_3 \in \mathbb{R})$$

$$= \int_{x_1 \in A} \int_{x_2 \in \mathbb{R}} \int_{x_3 \in \mathbb{R}} \frac{x_1 + 2x_2 + 3x_3}{3} \mathbb{1}(x_1, x_2, x_3 \in [0, 1]) \, dx_3 \, dx_2 \, dx_1$$

$$= \int_{x_1 \in A} \int_{x_2 \in \mathbb{R}} \frac{x_1 x_3 + 2x_2 x_3 + (3/2)x_3^2}{3} \Big|_0^1 \mathbb{1}(x_1, x_2 \in [0, 1]) \, dx_2 \, dx_1$$

$$= \int_{x_1 \in A} \int_{x_2 \in [0,1]} (1/3)[x_1 + 2x_2 + 3/2] | \mathbb{1}(x_1 \in [0, 1]) \, dx_2 \, dx_1$$

$$= \int_{x_1 \in A} (1/3)[x_1 x_2 + x_2^2 + (3/2)x_2] |_0^1 \mathbb{1}(x_1 \in [0, 1]) \, dx_1$$

$$= \int_{x_1 \in A} (1/3)[x_1 + 5/2]\mathbb{1}(x_1 \in [0, 1]) \, dx_1.$$

Hence the density of X_1 must be

$$\boxed{f_{X_1}(x_1) = (1/3)[x_1 + 5/2]\mathbb{1}(x_1 \in [0, 1]).}$$

Problems

23.1 Suppose (X_1, X_2, X_3) has joint density

$$f_{(X_1, \ldots, X_n)} \propto (x_1 + x_2)(x_1 + x_3)(x_2 + x_3)\mathbb{1}((x_1, x_2, x_3) \in [0, 1]^3).$$

   a) Find the normalized density.
   b) Find the marginal density of X_1.
   c) Find the expected value of X_1.

23.2 Suppose (X_1, X_2, X_3) has joint density

$$f(x_1, x_2, x_3) = (x_1 + x_2 + x_3)\mathbb{1}(x_1, x_2, x_3 \in [0, 1]).$$

   a) Find the marginal density of X_1.
   b) Find $\mathrm{Cov}(X_1, X_3)$.

23.3 Suppose $X_1, \ldots, X_n$ have joint density

$$(1/10)^n \mathbb{1}((x_1, \ldots, x_n) \in [0, 10]^n).$$

   Show that the X_i are independent.

23.4 Suppose Z_1, Z_2, Z_3 are iid standard normal random variables. Find their joint density.

23.5 Suppose that X and Y have joint density:

$$f_{(X,Y)}(x, y) = (3/4)xy^2 \mathbb{1}(x \in [0, 1])\mathbb{1}(y \in [0, 2])$$

Show that X and Y are independent.

23.6 Suppose that X and Y have joint density

$$f_{(X,Y)}(x, y) = (2/3)(x + 2y)\mathbb{1}(x \in [0, 1], y \in [0, 1]).$$

a) Find $\mathbb{P}(X \leq 0.5)$.

b) Find $\mathbb{P}(Y \leq 0.5)$.

c) Find $\mathbb{P}(X \leq 0.5, Y \leq 0.5)$.

d) Prove that X and Y are not independent.

Chapter 24

Bayes' Rule for densities

Question of the Day A drug lowers cholesterol by 20 or more points with unknown probability p. A statistician models $p \sim \mathsf{Unif}([0, 1])$ and individuals as independent Bernoulli random variables. In a study of 17 individuals, the drug was effective in 4 of them. Conditioned on this information, what is the new distribution of p?

Summary
Suppose X_1 and X_2 have joint density $f_{X_1, X_2}(x_1, x_2)$. Then Bayes' Rule for densities is
$$f_{[X_1 | X_2 = x_2]}(x_1) = \frac{f_{[X_2 | X_1 = x_1]}(x_2) f_{X_1}(x_1)}{f_{X_2}(x_2)}.$$

Recall that Bayes' Rule for events of positive probability allows us to turn around conditioning. If we know the distribution of X given Y, then Bayes' Rule allows us to determine Y given X. The rule is (for $\mathbb{P}(Y \in B)$ nonnegative),

$$\mathbb{P}(X \in A | Y \in B) = \frac{\mathbb{P}(Y \in B | X \in A) \mathbb{P}(X \in A)}{\mathbb{P}(Y \in B)}.$$

When dealing with events such as $\{X = s\}$, for continuous functions this will be 0. So instead we look at events $\{X \in ds\}$, the event that X is in an infinitesitmally small interval around s. When we are conditioning, $\{X \in ds\}$ and $\{X = s\}$ are the same: knowledge that we are arbitrarily close to s and actually s gives us the same information. But when we are trying to find $\mathbb{P}(X \in ds)$, we know that this is $f_X(s) \, ds$, and so infinitesimally small, but still nonzero.

Plugging into Bayes' Rule then gives:

$$
\begin{aligned}
f_{X|Y=y}(x)\,dx &= \mathbb{P}(X \in dx | Y = y) \\
&= \mathbb{P}(X \in dx | Y \in dy) \\
&= \frac{\mathbb{P}(X \in dx, Y \in dy)}{\mathbb{P}(Y \in dy)} \\
&= \frac{\mathbb{P}(Y \in dy | X \in dx)\mathbb{P}(X \in dx)}{\mathbb{P}(Y \in dy)} \\
&= \frac{f_{Y|X=x}(y)\,dy\, f_X(x)\,dx}{f_Y(y)\,dy} \\
&= \frac{f_{Y|X=x}(y)f_X(x)\,dx}{f_Y(y)}
\end{aligned}
$$

The above is not a mathematical proof (you cannnot just cancel dy terms without justificiation in a formal proof), but it can be made precise, and the following result does hold for densities.

> **Theorem 7** (Bayes' Rule for densities)
> Suppose X has density f_X and Y has density f_Y (not necessarily with respect to the same measure.) Then for y such that $f_Y(y) > 0$,
>
> $$f_{X|Y=y}(x) = \frac{f_{Y|X=x}(y)f_X(x)}{f_Y(y)}.$$

Some remarks!

- The theorem states the results hold even if the densities are not with respect to the same measures. In particular, one of the random variables could be discrete and the other continuous, and the result would still hold.

- If you do not know $f_Y(y)$, this result states

$$f_{X|Y=y}(x) \propto f_{Y|X=x}(y)f_X(x).$$

  Remember that $\propto$ means *proportional to* here, and means that there is some factor that does not depend on x (it might depend on y though) multiplying the right hand side to make inequality. How to find that factor? Remember the left hand side is a density, so if we integrate both sides with respect to x, that should equal 1. That allows us to find the constant.

- Statisticians call the initial distribution of X before learning the value Y the *prior*. The density of Y given X is called the *likelihood*, and the distribution of X after learning the value of Y is the *posterior*. So Bayes' Rule for densities can be written with less formal notation as

  posterior density $\propto$ prior density $\cdot$ likelihood.

Question of the day.   Now let us illustrate these ideas with the qotd. Initially, $p \sim \mathsf{Unif}([0,1])$. This is the prior for p. This means $f_p(t) = \mathbb{1}(t \in [0,1])$.

Next, let N denote the number of individuals for which the drug worked. Then we know that since the trials were work/not work independently, that $[N|p] \sim \mathsf{Bin}(17, p)$. That means $[N|p]$ has density

$$f_{N|p=t}(i) = \binom{17}{i} t^i (1-t)^{17-i} \mathbb{1}(\{i \in \{0, \ldots, 17\}).$$

Therefore,

$$
\begin{aligned}
f_{p|N=i}(t) &\propto f_p(t) f_{N|p=t}(i) \\
&= \mathbb{1}(t \in [0,1]) \binom{17}{i} t^i (1-t)^{17-i} \mathbb{1}(\{i \in \{0, \ldots, 17\}) \\
&\propto \mathbb{1}(t \in [0,1]) t^i (1-t)^{17-i}.
\end{aligned}
$$

Since the binomial coefficient 17 choose i and $\mathbb{1}(i \in \{0, \ldots, 17\})$ do not depend on t, they get absorbed into the constant of proportionality.

Now, to find the constant of proportionality, we integrate the result with respect to t using the data that $N = 4$:

$$1 = C \int_{t \in \mathbb{R}} t^4 (1-t)^{17-4} \mathbb{1}(t \in [0,1]) \, dt = C \int_{t \in [0,1]} t^4 (1-t)^{17-4} \, dt = \frac{C}{42840}$$

so $C = 42840$.

Hence the final distribution is

$$\boxed{f_{p|N=4}(t) = 42840 t^4 (1-t)^{13} \mathbb{1}(t \in [0,1]).}$$

In fact, we did not need to do the integration if we had recognized that this is a *Beta* distribution with parameters 5 and 14. Therefore,

$$\boxed{[p|N = 4] \sim \mathsf{Beta}(5, 14).}$$

would also be an acceptable answer.

Note that $\mathsf{Unif}([0,1]) = \mathsf{Beta}(1, 1)$. Hence the prior is a Beta distribution and the posterior is a Beta distribution.

> **Definition 66**
> If the prior and posterior for a Bayesian analysis belong to the same family of distributions, call them **conjugate**.

Suppose the prior distribution is

$$p \sim \mathsf{Beta}(a, b),$$

and the data given p is $N \sim \mathsf{Bin}(n, p)$. Then the posterior distribution is

$$[p|N = i] \sim \mathsf{Beta}(a + i, b + (n - i)).$$

Therefore, we say that the Beta family is conjugate with a binomial likelihood.

There are a couple dozen families of distributions and likelihoods that are conjugate. When working with these particular families, calculations become very easy.

Example 59

Suppose that $Y \sim \text{Exp}(100)$, and $[X|Y] \sim \text{Exp}(Y)$. Given $X = 42$, what is the new distribution of Y?

Answer   Here $f_Y(s) = 100 \exp(-100s)\mathbb{1}(s \geq 0)$, $f_{X|Y=s}(t) = s\exp(-st)\mathbb{1}(t \geq 0)$, and so

$$f_{Y|X=t}(s) \propto 100 \exp(-100s)\mathbb{1}(s \geq 0)s\exp(-st)\mathbb{1}(t \geq 0)$$
$$\propto s\exp(-(100+t)s)\mathbb{1}(s \geq 0).$$

Integrating the right hand side for $s \in \mathbb{R}$ gives

$$\int_{s \geq 0} s\exp(-(100+t)s) \, ds = \frac{1}{(100+t)^2}.$$

So

$$f_{Y|X=42} = 142^2 s\exp(-142s)\mathbb{1}(s \geq 0).$$

That is, $\boxed{[Y|X = 42] \sim \text{Gamma}(2, 142)}$.

Problems

24.1 Suppose $A \sim \text{Unif}\{1, 2, 3, 4, 5, 6\}$ and $[B|A] \sim \text{Exp}(A)$. Given $B = 3.6$, what is the distribution of A with the normalizing constant given to four sig figs?

24.2 Suppose $X \sim \text{Unif}(\{1, 2, 3, 4\})$ and $[Y|X] \sim \text{Unif}([0, X])$. Given $Y = 2.4$, what is the distribution of X?

24.3 Suppose $X_1 \sim \text{Unif}([0, 10])$ and $X_2 \sim \text{Unif}([0, 20])$. Let $B \sim \text{Unif}\{1, 2\}$.

    a) Given $X_B = 15$, what is the chance that $B = 2$?

    b) Given $X_B = 7$, what is the chance that $B = 2$?

24.4 Suppose $Y_1 \sim \text{Exp}(1)$ and $Y_2 \sim \text{Exp}(2)$, and $B \sim \text{Unif}(\{1, 2\})$. Find

$$\mathbb{P}(B = 1|Y_B = 4.3).$$

24.5 A drug company believes that a new treatment is effective on patients with probability p, where p is uniform over $[0, 1]$. A drug trial keeps trying the drug on patients until it finds four patients where the drug is effective. The study needed to enroll $N = 21$ patients before they found four that the drug worked on.

Given this information, what is the new distribution of p?

24.6 Continuing the last problem, the drug company continues testing patients until two more are found where the drug is effective. In this second trial, 8 more patients were seen to find two where the drug was effective. Building on the information from the first trial, what is the new distribution of p?

Tail inequalities: Markov and Chebyshev

Question of the Day Suppose $\mathbb{E}[|X|] = 5$. Bound $\mathbb{P}(|X| \geq 10)$.

Summary

Markov's inequality states that for an integrable random variable X and $a > 0$,

$$\mathbb{P}(|X| \geq a) \leq \mathbb{E}[|X|]/a.$$

The Chebyshev inequality states that for a random variable X with finite variance and $a > 0$,

$$\mathbb{P}(|X - \mu| \geq a) \leq \mathbb{V}(X)/a^2.$$

Probability distributions with densities $f(a)$ must have the density go to zero as a becomes very large or very small. But what about the area under the density? *Tail inequalities* are a way of giving an upper bound on this sort of probability.

The first, and simplest, tail inequality is called Markov's inequality. It has limited use in applications, but serves as a building block for creating more powerful tail inequalities more often used in practice.

In essence, what Markov's inequality says is that in order for $\mathbb{E}[|X|]$ to be small, you cannot put too much weight out past a.

Fact 91 (Markov's inequality)

Let X be a random variable with finite mean. Then for all $a > 0$,

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}[|X|]}{a}.$$

Proof. Note that if we multiply $|X|$ by a number that is either 0 or 1, the product will be at most $|X|$. That is

$$|X| \geq |X|\mathbb{1}(|X| \geq a).$$

Whenever $\mathbb{1}(|X| \geq a) = 1$, $|X| \geq a$, so

$$|X| \geq |X|\mathbb{1}(|X| \geq a) \geq a\mathbb{1}(|X| \geq a).$$

Recall that expected value preserves inequalities so

$$\mathbb{E}[|X|] \geq \mathbb{E}[a\mathbb{1}(|X| \geq a)] = a\mathbb{E}[\mathbb{1}(|X| \geq a)] = a\mathbb{P}(|X| \geq a).$$

$\square$

Remark   Markov's inequality is sometime written as:

For any nonnegative random variable X with finite mean and $a > 0$, $\mathbb{P}(X \geq a) \leq \mathbb{E}[X]/a$.

This is equivalent to the other formulation. Form #2 implies form #1 since $|X|$ is a nonnegative random variable. Form #1 implies form #2 since for a nonnegative random variable, $|X| = X$.

Markov's inequality can be used to solve the qotd. Here we are given that $\mathbb{E}[|X|] = 5$, and the goal is to bound $\mathbb{P}(|X| \geq 10)$. Here $a = 10$, and Markov's inequality gives

$$\mathbb{P}(|X| \geq 10) \leq \frac{\mathbb{E}[|X|]}{10} = \frac{5}{10} = \boxed{0.5000}.$$

One of the nice things about Markov's inequality is that it can be applied without knowing the exact values of the parameters describing the distribution.

> **Example 60**
> Suppose $A \sim \text{Exp}(\lambda)$. Upper bound $\mathbb{P}(A \geq 5/\lambda)$ using Markov's inequality.
>
> **Answer**   Since $A \geq 0$, $|A| = A$. Also, $\mathbb{E}[A] = 1/\lambda$, so
>
> $$\mathbb{P}(A \geq 5/\lambda) = \frac{1/\lambda}{5/\lambda} = \frac{1}{5} = \boxed{0.2000}.$$

25.1   Chebyshev's inequality

Markov's inequality was actually first shown by Markov's Ph.D. advisor, Chebyshev. Markov reproved the result as part of his Ph.D. thesis. Chebyshev was interested in a stronger inequality that used not only the first moment of the random variable, but also the second. This inequality gives a way of bounding how far away the random variable is from its expected value.

> **Fact 92** (Chebyshev's inequality)
> Suppose that X has finite first and second moments. Then for all $a > 0$,
>
> $$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathbb{V}(X)}{a^2}.$$

Proof. Let $Y = (X - \mathbb{E}[X])^2$. Then $|Y| = Y$, so Markov's inequality says for all $a > 0$,

$$\mathbb{P}(Y \geq a^2) \leq \frac{\mathbb{E}[Y]}{a^2}.$$

But $\mathbb{E}[Y] = \mathbb{V}(X)$ by definition, and $\{Y \geq a^2\} \Leftrightarrow \{|X - \mathbb{E}[X]| \geq a\}$. Hence the inequality is shown. $\square$

Example 61
Suppose $A \sim \text{Exp}(\lambda)$. Upper bound $\mathbb{P}(A \geq 5/\lambda)$ using Chebyshev's inequality

Answer Since $A \geq 0$, $|A| = A$. Also, $\mathbb{E}[A] = 1/\lambda$, and $\mathbb{V}(A) = 1/\lambda^2$, so

$$
\begin{aligned}
\mathbb{P}(A \geq 5/\lambda) &= \mathbb{P}(A - 1/\lambda \geq 4/\lambda) \\
&\leq \mathbb{P}(|A - 1/\lambda| \geq 4/\lambda) \\
&\leq \frac{\mathbb{V}(A)}{(4/\lambda)^2} \\
&= \frac{1/\lambda^2}{16/\lambda^2} \\
&= \frac{1}{16} = \boxed{0.06250}.
\end{aligned}
$$

In this case, Chebyshev's inequality gave a better result than Markov's inequality, but that is not always the case. Because both inequalities are true, you can always find both bounds and take the smaller of the two results.

If we set $a = k\,\text{SD}(X)$, then $\mathbb{V}(X)/a^2 = 1/k^2$ and we obtain an alternate form of Chebyshev.

Fact 93 (Chebyshev's inequality (alternate form))
Suppose that X has finite mean and standard deviation. Then for all $k > 0$,

$$
\mathbb{P}(|X - \mathbb{E}[X]| \geq k\,\text{SD}(X)) \leq \frac{1}{k^2}.
$$

Another way to say this is, the chance that a random variable is at least k standard deviations from its mean goes down at least quadratically in k.

Sample averages Let $X_1, X_2, \ldots \sim X$ be iid and consider the sample average

$$
S_n = \frac{X_1 + \cdots + X_n}{n}
$$

From linearity of expectation we have

$$
\mathbb{E}[S_n] = \frac{\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]}{n} = \frac{n\mathbb{E}[X]}{n} = \mathbb{E}[X],
$$

so the sample average always has the same mean as the original function.
On the other hand,

$$
\text{SD}(S_n) = \sqrt{\mathbb{V}(S_n)} = \sqrt{\frac{\mathbb{V}(X_1) + \cdots + \mathbb{V}(X_n)}{n^2}} = \sqrt{\frac{n\mathbb{V}(X)}{n^2}} = \frac{\text{SD}(X)}{\sqrt{n}}.
$$

By Chebyshev, that tells us that

$$
\mathbb{P}(|S_n - \mathbb{E}(X)| \geq a) \leq \frac{\mathbb{V}(X)}{a^2} \cdot \frac{1}{n}.
$$

So Chebyshev's inequality gives us that the sample average will always get closer and closer to the mean. However, this convergence is only inversely linear in n. In practice, the sample average converges much more quickly to the true result.

To show exponentially fast convergence for sample averages, we need an even more powerful inequality, and that is Chernoff's inequality in the next section.

Problems

25.1 Suppose X is a random variable with mean 0.4, mean absolute deviation of 1.5, and standard deviation of 2.

   a) Give an upper bound on $\mathbb{P}(|X - 0.4| > 4)$ using Markov's inequality.

   b) Give an upper bound on $\mathbb{P}(|X - 0.4| > 4)$ using Chebyshev's inequality.

   c) Which is better? (Or equivalently, if you were asked to give the best upper bound on $\mathbb{P}(|X - 0.4| > 4)$, what would you report?)

25.2 Suppose Y has mean 2.3, mean absolute deviation 1.1, and standard deviation 1.8. Bound $\mathbb{P}(|Y - \mathbb{E}[Y]| > 3)$ as best you can using the Markov and Chebyshev inequalities.

25.3 A construction project will take an unknown amount of time. The builders believe that the mean will be fifty days with a standard deviation of ten days.

   a) Give an upper bound for the chance the project takes at least sixty days.

   b) Give an upper bound for the chance the project takes at least one hundred days.

25.4 Suppose $X \geq 0$ has $\mathbb{E}[X] = \mu$ and $\mathbb{V}(X) = 1.3\mu$.

   a) Bound $\mathbb{P}(X \geq 5\mu)$ using Markov's inequality.

   b) Bound $\mathbb{P}(X \geq 5\mu)$ using Chebyshev's inequality.

25.5 Outreach Solutions serves a number of clients each day that is uniform over $\{1, 2, 3, 4, 5\}$. Let N be the total number of clients served in a week of seven days.

   a) What is the expected value of N?

   b) What is the standard deviation of N?

   c) Using the fact that N is symmetric about its mean, give a lower bound on the probability that $N \leq 26$.

25.6 A manufacturing plant ships 100 boxes per day, each of which contains 300 items. If each individual item is defective with probability 0.01 independently of the others, upper bound the probability that more than 600 items are defective.

25.7 Suppose X has finite mean μ and standard deviation σ. All random variables have at least one median. Show that there must be a median of X somewhere strictly between $\mu - \sqrt{3}\sigma$ and $\mu + \sqrt{3}\sigma$.

25.8 Bound the probability that a random variable X lies more than 2.5 standard deviations away from its mean.

25.9 A construction project time T has the following mean, standard deviation, mean absolute deviation, and moment generating function at 0.5:

$$\mathbb{E}[T] = 100$$
$$\sqrt{\mathbb{E}[(T - \mathbb{E}[T])^2]} = 15$$
$$\mathbb{E}(|T - \mathbb{E}[T]|) = 12$$
$$\mathbb{E}(\exp(0.5T)) = \exp(63).$$

Using these facts together with Markov and Chebyshev, put as best an upper bound as you can on $\mathbb{P}(T > 130)$. Be sure to show all your work!

25.10 Suppose X is a nonnegative random variable with $\mathbb{E}[X^3] = 30$. Use this to bound the probability that $X > 10$.

Tail inequalities: Chernoff

Question of the Day Suppose $X_1, X_2, \ldots, X_{20} \sim X$ are iid where

$$\mathbb{P}(X = -0.5) = \mathbb{P}(X = 0.7) = 1/2.$$

Bound the probability that $X_1 + \cdots + X_{20} \geq 10$.

Summary
The Chernoff inequality states that for a random variable X with finite moment generating function $\mathrm{mgf}_X(t)$ that

$$(\forall t \geq 0)(\mathbb{P}(X \geq a) \leq \mathrm{mgf}_X(t) \exp(-ta),$$

and

$$(\forall t \leq 0)(\mathbb{P}(X \leq a) \leq \mathrm{mgf}_X(t) \exp(-ta),$$

provided the mgf_X exists for those values of t.

For Markov's inequality we used the expected value, for Chebyshev's inequality we used the variance. For *Chernoff's inequality* we will use the moment generating function. This will allow us to get inequalities for sums and sample averages that go down exponentially fast in the number of draws.

First, the inequality.

Fact 94 (Chernoff's inequality)
For a random variable and any $t > 0$ where $\mathrm{mgf}_X(t)$ exists,

$$\mathbb{P}(X \geq a) \leq \mathrm{mgf}_X(t) \exp(-ta).$$

For any $t < 0$ where $\mathrm{mgf}_X(t)$ exists,

$$\mathbb{P}(X \leq a) \leq \mathrm{mgf}_X(t) \exp(-ta).$$

Proof. Note that if $t \geq 0$

$$\mathbb{P}(X \geq a) = \mathbb{P}(tX \geq ta) = \mathbb{P}(\exp(tX) \geq \exp(ta)),$$

which is at most $\mathbb{E}[\exp(tX)]/\exp(ta)$ by Markov's inequality.

The other inequality is shown in a similar fashion.   □

Let's apply this inequality to the qotd. Recall that because the X_i are iid distributed as X,

$$\text{mgf}_{X_1+\cdots+X_{20}}(t) = \text{mgf}_{X_1}(t)\,\text{mgf}_{X_2}(t)\cdots\text{mgf}_{X_{20}}(t) = \text{mgf}_X(t)^{20}.$$

Hence

$$\mathbb{P}(X_1 + \cdots + X_{20} \geq 10) = \text{mgf}_X(t)^{20}\exp(-10t) = [\text{mgf}_X(t)\exp(-0.5t)]^{20}.$$

We took the exponent of 20 out of the expression to emphasize the probability bound given by Chernoff decreases exponentially as the number of random variables in the sum increases.

Now

$$\text{mgf}_X(t) = (1/2)\exp(-0.5t) + (1/2)\exp(0.7t),$$

so

$$g(t) = \text{mgf}_X(t)\exp(-0.5t) = [(1/2)\exp(-t) + (1/2)\exp(0.2t)].$$

When $t = 0$ this right hand side is 1, but as t grows, it dips slightly smaller before rising again.
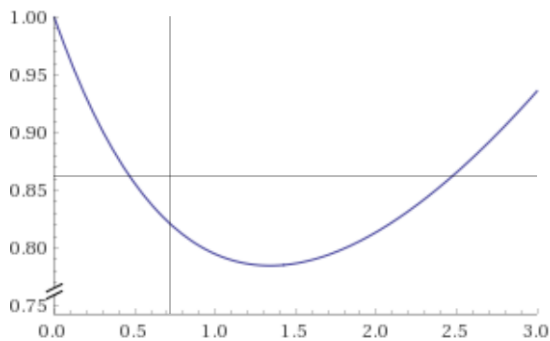
The derivative is

$$g'(t) = -(1/2)\exp(-t) + (0.2)(1/2)\exp(0.2t).$$

Note $g'(0) = (1/2)(-1) + (1/2)(0.2) < 0$ and $g'(3) > 0.9359 > 0$. Next find any critical points:

$$g'(t) = 0$$
$$-(1/2)\exp(-t_1) + (1/2)(0.2)\exp(0.2t) = 0$$
$$1/0.2 = \exp(1.2t)$$
$$t = \ln(5)/1.2.$$

So there is a unique value $t_1 = \ln(5)/1.2$ such that $g'(t_1) = 0$, and $g'(t)$ is continuous, therefore $g(t)$ has a global minimum value of $g(t_1)$. To find t_1:

Putting that back into g gives $\exp(t) = 5^{1/2}$, so

$$g(t_1) = (1/2)[5^{-1/1.2} + 5^{0.2/1.2}],$$

and raising to the 20th power gives

$$\mathbb{P}(X_1 + \cdots + X_{20} \geq 10) \leq \boxed{0.007815}$$

Given to more digits, the answer is $0.00781493\ldots$ We rounded up the last digit in our truncation because we are giving an upper bound.

26.1 Chernoff applied to Binomials

Now let's tackle a trickier problem: applying Chernoff bounds to a general distribution. In particular, let's consider the binomial distribution. Say $B \sim \text{Bin}(n, p)$, then we can view this as

$$B = B_1 + \cdots + B_n,$$

where each B_i is iid $\text{Bern}(p)$. Each B_i has moment generating function

$$p \exp(t) + (1 - p) \exp(0t) = p \exp(t) + 1 - p = 1 + p(\exp(t) - 1).$$

A useful fact is that since the exponential function is convex, it lies above any tangent line. The tangent line to $\exp(x)$ at 1 is $1 + x$. This means that $1 + x \leq \exp(x)$ for any real x, which gives

$$\text{mgf}_{B_i}(t) \leq \exp(p \exp(t) - 1).$$

On average, the binomial will be np, but often it will be larger. Set $\epsilon > 0$. Then Chernoff's bound says that

$$\begin{aligned}
\mathbb{P}(B > (1 + \epsilon)np) &\leq \text{mgf}_B(t) \exp(-t(1 + \epsilon)np). \\
&\leq [\exp(p \exp(t) - 1) \exp(-t(1 + \epsilon)p)]^n \\
&= \exp(p \exp(t) - p - t(1 + \epsilon)p)^n = \exp(g(t))^n
\end{aligned}$$

To make the right hand side as small as possible, make $g(t)$ as small as possible.

Differentiating gives

$$g'(t) = p \exp(t) - (1 + \epsilon)p,$$

which is increasing in t, making for a unique global minimum value at the critical point where $\exp(t) = 1 + \epsilon$. Plugging back in to $g(t)$ gives

$$\begin{aligned}
\mathbb{P}(B > (1 + \epsilon)np) &\leq \exp(p(1 + \epsilon) - p - \ln(1 + \epsilon)(1 + \epsilon)p)^n \\
&= \left(\frac{\exp(p\epsilon)}{(1 + \epsilon)^{(1+\epsilon)p}} \right)^n \\
&= \left(\frac{\exp(\epsilon)}{(1 + \epsilon)^{(1+\epsilon)}} \right)^{np}.
\end{aligned}$$

That's a little difficult to parse, it's easier if we write out the Taylor series of what's inside the parenthesis:

$$\begin{aligned}
\mathbb{P}(B > (1 + \epsilon)np) &\leq \exp(p(1 + \epsilon) - p - \ln(1 + \epsilon)(1 + \epsilon)p)^n \\
&= \left(1 - \frac{\epsilon^2}{2} + \frac{\epsilon^3}{6} + \frac{\epsilon^4}{24} - \cdots \right)^n.
\end{aligned}$$

The first two terms indicate that $n \approx 2\epsilon^{-2} \ln(1/\delta)$ will be required for a binomial to make

$$\mathbb{P}(B > (1 + \epsilon)np) \leq \delta.$$

A similar result holds for the lower bound.

Fact 95
For $B \sim \mathsf{Bin}(n, p)$ and $\epsilon > 0$,

$$\mathbb{P}(B \geq (1 + \epsilon)np) \leq \left(\frac{\exp(\epsilon)}{(1 + \epsilon)^{(1+\epsilon)}} \right)^{np}$$

$$\mathbb{P}(B \leq (1 - \epsilon)np) \leq \left(\frac{\exp(-\epsilon)}{(1 - \epsilon)^{(1-\epsilon)}} \right)^{np}.$$

Problems

26.1 Suppose that X has moment generating function $\mathrm{mgf}_X(t) = [(\exp(t) - 1)/t]^{10}$. Bound $\mathbb{P}(X \geq 8)$ with Chernoff using $t = 5$.

26.2 Suppose $U_1, \ldots, U_{20} \sim \mathsf{Unif}([0, 1])$.

    a) For $S = U_1 + \cdots + U_{20}$, find $\mathrm{mgf}_S(t)$.

    b) Use Chernoff to bound $\mathbb{P}(S \geq 13)$.

26.3 Use Chernoff's inequality to give the best upper bound you can on the probability that the sum of 12 iid random variables uniform over $[0, 1]$ is at least 9.

26.4 Markov Auditing Company completes either 0, 1, or 2 audits in a day independently each day with respective probabilities 20%, 40% and 40%. Let X_i denote the number of audits on day i.

    a) Find the mean and standard deviation of X_i.

    b) Approximate the probability with the CLT that on the first 25 days, at least 36 audits are completed.

    c) Use Chebyshev's Inequality to upper bound the probability that on the first 25 days, at least 36 audits are completed.

    d) Use Chernoff's bound with $t = 0.5$ to upper bound the same probability.

Heavy and light tailed distributions

Question of the Day How can I model data with no variance? with no mean?

Summary A random variable X is said to have a **heavy tailed** distribution if there is some i such that $\mathbb{E}[|X|^i] = \infty$. If $i = 2$ then X does not have a standard deviation, and if $i = 1$ then X does not have a mean.

The standard **Cauchy** distribution is a heavy-tailed distribution. This distribution does not have a mean. The density of a standard Cauchy is

$$\frac{2}{\tau} \cdot \frac{1}{1 + s^2}.$$

Another example of a heavy tailed distribution is the **Zeta** (also known as **Zipf**) distribution with parameter s. This distribution has density for $i \in \{1, 2, \ldots\}$ of

$$\frac{1}{\zeta(s)} \cdot \frac{1}{i^s},$$

where $\zeta(s)$ is the Riemann Zeta function $\zeta(s) = \sum_{i=1}^{\infty} 1/i^s$.

27.1   Light tailed distributions

Normal random variables are nice. The density $\tau^{-1/2} \exp(-x^2/2)$ goes down very, very fast as x gets large, which is why the moment generating function $\mathbb{E}[\exp(tZ)]$ is finite for any t. Note that we are exponentiating Z here, which you would think would on average make it pretty big. But Z is so unlikely to be far away from 0 that this moment generating function is defined for all t.

Exponential random variables are nice, but not quite as nice as normals. They have density $\lambda \exp(-\lambda s)\mathbb{1}(s \geq 0)$, and so they only have moment generating function that is finite for $t \in (-\infty, \lambda]$. For $A \sim \mathsf{Exp}(\lambda)$ with $\lambda > 0$, that means the moment generating function can be used to show $\mathbb{E}[A^k] < \infty$ for all integer k. That is, all moments of the random variable are finite. This motivates the definition of a *light tailed* distribution.

Definition 67
A random variable A is **light tailed** if for all $k \in \{1, 2, \ldots\}$,

$$\mathbb{E}[|A|^k] < \infty.$$

27.2   Heavy tailed distributions

The Central Limit Theorem is one of the most powerful theorems in mathematics, but it definitely does not mean "all random variables are normal". The CLT only applies when random variables are being added, it does not apply when random variables are being multiplied. This case happens in situations such as income distribution, population distribution, word usage, and many other contexts.

In these types of data, often *heavy tailed distributions* are seen. These do not have a moment generating function that is finite anywhere except $t = 0$. For these random variables, there exists some k such that $\mathbb{E}[|X|^k] = \infty$.

> **Definition 68**
> A random variable X is **heavy tailed** if there exists some k such that $\mathbb{E}[|X|^k] = \infty$.

The first random variable we saw with a heavy tail was the Cauchy distribution with density

$$\frac{2}{\tau} \cdot \frac{1}{1+s^2}$$

which makes for $X \sim$ Cauchy,

$$\begin{aligned}
{}^{\backprime}\mathbb{E}[|X|] &= \int_{x \in \mathbb{R}} \frac{2}{\tau} \cdot \frac{|s|}{1+s^2} \, ds \\
&= \int_{x \geq 0} \frac{4}{\tau} \cdot \frac{s}{1+s^2} \, ds \\
&= \frac{2}{\tau} \cdot \ln(1+s^2)|_0^\infty = \infty.
\end{aligned}$$

27.3   The Zeta distribution

Another heavy tailed distribution with more flexibility in the heaviness of the tail is the *Zeta* distribution.

> **Definition 69**
> Say that random variable $X \in \{1, 2, 3, \ldots\}$ has the **Zeta** or **Zipf** distribution with parameter $\alpha > 1$, if the density is
>
> $$f_X(i) = \frac{1}{\zeta(\alpha)} \cdot \frac{1}{i^\alpha},$$
>
> where
>
> $$\zeta(\alpha) = \sum_{i=1}^\infty \frac{1}{i^\alpha}$$
>
> is the Riemann Zeta function.

Remarks

- The parameter α must be greater than 1, since the Harmonic series

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots$$

diverges, while $\zeta(\alpha)$ is finite for $\alpha > 1$ by the Integral test.

- When $\alpha \leq 2$, the Zeta distribution has no mean. When $\alpha \leq 3$, the Zeta distribution has no variance.

Problems

27.1 For $X \sim$ Cauchy, find
$$\mathbb{P}(X \in [0,5]).$$

27.2 For $X \sim$ Cauchy, find
$$\mathbb{P}(3X + 5 \in [0,10]).$$

27.3 Estimate $\zeta(2.5)$ to four significant figures.

27.4 Suppose $X \sim$ Zeta(1.5). Find $\mathbb{P}(X \in [1,10])$ by summing a large enough number of terms.

27.5 For $X \sim$ Zeta(α) with $\alpha > 1$, prove that $\ln(X)$ always has finite mean.

27.6 For $X \sim$ Zeta(α), for what values of α is $\mathbb{E}[X^3]$ finite?

Chapter 28

Uniform and Bernoulli marginal distributions

Question of the Day A preserve has 53 animals, 24 of which are male and 29 of which are female. Five of the animals are chosen uniformly at random without replacement to be tagged. What is the average number tagged that are male?

nt>

Chapter 28

Uniform and Bernoulli marginal distributions

Question of the Day A preserve has 53 animals, 24 of which are male and 29 of which are female. Five of the animals are chosen uniformly at random without replacement to be tagged. What is the average number tagged that are male?

Summary
Consider drawing a subset of k objects uniformly at random without replacement from a set of n objects. Out of the n objects, m are marked in some way. Let X denote the number of marked objects chosen. Then X has a **hypergeometric** distribution with parameters n, m, and k. For $i \in \{0, \ldots, \min(k, m)\}$:

$$\mathbb{P}(X = i) = \frac{\binom{m}{i}\binom{n-m}{k-i}}{\binom{n}{k}}.$$

28.1 Drawing without replacement

Consider a small example. Suppose that four out of 12 items in a carton are defective. Two of the items are selected uniformly at random. What is the chance that exactly 1 of the items drawn out are defective?

This problem is an example of *sampling without replacement* and arises occasionally in statistical sampling from small populations. This type of problem is rare in practice, as small populations can usually be tested completely, and for large populations the difference in probabilities between sampling without replacement and with replacement become very small very quickly.

That being said, this type of problem does come up once in a while, so it is helpful to see how to tackle it.

For the example above, the numbers are small enough that we can deal with it directly. If I draw two items, the possible outcomes are DD, DN, ND, or NN, where N means not defective and D means defective.

We can calculate each of these probabilities using conditioning. For instance,

$$\mathbb{P}(DD) = \frac{4}{12} \cdot \frac{3}{11}.$$

175

The first 4 comes from the 4 defectives, but if the first draw is a D, that leaves only 3 defectives, giving the 3 in the numerator of the second fraction.

For DN, the calculation is

$$\mathbb{P}(DN) = \frac{4}{12} \cdot \frac{8}{11}$$

and for ND, the calculation is

$$\mathbb{P}(ND) = \frac{8}{12} \cdot \frac{4}{11}.$$

Note that the product of the denominators $(12 \cdot 11)$ is the same as for DD, only the numerator changes. Also note that for ND and DN, the numerator product is $8 \cdot 4$ and $4 \cdot 8$, which are the same number because multiplication does not care what order you multiply the numbers. More generally, for a sequence NNNDDNN, the product of the numerator is completely determined by the number of D's and N's, and not by their order.

Finally, $\mathbb{P}(\text{NN}) = (8/12)(7/11)$. Therefore, if X is the number of defective items in the sample of two,

$$\mathbb{P}(X = 0) = \frac{12}{132}, \ \mathbb{P}(X = 1) = \frac{64}{132}, \ \mathbb{P}(X = 2) = \frac{56}{132}$$

The probabilities have been left as fractions here to show that they add up to 1.

In general, we say that X has a **hypergeometric** distribution.

28.2   Theory

To find the average number of defectives in the draws (like in the question of the day), it helps to have a more systematic approach to hypergeometric random variables.

Suppose that

$$(U_1, \ldots, U_n) \sim \mathsf{Unif}(A^n).$$

One of our earliest results is that the U_i are independent random variables, each with marginal distribution $U_i \sim \mathsf{Unif}(A)$. However, it is possible to draw the U_i such that each is marginally uniform, but they are dependent random variables.

Consider a vector $(a_1, \ldots, a_n)$. A *permutation* is a reordering of the elements of the vector. For instance, $(3, 4, 1, 2)$ is a permutation of $(1, 2, 3, 4)$.

For a set of size n, there are $n!$ permutations. Let $\mathcal{S}_n$ denote the set of permutation vectors of $(1, 2, \ldots, n)$. Then suppose we draw uniformly from the set of distributions.

> **Definition 70**
> Say that
> $$(X_1, \ldots, X_n) \sim \mathsf{Unif}(\mathcal{S}_n)$$
> is a **draw without replacement** of the elements $\{1, \ldots, n\}$.

Each marginal distribution will be uniform.

> **Fact 96**
> Let $(X_1, \ldots, X_n) \sim \mathsf{Unif}(\mathcal{S}_n)$. Then for all $i \in \{1, \ldots, n\}$, $X_i \sim \mathsf{Unif}(\{1, \ldots, n\})$.

Recall that X_{-i} denotes the vector of values with X_i removed, so for instance if $(X_1, X_2, X_3, X_4) = (4, 2, 1, 3)$, $X_{-3} = (4, 2, 3)$.

Proof. Fix $i \in \{1, \ldots, n\}$ and let $j \in \{1, \ldots, n\}$ as well. Then

$$\mathbb{P}(X_i = j) = \frac{\#(\{x \in \mathcal{S}_n | x(i) = j\})}{n!} = \frac{(n-1)(n-2)\cdots(1)}{n!} = \frac{1}{n},$$

so $X_i \sim \mathsf{Unif}(\{1, \ldots, n\})$. □

Of course, the X_i are not independent! For instance, for $i \neq j$,

$$\mathbb{P}(X_i = X_j = 1) = 0,$$

while

$$\mathbb{P}(X_i = 1)\mathbb{P}(X_j = 1) = \frac{1}{n} \cdot \frac{1}{n} = \frac{1}{n^2}.$$

Now suppose that we take the X_i and use them to form indicator random variables. For some fixed $a \in \{1, \ldots, n\}$, let

$$B_i = \mathbb{1}(X_i \leq a).$$

Fact 97
The B_i created in this way have marginal distributions:

$$B_i \sim \mathsf{Bern}(a/n).$$

Proof. Since $X_i \sim \mathsf{Unif}(\{1, \ldots, n\})$, $\mathbb{P}(B_i = 1) = a/n$. □

As the X_i are not independent, the B_i are not independent as well!

Example 62
Suppose $n = 10$ and $a = 6$. What is the chance that $B_1 = B_2 = 1$?

Answer Consider choosing X_1 uniformly from $\{1, 2, \ldots, n\}$. Then $[X_2|X_1] \sim \mathsf{Unif}(\{1, \ldots, n\} \setminus \{X_1\})$. For $B_1 = B_2 = 1$, $X_1 \in \{1, \ldots, a\}$ and $X_2 \in \{1, \ldots, a\} \setminus \{X_1\}$. The chance that both these things happen is

$$\frac{a}{n} \cdot \frac{a-1}{n-1} = \frac{6}{10} \cdot \frac{5}{9} = \frac{3}{9} = \boxed{0.3333\ldots}.$$

Here is another way to state the same type of problems. If I have a group of n objects, a of which are marked in some way, and I draw out k objects without replacement uniform from the set, what is the chance that exactly i of the objects have the special mark?

Example 63
An urn contains fourteen balls, ten are red and four are blue. If two balls are drawn out uniformly without replacement, what is the chance that both are red?

Answer The chance that the first ball is red is $10/14$. Then the chance that the second ball is red given that the first is red (and remembering that we are drawing without replacement) is $(10/14)(9/13) = \boxed{0.4945\ldots}$.

Another way to approach this type of problem is using binomial coefficients. There are 14 choose 2 ways to pick two balls uniformly at random without replacement from the set of fourteen balls. How many subsets are red? Well, there are 10 red balls and so there are 10 choose 2 ways to pick two red balls to be our subset.

Therefore the answer is

$$\frac{\binom{10}{2}}{\binom{14}{2}} = \frac{\frac{10 \cdot 9}{2!}}{\frac{14 \cdot 13}{2!}} = \boxed{0.4945\ldots}.$$

We get the same answer no matter how we approach the problem!

Now suppose we do not want all of the balls to be the same color.

Example 64

An urn contains fourteen balls, ten are red and four are blue. Six are drawn out uniformly at random without replacement. Let X denote the number that are red. What is $\mathbb{P}(X = 3)$?

Answer   There are 14 choose 6 ways to draw six balls uniformly at random without replacement from the set of fourteen balls. For instance, if the balls are numbered $\{1, \ldots, 14\}$, then the red balls are $\{1, \ldots, 10\}$ and $\{11, \ldots, 14\}$ are blue. Then a subset with exactly three red balls is

$$\{2, 7, 9, 11, 12, 13\}.$$

How many such subsets are there? Well, the first three entries have to be red, and there are 10 choose 3 ways to pick the red balls. The last three entries have to be blue, and there are 4 choose 3 ways to pick the blue balls. Hence the total number of ways is $\binom{10}{3} \cdot \binom{4}{3}$.

That means the overall probability is

$$\frac{\binom{10}{3}\binom{4}{3}}{\binom{14}{6}} = \frac{160}{1001} = \boxed{0.1598\ldots}$$

For $B_1, B_2, \ldots, B_n$ iid $\mathsf{Bern}(p)$, we say that

$$B_1 + \cdots + B_k \sim \mathsf{Bin}(k, p).$$

For our $B_1, \ldots, B_n$ coming from the draws without replacement, we say that the sum of the first n of these random variables has a *hypergeometric distribution*.

Definition 71

Suppose $(X_1, \ldots, X_n) \sim \mathsf{Unif}(\mathcal{S}_n)$ and $B_i = \mathbb{1}(X_i \leq a)$. Then say

$$N = B_1 + \cdots + B_k$$

has a **hypergeometric distribution** with parameters n, k, and a.
   Write $X \sim \mathsf{Hypergeo}(n, k, a)$.

Fact 98
The density of $X \sim \text{Hypergeo}(n, k, a)$ is
$$f_X(i) = \frac{\binom{a}{i}\binom{n-a}{k-i}}{\binom{n}{k}} \mathbb{1}\,(i \in \{0, \dots, \min k, a\}).$$

Remark The hypergeometric distribution and the geometric distribution have nothing whatsoever to do with each other!

Since the hypergeometric distribution is the sum of k different Bernoulli random variables, the mean of the hypergeometric equals k times the mean of the Bernoulli random variables.

Fact 99
For $X \sim \text{Hypergeo}(n, k, a)$,
$$\mathbb{E}[X] = \frac{ka}{n}.$$

Example 65
This fact is enough to handle the Question of the Day: with five animals chosen uniformly from 53, where 24 are male, the expected number of male animals is $(5 \cdot 24)/(53) \approx$ $\boxed{2.264}$.

The variance is calculated similarly, although there is more work to be done because we must calculate the covariances between pairs of Bernoulli random varaibles. First, this is the result.

Fact 100
For $X \sim \text{Hypergeo}(n, k, a)$,
$$\mathbb{V}(X) = \frac{km}{n} \cdot \frac{(n-m)(n-k)}{n(n-1)}.$$

Proof. As before $X = B_1 + \cdots + B_k$. Hence
$$\mathbb{V}(X) = \sum_{i=1}^{k} \mathbb{V}(B_i) + 2\sum_{i<j} \text{Cov}(B_i, B_j) = k\mathbb{V}(B_1) + k(k-1)\,\text{Cov}(B_1, B_2).$$

Since the B_i are Bernoulli with parameter a/n, $\mathbb{V}(B_i) = (m/n)(1 - m/n)$ and
$$\text{Cov}(B_i, B_j) = \mathbb{E}[B_i B_j] - \mathbb{E}[B_i]\mathbb{E}[B_j] = \mathbb{P}(B_i = B_j = 1) - (a/n)^2$$
$$= \frac{a}{n} \cdot \frac{a-1}{n-1} - \frac{a}{n} \cdot \frac{a}{n}$$
$$= -\frac{a}{n}\left[\frac{n-a}{n(n-1)}\right].$$

Then

$$\begin{aligned}
\mathbb{V}(X) &= k\frac{a}{n}\left[1 - \frac{a}{n}\right] - k(k-1)\frac{a}{n}\left[\frac{n-a}{n(n-1)}\right] \\
&= \frac{ka}{n}\left[1 - \frac{a}{n} - \frac{n-a}{n}\cdot\frac{k-1}{n-1}\right] \\
&= \frac{ka}{n}\left[\frac{n(n-1) - a(n-1) - (n-a)(k-1)}{n(n-1)}\right] \\
&= \frac{ka}{n}\left[\frac{(n-a)((n-1) - (k-1))}{n(n-1)}\right] \\
&= \frac{ka}{n}\left[\frac{(n-a)(n-k)}{n(n-1)}\right]
\end{aligned}$$

□

Notice that the without replacement makes the variance slightly less than it is in the binomial, with replacement, case.

Problems

28.1 A small plastic bucket contains tiles with the letters MISSISSIPPI. Four of these tiles are drawn out of the bucket without replacement.

    a) What is the chance that all four S tiles are drawn?

    b) What is the chance that exactly two out of the four drawn tiles are S?

28.2 A jar contains five blue and ten green marbles. Seven marbles are drawn from the jar, what is the chance that exactly 3 are blue?

28.3 In the prodction output of a factory for a day are 500 screws, 10 of which are below standard.

    a) If 5 of the screws are drawn uniformly at random without replacement from the 500 screws, what is the chance that this sample does not contain a below standard screw?

    b) What is the expected number of below standard screws in this sample of 5?

28.4 Continuing the last problem, how many screws should be in the sample so that on average one of the sampled screws will be below standard?

28.5 Fifteen hundred of a particular bird are living in an ecosystem. Twenty are sampled without replacement, and five are found to be carrying a particular gene. What number of birds out of the 1500 should have this gene so that the expected number found in the sample would be 5? (This type of estimate is called a *method of moments estimate.*)

28.6 Out of a set of 316 college students, 160 of which are 20 years or older and 156 of which are younger than 20 years, 48 are chosen uniformly at random to complete a survey.

a) What is the expected number of students completing the survey who are 20 years or older?

b) What is the standard deviation of the expected number of students completing the survey who are 20 years or older?

Chapter 29

The Multinomial distribution

Question of the Day Suppose that in a survey residents are asked if they are unsatisfied, satisfied, or highly satisfied with their long distance service. If each resident is independently unsatisfied with probability 20%, satisfied with probability 70%, and highly satisfied with probability 10%, what is the correlation between the number of unsatisfied and highly satisfied participants?

Summary

The **multinomial distribution** arises from trials where there are two or more possible answers. If each of n trials has outcome $\{1, \ldots, k\}$, and the trials are iid, then $(X_1, \ldots, X_k) \sim \mathsf{Multinom}(n, p_1, \ldots, p_k)$, where p_i is the probability any given trial has outcome i.

For each $i \in \{1, \ldots, k\}$, $X_i \sim \mathsf{Bin}(n, p_i)$, and $\mathrm{Cov}(X_i, X_j) = -p_i p_j$.

In forming the binomial distribution, we considered Bernoulli experiments that had one of two outcomes, success or failure, 1 or 0.

What if there are more than two choices? Perhaps there are three choices for each trial. In this case, we could count the number of trials where choice 1 occurred, where choice 2 occurred, and where choice 3 occurred.

As an example, consider a survey of 100 people. The results of the survey are the random variables (X_1, X_2, X_3), where X_1 is the number under 18, X_2 the number 18 to 25, and X_3 the number over 25. Suppose the probability a person surveyed is under 18 is 0.2, 18 to 25 is 0.5, and over 25 is 0.3, Then

$$X_1 \sim \mathsf{Bin}(100, 0.2)$$
$$X_2 \sim \mathsf{Bin}(100, 0.5)$$
$$X_3 \sim \mathsf{Bin}(100, 0.3),$$

with the restriction that $X_1 + X_2 + X_3 = 100$. In general, we have the following.

Definition 72

Say that $(X_1, \ldots, X_k)$ has a **multinomial distribution with parameters** $n, p_1, \ldots, p_k$ (write $(X_1, \ldots, X_k) \sim \mathsf{Multinom}(n, p_1, \ldots, p_k)$) if the p_i are nonnegative parameters with $p_1 + \cdots + p_k = 1$ and for all i, $X_i \sim \mathsf{Bin}(n, p_i)$ and $X_1 + \cdots + X_k = n$.

Recall that if $B_i \sim \mathsf{Bern}(p)$ are independent for $i \in \{1, \ldots, n\}$, then

$$X = B_1 + \cdots + B_n \sim \mathsf{Bin}(n, p).$$

This same idea can be extended to the multinomial case.

> **Fact 101**
> Let W be a discrete random variable such that $\mathbb{P}(W = i) = p_i$ where $\sum_{i=1}^{n} p_i = 1$.
> Then for $W_1, \ldots, W_n \sim W$ iid, and $i \in \{1, 2, \ldots, k\}$
>
> $$X_i = \sum_{j=1}^{n} \mathbb{1}(W_j = i).$$
>
> Then $(X_1, \ldots, X_k) \sim \mathsf{Multinom}(n, p_1, \ldots, p_k)$.

Proof. Since each X_i is the sum of n independent indicator random variables that equal 1 with probability p_i. $X_i \sim \mathsf{Bin}(n, p_i)$. Also,

$$\sum_{i=1}^{k} X_i = \sum_{i=1}^{k} \sum_{j=1}^{n} \mathbb{1}(W_j = i) = \sum_{j=1}^{n} \sum_{i=1}^{k} \mathbb{1}(W_j = i)$$

$$= \sum_{j=1}^{n} \mathbb{1}(W_j = 1) + \mathbb{1}(W_j = 2) + \cdots + \mathbb{1}(W_j = k)$$

$$= \sum_{j=1}^{n} 1 = n.$$

$\square$

In particular, this fact implies that if $X \sim \mathsf{Bin}(n, p)$, then $(X, n - X) \sim \mathsf{Multinom}(n, p, 1 - p)$.

The density of a multinomial

It is possible to write down a density for $(X_1, \ldots, X_n)$ multinomial. This density requires the *multichoose* function.

> **Definition 73**
> Let $a_1, a_2, \ldots, a_k$ be a set of symbols. The number of ways to arrange $i_1\ a_1$ symbols, $i_2\ a_2$ symbols, and so on up to $i_k\ a_k$ symbols in a row is given by the **multichoose** function also known as the **multinomial coefficient**, written as
>
> $$\binom{i_1 + \cdots + i_n}{i_1, i_2, \ldots, i_n}.$$

The formula for calculating the multichoose expression is similar to that of a binomial coefficient.

> **Fact 102**
> For nonnegative integers $i_1, \ldots, i_k$,
>
> $$\binom{i_1 + \cdots + i_n}{i_1, i_2, \ldots, i_n} = \frac{(i_1 + \cdots + i_n)!}{i_1! i_2! \cdots i_n!}$$

Example 66

How many ways are there to arrange 4 a's, 5 b's, and 3 c's in a row?

Answer There are $4 + 5 + 3 = 12$ symbols, so this is

$$\binom{12}{4,5,3} = \frac{12!}{4!5!3!} = 27720$$

Example 67

For $(X_1, X_2, X_3) \sim \mathsf{Multinom}(12, 0.2, 0.5, 0.3)$, what is $\mathbb{P}(X_1 = 4, X_2 = 5, X_3 = 3)$?

Answer Let's say the outcome of an individual trial is a with probability 0.2, b with probability 0.5, and c with probability 0.3. Then one particular outcome with 4 a's, 5 b's, and 3 c's is $caaabccbabbb$. The probability of this outcome is $(0.3)(0.2)(0.2)(0.2(0.5)(0.3)(0.3)(0.5)(0.2)(0.5)(0.5)(0.5) = (0.2)^4(0.5)^5(0.3)^3$. Because multiplication commutes, this will be the probability of any outcome with 4 a's, 5 b's, and 3 c's. So the total probability is

$$\binom{12}{4,5,3}(0.2)^4(0.5)^5(0.3)^3 = \boxed{0.03742\dots}.$$

The last example can be generalized to get the density for arbitrary multinomials.

Fact 103

For $(X_1, \dots, X_k) \sim \mathsf{Multinom}(n, p_1, \dots, p_k)$. Then

$$f_{(X_1, \dots, X_n)}(i_1, \dots, i_n) = \binom{n}{i_1, i_2, \dots, i_k} p_1^{i_1} \cdots p_n^{i_n} \cdot \mathbb{1}(i_1, \dots, i_k \in \{0, \dots, n\}) \cdot$$
$$\mathbb{1}(i_1 + \cdots + i_k = n).$$

29.1 Covariance

The covariance between different components of a multinomial follows directly from the indicator representation. It is negative because when one component is higher, all other components are lower on average.

Fact 104

For $(X_1, \dots, X_k) \sim \mathsf{Multinom}(n, p_1, \dots, p_k)$, for all $i \neq j$, $\mathrm{Cov}(X_i, X_j) = -n p_i p_j$.

Proof. Let $i \neq j$. Then the covariance between X_i and X_j is

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j] = \mathbb{E}[X_i X_j] - (np_i)(np_j),$$

Remember for indicator functions $\mathbb{1}(A)\mathbb{1}(B) = \mathbb{1}(AB)$. Also, $X_i = \sum_k \mathbb{1}(W_k = i)$ and $X_j = \sum_\ell \mathbb{1}(W_\ell = j)$, so

$$X_i X_j = \sum_k \sum_\ell \mathbb{1}(W_k = i, W_\ell = j),$$

which means
$$\mathbb{E}[X_i X_j] = \sum_k \sum_\ell \mathbb{P}(W_k = i, W_\ell = j).$$

It would be nice if each of the probabilities were $p_i p_j$. If that was true, then the double sum would be $n^2 p_i p_j$. Unfortunately, the probabilities only hold this value when $k \neq \ell$. When $k = \ell$, the probability is zero. Since $k = \ell$ occurs for exactly n of the terms in the double sum, the number of terms where $\mathbb{P}(W_k = i, W_\ell = j) = p_i p_j$ is $n^2 - n$, with the rest being zero. Hence
$$\mathbb{E}[X_i X_j] = (n^2 - n)p_i p_j.$$

That makes
$$\mathrm{Cov}(X_i, X_j) = (n^2 - n)p_i p_j - n^2 p_i p_j = -n p_i p_j.$$

$\square$

Problems

29.1 Suppose $(W_1, W_2, W_3, W_4) \sim \mathsf{Multinom}(10, 0.3, 0.2, 0.4, 0.1)$. What is
$$\mathbb{P}((W_1, W_2, W_3, W_4) = (1, 3, 2, 4))?$$

29.2 Fifteen individuals are given a question with five answers. If the probability for each of the five answers is modeled as $(0.2, 0.2, 0.1, 0.25, 0.25)$, then what is the chance that the total counts for the answers comes to $(4, 3, 1, 5, 2)$?

29.3 Suppose $(X_1, X_2, X_3) \sim \mathsf{Multinom}(30, 0.5, 0.1, 0.4)$.

    a) What is the distribution of X_1?

    b) What is $\mathbb{E}[X_1]$?

    c) Find $\mathrm{Cov}(X_1, X_3)$.

29.4 Suppose $(Y_1, Y_2, Y_3, Y_4) \sim \mathsf{Multinom}(100, 0.2, 0.3, 0.4, 0.1)$.

    a) What is the distribution of Y_3?

    b) Find $\mathrm{Cov}(Y_3, Y_4)$.

    c) Find $\mathrm{Cov}(Y_3, Y_3)$.

29.5 A research project is studying an animal population. There are 30 animals collected. Each animal has a 10% chance of being genotype A, 20% of being genotype B, and 70% of being genotype C. Let (N_A, N_B, N_C) be the number of animals found of each genotype.

    a) What is the distribution of (N_A, N_B, N_C)?

    b) What is $\mathbb{E}(N_A)$?

    c) What is $\mathrm{Cov}(N_A, N_B)$?

    d) What is $\mathrm{Cov}(N_A, N_C)$?

29.6 A psychology experiment is studying the effects on positive messages on sleep. Suppose they survey 73 students. The model predicts that 25% will sleep well, 55% sleep moderately well, and 20% will sleep badly. They record the number of students for each of the three responses as (N_1, N_2, N_3). Answer the following.

a) What is the distribution of (N_1, N_2, N_3)?

b) What is $\mathbb{E}(N_3)$?

c) What is $\mathrm{Cov}(N_1, N_3)$?

d) What is $\mathrm{Cor}(N_1, N_3)$?

Multinormal random variables

Question of the Day Suppose

$$A = \begin{pmatrix} 1 & -1 \\ 0 & 3 \end{pmatrix}.$$

Let $Z = (Z_1, Z_2)$, where the Z_i are iid standard normal random variables. For $W = AZ$:

1. What is the distribution of $W = (W_1, W_2)$?

2. Find $\text{Cor}(W_1, W_2)$.

Summary
For $Z_1, \ldots, Z_n$ be iid standard normal random variables, $\mu \in \mathbb{R}^n$, A an n by n matrix, and $W = AZ$. Then W has a **multivariate normal** or **multinormal** distribution. Write $W \sim \text{Multinorm}(\mu, AA^T)$.
Call $\Sigma = AA^T$ the **covariance matrix**, and $\Sigma(i, j) = \text{Cov}(W_i, W_j)$. The density of W is
$$f_W(w) = \tau^{-n/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(w - \mu)^T \Sigma^{-1}(w - \mu) \right)$$

Suppose $Z_1, \ldots, Z_n$ are iid standard normal random variables. Then because the Z_i are independent, knowledge of one does not affect knowledge of the other.

This makes calculation easy, but it is terrible for modeling real data, where often knowledge of one factor changes the distribution of another.

For instance, for two tech stocks, if one is higher than average the other might also be higher. Birth weight of bears might be positively correlated with available nutrition. Time spent watching T.V. might be negatively correlated with crime rates, and so on.

Therefore it is helpful to have a distribution where each component has a marginal distribution that is normal, but we allow for positive or negative correlation between the different components.

To keep things simple, start with two iid standard normal random variables, Z_1 and Z_2. One of our rules for random variables is that the sum of independent normal random variables is also a normal random variables. The parameters (mean and variance) are just the sum of the parameters

for the original random variables. For instance, following those rules gives

$$Z_1 + Z_2 \sim \mathsf{N}(0, 2)$$
$$Z_1 - 2Z_2 \sim \mathsf{N}(0, 5),$$

where $-2Z_2$ has variance $(-2)^2$ and Z_1 has variance 1 giving their sum with variance 5.

We could write these equations using matrices:

$$\begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix}$$

or even more simply

$$AZ = W,$$

where $A(1, 1) = 1$, $A(1, 2) = 1$, $A(2, 1) = 1$, $A(2, 2) = -2$.

Now look at the covariance between W_1 and W_2.

$$\begin{aligned} \mathrm{Cov}(W_1, W_2) &= \mathrm{Cov}(A(1, 1)Z_1 + A(1, 2)Z_2, A(2, 1)Z_1 + A(2, 2)Z_2) \\ &= A(1, 1)A(2, 1)\,\mathrm{Cov}(Z_1, Z_1) + A(1, 1)A(2, 2)\,\mathrm{Cov}(Z_1, Z_2) \\ &\quad + A(1, 2)A(2, 1)\,\mathrm{Cov}(Z_2, Z_1) + A(1, 2)A(2, 2)\,\mathrm{Cov}(Z_2, Z_2) \\ &= A(1, 1)A(2, 1) + A(1, 2)A(2, 2). \end{aligned}$$

The simplification at the end comes from $\mathrm{Cov}(Z_i, Z_i) = \mathbb{V}(Z_i) = 1$ and $\mathrm{Cov}(Z_1, Z_2) = 0$ since Z_1 and Z_2 are independent. This final result has a special form. $A(1, 1)A(2, 1) + A(1, 2)A(2, 2)$ is the dot product of the first row of A and the second column of A.

This idea can be generalized to $Z_1, \dots, Z_n$ iid standard normal, and

$$AZ = W,$$

where A is an n by n matrix and W is then an n dimensional vector. Then a similar calculation as above gives

$$\mathrm{Cov}(W_i, W_j) = r_i \cdot c_j,$$

where r_i is the ith row of the matrix A and c_j is the jth column of the matrix A.

We often combine the covariances into a matrix where the (i, j) entry is $\mathrm{Cov}(W_i, W_j)$. Because standard deviation is usually denoted by a lowercase Greek letter sigma, σ, this covariance matrix is usually denoted with the capital Greek letter sigma. This looks like Σ.

From what we calculated earlier, we know that

$$\Sigma = AA^T.$$

Here A^T represents the *transpose* of the matrix, that flips the rows and columns. For example,

$$B = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \quad B^T = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}.$$

Once we have the covariances, we can add a constant to change the mean of the random variables.

> **Definition 74**
> Let $Z_1, \ldots, Z_n$ be iid standard normal random variables. For A an n by n real matrix and $\mu \in \mathbb{R}^n$, say
> $$W = AZ + \mu$$
> has a **multivariate normal** or **multinormal** distribution with mean μ and covariance $\Sigma = AA^T$. Write
> $$W \sim \text{Multinorm}(\mu, \Sigma).$$

Then the key fact about the multivariate normal is as follows.

> **Fact 105**
> For $W \sim \text{Multinorm}(\mu, \Sigma)$, $\mathbb{E}[W] = \mu$, and for $(i, j) \in \{1, 2, \ldots, n\}^2$, $\text{Cov}(W_i, W_j) = \Sigma(i, j)$.

> **Example 68**
> **Question of the Day.** Suppose
> $$A = \begin{pmatrix} 1 & -1 \\ 0 & 3 \end{pmatrix}.$$
>
> Let $Z = (Z_1, Z_2)$, where the Z_i are iid standard normal random variables. For $W = AZ$:
>
> 1. What is the distribution of $W = (W_1, W_2)$?
>
> 2. Find $\text{Cor}(W_1, W_2)$.
>
> **Answer** Because W is a matrix times a vector of iid standard normal random variables, it will have a multivariate normal distribution. The first parameter (the mean vector) is the zero vector. The second parameter (the covariance matrix) is
> $$\Sigma = \begin{pmatrix} 1 & -1 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 3 \end{pmatrix} = \begin{pmatrix} 2 & -3 \\ -3 & 9 \end{pmatrix}.$$
>
> We can use this information to solve the problems.
>
> 1. $\boxed{W \sim \text{Multinorm}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & -3 \\ -3 & 9 \end{pmatrix} \right)}.$
>
> 2.
> $$\text{Cor}(W_1, W_2) = \frac{\text{Cov}(W_1, W_2)}{\sqrt{\mathbb{V}(W_1)\mathbb{V}(W_2)}}$$
> $$= \frac{-3}{\sqrt{2}\sqrt{9}} = \boxed{-0.7071\ldots}.$$

Note that this means the variances of the individual components lie on the diagonal.

Example 69

Suppose (W_1, W_2, W_3) is multivariate normal with mean $(1.2, -2, 3.4)$ and covariance matrix

$$\begin{pmatrix} 6.97 & -0.64 & -3.52 \\ -0.64 & 6.8 & -1.52 \\ -3.52 & -1.52 & 14.24 \end{pmatrix}$$

What is the distribution of W_2?

Answer All the marginal distributions for a multivariate normal are themselves normal. The variance for W_2 is the $(2, 2)$ entry of the covariance matrix. Hence $\boxed{W_2 \sim \mathsf{N}(-2, 4.6)}$.

Often we are not told the matrix A. Instead, we are only given the covariance matrix Σ. This covariance matrix will always have a property from linear algebra called *positive definiteness*. It is always possible to figure out from such a positive definite matrix what the matrix A is so that $\Sigma = AA^T$. This is called the *Cholesky decomposition* of Σ.

Recall that for a standard normal random variable, the density is

$$f_Z(z) = \tau^{-1/2} \exp(-z^2/2).$$

When we take $X = \mu + \sigma Z$, we get density

$$f_X(x) = \tau^{-1/2}\sigma^{-1} \exp(-(1/2)((x - \mu)/\sigma)^2)$$
$$= \tau^{-1/2}\sigma^{-1} \exp\left(-\frac{1}{2}((x - \mu)\sigma^{-2}(x - \mu)\right)$$

A similar result holds for the multivariate normal.

Fact 106

For $W = (W_1, \ldots, W_n) \sim \mathsf{Multinorm}(\mu, \Sigma)$ and $w = (w_1, \ldots, w_n)$, W has joint density

$$f_W(w) = \tau^{-n/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(w - \mu)^T \Sigma^{-1}(w - \mu)\right)$$

Problems

30.1 Let Z_1, Z_2, Z_3 be iid normal.

What is the distribution of

    a) What is the distribution of $Z_1 + Z_2 + Z_3$?

    b) What is the distribution of $5 - Z_1 + 2Z_2 + 4Z_3$?

30.2 Let Z_1, Z_2 be iid normal.

    a) What is the distribution of $Z_1 + Z_2$?

    b) What is the distribution of $Z_1 - Z_2$?

c) What is the distribution of $4Z_1 - 2Z_2$?

30.3 For Z_1, Z_2, Z_3 iid normal let

$$W_1 = Z_1 + Z_2 - 2Z_3$$
$$W_2 = -Z_1 + Z_3$$
$$W_3 = Z_3.$$

a) Find $\text{Cov}(W_1, W_3)$.

b) What is the distribution of (W_1, W_2, W_3)?

30.4 Let Z_1, Z_2, Z_3, Z_4 be iid standard normal random variables. Let $W_1 = Z_1 + Z_2 + 2Z_3 - Z_4$, $W_2 = Z_2 - Z_3 + 4Z_4$.

a) Find $\mathbb{V}(W_1)$.

b) Find $\text{Cov}(W_1, W_2)$.

30.5 Suppose that (X_1, X_2, X_3) is a multivariate normal with mean $2.3, 1.8, -1.6$) and covariance matrix

$$\begin{pmatrix} 1.1 & 0 & 2.3 \\ 0 & 2.4 & 1.6 \\ 2.3 & 1.6 & 0.7 \end{pmatrix}$$

a) What is the distribution of X_1?

b) What is $\text{Cov}(X_1, X_3)$?

30.6 Suppose that (Y_1, Y_2, Y_3, Y_4) is a multivariate normal with mean $(0, 0, 0, 0)$ and covariance matrix

$$\begin{pmatrix} 4.3 & 6.2 & -2.3 & 2.6 \\ 6.2 & 3.4 & -1.7 & 10.5 \\ -2.3 & -1.7 & 1.0 & 4.7 \\ 2.6 & 10.5 & 4.7 & 4.2 \end{pmatrix}$$

a) What is the distribution of Y_2?

b) What is $\text{Cov}(Y_1, Y_4)$?

Chapter 31

Order Statistics

Question of the Day Let U_1, U_2, U_3, U_4 be iid uniform random variables over $[0, 1]$. What is the density of the second smallest number among the four?

Summary

The **order statistics** of random variables $(X_1, \ldots, X_n)$ is the vector $(X_{(1)}, \ldots, X_{(n)})$ such that there is a permutation $f : \{1, \ldots, n\} \to \{1, \ldots, n\}$ such that $X_{(i)} = X_{f(i)}$, and

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

Suppose $X_1, \ldots, X_n$ are iid with the same distribution as X. If X has density f_X with respect to Lebesgue measure, and cdf F_X, then the density of the ith order statistic is

$$f_{X_{(i)}}(s) = n \binom{n-1}{i-1} F_X(s)^{i-1} f_X(s)(1 - F_X(s))^{n-i}.$$

The *order statistics* of a vector are the same values as the vector, but listed smallest to largest. For instance, if the vector was

$$(3.1, 2.8, 1.7, 8.1, -1.2, 2.8, 3.6),$$

then the order statistics would be

$$(-1.2, 1.7, 2.8, 2.8, 3.1, 3.6, 8.1)$$

If we labeled the original components of the vector using subscripts

$$(x_1, x_2, \ldots, x_n),$$

then we use subscripts surrounded by parentheses to indicate that they are the order statistics. So the notation for the order statistics would be

$$(x_{(1)}, x_{(2)}, \ldots, x_{(n)}).$$

In the question of the day, (U_1, U_2, U_3, U_4) are random variables, and so the order statistics

$$(U_{(1)}, U_{(2)}, U_{(3)}, U_{(4)})$$

are random variables as well.

The question of the day asks about the second smallest number, which would be $U_{(2)}$ using order statistics notation.

Consider $\mathbb{P}(U_{(2)} \in da)$. This means that at least one of the uniforms is close to a, while exactly one uniform is below a, and 2 more uniforms are above a. For a standing uniform U, $\mathbb{P}(U \in da) = \mathbb{1}(a \in [0,1])\, da$. Therefore, for $a \in [0,1]$,

$$\mathbb{P}(U_{(2)} \in da) = (5)(\mathbb{1}(a \in [0,1])\, da)\binom{4}{1}(a)(1-a)^2 = 20a(1-a)^2 \mathbb{1}(a \in [0,1])\, da.$$

Here 5 is the number of choices among the uniforms to pick the one that will be near a, da is the probability that the uniform chosen is actually within an interval of width da that surrounds a. The factor $\binom{4}{1}$ is the number of ways to choose which of the four remaining uniforms is below a, and a is the chance that it actually is below a. Finally $(1-a)^2$ is the probability that the remaining two uniforms are above a.
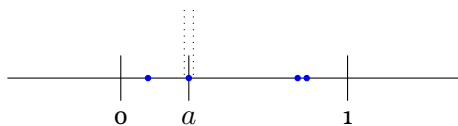
This is a *Beta* distribution with parameters 2 and 3. Write $U_{(2)} \sim \text{Beta}(2,3)$.
More generally, the following fact holds.

> **Fact 107**
> Suppose $U_1, \ldots, U_n$ are iid $\text{Unif}([0,1])$. Then
> $$U_{(i)} \sim \text{Beta}(i, n - i + 1).$$

31.1   Formula for the density of an order statistic

For continuous random variables, this same argument can be generalized to give the following method for calculating the density of order statistics.

> **Fact 108**
> For $X_1, \ldots, X_n$ iid as X that has density f_X with respect to Lebesgue measure, let F_X be the cdf of X. Then the ith order statistic has density
> $$f_{X_{(i)}}(s) = n\binom{n-1}{i-1} F_X(s)^{i-1} f_X(s)(1 - F_X(s))^{n-i}$$
> with respect to Lebesgue measure.

Proof. Suppose $X_1, \ldots, X_n$ are iid with cdf F_X, and let $i \in \{1, \ldots, n\}$. Then for the ith order statistic to be less than a, at least i of the values of the X_j have to be at most a. So

$$\mathbb{P}(X_{(i)} \le a) = \sum_{j=i}^{n} \binom{n}{j} F_X(a)^j (1 - F_X(a))^{n-j}.$$

Differentiating gives

$$f_{X_{(i)}}(a) = \sum_{j=1}^{n} \binom{n}{j} [j F_x(a)^{j-1} f_X(a)(1 - F_X(a))^{n-j} - (n-j)F_X(a)^j(1 - F_X(a))^{n-j} f_X(a)]$$

Consider now that

$$f(n,j) = j\binom{n}{j}F_X(a)^{j-1}F_X(a)^{n-j} = n\frac{(n-1)!}{(j-1)!(n-j)!}F_X(a)^{j-1}F_X(a)^{n-j},$$

and

$$(n-j)\binom{n}{j}F_X(a)^j(1-F_X(a))^{n-j-1} = n\frac{n!(n-j)}{j!(n-j)!}F_X(a)^j(1-F_X(a))^{n-j-1}$$

$$= n\frac{(n-1)!}{j!(n-j-1)!}F_X(a)^j(1-F_X(a))^{n-j-1}$$

$$= f(n,j+1),$$

So the sum for j from i to n is

$$[f(n,i) - f(n,i+1)] + [f(n,i+1) - f(n,i+2)] + \cdots + [f(n,n) - f(n,n+1)]$$

which is a telescoping sum. When $j = n$, $n - j = 0$, so the last term $f(n, n + 1) = 0$, leaving the sum as just $f(n, j)$. Hence

$$f_{X_{(i)}}(a) = \sum_{j=1}^{n}\binom{n}{j}[jF_x(a)^{j-1}f_X(a)(1-F_X(a))^{n-j} - (n-j)F_X(a)^j(1-F_X(a))^{n-j}f_X(a)]$$

$$= f(n,i)f_X(a)$$

$$= n\binom{n-1}{j-1}F_X(a)^if_X(a)(1-F_X(a))^{n-i}.$$

□

Problems

31.1 Suppose X_1, X_2, X_3 are iid with density $f(s) = s/2 \cdot \mathbb{1}(s \in [0, 2])$.

    a) What is the density of $X_{(1)}$?

    b) What is $\mathbb{E}[X_{(2)}]$?

31.2 Suppose T_1, T_2, T_3 are iid $\mathsf{Exp}(3)$. What is the density of $T_{(2)}$?

31.3 Suppose $\mathbb{P}(X = 0) = 0.3$, $\mathbb{P}(X = 1) = 0.5$, and $\mathbb{P}(X = 2) = 0.2$. Suppose that X_1, X_2, X_3 are iid with the same distribution as X.

    a) What is the distribution of $X_{(1)}$?

    b) What is $\mathbb{E}[X_{(2)}]$?

31.4 Suppose X has density

$$f_X(i) = 0.2\mathbb{1}(X = 1) + 0.7\mathbb{1}(X = 9) + 0.1\mathbb{1}(X = 13).$$

Let X_1, X_2, X_3 be iid as X. Draw the cdf of $X_{(2)}$

31.5 What is the chance that for three iid uniforms over $[0, 1]$, that the middle of the three numbers falls in the interval $[1/3, 2/3]$?

31.6 What is the chance that for eleven iid uniforms in $[0, 2]$, the middle number is between 0.9 and 1.1?

Measurable functions and Random variables

Question of the Day Can random variables be expressed as functions?

Summary
For a measurable space $(\Omega_1, \mathcal{F}_1)$ and measurable space $(\Omega_2, \mathcal{F}_2)$, a function $X : \Omega_1 \to \Omega_2$ is **measurable** if for all $A \in \mathcal{F}_2$, the set $\{a \in \Omega_1 : X(a) \in A\} \in \mathcal{F}_1$. If the measurable space $(\Omega_1, \mathcal{F}_1)$ has a probability distribution $\mathbb{P}_1$ on it, then X induces a probability distribution on Ω_2 called the **distribution of** X, and X is a **random variable**.

A random variable represented a number about which we have partial information. That is the essence of a random variable, but it is not all that helpful when it comes to rigorous proofs. So mathematicians have developed another view of random variables, that they are functions. Since functions can be defined in terms of sets, this reduces probability to set theory, which was a popular form of mathematical foundations in the first half of the twentieth century.

To explain this process begin with a measurable space. Remember that this is a set such as Ω_1, together with a collection of subsets of Ω_1 that form a σ-algebra. Call the σ-algebra $\mathcal{F}_1$. We call $\mathcal{F}_1$ the *measurable* sets. Because Ω_1 has measurable sets, now we can create a probability distribution $\mathbb{P}_1 : \mathcal{F}_1 \to [0, 1]$.

Next, suppose that we have a second set Ω_2 with its own set of measurable sets $\mathcal{F}_2$. Finally, assume that we have a function $X : \Omega_1 \to \Omega_2$ that takes elements of Ω_1 and maps them to elements of Ω_2.

So the value of $X(a)$ always lies in Ω_2. Let $A \subseteq \Omega_2$. Then an event like $\{X \in A\}$ is really mathematical shorthand for the following:

$$\{X \in A\} = \{a \in \Omega_1 : X(a) \in A\}.$$

(This set $\{X \in A\}$ is also known as the *inverse of A under X*.)

In order for this event $\{X \in A\}$ to mean something, we want this new event to be measurable back in $\mathcal{F}_1$. That is the motivation behind the following definition.

Definition 75
Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be a pair of measurable spaces. Let $X : \Omega_1 \to \Omega_2$. Then X is a **measurable function** if

$$(\forall A \in \mathcal{F}_2)(\{a : X(a) \in A\} \in \mathcal{F}_1).$$

This is equivalent to saying that for all measurable events A in $\mathcal{F}_2$, we want the inverse event $\{X \in A\}$ to be measurable back in $\mathcal{F}_1$.

Example 70

Suppose $\Omega_1 = [0, 1]$, $\Omega_2 = \{3, 4\}$, and $\mathcal{F}_2 = \{\emptyset, \{3\}, \{4\}, \{3, 4\}\}$.
    Consider the function
$$X = 3 + \mathbb{1}(a \leq 0.3).$$

What events does $\mathcal{F}_1$ have to contain for X to be a measurable function?

Answer   Looking at the inverse of measurable events in $\mathcal{F}_2$ gives

$$\{X = 4\} = \{X \in \{4\}\} = \{a : X(a) \in \{4\}\} = [0, 0.3].$$

Similarly

$$\{X = 3\} = (0.3, 1], \quad \{X \in \emptyset\} = \emptyset, \quad \{X \in \{3, 4\}\} = [0, 1].$$

So X is a measurable function if and only if

$$\{\emptyset, [0, 0.3], (0.3, 1], [0, 1]\} \subseteq \mathcal{F}_1.$$

So at this point if A is measurable in Ω_2, then $\{X \in A\}$ is measurable in Ω_1. Now suppose that we have a probability measure over Ω_1. Then $\{X \in A\}$ can be assigned a probability. That probability is of course the *distribution* of X since

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A).$$

Definition 76

If $X : \Omega_1 \to \Omega_2$ is a measurable function, and $\mathbb{P}$ is a probability measure over Ω_1, then $\mathbb{P}_X(A) = \mathbb{P}(X \in A)$ is a probability measure over Ω_2, and X is a **random variable**.

Remember that like all such definitions, this one helps us prove theorems and facts, but does not really help with our intuition. The intuition remains the same as always: any reasonable function of a uniform random variable is also a random variable.

Problems

32.1 Suppose $\Omega_1 = [0, 1]$ and $\Omega_2 = \{1, 2, 3\}$. Say The function $Y : \Omega_1 \to \Omega_2$ is defined as

$$Y(y) = 1 + \mathbb{1}(y \in [0.4, 0.6]) + \mathbb{1}(y \in [0.5, 0.7]).$$

What is $Y^{-1}(\{2\})$?

32.2 Suppose $\Omega_3 = [-1, 1]$ and $\Omega_4 = [0, 5]$. For $X : \Omega_3 \to \Omega_4$ defined as

$$X(x) = x^2,$$

find $X^{-1}([0.5, 1.5])$.

32.3 Suppose $\Omega_1 = [0, 1]$, $\Omega_2 = \{1, 2, 3\}$, and

$$\mathcal{F}_2 = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

For $Y : \Omega_1 \to \Omega_2$, let

$$Y(y) = 1 + \mathbb{1}(y \in [0.4, 0.6]) + \mathbb{1}(y \in [0.5, 0.7]).$$

Then what sets must $\mathcal{F}_1$ (a sigma algebra over Ω_1 contain in order for Y to be a measurable function?

32.4 Suppose $\Omega_3 = [-1, 1]$ and $\Omega_4 = [0, 5]$. For $X : \Omega_3 \to \Omega_4$ defined as

$$X(x) = x^2,$$

and

$$\mathcal{F}_4 = \{\emptyset, [0.5, 1.5], [0, 0.5) \cup (1.5, 5], [0, 5]\},$$

what sets must $\mathcal{F}_4$ (a sigma algebra over Ω_4) contain in order for X to be a measurable function?

32.5 Continuing with Y from earlier, if $\mathbb{P}_1$ over Ω_1 is uniform, what is $\mathbb{P}(Y \leq 2)$?

32.6 Continuing with X from earlier, if $\mathbb{P}_3$ over Ω_3 is uniform, what is $\mathbb{P}(X \leq 0.3)$?

Part II

Experiments in probability

Getting to know randomness

Summary In this lab you will get to understand the basic behavior of random variables through use of the R programming environment. The `sample` command can be used to draw random variables from a variety of distributions. The `hist` command can be used to summarize date

Instructions   In this first lab we will learn how to use the R programming environment to learn about probability. This lab is divided into two parts to help you gauge your progress.

Before you begin   Be sure that the latest version of R and RStudio are loaded onto your computer or laptop.

- You can find R at https://www.r-project.org/. Click the link to **download R** in the first paragraph, and proceed from there.

- RStudio is an IDE for R that you can download from posit at https://posit.co/products/open-source/rstudio/. Click the **DOWNLOAD RSTUDIO** button in the upper right corner of the website. Be sure to download RStudio Desktop and *not* RStudio Desktop Pro.

First part   To begin, start up RStudio, an integrated development environment (IDE) for R. In the lower left window will be the console. Here you can type commands and see what R does.

- We will start with some simple arithmetic. Type

   ```
   3 + 5
   ```

   Note that R returns `[1]  8`. The 8 is of course the answer to $3 + 5$, the `[1]` indicates that the 8 is the first number in the output. Try using R to calculate $141 + 232 - 14$ and record your answer.

- One thing that is useful is getting R to generate sequences. Try

   ```
   1:6
   ```

and report the result.

- Next let's roll a fair six sided die. Try

      sample(1:6, 1)

  and report the result.

- Now try the sample `sample(1:6, 1)` command three times. You do not have to retype the command three times, use the up arrow key to get back commands that you previously used in R.

- R can of course draw three numbers from the sequence by passing more parameters to `sample`. Try

      sample(1:6, 3)

  and report your result.

- Notice that the numbers that were returned were all different. That is the default behavior for the `sample` command. You can see this causes problems when you try to sample more values than are in the set. Try

      sample(1:6, 7)

  Report the last five words of the error message R gives.

- To find out more about `sample`, we can put a ? in front of the command to enter the help. The help screen appears in the right hand corner. Try

      ?**sample**

  What are the first two parameters for the `sample` command?

- In order to get sampling with replacement, we need to add another parameter. Try

      sample(1:6, 7, **replace** = TRUE)

  (Note that TRUE is all caps here.) Report your result.

- Now lets keep your sample in another variable. In R, the <− command assigns a value to a variable name. So try

```
x <- 5
y <- 3
x + y
```

  Record your answer.

- Now let's put our random sample in a variable. Try

```
results <- sample(1:6, 7, replace = TRUE)
print(results)
```

  and record your results.

- Let's generate a lot of fair six-sided die rolls, and look at a histogram of the results. Try

```
results <- sample(1:6, 10^7, replace = TRUE)
hist(results)
```

  and sketch the result.

- The sequence $1:6$ is an example of a *vector* in R. To create vectors manually, you can use the c command (for combine). Try the following

```
die <- c(1, 1, 1, 2, 2, 3, 4, 5, 6, 7)
print(die)
```

  Report the result.

- Now lets sample from this distribution. Try

```
results.die <- sample(die, 10^7, replace = TRUE)
hist(results.die)
```

  and sketch the result.

- If we want to see the seventh entry from results.die, we can use square brackets. Try

```
print(results.die[7])
```

and report your result.

- To see the first ten entries of `results.die`, combine the brackets with the sequence notation. Try

```
print(results.die[1:10])
```

and report your result.

- We can check which of the first 10 entries are equal to 1 by using the == command. Try

```
print(results.die[1:10] == 1)
```

Did the TRUE and FALSE pattern match what you thought it would?

- If we apply a numerical function to the output, then it will convert TRUE to a 1 and FALSE to a 0. Try

```
sum(results.die[1:10] == 1)
```

Record the result.

- We can estimate the probability that a 1 appears by taking the number of ones that appear and dividing by the number of draws. Try

```
sum(results.die == 1) / length(results.die)
```

Record your result. Was this what you expected?

Second Part   In this second part of the lab you will learn about the `min`, `max`, and `sum` functions in R

- The `min` function finds the minimum of a vector of values. Try

```
v <- c(10, 4, 7)
min(v)
```

Record the result.

- The `max` function finds the maximum of a vector of values. Try

```
max(v)
```

Record the result.

- The `sum` command finds the sum of a vector of values. Try

```
sum(v)
```

Record the result.

- We can use these commands on multiple rolls of our six sided die. Try

```
min(sample(1:6, 3, replace = TRUE))
```

and report your result. This is the smallest value among three rolls of a fair six-sided die.

- What is the probability that $\min\{X_1, X_2, X_3\} = 1$ where $X_1, X_2, X_3 \sim \mathsf{Unif}(\{1, \ldots, 6\})$ are independent and identically distributed (abbreviated iid)? (Hint: it might be easier to calculate the complement of this event.)

- Let's test your answer with a simulation. First, we must generate a bunch of draws from the minimum of the three rolls. We can use the `replicate` command in R to accomplish this. Try

```
results <- replicate(10^6, min(sample(1:6, 3, replace = TRUE)))
hist(results)
```

and sketch your results. Because we are using the minimum function, the result is closer to 0.

- Estimate the probability that a 1 occurred with

```
sum(results == 1) / length(results)
```

Report the result.

- Now try out the max function with

```
results <- replicate(10^6, max(sample(1:6, 3, replace = TRUE)))
hist(results)
```

Sketch the result.

- Last (but not least) let's try summing the three die rolls, and sketching the histogram.

```r
results <- replicate(10^6, sum(sample(1:6, 3, replace = TRUE)))
hist(results)
```

- Previously we used the `sample` command to draw from a set of objects with replacement, but what if we draw without replacement? Then we end up with a random permutation of the objects. Try

```r
letters <- c('a', 'a', 'a', 'b', 'b')
perm <- sample(letters)
print(perm)
```

  Report the result.

- Note that although we used the single quote ' in defining the variable `letters`, R prints out the result using a double quote ". In fact, you can use either the single quote or the double quote in defining letters (which are called strings in most computing languages.

  What should the probability that the letter 'a' falls into the third position be?

- Now let's test that through simulation. Try

```r
results <- replicate(10^5, sample(letters)[3] == 'a')
sum(results)/length(results)
```

  Report your estimate.

- We've been using `sum(results) / length(results)` to get the average value of the results vector, but there actually is a command in R that does both these simultaneously. Try

```r
\mean(results)
```

- Try estimating the probability that 'a' is in the fourth position. Does the position matter for the probability?

- Try estimating the probability that 'b' is in the 3rd position.

Chapter 34

Continuous random variables

Summary In the first part of the lab you will learn to use the `runif` commands to generate random variables that are uniform over $[0, 1]$. In the second part you will learn about *order statistics* and the `sort` command.

- Try the command

    ```
    runif(1)
    ```

    This generates a single uniform random number over the interval $[0, 1]$. Try the command three times and record your results. By the way, you don't have to retype the command three times: use the up arrow key to reissue commands that you previously used in R.

- Note that R could have generated all 3 numbers at once by changing the parameter given to `runif`. Try

    ```
    runif(3)
    ```

    and record your results.

- Now let's put 1000 random iid uniform $[0, 1]$ numbers into an array with

    ```
    a <- runif(1000)
    ```

    Try typing `a`. You can see that as it displays the numbers. Perhaps more useful is a plot. Try

    ```
    plot(a)
    ```

    On the x-axis are the numbers 1 to 1000, indicating which uniform they are plotting. On the y-axis are the actual numbers, which all fall between 0 and 1. Now try

    ```
    hist(a)
    ```

    in order to get a histogram of the numbers. Sketch the resulting plot.

- Unlike with the discrete random variables from our last lab, for continuous random variables there is no easy way to see how many bars should be in the histogram. The histogram will probably contain about 10 bars by default, but we can change that by passing parameters to the hist function. We can create more advanced sequences of numbers in R using the seq command. Try

```
seq(0, 1, by = 0.1)
```

What do you get?

- Let's make more bars by using the seq command to give the boundaries of the histogram. Try

```
hist(a, seq(0, 1, by = 0.05))
```

How many bars are there in the histogram?

- Our histogram is looking a bit ragged, so let's up the number of uniforms that we are using. Try

```
hist(runif(10^6), seq(0, 1, by = 0.05))
```

About what is the frequency of each bar?

- You can get the inbuilt help for the hist command by using ?hist in R. Always the ? in front of a command gives the help for that command. Of course, you can always learn about a command in R by Googling it. If you try ?hist, you will see that there is a freq parameter. By default this parameter is TRUE; by setting it to FALSE instead the density of the numbers will be given. Try the following command

```
hist(runif(100000), seq(0, 1, by = 0.05), freq = FALSE)
```

What is the label on the y-axis now?

- Unlike discrete uniform random variables, continuous uniform variables can be *shifted* and *scaled*. Start by scaling. This code takes the numbers that lie in the interval $[0, 1]$ and multiplies them by three so that they lie in the interval $[0, 3]$.

```
results <- 3 * runif(10^6)
```

Our x-axis will run from 0 up to 3 to hold these new numbers.

```
hist(results, seq(0, 3, by = 0.05), freq = FALSE)
```

Sketch the resulting histogram.

- The result of multiplying a uniform over $[0, 1]$ by 3 is to scale the interval. The resulting random variable is now uniform over $[0, 3]$. Now shift the uniform as well as scale by adding 2. This will make the new random variable uniform over $[2, 5]$. The following code does this and then makes the x-axis from 1 to 6. Note that from 1 up to 2 and 5 up to 6 there are no numbers, they are all uniform over $[2, 5]$.

```
results2 <- 3 * runif(10^6) + 2
hist(results2, seq(0, 6, by = 0.05), freq = FALSE)
```

  Sketch the resulting histogram.

- This is an estimate of the density function of the uniforms. Let's try looking at the density of functions like the square of uniforms. Recall that numbers can be squared using ^2.

```
seq(1, 4)
seq(1, 4)^2
```

  Now try it with random draws.

```
runif(5)^2
```

  Since numbers from 0 up to 1 when squared are still in $[0, 1]$ these do not look much different at first glance. But the histogram of lots of draws will be very different! Try

```
hist(runif(100000)^2, breaks = 20, freq = FALSE)
```

  and sketch the result. This is the density estimate for the square of uniforms. Remember that squaring a number between 0 and 1 will make it smaller, so this pushes the density towards smaller values.

- Repeat with the square root function instead.

```
hist(sqrt(runif(100000)), breaks = 20, freq = FALSE)
```

  and sketch the result. This is the density estimate for the square root of uniforms. Remember that taking the square root of a number between 0 and 1 will make it larger, so this pushes the density towards values closer to 1.

- Next do this for the negative log function

```
hist(-log(runif(100000)), breaks = 20, freq = FALSE)
```

and sketch the result.

- Now consider estimating something like $\mathbb{P}(U_1 \leq U_2^2)$, where U_1 and U_2 are independent uniform over $[0, 1]$. Try

```
mean(runif(10^6) <= runif(10^6)^2)
```

Report your result. Was this what you expected?

- Use the same idea to estimate $\mathbb{P}(U_1 \leq U_2^3)$.

Second part   This part of the lab studies what are called the *order statistics* of random variables. The first order statistics we will examine are the maximum and minimum of the random variables.

- Two more functions that are used a lot in probability are `max` and `min`. Try

  ```
  max(10, 14)
  ```

  and record the result.

- Now generate the maximum of two uniform random numbers.

  ```
  max(runif(2))
  ```

  To do this same action lots of times, use the `replicate` command

  ```
  a <- replicate(10000, max(runif(2)))
  hist(a, breaks = 20, freq = FALSE)
  ```

  Sketch the resulting density estimate.

- What is the range of the density (the y-axis)?

- Now generate an equal number of random variables that is the maximum of three independent uniforms, and sketch the resulting histogram.

- Another useful function is `sum`. Try the following.

  ```
  a <- replicate(10000, sum(runif(2)))
  hist(a, breaks = 20, freq = FALSE)
  ```

  Sketch the result.

- Try the same thing, but with 100000 replications and summing 10 uniforms.

- The maximum is the largest of the uniforms, and the minimum is the smallest. How do we get at the middle one? We can use the `sort` command in R. Try

  ```
  sort(runif(3))
  ```

  and report your results.

- We can pick out the second element of this vector using `sort(unif(3))[2]`. (Note that we use brackets `[2]` around the 2 when we are picking elements out of a vector.)

  ```
  b <- replicate(10000, sort(runif(3))[2])
  hist(b, breaks = 20, freq = FALSE)
  ```

  and sketch the result. Notice how using this middle value is more likely to be in the middle of $[0, 1]$ than a vanilla uniform.

- When you sort random variables X_1, X_2, X_3 you are creating what are called *order statistics*. They are written using parenthesis around the subscript, so

  $$X_{(1)} \leq X_{(2)} \leq X_{(3)}.$$

  For example, if $X_1 = 0.34$, $X_2 = 0.15$, and $X_3 = 0.75$, then $X_{(1)} = 0.15$, $X_{(2)} = 0.34$, and $X_{(3)} = 0.75$. Try generating 10^6 draws from the third order statistic of 10 uniforms, and sketch a histogram of the result.

- Let's consider how to calculate

  $$\mathbb{P}(X_{(2)} < 0.3, X_{(3)} \geq 0.3\}.$$

  This is saying that the second smallest of the numbers is less than 0.3, and the third smallest of the numbers is at least 0.3. For this to happen, exactly two of the uniforms must be at most 0.3, and exactly eight of the uniforms must be at least 0.7. There are 10 choose 2 ways to choose the small uniforms, and then the large uniforms are those that remain. The chance the chosen uniforms are small is 0.3^2, and the chance that the unchosen uniforms are large is 0.7^8. So altogether the chance is

  $$\binom{10}{2} 0.3^2 0.7^8.$$

  You can find this value in R using

  ```
  choose(10, 2) * 0.3^2 * 0.7^8
  ```

What is this probability?

- Now let's try to estimate this probability by simulation. We need to generate a random variable that is 1 if $X_{(2)} < 0.2$ and $X_{(3)} \geq 0.3$ and 0 otherwise. To do this we will need to give replicate two commands. Multiple commands can be combined in R in a single line using a semicolon symbol ; . We will put this in curly braces { and } to indicate that the commands should be combined.

```
{v <- sort(runif(10)); as.integer(v[2] < 0.3 & v[3] >= 0.3)}
```

Report your result.

- By the way, a random variable that is either 0 or 1 is called a *Bernoulli* or an *indicator* random variable. It is called an indicator random variable because it can be written using an indicator function.

$$Y = \mathbb{1}(X_{(2)} < 0.3, X_{(3)} \geq 0.3).$$

Now, one draw from the distribution of Y isn't very helpful. Let's do this a million times and record the results.

```
results <- replicate(10^6, {v <- sort(runif(10)); as.integer(v
    [2]<0.3 & v[3] >= 0.3)})
mean(results)
```

(Note this first command might take a while depending on the speed of your computer. Try it first with 10^4 and then 10^5 to get an idea of how long 10^6 will take.) Report your estimate.

- Calculate exactly the following probability for $X_1, \ldots, X_{10}$ iid $\mathsf{Unif}([0, 1])$.

$$\mathbb{P}(X_{(4)} < 0.5, X_{(5)} \geq 0.5).$$

- Estimate the above probability using 10^6 samples from $\mathbb{1}(X_{(4)} < 0.5, X_{(5)})$. Report your estimate.

- Consider a different problem. Suppose that $T_1 \sim \mathsf{Exp}(1)$ and $T_2 \sim \mathsf{Exp}(2)$ are independent exponential random variables. Then because T_2 has the higher rate, it will tend to be smaller than T_1. But what is the chance of that? To find out, first consider generating copies of the T_1 and T_2 random variables uniformly, and comparing one by one.

An exponential random variable of rate λ can be generated by taking the negative log of a uniform, and dividing by λ.

Try

```
n <- 10^6
t1 <- -log(runif(n))
t2 <- -log(runif(n)) / 2
mean(t2 <= t1)
```

and report your estimate. (Note that here each component of `t1` is $\mathrm{Exp}(1)$ and each component of `t2` is $\mathrm{Exp}(2)$.)

- Estimate the probability that $T_1 \sim \mathrm{Exp}(1)$ is at least $T_3 \sim \mathrm{Exp}(3)$ if the random variables are independent.

Chapter 35

Conditioning

Summary Partial information about a random variable can be encoded using *conditioning*. Write $\mathbb{P}(A|B)$ for the probability A occurs given B occurs. The conditional probability formula is (for $\mathbb{P}(B) > 0$)

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

First part One way of thinking about the probability of an event is that an experiment is being carried out simultaneously in an infinite number of parallel universes. The percentage of the universes where the event happens is the probability of the event. This way of thinking about probability is called the *frequentist* interpretation and gives us a way to estimate probabilities by performing independent experiments.

- Let's begin by generating some uniforms over $\{1, \ldots, 10\}$. Note that we will be explicitly setting parameters in the following command. This is so we do not have to remember what order the parameters are supposed to come in.

    ```
    sample(x = 1:10, size = 5, replace = TRUE)
    ```

    Report your sample.

- Try the same command, but with the parameters in a different order.

    ```
    sample(size = 5, replace = TRUE, x = 1:10)
    ```

- Now try generating 7 iid draws from $\{1, \ldots, 10\}$. Write down the command you used.

- Now get a bunch of these draws by changing the value of `size` parameter. Try

    ```
    results <- sample(x = 1:10, size = 10^6, replace = TRUE)
    head(results)
    length(results)
    ```

What is the number of elements of the results vector?

- Plot the histogram of these results with

```
hist(results, breaks = 0:10)
```

and sketch the result.

- For an event A, we can estimate the probability that A occurs by counting out of our trials, how many fell into A. For instance, for $A = \{1, 2, 3, 4\}$, typing

```
results[1:10] <= 4
```

returns a vector of entries that are TRUE or FALSE. If we use

```
sum(results[1:10] <= 4)
```

then all the TRUE values are converted to 1, the FALSE values are converted to 0, and then summed up to tell you the total number of true statements. How many TRUE values were there in your first ten uniforms?

- Now let's estimate $\mathbb{P}(U \in A)$ by

```
sum(results <= 4) / length(results)
```

That tells us what percentage of uniforms out of the 10^6 draws were at most 4. Instead of using **sum** and **length** and dividing, you can also use **mean**. Try

```
mean(results <= 4)
```

and report your result.

- To select some of the elements of `results`, we can use something like `results[1:10]`. Give this a try and report the result.

- We can use a logical statement to find elements that satisfy a certain criterion. We first put the first ten elements of results into x. Next, we return only those elements of x that are at most 6.

```
x <- results[1:10]
print(x)
print(x[x <= 6])
```

Report your results.

- The last command picked out the first ten entries of `results` that were at most 6. Now let's do that for the whole million entries.

```
r6 <- results[results <= 6]
```

How many entries are there in `r6`?

- Now let's plot the histogram of `r6`

```
hist(r6, breaks = 0:10)
```

and sketch the result.

- The resulting histogram looks a lot like that of the uniform distribution over the integers 1 through 6. This is not a coincidence! Using our conditioning notation, say that for $X \sim \mathsf{Unif}(\{1, \ldots, 10\})$ it holds that

$$[X|X \le 6] \sim \mathsf{Unif}(\{1, 2, \ldots, 6\}).$$

The idea is that if a random variable X is uniform over B, and $A \subseteq B$, then $[X|X \in A]$ is uniform over A. Using this idea, what is the distribution of $[X|X \le 3]$?

The conditional probability formula

- Okay, now let's use our uniforms to create events and test the conditional probability formula. Recall that for random variables X and Y, if $\mathbb{P}(X \in A) > 0$,

$$\mathbb{P}(Y \in B|X \in A) = \frac{\mathbb{P}(Y \in B, X \in A)}{\mathbb{P}(X \in A)}.$$

If $U \sim \mathsf{Unif}(\{1, \ldots, 10\})$, then let $A = \{1, 2, 3, 4\}$ and $B = \{3, 4, 5, 6, 7\}$. Then the conditional probability formula says that

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

We still have 10^6 draws from the uniform in `results`. First let's estimate $\mathbb{P}(U \in B)$ with

```
sum(results <= 7 & results >= 3) / length(results)
```

Note that the ampersand character `&` means *logical and* in R. That means that both the number must be at most 7 and at least 3 for it to be true. Report your resulting estimate.

- Next estimate the probability that our result is both in A and in B

```
sum(results <= 7 & results >=3 & results <= 4) / length(
    results)
```

Report your estimate of $\mathbb{P}(U \in A \cap B)$.

Note that the `results <= 7` part of the formula is redundant. If `results <= 4` is true then it must also be true that `results <= 7`.

- So our estimate for $\mathbb{P}(U \in A | U \in B)$ is the estimate for $\mathbb{P}(U \in A, B)$ divided by the esttimae for $\mathbb{P}(U \in B)$. Record this estimate. .

- Now let's try conditioning directly. First create a vector `ub` contains uniforms conditioned to lie in B. Then see how many of these fall into A.

```
ub <- results[results <= 7 & results >= 3]
sum(ub <= 4) / length(ub)
```

Report this estimate of $\mathbb{P}(U \in A | U \in B)$. How does this compare to our previous estimate?

Second Part

Bayes' Rule

- One of the classic errors in probability is to mix up $\mathbb{P}(A|B)$ (the chance A occurs given B occurs) and $\mathbb{P}(B|A)$ (the chance B occurs given A occurs.) Let's tackle this with an experiment. Suppose that $U \sim \mathsf{Unif}([0,1])$, and $A = \{U \in [0, 0.3]\}$, while $B = \{U \in [0.2, 0.4]\}$. First let's estimate $\mathbb{P}(A|B)$:

```
results <- runif(10^6)
mean(results[results <= 0.4 & results >= 0.2] <= 0.3)
```

Note that the `results <= 0.4 &results >= 0.2]` part only keeps uniforms in $[0.2, 0.4]$. This is the conditioning on B part of things. Then the `<= 0.3` part of things estimates the probability that A occurs given B. Report your estimate.

- Now let's try it the other way around, and estimate the probability of B given that A occurs.

```
results_A <- results[results <= 0.3]
mean(results_A <= 0.4 & results_A >= 0.2)
```

Report your estimate. Recall that your first estimate was for $\mathbb{P}(A|B)$, and this estimate is for $\mathbb{P}(B|A)$.

- Bayes' Rule says that there is a way to turn the conditioning around, that

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

Let's try that

```
prob.b <- mean((results <= 0.4) & (results >= 0.2))
prob.a <- mean(results <= 0.3)
prob.bgivena <- mean((results[results<=0.3] <= 0.4) & (results[
    results <= 0.3] >= 0.2))
print(prob.bgivena * prob.a / prob.b)
```

Report your result.

- A classic problem involving Bayes' Rule goes as follows. Suppose that there is a 1% chance that a particular person has a disease. A test for the disease is correct 95% of the time, but is wrong 5% of the time. If the test says that the person has the disease, what is the chance that the person actually has the disease?

To simulate this, we will use U_1 to determine if someone has the disease, and U_2 to determine if someone tests positive for the disease. Start with the following.

```
u1 <- runif(10^8)
u2 <- runif(10^8)
```

Someone has the disease if $U_1 \leq 0.01$. If $U_1 \leq 0.01$, the test is positive if $U_2 \leq 0.95$. But if $U_1 > 0.01$ (they do not have the disease) the test is still positive if the test is wrong, that is, if $U_2 > 0.95$. Recall that `&` is logical and, while `|` is logical or. So we are only keeping the results where the test is positive if we set

```
test.pos <- u1[((u1 <= 0.01) & (u2 <= 0.95)) | ((u1 > 0.01) & (
    u2 > 0.95))]
print(length(test.pos)/length(u1))
```

Report the result of these commands. Does this fit with what you thought the probability of a positive test would be?

- So `test.pos` now contains all the uniforms where the test was positive. Some of these were positive because $U_1 \leq 0.01$ and $U_2 \leq 0.95$, but some are positive because $U_1 > 0.01$ and $U_2 \leq 0.05$. To estimate the chance that someone actually has the disease, try

```
mean(test.pos <= 0.01)
```

Report your result.

- Note that even though the test reported positive, the chance of having the disease is still only about one in six. The reason is that it is much more likely that the person does not have the disease and the test was wrong ($0.99 \cdot 0.05 = 0.0495$) then it is that the person does have the disease and the test was right $0.01 \cdot 0.95 = 0.0095$). So given the test was positive, the chance of having the disease is only $0.0095/[0.0095 + 0.0495]$. Find this value.

Uniforms in more than two dimensions

- The principle that for $A \subseteq B$, if $X \sim \mathsf{Unif}(B)$, then $[X|X \in B] \sim \mathsf{Unif}(A)$ works in higher dimensions as well. To see this in action, start with 1000 points drawn uniformly from the unit square.

```
results_pts <- t(replicate(1000, runif(2)))
```

Break this down: the `runif(2)` command generates two uniform random variables over $[0, 1]$, together they make a single point in the unit square. Then the `replicate` command repeats this procedure 1000 times. Finally, the `t` command transposes a matrix. This leaves us with a matrix with two columns and 1000 rows, which is similar to how statistical data would be presented.

Report the results of using `head(results_pts)`.

- Note that column one is labeled `[,1]` and column two is labeled `[,2]`. R uses statistical notation where the comma is a wildcard character. So `[,1]` means all the items in any row and the first column. Let's plot the first column as the horizontal axis, and the second column as the vertical axis.

  ```
  plot(results_pts[,1], results_pts[,2])
  ```

  This is what completely uniform random data looks like. Now let's select those rows where the x coordinate is less than the y coordinate.

  ```
  r1 <- results_pts[results_pts[,1] <= results_pts[,2],]
  plot(r1[,1], r1[,2])
  ```

  What is the resulting shape of the points? In other words, for $(X, Y)\mathsf{Unif}([0,1] \times [0,1])$,

  $$[(X, Y)|X \leq Y] \sim \mathsf{Unif}(A),$$

  what is the region A?

Odds

- Another way to view conditioning is through the use of *odds*. The odds between two disjoint events is the ratio of the probabilities for each of the events. So for instance, say

  $$\mathbb{P}(X = 1) = \frac{2}{9}, \ \mathbb{P}(X = 2) = \frac{3}{9}, \ \mathbb{P}(X = 3) = \frac{4}{9}.$$

  Then the odds that $X = 1$ versus $X = 2$ is $(2/9)/(3/9) = 2/3$, which can be written as $2 : 3$ (read 2 to 3). What are the odds of $X = 1$ versus $X = 3$?

- The advantage of the colon notation versus the fraction notation is that odds can be given for more than two things simultaneously. For instance, the odds for $X = 1$ versus $X = 2$ versus $X = 3$ is

  $$2 : 3 : 4.$$

  This tells us that the odds for $X = 2$ versus $X = 3$ is $3 : 4$. Suppose that $Y \in \{1, 2, 3, 4\}$ has odds of

  $$3 : 1 : 2 : 7.$$

  What are the odds of $Y = 1$ versus $Y = 4$?

- The nice thing about odds is that when we condition on an event, the odds stay the same, we just keep the values given the information. For instance, for $[Y|Y \in \{1, 3, 4\}$, we just

keep the odds numbers associated with 1, 3, and 4. So the odds $Y = 1$ versus $Y = 3$ versus $Y = 4$ given $Y \in \{1, 3, 4\}$ is

$$3 : 2 : 7.$$

The odds originally where $3 : 1 : 2 : 7$ for numbers $(1, 2, 3, 4)$. Conditioned on $Y \neq 2$, remove the number associated with 2, so the new odds are $3 : 2 : 7$ for numbers $(1, 3, 4)$. What are the odds of Y being 1, 2, or 3 given $Y \in \{1, 2, 3\}$?

- To turn odds back into probabilities, simply divide by the *normalizing constant*. This is just the sum of the numbers in the odds. For X with odds $2 : 3 : 4$, add to get normalizing constant $2 + 3 + 4 = 9$. Then divide by 9 to get odds of $(2/9) : (3/9) : (4/9)$. Then the probability $X = 1$ is $2/9$, and so on. What are the probabilities for $Y \in \{1, 2, 3, 4\}$ for the odds given earlier?

- Now let's combine these ideas, We know the odds for Y being 1 versus 2 versus 3 given $Y \in \{1, 2, 3\}$. So find $\mathbb{P}(Y = i | Y \in \{1, 2, 3\})$ for $i \in \{1, 2, 3\}$.

Continuous distributions

So far we have been using the `hist` command in R to approximate the density, but for continuous random variables, the `plot(density)` command is much more useful. This command creates what is called a *kernel density plot*, and while it does not handle discontinuities in the distribution very well, overall it paints a clearer picture of what is happening with the distribution.

- **Uniforms** The first density that we have is for $U \sim \text{Unif}([0,1])$.

$$f_U(s) = \mathbb{1}(s \in [0,1]).$$

  Remember that the indicator function $\mathbb{1}$ evaluates to 1 when the argument inside is true and 0 otherwise. So this is a fast way to write that $f_U(s) = 1$ when $s \in [0,1]$ and $f_U(s) = 0$ otherwise. Sketch a graph of this function.

- Now R does have a built in command for evaluating the density of random variables. The name is d followed by the name of the distribution. Try the following commands

  ```
  dunif(1.1)
  dunif(0.6)
  dunif(-0.3)
  ```

  and report your results.

- By using a vector, R could have accomplished the same task in one line:

  ```
  dunif(c(1.1, 0.6, -0.3))
  ```

  returns all three values. Let's use this to plot the density of the uniform. Try

  ```
  x <- seq(-2, 2, by=0.1)
  plot(x, dunif(x), type = 'l')
  ```

and sketch the results. Be sure to type a lowercase ell (`l`) for line and not a one (`1`) for the `type` parameter. .

- Note that the plot did not really handle the discontinuities at 0 and 1 very well. Let's get an estimate of the density function by drawing samples from the uniform distribution. To do this, we will use the **density** command. Try the following

```
results <- runif(10^6)
plot(density(results))
```

  Sketch the result. This command tries to approximate what the density function is for the random variable draws.

- This is called the *kernel density plot.* As with the deterministic plot, the kernel density plot does not handle discontinuities very well. Let's try an estimate of the kernel density of the sum of two uniform random variables. Here the density will be continuous, and so the kernel density estimate does a pretty good job.

```
r2 <- replicate(10^5, sum(runif(2)))
plot(density(r2))
```

  Sketch your result.

- **Exponential** You can get an exponential random variable with parameter λ by taking the negative of the natural logarithm of a uniform, and then dividing by λ. The density is

$$f(s) = \lambda \exp(-\lambda s)\mathbb{1}(s \geq 0).$$

  Does this density have a discontinuity?

- Try

```
results <- -log(runif(10^6))
plot(density(results))
```

  and sketch the result.

- Let's add a line indicating the true density on top of things.

```
x <- seq(0, 8, by = 0.1)
lines(x, exp(-1*x), col = "blue", lwd = 2)
```

Note the use of the `lines` command rather than the the `plot` command. If you had used `plot`, R would have started over with a new plot. By using `lines`, R puts the result on top of the existing plot.

Next, repeat this experiment for $\lambda = 0.7$.

```
plot(density(results / 0.7))
lines(x, 0.7 * exp(-0.7 * (x)), col = "blue", lwd = 2)
```

Sketch the result.

- Notice that when we divided the `results` variable with our random results by 0.7, we had to multiply the density by 0.7 so that the area undeath the density curve stayed 1.

  R has built in commands for generating exponential random variables. Try

```
results2 <- rexp(10^6, rate = 0.7)
lines(density(results2), col = "gold", lwd = 2)
```

Does the R command `rexp` generate data from the same distribution as the negative log method?

- Is the kernel density estimate as accurate as before?

  Gamma/Erlang   When we sum exponential random variables together, we obtain a *gamma* distribution. Let's give it a try.

```
results <- replicate(10^5, sum(rexp(3, rate = 0.7)))
plot(density(results))
```

Sketch the result.

- For $X_1, \ldots, X_k$ iid $\mathsf{Exp}(\lambda)$, say that $X_1 + \cdots + X_k = X \sim \mathsf{Gamma}(k, \lambda)$, where X has density
$$f_X(s) = \frac{\lambda^k s^{k-1} \exp(-\lambda s)}{\Gamma(a)} \mathbb{1}(s \geq 0).$$

Here $\Gamma(a)$ is called the *gamma function*. A useful fact is when a is an integer, $\Gamma(a) = (a-1)!$. Test out this fact with

```
integrate 4^6*s^5*exp(-4*s) from 0 to infinity
```

in Wolfram Alpha. What is the result? Is the result a factorial of an integer? (For instance, if your result was 24, then $24 = 4!$.)

- The command for directly generating gamma random variables in R is `rgamma`. Use this command to generate 10^6 iid $\mathrm{Gamma}(3, 0.7)$ random variables and plot the kernel density estimate.

Second Part

- **Beta** The *beta* distribution comes from looking at the *order statistics* of uniform random variables over $[0, 1]$. Given a set of random variables $X_1, \ldots, X_n$, the order statistics are the same numbers, put in order. We use subscripts surrounded by parentheses to indicate order statistics. That is,

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

So for instance, `u <-runif(4)` generates $U_1, \ldots, U_4$ iid uniform over $[0, 1]$.

The `sort(u)` command then generates their order statistics. Give this a try, and report your four order statistics $U_{(1)}, U_{(2)}, U_{(3)}, U_{(4)}$.

- To get the ith order statistic from a vector v, use `sort(v)[i]`. Putting this all together, let's generate some betas:

```
results <- replicate(10^5, sort(runif(10))[4])
plot(density(results))
```

Sketch the result.

- Note that the density peaks at about $4/10$, this is because we used the fourth order statistic out of 10 uniforms. Try the seventh order statistic out of 10 uniforms and sketch the resulting density.

- If we use the ith order statistic out of n uniforms (so $X = U_{(i)}$) then

$$X \sim \text{Beta}(i, n - i + 1)$$

and the density is

$$f_X(s) = \frac{x^{i-1}(1 - x)^{n-i}}{B(i, n - i + 1)} \mathbb{1}(x \in [0, 1]).$$

where B is a function that gives the normalizing constant called the *beta function*. Note that while lowercase Greek letter beta (β) is easy to tell apart from the lowercase Roman letter b, the upper case Greek beta (B) looks exactly like the upper case Roman letter B. Use Wolfram Alpha to find $B(7, 10 - 7 + 1)$ with

```
integrate x^(7-1)*(1-x)^(10-7) from 0 to 1
```

What is $B(7, 10 - 7 + 1)$?

- The command for directly generating beta random variables in R is `rbeta`. Use this command to generate 10^6 Beta$(7, 4)$ random variables and plot the kernel density estimate.

The next distribution we consider is the *normal* distribution, also known as the *Gaussian.*

- Try

```
z <- rnorm(10^6)
plot(density(z))
```

Sketch the result.

- This is called the *standard normal distribution.* If we scale and shift the data, we get a different normal distribution.

```
plot(density(3 * z + 10))
```

Now the peak should be centered over 10 rather than over 0. Sketch the result.

- It turns out that multiplying a standard normal by 3 is the same as adding 9 independent standard normals together. Try

```
z2 <- 10 + replicate(10^5, sum(rnorm(9)))
lines(density(z2), col = "blue")
```

(By using `lines` instead of `plot` here we add the drawing to the existing plot rather than starting over.) Does the new kernel density estimate look like the old one?

- Now let's take a look at the square of a standard normal. This is called the *chi-squared distribution.* It is also written as χ^2, since χ is the Greek letter chi. Try

```
plot(density(z^2))
```

and sketch the result.

- The χ^2 distribution is related to the gamma distribution. Try

```
lines(density(rgamma(10^6, shape = 1/2, rate = 1/2)))
```

What would you say the relationship is?

For all of the distributions (uniform, exponential, gamma/Erlang, beta, normal) that we have looked at the kernel density plot has been very good. However, this is not always the case. In this part of the lab we will look at a situation where the kernel density plot fails, namely, when the density function is only going down polynomially rather than exponentially in the tails.

- **Cauchy** The kernel density plot by default in R smooths using a normal density. This works well for random variables where the density is declining quickly (like normals, gammas, betas, and exponentials) but not so well when the density is declining slowly (like Cauchys). To see this effect, first let's plot the density of a standard Cauchy.

```
x <- seq(-10, 10, by = 0.1)
plot(x, 1 / pi / (1 + x^2), type = "l", col = "blue", lwd =
    2)
```

Sketch the result.

- Now let's do the kernel density estimate. Cauchy random variables come from taking the tangent of a uniform number over $[-\tau/4, \tau/4]$.

```
results_C <- tan(pi * runif(10^6) - pi / 2)
plot(density(results_C))
```

Sketch the result.

- Why do they look so different? Because the Cauchy tails are only going down polynomially we say they have a *heavy tail*. Heavy tailed distributions have a lot of very large and very small values, which confuses the kernel density plot.

  Another example of a random variable with a heavy tail is $X = 1/U$, where $U \sim \mathsf{Unif}([0, 1])$. Try

```
results_X <- 1 / runif(10^6)
plot(density(results_X))
```

Sketch the result. Again the presence of extremely large values throws everything off and pushes everything into a spike.

Now we will look at the relationship between the beta and gamma functions.

- It turns out that the beta function and gamma function are related. In general:

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

Since $\Gamma(a) = (a-1)!$ for integers a, this makes $B(a,b)$ look like the inverse of a bi Recall that for integers a, $\Gamma(a) = (a-1)!$. Use this to find $\Gamma(7)$, $\Gamma(10-7+1)$, $\Gamma(11)$, and $B(7, 10-7+1)$.

Chapter 37

Expected value

Summary This lab will introduce you to the mean of random variables and the Strong Law of Large Numbers

- Suppose I have four values $(x_1, x_2, x_3, x_4) = (5, 1, 4, 2)$. Then the first four sample averages are

$$5/1, \ (5+1)/2, \ (5+1+4)/3, \ (5+1+4+2)/4.$$

You can find these using the `cumsum` (cumulative sum) function in R. Try

```
x <- c(5, 1, 4, 2)
cumsum(x)
```

and report the result.

- Divide these numbers by 1, 2, 3, and 4 with

```
cumsum(x) / 1:4
```

- Verify that the last entry is the average of the four numbers with

```
mean(x)
```

- Now let's try the same thing for 1000 uniforms over $[0, 1]$.

```
u <- runif(1000)
y <- cumsum(u) / 1:length(u)
plot(y, type = "l", col = "blue")
```

Make a rough sketch of the result.

- Look at just the last half of the values with

```
plot(y[501:1000], type = "l", col = "blue")
```

  Sketch the result.

- Now generate 10^6 uniforms and again plot the sample averages. Sketch the result.

- This is an example of a set of sample averages that converge to a particular value. That value is the *mean* (aka the *expected value*, *expectation* or *average*) of the random variable. Since uniforms have continuous density, we can find them using the formula

$$\mathbb{E}[U] = \int_{u \in \mathbb{R}} u \mathbb{1}\,(0 \le u \le 1)\, du,$$

  where u is there because we are trying to find the mean of U, and $\mathbb{1}\,(0 \le u \le 1)$ is the density of a uniform random variable.

  The indicator function can be used to modify the endpoints of the integral. That is,

$$\mathbb{E}[U] = \int_0^1 u\, du,$$

  In Wolfram Alpha, use

```
integrate x from 0 to 1
```

  and give the result. .

- Now let's try to estimate $\mathbb{E}[U^2]$ in R.

```
mean(u^2)
```

Report the result.

- Now try converting the indicator function in the integral to limits

$$\int_{x\in\mathbb{R}} x^2 \mathbb{1}(0 <= x <= 1)\, dx$$

then evaluate in Wolfram Alpha and report the result.

- The fact that the sample average of $U_1^2, U_2^2, \ldots, U_n^2$ converges to $\mathbb{E}[U^2]$ is called the *Strong Law of Large Numbers*, or SLLN, and is one of the major theorems in probability theory.

  We say that a random variable U is *integrable* if $\mathbb{E}[|U|]$ is a finite number. (It will always be either a finite nonnegative number or ∞.)

  The SLLN does not apply to all random variables! Try the following and sketch the result:

  ```
  w <- 1 / runif(1000)
  plot(cumsum(w) / 1:length(w), type = "l", col = "blue")
  ```

- Try it again but with 10^6 draws from $1/U$.

- The plots show that there are sudden bursts of increases followed by a slow decline. If the bursts are bigger than the decline parts, then the SLLN will not hold.

  To get an idea of why these bursts occur, let's look at the four largest values of w:

  ```
  sort(w)[(length(w)-3):length(w)]
  ```

  Report these values.

- Note that the change in the overall sample average caused by the largest of these values is the value divided by 10^6. How much did just this one value change the sample average?

- When the random variable occasionally has these super large values, it keeps making the total sample average larger and larger. One way to see if this happens is to look at the integral that gives $\mathbb{E}[|1/U|]$.

  What is

$$\int_{u \in \mathbb{R}} (1/u) \mathbb{1}\,(u \in [0, 1])\ du?$$

Second Part

- **Cauchy distribution** In the example of $1/U$, $|1/U| = 1/U$ because it is always positive. A different example is the Cauchy distribution, which has density

$$f_X(s) = \frac{2}{\tau} \cdot \frac{1}{1 + s^2}.$$

First let's draw the density of a Cauchy, and also a normal distribution for comparison

```
x <- seq(-5, 5, by = 0.1)
plot(x, dnorm(x), type = "l", col = "red")
lines(x, dcauchy(x), type = "l", col = "blue")
```

Sketch the result.

- The Cauchy and normal distributions have similar shapes, the difference being that the Cauchy is a bit lower near o and more of the probability has been pushed out to the tails. But that probability in the tails makes all the difference!

First let's try the SLLN for normal random variables, repeating the experiment four times

```
replicate(4, mean(rnorm(10^6)))
```

What is the result?

- Based on these results, would you say that the SLLN holds for normal random variables?

- Let's look at the mean of 10^6 Cauchy random variables four times.

```
replicate(4, mean(rcauchy(10^6)))
```

Record the results

- Based on these results, would you say that the SLLN holds for Cauchy random variables?

- Now let's generate 10^6 Cauchy random variables and look at the sample average as the number of samples grows.

```
r <- rcauchy(10^6)
plot(cumsum(r) / 1:length(r), type = "l", col = "blue")
```

Sketch the plot.

- The Cauchy random variable will sometimes jump up and sometimes jump down. That is because its density allows for both large positive and large negative numbers. Use

      ```
      max(r)
      min(r)
      ```

  to find the largest and smallest values for the Cauchy, and report your results.

- **Importance Sampling** In Monte Carlo simulation, we construct a random variable whose mean is equal to the target value.

  Consider the integral

  $$\int_0^1 \exp(-x^{1.5})\, dx.$$

  Find the integral to four significant figures using Wolfram Alpha.

- Now let's build a random variable with this integral as its mean. Start with U which has $\mathbb{1}(x \in [0, 1])$ as its density. Then

  $$\mathbb{E}[\exp(-U^{1.5})] = \int_0^1 \exp(-x^{1.5})\, dx,$$

  so the following should approximate the integral.

      ```
      u1000 <- runif(10^3)
      mean(exp(-u1000^(1.5)))
      ```

  Try this four times (to see what kind of variation in your answers you get) and report the results.

- This idea of using a function of the random variable to get an integral is called *importance sampling*. Generally speaking, importance sampling works better when the density of the random variable is as close as possible to the integrand. The uniform density is flat which does not match the shape of $\exp(-x^{1.5})$ at all.

  Something closer would be $\exp(-x)$, but we also want the density to only be positive over $[0, 1]$. Therefore, what we would like is an exponential random variable with rate 1 conditioned to lie in $[0, 1]$.

Recall that an exponential random variable can be found by using

$$T = -\ln(U)$$

If $T \in [0, 1]$, then what does that tell you about U? The probability that $T \sim \mathrm{Exp}(1)$ falls

into $[0, 1]$ is $1 - \exp(-1)$. Let $W = [T | T \in [0, 1]]$ Then

$$f_W(s) = \frac{\exp(-s)}{1 - \exp(-1)} \mathbb{1}(s \in [0, 1]).$$

To make the expectation match the integral then, we use

$$I = \int_{s \in \mathbb{R}} (1 - \exp(-1)) \exp(-s^{1.5} + s) \frac{\exp(-s)}{1 - \exp(-1)} \mathbb{1}(s \in [0, 1]) \, ds$$
$$= \mathbb{E}[(1 - \exp(-1)) \exp(-W^{1.5} + W)].$$

Let's put this all together and create our new random variables.

```
u_is <- runif(10^3, min = exp(-1), max = 1)
w_is <- -log(u_is)
plot(density(w_is))
x_is <- seq(0, 1, by = 0.01)
lines(x_is, exp(-x_is) / (1 - exp(-1)), type = "l", col = "red"
    )
```

Sketch the result.

- Now let's put everything together.

```
mean((1 - exp(-1)) * exp(-w_is^(1.5) + w_is))
```

Repeat this entire process four times and record the results to get an idea of the spread in the estimate.

- Did you see more or less variation than when you used the uniform over $[0, 1]$ variables?

- To see why this has less variation, let's plot the values that $(1 - \exp(-1)) \exp(-W^{1.5} + W)$ can take on.

```
y_is <- (1 - exp(-1)) * exp(-x_is^1.5 + x_is)
plot(x_is, y_is, type = "l", col = "blue")
```

Sketch the result

- Let's look at the maximum relative error that can occur. Try

```
max(y_is) / min(y_is) - 1
```

What is the result?

- Now let's try the same thing for the estimate using uniforms.

```
y2 <- exp(-x_is^1.5)
max(y2) / min(y2) - 1
```

Report the maximum relative error for the uniforms.

Joint densities

Summary In this lab we will plot points directly drawn from a distribution in order to visualize the density in one and two dimensions. You will also learn how to use an *auxiliary random variable* in order to create draws from different densities.

First Part

- Recall, that f_X is the density of X, if for all A,

$$\mathbb{P}(X \in A) = \int_{x \in A} f_X(x) \, d\mu = \int_x \mathbb{1}(x \in A) f_X(x) \, dx.$$

So whenever the density is large, we expect more points in that area, and when the density is small, we expect fewer points in that region.

The normal density $f_s = \tau^{-1/2} \exp(-s^2/2)$ is large in the middle value of 0 and smaller in the tails, so we expect draws to be concentrated towards the middle. Try

```
x <- rnorm(20)
stripchart(x)
```

Give a rough sketch of the result.

- Now try the same for 100 points and sketch the result.

- The Cauchy distribution has density $2\tau^{-1}(1 + s^2)^{-1}$. It is also big in the middle but has much larger tails. This means it is much more likely to have *outliers* that are very big or very small. Try

```
x1 <- rcauchy(20)
stripchart(x1)
```

and sketch the results.

- Now let's go back to uniforms over $[0, 1]$. Try

```
x2 <- runif(1000)
plot(density(x2))
```

Sketch the plot.

- Now we will use an *auxiliary* variable to draw samples from the density

$$f_X(s) \propto s(1 - s)\mathbb{1}(s \in [0, 1]).$$

Sketch this function, recalling that you can graph functions in Wolfram Alpha using the `plot` command.

- Note that the largest this density can be is $1/4$ as $s = 1/2$. So my auxiliary random variable will also be uniform from 0 to $1/4$. Try the following

```
y2 <- 0.25 * runif(length(x2)))
plot(x2, y2)
```

This gives points that are uniform over $[0, 1] \times [0, 1/4]$. Now let us restrict to points where $y < x(1 - x)$.

```
keep2 <- which(y2 < x2 * (1 - x2))
plot(x2[keep2], y2[keep2])
```

Sketch this region that the points (x, y) are uniform over.

- Because we started with `(x, y)` uniform over $[0, 1] \times [0, 1/4]$, and kept the points that fell into $A = \{(x, y) : x \in [0, 1], y \in [0, x(1 - x)]\}$, then `(x2[keep2], y2[keep2])` are uniform over A.

The area of $[0, 1] \times [0, 1/4]$ is $1/4$. So the percentage of points from $[0, 1] \times [0, 1/4]$ that fell into the region A times $1/4$ gives an estimate of the area of A.

```
0.25 * length(keep2) / length(x2)
```

and report the result.

- Now find $\int_{s \in \mathbb{R}} s(1-s)\mathbb{1}(s \in [0,1])\ ds$ exactly using Wolfram Alpha and compare to your estimate above.

- In two and higher dimensions we also have densities. We can again take random samples to get an idea of what various densities look like in practice. For instance, try

```
x3 <- runif(100)
y3 <- runif(100)
plot(x3, y3)
```

and sketch the result.

- Now let's create a set of values that have the density

$$f_{(X,Y)}(x,y) \propto (x+2y)\mathbb{1}(\{(x,y), x \in [0,1], y \in [0,1]).$$

Again we will use an auxiliary random variable z to accomplish this. The largest the function can be is 3, and so z will be uniform over $[0,3]$.

```
z3 <- 3 * runif(length(x3))
keep3 <- which(x3 + 2 * y3 < z3)
plot(x3[keep3], y3[keep3])
```

For x3 values in $[0,1]$ and y3 values in $[0,1]$, $x_3 + 2y_3 \in [0,3]$, which is why z3 is chosen to be uniform from 0 to 3. The code above is more likely to keep (x,y) values with large y and x values.

- Let's create a larger set of draws from this distribution.

```
n <- 10^6
x4 <- runif(n)
y4 <- runif(n)
z4 <- 3 * runif(n)
keep4 <- which(z4 < x4 + 2 * y4)
```

Now let's estimate the normalizing constant for the density by multiplying the volume of $[0,1] \times [0,1] \times [0,3]$ by the percentage of (x,y,z) that we kept.

```
3 * length(keep4) / n
```

Report your result. (Remember, the density is *divided* by this number to get a normalized density.)

- Now find the normalizing constant exactly using Wolfram Alpha and compare to your estimate.

- Let's check for independence.

```
print(sum(x4[keep4] < 0.5) / length(keep4))
print(sum(y4[keep4] < 0.5) / length(keep4))
print(sum((x4[keep4] < 0.5) & (y4[keep4] < 0.5)) / length(keep4
    ))
```

  Report your estimates.

- Would you say that $\mathbb{P}(X < 1/2)\mathbb{P}(Y < 1/2) = \mathbb{P}(X < 1/2, Y < 1/2)$?

- Find these values exactly using Wolfram Alpha and compare to your estimates. (Remember to include the normalizing constant that you found in the density!)

Second Part

- The correlation

$$\mathrm{Cor}(X, Y) = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\mathrm{SD}(X)\,\mathrm{SD}(Y)}$$

  will be 0 for random variables that are independent. If X and Y are independent, then $\mathbb{E}[X|Y] = \mathbb{E}[X]$, and

$$\mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[XY|Y]] = \mathbb{E}[Y\mathbb{E}[X|Y]] = \mathbb{E}[Y\mathbb{E}[X]] = \mathbb{E}[X]\mathbb{E}[Y].$$

  Try this with

```
z5 <- rnorm(10^6)
z6 <- rnorm(10^6)
mean(z5 * z6) - mean(z5) * mean(z6)
```

  Report your result.

- We can form more random variables that *are* correlated by taking linear combinations of $z5$ and $z6$. Try

```
z7 <- z5 + 2 * z6
z8 <- z5 - z6
```

  Note that if you add or subtract normal random variables, the result is still a normal random variable! Test this with

```
plot(density(z7))
plot(density(z8))
```

  and sketch the results.

- Now $z7$ is not determined completely by $z5$ because of the randomness in $z6$, but as $z5$ grows, so does $z7$ on average. The ratio of growth is one-to-one (also written 1-1), for every unit rise in $z5$, the average value of $z7$ grows by the same amount. Check this by estimating the covariance between $z7$ and $z5$ with

```
mean(z7 * z5) - mean(z7) * mean(z5)
```

  Report your result.

- When $z6$ goes up, the average of $z7$ grows by twice the change in $z6$. So now let us look at their covariance.

```
mean(z7 * z6) - mean(z7) * mean(z6)
```

Report your result.

- The standard statistical estimate for covariance is actually slightly different than what we have been using. Try

```
cov(z7, z6)
```

How does the result compare to the estimate we made earlier?

- Of course, covariance is symmetric in the variables. Try

```
cov(z6, z7)
```

Did you obtain the same estimate?

- To move from covariance to correlation, we must divide by the standard deviations of the variables. Try

```
cov(z7, z6) / sd(z6) / sd(z7)
cov(z7, z5) / sd(z5) / sd(z7)
```

and report the results.

- Note that since z7 is more strongly correlated with z6 than z5, the estimate of the correlation is much stronger. R has a shorthand command for finding correlation just like it has shorthand for estimating the covariance. Try

```
cor(z6, z7)
cor(z5, z7)
```

to estimate the correlation directly. Does this match your earlier estimates?

- Now see what positively correlated draws look like when they are plotted. Try

```
plot(z5[1:100], z7[1:100])
```

and sketch the result.

- The covariance of `z6` and `z7` is twice as great. Try

    ```
    plot(z6[1:100], z7[1:100])
    ```

    and sketch the result.

- What is the difference between this plot and the previous one?

- Next we turn our attention to random variables that are negatively correlated.

    ```
    cov(z5, z8)
    cor(z6, z8)
    ```

    Report your results.

- Try sketching a plot of the first 100 points of negatively correlated normals.

    ```
    plot(z6[1:100], z8[1:100])
    ```

    and sketch the result.

- Recall that correlation is symmetric, so flipping the variables around yields similar results. Try

    ```
    plot(z8[1:100], z6[1:100])
    ```

    and sketch the result.

Chapter 39

Transforming random variables

Summary

- Rotating points in $\mathbb{R}^2$.

- Rotational invariance of normal distribution.

- Functions in R.

First Part   First we will learn how to rotate points in the x, y plane.

- Given a vector of points

$$\begin{pmatrix} x \\ y \end{pmatrix},$$

the *rotation matrix $R(t)$* is

$$R(t) = \begin{pmatrix} \cos(t) & -\sin(t) \\ \sin(t) & \cos(t) \end{pmatrix}$$

Notice that the Jacobian of the matrix is 1: that means it does not alter the areas of regions. Then for a point $v \in \mathbb{R}^2$,

$$R(t)v$$

is the point v rotated counterclockwise by angle t.

In R, the `cos` and `sin` functions take an argument in radian measure. Multiply by $\tau/360 = 2\pi/360$ to convert from degrees to radians. For instance, an angle of 20 degrees can be converted to radians in R with

```
t <- 20 * 2 * pi / 360
```

Write down the rotation matrix for t equal to 20 degrees.

- Now multiply this matrix by the column vector to find the point that is $(0.5, 0.5)$ rotated by 20 degrees counterclockwise. Plot the original point and the rotated point.

- In R, one way we can create matrices using the `rbind` (which stands for row bind) function, or with the `cbind` (which stands for column bind). Remember that we can create vectors with c. Then we can bind these vectors together as follows.

```
t <- 20 * 2 * pi / 360
R <- rbind(c(cos(t), -sin(t)), c(sin(t), cos(t)))
x <- cbind(c(0.5, 0.5))
```

Report the values of `R` and x.

- The matrix multiplication operator in R is `%*%`. Try

```
R %*% x
```

and report the result.

- Now let's look how the rotational symmetry of standard normal random variables works. Try

```
z1 <- rnorm(10^6)
z2 <- rnorm(10^6)
plot(density(z1), col = "gold")
lines(density(z2), col = "blue")
```

and sketch the results.

- Now let's rotate all of these points 20 degrees to the left.

```
A <- rbind(z1, z2)
B <- R %*% A
```

Let's plot the first five points in the original and rotated points.

```
plot(A[1,1:5], A[2,1:5], xlim = c(-3, 3), ylim = c(-3, 3))
points(B[1,1:5], B[2,1:5], col = "blue")
```

- Now let's look at the densities of the rotated points. The notation `B[1,]` gives the first row of the matrix B, while `B[2,]` gives the second row.

```
plot(density(B[1,]), col = "gold")
lines(density(B[2,]), col = "blue")
```

Sketch the result.

- **Sidebar: functions in R**   Up until now, we have been restricting ourselves to using the built-in functions in R, but in fact, we can create our own functions!

In creating our rotation matrix, we used a value t that can change. In R, the best way to do this is with a function. Try the following.

```
rotate <- function(deg) {t <- deg * 2 * pi / 360; return(rbind(
    c(cos(t), -sin(t)), c(sin(t), cos(t)))) }
rotate(20)
```

What the rotate function does is first calculate the value of `t` given the argument `deg`, and then returns the appropriate matrix to the user. It executes these two lines every time the `rotate` function is called with an argument.

The `rotate(20)` command should return your rotation matrix for 20 degrees from earlier. Using the function `rotate`, calculate the rotation matrix for 110 degrees.

- Check that your new rotation matrix does what it is supposed to by calculating $(1, 0)$ and $(0, 1)$ rotated 110 degrees.

- We can use a function to turn Cartesian coordinates into polar coordinates. Try entering the following function.

```
car2pol <- function(v) {
    x <- v[1]
    y <- v[2]
    r <- sqrt(x^2 + y^2)
    t <- atan2(y, x)
    return(c(r, t))
}
```

Now `car2pol(c(1, 2))` returns the polar coordinates for the point $(1, 2)$. Find the polar coordinates for the point $(-2, 6)$.

- Now let's find the polar coordinates when we draw two iid standard normal random variables. We will repeat the experiment 1000 times with the `replicate` command.

```
results <- replicate(1000, car2pol(rnorm(2)))
```

Check to see the dimensions of `results` with the `dim` command.

```
dim(results)
```

What are the dimensions of `results`?

- The first coordinate is the R value, while the second coordinate is the θ value.

```
plot(density(results[2,]))
```

Sketch the result. (Note that here the θ values are forced to lie in $[-\tau/2, \tau/2]$.)

- Now let's sketch the density of the distances from the origin, and compare to the Rayleigh density.

```
plot(density(results[1,]))
x <- seq(0, 4, by = 0.1)
lines(x, x * exp(-x^2 / 2), col = "blue", lwd = 2)
plot(density(results[1,]))
```

Sketch the result.

Second Part

- If we draw $Z_1, \ldots, Z_d$ iid $\mathsf{N}(0, 1)$, then the points have a direction that is uniform over all possible directions. So we can create a function that generates uniformly from the surface of a sphere as follows.

```
sphere.surface <- function(d) {
    v <- rnorm(d)
    return(v / sqrt(sum(v^2)))
}
```

This is called *Haar measure* on the surface of a sphere. Let's generate 10^6 of these, and plot the first 100 of these.

```
results2 <- replicate(10^6, sphere.surface(2))
plot(results2[1, 1:100], results2[2, 1:100])
```

Sketch the result.

- Now let's plot the density of the x values

```
plot(density(results2[1,])))
```

Sketch the result

- Now let's move up to the surface of a 3-dimensional sphere.

```
results3 <- replicate(10^6, sphere.surface(3))
plot(density(results3[1,])))
```

Conjecture what the distribution of X is for (X, Y, Z) uniform over the surface of the 3-sphere.

- By symmetry, X, Y, and Z will have the same distribution. But because they lie on the surface of the sphere, $X^2 + Y^2 + Z^2 = 1$. So they are not independent. However, given X and Y, for instance, Z is equally likely to be $\sqrt{1 - X^2 - Y^2}$ and $-\sqrt{1 - X^2 - Y^2}$. So $\mathbb{E}[Z|X, Y] = 0$ for all X and Y, which means that (X, Y), (X, Z), and (Y, Z) are all uncorrelated. Check this with

```
cor(results3[1,], results3[2,])
cor(results3[1,], results3[3,])
cor(results3[2,], results3[3,])
```

and report your results.

- Now generate 10^6 points uniformly from the surface of the 4 dimensional sphere and plot the kernel density estimate of the first coordinate.

When we go higher in dimension, the first coordinate is more and more likely to be close to 0. In general, when you are on the surface of a high dimensional sphere, you are very likely to have all the coordinates very near 0. Of course, they are dependent, so if $X_1, \ldots, X_{d-1}$ are all 0, then X_d must equal 1.

- Now suppose (U_1, U_2) are iid uniform over $[0, 1]$. Let's generate 1000 of these and sketch the plot.

  ```
  u1 <- runif(1000)
  u2 <- runif(1000)
  plot(u1, u2)
  ```

- Now let's transform them using $W_1 = U_1 - U_2$, $W_2 = U_1 + U_2$. Then the Jacobian of this transform is a constant, so the transformed variables are still uniform. Test this with

  ```
  plot(u1 - u2, u1 + u2)
  ```

  Sketch the result.

- These provide another example of random variables that are dependent (the values that W_2 can take on depend on the value of W_1) but which are uncorrelated. Check this with

  ```
  cor(u1 - u2, u1 + u2)
  ```

  Report your result.

- **Moment generating functions** Recall that for a random variable, X, the moment generating function is $\mathrm{mgf}_X(t) = \mathbb{E}[\exp(tX)]$ whenever this is finite. For instance, if $X \sim \mathsf{Unif}(\{1, 2, 3\})$, then

$$\mathrm{mgf}_X(t) = (1/3)[e^t + e^{2t} + e^{3t}].$$

Then for $X_1, \ldots, X_n$ iid $\mathsf{Unif}(\{1, 2, 3\})$,

$$\mathrm{mgf}_{X_1 + \cdots + X_n}(t) = \mathrm{mgf}_X(t)^n.$$

Find

$$\mathbb{P}(X_1 + \cdots + X_{20} = 40)$$

using the moment generating function and Wolfram Alpha.

- Using Wolfram Alpha, find the moment generating function of $T \sim \mathsf{Exp}(1)$, which has density $f_T(s) = \exp(-s)\mathbb{1}(s \geq 0)$. Recall that

$$\mathrm{mgf}_T(t) = \mathbb{E}[\exp(tT)] = \int_s \exp(ts) f_T(s) \, ds.$$

Report the result.

- What is the moment generating function of $T_1 + T_2 + \cdots + T_n$ where T_i are iid $\mathsf{Exp}(1)$?

- Find the moment generating function of G with density $(s^4/24) \exp(-s)\mathbb{1}(s \geq 0)$.

- Do G and $T_1 + \cdots + T_5$ have the same density?

Chapter 40

Discrete Distributions

Summary This lab will introduce you to the common discrete distributions.

First Part

- Start with the simplest discrete distribution, the *Bernoulli* or *indicator* random variables. For $U \sim \mathsf{Unif}([0,1])$, $B = \mathbb{1}(U \le p)$ is a Bernoulli random variable with parameter p, and we write $B \sim \mathsf{Bern}(p)$. We can create such variables by using the `as.integer` function together with logical statements involving uniforms. Try the following.

  ```
  u <- runif(10)
  print(u <= 0.3)
  print(as.integer(u <= 0.3)).
  ```

  Report your results.

- Now let's do many of these and form a histogram.

  ```
  b <- as.integer(runif(10^6) <= 0.3)
  hist(b)
  ```

  Sketch the histogram.

- When $B_1, B_2, B_3, \ldots$ are an iid sequence of $\mathsf{Bern}(p)$ random variables, we call the sequence a *Bernoulli process*. A *Bernoulli process* can be used to create three other important distributions. The first is the *binomial* distribution with parameters n and p (write $N \sim \mathsf{Bin}(n,p)$. This is the sum of the first n random variables. For instance, if $n = 6$ and $p = 0.3$, then you can generate one Binomial by using

  ```
  b <- sum(runif(6) <= 0.3)
  ```

  Let's try creating 10 iid $\mathsf{Bin}(6, 0.3)$ random variables.

```
        b <- replicate(10, sum(runif(6) <= 0.3))
```

Report your results.

- Try generating 10^6 iid Bin$(6, 0.3)$ and sketch the histogram of the results.

- As with all built in distributions, R can generate random binomials by placing an `r` in from of the distribution name. Try

```
        c <- rbinom(10^6, size = 6, prob = 0.3)
```

and sketch a histogram of the results. .

- The next distribution that comes from the Bernoulli process is the *geometric.* Here we examine $G = \inf\{i : B_i = 1\}$. So for instance, if the Bernoulli process starts with $0, 1, 1, 0, 0, 1$, then $\{i : B_i = 1\} = \{2, 3, 6\}$. Suppose the Bernoulli process starts with $1, 0, 0, 0, 1, 1$. What are the first few elements of $\{i : B_i = 1\}$?

- The *infimum* of a set is the greatest lower bound. For a subset of integers, it is kind of like the minimum. So $\inf\{7, 4, 6\} = 4$, and $\inf\{5, 6, 7, \ldots\} = 5$. One difference between the infimum of a set and the minimum is that $\inf(\emptyset) = \infty$. With that in mind, what is $\inf\{2, 4, 6, 8, \ldots\}$?

- Now we combine these ideas. A geometric random variable is the smallest value of i such that $B_i = 1$. Formally, $G = \inf\{i : B_i = 1\}$. So if the sequence starts $0, 0, 0, 1, 1, 0$, then $\{i : B_i = 1\} = \{4, 5, \ldots\}$, and $G = \inf\{i : B_i = 1\} = 4$. Generate a Bernoulli process and find $\{i : B_i = 1\}$ using the `which` function.

```
        b <- as.integer(runif(10) <= 0.3)
        print(b)
        which(b == 1)
```

Report the result.

- Now let's generate some geometric random variables using this approach.

```
g <- replicate(10^4, min(which(runif(100) <= 0.3)))
hist(g, breaks=0:max(g))
```

Sketch the histogram of g.

- As with binomials, R has built in functions for drawing random geometric random variables. Try

```
g <- rgeom(10^4, prob = 0.3)) + 1
hist(g, breaks=0:max(g))
```

and sketch the result. Notice that we had to add 1 to the random geometrics generated by R. That is because some authors define a geometric as $\inf\{i : B_i = 1\} - 1$. While both definitions are valid, in this course we will always use the $\inf\{i : B_i = 1\}$ version.

- A slightly different way to write the definition of geometric random variable is to say

$$G = \inf\left\{ i : \sum_{j=1}^{i} B_i = 1 \right\}.$$

Then we can extend this definition to *negative binomial distribution* with parameters k and p, by letting

$$G_k = \inf\left\{ i : \sum_{j=1}^{i} B_i = k \right\}$$

Note that if $B_1, B_2, \ldots$ are iid $\text{Bern}(p)$ and $G_1, G_2, \ldots$ are iid $\text{Geo}(p)$, and

$$N_n = B_1 + \cdots + B_n$$
$$G_k = G_1 + \cdots + G_k,$$

Then N_n is binomial with parameters n and p, and G_k is negative binomial with parameters k and p. Put another way, N_n is the random number of successes in a fixed number of trials n, and G_k is the random number of trials needed to obtain a fixed number of successes k.

Try $k = 1$, which gives a geometric random variable

```
g <- rnbinom(10^4, size = 1, prob = 0.3) + 1
hist(g, breaks=0:max(g))
```

Sketch the result.

- Next try adding together multiple random variables

```r
g <- rnbinom(10^4, size = 4, prob = 0.3) + 1
hist(g, breaks=0:max(g))
```

Sketch the result.

Second Part

- **Poisson** There are multiple ways to view the Poisson random variable. First, consider it as the limiting distribution of a binomial random variable, where np equals a constant μ, while n goes to infinity (which makes p go to zero.) For instance, let $\mu = 0$. Try

  ```
  n <- rbinom(10^6, size = 10, prob = 0.3)
  hist(n, breaks=0:max(n))
  ```

  Sketch the result. Notice that the most common value is $3 = (10)(0.3)$.

- Now sketch the histogram of 10^6 draws from $\mathsf{Bin}(100, 0.03)$. Note that $np = (100)(0.03) = 3$ remains the same.

- Now sketch the histogram of 10^6 draws from $\mathsf{Bin}(10000, 0.0003)$. Again $np = (10000)(0.0003) = 3$ remains the same.

- At this point the distribution is very close to a *Poisson* distribution with parameter $\mu = 3$ (write $N \sim \mathsf{Pois}(\mu)$). Try the following.

  ```
  n <- rbinom(10^6, size = 10, prob = 0.3)
  hist(n, breaks=0:max(n))
  ```

  Sketch the result.

- Poisson random variables are useful for modeling phenomena where there are lots of experiments, each with a low chance of success. For instance, the number of defects in a large assembly line with very low chance of failure, or the number of typos in words in a book, typically follow a Poisson distribution.

  Let's calculate the probability that $N = 0$ for $N \sim \mathsf{Pois}(\mu)$. This is the limit as $n \to \infty$ of $\mathbb{P}(N_n = 0)$ where $N_n = \mathsf{Bin}(n, \mu/n)$. So

  $$\mathbb{P}(N = 0) = \lim_{n \to \infty} \binom{n}{0} (\mu/n)^0 (1 - \mu/n)^n.$$

Find this limit. Recall that the exponential function $\exp$ satisfies

$$\exp(x) = \lim_{n \to \infty} (1 + x/n)^n.$$

- Now try

```
n <- rbinom(10^6, size = 10000, prob = 0.0003)
mean(n == 0)
```

  to verify your calculation above.

- To study the rest of the distribution, consider for a binomial random variable,

$$
\begin{aligned}
\frac{\mathbb{P}(N_n = i + 1)}{\mathbb{P}(N_n = i)} &= \frac{\binom{n}{i+1} p^{i+1} (1-p)^{n-(i+1)}}{\binom{n}{i} p^i (1-p)^{n-i}} \\
&= \frac{n!}{(i+1)!(n-i-1)!} \cdot \frac{i!(n-i)!}{n!} \cdot \frac{p}{1-p} \\
&= \frac{n-i}{i+1} \cdot \frac{p}{1-p}.
\end{aligned}
$$

  For $X \sim \mathrm{Bin}(100, 0.6)$, what is $\mathbb{P}(X = 51)/\mathbb{P}(X = 50)$?

- Now estimate this value with

```
n <- rbinom(10^6, size = 100, prob = 0.6)
mean(n == 51) / mean(n == 50)
```

- Note that for $\mu = np$,

$$
\frac{\mathbb{P}(N_n = i + 1)}{\mathbb{P}(N_n = i)} = \frac{n-i}{i+1} \cdot \frac{p}{1-p} = \frac{\mu - i(\mu/n)}{(i+1)(1 - \mu/n)}.
$$

  Go ahead and take the limit of this right hand side as $n \to \infty$, keeping μ as a constant.

- Use your limit to find

$$\frac{\mathbb{P}(N=2)}{\mathbb{P}(N=1)}$$

  for $N \sim \text{Pois}(3)$.

- Verify your estimate with

```
n <- rpois(10^6, lambda = 3)
mean(n == 2) / mean(n == 1)
```

  Report your result.

- So at this point, we know how to find $\mathbb{P}(N=0)$, and how to find $\mathbb{P}(N=i+1)/\mathbb{P}(N=i)$. Using

$$\mathbb{P}(N=3) = \mathbb{P}(N=0) \cdot \frac{\mathbb{P}(N=1)}{\mathbb{P}(N=0)} \cdot \frac{\mathbb{P}(N=2)}{\mathbb{P}(N=1)} \cdot \frac{\mathbb{P}(N=3)}{\mathbb{P}(N=2)},$$

  find $\mathbb{P}(N=3)$ for $N \sim \text{Pois}(3)$. By the way, this is known as a telescoping product.

- **Poisson process** The second way to think of a Poisson random variable is as coming from a Poisson point process. In this case, if we generate a standard PPP (so it has rate 1) on $[0, \infty)$, then
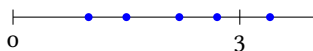
  the number of points that fall into the interval $[0, 3)$ will have a Poisson distribution. In a standard PPP, the distance until the first point is an exponential random variable with mean 1.

  In general, $\text{Pois}(\mu)$ is the distribution of the number of points that fall in the interval $[0, \mu)$.

  The following code generates the first point in the process T_1. If T_1 is above μ, then 0 points fell in $[0, \mu)$. Otherwise, we generate the number that fell in $[T_1, \mu)$, or equivalently $[0, \mu - T_1)$, and add that to the first point.

```
rpoisson <- function(mu = 3) {
    t1 <- rexp(1)
    if (t1 > mu) return(0)
    else {x <- rpoisson(mu - t1); return(1 + x)}
}
```

Replicate 10^5 iid Pois(3) draws with the function and plot the histogram.

- Having a function call itself during the execution is called *recursion*, and when recursion is used as part of a simulation algorithm it becomes a *perfect simulation* algorithm.

  Generally speaking, recursive algorithms tend to be slower than nonrecursive ones. You can test this out with the **system.time** function, which times how long commands take to execute. Try

  ```
  system.time(results <- replicate(10^5, rpoisson(3)))
  system.time(results <- rpois(10^5, 3))
  ```

  and report your results.

Chapter 41

Central Limit Theorem

First Part

- Consider adding up a bunch of random variables. For instance, if we add up 10 iid uniforms over $[0, 1]$, we end up with a random variable S that has mean 5 and variance $10/12$.

```
results <- replicate(1000, sum(runif(10)))
mean(results)
var(results)
```

Report your results.

- Look at a density plot for this sum of variables.

```
plot(density(results))
```

Sketch the kernel density estimate of S.

- The *Central Limit Theorem* says that as more and more random variables are added together, the result looks more and more like the density of a normal random variable. It helps to standardize the random variable by subtracting off the mean and dividing by the standard deviation. The result will then have mean 0 and standard deviation 1, and will look similar to the density of a standard normal random variable.

```
plot(density((results - 5) / sqrt(10 / 12)))
x <- seq(-2, 2, by = 0.1)
lines(x, dnorm(x), lwd = 2, col = "blue")
```

Sketch the result.

- Consider the same procedure but this time with exponential rate 2 random variables.

```
results2 <- replicate(1000, sum(rexp(10, rate = 2)))
plot(density((results - 10 * 1 / 2) / sqrt(10 * (1 / 2)^2)))
lines(x, dnorm(x), lwd = 2, col = "blue")
```

Sketch the result.

- Notice that the result is a bit off compared to adding 10 uniforms. That is because like the normal distribution, the uniform distribution is symmetric around its mean, while the exponential distribution is not. We say that the exponential distribution is *skewed*. In general, the skewness of a random variable X with mean μ and standard deviation σ is

$$\text{skew}(X) = \mathbb{E}\left[\left(\frac{X - \mu}{\sigma}\right)^3\right].$$

Estimate the skewness for $S = U_1 + \cdots + U_{10}$ where the U_i are iid $\mathsf{Unif}([0, 1])$ and $R = T_1 + \cdots + T_{10}$ are iid $\mathsf{Exp}(2)$ with

```
mean(((results - 10 * 1 / 2) / sqrt(10 / 12))^3)
mean(((results2 - 10 * 1 / 2)/sqrt(10 * (1 / 2)^2))^3)
```

Report your estimates for the uniforms and exponentials.

Second part

- The CLT also holds for discrete random variables. Consider the rolls of a six sided die.

```
results3 <- sample(1:6, size = 1000, replace = TRUE)
hist(results3, breaks = 0:6)
```

- Now consider the rolls of the sum of a pair of dice.

```
results4 <- replicate(1000, sum(sample(1:6, size = 2, replace =
    TRUE)))
hist(results4, breaks = 0:12)
```

- Now sum together 20 dice.

```
results5 <- replicate(1000, sum(sample(1:6, size = 20, replace
    = TRUE)))
hist(results5, breaks = 0:120)
```

Sketch the resulting plot.

- This will continue, with more and more dice rolls resulting in a better fit. Let X be the sum of 20 iid fair six-sided dice rolls. Now let's look at how something like $\mathbb{P}(X \leq 60)$. Subtracting off the mean and dividing by the standard deviation gives something approximately normal.

  The mean formula for X is $(20)(1+6)/2 = 70$, and the variance is $(20)(6-1)(6+1)/12 = 700/12 = 58.333\ldots$.

  Estimate the mean and variance of `results5` with `mean` and `var` an report your results.

- So

$$\mathbb{P}(X \leq 60) = \mathbb{P}\left(\frac{X - 70}{\sqrt{700/12}} \leq \frac{60 - 70}{\sqrt{700/12}}\right)$$
$$\approx \mathbb{P}(Z \leq -1.309\ldots).$$

  This can be found with `pnorm((60 - 70) / sqrt(700 / 12))`. Report this estimate of the probability.

- The probability can also be estimated from the sample die rolls.

  ```
  mean(results5 <= 60)
  ```

  Report this Monte Carlo estimate.

- The last estimate is in some way better than the CLT estimate. The CLT estimate is a fixed thing, it does not get any better. Increasing the number of samples from 1000 to a larger number will give a better estimate. Try estimating the probability using 100000 samples and report your result.

Part III

MATHEMATICS NEEDED FOR PROBABILITY

Chapter 42

Logic notation

Summary The symbol $\forall$ means *for all* or *for every*. The symbol $\exists$ means *there exists* or *there is at least one*. The symbol $\wedge$ means *logical and*, the symbol $\vee$ means *logical or*, and $\neg$ means *logical negation*.

42.1   True and false

Much of this text is about determining the probability that events are true or false. Start with *logical statements*.

> **Definition 77**
> A **logical statement** is a statement that is either true or false.

For instance 3 is greater than 1 is true and 7 is greater than 10 is false. Once we introduce variables to the mix, the truth or falseness of a statement depends on the actual value of the variable. For instance $x > 1$ is true when $x = 2$ but false when $x = 1/2$.

42.2   For all and for every

In the last example, there exists some value of x such that $x > 1$ is true. In logic notation, there exists is represented by $\exists$. So the statement

$$(\exists \text{ real numbers } x)(x > 1)$$

is true because there does exist at least one real value of x in the set of reals that makes the statement at the end true.

On the other hand, some statements are true no matter what the value of the variable is. In

$$(\forall \text{ real numbers } x)(x^2 \geq 0),$$

the symbol $\forall$ means *for all*, or *for every*, and means that the statement at the end ($x^2 \geq 0$) is true for every possible of choice for the variable x as a real number.

Things really get interesting once we start to combine the two. Consider the logical statement

$$(\forall x \in \mathbb{R})(\exists y \in \mathbb{R})(x + y \geq 10).$$

This can be read as follows:

For every real number x, there exists a real number y such that $x + y$ is at least 10.

This statement is true. However, the statement

$$(\exists y \in \mathbb{R})(\forall x \in \mathbb{R})(x + y \geq 10)$$

is false. Order matters in these logical statements!

In logic, "there exists" and "for all" are known as *universal quantifiers*. In this course, since we are usually dealing with real numbers, we will use the shorthand $(\forall x)$ to mean $(\forall$ real numbers $x)$.

42.3   Proving logical statements

How can I be sure that $(\forall x)(\exists y)(x + y \geq 10)$ is true? I can use a *proof* to show the result. In order to prove such a statement, we begin by dealing with the quantifiers at the start of the proof.

For instance, let us try to prove $(\forall x)(x^2 \geq 0)$. The very first line of the proof comes from the fact that x is using a universal quantifier. When I see a $\forall$ statement I have to *instantiate the quantifier*. That means that in my first line of the proof, I let the variable be an arbitrary value.

So my first line of the proof that $(\forall x)(x^2 \geq 0)$. is

Let x be a real number.

Not very exciting! However, by doing this we are signaling that the value of x has been chosen, and so now we can talk about the value of x as a fixed quantity.

For instance, because x is a fixed quantity, we know that it is either greater than or equal to 0, or it is less than or equal to 0. Since the product of two positive numbers and two negative numbers is nonnegative, this gives us our proof.

Written out completely, our proof is as follows.

Fact $(\forall x)(x^2 \geq 0)$

Proof   Let x be a real number.
Suppose that $x \geq 0$, then $x^2 = x \cdot x \geq 0$.
Suppose that $x \leq 0$, then $x^2 = x \cdot x \geq 0$.
Either way, $x^2 \geq 0$, and the proof is complete. $\square$

A couple of remarks about the proof.

- In mathematical writing, we use complete sentences. Often when thinking about the proof we think in sentence fragments, but the final proof should always use complete sentences.

- We ended the proof with the symbol $\square$, which indicates that the proof is complete. Another common way to end a proof is with QED, which stands for the Latin phrase *quad erat demonstrandum* which means "what was to be demonstrated." Most areas of mathematics have removed Latin phrases from their everyday use, and so the simple symbol $\square$ is preferred today.

- We used "suppose" here to break the possible values of x into different cases. Another common way to say this is to use the word "case". So "Case 1: $x \geq 0$" and "Case 2: $x \leq 0$" could also have been used.

When instantiating $(\forall x)$, we have to allow for any value of x. When instantiating $(\exists x)$, we get to pick the value of the variable x. This can make these types of proofs very simple.

Fact $(\exists x)(x + 5 \geq 10)$

Proof Let $x = 5$.
Then $x + 5 = 5 + 5 = 10 \geq 10$. $\square$

Now let us return to $(\forall x)(\exists y)(x + y \geq 10)$. Our first line instantiates x

Let x be a real number.

Our next line should instantiate y. Because we have already instantiated x, we can now use x in defining y. Now I want to end with $x + y \geq 10$. That means I want $y \geq 10 - x$. I cannot just say $y \geq 10 - x$ because that is not a number, I need to write y is equal to something. For instance, $y = 10 - x$ works. Hence my proof ends up being as follows.

Fact $(\forall x)(\exists y)(x + y \geq 10)$

Proof Let $x = 5$.
Let $y = 10 - x$
Then $x + y = x + 10 - x = 10 \geq 10$. $\square$

It is important to note that my choice of y here was not unique. For instance,

Proof Let $x = 5$.
Let $y = 14 - x$
Then $x + y = x + 14 - x = 14 \geq 10$. $\square$

is a perfectly valid proof. There are an infinite number of possible proofs for this statement, and you should not worry about obtaining the "best" proof by trying to make y as small as possible. Whatever works for the proof is great!

42.4 Logical and and logical or

The logical and operator is written $\wedge$ and links true an false statements together. The logical and of two logical statements is true only if the statements that it connects are all true. For instance,

$$(3 > 5) \wedge (7 > 5)$$

is false because $(3 > 5)$ is false and at least one false statement is enough to make a logical and false. However,

$$(5 > 3) \wedge (7 > 5) = \mathsf{T}.$$

The logical or operator is written $\vee$ and is true if at least one of the two logical statements linked is true. For example,

$$(3 > 5) \vee (7 > 5)$$

is true because $(7 > 5)$ is true and at least one true statement is enough to make a logical or true.
These can be written using a *truth table*.

> **Definition 78**
> **Logical and** ($\wedge$) and **logical or** ($\vee$) can be defined as follows for logical statements p and q.
>
> | p | q | $p \wedge q$ | $p \vee q$ |
> |-----|-----|--------------|------------|
> | T | T | T | T |
> | T | F | F | T |
> | F | T | F | T |
> | F | F | F | F |

42.5   Negation and proving things false

The negation operator $\neg$ switches true to false and false to true. For example, $(3 > 5)$ is false, but $\neg(3 > 5)$ is true. Similarly, $(\forall x \geq 3)(2x \geq 5)$ is true, but $\neg(\forall x \geq 3)(2x \geq 5)$ is false.

> **Definition 79**
> **Logical negation** ($\neg$) can be defined using the following table.
>
> | p | $\neg p$ |
> |-----|----------|
> | T | F |
> | F | T |

Suppose that I want to prove that a statement p is false. Then we can use the same rules as earlier, we just want to prove that $\neg p$ is true. In order to use this, we need rules for how to negate a logical statement.

For the final statement, this depends on the type of statement. For instance,

$$\neg(x = 5) = (x \neq 5),$$

whereas

$$\neg(x \geq 5) = (x < 5).$$

Consider negating a for all statement. This is saying that not all of the variable values are okay. In other words, there exists a variable value that is not okay. So our rule is negation turns $\forall$ into $\exists$, that is,

$$\neg(\forall p)(q) = (\exists p)(\neg q).$$

The choice of p that makes q false is sometimes called a *counterexample.*

Similarly, when we negate a there exists statement, we are saying that no matter what value we pick, we fail. So negation turns $\exists$ into $\forall$, that is,

$$\neg(\exists p)(q) = (\forall p) \neq (q).$$

Consider our earlier example $(\exists y)(\forall x)(x + y \geq 10)$. I claimed that this was false, but now we have the tools to prove it. Instead of trying to prove it false, try to prove the negation is true.

$$\neg(\exists y)(\forall x)(x + y \geq 10) = (\forall y)\neg(\forall x)(x + y \geq 10)$$
$$= (\forall y)(\exists x)\neg(x + y \geq 10) = (\forall y)(\exists x)(x + y < 10).$$

Fact $(\forall y)(\exists x)(x + y < 10)$

Proof Let $y \in \mathbb{R}$.
Let $x = 9 - y$
Then $x + y = 9 - y + y = 9 < 10$.

42.6 If then statements

A common formulation in logic is "If something is true, then something else is true." This can be written using the if-then operator $\to$, so

$$p \to q$$

means if statement p is true, then q is true. This is also read as p implies q. It turns out that we can write this operator using our previous operators as

$$(p \to q) = (p \wedge q) \vee \neg p$$

In other words, either p is not true, or if it is true, then q had better be true as well. This means that when proving these sorts of statements, we do not need to worry about what happens when p is not true. Our first line of an if-then proof is always "Let p be true."

For instance,

Fact If $a > 2$, then $a^2 > 3$.

Proof Let $a > 2$.
Then $a \cdot a > 2 \cdot 2$ so $a^2 > 4 > 3$. $\square$

Problems

42.1 Prove that $(\exists x)(2x + 3 \geq 10)$.

42.2 Prove that $(\forall y)(y^2 + 1 > 0)$.

42.3 Prove that $(\forall x)(\exists y)(xy \leq 0)$

42.4 Write $\neg(\forall x \in \mathbb{R})(\exists y)(2x + y \geq 4)$ without the negation.

42.5 Prove that if $x > 3$ then $2x > 6$.

42.6 Prove that $(\forall \epsilon > 0)(\exists \delta > 0)(\forall x \in [-\delta, \delta])(x^2 \leq \epsilon)$.

Chapter 43

Sets and Measures

Question of the Day What is a set?

Summary **Sets** are an unordered collection of **elements**. Measures such as **counting measure** and **Lebesgue measure** tell us the size of a set. We can combine two sets A and B to form the **Cartesian product** $A \times B$, where $(a, b) \in A \times B$ if and only if $a \in A$ and $b \in B$. This can be extended to the Cartesian product of an arbitrary number of sets.

The first mathematics that most people learn is the concept of a number. But what is a number really? One way of thinking about numbers is that they represent the size of a set. For instance, if I have a set of objects

$$\{\text{paperclip, pen, stapler}\},$$

then anyone would say that I have three objects. If I have the following fruit:

$$\{\text{apple, orange, banana}\},$$

then I have three objects. The fact that the first set was office supplies and the second set was fruit does not matter.

About a century ago, several mathematicians tried to write the foundations of mathematics using sets.

43.1 Sets

A **set** is a collection of elements where order does not matter. Put curly brackets around your elements to indicate that it is a set. For instance,

$$\{a, b, c\}$$

is the set containing elements a, b, and c. It is the same as the set $\{b, c, a\}$. Write that $a \in \{a, b, c\}$ is true, $b \in \{a, b, c\}$ is true, $c \in \{a, b, c\}$ is true, and all other statements such as $d \in \{a, b, c\}$ is false.

> **Definition 80**
> A mathematical object A is a set if for another mathematical object a, it is possible to determine if the logical statement $a \in A$ is true or false.

As throughout, T will be used for true and F for false.

For a basic set theory, this determination can be done using a Turing machine, but for more advanced mathematics, more powerful oracle computers would be needed. Throughout this text, only the Turing machine approach is needed.

Definition 81

If $(a \in A) = \mathsf{T}$ for set A and object a, say that a is an **element** of the set A.

There is also notation for when an object is *not* an element of A.

Notation 5

If $(a \in A) = \mathsf{F}$, write $a \notin A$.

Definition 82

Suppose that

$$(a \in A) \rightarrow (a \in B)$$

for sets A and B. Then say that A is a **subset** of B, and write $A \subseteq B$.

Note that the subset notation looks a lot like the $\leq$ notation: always face the open end of the $\subseteq$ symbol towards the larger set. For instance,

$$\{a, b\} \subseteq \{a, b, c\}, \quad \{a\} \subseteq \{a, b, c\}, \quad \{b, d\} \nsubseteq \{a, b, c, \}.$$

It helps to have a set with no elements, we call that the *empty set*.

Definition 83

If

$$(\forall a)(\neg(a \in A))$$

say that A is the **empty set**.

Notation 6

The empty set is often denoted $\{\}$ or $\emptyset$.

There are two useful operations we would like to be able to do with sets. First, for a collection of sets, consider elements that are in every single one of the sets. This is the *intersection* of the sets.

Definition 84

The **intersection** of two sets A and B is a set $A \cap B$ such that

$$(x \in A \cap B) = (x \in A) \vee (x \in B).$$

This can be extended to an arbitrary number of sets.

Definition 85

Suppose there is a set S such that for any element $s \in S$, A_s is also a set. Then the **intersection** of the sets are

$$(x \in \cap_{s \in S} A_s) = (\forall s \in S)(x \in A_s).$$

Second, consider which objects are in at least one of a collection of sets. This is the *union* of the sets.

> **Definition 86**
> The **union** of two sets A and B is a set $A \cap B$ such that
> $$(x \in A \cup B) = (x \in A) \wedge (x \in B).$$

This can be extended to an arbitrary number of sets.

> **Definition 87**
> Suppose there is a set S such that for any element $s \in S$, A_s is also a set. Then the **union** of the sets are
> $$(x \in \cup_{s \in S} A_s) = (\exists s \in S)(x \in A_s).$$

Euler diagrams can be used to indicate properties of sets. For instance
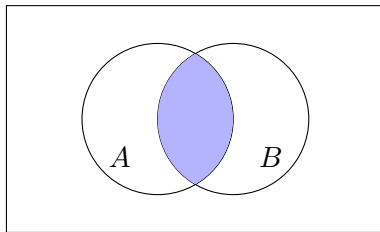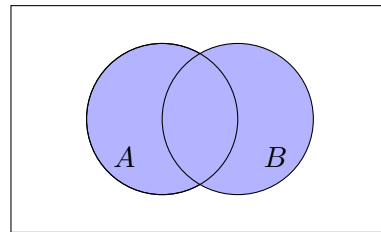
Shaded region is $A \cap B$.   Shaded region is $A \cup B$.

To summarize this notation:

$\emptyset$	the empty set
$\in$	is an element of
$\notin$	is not an element of
$\subseteq$	is a subset of
$\not\subseteq$	is not a subset of
$\cap$	intersection
$\cup$	union

Here is another Euler diagram.
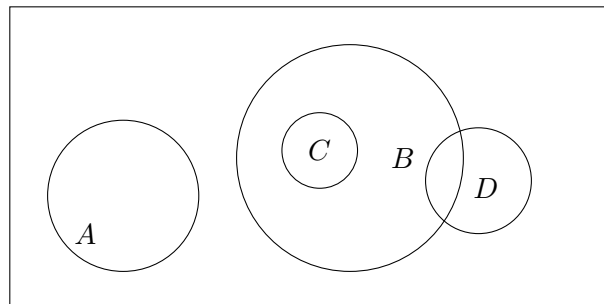
This picture tells us that $A \cap B$, $A \cap C$, $A \cap D$, and $C \cap D$ are all the empty set. It also indicates that $C \subseteq B$.

43.2   Some important sets

Important sets usually are denoted with blackboard boldface letters. The two most important sets of numbers used in this course are the following.

$$\mathbb{R} \quad \text{the real numbers}$$
$$\mathbb{Z} \quad \text{the integers}$$

43.3   Measures

A measure is another way of measuring the size of a set. There are two important measures that will be used constantly throughout this course.

1. Counting measure. This counts the number of objects in a set. For instance, the counting measure of $\{a, b, c\}$ is 3 because it contains three distinct elements. We will write $\#$ for counting measure, so $\#(\{a, b, c\}) = 3$. The counting measure of the empty set is 0, so $\#(\emptyset) = 0$.

   We can also write it as a sum

   $$\#(A) = \sum_{i \in A} 1,$$

   or using indicator functions

   $$\#(A) = \sum_{i} \mathbb{1}(i \in A).$$

2. Lebesgue measure. This is the measure that is the same as length in one dimension, area in two dimensions, volume in three dimensions and so on. For instance, the Lebesgue measure of the interval $[3.5, 7.2]$ is $7.2 - 3.5 = 3.7$. We will use ℓ for Lebesgue measure, so $\ell([3.5, 7.2]) = 3.7$.

   Just like you can find counting measure by summing up 1's, you can find Lebesgue measure for all of the sets considered in this course by integrating the constant function that is always 1.

   $$\ell(A) = \int_A 1(s) \, dA,$$

   or we can write it using indicator functions

   $$\ell(A) = \int_s \mathbb{1}(s \in A) \, dA.$$

For instance,

$$\ell([3.5, 7.2]) = \int_{s \in [3.5, 7.2]} 1 \, ds = s|_{3.5}^{7.2} = 7.2 - 3.5 = 3.7.$$

Let A be the triangle connecting vertices $(0, 0)$, $(0, 1)$, and $(1, 0)$ in $\mathbb{R}^2$. Then

$$\begin{aligned}
\ell(A) &= \int_{s \in A} 1 \, d\mathbb{R}^2 \\
&= \int_{x=0}^{1} \int_{y=0}^{1-x} 1 \, dy \, dx \\
&= \int_{x=0}^{1} 1 - x \, dx \\
&= -(1 - x)^2 / 2 |_0^1 = 1/2.
\end{aligned}$$

See the later chapter on Integrals for more details about iterated integrals.

A useful fact about a measure m is that if two sets do not intersect, then the measure of the union is the sum of the measure of the two sets. So in our second Euler diagram

$$m(A \cup B) = m(A) + m(B).$$

43.4 The Cartesian product of sets

If I have two sets A and B, then the *Cartesian product*, written $A \times B$ consists of ordered pairs (also called 2-tuples) (a, b) where $a \in A$ and $b \in B$.

> **Definition 88**
> The **Cartesian product** of two sets A and B (written $A \times B$) has
> $$(a, b) \in A \times B \text{ if and only if } a \in A \text{ and } b \in B.$$

For instance

$$\{a, b\} \times \{c, d, e\} = \{(a, c), (a, d), (a, e), (b, c), (b, d), (b, e)\}.$$

Both counting measure and Lebesgue measure are examples of *product measures*.

> **Definition 89**
> The measure μ is a **product measure** if for all measurable A and B,
> $$\mu(A \times B) = \mu(A) \cdot \mu(B).$$

> **Fact 109**
> Both counting measure and Lebesgue measure are product measures.

For example, let $A = \{a, b\}$ and $B = \{c, d, e\}$. Then

$$A \times B = \{(a, c), (a, d), (a, e), (b, c), (b, d), (b, e)\}.$$

Also,

$$\#(A) \cdot \#(B) = (2)(3) = 6 = \#(A \times B).$$

An example with Lebesgue measure, let $A = [2, 3]$ and $B = [6, 7.5]$. Then

$$A \times B = \{(x, y) : x \in [2, 3], y \in [6, 7.5]\},$$

here $\ell(A) = 3 - 2 = 1$, $\ell(B) = 7.5 - 6$ and $\ell(A \times B) = (1)(1.5) = 1.5$.

In other words, the area of a rectangle equals the product of the lengths of the sides!

Problems

43.1 What is the counting measure of $\{1, 2, \dots, 10\}$?

43.2 What is the counting measure of $\{1, 2, \dots, 10\} \cap \{6, 7, \dots, 15\}$

43.3 a) What is $\{r, g, b\} \cap \{g, b, y\}$?

b) What is $\{r, g, b\} \cup \{g, b, y\}$?

43.4 What is $\{1, 2, 3, 4, 5\} \cup \{3, 4, 5, 6, 7\}$?

43.5 What is the counting measure of $\{r, g, b\}$?

43.6 What is the counting measure of $\{x_1, x_2, x_3\}$?

43.7 What is the counting measure of $\{1, 3, 5\} \times \{7, 9\}$?

43.8 What is the counting measure of $\{r, g, b\} \times \{y, m, r\}$?

43.9 Let $A = \{r, g, b\}$. What is the counting measure of $A \times A \times A \times A$?

43.10 What is the counting measure of $\{0, 1\}^{10}$?

43.11   a) What is the Lebesgue measure of $[2, 10]$?

b) What is the Lebesgue measure of $[-6, 2]$?

43.12 What is the Lebesgue measure of $[-1, 1] \cap [0, 4]$?

43.13 What is the Lebesgue measure of $[3, 4.5] \times [0, 6]$?

43.14 What is the Lebesgue measure of $[0, 2]^4$?

43.15 *De Morgan's Laws* say that

$$(A \cup B)^C = A^C \cap B^C$$
$$(A \cap B)^C = A^C \cup B^C.$$

Assume this law hold for two sets, and then prove that

$$(A \cup B \cup C)^C = A^C \cap B^C \cap C^C.$$

Chapter 44

Functions

Summary

- A **function** $f : A \to B$ takes as input an element of A and returns as output a single element of B.

- A function is **one to one** (also written 1-1) if

$$(f(a) = f(a')) \to (a = a').$$

- A function is **onto** if

$$(\forall b \in B)(\exists a \in A)(f(a) = b).$$

Functions are the heart and soul of modern mathematics. Much of the appeal of mathematics as a tool is its ability to transform problems from a complicated formulation to a simpler result. And these transformations can often be written as functions.

This text will only be dealing with **computable functions**. These are essentially machines that take one or more variables as **input**. These input parameters are also called **arguments** of the function. The function then performs some computations and returns one or more **output** variables. There is a formal definition involving a model called *Turing machines*, but that's more detail than is necessary here.
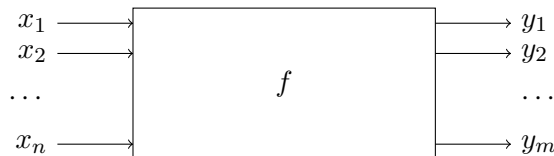
For instance, the computable function $f(x_1, x_2) = x_1 + x_2 = y_1$ has as input (x_1, x_2) and returns one output variable y_1.

Definition 90
The set of possible inputs to a function f is called the **domain**. The set of possible outputs of a function is called the **codomain**.

In other words, a function takes inputs from its domain and transforms them into elements of the codomain. Note that the codomain is not unique. For instance, I could say that $f(x) = x^2$ has domain and codomain equal to the set of real numbers, or I could say that the codomain consists only of nonnegative numbers.

> **Notation 7**
> Write $f : A \to B$ if A is the set of inputs to f, and all outputs lie in B.

For a given $b \in B$, there might be 0, 1, or an infinite number of values $a \in A$ such that $f(a) = b$. When there is at least 1, we call the function *onto*.

> **Definition 91**
> For function $f : A \to B$, the function is **onto** if
> $$(\forall b \in B)(\exists a \in A)(f(a) = b).$$

Mathematicians use the term *image* to describe the output that comes from applying the function to one or more inputs.

> **Definition 92**
> The **image** of $a \in A$ under the function f is $f(a)$. The **image** of $A' \subseteq A$ under the function f is $\{b : (\exists a \in A')(f(a) = b)\}$.

In words, the image of a set is all possible outputs of the function when the input comes from the set. We can characterize the onto property in terms of images.

> **Fact 110**
> Function $f : A \to B$ is **onto** or **surjective** if the image of A is B.

Onto functions have at least one input for every output. If a function has at most one input for every output, the function is one-to-one.

> **Definition 93**
> A function $f : A \to B$ is **one-to-one** (also written **1-1**) or **injective** if
> $$(\forall a, a' \in A)((f(a) = f(a')) \to (a = a')).$$

In words: if two inputs have the same output, the two inputs must have been the same input after all! Now, if a function is both 1-1 and onto than it is called (big surprise) *one-to-one and onto*.

> **Definition 94**
> A function f is **1-1 and onto** or a **bijection** if it is both one-to-one and onto.

Problems

44.1 Given a function $f : \{a, b, c\} \to \{b, c, d\}$, answer the following.

   a) What are the possible inputs to f?

b) What are the possible outputs from f?

44.2 Given a function $t : \{[0, 3] \to [0, 10]\}$, answer the following.

a) What are the possible inputs to t?

b) What are the possible output from t?

44.3 Suppose $g(a) = b$, $g(b) = b$, and $g(c) = d$. Is g one-to-one?

44.4 Suppose $h(2) = h(3) = 4$, $h(4) = 5$. Is h one-to-one?

44.5 Consider the function $f(x) = x^2$.

a) Say $f : [0, 1] \to [0, 1]$. Is f onto? Is it 1-1?

b) Say $f : [-1, 1] \to [0, 1]$. Is f onto? Is it 1-1?

c) Say $f : [-1, 1] \to [0, 2]$. Is f onto? Is it 1-1?

44.6 Suppose $r(a) = b$, $r(b) = c$, $r(c) = c$.

a) If $r : \{a, b, c\} \to \{a, b, c\}$, is r onto?

b) If $r : \{a, b, c\} \to \{b, c\}$, is r onto?

Chapter 45

Integration

Summary Integrals with respect to Lebesgue measure are the classic type of integral that you encounter in Calculus. Integrals over counting measure are the same as sums. In double (and triple and higher) integrals and sums, you can change the order of integration (or summation) using **Fubini's theorem** when the overall integral or sum is finite, or using **Tonelli's theorem** when the integrand or summand is nonnegative.

In Calculus you learned about integrals such as

$$A_1 = \int_{x=0}^{1} x^2 \, dx,$$

or sums like

$$A_2 = \sum_{i=1}^{\infty} 1/i^2.$$

These both involve summing over objects, and to a mathematician these are *both* examples of integrals. The first integral is with respect to Lebesgue measure, and the second is with respect to counting measure. For μ equal to Lebesgue measure, we can write

$$A_1 = \int_{x \in [0,1]} x^2 \, d\mu = \int_{x \in [0,1]} x^2 \, dx.$$

This is a *continuous* integral.

The sum is also an integral, but with respect to counting measure.

$$A_2 = \int_{i \in \{1,2,\dots\}} 1/i^2 \, d\# = \sum_{i \in \{1,2,\dots\}} 1/i^2.$$

Such an integral is a *discrete* integral.

Why is it useful to consider both integrals and sums as integrals (just with respect to different measures)? Well, it allows us to write theorems about integrals and sums just using one notation, rather than twice, once for continuous integrals and once for discrete integrals. Later on in this chapter we will see Tonelli and Fubini's theorems. These provide a nice example of this in action, because they apply to general integrals whether the measure is Lebesgue or counting measure.

45.1   Integrating over a measure

An integral consists of three parts:

1. The limits of integration.

2. The integrand.

3. The differential that tells us the measure being used for the integral.

These three parts can be remembered through the acronym LID: Limits, Integrand, Differential.
When μ is counting measure, an integral just becomes into a sum:

$$\int_{x\in A} f(x)\, d\mu = \sum_{x\in A} f(x).$$

Example 71
Let $f(i) = i$ for $i \in \{1, 2, 3, \ldots\}$. Find

$$\int_{i\in\{1,2,3,4\}} f(i)\, d\mu,$$

where μ is counting measure.

Answer   Since μ is counting measure, this becomes

$$\int_{i\in\{1,2,3,4\}} f(i)\, d\mu = f(1) + f(2) + f(3) + f(4) = 1 + 2 + 3 + 4 = \boxed{10}.$$

Example 72
For $f(i, j) = i + j$, what is $\int_{(i,j)\in A} f(i,j)\, d\#$ for $A = \{1, 2, 3\} \times \{1, 2\}$?

Answer   There are six elements in A, so the integral is the sum of the function over these elements:

$$\int_{(i,j)\in A} f(i,j) = f(1,1) + f(1,2) + f(2,1) + f(2,2) + f(3,1) + f(3,2).$$

When μ is Lebesgue measure, first try to calculate the Riemann integral that you learned about in your Calculus course. This is typically evaluated (when possible) by finding an antiderivative and using the Fundamental Theorem of Calculus. If the Riemann integral exists and is finite, the Lebesgue integral equals the same value.

Example 73
Let $f(x) = x$. Find

$$\int_{x \in [1,4]} f(x)\, dx.$$

Answer   Since $[x^2/2]' = x$ which is a continuous function over $[1, 4]$, the Fundamental
Theorem of Calculus tells us

$$\int_{x \in [1,4]} f(x)\, dx = \left.\frac{x^2}{2}\right|_1^4 = 8 - 1/2 = \boxed{7.500}.$$

45.2   Iterated integrals

When faced with an integral over 2 or more dimensions, it would be nice to be able to turn it into a
sequence of one dimensional integrals. For instance, we would like to be able to say that

$$\int_{(x,y) \in [0,1] \times [0,2]} f(x,y)\, d\mu(x,y) = \int_{x \in [0,1]} \int_{y \in [0,2]} f(x,y)\, d\mu(y)\, d\mu(x).$$

Unfortunately, this equality does not always hold. That being said, there are a couple simple
conditions under which this equality does hold.

1. Tonelli says that the equality holds when the integrand is nonnegative.

2. Fubini says that the equality holds when

$$\int_{(x,y) \in A} |f(x,y)|\, d\mu < \infty.$$

Note that if the integrand $f(x,y)$ is both positive and negative for points $(x,y) \in A$, then one
approach is to first calculate

$$\int_{(x,y) \in A} |f(x,y)|\, d\mu$$

using Tonelli, then if it is finite you can use Fubini on the original problem.

Formally, we can state these theorems as follows.

Theorem 8 (Fubini and Tonelli)

Suppose $A \subseteq \mathbb{R}^2$ and we wish to calculate

$$I = \int_{(x,y) \in A} f(x,y) \, d\mu.$$

where $\mu = \mu_1 \times \mu_2$ is a product measure. Suppose one of the following conditions holds:

1. Tonelli: $f(x,y) \geq 0$ for all $(x,y) \in A$.

2. Fubini: $\int_{(x,y) \in A} |f(x,y)| \, d\mu < \infty$.

Then

$$I = \int_{\{x|(\exists y)((x,y) \in A)\}} \left[\int_{\{y|(x,y) \in A\}} f(x,y) \, d\mu_2 \right] d\mu_1$$

$$= \int_{\{y|(\exists x)((x,y) \in A)\}} \left[\int_{\{x|(x,y) \in A\}} f(x,y) \, d\mu_1 \right] d\mu_2.$$

Let's start with an example using Lebesgue measure.

Example 74

Find

$$\int_{(x,y) \in [0,1] \times [0,2]} x^2 y \, d\mu,$$

where μ is Lebesgue measure.

Answer   Since the integrand is nonnegative for all $(x,y) \in [0,1] \times [0,2]$, Tonelli applies, and

$$\int_{(x,y) \in [0,1] \times [0,2]} x^2 y \, d\mathbb{R}^2 = \int_{x \in [0,1]} \int_{y \in [0,2]} x^2 y \, dy \, dx$$

$$= \int_{x \in [0,1]} x^2 y^2 / 2 \big|_0^2 \, dx$$

$$= \int_{x \in [0,1]} 2x^2 \, dx$$

$$= (2/3) x^3 \big|_0^1 = 2/3 \approx \boxed{0.6666}.$$

Here's an example that uses counting measure.

Example 75
Find
$$\sum_{i=1}^{\infty}\sum_{j=1}^{i}(1/2)^i.$$

Answer This is
$$\int_{(i,j):i\in\{1,2,\dots\},j\in\{1,\dots,i\}}(1/2)^i\,d\mu,$$
where μ is counting measure. Since the integrand is nonnegative, we can apply Tonelli. Note that the event
$$\{(i,j):j\in\{1,\dots,i\},i\in\{1,2,\dots\}\}$$
is the same as the event
$$\{(i,j):j\in\{1,2,3\dots\},i\in\{j,j+1,j+2,\dots\}\}$$
So by Tonelli
$$\sum_{i=1}^{\infty}\sum_{j=1}^{i}(1/2)^i=\sum_{j=1}^{\infty}\sum_{i=j}^{\infty}(1/2)^i$$
$$=\sum_{j=1}^{\infty}(1/2)^j/(1-1/2)$$
$$=1/(1-1/2)=\boxed{2}.$$

Note that

$$\sum_{i=1}^{\infty}\sum_{j=1}^{i}(1/2)^i=\sum_{i=1}^{\infty}i(1/2)^i,$$

so the above example also shows that Tonelli can be useful in evaluating sums over a single variable.

Example 76
Find
$$I = \int_{(x,y):x\in[0,1],y\in[x,1]} 1 - xy \, dy \, dx.$$

Answer   Because the integrand is nonnegative for all (x, y) such that $x \in [0, 1]$ and $y \in [x, 1]$, this is by Tonelli

$$I = \int_{x\in[0,1]} \int_{y\in[x,1]} 1 - xy \, dy \, dx$$
$$= \int_{x\in[0,1]} y - xy^2/2|_x^1 \, dx$$
$$= \int_{x\in[0,1]} 1 - x/2 - (x - x^3/2) \, dx$$
$$= \int_{x\in[0,1]} 1 - (3/2)x + x^3/2 \, dx$$
$$= x - (3/4)x^2 + x^4/8|_0^1$$
$$= 1 - 3/4 + 1/8 = 3/8 = \boxed{0.3750}.$$

In the example above, the limits for the y variable were allowed to depend on x because x had already been set by the integral on the outside.

What if we had wanted to do y on the outside? The set A looks like
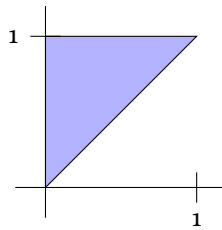
The smallest y can be in this region is 0, and the largest is 1. Once we pick y, then x can range from a low value of 0 to a high value of y. Hence

$$I = \int_{y\in[0,1]} \int_{x\in[0,y]} 1 - xy \, dx \, dy$$
$$= \int_{y\in[0,1]} x - x^2y/2|_0^y \, dy$$
$$= \int_{y\in[0,1]} y - y^3/2 \, dy$$
$$= y^2/2 - y^3/8|_0^1 \, dy$$
$$= 1/2 - 1/8 = \boxed{0.3750}.$$

and we end up with the same answer. (Thank goodness!)

The Tonelli and Fubini conditions were written for two-dimensional integrals, but actually applies to any integral of finite dimension.

45.3 Integration by parts

Integration by parts (IBP) allows us to slide a derivative from one factor inside an integral over to another factor. Its form is the same as the product rule for integration.

$$fg' = [fg]' - f'g.$$

You can see that this formula allows us to slide the derivative from g over to f, at the cost of having to add a negative sign and another term $[fg]'$. There are three steps to using IBP

1. Write your integrand as $f \cdot g'$.

2. Use the product rule formula to slide the derivative from g over to f.

3. Solve the simpler integral.

Usually you want to slide the derivative over to a term that becomes simpler when a derivative is applied. For instance, $[x]' = 1$, so often $f(x) = x$. Also $[\ln(x)]' = 1/x$, so we also try to slide derivatives over natural logs.

Example 77
Find

$$\int_0^1 x \exp(-x) \, dx$$

using IBP.

Answer Here $f(x) = x$ and $g'(x) = \exp(-x)$. Hence $g(x) = -\exp(-x)$. Plugging into the formula for IBP gives

$$\int_0^1 x \exp(-x) \, dx = \int_0^1 x[-\exp(-x)]' \, dx$$
$$= \int_0^1 [x(-\exp(-x))]' - [x]'(-\exp(-x)) \, dx$$
$$= -x \exp(-x)|_0^1 + \int_0^1 \exp(-x) \, dx$$
$$= -\exp(-1) - (0) + (-\exp(-x))|_0^1$$
$$= 1 - 2 \exp(-1) \approx \boxed{0.2642}.$$

Problems

45.1 Evaluate the following integrals:

$$\int_0^3 x^3 \, dx, \quad \int_{-\infty}^0 x \exp(x) \, dx, \quad \int_{-\infty}^\infty x \exp(-x^2/2) \, dx.$$

(Note, after you have worked out problems like this out, I encourage you to use tools like Wolfram Alpha to *check* your answers. For instance, type

```
integrate x^3 from 0 to 3
```

at the website `www.wolframalpha.com` to check your answer to the first integral.)

45.2 Find $\int_0^\infty x^2 \exp(-x) \, dx$.

45.3 Find

$$\int_1^2 x \ln(x) \, dx$$

by moving a derivative from $x = [x^2/2]'$ over to $\ln(x)$ to get rid of it.

45.4 Find $\int_0^1 -\ln(s) \, ds$.

45.5 Suppose

$$I = \int_0^\infty \int_0^x |f(x,y)| \, dy \, dx < \infty$$

so Fubini can be applied. Replace the question marks with the appropriate function of y.

$$I = \int_0^\infty \int_?^? f(x,y) \, dx \, dy.$$

45.6 Suppose

$$\sum_{i=1}^\infty \sum_{j=1}^i |w(i)| < \infty.$$

Replace the question marks with the appropriate function of j.

$$\sum_{i=1}^\infty \sum_{j=1}^i w(i) = \sum_{j=1}^\infty \sum_{i=?}^? w(i) < \infty.$$

Part IV

PROBLEM SOLUTIONS

Worked problems

1.1 For $A \in \mathcal{F}$, find $\mathbb{P}(A \cup A^C)$.

Solution Since $A \cup A^C = \Omega$, this is just $\mathbb{P}(\Omega) = \boxed{1}$.

1.3 Prove that if the state space Ω is measurable, so is $\emptyset$.

Solution The empty set $\emptyset$ is the complement of Ω, so if $\Omega \in \mathcal{F}$, that implies $\Omega^C = \emptyset \in \mathcal{F}$.

1.5 If $[0, 1 - 1/n]$ is measurable for every $n \geq 2$, show that the interval $[0, 1)$ is measurable.

Solution Since measurable sets are closed under countable union, it suffices to show that

$$[0, 1) = \cup_{n=1}^{\infty} [0, 1 - 1/n].$$

Let $x \in \cup_{n=1}^{\infty} [0, 1 - 1/n]$. Hence there is an n such that $x \in [0, 1 - 1/n]$ so $0 \leq x \leq 1 - 1/n < 1$. Hence $x \in [0, 1)$.

Let $x \in [0, 1)$. Then $0 \leq x < 1$. Let $i = \lceil 1/(1 - x) \rceil$. Then since $f(x) = \lceil x \rceil$ is an increasing function,

$$\frac{1}{1 - x} \leq \left\lceil \frac{1}{1 - x} \right\rceil = i.$$

So

$$1 - x \geq 1/i$$

and

$$-(1 - x) \leq -1/i.$$

Adding 1 to both sides then gives

$$1 - (1 - x) = x \leq 1 - 1/i.$$

So $x \in [0, 1 - 1/i]$ and must be in $x \in \cup_{n=1}^{\infty} [0, 1 - 1/n]$.

1.7 A *partition* of a set Ω is a collection of sets that are disjoint whose union is Ω. Suppose A, B, and C partition Ω. What is $\mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C)$?

Solution Since A, B, and C form a partition they are disjoint, so

$$\mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) = \mathbb{P}(A \cup B \cup C) = \mathbb{P}(\Omega) = \boxed{1}.$$

1.9 Suppose for $i \in \{0, 1, 2, \ldots\}$,

$$\mathbb{P}([i, i+1)) = (1/3)^{i+1}.$$

What is $\mathbb{P}([0, \infty))$?

Solution Since $A_i = [i, i+1)$ are disjoint, and

$$\cup_{i=1}^{\infty} [i, i+1) = [0, \infty),$$

we have

$$\mathbb{P}([0, \infty)) = \mathbb{P}(\cup A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) = \sum_{i=1}^{\infty} (1/3)^i = \frac{1/3}{1 - 1/3} = \boxed{0.5000}.$$

2.1 Suppose A and B are disjoint events, $\mathbb{P}(A) = 0.1$ and $\mathbb{P}(B) = 0.7$. What is $\mathbb{P}(A \cup B)$?

Solution Since A and B are disjoint, the probability of the union is the sum of the probabilities, so

$$0.1 + 0.7 = \boxed{0.8000}$$

2.3 Suppose $\mathbb{P}(A) = 0.4$, $\mathbb{P}(B) = 0.8$ and $\mathbb{P}(AB) = 0.3$. What is $\mathbb{P}(A \cup B)$?

Solution Using the principle of inclusion/exclusion,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB) = 0.4 + 0.8 - 0.3 = \boxed{0.9000}$$

2.5 If $\mathbb{P}([0, 3]) = 0.3$ and $\mathbb{P}([5, 9]) = 0.6$, what is $\mathbb{P}([0, 3] \cup [5, 9])$?

Solution Since probability is a measure and $[0, 3]$ and $[5, 9]$ are disjoint, this is

$$\mathbb{P}([0, 3]) + \mathbb{P}([5, 9]) = 0.3 + 0.6 = \boxed{0.9000}.$$

2.7 Say $\mathbb{P}(A_1) = \mathbb{P}(A_2) = \mathbb{P}(A_3) = 0.2$. Give an upper bound for $\mathbb{P}(A_1 \cup A_2 \cup A_3)$.

Solution From the union bound, $\mathbb{P}(A_1 \cup A_2 \cup A_3) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) = 0.2 + 0.2 + 0.2 = \boxed{0.6000}$.

2.9 Suppose a fair six sided die with sides labeled $\{1, 2, \ldots, 6\}$ is rolled three times. There are many possible outcomes, for instance, $(2, 3, 3)$ is one possible outcome.

    a) How many possible outcomes are there?

    b) If each outcome is equally likely, what must the probability of each outcome be?

    c) What is the chance of getting all 6's on the three rolls?

    d) What is the chance of not getting all 6's on the three rolls.

Solution

    a) There are $6 \cdot 6 \cdot 6 = \boxed{216}$ such outcomes.

    b) Note that

$$\mathbb{P}(1, 1, 1) + \mathbb{P}(1, 1, 2) + \cdots + \mathbb{P}(6, 6, 6) = 1.$$

    If each of the 216 outcomes have the same probability, that means each outcome must have probability $1/216$. Hence $1/216 = \boxed{0.004629\ldots}$.

c) From the last argument, $\mathbb{P}(6, 6, 6) = \boxed{0.004629\ldots}$.

d) From the complement rule, this is

$$1 - \frac{1}{216} = \frac{215}{216} = \boxed{0.9953\ldots}.$$

2.11 $\mathbb{P}(A \cup B) = 0.3$. What is $\mathbb{P}(A^C B^C)$?

Solution Recall De Morgan's Law: $(A \cup B)^C = A^C B^C$. Hence

$$\mathbb{P}(A^C B^C) = \mathbb{P}((A \cup B)^C) = 1 - \mathbb{P}(A \cup B) = 1 - 0.3 = \boxed{0.7000}.$$

2.13 Suppose $\mathbb{P}(A \in [0, 3]) = 1$, $\mathbb{P}(A \in [1, 2]) = 0.3$ and $\mathbb{P}(A \in [2, 3]) = 0.6$. What is $\mathbb{P}(A \in [2, 5])$?

Solution Since $\mathbb{P}(A \in [0, 3]) = 1$,

$$\mathbb{P}(A \in [2, 5]) = \mathbb{P}(A \in [2, 5] \cap [0, 3]) = \mathbb{P}(A \in [2, 3]) = \boxed{0.6000}.$$

3.1 Let $U \sim \mathsf{Unif}(\{1, 2, 3, 4, 5, 6\})$. What is $\mathbb{P}(U \leq 4)$?

Solution There are four values ($\{1, 2, 3, 4\}$) that are less than or equal to 4. There are six total values. So because it is uniform, this is $4/6 = \boxed{0.6666\ldots}$.

3.3 Let $A = \{a, b, c\}$ and $B = \{d, e\}$. What is $A \times B$?

Solution This set consists of six vectors:

$$\boxed{A \times B = \{(a, d), (a, e), (b, d), (b, e), (c, d), (c, e)\}.}$$

3.5 Let $W \sim \mathsf{Unif}(\{a, b, c, d\})$. What is $\mathbb{P}(W \in \{a, c\})$?

Solution Since W is uniform over a finite set,

$$\mathbb{P}(W \in \{a, c\}) = \frac{\#(\{a, c\})}{\#(\{a, b, c, d\})} = \frac{2}{4} = \boxed{0.5000}$$

3.7 Let $X_1 \sim \mathsf{Unif}(\{1, \ldots, 6\})$ and $X_2 \sim \mathsf{Unif}(\{1, \ldots, 6\})$ be independent. Then what is $\mathbb{P}(X_1 + X_2 = 6)$?

Solution Out of the 36 possibilities in $\{1, \ldots, 6\} \times \{1, \ldots, 6\}$, there are 5 that add up to 6:

$$(1, 5), (2, 4), (3, 3), (4, 2), (5, 1).$$

Since both X_1 and X_2 are uniform, each of these 36 possibilities are equally likely, so $5/36 \approx \boxed{0.1388}$ is the answer.

3.9 Suppose I roll three fair six sided dice so that each outcome is equally likely, and call the result (X_1, X_2, X_3). Let S be the smallest value showing on the dice. For $i \in \{1, 2, 3, 4, 5, 6\}$, find $\mathbb{P}(S = i)$.

Solution A trick for solving minimum problems is to consider $\mathbb{P}(S \geq 2)$. For this to happen, X_1, X_2 and X_3 must all be at least 2. Each one is at least 2 with probability $5/6$. Because the three random variables are independent, the probability that all are at least 2

is the product of these values, so this happens with probability $(5/6)^3$. Now consider $\mathbb{P}(S \geq 3)$. Here

$$\mathbb{P}(S \geq 3) = \mathbb{P}(X_1, X_2, X_3 \geq 3) = \left(\frac{4}{6}\right)^3.$$

Next we note that

$$\{S \geq 2\} = \{S \geq 3\} \cup \{S = 2\}.$$

The last two events are disjoint, so

$$\mathbb{P}(S \geq 2) = \mathbb{P}(S \geq 3) + \mathbb{P}(S = 2).$$

Hence

$$\mathbb{P}(S = 2) = \mathbb{P}(S \geq 2) - \mathbb{P}(S \geq 3) = \left(\frac{5}{6}\right)^3 - \left(\frac{4}{6}\right)^3.$$

In general, for $i \in \{1, 2, \ldots, 6\}$,

$$\mathbb{P}(S = i) = \left(\frac{6 - i + 1}{6}\right)^3 - \left(\frac{6 - i}{6}\right)^3$$

which gives the answer:

i	$\mathbb{P}(S = i)$
1	0.4212
2	0.2824
3	0.1712
4	0.08796
5	0.03240
6	0.004629

3.11 Prove that $\{2, 3, 4, 5, \ldots\}$ is a discrete set.

Solution Let $f(i) = i + 1$. Then let $j \in \{2, 3, 4, 5 \ldots\}$. Since $f(j - 1) = j$ and $j - 1 \in \{1, 2, 3, \ldots\}$, f is onto $\{2, 3, \ldots\}$. Hence $\{2, 3, \ldots\}$ is discrete.

4.1 Suppose $W \sim \mathsf{Unif}([-3, 3])$.

    a) What is $\mathbb{P}(W \in [-1, 2])$?

    b) What is $\mathbb{P}(W \in [-5, 0])$?

Solution

    a) Here $[-1, 2] \subseteq [-3, 3]$, so the answer is

$$\mathbb{P}(W \in [-1, 2]) = \frac{m([-1, 2])}{m([-3, 3])} = \frac{2 - (-1)}{3 - (-3)} = \frac{3}{6} = \boxed{0.5000}.$$

    b) Here $[-5, 0] = [-5, -3) \cup [-3, 3]$. Since $[-5, -3) \cap [-3, 3] = \emptyset$, it holds that $\mathbb{P}(W \in [-5, -3)) = 0$. Hence

$$\mathbb{P}(W \in [-5, 0]) = \mathbb{P}(W \in [-5, -3)) + \mathbb{P}(W \in [-3, 0])$$

$$= 0 + \frac{0 - (-3)}{3 - (-3)} = \frac{3}{6} = \boxed{0.5000}.$$

4.3 Suppose (U_1, U_2) is uniformly chosen over the unit circle

$$\{(x, y) : x^2 + y^2 \leq 1\}.$$

What is the chance that $|U_1| \geq U_2$?

Solution If $|x| \geq y$, then $x \geq y$ or $-x \geq y$. So the region looks like
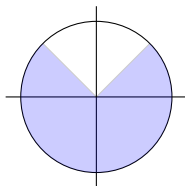
The shaded area is three quarters of the circle, so the probability is $\boxed{0.7500}$.

4.5 Let U_1 and U_2 be independent uniform random variables over $[0, 1]$. What is the chance that $U_2 \geq 3U_1$?

Solution Putting U_1 on the horizontal axis and U_2 on the vertical, the region where $U_2 \geq 3U_1$ is the shaded area.
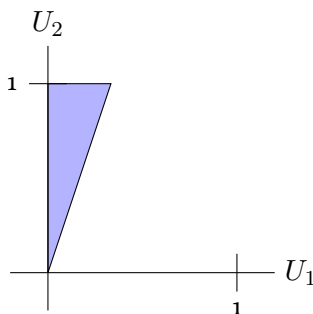
The area of the shaded region is $(1/2)(1/3)(1) = 1/6$ and the area of the whole unit square is 1, so the probability is $(1/6)/1 \approx \boxed{0.1666}$.

4.7 Say that $R \sim \mathsf{Unif}([0, 1])$.

    a) What is $\mathbb{P}(R \leq 0.4)$?

    b) What is $\mathbb{P}(R \leq 1.4)$?

    c) What is $\mathbb{P}(R \leq -0.4)$?

Solution

    a) This is the length of $[0, 0.4]$ (which is 0.4) divided by the length of $[0, 1]$ (which is 1). So $\boxed{0.4000}$.

    b) Because $R \in [0, 1]$ always, this is $\boxed{1}$.

    c) Since R is at least 0, this is $\boxed{0}$.

5.1 Suppose $U \sim \mathsf{Unif}([0, 1])$ and $A = -\ln(U)/2$.

    a) Find $\mathbb{P}(A \geq 2)$.

    b) Find $\mathbb{P}(A \geq -2)$.

    c) For $a \geq 0$, find $\mathbb{P}(A \geq a)$.

    d) For $a < 0$, find $\mathbb{P}(A \geq a)$.

Solution

    a) We need that $x \mapsto -x$ is a decreasing function, and $x \mapsto \exp(x)$ is increasing. So

$$\begin{aligned}
\mathbb{P}(A \geq 2) &= \mathbb{P}(-\ln(U)/2 \geq 2) \\
&= \mathbb{P}(\ln(U) \leq -4) \\
&= \mathbb{P}(U \leq \exp(-4)) \\
&= \mathbb{P}(U \in (-\infty, \exp(-4)] \cap [0,1]) \\
&= \mathbb{P}(U \in [0, \exp(-4)]) \approx \boxed{0.01831}.
\end{aligned}$$

    b) As in the last problem

$$\begin{aligned}
\mathbb{P}(A \geq -2) &= \mathbb{P}(-\ln(U)/2 \geq -2) \\
&= \mathbb{P}(\ln(U) \leq 4) \\
&= \mathbb{P}(U \leq \exp(4)) \\
&= \mathbb{P}(U \in (-\infty, e^4] \cap [0,1]) = \mathbb{P}(U \in [0,1]) = \boxed{1}
\end{aligned}$$

    c) For $a \geq 0$,

$$\begin{aligned}
\mathbb{P}(A \geq a) &= \mathbb{P}(-\ln(U)/2 \geq a) \\
&= \mathbb{P}(\ln(U) \geq -2a) \\
&= \mathbb{P}(U \geq \exp(-2a)) = \boxed{\exp(-2a)}.
\end{aligned}$$

    d) For $a \leq 0$,

$$\begin{aligned}
\mathbb{P}(A \geq a) &= \mathbb{P}(-\ln(U)/2 \geq a) \\
&= \mathbb{P}(\ln(U) \leq -2a) \\
&= \mathbb{P}(U \leq \exp(-2a)) = \boxed{1}
\end{aligned}$$

5.3 Let $U \sim \mathsf{Unif}([-1,1])$. Find the cdf of $1 - U^2$.

Solution For $U \in [-1,1]$, $U^2 \in [0,1]$ and $1 - U^2 \in [0,1]$. Hence $\mathrm{cdf}_{1-U^2}(a) = 0$ for $a < 0$ and $\mathrm{cdf}_{1-U^2}(a) = 1$ for $a > 1$. Suppose $a \in [0,1]$. Then

$$\begin{aligned}
\mathbb{P}(1 - U^2 \leq a) &= \mathbb{P}(1 - a \leq U^2) \\
&= \mathbb{P}((\sqrt{1-a} \leq U) \vee (U \leq -\sqrt{1-a})) \\
&= \mathbb{P}((\sqrt{1-a} \leq U) \vee (U \leq -\sqrt{1-a}), -1 \leq U \leq 1) \\
&= \mathbb{P}((\sqrt{1-a} \leq U \leq 1) \vee (-1 \leq U \leq -\sqrt{1-a})) \\
&= \mathbb{P}(\sqrt{1-a} \leq U \leq 1) + \mathbb{P}(-1 \leq U \leq -\sqrt{1-a})) \\
&= \frac{1 - \sqrt{1-a}}{1 - (-1)} + \frac{-\sqrt{1-a} - (-1)}{1 - (-1)} \\
&= 1 - \sqrt{1-a}.
\end{aligned}$$

Therefore, the entire cdf is

$$\boxed{\text{cdf}_{1-U^2}(a) = (1 - \sqrt{1-a})\mathbb{1}(a \in [0,1]) + \mathbb{1}(a > 1)}.$$

5.5 Let ω be uniform over $[0, 1]$, and suppose $X(\omega) = 2\omega + 3$. Find

    a) $\mathbb{P}(X \in [3.5, 4.7])$.

    b) $\mathbb{P}(X \in [0, 1])$.

    c) $\mathbb{P}(X^2 \le 10)$.

Solution

    a) Since $X = 2\omega + 3$, we can solve for ω to get

$$3.5 \le X \le 4.7$$
$$3.5 \le 2\omega + 3 \le 4.7$$
$$0.5 \le 2\omega \le 1.7$$
$$0.25 \le \omega \le 0.85,$$

    which since $\omega \sim \text{Unif}([0,1])$, the chance that happens is

$$\mathbb{P}(\omega \in [0.25, 0.85]) = 0.85 - 0.25 = \boxed{0.6000}.$$

    b) For $X \in [0, 1]$, solving for ω gives $\omega \in [-3/2, -1]$. The chance of this happening is $\boxed{0}$.

    c) Since $X \ge 0$, we have

$$X^2 \le 10$$
$$X \le \sqrt{10}$$
$$2\omega + 3 \le \sqrt{10}$$
$$\omega \le (\sqrt{10} - 3)/2 \approx \boxed{0.05409}.$$

5.7 Suppose $U \sim \text{Unif}([-1, 0])$.

    a) Let $X = U^2$. Find the cdf of X.

    b) Find the cdf of U.

Solution

    a) As always, begin with substituting the function of U for the random variable X.

$$\mathbb{P}(X \le a) = \mathbb{P}(U^2 \le a)$$

    This is 0 if $a < 0$ since $U^2 \ge 0$. Assume $a \ge 0$. Then

$$\begin{aligned}
\mathbb{P}(X \le a) &= \mathbb{P}(-\sqrt{a} \le U \le \sqrt{a}) \\
&= \mathbb{P}(-\sqrt{a} \le U \le \sqrt{a}, -1 \le U \le 0) \\
&= \mathbb{P}(\max(-1, -\sqrt{a}) \le U \le 0) \\
&= \frac{0 - \max(-1, -\sqrt{a})}{0 - (-1)} \\
&= \min(1, \sqrt{a}).
\end{aligned}$$

Putting this together gives

$$\mathrm{cdf}_X(a) = \sqrt{a}\mathbb{1}(a \in [0,1]) + \mathbb{1}(a > 1).$$

b) When $a > 0$ it holds that $\mathbb{P}(U \le a) = 1$ and for $a < -1$ it holds that $\mathbb{P}(U \le a) = 0$. For $a \in (-1, 0)$,

$$\mathbb{P}(U \le a) = \frac{a - (-1)}{0 - (-1)} = 1 + a$$

Hence

$$\mathrm{cdf}_U(a) = (1 + a)\mathbb{1}(a \in [0,1]) + \mathbb{1}(a > 1).$$

5.9 Let $G \sim \mathrm{Geo}(p)$. For i a value in $\{1, 2, 3 \ldots\}$, what is $\mathbb{P}(G = i)$?

Solution Recall that our Bernoulli random variables can be built as $B_i = \mathbb{1}(U_i \le p)$, where the $\{U_i\}$ are iid uniform over $[0, 1]$. For G to equal i, all of $U_1, \ldots, U_{i-1}$ must be greater than p, and $U_i \le p$. So

$$\begin{aligned}
\mathbb{P}(G = i) &= \mathbb{P}(U_1 > p, U_2 > p, \ldots, U_{i-1} > p, U_i \le p) \\
&= (1 - p)(1 - p) \cdots (1 - p)p \\
&= (1 - p)^{i-1}p.
\end{aligned}$$

5.11 Let $U \in [-1, 1]$. What is $\mathbb{P}(U^2 \ge 0.6)$?

Solution The graph of U^2 looks like
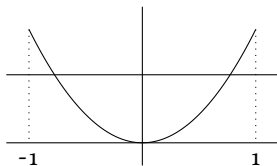
Note that $U^2 \ge 0.6$ when $U \ge \sqrt{0.6}$ and $U \le -\sqrt{0.6}$. So

$$\begin{aligned}
\mathbb{P}(U^2 \ge 0.6) &= \mathbb{P}(U \ge \sqrt{0.6}) + \mathbb{P}(U \le -\sqrt{0.6}) \\
&= \frac{1 - \sqrt{0.6}}{1 - (-1)} + \frac{-\sqrt{0.6} - (-1)}{1 - (-1)} \\
&= \boxed{0.2254}.
\end{aligned}$$

5.13 Consider the probability that for $\mathsf{Exp}(1)$ and $\mathsf{Unif}([0, 1])$ random variables drawn independently, that the second is bigger than the first. To find this, let U_1 and U_2 be iid $\mathsf{Unif}([0, 1])$. Then set $T = -\ln(U_2)$. Then find $\mathbb{P}(U_1 \ge T)$.

Solution This the same as finding the area of the region $\{U_1 \ge -\ln(U_2)\}$, or $\{-U_1 \le \ln(U_2)\}$, or $\{U_2 \ge \exp(-U_1)\}$. Graphically, this region looks like:
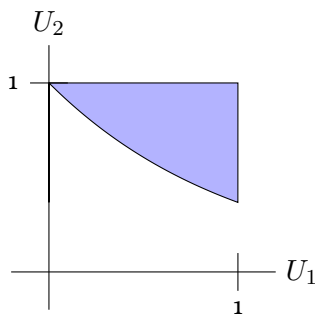
Then the area is

$$\text{area} = \int_{x=0}^{1} \int_{y=\exp(-x)}^{1} 1 \, dy \, dx$$
$$= \int_{x=0}^{1} (1 - \exp(-x)) \, dx$$
$$= x + \exp(-x)|_0^1$$
$$= 1/e = \boxed{0.3678\dots}.$$

5.15 Let $B \sim \text{Bern}(p)$ and $T \sim \text{Exp}(1)$ be independent random variables. Find $\mathbb{P}(T \geq B)$.
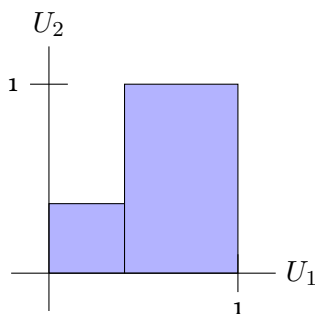
Solution Let $B = \mathbb{1}(U_1 \leq p)$, and $T = -\ln(U_2)$. Then if $U_1 \leq p$, we are looking at

$$\mathbb{P}(-\ln(U_2) \geq 1) = \mathbb{P}(U_2 \leq \exp(-1)) = \exp(-1),$$

and if $U_1 > p$, we are looking at

$$\mathbb{P}(-\ln(U_2) \geq 0) = \mathbb{P}(U_2 \leq \exp(-0)) = 1.$$

Hence the region looks like

and the area is

$$(1 - p)(1) + (p)(\exp(-1)) = \boxed{1 - p - p/e}.$$

5.17 The time until radioactive decay of a single atom is exponentially distributed with rate λ. If T is the time until the particle decays, the half-life t_{hl} is the time such that $\mathbb{P}(T \geq t_{\text{hl}}) = 1/2$. The half-life for an atom of uranium 238 is 4.5 billion years.

a) What is λ?

b) If the Earth is 4.2 billion years old, what is the chance that an atom of U-238 present at the birth of the planet is still intact?

Solution

a) Recall that for an exponential random variable, the median is $\ln(2)/\lambda$. So $\ln(2)/\lambda = 4.5$ billion years, which makes $\boxed{\lambda \approx 0.1540 \text{ per billion years.}}$

b) The cdf of an exponential random variable with rate λ is $F_T(a) = (1 - \exp(-\lambda a))\mathbb{1}(T \geq 0)$. So $\mathbb{P}(T > 4.5 \text{ billion years})$ is

$$\mathbb{P}(T > 4.5) = \exp(-4.2(\ln(2)/4.5)) = 2^{-4.2/4.5} = \boxed{0.5236\ldots.}$$

(What this implies is that roughly 52.36% of the original U-238 in the Earth is still intact. The rest has undergone radioactive decay.)

6.1 Suppose $\mathbb{P}(A|B) = 0.3$ and $\mathbb{P}(B) = 0.8$. What is $\mathbb{P}(AB)$?

Solution From the conditional probability formula, this is

$$\mathbb{P}(A|B)\mathbb{P}(B) = (0.3)(0.8) = \boxed{0.2400}.$$

6.3 Suppose $\mathbb{P}(A) = 0.3$ and $\mathbb{P}(B) = 0.5$, and we know that A and B are independent. What is $\mathbb{P}(A|B)$?

Solution Since A and B are independent,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A) = \boxed{0.3000}.$$

6.5 Let $X \sim \mathsf{Unif}(\{1, 2, 3, 4, 5, 6\})$.

a) What is $\mathbb{P}(X = 5|X \geq 3)$?

b) What is $\mathbb{P}(X = 5|X \geq 6)$?

Solution

a) Using the rule that $\mathbb{P}(A|B) = \mathbb{P}(A \text{ and } B)/\mathbb{P}(B)$,

$$
\begin{aligned}
\mathbb{P}(X = 5|X \geq 3) &= \frac{\mathbb{P}(X = 5, X \geq 3)}{\mathbb{P}(X \geq 3)} \\
&= \frac{\mathbb{P}(X = 5)}{\mathbb{P}(X = 3) + \mathbb{P}(X = 4) + \mathbb{P}(X = 5) + \mathbb{P}(X = 6)} \\
&= \frac{1/6}{1/6 + 1/6 + 1/6 + 1/6} \\
&= \frac{1}{4} = \boxed{0.2500}.
\end{aligned}
$$

b) Similar to the last part:

$$\mathbb{P}(X = 5|X \geq 6) = \frac{\mathbb{P}(X = 5, X \geq 6)}{\mathbb{P}(X \geq 3)}$$

But here X cannot both be 5 and be at least 6 at the same time! So $\mathbb{P}(X = 5, X \geq 6) = 0$, so the overall probability is $\boxed{0}$.

6.7 Let $X \sim \mathsf{Unif}(\Omega)$, where Ω is a finite set. Let $A \subseteq \Omega$. Let Y have the same distribution as X conditioned on $X \in A$. Prove that $Y \sim \mathsf{Unif}(A)$.

Solution To show that $Y \sim \mathsf{Unif}(A)$, it suffices to show that $(\forall a \in A)(\mathbb{P}(Y = a) = 1/\#(A))$.

Proof. Let $a \in A$. Then

$$\begin{aligned}
\mathbb{P}(Y = a) &= \mathbb{P}(X = a | X \in A) \\
&= \frac{\mathbb{P}(X = a, X \in A)}{\mathbb{P}(X \in A)} \\
&= \frac{1/\#(\Omega)}{\#(A)/\#(\Omega)} \\
&= \frac{1}{\#(A)},
\end{aligned}$$

which completes the proof. $\square$.

6.9 A lab occasionally has small leaks of chemicals in the experimental space. Each leak is independent of the others and has a 90% chance of being benign, and a 10% chance of being toxic. The lab director has two drones at her disposal. The first drone can detect whether or not any toxic leaks are in the lab. The second drone can count the number of leaks present in the lab.

The drones are sent in: the first reports that yes, there is at least one toxic leak in the lab. The second drone reports there are exactly three leaks in the lab.

Conditioned on this information, what is the chance that there is exactly one toxic leak, and two benign leaks?

Solution The second drone reported that there are three anomalies. Let $X_1, X_2, X_3 \sim \mathsf{Bern}(0.1)$ be iid. Then $X_i = 1$ indicates that leak i is toxic, while $X_i = 0$ indicates that it is benign.

Then the question is: what is $\mathbb{P}(X_1 + X_2 + X_3 = 1 | X_1 + X_2 + X_3 \geq 0)$? We use the conditional probability formula: $\mathbb{P}(A|B) = \mathbb{P}(A, B)/\mathbb{P}(B)$. In this case, $X_1 + X_2 + X_3 = 1$ implies $X_1 + X_2 + X_3 \geq 0$, so

$$\begin{aligned}
\mathbb{P}(X_1 + X_2 + X_3 = 1 | X_1 + X_2 + X_3 \geq 0) &= \frac{\mathbb{P}(X_1 + X_2 + X_3 = 1)}{\mathbb{P}(X_1 + X_2 + X_3 \geq 0)} \\
&= \frac{\binom{3}{1}(0.1)^1(0.9)^2}{1 - \mathbb{P}(X_1 + X_2 + X_3 = 0)} \\
&= \frac{3(0.1)^1(0.9)^2}{1 - (0.9)^3} \\
&\approx \boxed{0.8966}.
\end{aligned}$$

6.11 For $U \sim \mathsf{Unif}([2, 10]$, what is $\mathbb{P}(U \leq 3 | U \leq 5)$?

Solution Using our formula

$$\begin{aligned}
\mathbb{P}(U \leq 3 | U \leq 5) &= \frac{\mathbb{P}(U \leq 3, U \leq 5)}{\mathbb{P}(U \leq 5} \\
&= \frac{(3-2)/(10-2)}{(5-2)/(10-2)} = \frac{1}{3} \approx \boxed{0.3333}.
\end{aligned}$$

7.1 Suppose $X \sim \text{Bin}(10, 0.2)$. What is $\mathbb{P}(X \geq 2)$?

Solution This is a case where using the complement helps:

$$
\begin{aligned}
\mathbb{P}(X \geq 2) &= 1 - \mathbb{P}(X \leq 1) \\
&= 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) \\
&= 1 - \binom{10}{0}0.2^0 0.8^{10} - \binom{10}{1}0.2^1 0.8^9 \\
&\approx \boxed{0.6241 \ldots}
\end{aligned}
$$

7.3   a) What is 5 choose 2?

b) How many ways are there to arrange the letters AABBB?

Solution

a) Following our formula, this is

$$
\binom{5}{2} = \frac{5!}{2!3!} = \frac{5 \cdot 4}{1 \cdot 2} = \boxed{10}.
$$

b) We have to choose which two out of the five spaces receive the letter A, which is 5 choose 2, or $\boxed{10}$ as before.

7.5 How many sequences using letters F and S are of length 10 and have exactly 8 S letters?

Solution This is just the binomial coefficient 10 choose 8. Note that when writing the factorials, lots of factors cancel, which makes things easier in the calculation.

$$
\binom{10}{8} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 2 \cdot 1} = \frac{10 \cdot 9}{2 \cdot 1} = 5 \cdot 9 = \boxed{45}.
$$

I can check this in Wolfram Alpha by using `10 choose 8`.

7.7 Suppose $N \sim \text{Bin}(10, 0.3)$. What is $\mathbb{P}(N = 8)$?

Solution Using our formula for binomial probabilities,

$$
\mathbb{P}(N = 8) = \binom{10}{8}0.3^8 (0.7)^{10-8} \approx \boxed{0.001446}.
$$

I can check this answer in R with `dbinom(8,10,0.3)`.

7.9 Suppose that $[X|N] \sim \text{Unif}(\{1, 2, 3, \ldots, N\})$ and $N \sim \text{Unif}(\{1, 2, 3, 4, 5, 6\})$. What is

$$
\mathbb{P}(N = 3|X = 2)?
$$

Solution Using Bayes' Rule, this is

$$
\mathbb{P}(N = 3|X = 2) = \frac{\mathbb{P}(X = 2|N = 3)\mathbb{P}(N = 3)}{\mathbb{P}(X = 2)}
$$

The numerator is $(1/3)(1/6)$, while the denominator is

$$
\begin{aligned}
\mathbb{P}(X=2) = {} & \mathbb{P}(X=2|N=1)\mathbb{P}(N=1) + \mathbb{P}(X=2|N=2)\mathbb{P}(N=2) + \cdots + \\
& \mathbb{P}(X=2|N=6)\mathbb{P}(N=6) \\
= {} & 0 + (1/2)(1/6) + (1/3)(1/6) + (1/4)(1/6) + (1/5)(1/6) + (1/6)(1/6).
\end{aligned}
$$

Canceling the $1/6$ factors, the solution is

$$
\begin{aligned}
\mathbb{P}(N=3|X=2) &= \frac{(1/3)}{(1/2)+(1/3)+\cdots+(1/6)} \\
&= \frac{20}{87} \approx \boxed{0.2298}.
\end{aligned}
$$

7.11 Autotomic Industries produces two types of pain relievers that here we will call A and B for simplicity. Type A relieves pain in 40% of patients, while type B relieves pain in 20% of patients.

A patient takes one of the painkillers (they do not know which type) and relieves their pain. What is the chance that they used type A?

Solution Let P be the event that the pain is relieved. Then

$$
\mathbb{P}(P|A) = 0.4, \quad \mathbb{P}(P|B) = 0.2.
$$

So by Bayes' Rule,

$$
\begin{aligned}
\mathbb{P}(A|P) &= \frac{\mathbb{P}(P|A)\mathbb{P}(A)}{\mathbb{P}(P)} \\
&= \frac{\mathbb{P}(P|A)\mathbb{P}(A)}{\mathbb{P}(AP)+\mathbb{P}(BP)} \\
&= \frac{\mathbb{P}(P|A)\mathbb{P}(A)}{\mathbb{P}(P|A)\mathbb{P}(A)+\mathbb{P}(P|B)\mathbb{P}(B)}
\end{aligned}
$$

Since we have no information indicating otherwise, assume that the person was equally likely to have taken either A or B,

$$
\begin{aligned}
\mathbb{P}(A|P) &= \frac{(0.4)(1/2)}{(0.4)(1/2)+(0.2)(1/2)} \\
&= \frac{2}{3} = \boxed{0.6666}
\end{aligned}
$$

7.13 Bets on red and black on a roulette table pay even odds, which means if you bet x dollars and win, you get back your x dollar bet plus x more dollars. If you lose, then you lose your x dollar bet.

Suppose you repeatedly bet the same amount of money on red at a roulette table for twenty spins of the wheel. On an American Roulette wheel there are 18 out of 38 spaces that are red, and the ball is equally likely to land in any of the spaces.

a) Find the probability that at the end of the twenty games you are ahead (so you have more money than when you started.)

b) Find the probability that at the end of the twenty games you are behind (so you have less money than when you started.)

c) Find the probability that at the end of the twenty games you have broken even.

Solution

a) Let B be the number of times you win. Then since each space is equally likely, there is an $18/38$ chance of winning, which means $B \sim \mathsf{Bin}(20, 18/38)$. In order to be ahead, you must have won more than 10 games, so

$$\mathbb{P}(B > 10) = 1 - \mathbb{P}(B \leq 10) = \boxed{0.3223} \text{ (from R).}$$

b) To be behind, you must have won at most 9 games, and

$$\mathbb{P}(B \leq 9) \approx \boxed{0.5062} \text{ (from R).}$$

c) Last but not least, to break even you must have won exactly 10 games, which occurs with probability

$$\mathbb{P}(B = 10) \approx \boxed{0.1713} \text{ (from R).}$$

8.1 Suppose $X = \sqrt{U}$ where $U \sim \mathsf{Unif}([0, 1])$. Find the density of X.

Solution First find the cdf: $U \in [0, 1] \Rightarrow X \in [0, 1]$, so $\mathrm{cdf}_X(a) = 0$ for $a < 0$ and $\mathrm{cdf}_X(a) = 1$ for $a > 1$. In both these regions, differentiating gives 0.

Let $a \in [0, 1]$. Then

$$\begin{aligned}
\mathbb{P}(X \leq a) &= \mathbb{P}(\sqrt{U} \leq a) \\
&= \mathbb{P}(U \leq a^2) \\
&= a^2.
\end{aligned}$$

Therefore, the density for $a \in [0, 1]$ is $[a^2]' = 2a$. The overall answer is then

$$\boxed{f_X(a) = 2a\mathbb{1}(a \in [0, 1])}$$

8.3 Suppose $f_X(s) = \exp(-s)[1 - \exp(-2)]^{-1}\mathbb{1}(s \in [0, 2])$.

a) What is $\mathbb{P}(X \geq 1.1)$?

b) What is $\mathbb{P}(X \leq -0.5)$?

c) Graph F_X, the cdf of X.

Solution

a) This is found by the integral

$$\begin{aligned}
\mathbb{P}(X \geq 1.1) &= \int_{1.1}^{\infty} \frac{\exp(-s)}{1 - \exp(-2)}\mathbb{1}(s \in [0, 2]) \, ds \\
&= \int_{1.1}^{2} \frac{\exp(-s)}{1 - \exp(-2)} \, ds \\
&= \frac{\exp(-1.1) - \exp(-2)}{1 - \exp(-2)} \approx \boxed{0.2284}
\end{aligned}$$

b) Once the indicator function is included, the integral disappears! So the answer is 0.

$$\mathbb{P}(X \le -0.5) = \int_{1.1}^{\infty} \frac{\exp(-s)}{1 - \exp(-2)} \mathbb{1}(s \in [0, 2]) \, ds = \boxed{0}$$
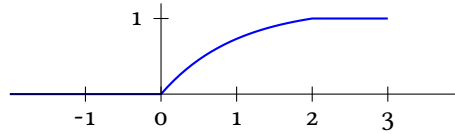
c) In fact, for any $a < 0$, $F_X(a) = 0$. For any $a > 2$,

$$\mathbb{P}(X \le a) = \int_{-\infty}^{a} \frac{\exp(-s)}{1 - \exp(-2)} \mathbb{1}(s \in [0, 2]) \, ds$$
$$= \int_{0}^{2} \frac{\exp(-s)}{1 - \exp(-2)} \, ds = 1.$$

For $a \in [0, 2]$,

$$\mathbb{P}(X \le a) = \int_{-\infty}^{a} \frac{\exp(-s)}{1 - \exp(-2)} \mathbb{1}(s \in [0, 2]) \, ds$$
$$= \int_{0}^{a} \frac{\exp(-s)}{1 - \exp(-2)} \, ds = \frac{1 - \exp(-a)}{1 - \exp(-2)}.$$

So the graph of the cdf looks like

8.5 Suppose W has density $f_W(x) = 3x^2 \mathbb{1}[x \in [0, 1])$. What is the density of $Y = 3W + 2$?

Solution The random variable Y is a shifted and scaled version of W. Hence we use the shifting and scaling result. Note that we have to figure out both the new functional form and solve inside the indicator function. Recall that $x \in [0, 1]$ is the same event as $0 \le x \le 1$.

$$f_{3W+2}(x) = f_W((x - 2)/3)$$
$$= (1/|3|)3((x - 2)/3)^2 \mathbb{1}(0 \le (x - 2)/3 \le 1)$$
$$= [(x - 2)^2/9] \mathbb{1}(0 \le x - 2 \le 3)$$
$$= \boxed{[(x - 2)^2/9] \mathbb{1}(2 \le x \le 5).}$$

Remark: the indicator reflects the fact that if I take a random variable that is somewhere in $[0, 1]$, multiply it by 3 and add 2, then the resulting number is somewhere in $[2, 5]$.

8.7 Suppose X has density $f_X(x) = C/(1 + x^2)$. What is C?

Solution Recall that the antiderivative of $1/(1 + x^2)$ is $\arctan(x)$. Hence

$$\mathbb{P}(X \in (-\infty, \infty)) = \int_{-\infty}^{\infty} \frac{C}{1 + x^2} \, dx$$
$$= C \arctan(x)|_{-\infty}^{\infty} = C[\tau/4 - (-\tau/4)],$$

So $C\tau/2 = 1$, and $C = 2/\tau \approx \boxed{0.3183}$.

8.9 Suppose $f_T(t) = 2\exp(-2t)\mathbb{1}(t \geq 0)$. Find the density of $2T + 1$.

Solution From the density formula, divide by $|2| = 2$, and replace t with $(t-1)/2$ everywhere it appears:

$$f_{2T+1}(t) = \frac{1}{2}2\exp(-2(t-1)/2)\mathbb{1}((t-1)/2 \geq 0)$$
$$= \boxed{\exp(-(t-1)))\mathbb{1}(t \geq 1)}.$$

8.11 Let $U \sim \text{Unif}([-2, 2])$.

a) Let $T = U^3$. What is the density of T?

b) Let $V = U^4$. What is the density of V?

Solution

a) First find the cdf:

$$F_T(a) = \mathbb{P}(T \leq a)$$
$$= \mathbb{P}(U^3 \leq a)$$
$$= \mathbb{P}(U \leq a^{1/3}).$$

When $a^{1/3} \in [-2, 2]$, then $a \in [-8, 8]$. For $a \in [-8, 8]$ $\mathbb{P}(U \in [-2, a^{1/3}]) = (a^{1/3} - (-2))/(2 - (-2))$. So

$$F_T(a) = (1/4)(2 + a^{1/3})\mathbb{1}(a \in [-8, 8]) + \mathbb{1}(a > 1).$$

Now differentiate to get the density

$$\boxed{f_T(a) = (1/12)a^{-2/3}\mathbb{1}(a \in [-8, 8])}.$$

b) Again we begin by finding the cdf of V:

$$F_T(a) = \mathbb{P}(V \leq a)$$
$$= \mathbb{P}(U^4 \leq a)$$
$$= \mathbb{P}(-a^{1/4} \leq U \leq a^{1/4})$$
$$= F_U(a^{1/4}) - F_U(-a^{1/4})$$

Now differentiate both sides to get:

$$f_V(a) = f_U(a^{1/4})(1/4)a^{-3/4} - f_U(-a^{1/4})(-1/4)a^{-3/4}.$$

The density of the uniform is $f_U(b) = (1/4)\mathbb{1}(b \in [-2, 2])$, so

$$f_V(a) = [(1/4)(1/4)a^{-3/4} + (1/4)(1/4)a^{-3/4}]\mathbb{1}(-a^{1/4} \in [-2, 2])$$
$$= (1/16)a^{-3/4}\mathbb{1}(a \in [0, 16]) + (1/16)a^{-3/4}\mathbb{1}(a \in [0, 16]$$
$$= \boxed{(1/8)a^{-3/4}\mathbb{1}(a \in [0, 16])}$$

8.13 Show that if T has an exponential distribution with rate λ, then $\lfloor T \rfloor + 1$ has a geometric distribution and find the parameter p of the geometric as a function of λ.

Solution First solve for T, then simplify, then compare to the probabilities for the geometric random variable:

$$\begin{aligned}
\mathbb{P}(\lfloor T \rfloor + 1 = i) &= \mathbb{P}(\lfloor T \rfloor = i - 1) \\
&= \mathbb{P}(i - 1 \leq T < i) \\
&= \mathbb{P}(T < i) - \mathbb{P}(T \leq i - 1) \\
&= (1 - e^{-i\lambda}) - (1 - e^{-(i-1)\lambda}) \\
&= e^{\lambda(i-1)}[1 - e^{-\lambda}] \\
&= [1 - e^{-\lambda}](e^\lambda)^{i-1}
\end{aligned}$$

when $i \in \{1, 2, 3, \ldots\}$. This is the same as a geometric random variable when $\boxed{p = 1 - e^{-\lambda}.}$

9.1 For X with density $f_X(i) = 0.3\mathbb{1}(i = 1) + 0.7\mathbb{1}(i = 4)$, what is $\mathbb{P}(X \leq 2)$?

Solution Here the density is with respect to counting measure, and so $\mathbb{P}(X \leq 2) = \mathbb{P}(X = 1) = \boxed{0.3000}$.

9.3 Let U_1 and U_2 be iid $\mathsf{Unif}(\{1, 2, 3, 4\})$. Find the density of $U_1 + U_2$.

Solution Since U_1 and U_2 are (with probability 1) always in $\Omega = \{1, 2, 3, 4\}$, their sum will be in the set of numbers $\{2, \ldots, 8\}$.

Now consider $i \in \{2, \ldots, 8\}$, for instance, $i = 4$. There are three ways that this event could occur:

$$\{i = 4\} = \{U_1 = 1, U_2 = 3\} \cup \{U_1 = 2, U_2 = 2\} \cup \{U_1 = 3, U_2 = 1\}.$$

There are sixteen possible outcomes for U_1 and U_2, so $\mathbb{P}(i = 4) = 3/16$. The other probabilities can be calculated in the same fashion, which gives:

$$\boxed{\begin{aligned}
f_{U_1 + U_2}(i) = \frac{1}{16} &[\mathbb{1}(i = 2) + 2\mathbb{1}(i = 3) + 3\mathbb{1}(i = 4) + 4\mathbb{1}(i = 5) \\
&+ 3\mathbb{1}(i = 6) + 2\mathbb{1}(i = 7) + \mathbb{1}(i = 8)]
\end{aligned}}$$

9.5 Suppose $X \sim \mathsf{Unif}(\{1, \ldots, 10\})$. What is the mode set of X?

Solution The density of X is $1/10$ for $i \in \{1, \ldots, 10\}$ and 0 otherwise. Hence the mode set is just $\boxed{\{1, \ldots, 10\}}$.

9.7 Suppose X has density $x^2 \exp(-x)\mathbb{1}(x \geq 0)$. Find the mode(s) of X.

Solution For $x < 0$, $f_X(x) = 0$, so the mode(s) cannot be there.

For $x \geq 0$,

$$[f_X(x)]' = [x^2 \exp(-x)]' = [x^2]' \exp(-x) + x^2[\exp(-x)]' = \exp(-x)(2x - x^2).$$

Since $\exp(-x) > 0$ for all x, this expression is positive when $2x - x^2 = x(2 - x) > 0$ and negative when $2x - x^2 < 0$. So it is positive for $x < 2$ and negative for $x > 2$. Hence the unique maximum (and so mode) is at $\boxed{2}$.

9.9 Suppose $X \sim \text{Exp}(1)$ and $Y \sim \text{Exp}(2)$ are independent.

    a) Find the survival function of X.

    b) Find the survival function of Y.

    c) Find $\mathbb{P}(\min(X, Y) \geq 2)$.

Solution

    a) This is

$$\begin{aligned}
S_X(a) &= \mathbb{P}(X > a) \\
&= \mathbb{P}(-\ln(U) > a) \\
&= \mathbb{P}(U < \exp(-a))
\end{aligned}$$

which is 1 for $a < 0$ and $\exp(-a)$ for $a \geq 0$. Hence

$$\boxed{S_X(a) = \mathbb{1}(a < 0) + \exp(-a)\mathbb{1}(a \geq 0).}$$

    b) This is similar to the X calculation:

$$\begin{aligned}
S_Y(a) &= \mathbb{P}(Y > a) \\
&= \mathbb{P}(-\ln(U)/2 > a) \\
&= \mathbb{P}(U < \exp(-2a))
\end{aligned}$$

which is 1 for $a < 0$ and $\exp(-2a)$ for $a \geq 0$. Hence

$$\boxed{S_Y(a) = \mathbb{1}(a < 0) + \exp(-2a)\mathbb{1}(a \geq 0).}$$

    c) Note that since X and Y are continuous, so is $\min(X, Y)$, and

$$\mathbb{P}(\min(X, Y) \geq 2) = \mathbb{P}(\min(X, Y) > 2) = S_{\min(X,Y)}(2).$$

Using

$$S_{\min(X,Y)}(a) = S_X(a)S_Y(a)$$

gives

$$\mathbb{P}(\min(X, Y) \geq 2) = \exp(-2)\exp(-2 \cdot 2) = \exp(-6) \approx \boxed{0.002478}.$$

10.1 Given that $\mathbb{P}(Y = 2) = 0.4$ and $\mathbb{P}(Y = -1) = 0.6$, what is $\mathbb{E}[Y]$?

Solution Sum the product of the outcomes times the probability of those outcomes to get

$$\mathbb{E}[Y] = (2)(0.4) + (-1)(0.6) = \boxed{0.2000}.$$

10.3 Suppose $\mathbb{P}(X = 2) = 0.3$, $\mathbb{P}(X = 4) = 0.2$ and $\mathbb{P}(X = 5) = 0.5$. What is $\mathbb{E}[X]$?

Solution Since X is a discrete random variable we sum the product of the different outcomes with the probabilities formed by those outcomes:

$$\mathbb{E}[X] = \sum_{x \in \{2,4,5\}} x\mathbb{P}(X = x) = (2)(0.3) + (4)(0.2) + (5)(0.5) = \boxed{3.900}.$$

10.5 Suppose $\mathbb{E}[X] = 34$. What is $\mathbb{E}[2X - 5]$?

Solution By the linearity of expectations, this is $2\mathbb{E}[X] - 5$, or $\boxed{63}$.

10.7 Say $X \sim \mathsf{Unif}(\{-2, -1, 0, 1, 2\})$. What is $\mathbb{E}[X]$?

Solution Since this distribution is symmetric about 0, and $\sum_{i=-2}^{2} i\mathbb{P}(X = i)$ is finite, the mean is $\boxed{0}$.

10.9 Say $\mathbb{E}[R] = 3$ and $\mathbb{E}[S] = 6$. What is $\mathbb{E}[R - S]$?

Solution By linearity, this is

$$\mathbb{E}[R - S] = \mathbb{E}[R] - \mathbb{E}[S] = 3 - 6 = \boxed{-3}.$$

10.11 Suppose $U_1, U_2, \ldots \sim \mathsf{Unif}(\{1, 2, 3, 4\})$. Show that $\lim_{n \to \infty}(U_1 + \cdots + U_n)/n = 2.5$ with probability 1.

Solution Note

$$\mathbb{E}[U] = (1/4)(1) + (1/4)(2) + (1/4)(3) + (1/4)(4) = 10/4 = 2.5.$$

Since each U_i has mean 2.5, the Strong Law of Large Numbers immediately gives us that

$$\lim_{n \to \infty} \frac{U_1 + \cdots + U_n}{n} = 2.5$$

with probability 1.

11.1 For X with density $12s^2(1 - s)\mathbb{1}(s \in [0, 1])$, find $\mathbb{E}[X]$.

Solution This is

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} s f_X(s) \, ds \\
&= \int_{-\infty}^{\infty} s \cdot 12s^2(1 - s)\mathbb{1}(0 \le s \le 1) \, ds \\
&= \int_0^1 12(s^3 - s^4) \, ds \\
&= 12(s^4/4 - s^5/5) \\
&= 12/20 = \boxed{0.6000}.
\end{aligned}$$

11.3 Suppose $U_1, U_2, \ldots \sim \mathsf{Unif}([0, 4])$. Show that $\lim_{n \to \infty}(U_1 + \cdots + U_n)/n = 2$ with probability 1.

Solution Since each U_i has mean 2, the Strong Law of Large Numbers immediately gives us that

$$\lim_{n \to \infty} \frac{U_1 + \cdots + U_n}{n} = 2$$

with probability 1.

11.5 For Z with density
$$f_Z(z) = \tau^{-1/2} \exp(-z^2/2),$$
verify using the integral that $\mathbb{E}[Z] = 0$.

Solution Here
$$\mathbb{E}[Z] = \int_{z=-\infty}^{\infty} z\tau^{-1/2} \exp(-z^2/2) \, dz$$
$$= \lim_{a \to -\infty} \lim_{b \to \infty} \int_a^b \tau^{-1/2} z \exp(-z^2/2) \, dz$$
$$= \lim_{a \to -\infty} \lim_{b \to \infty} \tau^{-1/2}(-\exp(-z^2/2))|_a^b$$
$$= \lim_{a \to -\infty} \lim_{b \to \infty} \tau^{-1/2}(\exp(-a^2/2) - \exp(-b^2/2))$$
$$= 0$$

and the result is shown.

11.7 Suppose $\mathbb{P}(X = -1) = 0.3$ and $\mathbb{P}(X = 1) = 0.7$. What is $\mathbb{E}[X^2]$?

Solution This is $(-1)^2(0.3) + 1^2(0.7) = \boxed{1}$. Of course, since X^2 always equals 1, this is trivially true!.

11.9 Let X have density $s \exp(-s^2/2)\mathbb{1}(s \geq 0)$. Find $\mathbb{E}[X^2]$.

Solution The integral is
$$\mathbb{E}[X^2] = \int_{s\in\mathbb{R}} s^2 \cdot s \exp(-s^2/2)\mathbb{1}(s \geq 0) \, ds = \int_{s\geq 0} s^3 \exp(-s^2/2) \, ds$$

To solve, first we want to get rid of the nonlinearity inside the exponential function. So let $t = s^2/2$, then $dt = s \, ds$, and
$$\mathbb{E}[X^2] = \int_{t\geq 0} s^2 \exp(-t) \, dt$$
$$= \int_{t\geq 0} 2t \exp(-t) \, dt$$
$$= 2 \int_{t\geq 0} t \exp(-t) \, dt.$$

We could use integration by parts at this point on this last integral , or we could recognize that this is the expected value of an exponential random variable of rate 1, so the overall answer is $\boxed{2}$.

11.11 Build a random variable W such that $\mathbb{E}[W] = I$, where
$$I = \int_{-1}^{1} 2x^2 \, dx.$$

Solution Let $U \sim \mathsf{Unif}([-1, 1])$. Then $f_U(x) = (1/2)\mathbb{1}(x \in [-1, 1])$. So
$$\mathbb{E}[4U^2] = \int_{-1}^{1} (4x^2)(1/2) \, dx = I.$$

Hence $\boxed{4U^2}$ works here.

11.13 For a random variable A, the *mean absolute deviation* of A is defined as

$$\mathrm{MAD}(A) = \mathbb{E}\left[|A - \mathbb{E}[A]|\right].$$

Let $A \sim \mathsf{Exp}(\lambda)$. Find $\mathrm{MAD}(A)$.

Solution The density of A is $\lambda \exp(-\lambda a)\mathbb{1}(a \geq 0)$ and these random variables have mean $1/\lambda$. Hence

$$\mathrm{MAD}(A) = \int_{s \in \mathbb{R}} |s - 1/\lambda|\lambda \exp(-\lambda s)\mathbb{1}(s \geq 0)\, ds$$

Since $|s - 1/\lambda| = s - 1/\lambda$ for $s \in [1/\lambda, \infty)$, and $|s - 1/\lambda| = 1/\lambda - s$ for $s \in [0, 1/\lambda)$, we have

$$\mathrm{MAD}(A) = I_1 + I_2,$$
$$I_1 = \int_{s \in [0,1/\lambda]} (1/\lambda - s)\lambda \exp(-\lambda s)\, ds,$$
$$I_2 = \int_{s \in [1/\lambda,\infty)} (s - 1/\lambda)\lambda \exp(-\lambda s)\, ds.$$

Using integration by parts on I_1 gives

$$I_1 = \int_{s \in [0,1/\lambda]} (1/\lambda - s)[-\exp(-\lambda s)]'\, ds$$
$$= \int_{s \in [0,1/\lambda]} [(1/\lambda - s)(-\exp(-\lambda s))]' - [(1/\lambda - s)]'[-\exp(-\lambda s)]\, ds$$
$$= \int_{s \in [0,1/\lambda]} [(1/\lambda - s)(-\exp(-\lambda s))]' - \exp(-\lambda s)\, ds$$
$$= [(1/\lambda - s)(-\exp(-\lambda s)) + \exp(-\lambda s)/\lambda]\,|_0^{1/\lambda}$$
$$= s\exp(-\lambda s))|_0^{1/\lambda}$$
$$= \exp(-1)/\lambda.$$

Using Integration by parts on I_2:

$$I_2 = \int_{s \in [1/\lambda,\infty)} -(1/\lambda - s)\lambda \exp(-\lambda s)\, ds$$
$$= -s\exp(-\lambda s))|_{1/\lambda}^{\infty}$$
$$= \exp(-1)/\lambda.$$

Note that we did not need to work through the antiderivative again because the integrand was the negative of the integrand that we found to compute I_1.

Putting it together, we get

$$\mathrm{MAD}(A) = \boxed{2\lambda^{-1}\exp(-1)}$$

11.15 Three zombies are chasing you. Each runs at a speed that is independent of the other, and a continuous uniform between 6.0 and 11.0 miles per hour.

a) If you can run at 10.0 miles per hour, what is the chance that you will get away from the zombies?

b) What is the expected speed of the fastest zombie?

Solution

a) If Z_1, Z_2, Z_3 are iid $\mathsf{Unif}([6, 11])$, then

$$
\begin{aligned}
\mathbb{P}(\max\{Z_1, Z_2, Z_3\} < 10) &= \mathbb{P}(Z_1 < 10, Z_2 < 10, Z_3 < 10) \\
&= \mathbb{P}(Z_1 \le 10)\mathbb{P}(Z_2 \le 10)\mathbb{P}(Z_3 \le 10) \\
&= \left(\frac{10 - 6}{11 - 6}\right)^3 = \boxed{0.5120}.
\end{aligned}
$$

b) It will be easier if we make

$$Z_i = 5U_i + 6,$$

where U_1, U_2, U_3 are iid $\mathsf{Unif}([0, 1])$. Then

$$\max\{Z_1, Z_2, Z_3\} = \max\{5U_1 + 6, 5U_2 + 6, 5U_3 + 6\} = 5 \max\{U_1, U_2, U_3\} + 6.$$

Linearity of expectation gives

$$\mathbb{E}[\max\{Z_1, Z_2, Z_3\}] = 5\mathbb{E}[\max\{U_1, U_2, U_3\}] + 6.$$

Now $Y = \max\{U_1, U_2, U_3\}$ has cdf

$$\mathrm{cdf}_Y(a) = \mathbb{P}(Y \le a) = \mathbb{P}(U_1 \le a)^3 = a^3$$

for $a \in [0, 1]$. Hence the derivative is

$$\mathrm{pdf}_Y(a) = 3a^2 \mathbb{1}(a \in [0, 1]),$$

and the expected value is

$$\mathbb{E}[Y] = \int_{\mathbb{R}} a \cdot 3a^2 \mathbb{1}(a \in [0, 1])\, da = \int_0^1 3a^3\, da = \frac{3}{4}a^4\big|_0^1 = \frac{3}{4}.$$

Therefore,
$$\mathbb{E}[\max\{Z_1, Z_2, Z_3\}] = 5 \cdot 34 + 6 = \boxed{9.750}$$

11.17 Let $U \sim \mathsf{Unif}([0, 2])$.

a) Find the cdf of $X = U^3$.

b) Find the density of X.

c) Find $\mathbb{E}[X]$.

Solution

a) The cdf of X is

$$\text{cdf}_X(a) = \mathbb{P}(X \leq a) = \mathbb{P}(U^3 \leq a) = \mathbb{P}(U \leq a^{1/3})$$

$$= \frac{a^{1/3}}{2-0} \mathbb{1}(a^{1/3} \in [0,2]) + \mathbb{1}(a^{1/3} > 2)$$

$$= \boxed{\frac{a^{1/3}}{2} \mathbb{1}(a \in [0,8]) + \mathbb{1}(a > 8).}$$

b) Differentiating gives the density

$$\boxed{\text{pdf}_X(a) = \frac{1}{6} a^{-2/3} \mathbb{1}(a \in [0, 0.8])}.$$

c) Once you set up the integral, it becomes a question of Calculus.

$$\mathbb{E}(X) = \int_{\mathbb{R}} a \cdot \frac{1}{6} a^{-2/3} \mathbb{1}(a \in [0,8]) \, da$$

$$= \int_0^8 \frac{1}{6} a^{1/3} \, da$$

$$= \frac{1}{6} \frac{a^{4/3}}{4/3} \bigg|_0^8 = \frac{1}{8} 8^{4/3} = 8^{1/3} = \boxed{2}.$$

11.19 Suppose $A \sim \text{Exp}(3)$, so A has density

$$f_A(s) = 3 \exp(-3s) \mathbb{1}(s \geq 0).$$

The density of an exponential is the multiplicative inverse of the rate, so $\mathbb{E}[A] = 1/3$.

a) What is $\mathbb{E}[2A - 1]$?
b) What is $\mathbb{E}[\exp(1.5A)]$?
c) What is the density of $2A - 1$?

Solution

a) Using linearity $2(1/3) - 1 = -1/3 = \boxed{-0.3333\ldots}$.
b) Using the law of the unconscious statistician

$$\mathbb{E}[\exp(1.5A)] = \int_{-\infty}^{\infty} \exp(1.5a) 3 \exp(-3a) \mathbb{1}(a \geq 0) \, da$$

$$= \int_0^{\infty} 3 \exp(-1.5a) \, da$$

$$= -(3/1.5) \exp(-1.5a)|_0^{\infty} = \boxed{2}.$$

c) Using the rules for shifting and scaling:

$$f_{2A-1}(s) = (1/|2|) f_A((s - (-1))/2).$$

Note $\mathbb{1}((s+1)/2 \geq 0) = \mathbb{1}(s \geq -1)$, so

$$f_{2A-1}(s) = \boxed{(3/2) \exp(-(3/2)(s+1)) \mathbb{1}(s \geq -1)}$$

12.1 Suppose that B_1, B_2 are iid $\text{Bern}(0.3)$. Say $\mathbb{P}(N = 1) = 0.6$ and $\mathbb{P}(N = 2) = 0.4$.

a) Find the density of
$$S = \sum_{i=1}^{N} B_i.$$

b) Find $\mathbb{E}[S]$ using the density.

c) Find $\mathbb{E}[S]$ using the Fundamental Theorem of Probability.

Solution

a) Adding one or two Bernoulli's gives either 0, 1, or 2. Hence we need to find $\mathbb{P}(S = i)$ for 0, 1, or 2. Breaking up the cases gives
$$\mathbb{P}(S = 0) = \mathbb{P}(N = 1)\mathbb{P}(B_1 = 0) + \mathbb{P}(N = 2)\mathbb{P}(B_1 = B_2 = 0)$$
$$= (0.6)(0.7) + (0.4)(0.7)^2 = 0.616$$
$$\mathbb{P}(S = 1) = \mathbb{P}(N = 1)\mathbb{P}(B_1 = 1) + \mathbb{P}(N = 1)\mathbb{P}(B_1 + B_2 = 1)$$
$$= (0.6)(0.3) + (0.4)(2(0.3)(0.7)) = 0.348$$
$$\mathbb{P}(S = 2) = \mathbb{P}(N = 2)\mathbb{P}(B_1 = B_2 = 1)$$
$$= (0.4)(0.3)^2 = 0.036$$

Hence S has density
$$\boxed{f_S(i) = 0.616 \cdot \mathbb{1}(i = 0) + 0.348 \cdot \mathbb{1}(i = 1) + 0.036 \cdot \mathbb{1}(i = 2)}.$$

b) That makes the mean
$$0.616(0) + 0.348(1) + 0.036(2) = \boxed{0.4200}.$$

c) Note $\mathbb{E}[B_i] = 0.3$, so $\mathbb{E}[S|N] = 0.3N$. Hence
$$\mathbb{E}[S] = \mathbb{E}[\mathbb{E}[S|N]]$$
$$= \mathbb{E}[0.3N]$$
$$= 0.3(1(0.6) + 2(0.4))$$
$$= \boxed{0.4200}.$$

12.3 A party has either low attendance (20% chance), medium attendance (40% chance) or high attendance (40% chance). With low attendance the average revenue for the party is $-\$300$, with medium $\$500$, and with high $\$1000$.

Draw an expectation tree to calculate the average revenue from the party.

Solution The tree looks like this:
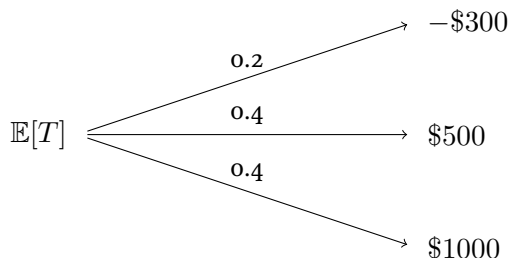
This makes the average revenue

$$(-300)(0.2) + (500)(0.4) + (1000)(0.4) = \boxed{540}.$$

12.5 Suppose the time until arrival of a customer (call it T) is an exponential random variable with rate parameter A (so $[T|A] \sim \text{Exp}(A)$.) A is a random variable that is uniform over the interval $[5, 10]$. What is $\mathbb{E}[T]$?

Solution Here $\mathbb{E}[T|A] = 1/A$, Hence

$$\begin{aligned}
\mathbb{E}[T] &= \mathbb{E}[\mathbb{E}[T|A]] \\
&= \mathbb{E}[1/A] \\
&= \int_{a \in \mathbb{R}} (1/a)(1/5)\mathbb{1}(a \in [5, 10]) \, da \\
&= \int_5^{10} 1/(5a) \, da \\
&= \ln(10) - \ln(5) = \ln(2) \approx \boxed{0.6931\dots}.
\end{aligned}$$

13.1 Suppose (X, Y) has density $(1/60)(x + 2y)\mathbb{1}(x \in [0, 2], y \in [0, 5])$.

    a) Find the marginal density of X.

    b) Find the marginal density of Y.

    c) Find $\mathbb{E}[XY]$.

Solution

    a) To find this, we integrate out y from the joint density:

$$\begin{aligned}
f_X(x) &= \int_{y \in \mathbb{R}} (1/60)(x + 2y)\mathbb{1}(x \in [0, 2], y \in [0, 5]) \, dy \\
&= \int_{y=0}^{5} (1/60)(x + 2y)\mathbb{1}(x \in [0, 2]) \, dy \\
&= (1/60)(xy + y^2)\mathbb{1}(x \in [0, 2])|_0^5 \\
&= (1/60)(5x + 25)\mathbb{1}(x \in [0, 2]) \\
&= \boxed{(1/12)(x + 5)\mathbb{1}(x \in [0, 2])}.
\end{aligned}$$

    b) To find this, we integrate out x from the joint density:

$$\begin{aligned}
f_Y(y) &= \int_{x \in \mathbb{R}} (1/60)(x + 2y)\mathbb{1}(x \in [0, 2], y \in [0, 5]) \, dy \\
&= \int_{x=0}^{2} (1/60)(x + 2y)\mathbb{1}(y \in [0, 5]) \, dx \\
&= (1/60)(x^2/2 + 2yx)\mathbb{1}(y \in [0, 5])|_0^2 \\
&= (1/60)(2 + 4y)\mathbb{1}(y \in [0, 5]) \\
&= \boxed{(1/30)(1 + 2y)\mathbb{1}(y \in [0, 5])}.
\end{aligned}$$

c) This integral is

$$\mathbb{E}[XY] = \int_{(x,y)\in\mathbb{R}^2} xy \cdot (1/60)(x+2y)\mathbb{1}(x \in [0,2], y \in [0,5]) \, d\mathbb{R}^2.$$

In this case, the integrand in always nonnegative, and so Tonelli allows us to break this into iterated integrals:

$$\begin{aligned}
\mathbb{E}[XY] &= \int_{x\in[0,2]} \int_{y\in[0,5]} (1/60)(x^2y + 2xy^2) \, dy \, dx \\
&= (1/60) \int_{x\in[0,2]} x^2y^2/2 + (2/3)(xy^3)|_0^5 \, dx \\
&= (1/60) \int_{x\in[0,2]} (25/2)x^2 + (250/3)x \, dx \\
&= (1/60) \left[(25/6)x^3 + (125/3)x^2 \right]\Big|_0^2 \\
&= (1/60) \left[(100/3) + (500/3) \right] \\
&= (1/60)(600/3) = 10/3 \approx \boxed{3.3333}
\end{aligned}$$

13.3 Suppose (X, Y) has density

$$f_{X,Y}(x, y) = (1/1260)x^3y^2\mathbb{1}(x \in \{1, 2, 3\})\mathbb{1}(y \in \{1, 3, 5\}).$$

a) Prove that X and Y are independent.

b) What is $\mathbb{P}(X = 2)$?

Solution

a) Note that

$$\begin{aligned}
f_{X,Y}(x, y) &= \frac{1}{1260}x^3y^2\mathbb{1}(x \in \{1, 2, 3\})\mathbb{1}(y \in \{1, 3, 5\} \\
&= \left[\frac{x^3}{36}\mathbb{1}(x \in \{1, 2, 3\}) \right] \left[\frac{y^2}{35}\mathbb{1}(y \in \{1, 3, 5\}) \right].
\end{aligned}$$

Since each factor inside brackets is a density, one only involving x and the other only involving y, then X and Y must be independent.

b) From the density of X, this is $2^3/36 = 2/9 = \boxed{0.2222\dots}$.

13.5 Suppose $(X_1, X_2) = (7.314, 2.103)$. What are the order statistics?

Solution These will be

$$\boxed{X_{(1)} = 2.103, \ X_{(2)} = 7.314}.$$

13.7 Suppose $(X_1, X_2) = (5.623, 5.623)$. What are the order statistics?

Solution These will be

$$\boxed{X_{(1)} = 5.623, \ X_{(2)} = 5.623}.$$

13.9 Suppose the order statistics $X_{(1)} = 1.3$ and $X_{(2)} = 3.4$. What possible states could the original vector (X_1, X_2) take on?

Solution The original vector could have been either of the two permutations of the numbers, so

$$\boxed{\{(1.3, 3.4), (3.4, 1.3)\}}.$$

14.1 Let $v = (-1, -1, 2)$ and $w = (5, 2, -3)$.

a) What is $v \cdot w$?

b) What is $\|v\|$?

Solution

a) This is

$$(-1)(5) + (-1)(2) + (2)(-3) = -5 - 2 - 6 = \boxed{-13}.$$

b) This is

$$\sqrt{(-1)^2 + (-1)^2 + (2)^2} = \sqrt{6} = \boxed{2.449\ldots}.$$

14.3 Say X is discrete with density $f_X(1) = 0.7$, $f_X(5) = 0.2$, $f_X(10) = 0.1$.

a) Find $\mathbb{E}[X]$.

b) Find $\mathrm{SD}[X]$.

Solution

a) This is

$$\mathbb{E}[X] = (0.7)(1) + (0.2)(5) + (0.1)(10) = \boxed{2.700}.$$

b) We need the second moment:

$$\mathbb{E}[X^2] = (0.7)(1)^2 + (0.2)(5)^2 + (0.1)(10)^2 = 15.7$$

Hence the standard deviation is

$$\mathrm{SD}(X) = \sqrt{15.7 - 2.7^2} = \boxed{2.900}.$$

14.5 Suppose $U \sim \mathsf{Unif}([0, 10])$.

a) What is the centered random variable U_c as a function of U?

b) What is the variance of U_c?

Solution

a) The mean of a uniform over an interval is the average of the endpoints of the interval, so $(0 + 10)/2 = 5$. That makes the centered random variable $\boxed{U_c = U - 5}$.

b) Note the variance of U_c is the same as the variance of U. The variance is the expected value of the square of the random variable minus the square of the expected value. Here

$$\mathbb{E}[U^2] = \int_{x \in \mathbb{R}} x^2 (1/10) \mathbb{1}(x \in [0, 10]) \, dx = 10^3 / [3(10)] = 100/3$$

so

$$\mathbb{V}(U) = 100/3 - 5^2 = 25/3 = \boxed{8.333\ldots}$$

14.7 Let (X, Y) be uniform over

$$A = \{(x, y) : x \ge 0, y \ge 0, x + y \le 1\}.$$

Find $\mathrm{Cov}(X, Y)$.

Solution To solve this, we need $\mathbb{E}[XY]$, $\mathbb{E}[X]$ and $\mathbb{E}[Y]$. This leads to four integrals.

The first integral is needed to find the measure of the set A.

$$\begin{aligned}
\ell(A) &= \int_{(x,y)} \mathbb{1}\mathbb{1}(x \ge 0, y \ge 0, x + y \le 1) \, d\mathbb{R}^2 \\
&= \int_{x \in [0,1]} \int_{y \in [0, 1-x]} 1 \, dy \, dx \\
&= \int_{x \in [0,1]} 1 - x \, dx \\
&= x - x^2/2 \big|_0^1 = 1/2,
\end{aligned}$$

therefore the joint density of X and Y is

$$f_{(X,Y)}(x, y) = 2 \cdot \mathbb{1}((x, y) \in A).$$

Now to find the expectations.

$$\begin{aligned}
\mathbb{E}[XY] &= \int_{(x,y)} xy \cdot 2 \mathbb{1}(x \ge 0, y \ge 0, x + y \le 1) \, d\mathbb{R}^2 \\
&= \int_{x \in [0,1]} \int_{y \in [0, 1-x]} 2xy \, dy \, dx \\
&= \int_{x \in [0,1]} xy^2 \big|_0^{1-x} \, dx \\
&= \int_{x \in [0,1]} [x - 2x^2 + x^3] \, dx \\
&= 1/2 - 2/3 + 1/4 = 1/12.
\end{aligned}$$

The next integral is

$$\mathbb{E}[X] = \int_{(x,y)} x \cdot 2\mathbb{1}\left(x \geq 0, y \geq 0, x + y \leq 1\right) d\mathbb{R}^2$$

$$= \int_{x\in[0,1]} \int_{y\in[0,1-x]} 2x \, dy \, dx$$

$$= \int_{x\in[0,1]} 2x(1 - x) \, dx$$

$$= \int_{x\in[0,1]} 2x - 2x^2 \, dx$$

$$= 1 - 2/3 = 1/3.$$

The last integral is

$$\mathbb{E}[Y] = \int_{(x,y)} y \cdot 2\mathbb{1}\left(x \geq 0, y \geq 0, x + y \leq 1\right) d\mathbb{R}^2$$

$$= \int_{x\in[0,1]} \int_{y\in[0,1-x]} 2y \, dy \, dx$$

$$= \int_{x\in[0,1]} (1 - x)^2 \, dx$$

$$= (1 - x)^3/(-3)|_0^1 \, dx$$

$$= 1/3.$$

Hence the covariance is

$$\frac{1}{12} - \frac{1}{9} = -\frac{1}{36} = \boxed{-0.02777\ldots}.$$

14.9 True or false: a random variable with finite mean always has a finite standard deviation.

Solution This is $\boxed{\text{false}}$. Consider $X = 1/\sqrt{U}$, where $U \sim \mathsf{Unif}([0,1])$. Then

$$\mathbb{E}[X] = \int_{u\in\mathbb{R}} u^{-1/2}\mathbb{1}\left(u \in [0,1]\right) du$$

$$= \int_0^1 u^{-1/2} \, du = u^{1/2}/(1/2)|_0^1 = 2.$$

However,

$$\mathbb{E}[X^2] = \int_{u\in real} (u^{-1/2})^2\mathbb{1}\left(u \in [0,1]\right) du$$

$$= \int_0^1 u^{-1} \, du = \ln(u)|_0^1 = \infty.$$

So while the mean is finite, the standard devation of X is not.

One note: it is true that whenever X has finite mean, the mean absolute deviation $\mathbb{E}[|X - \mathbb{E}[X]|]$ will be finite.

14.11 For a random variable with finite mean μ, finite standard deviation σ, and finite third
moment $\mathbb{E}[X^3]$, the *skewness* of the random variable is defined as

$$\text{skew}(X) = \mathbb{E}\left[\frac{(X - \mu)^3}{\sigma^3}\right].$$

    a) If X has skewness 3, what is the skewness of $2X$?

    b) What is the skewness of $-2X$?

Solution

    a) The mean of $2X$ is 2μ, and the standard deviation is $|2|\sigma = 2\sigma$. Hence the skewness
is

$$\text{skew}(2X) = \mathbb{E}\left[\frac{(2X - 2\mu)^3}{(2\sigma)^3}\right]$$
$$= \mathbb{E}\left[\frac{(X - \mu)^3}{\sigma^3}\right] = \boxed{3}.$$

    b) Here $\mathbb{E}[-2X] = -2\mu$, and the standard deviation is $|-2|\sigma = 2\sigma$, so

$$\text{skew}(-2X) = \mathbb{E}\left[\frac{(-2X + 2\mu)^3}{(2\sigma)^3}\right]$$
$$= -\mathbb{E}\left[\frac{(X - \mu)^3}{\sigma^3}\right] = \boxed{-3}.$$

14.13 Find the skewness of $U \sim \text{Unif}([0, 1])$.

Solution The skewness is

$$\mathbb{E}\left[\left(\frac{U - \mu}{\sigma}\right)^3\right] = \sigma^{-3}\mathbb{E}[(U - \mu)^3].$$

For a uniform on $[0, 1]$, $\mu = 1/2$ and $\sigma = 12^{1/2}$. Also, $Y = (U - 1/2) \sim$
$\text{Unif}([-1/2, 1/2])$ so

$$12^{-3/2}\mathbb{E}[Y^3] = 12^{-3/2}\int_{-1/2}^{1/2} s^3\,ds = 12^{-3/2}s^4/4\big|_{-1/2}^{1/2} = 12^{-3/2}[2^{-5} - 2^{-5}] = \boxed{0}.$$

The skewness is $\boxed{0}$ because the uniform distribution is symmetric around its mean.

14.15 Topper Building Co. suffers a number of delays that is uniform over $\{0, 1, 2, 3, 4\}$. Each
delay costs the builder an amount of time that is exponential with parameter 0.3 per
month. Find the expectation and variance of the total delay time.

Solution There are two random variables of interest here:

$$N := \text{the number of delays}$$
$$T := \text{the total sum of the delays.}$$

The problem states that $N \sim \mathsf{Unif}(\{0, 1, 2, 3, 4\})$. Given N, T is the sum of N independent exponential random variables. This gives a gamma distribution:

$$[T|N] \sim \mathsf{Gamma}(N, 0.3).$$

To find $\mathbb{E}[T]$, condition on N and then undo the conditioning by taking the expectation again:

$$\begin{aligned}
\mathbb{E}[T] &= \mathbb{E}[\mathbb{E}[T|N]] \\
&= \mathbb{E}[N/0.3] \\
&= (1/0.3)\mathbb{E}[N] \\
&= 2/0.3 \approx \boxed{6.667}.
\end{aligned}$$

To find the variance requires $\mathbb{E}[T^2]$. Another fact that comes in handy here is that for any random variable X, $\mathbb{E}[X^2] = \mathbb{V}(X) + \mathbb{E}[X]^2$.

So

$$\begin{aligned}
\mathbb{E}[T^2] &= \mathbb{E}[\mathbb{E}[T^2|N]] \\
&= \mathbb{E}[\mathbb{V}(T|N) + \mathbb{E}[T|N]^2] \\
&= \mathbb{E}[N/0.3^2 + (N/0.3)^2] \\
&= (1/0.3)^2 \mathbb{E}[N + N^2] \\
&= (1/0.3)^2[2 + (1/5)0^2 + (1/5)1^2 + (1/5)2^2 + (1/5)3^2 + (1/5)4^2] \\
&= 88.89.
\end{aligned}$$

Then

$$\mathbb{V}(T) = \mathbb{E}[T^2] - \mathbb{E}[T]^2 \approx \boxed{44.44}.$$

15.1 For $(X, Y) \sim \mathsf{Unif}(\{(0, 0), (0, 2), (1, 2)\})$, find the correlation between X and Y.

Solution We need $\mathbb{E}[XY]$, $\mathbb{E}[X]$, $\mathbb{E}[Y]$, $\mathbb{E}[X^2]$ and $\mathbb{E}[Y^2]$ to solve this problem.

$$\begin{aligned}
\mathbb{E}[XY] &= (1/3)[(0)(0) + (0)(2) + (1)(2)] = 2/3 \\
\mathbb{E}[X] &= (1/3)[(0) + (0) + (1)] = 1/3 \\
\mathbb{E}[Y] &= (1/3)[(0) + (2) + (2)] = 4/3 \\
\mathbb{E}[X^2] &= (1/3)[(0)^2 + (0)^2 + (1)^2] = 1/3 \\
\mathbb{E}[Y^2] &= (1/3)[(0)^2 + (2)^2 + (2)^2] = 8/3
\end{aligned}$$

Hence

$$\mathrm{Cor}(X, Y) = \frac{(2/3) - (1/3)(4/3)}{\sqrt{(1/3) - (1/3)^2} \cdot \sqrt{(8/3) - (4/3)^2}} = 1/2 = \boxed{0.5000}.$$

15.3 For (A, B) with density

$$f_{(A,B)}(a, b) = (2/3)(a + 2b)\mathbb{1}(a \in [0, 1])\mathbb{1}(b \in [0, 1]),$$

find $\mathrm{Cor}(A, B)$.

Solution We need $\mathbb{E}[AB]$, $\mathbb{E}[A]$, $\mathbb{E}[B]$, $\mathbb{E}[A^2]$ and $\mathbb{E}[B^2]$ to solve this problem. Fortunately, these integrands are all positive, and so Tonelli can be used to do the calculations with iterated integrals.

$$\mathbb{E}[AB] = \int_{a=0}^{1} \int_{b=0}^{1} (ab)(2/3)(a+b)\, db\, da = 1/3$$

$$\mathbb{E}[A] = \int_{a=0}^{1} \int_{b=0}^{1} a(2/3)(a+b)\, db\, da = 5/9$$

$$\mathbb{E}[B] = \int_{a=0}^{1} \int_{b=0}^{1} b(2/3)(a+b)\, db\, da = 11/18$$

$$\mathbb{E}[A^2] = \int_{a=0}^{1} \int_{b=0}^{1} a^2(2/3)(a+b)\, db\, da = 7/18$$

$$\mathbb{E}[B^2] = \int_{a=0}^{1} \int_{b=0}^{1} b^2(2/3)(a+b)\, db\, da = 4/9$$

Hence

$$\mathrm{Cor}(X, Y) = \frac{1/3 - (5/9)(11/18)}{\sqrt{(7/18) - (5/9)^2} \cdot \sqrt{(4/9) - (11/18)^2}} = \boxed{-0.08178\dots}.$$

15.5 Consider random variables X and Y with joint density

$$f_{(X,Y)}(x, y) = C \exp(-x - xy - y)\mathbb{1}(x \geq 0, y \geq 0).$$

a) Numerically find the covariance of X and Y?

b) Numerically find is the correlation of X and Y?

Solution Before finding the expected values needed for parts a and b, it is necessary to find C. The integrand is positive, so by Tonelli's theorem,

$$C^{-1} = \int_{x=0}^{\infty} \int_{y=0}^{\infty} \exp(-x - xy - y)\, dy\, dx.$$

This integral does not have a solution with elementary functions, but with the exponential integral function, it evaluates to $-e\,\mathrm{Ei}(-1)$, so

$$C \approx 1.67688.$$

Now numerically solve the integrals needed for the problem.

a) The integrals needed for this problem are as follows (all solved via Wolfram Alpha)

$$\mathbb{E}[XY] = \int_{x=0}^{\infty} \int_{y=0}^{\infty} xy1.67688 \exp(-x - xy - x)\, dx\, dy \approx 0.323126$$

$$\mathbb{E}[X] = \int_{x=0}^{\infty} \int_{y=0}^{\infty} x1.67688 \exp(-x - xy - x)\, dx\, dy \approx 0.676877$$

$$\mathbb{E}[Y] = \int_{x=0}^{\infty} \int_{y=0}^{\infty} y1.67688 \exp(-x - xy - x)\, dx\, dy \approx 0.676877$$

So

$$\mathrm{Cov}(X, Y) \approx 0.323126 - (0.676877)^2 \approx \boxed{-1.350}.$$

b) For the correlation, we need a few more facts..

$$\mathbb{E}[X^2] = \int_{x=0}^{\infty} \int_{y=0}^{\infty} x^2 1.67688 \exp(-x - xy - x) \, dx \, dy = 1$$

$$\mathbb{E}[Y^2] = \int_{x=0}^{\infty} \int_{y=0}^{\infty} y^2 1.67688 \exp(-x - xy - x) \, dx \, dy = 1$$

Hence $\text{SD}(X) = \text{SD}(Y) = (1 - 0.676877^2)^{1/2}$, which makes the correlation

$$\text{Cor}(X, Y) = \frac{-1.35037}{1 - 0.676877^2} \approx \boxed{-0.2492}.$$

16.1 Let $A \sim \text{Unif}([0, 1])$ and $B \sim \text{Unif}([0, 2])$ be independent. Find the density of $A + B$.

Solution Here $f_A(a) = \mathbb{1}(a \in [0, 1])$, and $f_B(b) = (1/2)\mathbb{1}(b \in [0, 2])$. Hence the convolution of the densities is

$$\begin{aligned} f_{A+B}(s) &= \int_a f_A(a) f_B(s - a) \, da \\ &= \int_a (1/2)\mathbb{1}(s - a \in [0, 2], a \in [0, 1]) \, da \\ &= \int_a (1/2)\mathbb{1}(a \le s, a \ge s - 2, a \in [0, 1]) \, da. \end{aligned}$$

When $s \le 0$ or $s \ge 3$ the indicator is always 0. When $s \in [0, 1]$,

$$\mathbb{1}(a \le s, a \ge s - 2, a \in [0, 1]) = \mathbb{1}(a \in [0, s])$$

and the integral is $s/2$.

When $s \in (1, 3]$,

$$\mathbb{1}(a \le s, a \ge s - 2, a \in [0, 1]) = \mathbb{1}(a \in [s - 2, 1])$$

and the integral is $[1 - (s - 2)]/2 = (3 - s)/2$. Therefore the final density is

$$f_{A+B}(s) = (s/2)\mathbb{1}(s \in [0, 1]) + [(3 - s)/2]\mathbb{1}(s \in (1, 3]).$$

16.3 Suppose $X \sim \text{Unif}(\{1, 2, 3\})$ and $Y \sim \text{Unif}(\{3, 5\})$ are independent. What is the density of $X + Y$?

Solution This will be

$$\begin{aligned} [f * g](i) &= \sum_j f_X(j) f_Y(i - j) \\ &= \sum_{j \in \{1,2,3\}} (1/3)\mathbb{1}(j \in \{1, 2, 3\})(1/2)\mathbb{1}(i - j \in \{3, 5\}) \\ &= (1/6)[\mathbb{1}(i - 1 \in \{3, 5\})) + \mathbb{1}(i - 2 \in \{3, 5\}) + \mathbb{1}(i - 3 \in \{3, 5\})] \\ &= \boxed{(1/6)(\mathbb{1}(i = 4) + \mathbb{1}(i = 5) + 2\mathbb{1}(i = 6) + \mathbb{1}(i = 7) + \mathbb{1}(i = 8))} \end{aligned}$$

16.5 Suppose R and G are discrete random variables where $R \sim \text{Bern}(0.3)$ and $G \sim \text{Geo}(0.6)$. So

$$f_R(i) = (0.3)\mathbb{1}(i = 1) + (0.7)\mathbb{1}(i = 0), \ \ f_G(i) = (0.6)(0.4)^{i-1}\mathbb{1}(i \in \{1, 2, \ldots\}.$$

Find the density of $R + G$.

Solution We know $R + G$ has density equal to the convolution of their densities:

$$\begin{aligned}
f_{R+G}(i) &= [f_R * f_G](i) \\
&= \sum_a f_R(a) f_G(i - a) \\
&= \sum_{a \in \{0,1\}} [0.3\mathbb{1}(a = 1) + 0.7\mathbb{1}(a = 0)](0.6)(0.4)^{i-a-1}\mathbb{1}(i - a \in \{1, 2, \ldots\}) \\
&= f_{a=0}(i) + f_{a=1}(i),
\end{aligned}$$

where

$$\begin{aligned}
f_{a=0}(i) &= (0.3)(0.6)(0.4)^{i-2}\mathbb{1}(i \in \{2, 3, \ldots\}) \\
f_{a=1}(i) &= (0.7)(0.6)(0.4)^{i-1}\mathbb{1}(i \in \{1, 2, 3, \ldots\}.
\end{aligned}$$

Note $f_{a=0}(1) = 0$, and $f_{a=1}(1) = (0.7)(0.6) = 0.42$, so

$$f_{R+G}(i) = 0.42\mathbb{1}(i = 1) + h(i)\mathbb{1}(i \in \{2, 3, \ldots\}),$$

where for $i \in \{2, 3, \ldots\}$,

$$\begin{aligned}
h(i) &= (0.3)(0.6)(0.4)^{i-2} + (0.7)(0.6)(0.4)^{i-1} \\
&= (0.3)(0.6)(0.4)^{i-2} + (0.7)(0.6)(0.4)(0.4)^{i-2} \\
&= 0.348(0.4)^{i-2}.
\end{aligned}$$

Putting this all together gives

$$\boxed{f_{R+G}(i) = 0.42\mathbb{1}(i = 1) + 0.348(0.4)^{i-2}\mathbb{1}(i \in \{2, 3, \ldots\}).}$$

17.1 Suppose $\mathbb{P}(X = 1) = \mathbb{P}(X = 2) = 0.5$. What is $\text{mgf}_X(t)$?

Solution This is

$$\mathbb{E}[\exp(tX)] = \boxed{(1/2)e^t + (1/2)e^{2t}}.$$

17.3 Suppose X has moment generating function

$$\text{mgf}_X(t) = 0.1 \exp(10t) + 0.9 \exp(-5t).$$

What is the density of X?

Solution Here $X \in \{-5, 10\}$ since there is an $\exp(-5t)$ and $\exp(10t)$ term. The coefficients then give the probabilities for a density of

$$\boxed{f_X(i) = 0.9\mathbb{1}(i = -5) + 0.1\mathbb{1}(i = 10).}$$

17.5 Let $Z_1, Z_2, \ldots, Z_n$ be iid $\mathsf{N}(0,1)$. Recall for $Z \sim \mathsf{N}(0,1)$, $\mathrm{mgf}_Z(t) = \exp(t^2/2)$.

    a) What is the the moment generating function of $Z_1 + Z_2$?

    b) What is the moment generating function of:

$$\frac{Z_1 + \cdots + Z_n}{n}.$$

    c) What is the moment generating function of:

$$\frac{Z_1 + \cdots + Z_n}{\sqrt{n}}.$$

Solution

    a) Use the fact that

$$\mathrm{mgf}_{Z_1+Z_2}(t) = \mathrm{mgf}_{Z_1}(t)\,\mathrm{mgf}_{Z_2}(t) = \exp(t^2/2)\exp(t^2/2) = \boxed{\exp(t^2)}.$$

    b) Consider the moment generating function of a scaled random variable:

$$\mathrm{mgf}_{\alpha X}(t) = \mathbb{E}[e^{t\alpha X}] = \mathrm{mgf}_X(\alpha t).$$

So for our problem:

$$\mathrm{mgf}_{Z_1/n}(t) = \mathrm{mgf}_{Z_1}(t/n) = \exp((t/n)^2/2) = \exp(t^2/2n^2).$$

Now add n of them together:

$$\mathrm{mgf}_{(Z_1/n)+\cdots+(Z_n/n)}(t) = \mathrm{mgf}_{Z_1/n}(t)^n = \mathrm{mgf}_{Z_1}(t/n)^n = [\exp(t^2/2n^2)]^n$$
$$= \boxed{\exp(t^2/2n)}.$$

    c) The same as (b), but scaling by $1/\sqrt{n}$:

$$\mathrm{mgf}_{(Z_1/\sqrt{n})+\cdots+(Z_n/\sqrt{n})}(t) = \mathrm{mgf}_{Z_1/\sqrt{n}}(t)^n = \mathrm{mgf}_{Z_1}(t/\sqrt{n})^n = \exp(t^2/2n)^n$$
$$= \boxed{\exp(t^2/2)}.$$

This means when you add up n normal random variables with mean 0 and standard deviation 1, and divide by $\sqrt{n}$, you get a normal random variable with mean 0 and standard deviation 1. A normal random variable is a *fixed point* with respect to this operation.

The CLT says that if you add up any n random variables with mean 0 and standard deviation 1, and divide by $\sqrt{n}$, you get approximately a normal random variable with mean 0 and standard deviation 1. This is because such operations tend to converge to the fixed point, in this case, normal.

17.7 Suppose that X has the following density:

$$f_X(r) = \frac{3}{8}(r^3 - 8r^2 + 19r - 12)\mathbf{1}(r \in [1,3]).$$

a) Find the mode(s) of X.

b) Find the median(s) of X.

c) Find the mean of X.

d) Find $\mathbb{E}[e^{tX}]$.

Solution A plot of $f_X(r)$ is:
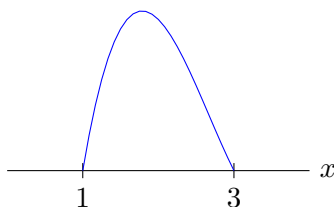
Note that it is skewed towards the left, this should make the mean, mode, and median a little to the left of 2 (because that is the middle of the interval $[1, 3]$ where the density is positive.)

a) To find the mode, maximize. The derivative of $f_X(r)$ with respect to r is $(9/8)r^2 - 6r + (57/8)$ which is a quadratic with exactly one zero in $[1, 3]$. This zero occurs at $\boxed{1.785}$, and since $f_X(r)$ is zero at $r = 1$ and $r = 3$, and positive at $r = 1.785$, this must be a maximum, and so is the mode.

b) The median will occur at the point where half the area under the density lies to the left of the point, and half lies to the right. There are many packages that do integrations and root finding, as well as most calculators. For Wolfram Alpha, the command:

```
1/2 = integrate (3/8)*(r^3-8*r^2+19*r-12) from 1 to a
```

yields $a \approx 0.42551$ and $a \approx 1.878$. Since the median must be in $[1, 2]$, $\boxed{1.878}$ is the answer.

c) The mean is even easier, since it just involves one integration:

$$\int_1^3 r(3/8)(r^3 - 8r^2 + 19r - 1)\, dr.$$

which yields $\boxed{1.900}$ as the answer.

Summarizing these first three parts:

$$
\begin{array}{rcl}
\text{(a)} & \text{mode} & 1.785 \\
\text{(b)} & \text{median} & 1.878 \\
\text{(c)} & \text{mean} & 1.900
\end{array}
$$

They are different because the density is not symmetric as in the case of a normal density.

d) The integral to be computed is

$$\int_1^3 e^{rt}(3/8)(r^3 - 8r^2 + 19r - 12)\, dt.$$

Wolfram Alpha returns the result

$$\text{mgf}_X(t) = \boxed{\frac{9 + 15t + 9e^t t^2 + e^{3t}(3t^2 + 3t - 9)}{4t^4}}.$$

Note that the Taylor series expansion of this expression is:

$$1 + \frac{19}{10}t + \frac{19}{10}t^2 + \cdots,$$

and so the value of $19/10 = 1.9$ matches the mean found earlier.

18.1 For Z a standard normal, find

$$\mathbb{P}(Z \in [-2, 2]).$$

Solution This is

$$\int_{-2}^{2} \frac{1}{\sqrt{\tau}} \exp(-s^2/2)\, ds$$

which can be approximately evaluated by WolframAlpha to be $\boxed{0.9545}$.

18.3 The Digital Life conference draws a number of attendees each year that is normally distributed with mean 59 000 and standard deviation 10 000. Independently, E3 draws a number of attendees that is normally distributed with mean 75 000 and standard deviation 5 000.

a) Suppose I average the two numbers. What is the distribution of the average?

b) What is the chance that the average of the two conferences is greater than 70 000?

c) What is the distribution of the number attending Digital Life minus the number attending E3?

d) What is the chance that more people attend Digital Life than E3?

Solution

a) Let A be the number of attendees at the Digital Life Conference, and B the number of attendees at E3. Then because they are independent normals, their sum and average and difference will also be normal. For convenience, we represent everything in units of 1000's. Then

$$\frac{A + B}{2} \sim \mathsf{N}((59 + 75)/2, (10^2 + 5^2)/4) \sim \boxed{\mathsf{N}(67, 31.25)}.$$

b) If we let $C \sim \mathsf{N}(67, 31.25)$, then

$$\mathbb{P}(C \geq 70) = \mathbb{P}(C - 67 \geq 3)$$
$$= \mathbb{P}\left(\frac{C - 67}{\sqrt{31.25}} \geq \frac{3}{\sqrt{31.25}}\right)$$
$$= \mathbb{P}\left(Z \geq \frac{3}{\sqrt{31.25}}\right)$$
$$= \boxed{29.57\%}.$$

c) For the difference, subtract the means (but still add the variances!)

$$A - B \sim \mathsf{N}(59 - 75, 10^2 + 5^2) \sim \boxed{\mathsf{N}(-16, 125).}$$

d) For this to happen $A - B \geq 0$.

$$\mathbb{P}(A - B \geq 0) = \mathbb{P}(A - B + 16 \geq 16)$$
$$= \mathbb{P}\left(\frac{A - B + 16}{\sqrt{125}} \geq \frac{16}{\sqrt{125}}\right)$$
$$= \mathbb{P}\left(Z \geq \frac{16}{\sqrt{125}}\right)$$
$$= \boxed{7.620\%}.$$

18.5 Suppose that $W_1, \ldots, W_n$ are iid standard normal random variables. What is the distribution of $W_1 + \cdots + W_n$?

Solution Any sum of independent normal random variables is another normal random variable. The new mean is the sum of n zero values, and the new variance is the sum of n one values, so
$$W_1 + \cdots + W_n \sim \boxed{\mathsf{N}(0, n)}.$$

19.1 Let $D_1, \ldots, D_8$ be iid rolls of a fair eight-sided die. Approximate the probability that $\sum D_i \geq 30$ using the CLT.

Solution First we calculate the mean and variance:

$$\mathbb{E}[D_i] = (1 + 8)/2 = 4.5.$$
$$\mathbb{V}[D_i] = ((8 - 1 + 1)^2 - 1)/12 = 63/12.$$

Next we standardize the sum

$$\mathbb{P}\left(\sum D_i \geq 17\right) = \mathbb{P}\left(\sum \frac{D_i - 4.5}{\sqrt{(8)(63/12)}} \geq \frac{30 - (4.5)(8)}{\sqrt{(8)(63/12)}}\right)$$
$$\approx \mathbb{P}\left(Z \geq \frac{30 - (4.5)(8)}{\sqrt{(8)(63/12)}}\right)$$
$$= \boxed{0.8227}$$

19.3 Suppose that R has density

$$f_R(r) = 2r \cdot \mathbb{1}(r \in [0, 1]).$$

a) What is the expected value of R?

b) What is the variance of R?

c) Say that $R_1, R_2, \ldots$ are independent random variables with the same distribution as R. Using the CLT, approximately what is

$$\mathbb{P}(R_1 + \cdots + R_{100} \geq 70)?$$

d) What is the expected value of R conditioned on $R \in [0.3, 0.5]$?

Solution

a) This is

$$\mathbb{E}[R] = \int_r r \cdot 2r \mathbb{1}(r \in [0,1]) \, dr = \int_{r=0}^1 2r^2 \, dr = 2/3 = \boxed{0.6666\dots}.$$

b) We will need the second moment of R for this:

$$\mathbb{E}[R^2] = \int_r r^2 \cdot 2r \mathbb{1}(r \in [0,1]) \, dr = \int_{r=0}^1 2r^3 \, dr = 1/2.$$

Hence the variance is

$$\mathbb{E}[R^2] - \mathbb{E}[R]^2 = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18} = \boxed{0.05555\dots}.$$

c) The CLT says that $p = \mathbb{P}(R_1 + \cdots + R_{100} \ge 70)$ is

$$p = \mathbb{P}\left(\frac{R_1 + \cdots + R_{100} - 100(2/3)}{\sqrt{1/18}\sqrt{100}} = \frac{70 - 100(2/3)}{\sqrt{100 \cdot 1/18}}\right)$$
$$\approx \mathbb{P}(Z \ge \sqrt{2}))$$
$$= \boxed{0.07864\dots}.$$

d) Conditioned to lie in $[0.3, 0.5]$, the density becomes unnormalized.

$$f_{R|R\in[0.3,0.5]}(r) \propto f_R(r)\mathbb{1}(r \in [0.3, 0.5]).$$

Normalizing gives

$$f_{R|R\in[0.3,0.5]}(r) = \frac{f_R(r)\mathbb{1}(r \in [0.3, 0.5])}{\int_{s\in[0.3,0.5]} f_R(s) \, ds} = \frac{2r}{0.16}\mathbb{1}(r \in [0.3, 0.5]).$$

That makes the conditional expectation

$$\mathbb{E}[R|R \in [0.3, 0.5]] = \int_r r \cdot \frac{2r}{0.16}\mathbb{1}(r \in [0.3, 0.5]) = \boxed{0.4083\dots}.$$

19.5 Let $U_1, U_2, \dots, U_{12}$ be standard uniform random variables (so uniform over $[0, 1]$.) Then approximate

$$\mathbb{P}(U_1 + \cdots + U_{12} < 7)$$

using the CLT.

Solution The mean of a standard uniform is

$$\mathbb{E}[U] = \int x\mathbb{1}(x \in [0,1]) \, dx = \int_0^1 x \, dx = 1/2,$$

while the second moment is

$$\mathbb{E}[U^2] = \int x^2\mathbb{1}(x \in [0,1]) \, dx = \int_0^1 x^2 \, dx = 1/3.$$

This makes the variance $1/3 - (1/2)^2 = 1/12$. Hence the variance of the sum is $12 \cdot (1/12) = 1$, and the mean is $12 \cdot (1/2) = 6$. Hence

$$\mathbb{P}(U_1 + \cdots + U_{12} < 7) = \mathbb{P}\left(\frac{U_1 + \cdots + U_{12} - 6}{\sqrt{1}} < \frac{7 - 6}{\sqrt{1}}\right)$$
$$\approx \mathbb{P}(Z < 1)$$
$$= \boxed{0.8413\dots}.$$

20.1 Suppose $X \sim \mathsf{Bin}(34, 0.23)$. What is $\mathbb{E}[X]$?

Solution The mean of a binomial random variable is the product of the parameters, so $(34)(0.23) = \boxed{7.820}$.

20.3 Let $G \sim \mathsf{Geo}(0.38)$.

    a) What is $\mathbb{E}[G]$?

    b) What is $\mathbb{V}[G]$

Solution

    a) The mean is $1/0.38 \approx \boxed{2.631}$.

    b) The variance is $(1 - 0.38)/0.38^2 \approx \boxed{4.293}$.

20.5 Let $N \sim \mathsf{NegBin}(20, 0.38)$.

    a) What is $\mathbb{E}[N]$?

    b) What is $\mathbb{V}[N]$

Solution

    a) The mean is $20/0.38 \approx \boxed{52.63}$.

    b) The variance is $20(1 - 0.38)/0.38^2 \approx \boxed{85.87}$.

20.7 Suppose $X \sim \mathsf{Bin}(13, 0.2)$ and $Y \sim \mathsf{Bin}(27, 0.2)$ are independent. What is the distribution of $X + Y$?

Solution X represents the number of successes on 13 independent trials (that are a success with probability 0.2). Y represents the number of successes on 27 independent trials (that are a success with probability 0.2). Together $X + Y$ represents the number of successes on 40 independent trials. So $X + Y \sim \mathsf{Bin}(40, 0.2)$.

20.9 Let Y be a positive integer valued random variable with $\mathbb{E}[Y] = 4.2$, and $[X|Y] = \mathsf{Bin}(Y, 0.3)$. Then what is $\mathbb{E}[X]$?

Solution Since $\mathbb{E}[X|Y] = (Y)(0.3)$,

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[0.3Y] = 0.3(4.2) = \boxed{1.260}.$$

20.11 Find $\mathbb{E}[G^2]$ for a geometric random variable by conditioning on B_1 and taking the expectation again.

Solution By the FTP

$$\begin{aligned}
\mathbb{E}[G^2] &= \mathbb{E}[\mathbb{E}[G^2|B_1]] \\
&= p(1)^2 + (1-p)\mathbb{E}[(1+G)^2] \\
&= p + (1-p)\mathbb{E}[1 + 2G + G^2] \\
&= p + (1-p)(1 + 2/p) + (1-p)\mathbb{E}[G^2].
\end{aligned}$$

Bringing the $(1-p)\mathbb{E}[G^2]$ over to the other side and dividing by p gives

$$\begin{aligned}
\mathbb{E}[G^2] &= 1 + \frac{1-p}{p}\left(1 + \frac{2}{p}\right) \\
&= \boxed{\frac{2-p}{p^2}}.
\end{aligned}$$

21.1 For P a PPP over $[0, \infty)$ of rate 2, what is the distribution of $\inf(P)$?

Solution The infimum of a set of points is the closest one to 0. So this is the distance from 0 to the first point, which is $\boxed{\mathsf{Exp}(2)}$ for this type of process.

21.3 Let P be a Poisson point process over $[0, \infty)$ of rate 1.8, and $P_1 = \inf(P)$. What is $\mathbb{P}(P_1 \le 1)$?

Solution Since $P_1 \sim \mathsf{Exp}(1.8)$, this is

$$\int_0^1 1.8\exp(-1.8s)\mathbb{1}(s \ge 0)\,ds = -\exp(-1.8s)|_0^1 = 1 - \exp(-1.8) \approx \boxed{0.8347}.$$

21.5 Suppose that times that a bus arrives at a stop over one hour is modeled as a Poisson point process of rate 1.4/hr.

a) What is the chance that exactly one bus arrives in the hour?

b) What is the expected number of buses that arrive in the hour?

c) What is the expected number of buses that arrive in the first half hour?

Solution

a) The number of buses will be Poisson with parameter 1.4, hence the chance of exactly one bus is
$$\mathbb{P}(N = 1) = \exp(-1.4)(1.4)^1/1! \approx \boxed{0.3452}.$$

b) The expected number in the first hour is $1(1.4) = \boxed{1.400}$.

c) The expected number in the first half hour is $(1/2)(1/4) = \boxed{0.7000}$.

21.7 For a Poisson point process over $[0, \infty)$ of rate λ, let $N_A = \#(P \cap A)$. Then find $\mathrm{Cov}(N_{[0,2)}, N_{[0,3)})$.

Solution Note that

$$N_{[0,3)} = N_{[0,2)} + N_{[2,3)},$$

so
$$\text{Cov}(N_{[0,2)}, N_{[0,3)}) = \text{Cov}(N_{[0,2)}, N_{[0,2)}) + \text{Cov}(N_{[0,2)}, N_{[2,3)})$$
$$= \mathbb{V}(N_{[0,2)}) + 0 = \boxed{2\lambda}.$$

22.1 Suppose N_1 and N_2 are independent Poisson random variables with means 2 and 3 respectively. What is the chance that $N_1 + N_2 = 5$?

Solution Their sum will be Poisson distributed with mean 5. Hence
$$\mathbb{P}(N_1 + N_2 = 5) = \exp(-5)\frac{5^5}{5!} = \boxed{0.1754\dots}.$$

22.3 EPA clean-up sites in a county are modeled as a Poisson point process of rate $\lambda = 3/\text{mi}^2$.

a) If the region has an area of 9 square miles, what is the expected number of clean-up sites?

b) If the region is known to have at least 25 clean up sites, what is the chance that it has at least 30 such sites? (Probably want to use a computer for the calculations on this one.)

Solution

a) The average number of clean-up sites will be the rate times the measure of the area, or
$$\frac{3}{\text{mi}^2} \cdot 9\text{mi}^2 = \boxed{27}.$$

b) This is $\mathbb{P}(N \geq 30 | N \geq 25)$, where $N \sim \text{Pois}(27)$. Using the conditional probability formula, this is
$$\frac{\mathbb{P}(N \geq 30, N \geq 25)}{\mathbb{P}(N \geq 25)} = \frac{\mathbb{P}(N \geq 30)}{\mathbb{P}(N \geq 25)} = \boxed{0.4535\dots},$$

where the last expression was evaluated in R using
```
(1 - ppois(29, 27))/(1 - ppois(24, 27))
```

22.5 Suppose that $P \sim \text{Pois}([0, 2], \lambda \cdot \ell)$, where $\lambda > 0$ is a constant and ℓ is Lebesgue measure. If $N_{[0,2]} = 10$, what is the chance that $N_{[0,1]} = 4$?

Solution We know that
$$\frac{\mu([0, 1]}{\mu[0, 2]} = \frac{\lambda(1 - 0)}{\lambda(2 - 0)} = \frac{1}{2}.$$

So
$$[N_{[0,1]} | N_{[0,2]} = 10] \sim \text{Bin}(10, 1/2),$$

and
$$\mathbb{P}(N_{[0,1]} | N_{[0,2]} = 10) = \binom{10}{4}\left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^6 = \boxed{0.2050\dots}.$$

22.7 Outbreaks of a disease are modeled as coming from a Poisson point process with rate 2.3 per square mile.

a) If the city is 3 square miles, what is the chance that there are exactly 6 outbreaks?

b) Suppose the part of the city west of the river is 1.2 square miles (leaving 1.8 square miles east of the river). If there are exactly 8 outbreaks across the city, what is the chance that at least 3 of them are on the west side of the river?

Solution

a) The number of points in the process will be Poisson with parameter $3 \cdot 2.3 = 6.9$. Hence

$$\mathbb{P}(N = 6) = \exp(-6.9)6.9^6/6! \approx \boxed{0.1510\ldots}.$$

b) If X is the number on the west side and N is the total number then

$$[X|N = 8] \sim \mathsf{Bin}(8, 1.2/(1.2 + 1.8)).$$

Let $p = 1.2/(1.2 + 1.8) = 0.4$.

$$\mathbb{P}(X \geq 3|N = 8) = 1 - \mathbb{P}(X \in \{0, 1, 2\}|N = 8)$$
$$= 1 - \binom{8}{0}(0.6)^8 + \binom{8}{1}(0.4)^1(0.6)^7 + \binom{8}{2}(0.4)^2(0.6)^6$$
$$= \boxed{0.6846\ldots}.$$

23.1 Suppose (X_1, X_2, X_3) has joint density

$$f_{(X_1,\ldots,X_n)} \propto (x_1 + x_2)(x_1 + x_3)(x_2 + x_3)\mathbb{1}((x_1, x_2, x_3) \in [0, 1]^3).$$

a) Find the normalized density.

b) Find the marginal density of X_1.

c) Find the expected value of X_1.

Solution

a) First

$$\int_{x_1 \in [0,1]} \int_{x_2 \in [0,1]} \int_{x_3 \in [0,1} (x_1 + x_2)(x_1 + x_3)(x_2 + x_3)\, dx_3\, dx_2\, dx_1$$

has value $5/4$, so the normalized density is

$$\boxed{\frac{4}{5}(x_1 + x_2)(x_1 + x_3)(x_2 + x_3)\mathbb{1}(x_1, x_2, x_3 \in [0, 1])}.$$

b) To get the marginal density we need to integrate out the x_2 and x_3 variables which gives:

$$\int_{x_2 \in [0,1]} \int_{x_3 \in [0,1} (4/5)(x_1 + x_2)(x_1 + x_3)(x_2 + x_3)\mathbb{1}(x_1 \in [0, 1])\, dx_3\, dx_2,$$

which evaluates to

$$\boxed{f_{X_1}(x_1) = \frac{2}{15}(6x_1^2 + 7x_1 + 2)\mathbb{1}(x_1 \in [0, 1])}$$

c) We can either go back to the beginning and integrate with the density multiplied by x_1, so just the marginal density times x_1 to get

$$\int_{x_1=0}^{1} x_1 \frac{2}{15}(6x_1^2 + 7x_1 + 2)\, dx_1 = \frac{29}{45} = \boxed{0.6444\ldots}$$

23.3 Suppose $X_1, \ldots, X_n$ have joint density

$$(1/10)^n \mathbb{1}((x_1, \ldots, x_n) \in [0, 10]^n).$$

Show that the X_i are independent.

Solution Since

$$(1/10)^n \mathbb{1}((x_1, \ldots, x_n) \in [0, 10]^n) = \prod_{i=1}^{n}(1/10)\mathbb{1}(x_i \in [0, 10]),$$

the joint density is the product of n marginal densities (each of which is the density of a uniform over $[0, 10]$), and the random variables are independent.

23.5 Suppose that X and Y have joint density:

$$f_{(X,Y)}(x, y) = (3/4)xy^2 \mathbb{1}(x \in [0, 1])\mathbb{1}(y \in [0, 2])$$

Show that X and Y are independent.

Solution Note that the density factors into a part that just depends on x, and a part that just depends on y:

$$f_{(X,Y)}(x, y) = [C_x x \mathbb{1}(x \in [0, 1))][C_y y^2 \mathbb{1}(y \in [0, 2])].$$

Hence X and Y are independent.

24.1 Suppose $A \sim \mathsf{Unif}\{1, 2, 3, 4, 5, 6\}$ and $[B|A] \sim \mathsf{Exp}(A)$. Given $B = 3.6$, what is the distribution of A with the normalizing constant given to four sig figs?

Solution First let's find the posterior density of A given $B = b$, where $b \geq 0$:

$$\begin{aligned}
f_{A|B=b}(a) &\propto f_A(a) f_{B|A=a}(b) \\
&= (1/6)\mathbb{1}(a \in \{1, \ldots, 6\})a \exp(-ab). \\
&\propto a \exp(-ab)\mathbb{1}(a \in \{1, \ldots, 6\}).
\end{aligned}$$

The normalizing constant is

$$C = \sum_{a=1}^{6} a \exp(-ab) = \exp(-b) - 2\exp(-2b) - \cdots - 6\exp(-6b).$$

For $b = 3.6$, this is $C = 0.028804\ldots$. Hence

$$\boxed{f_{A|B=3.6}(a) = 34.62a \exp(-3.6a)\mathbb{1}(a \in \{1, \ldots, 6\})}.$$

24.3 Suppose $X_1 \sim \mathsf{Unif}([0, 10])$ and $X_2 \sim \mathsf{Unif}([0, 20])$. Let $B \sim \mathsf{Unif}\{1, 2\}$.

a) Given $X_B = 15$, what is the chance that $B = 2$?

b) Given $X_B = 7$, what is the chance that $B = 2$?

Solution

a) Since $\mathbb{P}(X_1 > 10) = 0$, given $X_B = 15$, the chance that $B = 2$ is $\boxed{1}$.

b) Given $X_B = 7$, it is possible that B could be either 1 or 2. To Bayes' Rule!

We want to find the density of B given the value of X_B. So we will need the prior density of B and the conditional density of $X_B | B$. These are

$$f_B(b) = \frac{1}{2}\mathbb{1}(b \in \{1, 2\})$$

$$f_{X_B|B=1}(x) = (1/10)\mathbb{1}(x \in [0, 10])$$

$$f_{X_B|B=2}(x) = (1/20)\mathbb{1}(x \in [0, 20]).$$

We could also write this as

$$f_{X_B|B=b}(x) = 1/(a(b))\mathbb{1}(x \in [0, a(b)]),$$

where $a(1) = 10$ and $a(2) = 20$. Then Bayes' Rule says

$$f_{B|X_B=7}(b) \propto \frac{1}{a(b)}\mathbb{1}(x \in [0, a(b)])\frac{1}{2}\mathbb{1}(b \in \{1, 2\})$$

$$\propto \frac{1}{a(b)}\mathbb{1}(b \in \{1, 2\}).$$

Then to find the constant of proportionality, we just integrate the right hand side with respect to counting measure, which gives us the sum

$$C = \frac{1}{10} + \frac{1}{20}.$$

Hence

$$f_{B|X_b=7}(b) = \frac{1}{a(b)}\left[\frac{1}{10} + \frac{1}{20}\right]^{-1},$$

and

$$\mathbb{P}(B = 2|X_b = 7) = f_{B|X_b=7}(2) = \frac{1}{20}\left[\frac{1}{10} + \frac{1}{20}\right] = \frac{1}{3} = \boxed{0.3333\dots}.$$

24.5 A drug company believes that a new treatment is effective on patients with probability p, where p is uniform over $[0, 1]$. A drug trial keeps trying the drug on patients until it finds four patients where the drug is effective. The study needed to enroll $N = 21$ patients before they found four that the drug worked on.

Given this information, what is the new distribution of p?

Solution The posterior density is proportional to the prior density on p times the likelihood. Here $N|p = s$ is negative binomial with parameters 4 and s. Hence

$$f_p(s) = \mathbb{1}(s \in [0, 1])$$

$$f_{N|p=s}(i) = \binom{i-1}{3}s^4(1-s)^{i-4}\mathbb{1}(i \in \{4, 5, \dots\})$$

so the posterior is

$$f_{p|N=21}(s) \propto \binom{21}{3} s^4 (1-s)^{17} \mathbb{1}\,(s \in [0,1])$$

This is the distribution of a beta random variable with parameters 5 and 18. Hence

$$\boxed{[p|N=21] \sim \text{Beta}(5,18)}.$$

25.1 Suppose X is a random variable with mean 0.4, mean absolute deviation of 1.5, and standard deviation of 2.

   a) Give an upper bound on $\mathbb{P}(|X - 0.4| > 4)$ using Markov's inequality.

   b) Give an upper bound on $\mathbb{P}(|X - 0.4| > 4)$ using Chebyshev's inequality.

   c) Which is better? (Or equivalently, if you were asked to give the best upper bound on $\mathbb{P}(|X - 0.4| > 4)$, what would you report?)

Solution

   a) By Markov's inequality

$$\mathbb{P}(|X - 0.4| > 4) \leq \frac{\mathbb{E}[|X - 0.4|]}{4} = \frac{1.5}{4} = \boxed{0.3750}$$

   b) By Chebyshev's inequality

$$\mathbb{P}(|X - 0.4| > 4) \leq \frac{\mathbb{V}(X)}{4^2} = \frac{2^2}{4^2} = \boxed{0.2500}$$

   c) $\boxed{0.2500}$ is the better upper bound since is the smaller of the two values.

25.3 A construction project will take an unknown amount of time. The builders believe that the mean will be fifty days with a standard deviation of ten days.

   a) Give an upper bound for the chance the project takes at least sixty days.

   b) Give an upper bound for the chance the project takes at least one hundred days.

Solution Let T be the time needed for the construction project. Then T is nonnegative and the standard deviation is finite, so both Markov and Chebyshev apply here.

   a)

$$\mathbb{P}(T \geq 60) \leq 50/60 \text{ by Markov}$$
$$\mathbb{P}(T \geq 60) \leq 1/1^2 \text{ by Chebyshev}$$

So $\boxed{0.8333}$ is the best upper bound.

b)

$$\mathbb{P}(T \geq 100) \leq 50/100 \text{ by Markov}$$
$$\mathbb{P}(T \geq 100) \leq 1/5^2 \text{ by Chebyshev}$$

So $\boxed{0.04000}$ is the best upper bound.

25.5 Outreach Solutions serves a number of clients each day that is uniform over $\{1, 2, 3, 4, 5\}$. Let N be the total number of clients served in a week of seven days.

a) What is the expected value of N?

b) What is the standard deviation of N?

c) Using the fact that N is symmetric about its mean, give a lower bound on the probability that $N \leq 26$.

Solution

a) Let N_i be the number of clients served on day i. Then

$$N = N_1 + \cdots + N_7.$$

So

$$\mathbb{E}[N] = \mathbb{E}[N_1 + \cdots + N_7] = \mathbb{E}[N_1] + \cdots + \mathbb{E}[N_7] = 7\mathbb{E}[N_1]$$
$$= 7(1+5)/2 = \boxed{21}.$$

b) Similarly, since the N_i can be assumed to be iid:

$$\mathbb{V}[N] = \mathbb{V}[N_1 + \cdots + N_7] = \mathbb{V}[N_1] + \cdots + \mathbb{V}[N_7] = 7\mathbb{V}[N_1] = 7(5^2 - 1)/12 = 14.$$

Hence $\text{SD}(N) = \sqrt{14} \approx \boxed{3.741}$.

c) We need to find α such that $\mathbb{P}(N \leq 26) \geq \alpha$. But all we have are upper bounds? So apply the function $1 - x$ to both sides:

$$1 - \alpha \leq 1 - \mathbb{P}(N \leq 26) = \mathbb{P}(N \geq 27).$$

Using Markov's inequality gives

$$\mathbb{P}(N \geq 27) \leq \mathbb{E}[N]/27 = (7)(3)/27 = 21/27.$$

Using Chebyshev's inequality (and taking advantage of symmetry) gives:

$$\mathbb{P}(N \geq 27) = \mathbb{P}(N - 21 \geq 6)$$
$$= (1/2)\mathbb{P}(|N - 21| \geq 6)$$
$$\leq (1/2)\mathbb{V}(N)/6^2 = 7\mathbb{V}(N_1)/72.$$

Since $N_1 \sim \text{Unif}(\{1, 2, 3, 4, 5\})$, we have $\mathbb{V}(N_1) = (5^2 - 1)/12 = 2$.
So $1 - \alpha = 7/36$, which means $\alpha = 29/36 > \boxed{0.8055}$.

25.7 Suppose X has finite mean μ and standard deviation σ. All random variables have at least one median. Show that there must be a median of X somewhere strictly between $\mu - \sqrt{3}\sigma$ and $\mu + \sqrt{3}\sigma$.

Solution

Proof. Since X has a finite mean and variance, Chebyshev's inequality applies and

$$\mathbb{P}(|X - \mu| \geq \sqrt{3}\sigma) \leq \frac{1}{3}.$$

In particular, let $m \geq \mu + \sqrt{3}\sigma$. Then

$$\begin{aligned}
\mathbb{P}(X > m) &\leq \mathbb{P}(X \geq \mu + \sqrt{3}\sigma) \\
&= \mathbb{P}(X - \mu \geq \sqrt{3}\sigma) \\
&\leq \mathbb{P}(|X - \mu| \geq \sqrt{3}\sigma) \\
&\leq 1/3 \text{ by Chebyshev's inequality}
\end{aligned}$$

So if $m \geq \mu + \sqrt{3}\sigma$, then m is not a median.

Now suppose $m \leq \mu - \sqrt{3}\sigma$. Then

$$\begin{aligned}
\mathbb{P}(X \leq m) &\leq \mathbb{P}(X \leq \mu - \sqrt{3}\sigma) \\
&= \mathbb{P}(X - \mu \leq -\sqrt{3}\sigma) \\
&\leq \mathbb{P}(|X - \mu| \geq \sqrt{3}\sigma) \\
&\leq 1/3 \text{ by Chebyshev's inequality}
\end{aligned}$$

So again m cannot be a median!

If there is no median in $(-\infty, \mu - \sqrt{3}\sigma]$, and no median in $[\mu + \sqrt{3}\sigma, \infty)$, then there must be a median in $(\mu - \sqrt{3}\sigma, \mu + \sqrt{3}\sigma)$, and the proof is complete. $\square$

25.9 A construction project time T has the following mean, standard deviation, mean absolute deviation, and moment generating function at 0.5:

$$\begin{aligned}
\mathbb{E}[T] &= 100 \\
\sqrt{\mathbb{E}[(T - \mathbb{E}[T])^2]} &= 15 \\
\mathbb{E}(|T - \mathbb{E}[T]|) &= 12 \\
\mathbb{E}(\exp(0.5T)) &= \exp(63).
\end{aligned}$$

Using these facts together with Markov and Chebyshev, put as best an upper bound as you can on $\mathbb{P}(T > 130)$. Be sure to show all your work!

Solution Note $T \geq 0$ because it is measuring time. Markov's inequality gives

$$\mathbb{P}(T > 130) \leq \mathbb{E}[T]/130 = 100/130 = 0.7692\ldots.$$

Not very good! Knowing the standard deviation allows us to use Chebyshev:

$$\mathbb{P}(T > 130) = \mathbb{P}(T - 100 > 30) \leq \mathbb{P}(|T - 100| > 30) \leq \frac{15^2}{30^2} = \frac{1}{2^2} = 0.25.$$

Next up is the mean absolute deviation, which Markov tells us gives

$$\mathbb{P}(T > 130) = \mathbb{P}(T-100 > 30) \le \mathbb{P}(|T-100| > 30) \le \frac{\mathbb{E}[|T - 100|]}{30} = \frac{12}{30} = \frac{2}{5} = 0.4.$$

Finally,

$$\mathbb{P}(T > 130) = \mathbb{P}(\exp(0.5T) > \exp(0.5 \cdot 130)) \le \frac{\exp(63)}{\exp(65)} = \exp(-2) = 0.1353\ldots.$$

So our best upper bound is $\boxed{0.1353}$.

26.1 Suppose that X has moment generating function $\mathrm{mgf}_X(t) = [(\exp(t) - 1)/t]^{10}$. Bound $\mathbb{P}(X \ge 8)$ with Chernoff using $t = 5$.

Solution Chernoff says that

$$\mathbb{P}(X \ge 8) \le \mathrm{mgf}_X(t) \exp(-8t)$$

for all $t > 0$. In particular, for the moment generating function for this problem and this value of t,

$$\mathbb{P}(X \ge 8) \le [(e^5 - 1)/5]^{10} \exp(-8 \cdot 5) \approx \boxed{0.002108\ldots}$$

26.3 Use Chernoff's inequality to give the best upper bound you can on the probability that the sum of 12 iid random variables uniform over $[0, 1]$ is at least 9.

Solution A uniform over $[0, 1]$ has moment generating function when $t > 0$

$$\begin{aligned}
\mathbb{E}[\exp(tU)] &= \int_{-\infty}^{\infty} \exp(tu)\mathbb{1}(u \in [0, 1])\, du \\
&= \int_0^1 \exp(tu)\, du \\
&= \left.\frac{\exp(tu)}{t}\right|_0^1 \\
&= \frac{\exp(t) - 1}{t}.
\end{aligned}$$

Therefore the sum of 12 iid uniforms over $[0, 1]$ has moment generating function $[(e^t - 1)/t]^{12}$.

Putting that into Chernoff's bound gives

$$\begin{aligned}
\mathbb{P}(U_1 + \cdots + U_{12} \ge 9) &\le \mathrm{mgf}_{U_1 + \cdots + U_{12}}(t) \exp(-9t) \\
&= \left[\frac{e^t - 1}{t}\right]^{10} \exp(-0.75t)^{12} \\
&= \left[\frac{e^{0.25t} - e^{-0.75t}}{t}\right]^{12}
\end{aligned}$$

Numerically minimizing the term inside the brackets gives $0.664554\ldots$, which makes the best upper limit $\boxed{0.007419\ldots}$.

27.1 For $X \sim$ Cauchy, find
$$\mathbb{P}(X \in [0,5]).$$

Solution The density of a Cauchy is $(2/\tau)(1 + s^2)^{-1}$, so this is

$$\int_0^5 \frac{2}{\tau} \cdot 11 + s^2 \, ds = \frac{2}{\tau}[\arctan(5) - \arctan(0)] \approx \boxed{0.4371}.$$

27.3 Estimate $\zeta(2.5)$ to four significant figures.

Solution Note

$$\zeta(2.5) = \sum_{i=1}^{n} \frac{1}{i^{2.5}} + \sum_{i=n+1}^{\infty} \frac{1}{i^{2.5}}.$$

The key is figuring out how large n has to be to make our result accurate to four sig figs.
We can upper bound the sum using an integral.

$$\sum_{i=n+1}^{\infty} \frac{1}{i^{2.5}} = \int_{x=n+1}^{\infty} \lfloor x \rfloor^{-2.5} \, dx$$

$$\leq \int_{x=n+1}^{\infty} (x-1)^{-2.5} \, dx$$

$$= (2/3)n^{-1.5}.$$

By setting $n = 1645$, we make this sum at most 0.00001, and so

$$\zeta(2.5) \in \left[\sum_{i=1}^{1645} \frac{1}{i^{2.5}}, \sum_{i=1}^{1645} \frac{1}{i^{2.5}} + 0.00001 \right] = [1.34147, 1.34179],$$

so to 4 sig figs, $\zeta(2.5) \approx \boxed{1.341}$.

27.5 For $X \sim$ Zeta(α) with $\alpha > 1$, prove that $\ln(X)$ always has finite mean.

Solution The sum for the mean of $\ln(X)$ is

$$\mathbb{E}[\ln(X)] = \sum_{i=1}^{\infty} \ln(i) \frac{1}{i^{\alpha}}$$

$$= \sum_{i=2}^{\infty} \ln(i)/i^{\alpha}$$

since $\ln(1) = 0$.

Note that for $x \in [1, 2)$, $\ln(1 + x) \geq \ln(2)$ and $1/x^{\alpha} \geq 1/2^{\alpha}$. Similarly, for $x \in [2, 3)$, $\ln(1 + x) \geq \ln(3)$ and $1/x^{\alpha} \geq 1/3^{\alpha}$. Using this idea for all width 1 intervals gives

$$\mathbb{E}[\ln(X)] = \zeta(\alpha)^{-1} \sum_{i=2}^{\infty} \ln(i)/i^{\alpha}$$

$$\leq \zeta(\alpha)^{-1} \int_1^{\infty} \ln(1 + x)/x^{\alpha} \, dx$$

This is a tricky integral to solve. It can be made easier by noting that for $x \geq 1$,

$$\ln(1+x) - \ln(x) = \ln(1+1/x) \leq 1 + 1/x,$$

which means

$$\ln(1+x) \leq \ln(x) + 2$$

for $x \geq 1$. So

$$\mathbb{E}[\ln(X)] = \zeta(\alpha)^{-1} \sum_{i=2}^{\infty} \ln(i)/i^\alpha$$

$$\leq \zeta(\alpha)^{-1} \int_1^\infty [2 + \ln(x)]/x^\alpha \, dx$$

$$\leq 2 + \zeta(\alpha)^{-1} \int_1^\infty \ln(x)/x^\alpha \, dx.$$

To solve the integral now, integration by parts can be used

$$f(x) = \ln(x), \ g'(x) = x^{-\alpha}, \ f'(x) = 1/x, \ g(x) = -x^{-\alpha+1}/(\alpha-1),$$

so

$$\int_1^\infty \ln(x)/x^\alpha \, dx = -\ln(x)x^{-\alpha+1}/(\alpha-1)|_1^\infty + \int_1^\infty 1/x^\alpha \, dx = \zeta(\alpha).$$

Hence

$$\mathbb{E}[\ln(X)] \leq 2 + 1/(\alpha-1).$$

which is finite no matter the value of $\alpha > 1$!

28.1 A small plastic bucket contains tiles with the letters MISSISSIPPI. Four of these tiles are drawn out of the bucket without replacement.

   a) What is the chance that all four S tiles are drawn?

   b) What is the chance that exactly two out of the four drawn tiles are S?

Solution

   a) Let A_i be the event that the ith tile is an S. Then we want $\mathbb{P}(A_1A_2A_3A_4)$. Note that

$$\mathbb{P}(A_1A_2A_3A_4) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1A_2)\mathbb{P}(A_4|A_1A_2A_3)$$

$$= \frac{4}{11} \cdot \frac{3}{10} \cdot \frac{2}{9} \cdot \frac{1}{8}$$

$$\approx \boxed{0.003030}.$$

28.3 In the prodction output of a factory for a day are 500 screws, 10 of which are below standard.

   a) If 5 of the screws are drawn uniformly at random without replacement from the 500 screws, what is the chance that this sample does not contain a below standard screw?

   b) What is the expected number of below standard screws in this sample of 5?

Solution

a) This will be

$$\frac{490}{500} \cdot \frac{489}{499} \cdot \frac{488}{498} \cdot \frac{487}{497} \cdot \frac{486}{496} = \boxed{0.9035\dots}.$$

b) This would be $(10/500)(5) = \boxed{0.1000}$. Note that Markov's inequality means that for Y the number sampled that are below standard,

$$\mathbb{P}(Y \geq 1) \leq \mathbb{E}[Y]/1 = 0.1,$$

so $\mathbb{P}(Y = 0) \geq 1 - 0.1 = 0.9$ which is a fair approximation to the true answer.

28.5 Fifteen hundred of a particular bird are living in an ecosystem. Twenty are sampled without replacement, and five are found to be carrying a particular gene. What number of birds out of the 1500 should have this gene so that the expected number found in the sample would be 5? (This type of estimate is called a *method of moments estimate*.)

Solution If n birds carried the gene, the expected number found would be

$$\frac{n}{1500} \cdot 20.$$

Setting this equal to 5 and solving gives

$$\hat{n} = 5(1500)/20 = \boxed{375}.$$

29.1 Suppose $(W_1, W_2, W_3, W_4) \sim \mathsf{Multinom}(10, 0.3, 0.2, 0.4, 0.1)$. What is

$$\mathbb{P}((W_1, W_2, W_3, W_4) = (1, 3, 2, 4))?$$

Solution This is like when there are four answers to a question, and ten individuals are surveyed. From our formula:

$$
\begin{aligned}
\mathbb{P}((W_1, W_2, W_3, W_4) = (1, 3, 2, 4)) &= \binom{10}{1, 3, 2, 4} (0.3)^1 (0.2)^3 (0.4)^2 (0.1)^4 \\
&= \frac{10!}{1!3!2!4!} (0.3)^1 (0.2)^3 (0.4)^2 (0.1)^4 \\
&= \boxed{0.0004838\dots}.
\end{aligned}
$$

29.3 Suppose $(X_1, X_2, X_3) \sim \mathsf{Multinom}(30, 0.5, 0.1, 0.4)$.

a) What is the distribution of X_1?

b) What is $\mathbb{E}[X_1]$?

c) Find $\mathrm{Cov}(X_1, X_3)$.

Solution

a) The marginals of a multinomial are binomial, so $\boxed{X_1 \sim \mathsf{Bin}(30, 0.5)}$.

b) The mean of a binomial is the number of trials times the probability of success on each trial, so $30 \cdot 0.5 = \boxed{15}$.

c) The covariance between margins of a multinomial is the negative of the number of trials times the probabilities for the marginals involved. In this case, $30 \cdot 0.5 \cdot 0.4 = \boxed{6}$.

29.5 A research project is studying an animal population. There are 30 animals collected. Each animal has a 10% chance of being genotype A, 20% of being genotype B, and 70% of being genotype C. Let (N_A, N_B, N_C) be the number of animals found of each genotype.

a) What is the distribution of (N_A, N_B, N_C)?

b) What is $\mathbb{E}(N_A)$?

c) What is $\mathrm{Cov}(N_A, N_B)$?

d) What is $\mathrm{Cov}(N_A, N_C)$?

Solution

a) The numbers of each type will be multinomial. In this project:

$$\boxed{\mathsf{Multinom}(30, 0.1, 0.2, 0.7)}.$$

b) The mean is the probability of picking on of that type times the number of samples drawn. So
$$\mathbb{E}(N_A) = 0.1(30) = \boxed{3}.$$

c) The covariance is the negative of the product of the probabilities times the number being drawn. So
$$\mathrm{Cov}(N_A, N_B) = -30 \cdot 0.1 \cdot 0.2 = \boxed{-0.6000}.$$

d) Similarly, this will be
$$\mathrm{Cov}(N_A, N_C) = -30 \cdot 0.1 \cdot 0.7 = \boxed{-2.100}.$$

30.1 Let Z_1, Z_2, Z_3 be iid normal.

What is the distribution of

a) What is the distribution of $Z_1 + Z_2 + Z_3$?

b) What is the distribution of $5 - Z_1 + 2Z_2 + 4Z_3$?

Solution Remember that any linear combination of iid normals will be a normal family distribution. Just match the mean and variance.

a) This will be $\boxed{\mathsf{N}(0, 3)}$.

b) This will be $\boxed{\mathsf{N}(5, 21)}$.

30.3 For Z_1, Z_2, Z_3 iid normal let

$$W_1 = Z_1 + Z_2 - 2Z_3$$
$$W_2 = -Z_1 + Z_3$$
$$W_3 = Z_3.$$

a) Find $\text{Cov}(W_1, W_3)$.

b) What is the distribution of (W_1, W_2, W_3)?

Solution

a) Here

$$\begin{aligned}
\text{Cov}(W_1, W_3) &= \text{Cov}(Z_1 + Z_2 - 2Z_3, Z_3) \\
&= \text{Cov}(2Z_e, Z_3) \\
&= 2\mathbb{V}(Z_3) = \boxed{2}.
\end{aligned}$$

b) Since $W = AZ$, W has a multivariate normal (or multinormal) distribution. To find the parameters, we take

$$AA^T = \begin{pmatrix} 1 & 1 & -2 \\ -1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & 0 \\ 2 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 6 & -3 & -2 \\ -3 & 2 & 1 \\ -2 & 1 & 1 \end{pmatrix}$$

The mean is the 0 vector, so

$$\boxed{\begin{pmatrix} W_1 \\ W_2 \\ W_3 \end{pmatrix} \sim \text{Multinorm}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 6 & -3 & -2 \\ -3 & 2 & 1 \\ -2 & 1 & 1 \end{pmatrix} \right)}$$

30.5 Suppose that (X_1, X_2, X_3) is a multivariate normal with mean $2.3, 1.8, -1.6$ and covariance matrix

$$\begin{pmatrix} 1.1 & 0 & 2.3 \\ 0 & 2.4 & 1.6 \\ 2.3 & 1.6 & 0.7 \end{pmatrix}$$

a) What is the distribution of X_1?

b) What is $\text{Cov}(X_1, X_3)$?

Solution

a) This is $\boxed{\text{N}(2.3, 1.1)}$.

b) This is $\boxed{2.300}$.

31.1 Suppose X_1, X_2, X_3 are iid with density $f(s) = s/2 \cdot \mathbb{1}(s \in [0, 2])$.

a) What is the density of $X_{(1)}$?

b) What is $\mathbb{E}[X_{(2)}]$?

Solution

a) The density of $X_{(1)}$ is (from the order statistic formula):

$$f_{X_{(1)}}(s) = 3\binom{2}{0}f(s)F(s)^0(1 - F(s))^{3-1}.$$

The cdf of X_1 is for $s \in [0, 2]$:

$$F(a) = \int_{-\infty}^{a} f(s)\,ds = \int_{0}^{a} s/2\,ds = a^2/4.$$

When $a > 2$, then $F(a) = 1$, and for $a < 0$, $F(a) = 0$. Hence

$$f_{X_{(1)}}(s) = 3(s/2) \cdot \mathbb{1}(s \in [0, 2])(1 - s^2/4)^2 = \boxed{(3/2)s(1 - s^2/4)^2 \cdot \mathbb{1}(s \in [0, 2])}.$$

b) To find $\mathbb{E}[X_{(2)}]$, we need the density. Again using the order statistics density formula:

$$f_{X(2)}(s) = 3\binom{2}{1}f(s)F(s)^1(1 - F(s))^{3-2} = 3s(s^2/4)(1 - s^2/4) \cdot \mathbb{1}(s \in [0, 2]).$$

This makes the expected value

$$\begin{aligned}
\mathbb{E}[X_{(2)}] &= \int_{\mathbb{R}} s f_{X(2)}(s)\,ds \\
&= \int_{\mathbb{R}} s \cdot (3s^3/4)(1 - s^2/4)\mathbb{1}(s \in [0, 2])\,ds \\
&= \int_{0}^{2} (3s^4/4)(1 - s^2/2)\,ds \\
&= 48/35 \approx \boxed{1.371}.
\end{aligned}$$

31.3 Suppose $\mathbb{P}(X = 0) = 0.3$, $\mathbb{P}(X = 1) = 0.5$, and $\mathbb{P}(X = 2) = 0.2$. Suppose that X_1, X_2, X_3 are iid with the same distribution as X.

   a) What is the distribution of $X_{(1)}$?
   b) What is $\mathbb{E}[X_{(2)}]$?

Solution

   a) Since X is not continuous, we cannot use the order statistic density formula. What we can do is use our minimum idea from earlier in the course:

$$\begin{aligned}
\mathbb{P}(X_{(1)} \geq 0) &= 1^3 \\
\mathbb{P}(X_{(1)} \geq 1) &= 0.7^3 \\
\mathbb{P}(X_{(1)} \geq 2) &= 0.2^3,
\end{aligned}$$

which leads to

$$\boxed{\mathbb{P}(X_{(1)} = 0) = 0.6570,\ \mathbb{P}(X_{(1)} = 1) = 0.3350,\ \mathbb{P}(X_{(1)} = 2) = 0.008000.}$$

b) To find the expected value of $X_{(2)}$, it is necessary to find its distribution. This number is 0 if at least two of the X_1, X_2, X_3 are 0, and 2 if at least two of the X_1, X_2, X_3 are two. The fact that it is either 0, 1, or 2 then allows calculation of the final part of the distribution:

$$\mathbb{P}(X_{(2)} = 0) = \binom{3}{2} 0.3^2 0.7 + 0.3^3 = 0.216$$

$$\mathbb{P}(X_{(2)} = 3) = \binom{3}{2} s 0.2^2 0.8 + 0.2^3 = 0.104$$

Hence
$$\mathbb{E}[X_{(2)}] = 0.216(0) + 0.68(1) + 0.104(2) = \boxed{0.8880}.$$

31.5 What is the chance that for three iid uniforms over $[0, 1]$, that the middle of the three numbers falls in the interval $[1/3, 2/3]$?

Solution Since the random variables are uniform, the middle of the three numbers (aka the second order statistic) will have a beta distribution with parameters 1 and 1:

$$f_{U_{(2)}}(s) = \frac{s(1-s)}{B(2,2)} \cdot \mathbb{1}(s \in [0,1])$$

$$= s(1-s)\frac{\Gamma(4)}{\Gamma(1)\Gamma(1)}\mathbb{1}(s \in [0,1])$$

$$= 3! s(1-s) \cdot \mathbb{1}(s \in [0,1]).$$

So to find the probability this falls in $[1/3, 2/3]$, just integrate over this interval:

$$\mathbb{P}(U_{(2)} \in [1/3, 2/3]) = \int_{[1/3,2/3]} \mathbb{P}(U_{(2)} \in du) = \int_{1/3}^{2/3} 6u(1-u)\, du = \frac{13}{27} \approx \boxed{0.4814}.$$

32.1 Suppose $\Omega_1 = [0, 1]$ and $\Omega_2 = \{1, 2, 3\}$. Say The function $Y : \Omega_1 \to \Omega_2$ is defined as

$$Y(y) = 1 + \mathbb{1}(y \in [0.4, 0.6]) + \mathbb{1}(y \in [0.5, 0.7]).$$

What is $Y^{-1}(\{2\})$?

Solution The question is asking what inputs y to the Y function result in an output of 2. The way the function is constructed using indicators, this happens when exactly one of the statements $y \in [0.4, 0.6]$ and $y \in [0.5, 0.7]$ is true. This is exactly

$$\boxed{y \in [0.4, 0.5) \cup (0.6, 0.7]}.$$

32.3 Suppose $\Omega_1 = [0, 1]$, $\Omega_2 = \{1, 2, 3\}$, and

$$\mathcal{F}_2 = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}.$$

For $Y : \Omega_1 \to \Omega_2$, let

$$Y(y) = 1 + \mathbb{1}(y \in [0.4, 0.6]) + \mathbb{1}(y \in [0.5, 0.7]).$$

Then what sets must $\mathcal{F}_1$ (a sigma algebra over Ω_1 contain in order for Y to be a measurable function?

Solution Look at the inverse under Y of each of the sets in $\mathcal{F}_2$.

$$
\begin{aligned}
Y^{-1}(\emptyset) &= S_0 = \emptyset \\
Y^{-1}(\{1\}) &= S_1 = [0, 0.4) \cup (0.7, 1] \\
Y^{-1}(\{2\}) &= S_2 = [0.4, 0.5) \cup (0.6, 0.7] \\
Y^{-1}(\{3\}) &= S_3 = [0.5, 0.6] \\
Y^{-1}(\{1, 2\}) &= S_4 = [0, 0.5) \cup (0.6, 1] \\
Y^{-1}(\{1, 3\}) &= S_5 = [0, 0.4) \cup [0.5, 0.6] \cup (0.7.1] \\
Y^{-1}(\{2, 3\}) &= S_6 = [0.4, 0.7] \\
Y^{-1}(\{1, 2, 3\}) &= S_7 = [0, 1]
\end{aligned}
$$

So the measure sets need to include all eight of these possibilities.

$$
\mathcal{F}_1 = \boxed{\{S_0, \ldots, S_7\}}.
$$

32.5 Continuing with Y from earlier, if $\mathbb{P}_1$ over Ω_1 is uniform, what is $\mathbb{P}(Y \leq 2)$?

Solution

$$
\begin{aligned}
\mathbb{P}(Y \leq 2) &= \mathbb{P}_1(Y^{-1}(\{1, 2\})) \\
&= \mathbb{P}_1([0, 0.5) \cup (0.6, 1]) \\
&= \boxed{0.9000}.
\end{aligned}
$$

42.1 Prove that $(\exists x)(2x + 3 \geq 10)$.

Solution

Proof. Let $x = 4$. Then $2x + 3 = 11 \geq 10$ $\Box$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\Box$

42.3 Prove that $(\forall x)(\exists y)(xy \leq 0)$

Solution

Proof. Let $x \in \mathbb{R}$. Let $y = -|x|$. Suppose $x \geq 0$, then $y \leq 0$ so $xy \leq 0$. Suppose $x \leq 0$, then $y \geq 0$, so $xy \leq 0$. Either way $xy \leq 0$. $\Box$. $\qquad\qquad\qquad\qquad$ $\Box$

42.5 Prove that if $x > 3$ then $2x > 6$.

Solution

Proof. Let $x > 3$. Then multiplying by 2 gives $2x > 6$. $\Box$ $\qquad\qquad\qquad\qquad\qquad\qquad$ $\Box$

43.1 What is the counting measure of $\{1, 2, \ldots, 10\}$?

Solution Since there are ten elements in the set, $\boxed{10}$.

43.3   a) What is $\{r, g, b\} \cap \{g, b, y\}$?

b) What is $\{r, g, b\} \cup \{g, b, y\}$?

Solution

a) The intersection of these sets consists of the elements that are in both sets, so $\boxed{\{g, b\}}$.

b) The union of these sets consists of the elements that are in either one or two of these sets, so $\boxed{\{r, g, b, y\}}$.

43.5 What is the counting measure of $\{r, g, b\}$?

Solution Since there are three elements of the set, $\boxed{\#(\{r, g, b\}) = 3}$.

43.7 What is the counting measure of $\{1, 3, 5\} \times \{7, 9\}$?

Solution The counting measure is the product of the individual measures, so $3 \cdot 2 = \boxed{6}$.

43.9 Let $A = \{r, g, b\}$. What is the counting measure of $A \times A \times A \times A$?

Solution Using the multiplicative property of counting meausure for direct product, this is
$$3 \cdot 3 \cdot 3 \cdot 3 = \boxed{81}.$$

43.11   a) What is the Lebesgue measure of $[2, 10]$?

b) What is the Lebesgue measure of $[-6, 2]$?

Solution

a) This is the length of the interval $10 - 2 = \boxed{8}$.

b) This is the length of the interval $2 - (-6) = \boxed{8}$.

43.13 What is the Lebesgue measure of $[3, 4.5] \times [0, 6]$?

Solution Since the first set has Lebesgue measure $4.5 - 3 = 1.5$ and the second has Lebegue measure (length) of $6 - 0 = 6$, the Cartesian product has Lebesgue measure (area) of $(6)(1.5) = \boxed{9}$.

43.15 *De Morgan's Laws* say that
$$(A \cup B)^C = A^C \cap B^C$$
$$(A \cap B)^C = A^C \cup B^C.$$

Assume this law hold for two sets, and then prove that
$$(A \cup B \cup C)^C = A^C \cap B^C \cap C^C.$$

Solution Let $R = A \cup B$. Then
$$(A \cup B \cup C)^C = (R \cup C)^C = R^C \cap C^C$$

by the two set De Morgan's law. Also by the two set De Morgan's law, $R^C = A^C \cap B^C$. Putting this together gives

$$(A \cup B \cup C)^C = A^C \cap B^C \cap C^C$$

as desired.

44.1 Given a function $f : \{a, b, c\} \to \{b, c, d\}$, answer the following.

a) What are the possible inputs to f?

b) What are the possible outputs from f?

Solution

a) Possible inputs are $\boxed{\{a, b, c\}}$.

b) Possible outputs are $\boxed{\{b, c, d\}}$.

44.3 Suppose $g(a) = b$, $g(b) = b$, and $g(c) = d$. Is g one-to-one?

Solution It is $\boxed{\text{not one-to-one}}$ since there are two inputs that lead to the same output.

44.5 Consider the function $f(x) = x^2$.

a) Say $f : [0, 1] \to [0, 1]$. Is f onto? Is it 1-1?

b) Say $f : [-1, 1] \to [0, 1]$. Is f onto? Is it 1-1?

c) Say $f : [-1, 1] \to [0, 2]$. Is f onto? Is it 1-1?

Solution

a) Let $y \in [0, 1]$. Then $\sqrt{y} \in [0, 1]$ and $f(\sqrt{y}) = y$, so f is $\boxed{\text{onto}}$. If $a^2 = b^2$ then $|a| = |b|$ since $\sqrt{a^2} = |a|$. So for a and b both in $[0, 1]$, $a = b$, so f is $\boxed{\text{one-to-one}}$.

b) If it was onto before, increasing the domain keeps the function $\boxed{\text{onto}}$. However, now $f(-1/2) = f(1/2) = 1/4$, which shows (as a counterexample) that the function f is $\boxed{\text{not one-to-one}}$.

c) If $f(a) = 2$, then $a^2 = 2$, so $|a| = \sqrt{2}$ and $a \in \{-\sqrt{2}, \sqrt{2}\}$, neither of which is in $[-1, 1]$. Hence f is $\boxed{\text{not onto}}$. The example $f(-1/2) = f(1/2) = 1/4$ still works to show that it is $\boxed{\text{not one-to-one}}$.

45.1 Evaluate the following integrals:

$$\int_0^3 x^3 \, dx, \quad \int_{-\infty}^0 x \exp(x) \, dx, \quad \int_{-\infty}^\infty x \exp(-x^2/2) \, dx.$$

(Note, after you have worked problems like this out, I encourage you to use tools like Wolfram Alpha to *check* your answers. For instance, type

```
integrate x^3 from 0 to 3
```

at the website `www.wolframalpha.com` to check your answer to the first integral.)

Solution The antiderivative of x^3 is $x^4/4$, so using the Fundamental Theorem of Calculus gives

$$\int_0^3 x^3 \, dx = \frac{x^4}{4}\Big|_0^3 = \frac{81}{4} = \boxed{20.25}\,.$$

The next integral requires integration by parts. Recall:

$$\int_a^b f(x)g'(x) \, dx = f(x)g(x)|_a^b - \int_a^b f'(x)g(x) \, dx.$$

Setting $f(x) = x$ and $g'(x) = \exp(x)$ gives $f'(x) = 1$ and $g(x) = \exp(x)$, so

$$\int_{-\infty}^0 x\exp(x) \, dx = x\exp(x)|_{-\infty}^0 - \int_{-\infty}^0 \exp(x) \, dx$$
$$= x\exp(x)|_{-\infty}^0 - \exp(x)|_{-\infty}^0.$$

Recall you cannot just plug in negative infinity, you need to take the limit as the function goes to negative infinity.

Also remember polynomials like x grow much more slowly than exponentials, and $x\exp(-x)$ is just

$$\frac{x}{\exp(x)}.$$

So the numerator is growing polynomially, while the denominator is growing exponentially, so this limit is 0. More generally, the rule is

$$\text{logarithms} \ll \text{polynomials} \ll \text{exponentials} \ll \text{factorials}$$

Another way to tackle problems like this is with L'Hôpital's Rule: suppose that $\lim_{x\to a} f(x) = \lim_{x\to a} g(x)$ and that limit is either 0 or infinity. L'Hôpital's Rule then states that if $\lim_{x\to a} f'(x)/g'(x)$ exists and is finite, then $\lim_{x\to a} f(x)/g(x) = \lim_{x\to a} f'(x)/g'(x)$.

Now $x\exp(x) = x/\exp(-x)$, derivative of x is 1, and derivative of $\exp(-x)$ is $-\exp(-x)$, so

$$\lim_{x\to-\infty} \frac{1}{-\exp(-x)} = 0,$$

which means $\lim_{x\to-\infty} x\exp(x) = 0$ as well. Hence

$$\int_{-\infty}^0 x\exp(x) \, dx = x\exp(x)|_{-\infty}^0 - \exp(x)|_{-\infty}^0 = 0 - 0 - (1 - 0) = \boxed{-1}\,.$$

The third integral requires substitution. Recall that

$$\int_a^b f(g(x))g'(x) \, dx = \int_{g(a)}^{g(b)} g(y) \, dy.$$

In this case, $g(x) = -x^2/2$ so $g'(x) = -x$. Now, that $-x$ doesn't quite match the x in the integral, but just multiply and divide by -1 to get it in the right form:

$$\int_a^b x \exp(-x^2/2)\, dx = \int_a^b \frac{-x \exp(-x^2/2)}{-1}\, dx$$

$$= \int_{-a^2/2}^{-b^2/2} \frac{\exp(y)}{-1}\, dy = - \exp(y)|_{-a^2/2}^{-b^2/2}.$$

Hence

$$\int_{-\infty}^{\infty} x \exp(-x^2/2)]dx = \lim_{a\to-\infty} \lim_{b\to\infty} - \exp(-b^2/2) + \exp(-a^2/2) = \boxed{0}.$$

45.3 Find

$$\int_1^2 x \ln(x)\, dx$$

by moving a derivative from $x = [x^2/2]'$ over to $\ln(x)$ to get rid of it.

Solution Integration by parts gives us

$$\int_1^2 x \ln(x)\, dx = \int_{-1}^1 [x^2/2]' \ln(x)\, dx$$

$$= \int_1^2 -(x^2/2)[\ln(x)]' + [x^2 \ln(x)/2]'\, dx$$

$$= \int_1^2 -(x^2/2)(1/x) + [x^2 \ln(x)/2]'\, dx$$

$$= \int_1^2 -x/2 + [x^2 \ln(x)/2]'\, dx$$

$$= \int_1^2 -[x^2/4]' + [x^2 \ln(x)/2]'\, dx$$

$$= x^2[\ln(x)/2 - (1/4)]_1^2$$

$$= 4[\ln(2)/2 - (1/4)] - (1)[0 - (1/4)] = 2\ln(2) - 3/4 \approx \boxed{0.6362}.$$

We can check our answer by using `integrate x*ln(x) from 1 to 2` in Wolfram Alpha.

45.5 Suppose

$$I = \int_0^\infty \int_0^x |f(x,y)|\, dy\, dx < \infty$$

so Fubini can be applied. Replace the question marks with the appropriate function of y.

$$I = \int_0^\infty \int_?^? f(x,y)\, dx\, dy.$$

Solution In I, x can go anywhere from 0 to ∞, while y runs from 0 up to x. Writing this as an inequality gives

$$0 \le y \le x.$$

In terms of x, this means $x \geq y$. There is no upper bound on x, so replacing the ?s with functions of y gives:

$$I = \boxed{\int_0^\infty \int_x^\infty f(x)\, dx\, dy.}$$

Note that ∞ is a constant function of y.

Index